# Principal component analysis in treespace

Sean Cleary Aasa Feragen Megan Owen Daniel Vargas

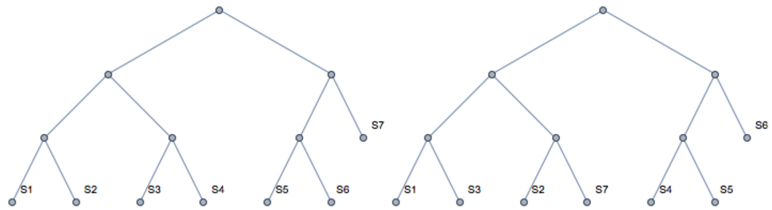
The City College of New York and the CUNY Graduate Center

NZ Phylogenomics, Waiheke Feb 2020

This material is based upon work supported by the National Science Foundation under grant no. DMS-1417820

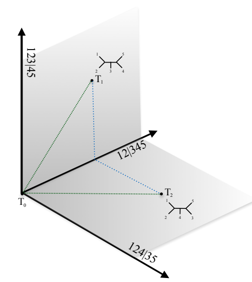
# Distances between phylogenetic trees

- Given two trees (rooted or unrooted) on the same set of leaves (taxa), to what extent to they differ?




- Metrics on trees: Nearest Neighbor Interchange (NNI), Robinson-Foulds (RF) $\Delta$, Subtree-Prune-Regraft (SPR), Tree Bisection and Reconnection (TBR), Billera-Holmes-Vogtmann (BHV), ...

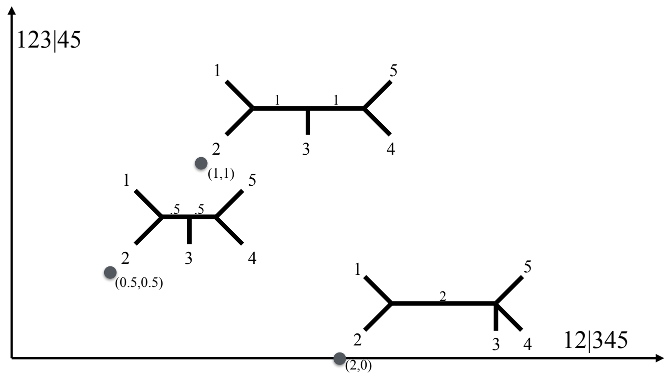
# Trees as Vectors




| | $1 \mid 2345$ | $2 \mid 1345$ | $3 \mid 1245$ | $4 \mid 1235$ | $5 \mid 1234$ | $12 \mid 345$ | $13 \mid 245$ | $14 \mid 235$ | $15 \mid 234$ | $23 \mid 145$ | $24 \mid 135$ | $25 \mid 134$ | $34 \mid 125$ | $35 \mid 124$ | $45 \mid 123$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $T_0 = (1, 2, 3, 4, 5)$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $T_1 = ((1, 2), (3, (4, 5))$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $T_2 = ((1, 2), (4, (3, 5))$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

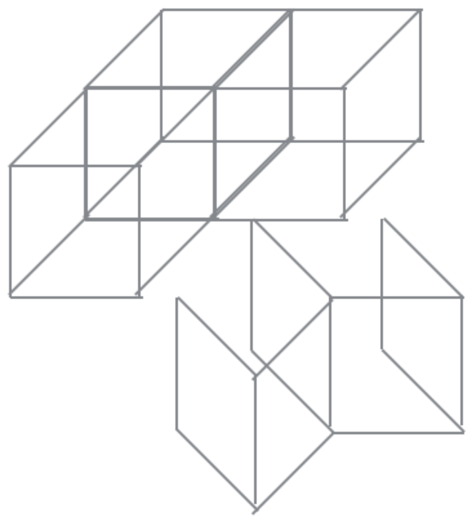
## Billeara-Holmes-Vogtmann space

- Trees on fixed *n* taxa, edge lengths positive reals
- Edges to leaves have fixed length 1
- Form a space with a metric which is locally Euclidean



