Phylogenetic Algebraic Geometry and Syntax

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

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

this lecture based on:

- Kevin Shu, Andrew Ortegaray, Robert Berwick, Matilde Marcolli, *Phylogenetics of Indo-European Language families via an Algebro-Geometric Analysis of their Syntactic Structures*, arXiv:1712.01719

Databases of syntactic structures of world languages

1. Syntactic Structures of World Languages (SSWL) `http://sswl.railsplayground.net/`
2. TerraLing `http://www.terraling.com/`
3. World Atlas of Language Structures (WALS) `http://wals.info/`
4. another set of data from Longobardi–Guardiano, Lingua 119 (2009) 1679-1706
5. more complete set of data by Longobardi, Linguistic Analysis, Vol.41 (2017) N.3-4, 517–556.

• First Step: data analysis of syntax of world languages with various mathematical tools (persistent topology, etc.)

• we used the most extensive database currently available: SSWL with 116 "variables" (syntactic "parameters") and 253 world languages (but... some problems with SSWL)

Problems of SSWL data

- Very non-uniformly mapped across the languages of the database: some are 100% mapped, while for some only very few of the 116 parameters are mapped
- Linguists criticize the choice of binary variable (not all of them should count as "true" parameters)

• the data of Longobardi–Guardiano are more reliable, with 62 languages (mostly Indo-European) and 83 parameters

• linguistic question: can languages that are far away in terms of historical linguistics end up being close in terms of syntactic parameters? (homoplasy?)

• Guideline for data use: given what is available at present, use SSWL and Longobardi data (two independent set of syntactic features) keeping limitations in mind and comparing structures of the two datasets

Phylogenetic questions:
syntactic parameters as dynamical variables

Two main types of questions on dynamical behavior of syntax:

1. Reconstruct the past: phylogenetic trees of language families, historical linguistics
2. Predict the future: dynamical models of language change due to language interaction (bilingualism, code switching), dynamical models of language acquisition

• Phylogenetic tree reconstruction: applying methods of phylogenetic algebraic geometry (how to account for the presence of relations?)
• Predictive dynamical models: syntactic parameters as spin-glass models, crucial role of relations in dynamics and equilibrium states

Phylogenetic reconstruction in Linguistics

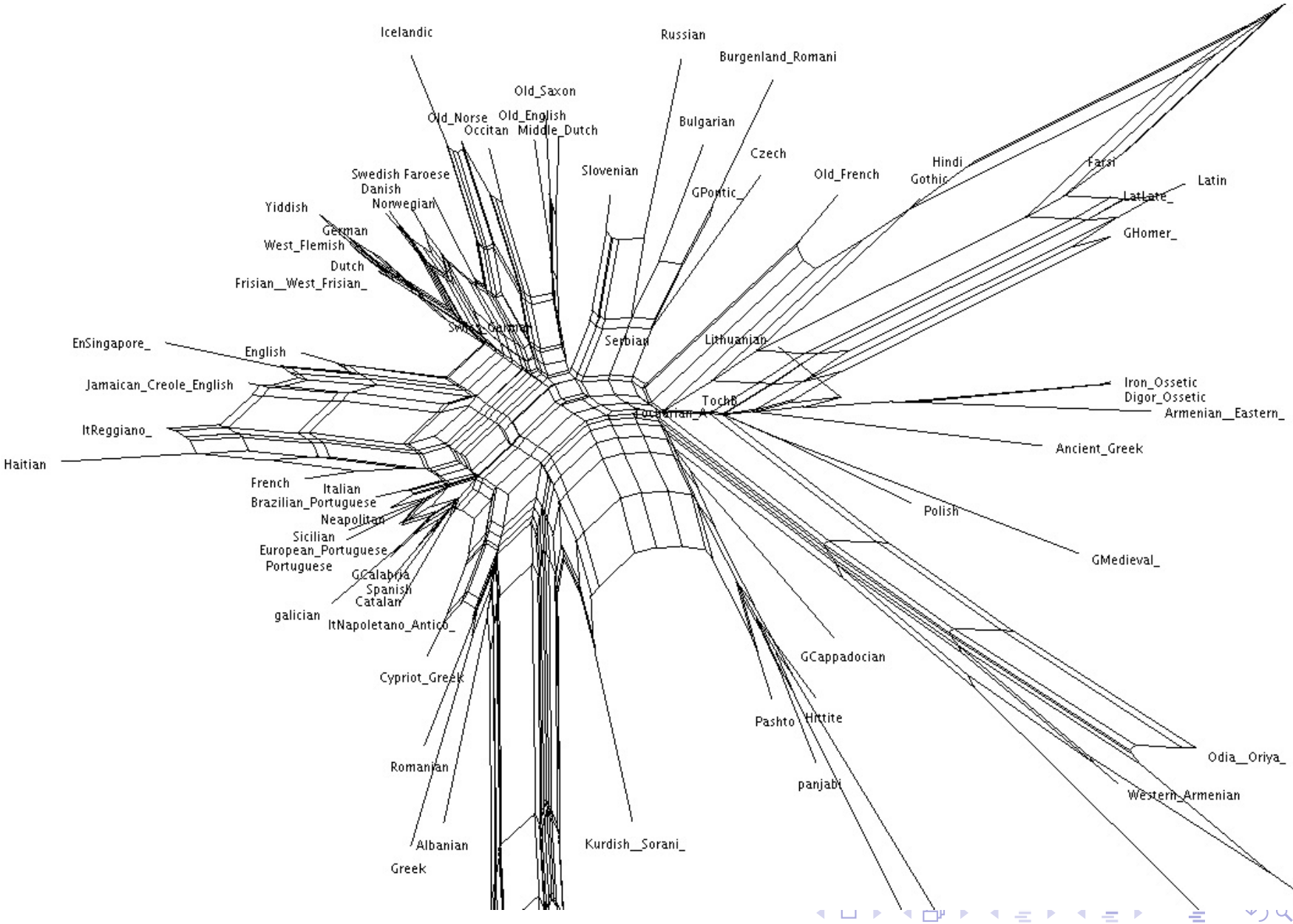
- Can one reconstruct phylogenetic trees computationally using only information on the modern languages?
- Can one reconstruct phylogenetic trees using syntactic parameters data? (Syntax is more stable than lexicon, slower changes, rare borrowing...)
- Long standing open problems: for example the question of the early Indo-European tree
- Linguistics has studied in depth how languages change over time (Philology, Historical Linguistics) usually via lexical and morphological analysis
- Goal: understand the historical relatedness of different languages, subdivisions into families and sub-families, phylogenetic trees of language families

Expect problems: SSWL data and phylogenetic reconstructions

- commonly used methods for computational phylogenetics (Hamming distance, etc.)
- known problems related to the use of Hamming metric for phylogenetic reconstruction
- SSWL problems mentioned above (especially non-uniform mapping)
- dependence among parameters (not independent random variables)
- syntactic proximity of some unrelated languages

- Phylogeny Programs for trees and networks
  - PHYLIP
  - Splittree 4
  - Network 5

Checking on the Indo-European tree where good Historical-Linguistics



Indeed Problems

- misplacement of languages within the correct family subtree
- placement of languages in the wrong subfamily tree
- proximity of languages from unrelated families (all SSWL)
- incorrect position of the ancient languages

- different approach:
  - subdivide into subfamilies (some a priori knowledge from morpholexical linguistic data, or use of topological H_0-method on syntactic data)
  - use syntactic data that are fully mapped over the subfamily
  - then use Phylogenetic Algebraic Geometry (Pachter, Sturmfels et al.) for statistical inference of phylogenetic reconstruction

Phylogenetic Algebraic Geometry: Main References

- L. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005
- L. Pachter, B. Sturmfels, *The Mathematics of Phylogenomics*, SIAM Review, Vol.49 (2007) N.1, 3–31
- B. Sturmfels, S. Sullivant, *Toric Ideals of Phylogenetic Invariants*, Journal of Computational Biology, Vol. 12 (2005) No. 2, 204–228
- N. Eriksson, K. Ranestad, B. Sturmfels, S. Sullivant, *Phylogenetic Algebraic Geometry*, in "Projective varieties with unexpected properties", pp.237–255, Walter de Gruyter, 2005.
- M. Drton, B. Sturmfels, S. Sullivant, *Lectures on Algebraic Statistics*, Birkhäuser, 2009.

Other Phylogenetic Algebraic Geometry References

- E. Allman, J. Rhodes, *Phylogenetic ideals and varieties for general Markov models*, Adv. Appl. Math. Vol.40 (2008) 127–148
- M. Casanellas, J. Fernández–Sánchez, *Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees*, Mol. Biol. Evol. 24 (2007) N.1, 288–293
- J. Draisma, E. Horobeț, G. Ottaviani, B. Sturmfels, R. Thomas, *The Euclidean distance degree of an algebraic variety*, Found. Comput. Math. 16 (2016), no. 1, 99–149
- N. Eriksson, *Using invariants for phylogenetic tree construction*, in "Emerging applications of algebraic geometry", pp. 89–108, IMA Vol. Math. Appl., 149, Springer, 2009

General Idea of Phylogenetic Algebraic Geometry

- Markov process on a binary rooted tree (gen. Jukes-Cantor model)
- probability distribution at the root $(\pi, 1-\pi)$

(frequency of 0/1 for parameters at root vertex) and transition matrices along edges M^e bistochastic

$$M^e = \begin{pmatrix} 1-p_e & p_e \\ p_e & 1-p_e \end{pmatrix}$$

- observed distribution at the n leaves polynomial function

$$p_{i_1,\dots,i_n} = \Phi(\pi, M^e) = \sum_{w_v \in \{0,1\}} \pi_{w_{v_r}} \prod_e M^e_{w_{s(e)}, w_{t(e)}}$$

with sum over "histories" consistent with data at leaves

- polynomial map that assigns

$$\Phi : \mathbb{C}^{4n-5} \to \mathbb{C}^{2^n}, \quad \Phi(\pi, M^e) = p_{i_1,\dots,i_n}$$

defines an *algebraic variety*

$$V_T = \overline{\Phi(\mathbb{C}^{4n-5})} \subset \mathbb{C}^{2^n}$$

Hidden Markov Models

- n observed states $Y_1, \ldots, Y_n$, each taking ℓ possible values
- n hidden states $X_1, \ldots, X_n$, each taking k possible values
- conditional independence

$$\mathbb{P}(X_i | X_1, \ldots, X_{i-1}) = \mathbb{P}(X_i | X_{i-1})$$

$$\mathbb{P}(Y_i | X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) = \mathbb{P}(Y_i | X_i)$$

