The path sets of weighted partially labelled trees

M.D. Hendy

Department of Mathematics and Statistics

Massey University

Palmerston North

New Zealand

Abstract

In the construction of evolutionary trees for a set of n species, we are sometimes given a set of measures of differences between each pair of species. The underlying assumption is that these differences are the result of a stochastic process of changes occurring on the edges of the historical tree T which incorporates the evolutionary relationships of the species. We quantify the "changes" on the edges of T as additive edge weights where the sum of edge weights on the path between two species is their difference value. A fundamental problem of biology is to determine T and its edge weight function w, given only the set of difference values between them. For real data an exact fit cannot be expected. A number of algorithms exist, of varying complexity, that estimate a tree which purports to represent T. Although the biological model implies that the edge weights are non-negative, a theoretical analysis of the effect of data errors requires the consideration of negative weights.

In this paper we establish the precise conditions under which the weighted tree (T, w) can be uniquely determined, that is the conditions for the mapping from (T, w) to the set of differences to be one to one. This is done in the general case where the differences can be negative, extending the classical result for positive differences. We prove that inversion is possible if and only if the weight of each edge is non-zero and every vertex of degree 1 and 2 of the tree represents one of the species being considered.

The proof of this result uses a natural application of a Hadamard matrix and provides us with an algorithm for reconstructing (T, w) from the set of difference values. However this algorithm is not practical for real data, but a closely related practical algorithm exists when the edge weights are positive.

Introduction

Let $T = (V, E)$ be a tree with edge set E and vertex set V. We say T is an (edge) **weighted tree** $\Leftrightarrow \exists\ w: E \to \mathbb{R}$. We say T is **partially** (vertex) **labelled** by the set $L \Leftrightarrow \exists\ l: L \to V$. Let (T, w) be a weighted tree partially labelled by $N = \{1, 2, ..., n\}$. For $i, j \in N$, let Π_{ij} be the (i, j)th **path**, the set of edges of E connecting vertices $l(i)$ and $l(j)$. For $i, j \in N$, let $D_{ij} = \sum_{e \in \Pi_{ij}} w(e)$, $D_{ii} = 0$ and $D(T, w) = [D_{ij} \mid i, j \in N]$ be the **difference matrix** of (T, w).

The major result of this paper is:

Theorem

Given the difference matrix D(T, w) we can reconstruct (T, w)

$\Leftrightarrow$ 1. $\forall v \in V, d(v) < 3 \Rightarrow v \in \ell(N)$;

and 2. $w(e) \neq 0, \forall e \in E$.

Note

The corresponding result for positive weights was established by Buneman [1971]. This extended result can also be shown to follow from the linear independence of split metrics independently developed by Bandelt and Dress [1990].

Example 1

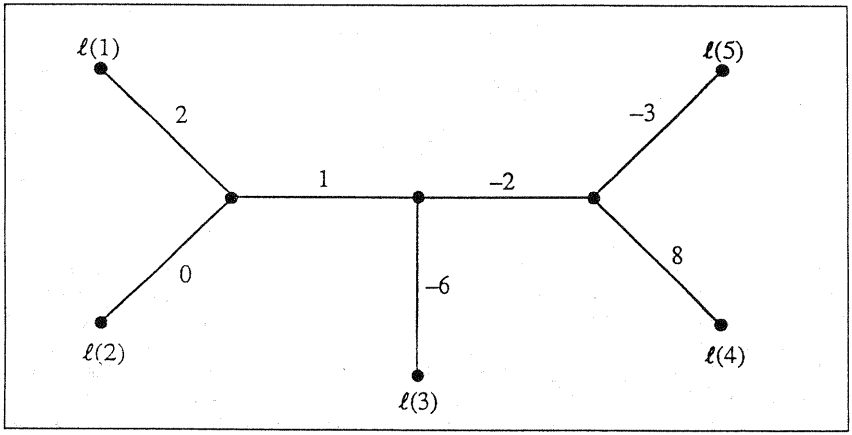

figure 1

A weighted tree partially labelled by N = {1, 2, 3, 4, 5}.

