Graphical inequality diagnostics for phylogenetic trees

Nathaniel Shiers
University of Warwick, UK
n.l.shiers@warwick.ac.uk

Jim Q. Smith
University of Warwick, UK
j.q.smith@warwick.ac.uk

Abstract

Phylogenetic trees can be studied as discrete graphical models with interior vertices hidden and leaves observed. In the case of binary random variables it has long been known that this class of models constrains the parameter space of the leaf margin. However, it was only last year that the inequality constraints were fully characterised for binary phylogenetic trees. The inferential consequences of these implicit inequality constraints have so far attracted little attention. However, this paper uses a subset of these constraints to demonstrate their usefulness as simple and transparent diagnostics that can be used at the preliminary stages of a phylogenetic analysis, not only to check whether any phylogenetic model is consistent with the data but also to guide the ensuing inference.

1 Review of implicit constraints

The estimation of Bayesian networks when interior vertices are hidden is challenging since the implicit geometry of the associated probability mass function of the observed variables, and hence also the likelihood, is complicated.

The simplest such model is the phylogenetic tree where all variables are binary. A phylogenetic tree is a graphical method for describing the evolutionary relationships between species both extant and extinct. The geometry of this space of models was expanded in Settimi and Smith (2000) and has recently attracted considerable interest (for example Allman et al. (2009) and Zwiernik and Smith (2012)). These advances have encouraged some authors to proceed to use the known geometry of these spaces to support inference and learning over the space of tree models (see Drton and Sullivant (2007) for example). These focus on the polynomial constraints that are implicit in these models. However, it is well known that not only these functional relationships but also additional inequality constraints are active. Last year Zwiernik and Smith (2011) fully characterised these inequality constraints for binary phylogenetic trees. So we are at last able, at least for this important class of graphical models, to explore the inferential use for learning of these derived inequality constraints.

In this paper we demonstrate how some of these constraints can be used for inference by providing the basis for various diagnostics. These are primarily designed for the early stages of a phylogenetic analysis. The first point we note is that all trees must satisfy certain cubic inequalities associated with all the triples of its observed variables. A simple diagnostic is produced which therefore simply checks whether data appears to fit with any tree structure or whether some of these inequalities appear to be violated. For if these constraints are significantly broken then we know that no tree will fit the data well and a search for a more elaborate graphical model of the data is needed. We also indicate how functions of the associated statistics can be used as a guide to which candidate tree models are likely to score highly in model selection. In this way, these preprocessing tools can be used to

CRiSM Paper No. 12-11v2, www.warwick.ac.uk/go/crism

help speed up and stabilise numerical scoring methods over the class.

A phylogenetic tree is often studied as the pattern (see Verma and Pearl (1990)) of a particular class of Bayesian networks, specifically trees (e.g. Dutkowski and Tiuryn (2007)). This paper will consider trees labelled by binary random variables each taking a value of 0 or 1 (for the results we use, we only require that the observed varaibles are binary). The variables on the trees considered will be classified into two sets: the $m-2$ hidden nodes $H \in \mathcal{H}$ and m manifest nodes $X \in \mathcal{X}$ (denoted by white and black circles respectively). We will consider trees with manifest variables as the leaves and all remaining variables hidden (i.e. never directly observed). It was shown in Settimi and Smith (2000) that all phylogenetic tree models are contained in the class of strictly trivalent trees, that is trees where all hidden vertices have degree 3. The results of this paper therefore focus on this class. To relate a trivalent tree to a phylogenetic tree rooted on a selected hidden node, a further hidden node can be inserted to produce a bifurcating tree from the root whilst remaining Markov equivalent. In Settimi and Smith (2000) it is shown that all the conditional independence statements implicit in a tree in terms of algebraic functions in their moments can be equivalently re-expressed in terms of conditions on moments. However, it is the novel parametrisation to tree cumulants by Zwiernik and Smith (2012, Section 3)) that allowed the authors to specify explicitly and elegantly all the constraints and conditions imposed by a strictly trivalent tree structure (Zwiernik and Smith (2011, Theorem 4.7)).