- special case: all transitions $X_{i-1} \mapsto X_i$ same $k \times k$-stochastic matrix $P = (p_{ij})$; all transitions $X_i \mapsto Y_i$ same $k \times \ell$-stochastic matrix $T = (t_{ij})$

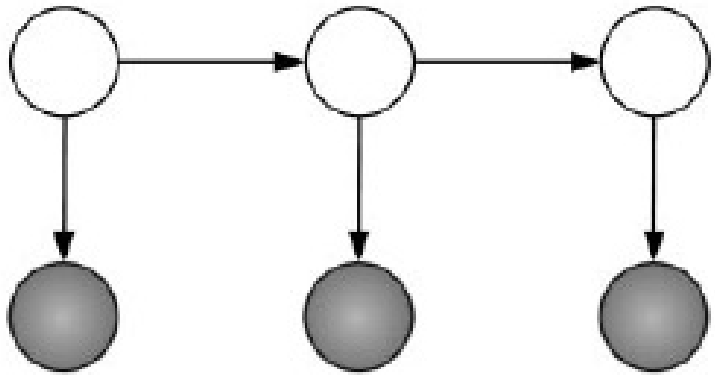
- a HMM described by the image of a polynomial map

$$\Phi : \mathbb{R}^{k(k+1)} \to \mathbb{R}^{\ell^n}$$

of degree $n-1$ bi-homogeneous in the coordinates p_{ij} and t_{ij}

- plus added positivity and normalization conditions (stochastic matrices and probability distributions)

- Example with $k = \ell = 2$ and $n = 3$, $\Phi = (\Phi_{ijk}) : \mathbb{R}^8 \to \mathbb{R}^8$



$$\begin{aligned}\Phi_{ijk} = \quad & p_{00}p_{00}t_{0i}t_{0j}t_{0k} + p_{00}p_{01}t_{0i}t_{0j}t_{1k} + p_{01}p_{10}t_{0i}t_{1j}t_{0k} + p_{01}p_{11}t_{0i}t_{1j}t_{1k} \\ + \quad & p_{10}p_{00}t_{1i}t_{0j}t_{0k} + p_{10}p_{01}t_{1i}t_{0j}t_{1k} + p_{11}p_{10}t_{1i}t_{1j}t_{0k} + p_{11}p_{11}t_{1i}t_{1j}t_{1k}\end{aligned}$$

• invariants of the HMM: polynomial functions on $\mathbb{R}^{\ell^n}$ that vanish on the image of Φ

• ideal $\mathcal{I}_\Phi$ generated by invariants? small k, ℓ, n Gröbner bases; larger computationally hard

Questions

• Viterbi sequence: find the most likely hidden data given observed data

• find all parameter values for a model that result in the same observed distribution

• find what parameter-independent relations hold between the observed probabilities $p_{i_1,\ldots,i_n} = \Phi_{i_1,\ldots,i_n}$

Setting of Phylogenetic Algebraic Geometry

- $\mathcal{T}$ a rooted binary tree with n leaves (hence $2n-2$ edges)
- At each vertex a binary random variable (e.g. one of the syntactic parameters)
- Probability distribution at the root vertex $\pi = (p, 1-p)$
- Along each edge e transition matrix: stochastic matrix $P_e = (p_{ij}^{(e)})$ with $\sum_i p_{ij}^{(e)} = 1$
- these represent the probabilities that a mutation in the parameter happens along that edge

Model Parameters

• the random variables at the leaves of the tree are *observed*; the random variables at the interior nodes are *hidden* (assuming no direct knowledge of the "ancient languages" in the family)

• matrix entries of transition matrices P_e and probability π at root vertex are model parameters

• number of parameters $N = (2n-2)k^2 + k$
(binary variable $k = 2$)

Polynomial Map

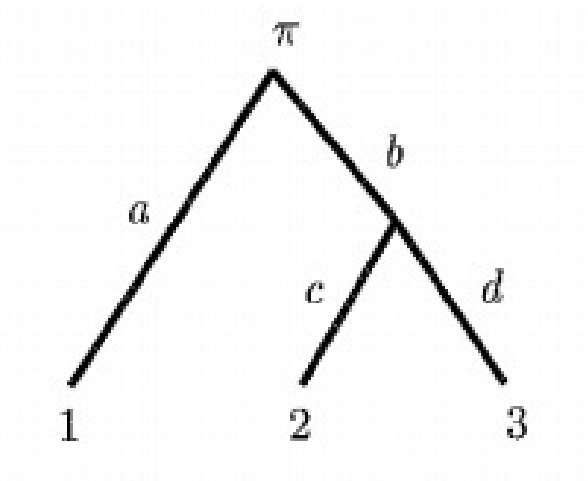
- at the n leaves there are $k^n = 2^n$ possible observations
- the probability of an observation at the leaves is a polynomial function of the parameters
- can view this as a complex polynomial

$$\Phi : \mathbb{C}^N \to \mathbb{C}^{2^n}$$

plus some (real) normalization conditions

- polytope $\Delta \subset \mathbb{R}^N_+ \subset \mathbb{C}^N$ determined by the conditions $\pi_1 + \pi_2 = 1$ and $\sum_i p_{ij}^{(e)} = 1$ with $\pi_i \geq 0$ and $p_{ij}^{(e)} \geq 0$
- Φ should map Δ to a cube $\mathcal{I}^n$ in $\mathbb{C}^{2^n}$ where $[0,1] \simeq \mathcal{I} \subset \mathbb{C}^2$ is $\mathcal{I} = \{(p_1, p_2) \,|\, p_1 + p_2 = 1, p_i \geq 0\}$

Example



$$M_a = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \qquad M_b = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix}$$

$$M_c = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix} \qquad M_d = \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix}.$$

$$\Phi_{ijk} = \pi_0 a_{0i} b_{00} c_{0j} d_{0k} + \pi_0 a_{0i} b_{01} c_{1j} d_{1k} + \pi_1 a_{1i} b_{10} c_{0j} d_{0k} + \pi_1 a_{1i} b_{11} c_{1j} d_{1k}$$

there are 8 such polynomials: $i, j, k \in \{0, 1\}$

- polynomial Φ is homogeneous in the parameters
- can view Φ as a map of projective spaces
- in the previous example

$$\Phi : \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^2 \to \mathbb{C}^8$$

$$\Phi : \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^1(\mathbb{C}) \to \mathbb{P}^7(\mathbb{C})$$

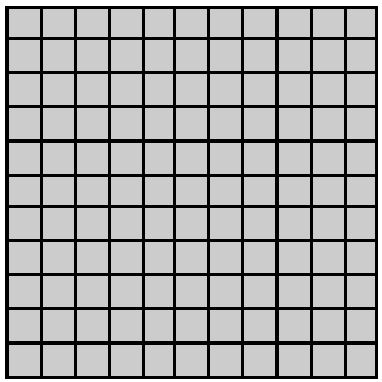
homogeneous with respect to each group of variables a, b, c, d, π

- the fibers of this morphism give all possible values of parameters (before imposing real normalization conditions) that give a certain probability at the leaves

What kinds of algebraic varieties occur in these models?

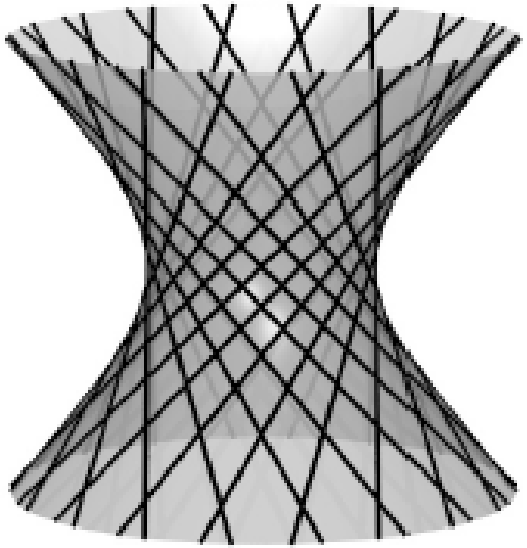
- Toric varieties (including Segre varieties and Veronese varieties)
- Determinantal varieties: the tree structure imposes rank constraints on matrices built starting from observed probabilities at the leaves
- Focus on following cases:
  1. Segre embeddings
  2. Secant varieties
- in our application to syntax we will encounter other varieties (defined as intersections of determinantal varieties in an ambient projective space) but we will use estimates of distance based on these simpler pieces