Consider the weighted partially labelled tree (T, w) of figure 1. For any pair of labelled vertices $\ell(i)$, $\ell(j)$, D_{ij}, the sum of the weights along the path Π_{ij} is easily determined. Hence we find the distance matrix $D = \begin{pmatrix} - & 2 & -3 & 9 & -2 \\ & - & -5 & 7 & -4 \\ & & - & 0 & -11 \\ & & & - & 5 \\ & & & & - \end{pmatrix}$.

Note

We note that for n labelled vertices, D has $\binom{n}{2} = \frac{n(n-1)}{2}$ free parameters, whereas T has no more than 2n-3 edges. Thus for $n > 3$, the mapping to D cannot be onto. Inversion must be restricted to the image space.

Proof

We begin the proof of the theorem by establishing the necessity of the conditions. Figures 2(a) and 2(b) are trees with unlabelled vertices of degrees 2 and 1. We see that for any real value x, $D = (D_{12}) = (1)$, so that condition (1) is necessary. Figures 2(c) and 2(d) are distinct trees with an edge of weight 0, while the corresponding distance matrices are both equal to that of figure 2(e) which has no edges with weight 0. In some applications these trees are regarded as equal, i.e., an edge of 0 weight is regarded as equivalent to the deletion of that edge with its endpoints identified. The same argument can be applied to any tree to establish the necessity of condition (2).

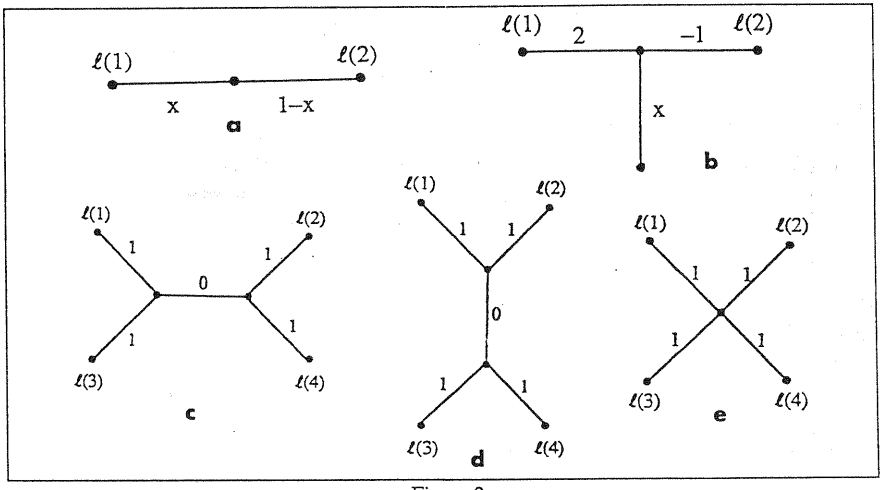

Figure 2

Examples of weighted trees showing the necessity of conditions 1 and 2 of the theorem. The trees of figures 2a and 2b violate the first condition and have the same difference matrix $D = \begin{pmatrix} - & 1 \\ & - \end{pmatrix}$ for any value x. The trees of figures 2c and 2d violate the second condition and have the difference matrix $D = \begin{pmatrix} - & 2 & 2 & 2 \\ & - & 2 & 2 \\ & & - & 2 \\ & & & - \end{pmatrix}$, the same as that of the tree of figure 2e.

The sufficiency of the conditions of the theorem requires further definitions and lemmas.

Definition

Let T be a tree partially labelled by N. Let $\alpha : E \rightarrow 2^N$ be a map which assigns the subset $\alpha(e) = \{i \mid e \in \Pi_{in}\}$ of N to the edge e of T. An example is given in figure 3.