1.1 Constraints on a tripod tree

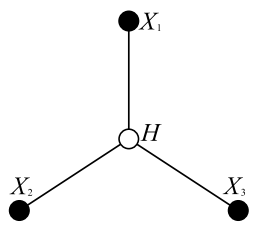


Figure 1: Tripod tree.

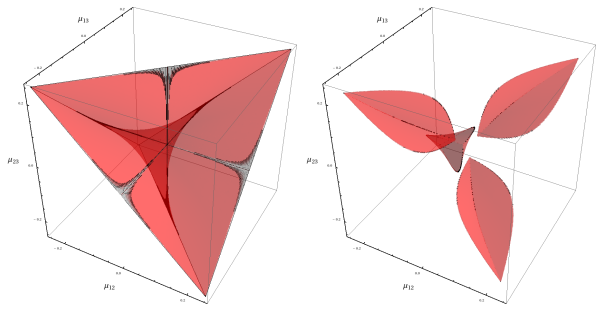
The inequality constraints imposed by the implicit conditional independence on the tripod tree with binary random variables X_1, X_2, X_3 (Figure 1), were originally given in Settimi and Smith (1998). Here we adopt the equivalent formulation described in Zwiernik and Smith (2011, Proposition 2.5). Using the definition of k^{th} central moment as $\mathbb{E}[(X - \mathbb{E}[X])^k]$ we denote the applicable central moments as μ_{12}, μ_{13}, μ_{23}, μ_{123} (equivalent to tree moments for order 2 and 3), and the means as λ_1, λ_2, λ_3. Then given an observed joint probability table P $(2 \times 2 \times 2)$, the data is consistent with the tripod tree structure if and only if:

(i) $\mu_{123} = 0$ and at least two of the three covariances μ_{12}, μ_{13}, μ_{23} vanish *or*

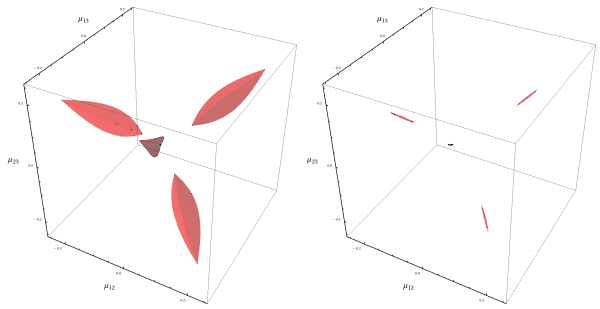
(ii)
$$\mu_{12}\mu_{13}\mu_{23} > 0 \tag{1.1}$$
$$|\mu_{jk}|\sqrt{\det(P)} + \mu_{123}\mu_{jk} \leq (1 + \bar{\mu}_i)\mu_{jk}^2 \tag{1.2}$$
$$|\mu_{jk}|\sqrt{\det(P)} - \mu_{123}\mu_{jk} \leq (1 - \bar{\mu}_i)\mu_{jk}^2 \tag{1.3}$$

where $\det(P) = \mu_{123}^2 + 4\mu_{12}\mu_{13}\mu_{23}$ and $\bar{\mu}_i = 1 - 2\lambda_i$ for all $i \in \{1, 2, 3\}$. We can view these admissible regions by fixing μ_{123} and $\bar{\mu}_i$, and then setting the covariances as the axes in three-dimensional space. Fixing $\bar{\mu}_i = 0.5$ for all i produces the largest regions of tree admissible values of moments (all else equal). We can then view separate plots for different values of μ_{123}, namely 0, 0.01, 0.02, 0.03 (see Figure 2). It is clear that as μ_{123} increases the admissible regions implied by the tree structure decrease. In fact, in the example above, as $|\mu_{123}|$ increases towards 0.03125 the region reduces to 4 singular points, and beyond this disappears.

