Graphical inequality diagnostics for phylogenetic trees

Nathaniel Shiers, Jim Q. Smith

Department of Statistics, University of Warwick, UK

THE UNIVERSITY OF WARWICK

Background

Conditional inequality statements encoded in graphical models are known to produce active constraints on the parameters of the models. Zwiernik and Smith (2011) fully characterised the inequality constraints for binary phylogenetic trees with manifest leaf variables and hidden interior variables.

Aim

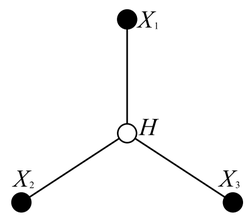
To demonstrate two graphical diagnostics complementary to existing techniques:

- A diagnostic which checks whether data appears to fit with any tree structure.
- A diagnostic which utilises functions of associated statistics to guide candidate tree model selection.

Constraints on a tripod tree

The inequality constraints associated with the tripod tree are of particular interest, as these inequalities must hold for any triple of manifest variables in a strictly trivalent tree (that is trees where all hidden vertices have degree 3). Using the notation of Zwiernik and Smith (2011, Proposition 2.5) we note that an observed joint probability table $\boldsymbol{P}$ $(\mathbf{2 \times 2 \times 2})$ is consistent with the tripod tree structure if and only if:
$\boldsymbol{\mu_{123} = 0}$ and at least two of $\boldsymbol{\mu_{12}}$, $\boldsymbol{\mu_{13}}$, $\boldsymbol{\mu_{23}}$ vanish ***or***

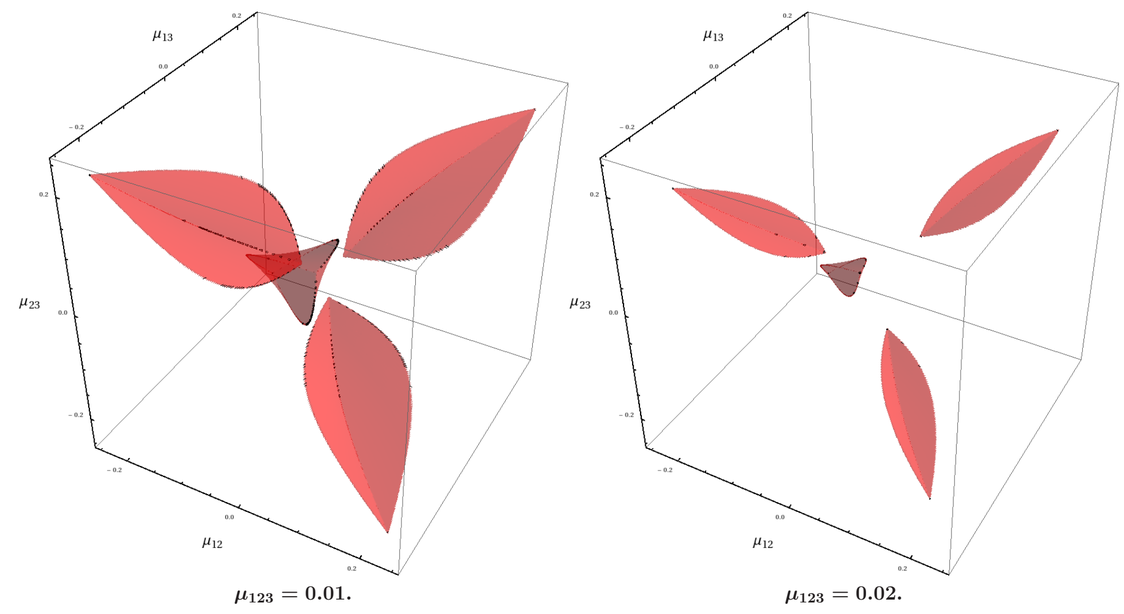
$$\begin{aligned} \boldsymbol{\mu_{12}\mu_{13}\mu_{23}} &> \mathbf{0} \\ |\boldsymbol{\mu_{jk}}|\sqrt{\det(\boldsymbol{P})} + \boldsymbol{\mu_{123}\mu_{jk}} &\leq (\mathbf{1} + \bar{\boldsymbol{\mu}}_{\boldsymbol{i}})\boldsymbol{\mu}_{\boldsymbol{jk}}^{\mathbf{2}} \\ |\boldsymbol{\mu_{jk}}|\sqrt{\det(\boldsymbol{P})} - \boldsymbol{\mu_{123}\mu_{jk}} &\leq (\mathbf{1} - \bar{\boldsymbol{\mu}}_{\boldsymbol{i}})\boldsymbol{\mu}_{\boldsymbol{jk}}^{\mathbf{2}} \end{aligned}$$



