Principal Components Analysis in Tree Space

Tom M. W. Nye, Newcastle University

tom.nye@ncl.ac.uk

Isaac Newton Institute, 21 June 2011
Summary

The problem

- Take a sample of trees on a fixed set of taxa
  - gene trees
  - bootstrap or posterior sample
- How can you characterize the data set? Quantify variability?
- Standard tools of multivariate analysis: clustering, Multi-Dimensional Scaling (MDS), Principal Components Analysis (PCA)
- Trees aren’t vectors – so how will PCA work?

A solution

- Regard PCA as a geometrical procedure
- Re-frame this procedure in tree space
Summary

The problem

- Take a sample of trees on a fixed set of taxa
  - gene trees
  - bootstrap or posterior sample

- How can you characterize the data set? Quantify variability?
- Standard tools of multivariate analysis: clustering, Multi-Dimensional Scaling (MDS), Principal Components Analysis (PCA)
- Trees aren’t vectors – so how will PCA work?

A solution

- Regard PCA as a geometrical procedure
- Re-frame this procedure in tree space
Data thanks to Anne Zupczok and Arndt von Haeseler
Why PCA?

Potentially have 1000’s of trees on 100’s of species. How do the data vary?

Principal components analysis:-

- Which features vary the most?
- How are features correlated?
- Quantification: proportion of variance explained

Analysing tree geometry (rather than topology) has advantages:-

- Geometry and topology interdependent
- Differences in geometry offer finer resolution of comparison
Why PCA?

Potentially have 1000’s of trees on 100’s of species. How do the data vary?

Principal components analysis:-

- Which features vary the most?
- How are features correlated?
- Quantification: proportion of variance explained

Analysing tree geometry (rather than topology) has advantages:-

- Geometry and topology interdependent
- Differences in geometry offer finer resolution of comparison
Why PCA?

Potentially have 1000’s of trees on 100’s of species. How do the data vary?

Principal components analysis:-

- Which features vary the most?
- How are features correlated?
- Quantification: proportion of variance explained

Analysing tree geometry (rather than topology) has advantages:-

- Geometry and topology interdependent
- Differences in geometry offer finer resolution of comparison
A Naive Approach

A split is a bi-partition of the leaves induced by cutting a branch:

![Diagram of a tree with branches AB and CDE]

Trees consist of sets of compatible splits
- e.g. $AB|CDE$ and $AC|BDE$ cannot both be in a tree

Embed tree-space in $\mathbb{R}^M$:-

- On $m$ leaves there are $M = (2^{m-1} - 1)$ possible splits
- Associate each split with a basis vector of $\mathbb{R}^M$ and represent tree as vector of branch lengths
- Each tree contains at most $2m - 3 \ll M$ splits

Principal components in $\mathbb{R}^M$ usually do not correspond to trees
A Naive Approach

A split is a bi-partition of the leaves induced by cutting a branch:

Trees consist of sets of compatible splits
e.g. $AB|CDE$ and $AC|BDE$ cannot both be in a tree

Embed tree-space in $\mathbb{R}^M$:

- On $m$ leaves there are $M = (2^{m-1} - 1)$ possible splits
- Associate each split with a basis vector of $\mathbb{R}^M$ and represent tree as vector of branch lengths
- Each tree contains at most $2m - 3 \ll M$ splits

Principal components in $\mathbb{R}^M$ usually do not correspond to trees
Geometry of PCA

1. Find centre of data (point which minimizes sum of squared distances)
2. Consider line through centre
3. Project points onto line
4. Identify line that maximizes variance (or minimizes sum of squared distances)
1. Find centre of data (point which minimizes sum of squared distances)
1. Find centre of data (point which minimizes sum of squared distances)
2. Consider line through centre
1. Find **centre** of data (point which minimizes sum of squared distances)
2. Consider line through centre
3. **Project** points onto line
1. Find centre of data (point which minimizes sum of squared distances)
2. Consider line through centre
3. Project points onto line
4. Identify line that maximizes variance (or minimizes sum of squared distances)
Topological Decomposition of Tree-Space (1)

Trees with same fixed topology represented as points in positive orthant of $\mathbb{R}^{m-3}$ (ignore leaves)
e.g. $m = 5$ taxa represented by quadrant in $\mathbb{R}^2$
Topological Decomposition of Tree-Space (2)