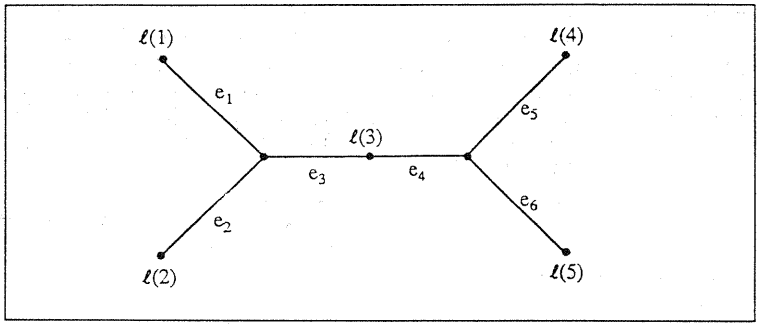

figure 3

A tree T partially labelled by {1, 2, 3, 4, 5}, and edges indexed. We find $\alpha(e_1) = \{1\}$, $\alpha(e_3) = \{1, 2\}$, $\alpha(e_4) = \{1, 2, 3\}$, etc.

We see below that α is injective, so that each edge e of T is identified by its α value.

Lemma 1

$\alpha(e) = \alpha(e') \Rightarrow e = e'$.

Proof

Let Π be a path containing e, beginning at $\ell(n)$ and extended maximally, so the other end is a vertex of degree1 and hence labelled, $\ell(j)$ say for some $j \in N$. Hence $j \in \alpha(e) \Rightarrow \alpha(e) \neq \varnothing$.

Suppose e' is another edge of T, and that $i \in \alpha(e) \cap \alpha(e')$. Thus e, e' $\in \Pi_{in}$. If $e \neq e'$, suppose that e' is further from $\ell(n)$ than e and that v is a vertex between them. If d(v) = 2, then v is labelled , so $v = \ell(j)$ for some $j \in N$. $j \in \alpha(e) - \alpha(e') \Rightarrow \alpha(e) \neq \alpha(e')$. If d(v) > 2, let e" be an edge incident at v, but not in Π_{in}. $\alpha(e") \neq \varnothing$, so for $j \in \alpha(e")$, $j \in \alpha(e) - \alpha(e') \Rightarrow \alpha(e) \neq \alpha(e')$.

Thus $\alpha(e) = \alpha(e') \Rightarrow e = e'$.

Corollary 2

$e, e' \in E \Rightarrow \alpha(e) \cap \alpha(e') \in \{\alpha(e), \alpha(e'), \varnothing\}$.

Definitions

$A = \{\alpha_0, \alpha_1, ..., \alpha_k\} \subset 2^N$ is a set of **partially nested subsets** of N

$\Leftrightarrow$

1. $\alpha_0 = N$;
2. $n \notin \alpha_i$, $1 \leq i \leq k$;
3. $\varnothing \notin A$;
4. $\alpha_i, \alpha_j \in A \Rightarrow \alpha_i \cap \alpha_j \in \{\alpha_i, \alpha_j, \varnothing\}$.

(Thus $\{\alpha(e) \mid e \in E(T)$ is a set of partially nested subsets of N.)

A **hierarchy** on N is any tree T partially labelled by N with every vertex of degree 1 or 2 labelled.

Lemma 3

There is a natural bijection between the set of hierarchies on N and the collection of sets of partially nested subsets of N.

Proof

If T is a hierarchy, then $A(T) = \{\alpha(e) \mid e \in E(T)\} \cup \{N\}$ is a set of partially nested subsets of N. If A is a set of partially nested subsets of N, then let T(A) be the graph with vertex set V = A, and edges $e = (\alpha_i, \alpha_j)$ whenever α_j is a maximal subset of α_i. $\forall\ \alpha_i \in A$, $\alpha_i \subseteq N \in A \Rightarrow$ T(A) is connected. $\alpha_i \cap \alpha_j \in \{\alpha_i, \alpha_j, \varnothing\} \Rightarrow$ T(A) is acyclic, so T(A) is a tree.

For $\alpha_i \in V$, let $t^{-1}(\alpha_i) = \alpha_i - \bigcup_{(\alpha i, \alpha j) \in E} \alpha_j \subseteq N$.