The boundary points of these graphical regions typically represent zeros in the marginal tables of each pair comprising a manifest node and H. This is equivalent to having a manifest node equal to H. In the space of the three second-order central moments, the edges represent having zero entries in two tables, and each vertex corresponds to zeros in all three tables. Thus, the boundary points of the geometry of the parameter spaces all relate to varying degrees of degenerate distributions. These degeneracies in hidden variable models have been studied for some time and it is understood that when data does not obey the conditional independence



$\mu_{123} = 0.00$ $\mu_{123} = 0.01$



$\mu_{123} = 0.02$ $\mu_{123} = 0.03$

Figure 2: Admissible regions of a tripod tree.

assumptions of a tree structure, the model with highest likelihood will contain some zeroes. This in turn implies particular functional relationships between the variables must hold (see, for example, de Leeuw et al. (1990)).

It is important to state that if the marginal data on the manifest nodes is inconsistent with a tree model, then in the limit (consistent) sample covariance estimates of the the manifest variables will correspond to points outside of these feasible regions. The geometries of such admissible regions are later illustrated in Section 3.

2 Two useful graphical results

The following theorem which has appeared previously, albeit as a corollary of more general theory in different forms and different contexts, has yet to be exploited but is particularly useful:

Theorem 1. *The constraints we can use for a tree diagnostic can be derived simply in the following graphical way:*

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

See Appendix A for proof.

This result allows us to construct a diagnostic test to check whether any tree could be consistent with a sample data set (see Section 4). This method is not the only means of assessing tree structures (e.g. the retention index Farris (1989)) but has the advantage of being very simple to implement and being grounded in theory. When the diagnostic does not reject the tree class, there is a second way in which the distributions of triples can be used to guide the search for promising candidate models.

First note that, by its definition, associated with every hidden variable $H \in \mathcal{H}$ of a strictly trivalent tree $\mathcal{T}$ is a partition $\Lambda(H, \mathcal{T})$ of the manifest variables into 3 subsets, each subset being the leaves of a subtree rooted at H. Interestingly, these partitions uniquely define a tree $\mathcal{T}$. Thus we have:

Theorem 2. *Each strictly trivalent tree $\mathcal{T}$ is uniquely identified by its set of partitions and*

$$\mathcal{X}(\mathcal{T}) \triangleq \{\Lambda(H, \mathcal{T}) : H \in \mathcal{H}\}$$

acts as an identifier, under the assumption of faithfulness (see Spirtes et al. (2001)).

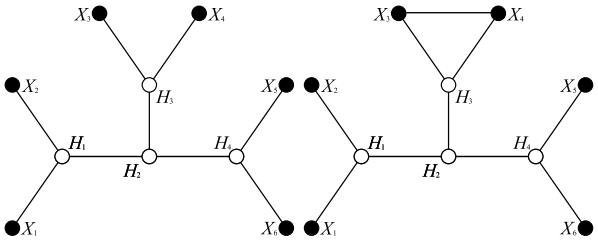
See Appendix B for proof.

Thus if there exists a unique phylogenetic tree, then the estimated moments of the triple will identify it. We will show that this simple result allows us to preselect good candidate trees for model selection. See Section 5 for more details.

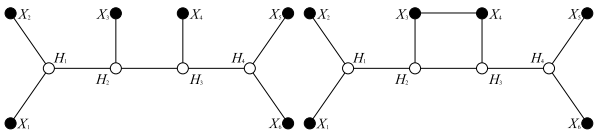
3 An illustration of the constraints

In this section we simulated binary data from the two non-isomorphic strictly trivalent trees with 6 leaves. Then from two other graphical models chosen to mirror typical variations which, for scientific reasons, we might expect of the tree. Unsurprisingly the emprical moments derived from the tree-generated data satisfy all the constraints whilst the emprical moments calculated from the non-tree data violates some of the constraints. The four graphs are shown in Figure 3.

The variations of the two trees were chosen to be similar in order to highlight any other differences. The probability distributions of the graphs were simulated so expected means for each X_i were the same across graphs, thus the



(a) 6-leaf Tree I (b) Non-tree I

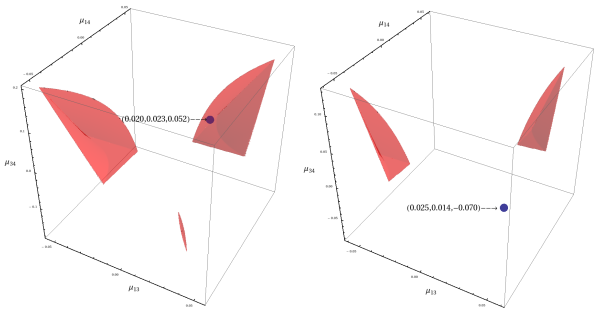


(c) 6-leaf Tree II (d) Non-tree II

Figure 3: 6-leafed trees and non-trees.

differences in the graphs may be expected to be highlighted in the higher order moments.