Orthants are stitched together along their faces

Nearest neighbour interchange:
Billera, Holmes, Vogtmann (2001) proved existence of the geodesic metric

- Consider paths composed of straight line segments in each orthant
- Define length of path to be sum segment of lengths

Then:

- The geodesic between two points $x, y$ is the shortest path between points
- Geodesic metric $d(x, y) = \text{length of geodesic}$
- Within each orthant, geodesic metric $=$ Euclidean metric

Calculation:

- Anne Kupczok and Megan Owen presented algorithms (2007)
- Owen and Provan (2009) developed $O(m^3)$ algorithm
Geodesic Metric

Billera, Holmes, Vogtmann (2001) proved existence of the geodesic metric

- Consider paths composed of straight line segments in each orthant
- Define length of path to be sum segment of lengths

Then:

- The geodesic between two points $x, y$ is the shortest path between points
- Geodesic metric $d(x, y) = \text{length of geodesic}$
- Within each orthant, geodesic metric $= \text{Euclidean metric}$

Calculation:

- Anne Kupczok and Megan Owen presented algorithms (2007)
- Owen and Provan (2009) developed $O(m^3)$ algorithm
Billera, Holmes, Vogtmann (2001) proved existence of the geodesic metric

- Consider paths composed of straight line segments in each orthant
- Define length of path to be sum segment of lengths

Then:-

- The geodesic between two points \( x, y \) is the shortest path between points
- Geodesic metric \( d(x, y) = \) length of geodesic
- Within each orthant, geodesic metric = Euclidean metric

Calculation:-

- Anne Kupczok and Megan Owen presented algorithms (2007)
- Owen and Provan (2009) developed \( O(m^3) \) algorithm
Example
ΦPCA: PCA with the geodesic metric

For first principal component:-

1. Find centre $x_0$ of data $x_1, \ldots, x_n$
2. Consider each line through $x_0$:
   - line $L$ is geodesic between all points $x, y \in L$
   - lines extend to infinity in two directions
3. Project $x_1, \ldots, x_n$ onto line by finding points $y_i$ on geodesic closest to $x_i$
4. Identify line that maximizes variance of $y_1, \ldots, y_n$ (or minimizes $\sum d(x_i, y_i)^2$)
Finding center $x_0$:-

- Cannot ‘add’ trees – so cannot compute a mean
- Minimizing $\sum d(x_0, x_i)^2$ computationally expensive
- Use the majority consensus tree

Projection:-

- Find closest point on $L$ to each point $x_i$
- Numerical search made efficient by
  - Euclidean approximation
  - shift end points slightly – geodesic doesn’t change much
Searching for the Optimal Line

No analytic solution – need to search for optimal line

- Geometrical problem: direction vector in initial quadrant
- Topological problem: which orthants?

ΦPCA algorithm:-

- Add one split $p$ and its NNI replacement $p'$ at a time
  - $p, p'$ must satisfy compatibility constraint with $L$
  - optimize direction vector given splits $p, p'$
  - for each pair $p, p'$ and direction vector, perform projection

- Analogous to one coord at a time in regular PCA

- Search algorithms for $p, p'$:
  - greedy approach
  - simulated annealing: birth / death moves
Searching for the Optimal Line

No analytic solution – need to search for optimal line

- Geometrical problem: direction vector in initial quadrant
- Topological problem: which orthants?

ΦPCA algorithm:-

- Add one split $p$ and its NNI replacement $p'$ at a time
  - $p, p'$ must satisfy compatibility constraint with $L$
  - optimize direction vector given splits $p, p'$
  - for each pair $p, p'$ and direction vector, perform projection

- Analogous to one coord at a time in regular PCA

- Search algorithms for $p, p'$:
  - greedy approach
  - simulated annealing: birth / death moves
Results

1. Application to metazoan (animal) data set
2. Long branch attraction simulation example

Metazoan data set (von Haeseler group, J. Comp. Biol. 2008)