$n \in \alpha_0$, $n \notin \alpha_i$, for $i > 0 \Rightarrow n \in t^{-1}(\alpha_0) \neq \varnothing$.

For $i > 0$, $d(\alpha_i) = 1 \Rightarrow t^{-1}(\alpha_i) = \alpha_i \neq \varnothing$, $d(\alpha_i) = 2 \Rightarrow \alpha_i$ has exactly one maximal subset α_j,

$\Rightarrow t^{-1}(\alpha_i) = \alpha_i - \alpha_j \neq \varnothing$,

therefore T(A) is a hierarchy.

Using the constructions above we find A(T(A)) = A, and T(A(T)) = T, establishing the natural bijection.

Definition

For (T, w) partially labelled by $N = \{1, 2, ..., n\}$ let $N' = \{1, 2, ..., n-1\}$, and $m = 2^{n-1} = 2^{|N'|}$. We define the **spectrum** of (T, w) as $q(T,w) \in \mathbb{R}^m$, where $q = (q_S \mid S \subseteq N')$ is indexed by the subsets of N', with:

$$q_\varnothing = -\sum_{e \in E} w(e);$$

$$q_{\alpha(e)} = w(e), \quad \forall e \in E;$$

$$q_S = 0, \quad \text{otherwise.}$$

Example 2

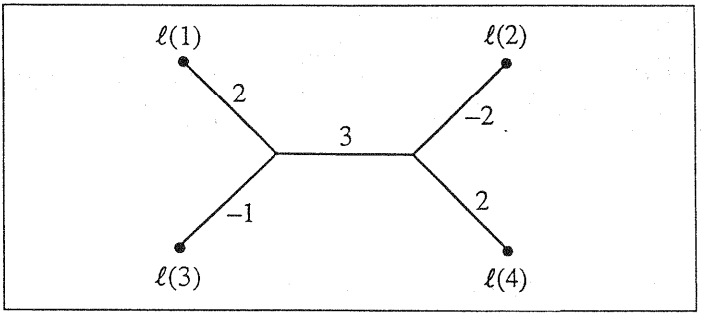

figure 4
A weighted partially labelled tree (T, w).

The partially labelled weighted tree (T, w) of figure 4, with n = 4, N' = {1, 2, 3}, m = 8 has spectrum q(T, w) = (-4, 2, -2, 0, -1, 3, 0, 2), where the components are indexed with the subsets {∅, {1}, {2}, {1, 2}, {3}, {1, 3}, {2, 3}, {1, 2, 3}} ordered recursively.

In general we find provided (T, w) satisfies the conditions of the theorem, then (T, w) is uniquely determined by its spectrum.

Lemma 4

If (T, w) and (T', w') are weighted partially labelled trees satisfying the conditions of the theorem and if $q(T, w) = q(T', w')$ then $T = T'$ and $w = w'$.

Proof

$\{S \mid q_S \neq 0, S \neq \varnothing\}$ is a set of nested subsets of N', which by lemma 3 uniquely specifies the edges, so $T = T'$. Then $w(e) = w'(e) = q_{\alpha(e)}$ so $w = w'$.

Definition

For $S \neq \varnothing$, the **power group** of S, P(S) is the group of all subsets of S under disjoint union. For $A \subseteq S$, let $h_A : P(S) \to (\{1, -1\} \times)$ where $h_A(B) = (-1)^{|A \cap B|}$.

Lemma 5

$\forall\ A \subseteq S$, h_A is a group homomorphism.

Proof

$|A \cap (B \nabla C)| = |A \cap B| + |A \cap C| - 2|A \cap B \cap C|$,
therefore $h_A(B \nabla C) = h_A(B)h_A(C)$. (Here ∇ is used as the symmetric difference symbol.)

Lemma 6

$H = [h_A(B) \mid A, B \subseteq S]$ is a symmetric Hadamard matrix of order $2^{|S|}$.