where $\mathbf{det}(\boldsymbol{P}) = \boldsymbol{\mu}_{\mathbf{123}}^{\mathbf{2}} + \mathbf{4}\boldsymbol{\mu_{12}\mu_{13}\mu_{23}}$, $\bar{\boldsymbol{\mu}}_{\boldsymbol{i}} = \mathbf{1} - \mathbf{2}\boldsymbol{\lambda_i}$ for all $\boldsymbol{i} \in \{\mathbf{1,2,3}\}$, $\boldsymbol{\mu}$ are central moments, and $\boldsymbol{\lambda}$ means.

Admissible covariance regions

We can view these admissible regions (indicated below in red) by fixing $\boldsymbol{\mu_{123}}$ and $\bar{\boldsymbol{\mu}}_{\boldsymbol{i}}$, and then setting the covariances as the axes in three-dimensional space.



$\boldsymbol{\mu_{123}} = \mathbf{0.01}$. $\boldsymbol{\mu_{123}} = \mathbf{0.02}$.

Theorem 1

The constraints that can be used for a tree diagnostic can be derived simply in the following graphical way:

Given any strictly trivalent phylogenetic tree $\boldsymbol{T}$, for all triples $\boldsymbol{X_i, X_j, X_k}$ there exists a unique hidden variable $\boldsymbol{H_{ijk}}$ such that $\boldsymbol{H_{ijk}}$ separates $\boldsymbol{(X_i, X_j, X_k)}$ in $\boldsymbol{T}$. i.e. $\perp\!\!\!\perp\ \boldsymbol{(X_i, X_j, X_k)|H_{ijk}}$