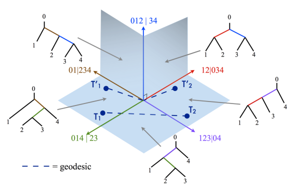
## Billera-Holmes-Vogtmann treespace

- Adjacency of orthants: two orthants of dimension $k$ share a face of dimension $l$ if they have $l$ edges in common.
- Gives gluing of orthants to obtain space



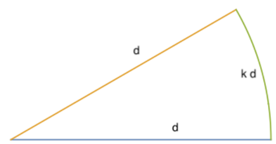
## Finding geodesics in treespace

- Piecewise Euclidean segments in a sequence of orthants
- Many cells to consider
- First algorithms exponential
- Polynomial-time algorithm of order $O(n^4)$ of Owen-Provan (2011) uses incompatibility graph of edges to compute which edges to drop and introduce successively.
- Owen-Provan algorithm foundational for computing means, interpolations, variances, principal components, ...

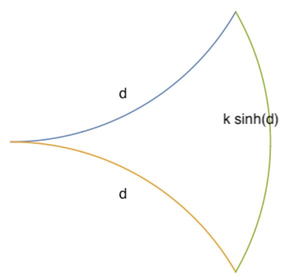


## Divergence of geodesics in hyperbolic space

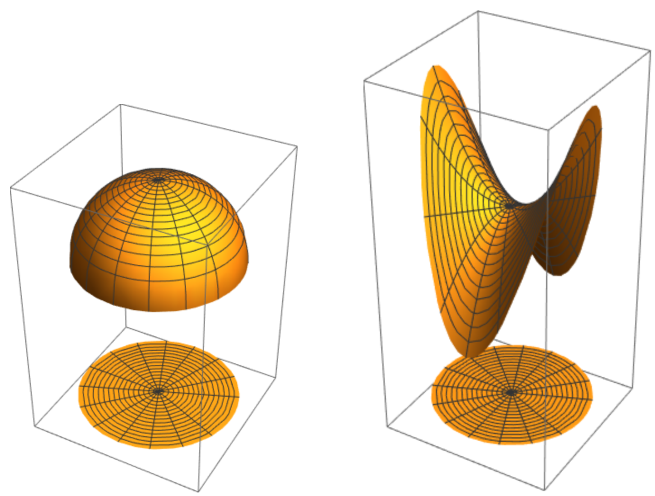
Euclidean: lengths of paths between points on two geodesic rays from $X$ at angle $\theta$ at distance $d$ increases linearly in $d$:




Hyperbolic: lengths of paths between points on two geodesic rays from $X$ at angle $\theta$ at distance $d$ outside the ball increases exponentially



## Projections from non-positive to Euclidean have distortion



# Capturing sets of trees via dimensionality reduction

Hillis, Heath, St. John (2005)
Given a set of trees $T_i$:

- Can construct pair-wise distance matrix wrt. preferred metric
- Embed this approximately in high-dimensional Euclidean space
- Take traditional Euclidean MDS to project to lower-dimensional space that best captures variances

Advantages: uses effective tools, gives valuable visualizations
Disadvantages: assumes a Euclidean structure

## Finding geodesics capturing the first principal component

Given a set of trees $\{T_i\}$:

- Nye (2011, 2014): efficient algorithm to find endpoints of a geodesic $\gamma$ where projections via BHV onto $\gamma$ are maximally correlated with pairwise BHV distances
- Feragen, Owen (2013, 2015): generalized Nye's algorithm for medical imaging of airway trees in healthy and diseased lungs
- Cleary, Feragen, Owen, Vargas (2015, 2020): iterate projections onto geodesics
- Nye et. al (2017): projections onto 2-dimensional subsurfaces of BHV space.