For large enough samples, the sample moments of the trees always satisfied the constraints demanded of their population analogues whilst the non-tree modifications did not. Both non-trees experienced the same violations: (X_1, X_3, X_4) and (X_2, X_3, X_4) for Inequality (1.2), (X_3, X_4, X_5) and (X_3, X_4, X_6) for Inequality (1.3), and all versions of Inequality (1.1) which involve X_{34} (e.g. $X_{13}X_{14}X_{34} < 0$). Note that all violated constraints include both indices 3 and 4 (where the extra edge was added). Moreover, every inequality involving indices 3 and 4 is broken. This suggests that these diagnostics could provide more insight than a binary acceptance or rejection of the tree structure; the violated constraints in these cases hint that X_3 and X_4 may be responsible for the exceptions.



(a) Triple 134 - Tree I (b) Triple 134 - Non-tree I

Figure 4: Plots of covariance point estimates.

The geometries of the admissible regions of the covariances $\mu_{13}, \mu_{14}, \mu_{34}$ are similar between the trees, and similar between the non-trees. The graphs of the admissible covariance regions are shown for Figure 3a and Figure 3b in Figure 4a and Figure 4b respectively, along with the estimates of the observed covariances plotted. The sample covariances are revisited in Section 6.

4 Application of diagnostics

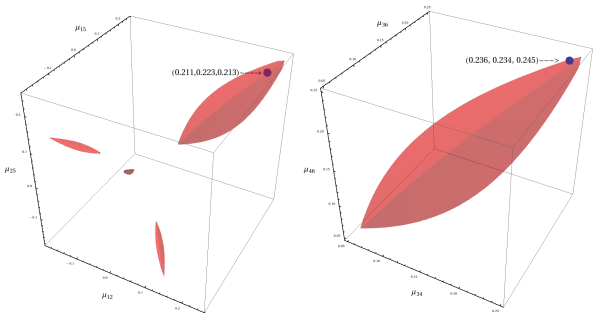
Phylogenetic trees can be constructed using gene sequences where the vectors of data are obtained from DNA sequences coded into binary. Each base in a sequence has one of four chemicals either T or C (pyrimidines) or A or G (purines) (Yang (2007)). We thus encode T and C as 1, and A and G as 0. To illustrate the use of diagnostics we use mitochondrial genetic data each of length 883 base pairs from BOLD Systems 3 (http://boldsystems.org) for six species from the class Mammalia:

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

We here assume that the selected region of DNA is appropriate and use the sampled moments to check whether any tree structure may be appropriate given this data.

The values of the constraints were calculated using the data to estimate the parameters, where crude sample estimates were used. Bahadur expansions can be used to produce estimators of the higher order terms (see Streitberg (1990)).

The genetic data did violate some constraints for five of the triples: (X_4,X_6,X_k) for $k = \{1, 2, 3, 5\}$, and (X_2,X_4,X_5). Figure 5a shows a covariance triple point within the admissible region, and contrastingly Figure 5b shows a different sample covariance triple outside the region associated with a tree. Superficially, this suggests that a tree is not a suitable model for the data. However, because the sample size is not large these violations could be attributed to random error, particularly given the closeness of the point in Figure 5b to the boundary of the admissible region. This is explored in Section 5.