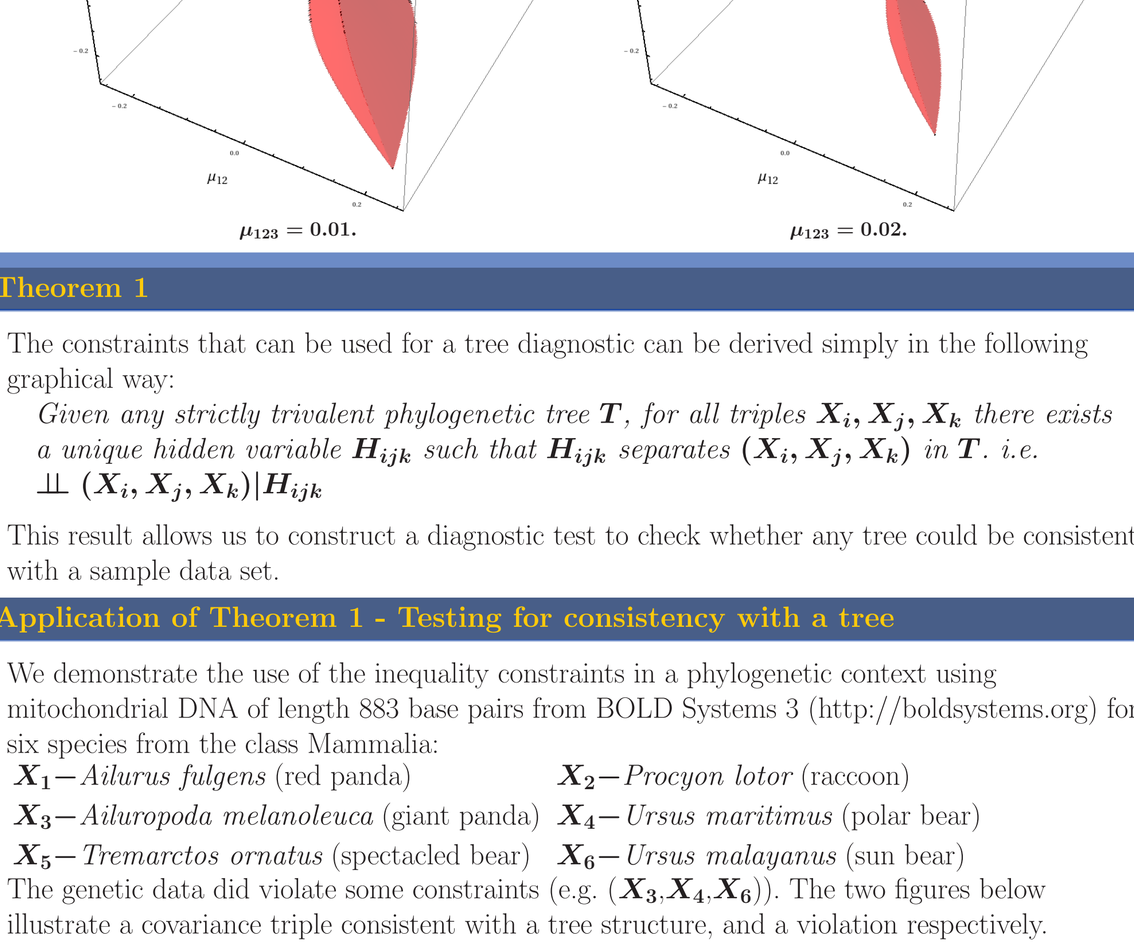
This result allows us to construct a diagnostic test to check whether any tree could be consistent with a sample data set.

Application of Theorem 1 - Testing for consistency with a tree

We demonstrate the use of the inequality constraints in a phylogenetic context using mitochondrial DNA of length 883 base pairs from BOLD Systems 3 (http://boldsystems.org) for six species from the class Mammalia:

$\boldsymbol{X_1}$—*Ailurus fulgens* (red panda) $\boldsymbol{X_2}$—*Procyon lotor* (raccoon)
$\boldsymbol{X_3}$—*Ailuropoda melanoleuca* (giant panda) $\boldsymbol{X_4}$—*Ursus maritimus* (polar bear)
$\boldsymbol{X_5}$—*Tremarctos ornatus* (spectacled bear) $\boldsymbol{X_6}$—*Ursus malayanus* (sun bear)

The genetic data did violate some constraints (e.g. ($\boldsymbol{X_3}$,$\boldsymbol{X_4}$,$\boldsymbol{X_6}$)). The two figures below illustrate a covariance triple consistent with a tree structure, and a violation respectively.



(a) Triple 125 (b) Triple 346

Figure: Point estimates of covariances.

A single violation does not automatically reject a tree structure, as sample errors can cause false positives. A discussion of sample size can be found in Shiers and Smith (2012).