Proof

If $A \neq \varnothing$, $\exists a \in A$, $h_A(\{a\}) = -1 \Rightarrow h_A$ is not trivial $\Rightarrow (G : \ker h_A) = 2$

$$\Rightarrow \sum_{B \subseteq S} h_A(B) = \sum_{B \subseteq S} h_B(A) = 0 .$$

If $A_1 \neq A_2$, $\sum_{B \subseteq S} h_{A_1}(B)h_{A_2}(B) = \sum_{B \subseteq S} h_B(A_1)h_B(A_2) = \sum_{B \subseteq S} h_B(A_1 \nabla A_2) = 0.$

Thus the rows and columns of H are orthogonal. (Using the recursive ordering of the subsets of S as in example 2, H can be shown to be the Kronecker product of $H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ with itself $|S| - 1$ times.)

Corollary 7

$H^{-1} = 2^{-|S|}H$.

Definition

For T partially labelled by N let $\Pi(T)$ be the subgroup of P(N') generated by $\{\Pi_{1n}, \Pi_{2n}, ..., \Pi_{(n-1)n}\}$, writing $\Pi_A = \nabla_{i \in A} \Pi_{in}$. The elements of $\Pi(T)$ are called **path sets** of T.

Lemma 8

Π_A is the union of $[(|A| + 1)/2]$ disjoint (no common edge) path sets Π_{ij}, for some $i, j \in A'$, where $A' = A$, if $|A|$ even, $A' = A \cup \{n\}$ if $|A|$ odd.

Proof

For any $i, j \in N$, $\Pi_{in} \nabla \Pi_{jn} = \Pi_{ij}$, and the internal vertices in the path Π_{ij} are of degree 2 in Π_{ij}. For $i, j, k, l \in N$, $\Pi_{ij} \nabla \Pi_{kl}$ the vertices of odd degree are $l(i)$, $l(j)$, $l(k)$ and $l(l)$. Thus the disjoint union is the union of two paths with no common edges. (Some paths could have 0 edges.) We complete the result by

induction on |A| taking the disjoint union of the paths of $\Pi_{A-\{i,j\}}$ with $\Pi_{\{i,j\}}$. For each successive pair, we increase the number of disjoint paths by one.

The decomposition of Π_A into these paths is not necessarily unique, as alternatives can exist at vertices of degree greater than 3 in T. However we can select the paths in lexicographic order to obtain a canonical decomposition.

Definition

Let $w_A = \sum_{e \in \Pi_A} w_e$. This is a generalisation of the weight of a path, and is computed in this sum using the edges of T. However, as we show, the value w_A can be determined without reference to T.

Suppose L is a list of numbers (x_i), with only one numerical value x_j say occurring an odd number of times, then we define ODD(L) = x_j, to be the unique value occurring an odd number of times.

Lemma 9

$w_A = ODD(w_{A-\{i,j\}} + D_{ij} \mid j \in A'-\{i\})$, for any $i \in A$.

Proof

If $A = \{i\}$, then $\Pi_A = \Pi_{\{i,n\}} = \Pi_{in}$ is a single path, so $w_A = D_{in}$, $w_{A-\{i,n\}} = \emptyset$, and $ODD(w_{A-\{i,j\}} + D_{ij} \mid j \in A'-\{i\}) = ODD(D_{in}) = D_{in}$.

Likewise when $A=\{i,j\}$, $ODD(w_{A-\{i,k\}} + D_{ik} \mid k \in A'-\{i\}) = ODD(D_{ij}) = D_{ij}$.

Suppose $|A| > 2$ and that the lemma applies to all subsets of order |A| -2. Let i be an element of A. By lemma 8, Π_A comprises |A'|/2 disjoint paths, so let $j \in A'$ be an index so that Π_{ij} is in the canonical decomposition of Π_A. For any $k \in A' - \{i, j\}$, let $t(l)$ be the vertex paired with $t(k)$ in the canonical decomposition of Π_A. Then $w_{A-\{i,k\}} + D_{ik} = w_{A-\{i,l\}} + D_{il}$. This gives a unique pairing of equal valued sums in the list $L = (w_{A-\{i,k\}} + D_{ik})$ as k runs through all indices in $A' - \{i, j\}$.Thus there will be an even number of values common to $D_{ik} + w_{A-\{i,k\}}$. However |A'| is even, so the number of elements of $A' - \{i\}$ is odd, and the value $D_{ij} + w_{A-\{i,j\}}$ occurs an odd number of times.