(a) Triple 125 (b) Triple 346

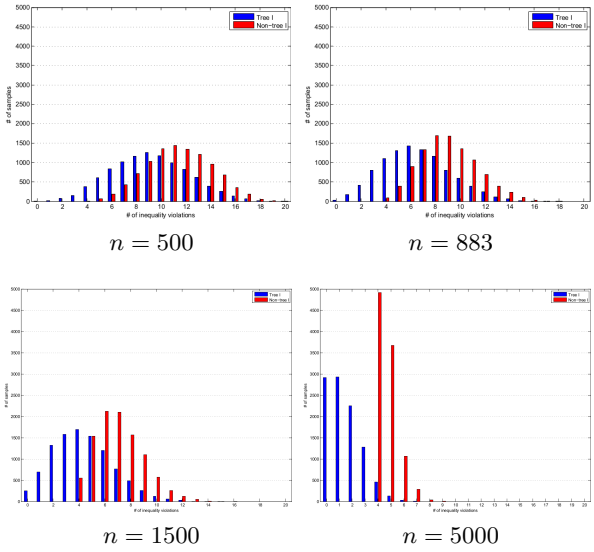
Figure 5: Point estimates of covariances.

5 The effect of sample size

We now examine the likely power of these diagnostics for studies like those given below using data generated from the graphs in Figure 3 for smaller samples. Since the structure of the graphs is known, for the trees any violations of the constraints are due to estimation error of the moments. The inequalities can be tested for each of the four graphs using different sample sizes (n = 500, 883, 1500, 5000), and repeated 10^4 times. For the 20 combinations of triples on each graph the number of triples which violate at least one constraint is recorded. This is relevant as even for a tree, some violations (depending on the sample size) will be expected due to noise. A comparison of these results is shown in Figure 6 for Tree I and Non-tree I from Figure 3.

It is noticeable at a sample size of 883 (as in Section 4) there is a clear shift between the modal violations in the non-tree model relative to the tree model. If 5 or fewer violations are observed Tree I seems the more likely model, and if 11 or more then Non-tree I appears the favourable hypothesis. Note that the frequencies for the non-tree are likely to be conservative (shifted to the left) compared to more complex non-trees, and so the comparison is also prudent.

We compared the two six-leaved tree models and noted that they performed similarly in terms of number of violations. The sample size required for these diagnostics to be useful is likely to vary depending on the topology of the models, as well as the underlying true joint probability table. In the examples we have looked at, a large number of violations would suggest a non-tree



$n = 500$ $n = 883$

$n = 1500$ $n = 5000$

Figure 6: Frequency of violations on a tree (black) and non-tree (white) for 4 sample sizes.

model even for a moderate sample size.

Returning to the application in Section 4 where 5 violations were observed, the frequency of violation plots can be studied. Figure 7 shows the plot for Tree I and Non-tree I, where the arrow indicates the frequencies for 5 violations. If the plot is thought to be typical of other distributions on trees of this size, the level of violations would suggest that the gene sequence data is more likely to conform to a tree than a non-tree. The compared non-tree is from the simplest subset of non-trees (only one additional edge). A more complex tree would be expected to produce more violations and so additional weight may be given to the hypothesised tree.

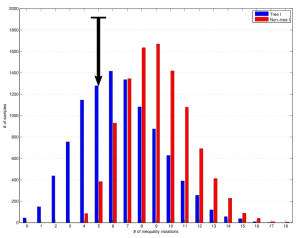


Figure 7: Frequency of violations (n = 883).

6 Inferring trees from moments

It can be shown that 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 X_i, X_j, X_k with $(X_i \perp\!\!\!\perp X_j \perp\!\!\!\perp X_k)|H_{ijk}$ $S_{ijk} = \ln(|\mu_{ij}|)+\ln(|\mu_{ik}|)+\ln(|\mu_{jk}|)-2\ln(|\mu_{ijk}|)$ where 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 $\mathcal{X}(\mathcal{T})$ and hence, from the theorem, candidates $\mathcal{T}$. Note that triples (i, j, k) and (i', j', k') share the same separator H in $\mathcal{T}$ if the pairs (i, i'), (j, j') and (k, k') all lie in different subsets in $\Lambda(H, \mathcal{T})$. So these $(n-2)$ partitions can be calculated by first clustering the signatures by magnitude into up to $(n-2)$ clusters and from this deducing $\Lambda(H, \mathcal{T})$. For small trees this method, simply using the statistics already calculated for our first diagnostics, allows us to identify some promising trees.