Segre embeddings



$\mathbb{P}^1 \times \mathbb{P}^1$

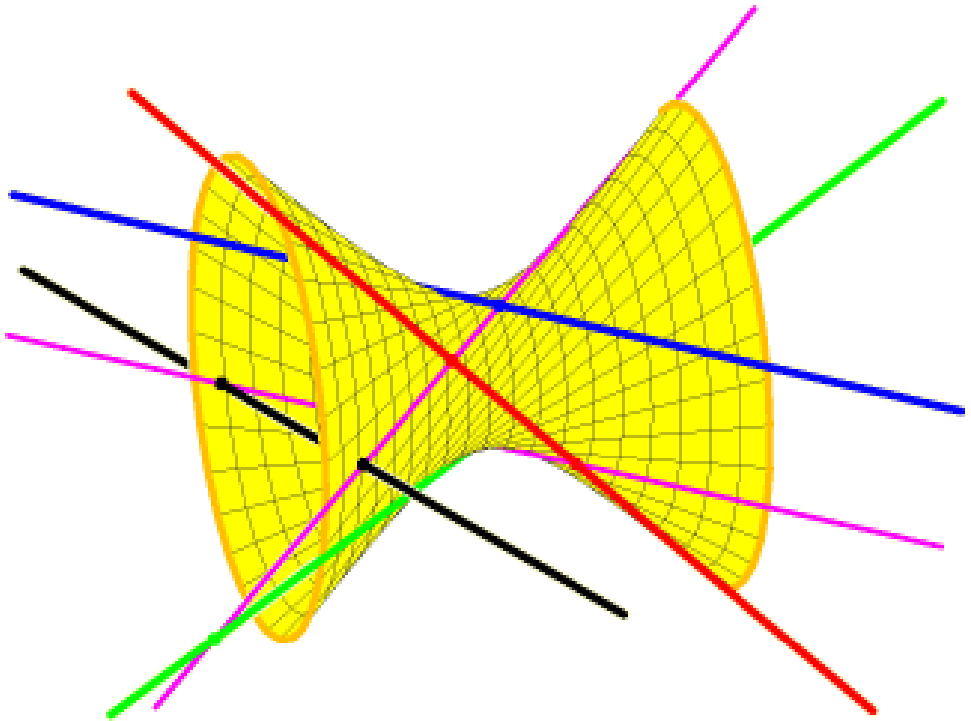
f



$X \subset \mathbb{P}^3$

$\mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^3$ with $((x_0 : x_1), (y_0 : y_1)) \mapsto (x_0y_0 : x_0y_1 : x_1y_0 : x_1y_1)$ and higher dimensional generalizations (embeddings of products of projective spaces inside larger projective spaces)

Secant varieties



variety of cords, closure (Zariski) of union of all secant lines of a variety V

Main Toolbox: Phylogenetic Invariants

• Allman–Rhodes theorem: ideal $\mathcal{I}_T$ defining V_T generated by all 3×3 minors of all *edge flattenings* of tensor $P = (p_{i_1,\ldots,i_n})$: $2^r \times 2^{n-r}$-matrix $Flat_{e,T}(P)$

$$Flat_{e,T}(P)(u, v) = P(u_1, \ldots, u_r, v_1, \ldots, v_{n-r})$$

where edge e removal separates boundary distribution into 2^r variable and 2^{n-r} variables

• phylogenetic invariants $\phi_T(P)$: 3×3 minors evaluated at boundary distribution $P = (p_{i_1,\ldots,i_n})$ given by data

- candidate trees T test by phylogenetic invariants
  - if T is the correct tree the phylogenetic invariants ϕ_T vanish when evaluated on the observed boundary distribution P (obtained from the data)

$$\phi_T(P) = 0$$

  - usually some noise in the data, so compare trees by how closely satisfied is the vanishing condition
  - closeness in some norm: ℓ^∞-norm or ℓ^1-norm

$$\|\phi_T(P)\|_{\ell^\infty} = \max_{M \in 3\times 3\text{-minors of } Flat_{e,T}(P)} |\det(M)|$$

$$\|\phi_T(P)\|_{\ell^1} = \sum_{M \in 3\times 3\text{-minors of } Flat_{e,T}(P)} |\det(M)|$$

  - ℓ^∞-norm is a weaker invariant of the ℓ^1-norm: loses information about the ϕ_T

Main Toolbox: Euclidean Distance

- Euclidean distance of the point P from the variety V_T (in ambient affine space)
- Eckrat–Young formula: for a determinantal variety

$$\mathcal{D}_r(n,m) = \{n \times m \text{ matrices of rank } \leq r\}$$

$$\text{dist}(M, \mathcal{D}_r(n,m)) = (\sum_{i=r+1}^{n} \sigma_i^2)^{1/2}$$

with σ_i singular values of the $n \times m$ flattenings M

Estimates of distance

- Euclidean distance of the point P from certain intersections $V_k \cap W$
- in general $\mathrm{dist}(P, V_1) < \mathrm{dist}(P, V_2)$ does not imply $\mathrm{dist}(P, V_1 \cap W) < \mathrm{dist}(P, V_2 \cap W)$
- assume established that $P \in W$
- then conditional case: if know that $P \in W$, then minimizing $\mathrm{dist}(P, V_k)$ suffices
- in more general cases, can use separate distances $\mathrm{dist}(P, V)$ and $\mathrm{dist}(P, W)$ as estimates from below of $\mathrm{dist}(P, V \cap W) \geq \max\{\mathrm{dist}(P, V), \mathrm{dist}(P, W)\}$ if easier to compute: if large can be used to rule out candidate trees

Procedure

- set of languages $\mathcal{L} = \{\ell_1, \dots, \ell_n\}$ (selected subfamily)
- set of syntactic parameters mapped for all: π_i, $i = 1, \dots, N$
- gives vectors $\pi_i = (\pi_i(\ell_j)) \in \mathbb{F}_2^n$
- compute frequencies

$$P = \{p_{i_1,\dots,i_n} = \frac{N_{i_1,\dots,i_n}}{N}\}$$

with $N_{i_1,\dots,i_n}$ = number of occurrences of binary string $(i_1, \dots, i_n) \in \mathbb{F}_2^n$ among the $\{\pi_i\}_{i=1}^N$

- Produce automatically a set of candidate trees (eg PHYLIP)
- Given a *candidate tree* T, compute all 3×3 minors of each flattening matrix $Flat_{e,T}(P)$, for each edge
- evaluate ℓ^∞ and ℓ^1 norm of $\phi_T(P)$ over all 3×3 minors of flattening matrices
- obtain estimates of Euclidean distance of P to V_T (or part of V_T that distinguishes candidate trees)
- select best fit tree on the basis of these tests

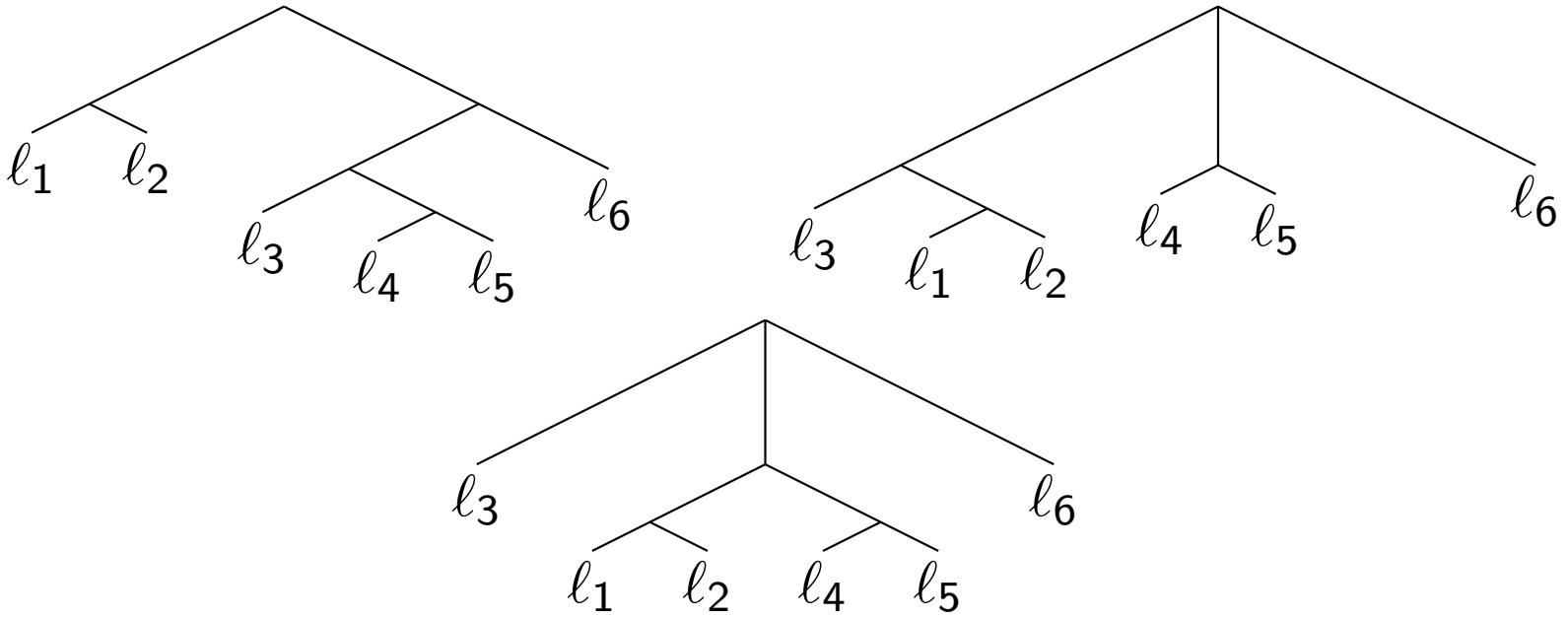
First Example: Germanic Languages

- small set of languages: ℓ_1 =Dutch, ℓ_2 =German, ℓ_3 =English, ℓ_4 =Faroese, ℓ_5 =Icelandic, ℓ_6 =Swedish
- candidate trees produced by PHYLIP on SSWL data

$$\begin{aligned} \texttt{pars1} &= ((\ell_1, \ell_2), (\ell_3, (\ell_4, \ell_5)), \ell_6) \\ \texttt{pars2} &= ((\ell_3, (\ell_1, \ell_2)), (\ell_4, \ell_5), \ell_6) \\ \texttt{pars3} &= (\ell_3, ((\ell_1, \ell_2), (\ell_4, \ell_5)), \ell_6) \end{aligned}$$

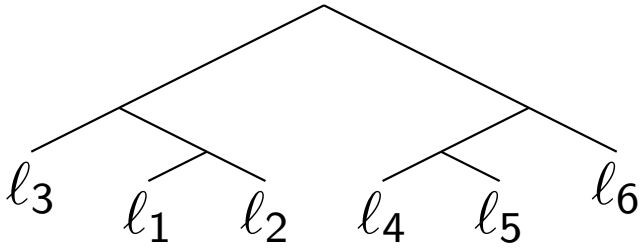
- compute flattenings for each of these trees (after resolving trivalent ambiguities into binary trees)