- 118 gene trees from 20 metazoan species
Metazoan Data Set

c.elegans
c.briggsae
c.anopheles
c.dmelanogaster
c.dpseudoobscura
c.apis
c.danio
c.takifugu
c.tetraodon
c.xtropicalis
c.gallus
c.monodelphis
c.mmusculus
c.rnorvegicus
c.btaurus
c.canis
c.mmulatta
c.human

c.pantro
c.ciona

cyeast
Metazoan Data Set

c.elegans
c.briggsae
can.
d.melanogaster
d.pseudoobscura
a.pis
d.anopheles
d.takifugu
d.tetraodon
x.tropicalis
gallus
m.musculus
r.norvegicus
b.taurus
c.anis
c.human
p.patto
m.mulatta
yeast

ciona
Metazoan Data Set

c. elegans
c. briggsae
c. anopheles
D. melanogaster
D. pseudoobscura
A. apis
d. danio
t. takifugu
t. tetraodon
x. tropicalis
G. gallus
M. monodelphis
M. musculus
M. norvegicus
B. taurus
C. canis
M. mulatta
H. sapiens
P. troglodytes
C. intestinalis
S. cerevisiae
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set

c.elegans
c.briggsae
c.anopheles
d.melanogaster
d.pseudoobscura
aapis
danio
takifugu
tetraodon
x.tropicalis
gallus
monodelphis
m.musculus
r.norvegicus
b.taurus
c.canis
m.mulatta
human
pantro
ciona
yeast
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set
Metazoan Data Set *

- *celegans*
- *c. briggsae*
- *yeast*
- *apis*
- *d. melanogaster*
- *d. pseudoobscura*
- *anopheles*
- *ciona*
- *danio*
- *takifugu*
- *tetraodon*
- *b. taurus*
- *c. canis*
- *p. pantro*
- *human*
- *m. mulatta*
- *m. musculus*
- *r. norvegicus*
- *m. domestica*
- *gallus*
- *x. tropicalis*
When the true tree contains long branches next to short branches, long branches are erroneously placed together in estimated phylogenies

Simulation:-

- Simulate 100 DNA alignments from ‘true tree’ (genes)
- Estimate trees from alignments (gene trees)
- Apply ΦPCA to resultant collection of alternative trees
Long Branch Attraction Tree

LBA Simulation
LBA Simulation
LBA Simulation
LBA Simulation
LBA Simulation
LBA Simulation
LBA Simulation
LBA Simulation
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a **geometrical procedure** in tree-space
- Incorporates geometrical and topological information about trees via the geodesic metric
- **Computational feasibility** is achieved by greedy or stochastic search for optimal line
- Method successfully identifies the most variable features and correlations between features
- Captures known variability in experimental data and reveals long branch attraction in simulated data
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a **geometrical procedure** in tree-space
  - Incorporates geometrical and topological information about trees via the geodesic metric
  - **Computational feasibility** is achieved by greedy or stochastic search for optimal line
  - Method successfully identifies the most variable features and correlations between features
  - Captures known variability in experimental data and reveals long branch attraction in simulated data
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a geometrical procedure in tree-space
- Incorporates geometrical and topological information about trees via the geodesic metric
- Computational feasibility is achieved by greedy or stochastic search for optimal line
- Method successfully identifies the most variable features and correlations between features
- Captures known variability in experimental data and reveals long branch attraction in simulated data
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a **geometrical procedure** in tree-space
- Incorporates geometrical and topological information about trees via the geodesic metric
- **Computational feasibility** is achieved by greedy or stochastic search for optimal line
- Method successfully identifies the most variable features and correlations between features
- Captures **known variability** in experimental data and reveals long branch attraction in simulated data
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a **geometrical procedure** in tree-space
- Incorporates geometrical and topological information about trees via the geodesic metric
- **Computational feasibility** is achieved by greedy or stochastic search for optimal line
- Method successfully identifies the most variable features and correlations between features
- Captures known variability in experimental data and reveals long branch attraction in simulated data
Summary

How can you understand variability in large sets of trees?

- Attempt to re-interpret PCA as a geometrical procedure in tree-space
- Incorporates geometrical and topological information about trees via the geodesic metric
- Computational feasibility is achieved by greedy or stochastic search for optimal line
- Method successfully identifies the most variable features and correlations between features
- Captures known variability in experimental data and reveals long branch attraction in simulated data
Further work:

- Higher components: approximation by ‘planes’
- Better theory of distributions on tree space

Thanks to:

- Anne Kupczok and Arndt von Haeseler
- Two anonymous reviewers