Figure 8 shows the standardised signatures S^*_{ijk} for both trees in Figure 3 with the signatures of interest labelled. The clustering of the signatures demonstrates the power of the diagnostic - there is remarkably clear separation of all S^*_{12k} and 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 X_3 X_4 cluster is indicative of this.

When applied to the genetic data we get strong clustering of signatures involving the raccoon and giant panda, and some clustering of polar bear and sun bear - the signatures for the remaining species (red panda and spectacled bear) are dispersed. This diagnostic would thus suggest Tree II as a starting point for an analysis fitting the data, with the latter species being the singletons.

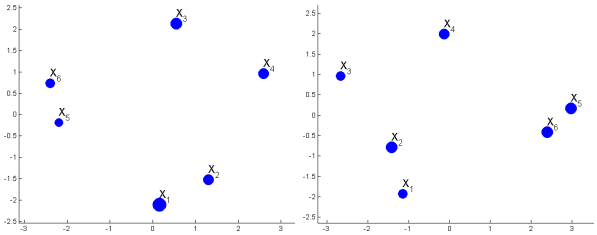
Finally, the sample covariances $\hat{\mu}_{ij}$ themselves can provide some indication of the topology of the tree if correlations between each manifest variable and its hidden parent are of about the same magnitude. For then $\hat{\mu}_{ij}$ tends to be higher when the number of edges between these vertices is fewer (as utilised in Harmeling and Williams (2011)). For example, considering Section 3, for Tree I $\hat{\mu}_{56}$, $\hat{\mu}_{12}$ and $\hat{\mu}_{34}$ have the greatest absolute values and for Tree II $\hat{\mu}_{56}$ and $\hat{\mu}_{12}$ have greatest magnitude, but now $\hat{\mu}_{23}$ and $\hat{\mu}_{24}$ rank above $\hat{\mu}_{34}$ in magnitude. This appears to reflect the structure of Tree II where X_3 and X_4 are not directly joined to the same vertex.

It follows that non-metric multidimensional scaling (MDS) can be used on a function of sample covariances, allowing the relationships to be displayed graphically. Here we use the function $\delta_{ij} = -2\ln(|\hat{\mu}_{ij}|)$

The resulting plots (Figure 9) for a 2D scaling relate to the trees in Figure 3. They were generated using a sample sizes of 883, which gives a relevant comparison with the gene data (see Figure 10). Inequality violation is expressed through the size of the plotted points. The relative sizes are determined via $(1+V_i)^{-1/2}$ where V_i is the number of violations for variable X_i.

For Tree I, the MDS indicates clear pairings (X_1, X_2), and (X_5, X_6), plus X_3 and X_4 are reasonably close which suggests these three pairs are a distance of 2 edges apart giving us the topology here. However in Tree II X_3 and X_4 are further apart so might be conjectured as singletons - (giving us Tree II). Note that this ambiguity may be caused in part by some $|\mu_{ij}|$ being close to zero.

Figure 10 is the result of performing the same MDS for the genetic data with the hope that the plot indicates one of the two six-leaved trees. Unlike the previous plots, the points are separated into two groups of three. This matches the form of Tree II, and although some of the distances are similar it could be hypothesised that $(X_2\ X_4)$, and $(X_5\ X_6)$ are pairs, with X_1



(a) Tree I (b) Tree II

Figure 9: MDS for trees in Figure 3.