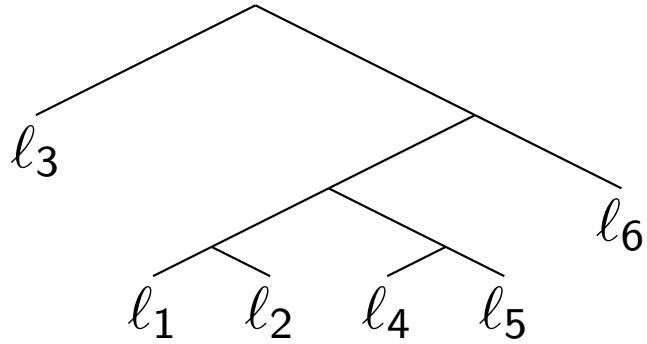
- `pars1`, `pars2`, and `pars3` trees



- resolve non-binary trees

• up to shifts in the position of the root vertex binary trees for `pars2` and `pars3`





• position of the root vertex not determined by this algorithm, only tree topology: need to use additional information to locate it

• note that all these candidate trees agree on the proximity of ℓ_1 and ℓ_2 (Dutch and German) and of ℓ_4 and ℓ_5 (Faroese and Icelandic) ...conditional case

- Flattenings:
  - `pars1`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P) \quad 8\times 8\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

  - `pars2`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P) \quad 8\times 8\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

  - `pars3`:

$$\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P) \quad 4\times 16\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P) \quad 16\times 4\text{matrix}$$

$$\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P) \quad 16\times 4\text{matrix}$$

- $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ contribute to all candidate trees, do not discriminate between them

- conditional problem: assuming that P lies on the varieties cut out by the phylogenetic invariants of $\mathrm{Flat}_{\{\ell_1,\ell_2\}\cup\{\ell_3,\ell_4,\ell_5,\ell_6\}}(P)$ and $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3,\ell_6\}\cup\{\ell_4,\ell_5\}}(P)$ select the most likely candidate that makes the remaining condition satisfied

- left with simpler setting:
  - $F_1 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ for `pars1`
  - $F_2 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ for `pars2`
  - $F_3 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ for `pars3`

- single flattening: phylogenetic ideal generated by its 3×3 minors

• the geometry involved consists of classical algebro-geometric spaces:

- `pars1`: secant variety $\mathrm{Sec}(\mathcal{S}(8,8))$ of Segre variety $\mathcal{S}(8,8) = \mathbb{P}^7 \times \mathbb{P}^7$ embedded in $\mathbb{P}^{63}$ via Segre embedding $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_6} y_{i_3,i_4,i_5}$
- `pars2`: $\mathrm{Sec}(\mathcal{S}(8,8))$, with $\mathcal{S}(8,8)$ embedded in $\mathbb{P}^{63}$ via $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_3} y_{i_4,i_5,i_6}$.
- `pars3`: secant variety $\mathrm{Sec}(\mathcal{S}(16,4))$ of Segre variety $\mathcal{S}(16,4) = \mathbb{P}^{15} \times \mathbb{P}^3$ embedded in $\mathbb{P}^{63}$ via Segre embedding $u_{i_1,\ldots,i_6} = x_{i_1,i_2,i_4,i_5} y_{i_3,i_6}$.

Boundary distribution

- 90 SSWL parameters are completely mapped for these languages
- for each binary string $(i_1, \dots, i_6)$ count occurrences as values of some syntactic parameter on the languages $\ell_1, \dots, \ell_6$
- frequency matrix:

$$\begin{array}{lll} n_{110111} = 3 & n_{000011} = 1 & n_{000010} = 4 \\ n_{000000} = 40 & n_{110000} = 2 & n_{001110} = 1 \\ n_{000100} = 2 & n_{111111} = 22 & n_{111110} = 1 \\ n_{000110} = 1 & n_{111101} = 3 & n_{100000} = 2 \\ n_{010000} = 1 & n_{111001} = 2 & n_{110110} = 1 \\ n_{010111} = 1 & n_{001000} = 2 & n_{000111} = 1 \end{array}$$

$n_{i_1,\dots,i_6} = 0$ otherwise; frequencies $p_{i_1,\dots,i_6} = n_{i_1,\dots,i_6}/90$

- from this compute the flattening matrices

Example:

flattening matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$

$$\begin{pmatrix}
\frac{4}{9} & \frac{1}{45} & \frac{1}{45} & 0 & \frac{2}{45} & \frac{1}{90} & 0 & \frac{1}{90} \\
\frac{1}{90} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & \frac{1}{90} \\
0 & 0 & 0 & 0 & \frac{1}{90} & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{45} & \frac{1}{30} & 0 & \frac{1}{30} & 0 & \frac{11}{45}
\end{pmatrix}$$

ℓ_1 =Dutch, ℓ_2 =German, ℓ_3 =English, ℓ_4 =Faroese, ℓ_5 =Icelandic, ℓ_6 =Swedish

flattening matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$

$$\begin{pmatrix}
\frac{4}{9} & \frac{2}{45} & \frac{1}{45} & \frac{1}{90} & 0 & \frac{1}{90} & 0 & \frac{1}{90} \\
\frac{1}{90} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{90} \\
\frac{1}{45} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & \frac{1}{90} & 0 & 0 & 0 & \frac{1}{30} \\
\frac{1}{45} & 0 & 0 & \frac{1}{90} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{90} & \frac{1}{45} & 0 & \frac{1}{30} & \frac{11}{45}
\end{pmatrix}$$

flattening matrix $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$

$$\begin{pmatrix}
\frac{4}{9} & 0 & \frac{1}{45} & 0 \\
\frac{1}{90} & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & \frac{1}{45} \\
\frac{2}{45} & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\frac{1}{45} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{30} \\
\frac{1}{90} & \frac{1}{90} & \frac{1}{90} & 0 \\
0 & \frac{1}{90} & 0 & 0 \\
0 & 0 & 0 & 0 \\
\frac{1}{90} & \frac{1}{30} & \frac{1}{90} & \frac{11}{45}
\end{pmatrix}$$

Phylogenetic invariants favor the tree binary tree pars2:

- for the tree pars1

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_1} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_1}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_1} |\phi(P)| = \frac{3707}{364500}$$

- for the tree pars2

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_2} |\phi(P)| = \frac{419}{364500}$$

$$\|\phi_{T_2}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_2} |\phi(P)| = \frac{2719}{364500}$$

- for the tree pars3

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max_{\phi \in 3\times 3 \text{ minors of } F_3} |\phi(P)| = \frac{22}{18225}$$

$$\|\phi_{T_3}(P)\|_{\ell^1} = \sum_{\phi \in 3\times 3 \text{ minors of } F_3} |\phi(P)| = \frac{949}{91125}$$

Euclidean distance

• varieties defined by the 3×3-minors of the three flattening matrices:

- $\mathcal{D}_2(8,8) = \mathrm{Sec}(\mathcal{S}(8,8))$: 28-dimensional determinantal variety of all 8×8 matrices of rank at most two
- $\mathcal{D}_2(16,4) = \mathrm{Sec}(\mathcal{S}(16,4))$: 36-dimensional determinantal variety of all 16×4 matrices of rank at most two

• phylogenetic algebraic variety of a candidate tree: intersection with remaining equations coming from the 3×3 minors of the other common flattenings (intersections of three different determinantal varieties inside a common ambient space $\mathbb{A}^{2^6}$)
• conditional case, assuming P on the common varieties, evaluate distance from the remaining one

- Euclidean distance of $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)$ from $\mathcal{D}_2(8,8)$
- Euclidean distance of $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)$ from $\mathcal{D}_2(8,8)$
- Euclidean distance of the point $\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)$ from $\mathcal{D}_2(16,4)$.

Euclidean distance favors the tree pars2

Eckart-Young theorem: compute singular values of these three matrices

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P)) \sim \mathrm{diag}(0.44940, 0.25001,$$

$$0.19237\times10^{-1}, 0.96007\times10^{-2}, 0.21595\times10^{-2}, 0.88079\times10^{-3}, 4.6239\times10^{-19}, 0)$$

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P)) \sim \mathrm{diag}(0.44956, 0.25018,$$

$$0.14729 \times 10^{-1}, 0.44229 \times 10^{-2}, 0.27802 \times 10^{-2}, 0.24881 \times 10^{-17}, 0)$$

$$\Sigma(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P)) \sim$$

$$\mathrm{diag}(0.44939, 0.24994, 0.20625 \times 10^{-1}, 0.94442 \times 10^{-2}).$$