Theorem 2

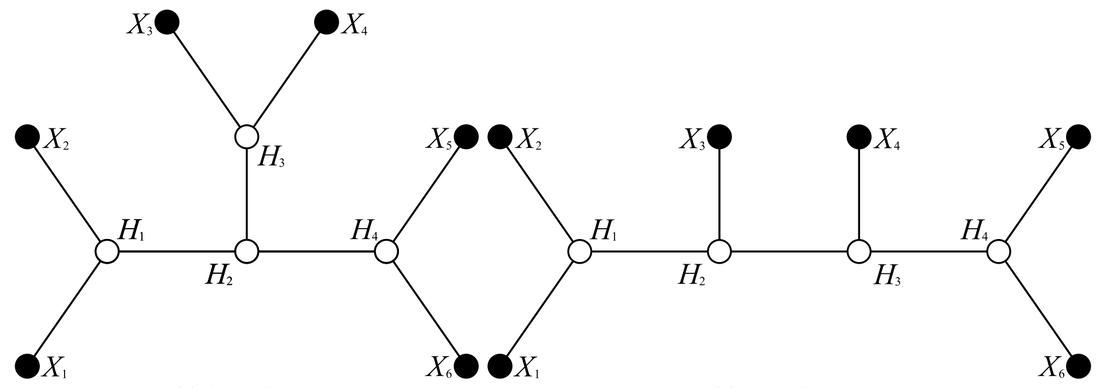
Note for every hidden variable $\boldsymbol{H} \in \boldsymbol{\mathcal{H}}$ of a strictly trivalent tree $\boldsymbol{\mathcal{T}}$, there is a partition $\boldsymbol{\Lambda(H, \mathcal{T})}$ of the manifest variables into 3 subsets, each subset being the leaves of a subtree rooted at $\boldsymbol{H}$. Usefully, these partitions uniquely define a tree $\boldsymbol{\mathcal{T}}$. Thus we have:

Each strictly trivalent tree $\boldsymbol{\mathcal{T}}$ is uniquely identified by its set of partitions and $\boldsymbol{\mathcal{X}(\mathcal{T})} \triangleq \boldsymbol{\{\Lambda(H, \mathcal{T}) : H \in \mathcal{H}\}}$ acts as an identifier, under the assumption of faithfulness (see Spirtes et al. (2001)).

Thus, if there exists a unique phylogenetic tree, then the estimated moments of the triple will identify it.

Application of Theorem 2 - Inferring trees from moments

We simulated binary data from the two non-isomorphic strictly trivalent trees with 6 leaves. The probability distributions of the graphs were simulated so expected means for each $\boldsymbol{X_i}$ were the same across graphs.



(a) 6-leaf Tree I (b) 6-leaf Tree II

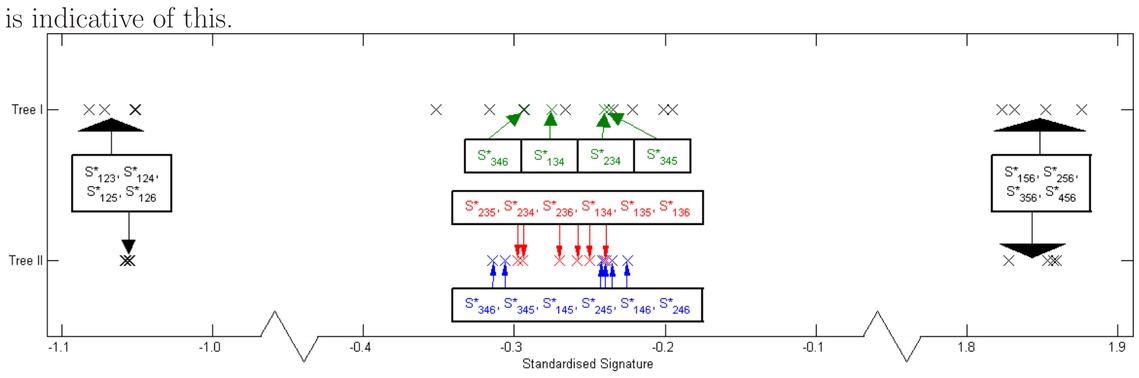
The first 3 moments provide us with consistent but inefficient estimates of a tree. In Settimi and Smith (1998) it was shown that (provided none of the terms is zero) for any triple $\boldsymbol{X_i}$, $\boldsymbol{X_j}$, $\boldsymbol{X_k}$ with $\boldsymbol{(X_i \perp\!\!\!\perp X_j \perp\!\!\!\perp X_k)|H_{ijk}}$