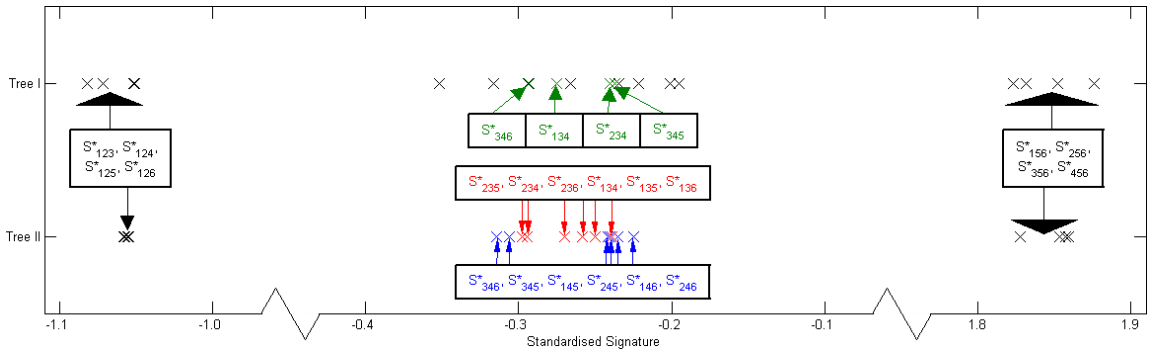


Figure 8: Plot of standardised signatures for Tree I and Tree II.

and X_3 as singletons.

It is interesting to note that the ambiguity is reflected in a more detailed analysis of this data set. We are currently using these simple methods described to preselect good trees with the hope of achieving a large time saving at little cost to accuracy. We then use the subset of trees as a starting point for MCMC likelihood-based model selection algorithms, which otherwise often get stuck within local maxima (e.g. see Chor et al. (2000)).

7 Discussion

In this paper we have illustrated how some simple graphical properties of trees allow us to construct useful diagnostics based on inequality violations and certain functions of sample moments up to degree 3. The diagnostics are trivial to calculate and for the sizes of data sets used in phylogenetic trees (now typically 1500 base pairs if using BOLD) provide a relatively powerful method. In particlar the inequality diagnostics are complementary to the algebraic methods developed by Drton and Sullivant (2007) which are based on different functional constraints. We are currently developing analogous inequality diagnostics using the same graphical properties

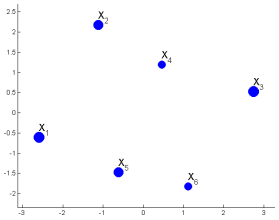


Figure 10: MDS for gene data.

but for differently distributed variables.

One appealing and unusual feature of our methods based on low order moments is that as we add more species to the putative tree we simply need to check the new triples introduced by the additional manifest variables. So in this sense it scales up. Furthermore if violations of the tree structure are discovered our methods also sometimes allow us to identify a subset of manifest variables on which a tree is valid. So for example in the two non-trees of Figure 3 we can still deduce that a tree might be valid on X_1, X_2, X_5, X_6 since no violations of the inequality constraints are apparent.

Of course, rather than simple singularities it is important to develop more general theory for tree diagnostics so that they can be exploited routinely. But even in this naive form, our methods appear surprisingly effective. This work is continuing and we will report our findings in a future paper.

We gratefully acknowledge the referees for their thorough and valuable feedback.

References

E S Allman, C Matias, and J A Rhodes. 2009. Identifiability of parameters in latent structure models with many observed variables. *Ann. Stat.*, 37(6A):3099–3132.

B Chor, M D Hendy, B R Holland, and D Penny. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Biol. Evol.*, 17(10):1529–1541.

J de Leeuw, P G M van der Heijden, and P Ver-

boon. 1990. A latent time-budget model. *Stat. Neerlandica*, 44(1):1–22.

M Drton and S Sullivant. 2007. Algebraic statistical models. *Stat. Sinica*, 17(4):1273–1297.

J Dutkowski and J Tiuryn. 2007. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinforma.*, 23(13):149–158.

J S Farris. 1989. The retention index and the rescaled consistency index. *Cladistics*, 5(4):417–419.

S Harmeling and C K I Williams. 2011. Greedy learning of binary latent trees. *Pattern Anal. Mach. Intell.*, 33(6):1087–1097.

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

R Settimi and J Q Smith. 2000. Geometry, moments and conditional independence trees with hidden variables. *Ann. Stat.*, 28(4):1179–1205.

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

B Streitberg. 1990. Lancaster interactions revisited. *Ann. Stat.*, 18(4):1878–1885.

T Verma and J Pearl. 1990. Equivalence and synthesis of causal models. In *Proc. 6th Conf. Uncertain. Artif. Intell.*, pages 255–270. Elsevier Science.