Then obtain

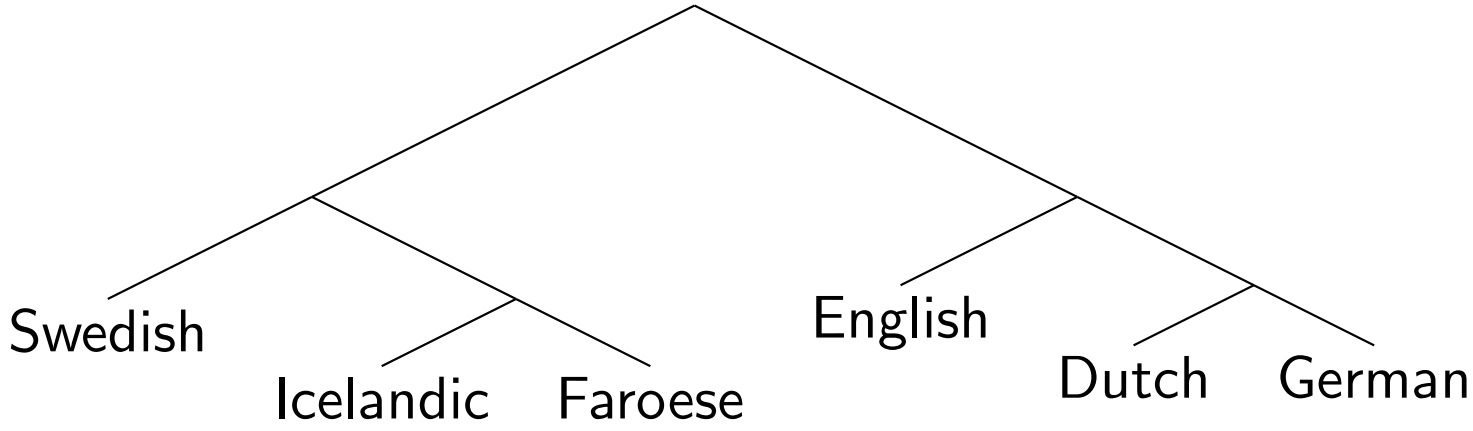
$$\mathrm{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_6\}\cup\{\ell_3,\ell_4,\ell_5\}}(P), \mathrm{Sec}(\mathcal{S}(8,8)))^2 = \sigma_3^2+\cdots+\sigma_8^2 = 0.46768\times10^{-3}$$

$$\mathrm{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6\}}(P), \mathrm{Sec}(\mathcal{S}(8,8)))^2 = \sigma_3^2+\cdots+\sigma_8^2 = 0.24424\times10^{-3}$$

$$\mathrm{dist}(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4,\ell_5\}\cup\{\ell_3,\ell_6\}}(P), \mathrm{Sec}(\mathcal{S}(16,4)))^2 = \sigma_3^2+\sigma_4^2 = 0.51457\times10^{-3}$$

Result

- correctly identifies the West Germanic/North Germanic split



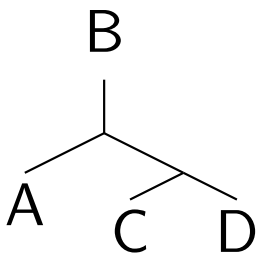
- other PHYLIP candidate trees misplaced it

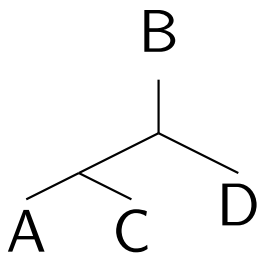
Other examples: Romance languages, Slavic languages

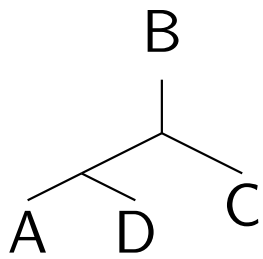
- same method; comparative use of SSWL and Longobardi data
- placement of ancient languages and root vertex
- estimates of Euclidean distance

Another example with Germanic languages:

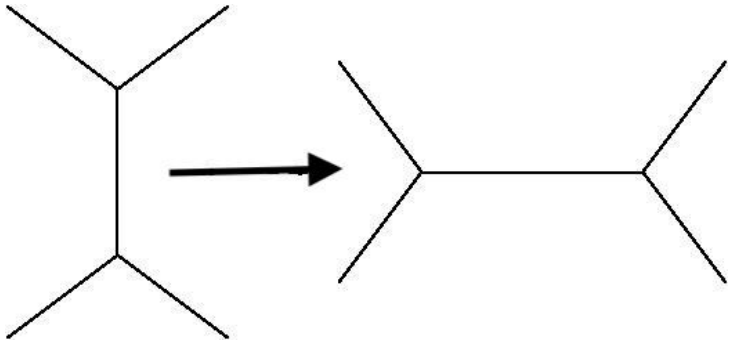
- $\ell_1 =$ Norwegian, $\ell_2 =$ Danish, $\ell_3 =$ Gothic, $\ell_4 =$ Old English, $\ell_5 =$ Icelandic, $\ell_6 =$ English, $\ell_7 =$ German
- for these both SSWL and Longobardi data available (compare data sets)
- PHYLIP produces single tree but with higher valence vertices: resolve in different candidate binary trees as before
- ancient languages Gothic and Old English treated as leaves creates problems
- with $A = \{\ell_1, \ell_2\}$, $B = \{\ell_3, \ell_4\}$, $C = \{\ell_5\}$, $D = \{\ell_6, \ell_7\}$ binary splittings





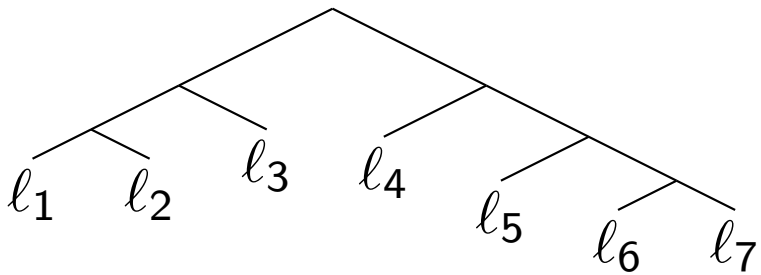


• topological move for placement of ancient languages preserving relation to rest of tree

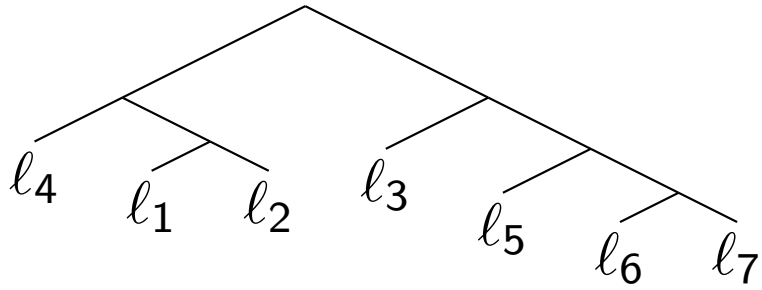


• resulting trees

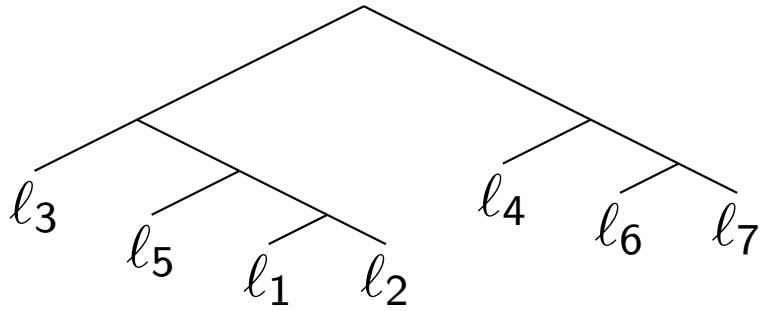
$T_1(G) =$



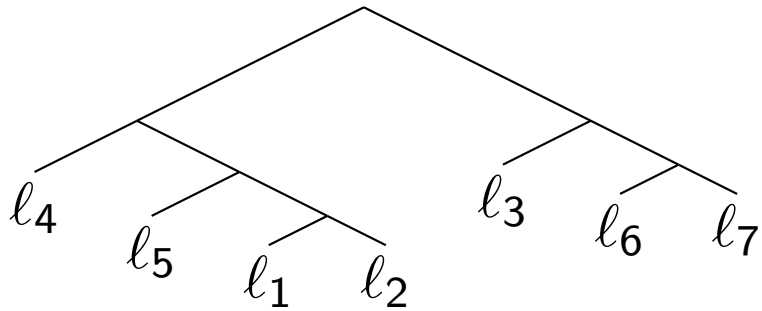
$T_2(G) =$



$T_3(G) =$



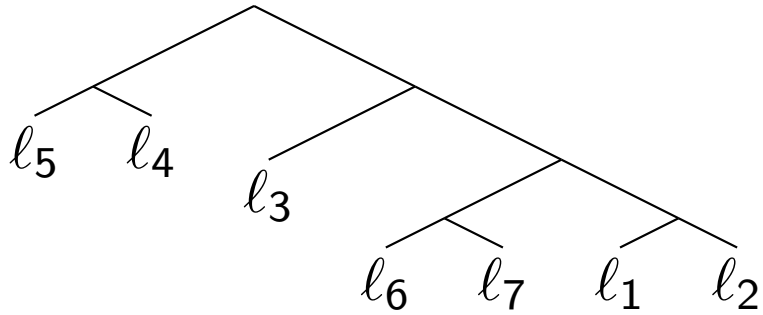
$T_4(G) =$



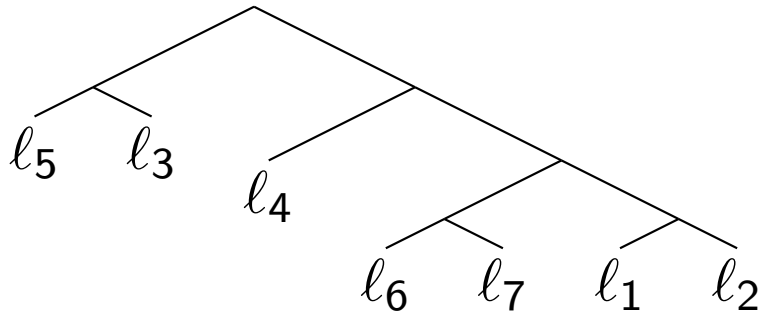
$\ell_1 =$ Norwegian, $\ell_2 =$ Danish, $\ell_3 =$ Gothic, $\ell_4 =$ Old English, $\ell_5 =$ Icelandic, $\ell_6 =$ English, $\ell_7 =$ German

- also two "worse" trees

$T_5(G) =$



$T_6(G) =$



T_5 incorrectly places $\{\ell_1, \ell_2\}$ and $\{\ell_6, \ell_7\}$ in closer proximity and ℓ_5 in a separate branch away from the ancient languages $\{\ell_3, \ell_4\}$, placing ℓ_4 as the ancient language in closer proximity to ℓ_5; T_6 has similar misplacements