$$\boldsymbol{S_{ijk}} = \mathbf{ln}(|\boldsymbol{\mu_{ij}}|) + \mathbf{ln}(|\boldsymbol{\mu_{ik}}|) + \mathbf{ln}(|\boldsymbol{\mu_{jk}}|) - \mathbf{2}\,\mathbf{ln}(|\boldsymbol{\mu_{ijk}}|)$$

where $\boldsymbol{S_{ijk}}$ (the **signature**) depends only on the margin distributions of the triple.
From Theorem 2, for large enough datasets the signatures of the corresponding sample quantities will indicate candidate tree partitions $\boldsymbol{\mathcal{X}(\mathcal{T})}$ and hence, from the theorem, candidates $\boldsymbol{\mathcal{T}}$. Note that triples $\boldsymbol{(i, j, k)}$ and $\boldsymbol{(i', j', k')}$ share the same separator $\boldsymbol{H}$ in $\boldsymbol{\mathcal{T}}$ if the pairs $\boldsymbol{(i, i')}$, $\boldsymbol{(j, j')}$ and $\boldsymbol{(k, k')}$ all lie in different subsets in $\boldsymbol{\Lambda(H, \mathcal{T})}$. So these $\boldsymbol{(n-2)}$ partitions can be calculated by first clustering the signatures by magnitude into up to $\boldsymbol{(n-2)}$ clusters and from this deducing $\boldsymbol{\Lambda(H, \mathcal{T})}$.

Standardised signature plot

The signature plot below shows the standardised signatures $\boldsymbol{S^*_{ijk}}$ for both non-isomorphic trees 6-leafed tree with the signatures of interest labelled. The clustering of the signatures demonstrates the power of the diagnostic - there is remarkably clear separation of all $\boldsymbol{S^*_{12k}}$ and $\boldsymbol{S^*_{i56}}$ which supports the topologies of the trees. The signatures involving **3** and **4** are less distinct, but this may be in part unavoidable overlapping of true clusters. However, trees like Tree I have an interior node and it can be shown that this leads to signatures with a higher variance. The wide spread in Tree I of the middle cluster (overlapping) with the $\boldsymbol{X_3}$ $\boldsymbol{X_4}$ cluster is indicative of this.



Plot of standardised signatures for Tree I and Tree II.

Summary & remarks

- Two graphical diagnostics based on inequality violations and low order sample moments.
- Simple to calculate, and scale-up to larger trees.
- Complementary to existing techniques.
- Useful in the preliminary stages of phylogenetic analysis to check consistency with a tree.
- Can be used to identify starting points for a tree-model searches.
- If non-tree structure detected, possible to identify subset of variables consistent with tree.
- There exist (albeit, in general not yet determined) equivalent inequalities for non-binary data, and in a similar manner alternative 'signatures'.

Future work

- Development of a general theory associated with sample sizes and with interpretation of signature plots for first and second diagnostics respectively.
- Development of analogous inequality diagnostics using the same graphical properties but for differently distributed variables.
- Application to functional acoustic language data in a language phylogenetic tree context.

References

Settimi, R and J Q Smith. 1998. On the geometry of bayesian graphical models with hidden variables. In *Proc. 14th Conf. Uncertain. Artif. Intell.*. Morgan Kaufmann, pages 472–479.

Shiers, N and J Q Smith. 2012. Graphical inequality diagnostics for phylogenetic trees. In *Proc. 6th European Workshop Probabilistic Graph. Model. (to appear)*.

Spirtes, P, C Glymour, and R Scheines. 2001. *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Zwiernik, P and J Q Smith. 2011. Implicit inequality constraints in a binary tree model. *Electron. J. Stat.* 5:1276–1312.

We gratefully acknowledge the referees of the corresponding paper (Shiers and Smith (2012)) for their thorough and valuable feedback.