Z Yang. 2007. *Computational Molecular Evolution*. Oxford University Press.

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

P Zwiernik and J Q Smith. 2012. Tree cumulants and the geometry of binary tree models. *Bernoulli*, 18(1):290–321.

Appendix A Proof of Theorem 1

Proof. Let X_i, X_j, X_k be any three manifest variables (leaves) on a phylogenetic tree T. Let X_a-X_b denote a path between X_a and X_b. Similarly, X_a-X_b-X_c denotes a path between X_a and X_c, and the path contains at least X_b.

By properties of a tree, there is exactly one path with no repeated edges X_i-X_j and similarly one such path X_i-X_k and X_j-X_k. Note that the intersection of the non-repeating paths X_i-X_j and X_i-X_k has at least one hidden vertex as it contains the hidden node adjacent to X_i. Denote the vertex in this intersection furthest from X_i as H^*, thus the intersection of X_j-H^* and X_k-H^* is the vertex H^*. Now consider the repeating path X_j-H^*-X_i-H^*-X_k. By removing the repeating nodes and their edges, a non-repeating subpath X_j-H^*-X_k is formed, with H^* being the only remaining node from the intersection of X_i-X_j and X_i-X_k . Thus $H^* = H_{ijk}$ and furthermore H_{ijk} is unique as no other vertex appears on all three non-repeating paths X_i-X_j, X_i-X_k and X_j-X_k. □

Appendix B Proof of Theorem 2

Proof. Let H' be a vertex in $\mathcal{T}$ which is a leaf of the subtree $\mathcal{T}^{\mathcal{H}}$ consisting of all hidden vertices and their connecting edges in $\mathcal{T}$. Then since $\mathcal{T}$ is strictly trivalent and H' is an interior vertex in $\mathcal{T}$, H' must be connected to two manifest vertices of $\mathcal{T}$ which we label X_{m-1} and X_m. Thus $\{X_{m-1}\}$ and $\{X_m\}$ are singletons in $\Lambda(H', \mathcal{T})$.

Suppose there exists m, the lowest number of leaves a strictly trivalent tree can have such that there exists nonisomorphic $\mathcal{T}_1$ and $\mathcal{T}_2$ with $\mathcal{X}(\mathcal{T}_1) = \mathcal{X}(\mathcal{T}_2)$. For $m = 3$ there is only one strictly trivalent tree, so $m \geq 4$. Now select $H' \in \mathcal{H}$ such that $\Lambda(H', \mathcal{T}_1)$ $(= \Lambda(H', \mathcal{T}_2))$ contains observed vertices X_{m-1}, X_m as singletons.

Note then that for all other $H \in \mathcal{H}$ by the definition of $\mathcal{X}(\mathcal{T}_1)$ (and $\mathcal{X}(\mathcal{T}_2)$) the pair $\{X_{m-1}, X_m\}$ are contained in the same subset of both partitions $\Lambda(H, \mathcal{T}_1)$ and $\Lambda(H, \mathcal{T}_2)$. So

$$\mathcal{X}(\mathcal{T}') = \{\Lambda(H, \mathcal{T}_i) : H \in \mathcal{H} \setminus \{H'\}, i = 1, 2\}$$

is isomorphic to a set of partitions of $\{X_1, \ldots, X'_{m-1}\}$ where X'_{m-1} is identified with $\{X_{m-1}, X_m\}$.

Now define trees $\mathcal{T}'_i$ from $\mathcal{T}_i$, $i = 1, 2$ each having $m - 1$ observed vertices: In $\mathcal{T}'_i$ replace $H' \in \mathcal{H}$ by a manifest variable X'_{m-1} then delete vertices X_{m-1}, X_m and their connecting edges.

By construction $\mathcal{X}(\mathcal{T}'_1) = \mathcal{X}(\mathcal{T}'_2)$, yet $\mathcal{T}'_1$ and $\mathcal{T}'_2$ have $m - 1$ manifest variables so by the definition of m they must be isomorphic. But then by construction $\mathcal{T}_1$ and $\mathcal{T}_2$ are also isomorphic. Thus no such m exists. Contradiction. □