Longobardi parameters data

$$\ell_1 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0]$$

$$\ell_2 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0]$$

$$\ell_3 = [1,1,1,1,0,1,1,0,0,0,0,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0]$$

$$\ell_4 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0]$$

$$\ell_5 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0]$$

$$\ell_6 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0]$$

$$\ell_7 = [1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0]$$

- boundary distribution at leaves

$$n_{1111111} = 12 \quad n_{0000000} = 24 \quad n_{1101111} = 1 \quad n_{1111101} = 1$$
$$n_{1111100} = 1 \quad n_{1111011} = 1 \quad n_{1100111} = 1 \quad n_{0011111} = 1$$

probabilities

$$p_{1111111} = \frac{2}{7} \quad p_{0000000} = \frac{4}{7} \quad p_{1101111} = \frac{1}{42} \quad p_{1111101} = \frac{1}{42}$$

$$p_{1111100} = \frac{1}{42} \quad p_{1111011} = \frac{1}{42} \quad p_{1100111} = \frac{1}{42} \quad p_{0011111} = \frac{1}{42}$$

all other $p_{a_1 \cdots a_7} = 0$

SSWL data

- 68 SSWL variables completely mapped for all these seven languages
- boundary probability distribution for SSWL data

$$
\begin{array}{lll}
p_{0,0,0,0,0,0,0} = \frac{13}{34} & p_{1,1,1,1,1,1,1} = \frac{4}{17} & p_{0,0,1,1,0,0,1} = \frac{1}{34} \\
p_{0,0,1,0,0,0,0} = \frac{3}{68} & p_{1,1,0,1,0,0,0} = \frac{1}{68} & p_{0,0,1,1,1,1,0} = \frac{1}{68} \\
p_{0,0,1,1,1,0,0} = \frac{1}{68} & p_{0,0,1,0,1,0,0} = \frac{1}{68} & p_{1,1,0,1,0,1,1} = \frac{1}{34} \\
p_{1,0,1,1,1,0,0} = \frac{1}{68} & p_{1,1,1,1,1,0,1} = \frac{1}{68} & p_{1,1,1,1,1,0,0} = \frac{1}{68} \\
p_{1,1,1,1,0,1,1} = \frac{3}{68} & p_{1,1,0,1,1,0,1} = \frac{1}{68} & p_{0,0,0,0,1,0,0} = \frac{1}{68} \\
p_{1,1,0,0,1,1,1} = \frac{1}{68} & p_{0,0,0,0,0,1,0} = \frac{1}{68} & p_{0,0,0,1,0,0,0} = \frac{1}{34} \\
p_{0,0,0,0,0,0,1} = \frac{1}{68} & p_{0,0,1,1,0,0,0} = \frac{1}{68} & p_{1,1,0,1,1,1,1} = \frac{1}{68}
\end{array}
$$

Flattenings

$$
\begin{array}{lll}
T_1 & M_1 = \mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}, & M_2 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}} \\
T_2 & M_1 = \mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}, & M_3 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}} \\
T_3 & M_4 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}, & M_5 = \mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}} \\
T_4 & M_4 = \mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}, & M_6 = \mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}} \\
T_5 & M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}, & F_5 = \mathrm{Flat}_{\{\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_3,\ell_6,\ell_7\}} \\
T_6 & M_7 = \mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}, & F_6 = \mathrm{Flat}_{\{\ell_3,\ell_5\}\cup\{\ell_1,\ell_2,\ell_4,\ell_6,\ell_7\}}
\end{array}
$$

Phylogenetic invariants from Longobardi data (favorite: T_3)

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \max_{3\times 3\text{minors}} |\phi(P)| = \frac{4}{1029}, \quad \|\phi_{T_1}(P)\|_{\ell^1} = \sum_{3\times 3\text{minors}} |\phi(P)| = \frac{83}{8232}$$

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \max |\phi(P)| = \frac{4}{1029}, \quad \|\phi_{T_2}(P)\|_{\ell^1} = \sum |\phi(P)| = \frac{233}{24696}$$

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \max |\phi(P)| = \frac{1}{3087}, \quad \|\phi_{T_3}(P)\|_{\ell^1} = \sum |\phi(P)| = \frac{16}{3087}$$

$$\|\phi_{T_4}(P)\|_{\ell^\infty} = \max |\phi(P)| = \frac{4}{1029}, \quad \|\phi_{T_4}(P)\|_{\ell^1} = \sum |\phi(P)| = \frac{181}{18522}$$

$$\|\phi_{T_5}(P)\|_{\ell^\infty} = \max |\phi(P)| = \frac{4}{1029}, \quad \|\phi_{T_5}(P)\|_{\ell^1} = \sum |\phi(P)| = \frac{233}{24696}$$

$$\|\phi_{T_6}(P)\|_{\ell^\infty} = \max |\phi(P)| = \frac{4}{1029}, \quad \|\phi_{T_6}(P)\|_{\ell^1} = \sum |\phi(P)|\,| = \frac{83}{8232}$$

Euclidean distance estimate from Longobardi data

- $\mathrm{dist}(P, V_{T_1}) \geq L_1$

  $L_1 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_3\}\cup\{\ell_4,\ell_5,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P), \mathcal{D}_2(8,16))\}$

- $\mathrm{dist}(P, V_{T_2}) \geq L_2$

  $L_2 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_4\}\cup\{\ell_3,\ell_5,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_5,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_4\}}(P), \mathcal{D}_2(8,16))\}$

- $\mathrm{dist}(P, V_{T_3}) \geq L_3$

  $L_3 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_4,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_3,\ell_5\}}(P), \mathcal{D}_2(8,16))\}$

- $\mathrm{dist}(P, V_{T_4}) \geq L_4$

  $L_4 = \max\{d(\mathrm{Flat}_{\{\ell_1,\ell_2,\ell_5\}\cup\{\ell_3,\ell_4,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16)), d(\mathrm{Flat}_{\{\ell_3,\ell_6,\ell_7\}\cup\{\ell_1,\ell_2,\ell_4,\ell_5\}}(P), \mathcal{D}_2(8,16))\}$

- $\mathrm{dist}(P, V_{T_5}) \geq L_5$

  $$L_5 = \max\{d(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2, d(F_5(P), \mathcal{D}_2(4,32))^2\}$$

- $\mathrm{dist}(P, V_{T_5}) \geq L_6$

  $$\max\{d(\mathrm{Flat}_{\{\ell_3,\ell_4,\ell_5\}\cup\{\ell_1,\ell_2,\ell_6,\ell_7\}}(P), \mathcal{D}_2(8,16))^2, d(F_6(P), \mathcal{D}_2(4,32))^2\}$$

$$L_1 = 0.58597\times10^{-3}, \quad L_2 = 0.57831\times10^{-3}, \quad L_3 = 0.30245\times10^{-4},$$
$$L_4 = 0.58595\times10^{-3}, \quad L_5 = 0.57831\times10^{-3}, \quad L_6 = 0.58597\times10^{-3}$$

- again favors T_3 (but lower bound only)

Phylogenetic invariants from SSWL data

$$\|\phi_{T_1}(P)\|_{\ell^\infty} = \frac{13}{4913}, \quad \|\phi_{T_1}(P)\|_{\ell^1} = \frac{8811}{157216}$$

$$\|\phi_{T_2}(P)\|_{\ell^\infty} = \frac{13}{4913}, \quad \|\phi_{T_2}(P)\|_{\ell^1} = \frac{7103}{157216}$$

$$\|\phi_{T_3}(P)\|_{\ell^\infty} = \frac{13}{4913}, \quad \|\phi_{T_3}(P)\|_{\ell^1} = \frac{5439}{157216}$$

$$\|\phi_{T_4}(P)\|_{\ell^\infty} = \frac{13}{4913}, \quad \|\phi_{T_4}(P)\|_{\ell^1} = \frac{5739}{157216}$$

$$\|\phi_{T_5}(P)\|_{\ell^\infty} = \frac{13}{4913}, \quad \|\phi_{T_5}(P)\|_{\ell^1} = \frac{25}{578}$$

$$\|\phi_{T_6}(P)\|_{\ell^\infty} = \frac{207}{78608}, \quad \|\phi_{T_6}(P)\|_{\ell^1} = \frac{11795}{314432}$$

- ℓ^∞ norm incorrectly picks $T_6(G)$ but ℓ^1 orrectly selects $T_3(G)$

Euclidean distance estimate from SSWL data

$$d(M_1, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.25068 \times 10^{-2}$$

$$d(M_2, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.30816 \times 10^{-2}$$

$$d(M_3, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.18155 \times 10^{-2}$$

$$d(M_4, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.38172 \times 10^{-2}$$

$$d(M_5, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.21780 \times 10^{-2}$$

$$d(M_6, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.27252 \times 10^{-2}$$

$$d(M_7, \mathcal{D}_2(8,16))^2 = \sigma_3^2 + \cdots + \sigma_8^2 = 0.18867 \times 10^{-2}$$

$$d(F_5, \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.21971 \times 10^{-2}$$

$$d(F_6, \mathcal{D}_2(4,32))^2 = \sigma_3^2 + \sigma_4^2 = 0.13615 \times 10^{-2}$$

• lower bound on the Euclidean distance not reliable: correctly excludes $T_1(G)$, $T_2(G)$, $T_4(G)$, $T_5(G)$ but gives lower value to $T_6(G)$ rather than correct tree $T_3(G)$

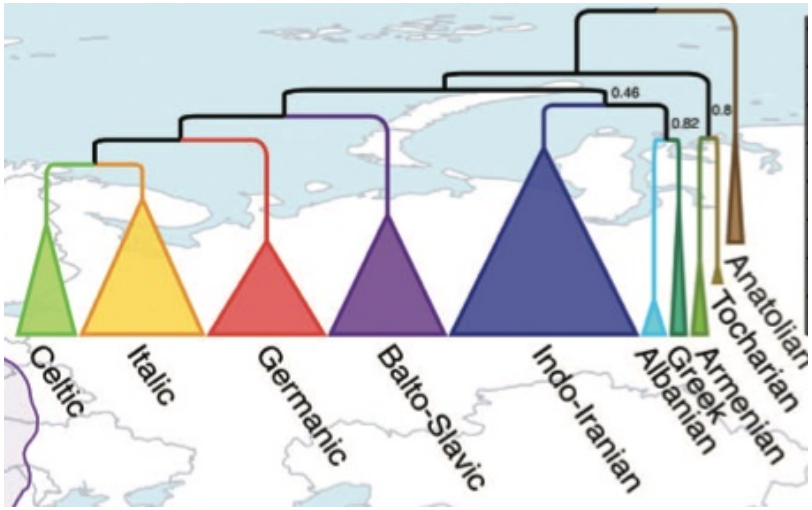
Early Indo-European tree: can one use this method to say something about the early branched of the Indo–European tree?