## Iterated projections

Given a set of trees $\{T_i\}$:

- Find $\gamma$ the geodesic best capturing the projected distances, parameterize via $[0, 1]$.
- For each $T_i$, find point on $\gamma$ closest, to give coordinates $t_i$
- Bin the trees into subsets $S_j$ by coordinates $t_i$
- Repeat the analysis on the smaller subsets $S_j$

Lives a collection of piecewise projections
Can iterate to get subsets $S_{j,k,\dots}$

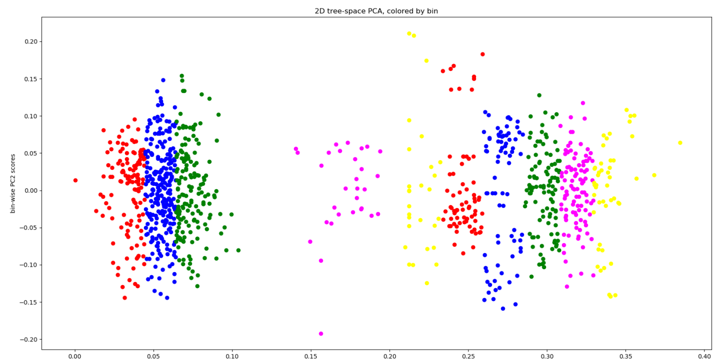
## Two dimensional case

Given a set of trees $\{T_i\}$:

- Find $\gamma$, backbone of the entire collection
- Bin trees into say 10 subsets along projections onto $\gamma$
- Find $\gamma_0$ to $\gamma_9$, geodesics for those 10 subsets
- Give coordinates $(x, y)$ to trees based upon their projections onto $\gamma$ and the relevant $\gamma_i$
- Use these coordinates to plot and visualize data

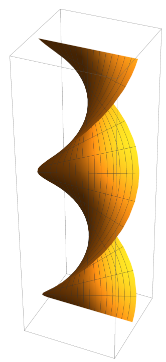
# Example

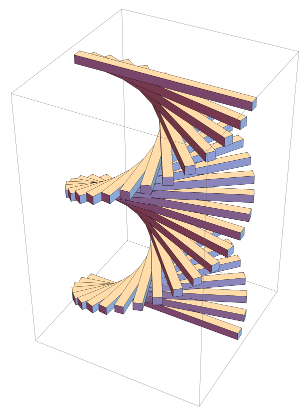
VL3 dataset: 442 trees on 50 taxa, GARTFase Beiko et al. (2006), alignments bacterial and archaeal sequences of protein-coding genes



## Piecewise approximation

Approximate helicoid via backbone and perpendicular parts of changing direction:



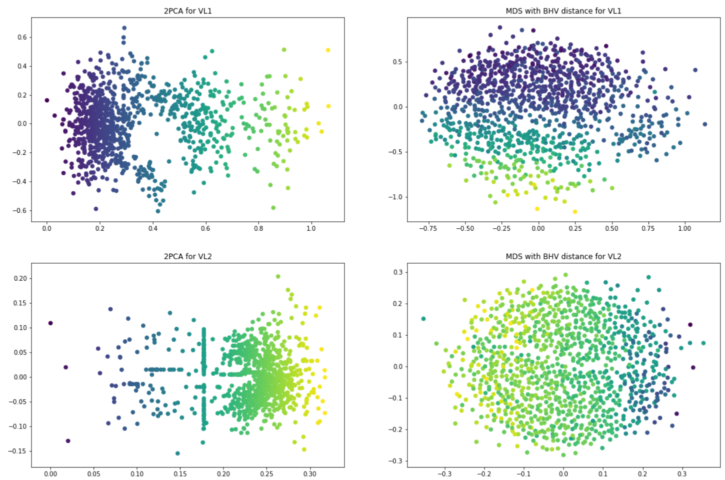


## Properties

- Projections are optimized locally on subsets, more robustness with respect to far off points
- Choices about orientation made to maximize correlation
- Uses: cluster identification, visualizing distribution