Example

For the tree of figure 4 we see $w_{\{1,2,3\}} = 2 + (-1) + (-2) + 2 = 1$, and $w_{\{1\}} + D_{23} = 7 + 0 = 7$, $w_{\{2\}} + D_{13} = 0 + 1 = 1$, $w_{\{3\}} + D_{12} = 4 + 3 = 7$. ODD(7, 1, 7) = 1.

The ODD function gives us a way of computing w_A for each $A \subseteq N'$, without reference to T. The next lemma gives another way of computing w_A directly from the q_e values again without direct reference to T. In corollary 11 this relation is inverted to show us how (T, w) can be recovered

Lemma 10

$$\forall A \subseteq N',\ w_A = -\frac{1}{2} \sum_{B \subseteq N'} h_A(B) q_B.$$

Proof

$e \in \Pi_A = \Pi_{(i_1,i_2)} \nabla \ldots \nabla \Pi_{(i_{2k-1},i_{2k})} \Rightarrow$ the number of these paths containing e is odd, and hence $h_{\alpha(e)}(A) = -1$.

$q_B = 0$ except for $B = \varnothing$ and $B = \alpha(e)$, for edges e of T. As $q_{\alpha(e)} = w(e)$, $\sum_{B \subseteq N'} h_B(A) q_B = q_\varnothing +$

$$\sum_{e \in E} h_{\alpha(e)}(A) q_{\alpha(e)} = - \sum_{e \in E} w(e) + \sum_{e \in E} h_{\alpha(e)}(A) w(e) = -2 \sum_{e \in \Pi_A} w(e) = -2w_A.$$

This gives the matrix product $w = -\frac{1}{2} Hq$, so multiplying by H^{-1} (corollary 7) we obtain:

Corollary 11

$$\forall B \subseteq N', \; q_B = -2^{2-n} \sum_{A \subseteq N'} h_A(B) w_A.$$

This completes the proof of the theorem. Given the values D derived from a partially labelled weighted tree (T, w) satisfying the conditions of the theorem, we can use lemma 9 to calculate $w_A \; \forall A \subseteq N'$, and then the inversion, corollary 11 to determine $q_B \; \forall B \subseteq N'$. Then the set $\{B \mid q_B \neq 0, B \neq \varnothing\}$ is the set of partially nested subsets of N that uniquely specifies the edges of T (lemma 3). We then recover the edge weights $w_e = q_{\alpha(e)}$.

Theoretically this describes an algorithm for recovering the evolutionary trees from sets of distance data D, however it does not represent a practical algorithm. Real data will contain a degree of uncertainty in the D_{ij} values, while the ODD function of lemma 9 requires exact data. It is not easy to see how it could be adapted to allow for imprecision. If the edge weights were all positive, then we find w_A is the maximum of that set. This forms part of a practical algorithm for evolutionary tree recovery. [Hendy and Penny, 1991]. There are a number of other algorithms which find a partially labelled weighted tree (T, w) which fits the data exactly, if there is an exact fit. [Buneman, 1971] Thus, although the procedure above is not in itself a practical method, it does provide the precise conditions for invertibility.

This analysis can be extended to any application of additive measures on the edges of a partially labelled tree, where for example the q_e values may be time intervals or expected numbers of events. [Hendy, 1989].

References

Bandelt, H.-J. and Dress, A.W.M. (1990). A canonical decomposition theory for metrics on a finite set. (Preprint)

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In "Mathematics in the archeological and historical sciences". F.R. Hodson, D.G. Kendall and P. Tautu (ed.s.). Edinburgh, University Press. 387-395.

Hendy, M.D. (1989). The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool., 38 310-321.

Hendy, M.D. and Penny, D. (1991). Spectral Analysis of phylogenetic data. (Preprint)