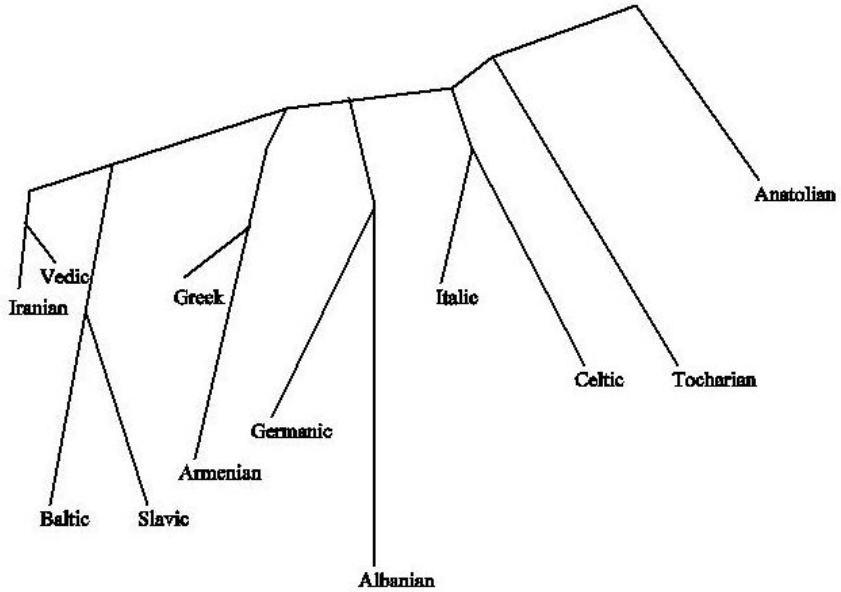
• Examples of questions about early branches of the tree of Indo-European languages:

- The relative positions of the Greco-Armenian subtrees;
- The position of Albanian in the tree;
- The relative positions of these languages with respect to the Anatolian-Tocharian subtrees.

• Try a comparison, based on SSWL data, between tree of Gray and Atkinson (Nature, 2003) and tree via morphological analysis (Ringe, Warnow, Taylor, 2002)

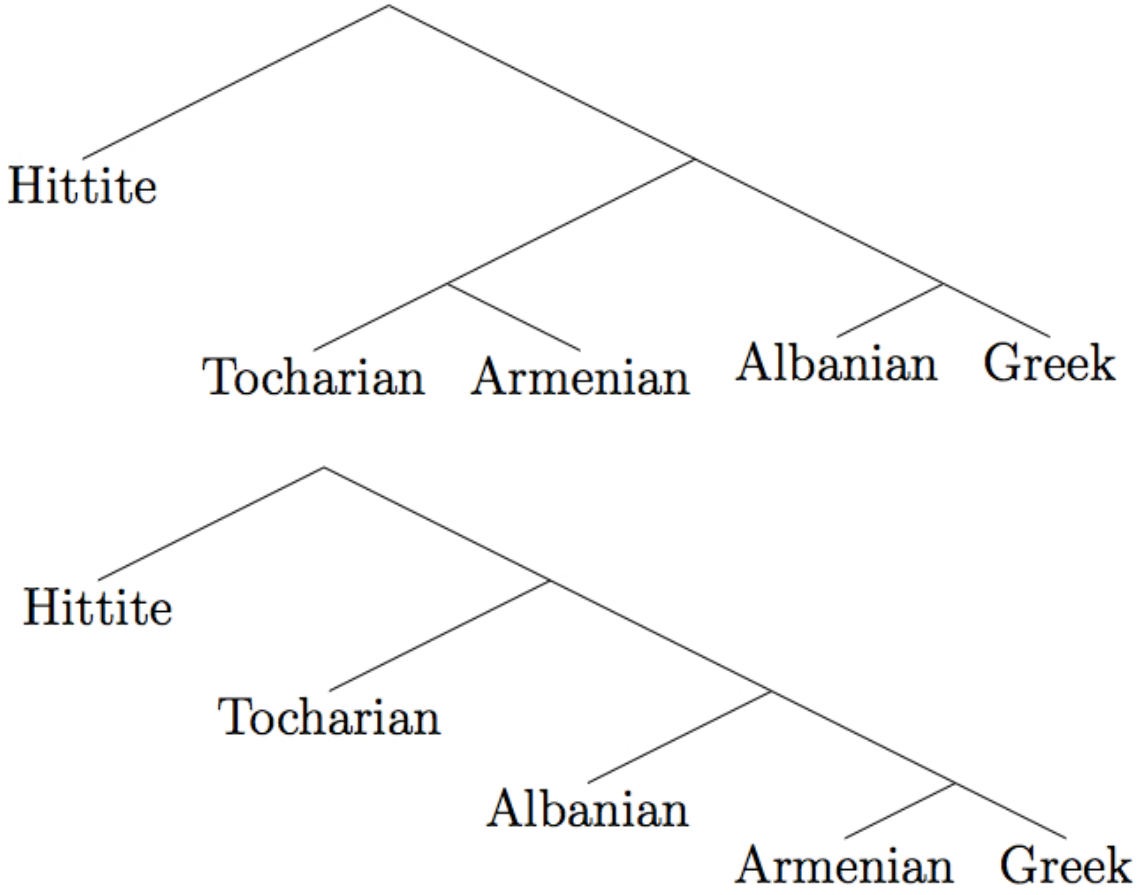
• A. Perelysvaig, M.W. Lewis, *The Indo-European controversy: facts and fallacies in Historical Linguistics*, Cambridge University Press, 2015.





The Atkinson–Gray early Indo-European tree and the Ringe–Warnow–Taylor tree

Focus on a smaller part of the tree: relative position of these languages



Can detect the difference from syntactic parameters? Using Phylogenetic Algebraic Geometry of Syntactic Parameters?

• Problem: SSWL data for Hittite, Tocharian, Albanian, Armenian, and Greek have a small number of parameters that is completely mapped for all these languages (and these parameters largely agree); Hittite and Tocharian not mapped in Longobardi's data.

• only 22 of the SSWL parameters are completely mapped for all of these languages

$$p_{00000} = 4/11, \quad p_{11111} = 3/11, \quad p_{11101} = 2/11,$$
$$p_{11011} = 1/22, \quad p_{10111} = 1/11, \quad p_{01000} = 1/22$$

with $p_{i_1,\dots,i_5} = 0$ for all the remaining binary vectors in $\{0,1\}^5$.

First Case: flattening matrices

$$\begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} & 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\begin{pmatrix} \frac{4}{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{11} \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & \frac{2}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

Second Case: flattening matrices

$$\begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} & 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

$$\begin{pmatrix} \frac{4}{11} & 0 & \frac{1}{22} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{22} \\ 0 & 0 & 0 & \frac{2}{11} \\ 0 & \frac{1}{11} & 0 & \frac{3}{11} \end{pmatrix}$$

Phylogenetic Invariants

1. For the Gray-Atkins tree T_1:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{8}{1331}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{61}{2662}$$

2. For the Ringe–Warnow–Taylor tree T_2:

$$\|\Phi_{T_1}(P)\|_{\ell^\infty} = \max_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{8}{1331}$$

$$\|\Phi_{T_1}(P)\|_{\ell^1} = \sum_{\substack{\phi \in 3\times 3 \text{minors} \\ \text{of flattenings of } T_1}} |\phi(P)| = \frac{18}{1331}$$

Estimates of Euclidean distance

- distances

$$D_{1,1} = \mathrm{dist}(\mathrm{Flat}_{e_1,T_1}(P), \mathcal{D}_2(4,8)), \quad D_{1,2} = \mathrm{dist}(\mathrm{Flat}_{e_2,T_2}(P), \mathcal{D}_2(8,4))$$

with Euclidean distance estimate for T_1 by $L_1 = \max\{D_{1,1}, D_{1,2}\}$

$$D_{2,1} = \mathrm{dist}(\mathrm{Flat}_{e_1,T_2}(P), \mathcal{D}_2(4,8)), \quad D_{2,1} = \mathrm{dist}(\mathrm{Flat}_{e_2,T_2}(P), \mathcal{D}_2(8,4))$$

with Euclidean distance estimate for T_2 by $L_2 = \max\{D_{2,1}, D_{2,2}\}$

- singular values

$$\Sigma(\mathrm{Flat}_{e_1,T_1}(P)) = \mathrm{diag}(0.3664662612, 0.3394847389, 0.5018672314\times 10^{-1}, 0)$$

$$\Sigma(\mathrm{Flat}_{e_2,T_1}(P)) = \mathrm{diag}(0.3664662612, 0.3388120907, 0.5454321492\times 10^{-1}, 0)$$