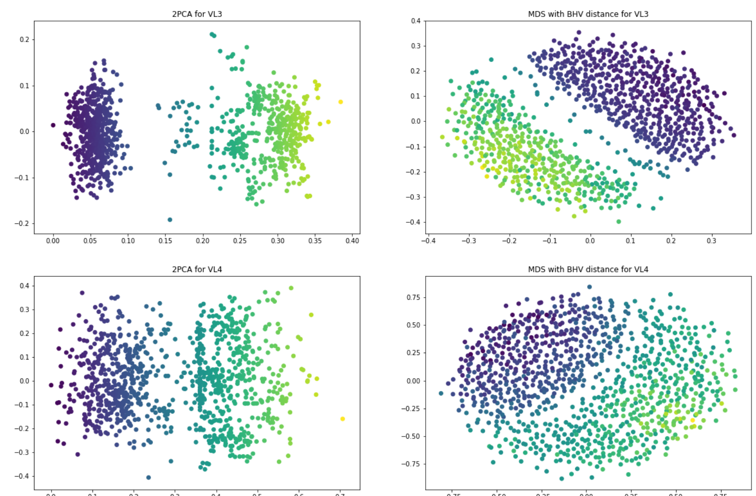
# Examples

Colored by first PCA coordinates:



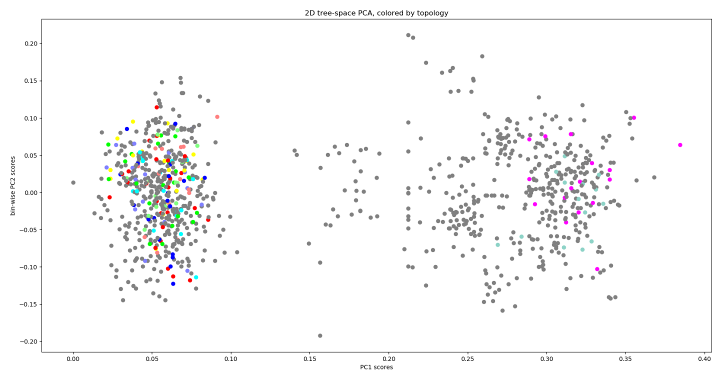
# Examples

Colored by first PCA coordinates:



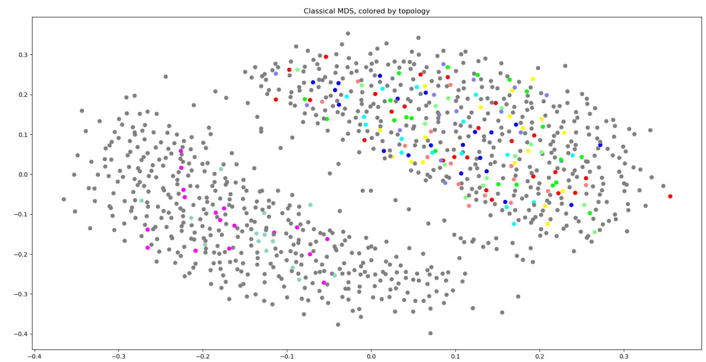
# Common topologies in 2PCA

VL3 dataset, Tree PCA colored by topology:



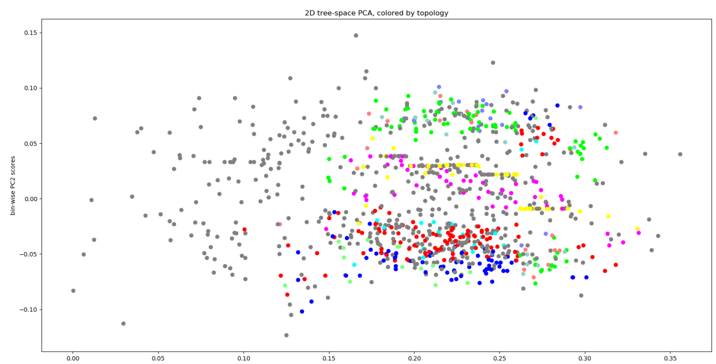
# Common topologies in MDS

VL3 dataset, MDS, colored by topology:

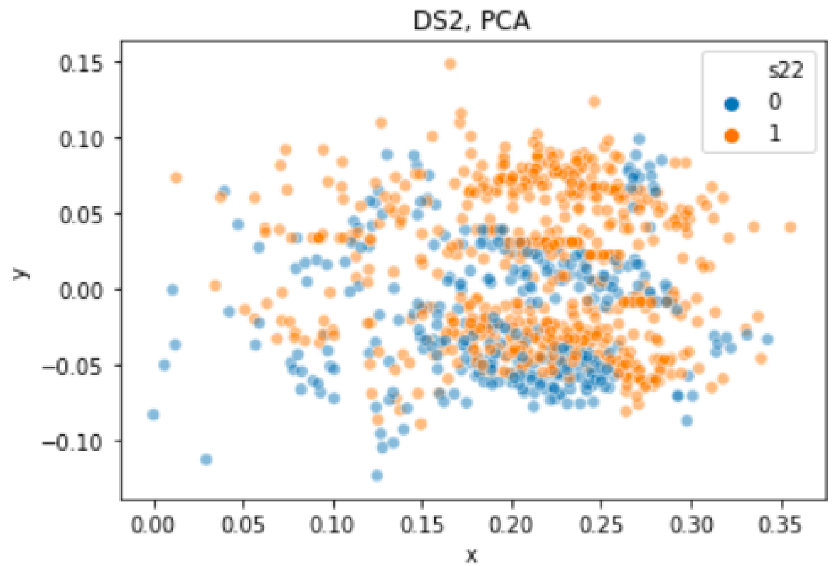


# Common topologies in Tree PCA

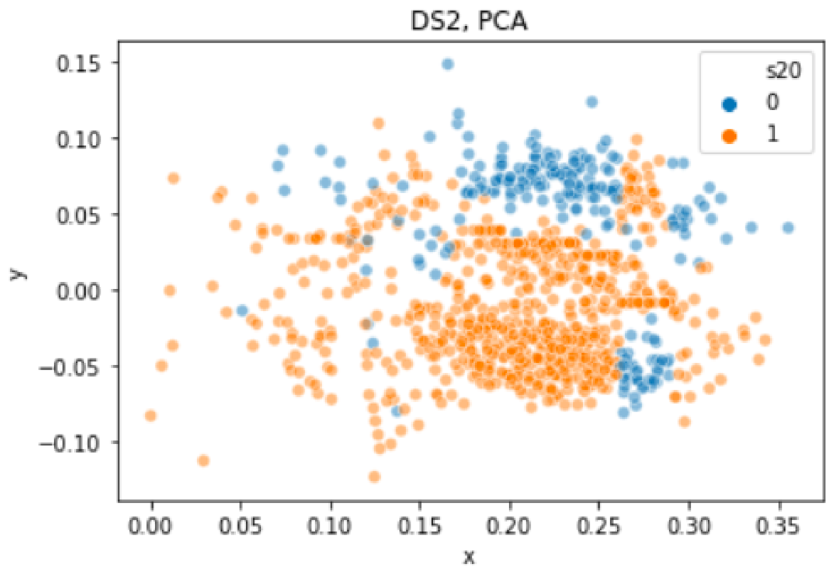
DS2 dataset: 2520 trees on 29 taxa, rDNA; 18s Garey et al. (2012) , sequences from eukaryote species, colored by topology:



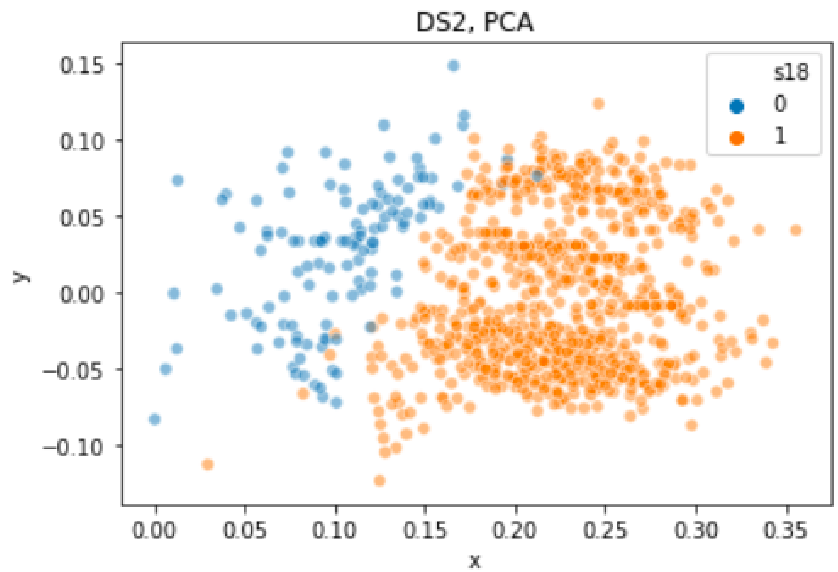
## Specific splits: edge 22 present



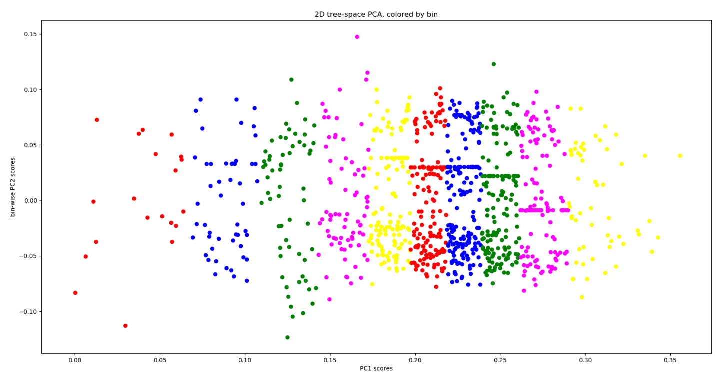
# Specific splits: edge 20 present



## Specific splits: edge 18 present

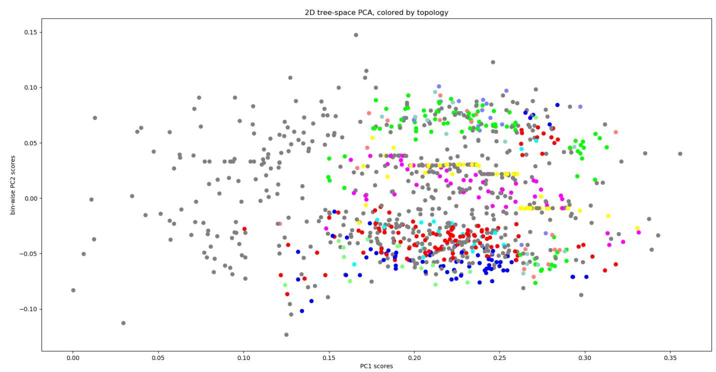


# DS2 by bin

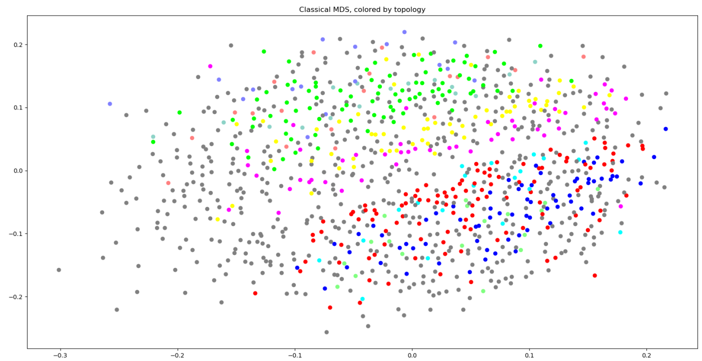


2D tree-space PCA, colored by bin
0.15
0.10
0.05
0.00
−0.05
−0.10
bin-wise PC2 scores
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
PC1 scores

# DS2 by topology



# DS2 by topology



## Observations:

Observations and concerns:

- Number of bins: preset or selected by KDE methods
- Danger of overfitting with many bins
- No guarantee second PCA is perpendicular to first PCA
- Third PCA generally sparse data to fit

## Contributors

Aasa Feragen:  Megan Owen: 

Treespace Working Group:

`sites.google.com/site/treespaceworkinggroup/`