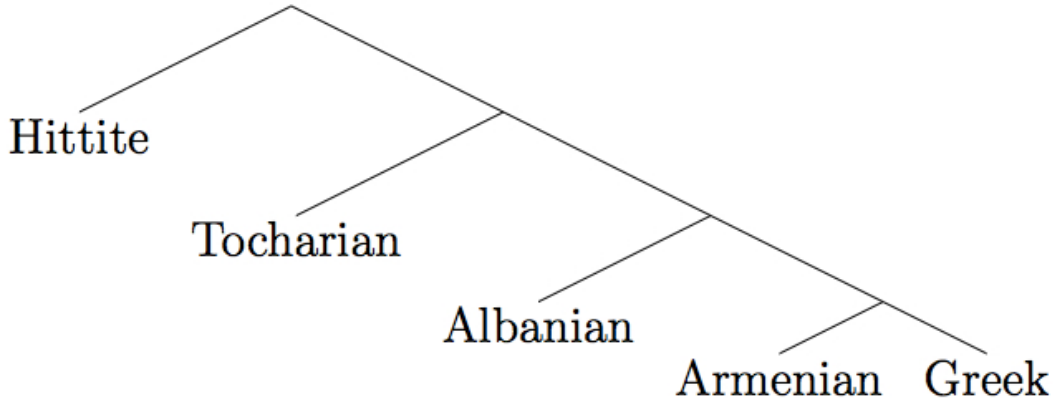
$$\Sigma(\mathrm{Flat}_{e_1,T_2}(P)) = \mathrm{diag}(0.3664662613, 0.3421098124, 0.2700872640\times 10^{-1}, 0)$$

$$\Sigma(\mathrm{Flat}_{e_2,T_2}(P)) = \mathrm{diag}(0.3664662613, 0.3394847388, 0.5018672301\times 10^{-1}, 0)$$

- $L_1 = 0.5454321492 \times 10^{-1}$ and $L_2 = 0.5018672301 \times 10^{-1}$

• the ℓ^∞ norm does not distinguish the two trees while the ℓ^1 norm prefers the Ringe–Warnow–Taylor tree T_2; Euclidean distance estimate also favors T_2

• the SSWL data favor the Ringe–Warnow–Taylor tree over the Atkinson–Gray tree, *but the SSWL data is problematic!* ...need better syntactic data on these languages (especially Hittite and Tocharian that are poorly mapped in databases)



Some more general facts about Phylogenetic Algebraic Geometry

- parameter inference from tropicalization of the algebraic variety
- min-plus (or tropical) semiring $\mathbb{T} = \mathbb{R} \cup \{\infty\}$, with operations $\oplus$ and $\odot$ given by

$$x \oplus y = \min\{x, y\},$$

with ∞ the identity element for $\oplus$ and with

$$x \odot y = x + y,$$

with 0 the identity element for $\odot$

- operations $\oplus$ and $\odot$ satisfy associativity and commutativity and distributivity of the product $\odot$ over the sum $\oplus$
- addition is no longer invertible and is idemponent
$x \oplus x = \min\{x, x\} = x$

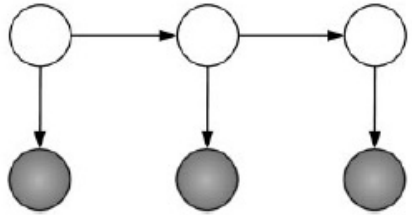
Tropical polynomials

- function $\phi : \mathbb{R}^n \to \mathbb{R}$ of the form

$$\phi(x_1, \ldots, x_n) = \oplus_{j=1}^{m} a_j \odot x_1^{k_{j1}} \odot \cdots \odot x_n^{k_{jn}}$$

$$= \min\{ \begin{array}{l} a_1 + k_{11}x_1 + \cdots + k_{1n}x_n, \\ a_2 + k_{21}x_1 + \cdots + k_{2n}x_n, \\ \cdots \\ a_m + k_{m1}x_1 + \cdots + k_{mn}x_n \ \}. \end{array}$$

- tropicalization: algebraic varieties become piecewise linear spaces
- can recover information about a variety from its tropicalization



• in HMM example with $n = 3$ and $k = \ell = 2$ the tropicalization of the polynomials Φ_{ijk}

$$\begin{aligned}\Phi_{ijk} = \quad & p_{00}p_{00}t_{0i}t_{0j}t_{0k} + p_{00}p_{01}t_{0i}t_{0j}t_{1k} + p_{01}p_{10}t_{0i}t_{1j}t_{0k} + p_{01}p_{11}t_{0i}t_{1j}t_{1k} \\ + \quad & p_{10}p_{00}t_{1i}t_{0j}t_{0k} + p_{10}p_{01}t_{1i}t_{0j}t_{1k} + p_{11}p_{10}t_{1i}t_{1j}t_{0k} + p_{11}p_{11}t_{1i}t_{1j}t_{1k}\end{aligned}$$

is given by

$$\tau_{ijk} = \min\{u_{h_1h_2} + u_{h_2h_3} + v_{h_1i} + v_{h_2j} + v_{h_3k} \,|\, (h_1, h_2, h_3) \in \{0,1\}^3\}$$

where $u_{ab} = -\log(p_{ab})$ and $v_{ab} = -\log(t_{ab})$

• Viterbi sequence: (h_1, h_2, h_3) realizing mimimum, given observed (i, j, k) is the Viterbi sequence of hidden data

Newton polytope

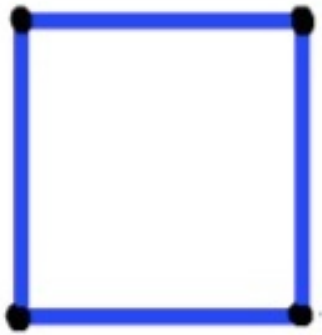
- polynomial $f = \sum_{\omega \in \mathbb{Z}^n} a_\omega x^\omega$ with $x^\omega = x_1^{\omega_1} \cdots x_n^{\omega_n}$
- Newton polytope

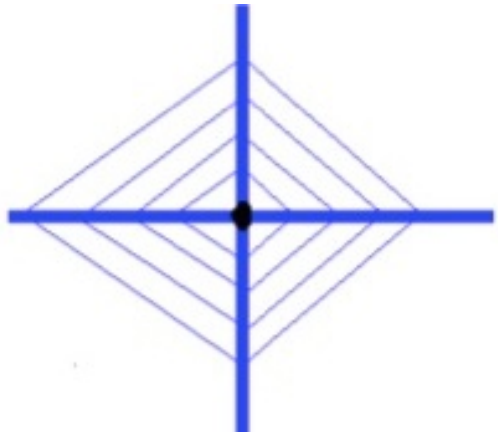
$$\mathcal{N}(f) = \text{Convex Hull}\{\omega \in \mathbb{Z}^n \,|\, a_\omega \neq 0\} \subset \mathbb{R}^n$$

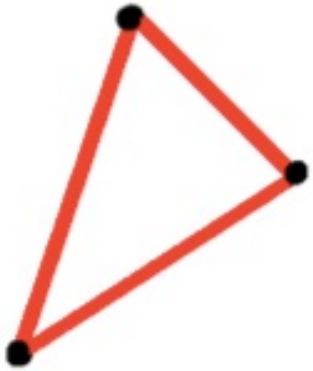
- $\mathcal{N}(f+g) = \mathcal{N}(f) \cup \mathcal{N}(g)$ and $\mathcal{N}(f \cdot g) = \mathcal{N}(f) + \mathcal{N}(g)$
(Minkowski sum of polytopes $\mathcal{P} + \mathcal{Q} = \{x + y \,|\, x \in \mathcal{P}, y \in \mathcal{Q}\}$
- normal fan $\mathcal{C}(\mathcal{N}(f))$: normal cones of all faces $\mathcal{C}_F(\mathcal{N}(f))$

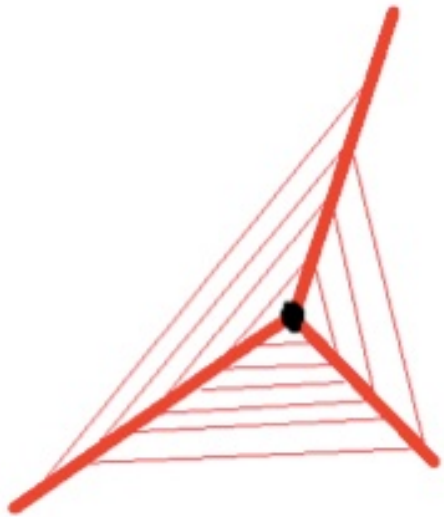
$$\mathcal{C}_F(\mathcal{N}(f)) = \{w \in \mathbb{R}^n \,|\, F = F_w(\mathcal{N}(f))\}$$

$$F_w(\mathcal{N}(f)) = \{x \in \mathcal{N}(f) \,|\, (x-y) \cdot w \leq 0 \ \forall y \in \mathcal{N}(f)\}$$









• the set of parameters $U = (u_{ab})$, $V = (v_{ab})$ in tropicalization τ_{ijk} of Φ_{ijk} that determine the Viterbi sequence (h_1, h_2, h_3) is the normal cone to a vertex of the Newton polygon $\mathcal{N}(\Phi_{ijk})$

• given observed data (i, j, k) and hidden data (h_1, h_2, h_3) the normal cones of $\mathcal{N}(\Phi_{ijk})$ give all parameter values for which (h_1, h_2, h_3) is the most likely explanation for the observed (i, j, k)

• domains of linearity of the piecewise linear tropical τ_{ijk} are the cones in the normal fan $\mathcal{C}_F(\mathcal{N}(\Phi_{ijk}))$; each maximal cone corresponds to one set of hidden data (h_1, h_2, h_3) maximizing probability

$$\tau_{ijk} = -\log \mathbb{P}((X_1, X_2, X_3) = (h_1, h_2, h_3) \,|\, (Y_1, Y_2, Y_3) = (i, j, k))$$

• each vertex of the Newton polygon $\mathcal{N}(\Phi_{ijk})$ determines an inference function: $(i, j, k) \mapsto (h_1, h_2, h_3)$ that realize min τ_{ijk}

How to improve the syntactic phylogenetic models?

- the hypothesis that individual syntactic parameters behave like identically distributed independent random variables for a Markov process on a tree needs to be revised: relations between parameters need to be included in the model
- part of the relations can only be detected statistically (more detailed discussion of this later)
- need to correct the boundary distribution at the leaves of the tree by a different weight for different parameters that corresponds to different amount of "recoverability" from other parameters (amount of independence)
- additional information not captured by trees: more topological information with non-trivial homology (unlike trees)