arXiv:0802.2395v1 [math.CO] 17 Feb 2008

Combinatorics of least squares trees

Radu Mihaescu* Lior Pachter*†

February 24, 2013

Abstract

A recurring theme in the least squares approach to phylogenetics has been the discovery of elegant combinatorial formulas for the least squares estimates of edge lengths. These formulas have proved useful for the development of efficient algorithms, and have also been important for understanding connections among popular phylogeny algorithms. For example, the selection criterion of the neighbor-joining algorithm is now understood in terms of the combinatorial formulas of Pauplin for estimating tree length.

We highlight a phylogenetically desirable property that weighted least squares methods should satisfy, and provide a complete characterization of methods that satisfy the property. The necessary and sufficient condition is a multiplicative four point condition that the the variance matrix needs to satisfy. The proof is based on the observation that the Lagrange multipliers in the proof of the Gauss–Markov theorem are tree-additive. Our results generalize and complete previous work on ordinary least squares, balanced minimum evolution and the taxon weighted variance model. They also provide a time optimal algorithm for computation.

1 Introduction

The least squares approach to phylogenetics was first suggested by Cavalli-Sforza & Edwards [3] and Fitch & Margoliash [9]. The precise problem formulated in [3] was Problem 1.1:

Definition 1.1. (Pair-edge incidence matrix) *Given a phylogenetic X-tree T with edge set E and $|X| = n$ (see [26] for basic definitions), the*

*Department of Mathematics, UC Berkeley

†lpachter@math.berkeley.edu

pair-edge incidence matrix *of T is the $\binom{n}{2} \times |E|$ matrix*

$$(S_T)_{ij,e} = \begin{cases} 1 \text{ if } e \in E \text{ is an edge on the path between } i \text{ and } j, \\ 0 \text{ otherwise.} \end{cases}$$

Definition 1.2. (Tree-additive map) *Let T be a phylogenetic X-tree. A dissimilarity map D is T-additive if for some vector $l \in \mathbf{R}^{|E|}$,*

$$D_{ij} = (S_T l)_{ij}. \tag{1}$$

Problem 1.1 (Ordinary least squares) *Find the phylogenetic X-tree T and T-additive map $\hat{D}$ that minimizes*

$$\sum_{i,j \in \binom{X}{2}} (D_{ij} - \hat{D}_{ij})^2. \tag{2}$$

For a fixed tree, the solution of Problem 1.1 is a linear algebra problem (Theorem 1.3.). However Rzhetsky & Nei [24] showed that the **O**rdinary **L**east **S**quares edge lengths could instead be computed using elegant and efficient combinatorial formulas. Their result was based on an observation of Vach [27], namely that OLS edge lengths obey the desirable **I***ndependence of* **I***rrelevant* **P***airs* property (our choice of terminology is inspired by social choice theory [23]):

Property 1.1 (IIP) *Let T be a phylogenetic X-tree and e an edge in T. A linear edge length estimator for e is a linear function from dissimilarity maps to the real numbers, i.e. $\hat{l}_e = \sum_{ij} p_{ij} D_{ij}$. We say that such an estimator satisfies the IIP property if $p_{ij} = 0$ when the path from i to j in T (denoted $\overline{i,j}$) does not contain either of e's endpoints.*

In other words, the IIP property is equivalent to the statement that the sufficient statistic for the least squares estimator of the length of e is a projection of the dissimilarity map onto the coordinates given by pairs of leaves whose joining path contains at least one endpoint of e. It has been shown that this crucial property is satisfied not only by ordinary least squares (OLS) estimators, but also by specific instances of **W***eighted* **L***east* **S***quares* estimators (e.g., [25]).

Problem 1.2 (Weighted least squares) *Let T be a phylogenetic X-tree and D be a dissimilarity map. Find the T-additive map $\hat{D}$ that minimizes*

$$\sum_{i,j \in \binom{X}{2}} \frac{1}{V_{ij}} \left(D_{ij} - \hat{D}_{ij}\right)^2. \tag{3}$$

The *variance matrix* for weighted least squares is the $\binom{n}{2} \times \binom{n}{2}$ diagonal matrix V whose diagonal entries are the V_{ij}. Note that V can also be regarded as a dissimilarity map and we will do so in this paper. Weighted least squares for trees was first suggested in [9] and [14], with the former proposing specifically $V_{ij} = D_{ij}^2$.

Theorem 1.3. (Least squares solution) *The solution to Problem 1.2 is given by* $\hat{D} = S_T\hat{l}$ *where*

$$\hat{l} = (S_T^t V^{-1} S_T)^{-1} S_T^t V^{-1} D. \tag{4}$$

We note that The OLS problem reduces to the case $V = I$. The statistical significance of the variance matrix together with a statistical interpretation of Theorem 1.3. is provided in Section 2.

It follows from (4) that the lengths of the edges in a weighted least squares tree are linear combinations of the entries of the dissimilarity map. A natural question is therefore which variances matrices V result in edge length estimators that satisfy the IIP property? Our main result is an answer to this question in the form of a characterization (Theorem 3.4.): a WLS model is IIP if and only if the variance matrix is semi-multiplicative. We show that such matrices are good approximations to the variances resulting from popular distance estimation procedures. Moreover, we provide combinatorial formulas that describe the WLS edge lengths under semi-multiplicative variances (Equation 20), and show that they lead to optimal algorithms for computing the lengths (Theorem 4.1.).

The key idea that leads to our results is a connection between Lagrange multipliers arising in the proof of the Gauss–Markov theorem and the weak fundamental theorem of phylogenetics that provides a combinatorial characterization of tree-additive maps (Remark 2.5.). This explains many isolated results in the literature on least squares in phylogenetics; in fact, as we show in the section "The multiplicative model and other corollaries", almost all the known theorems and algorithms about least squares estimates of edge lengths follow from our results.

2 BLUE Trees

The foundation of least squares theory in statistics is the Gauss–Markov theorem. This theorem states that the **B**est **L**inear **U**nbiased **E**stimator for a linear combination of the edge lengths, when the errors have zero expectation, is a least squares estimator. We explain this theorem in the context of Problem 1.2.

Lemma 2.1. *For any phylogenetic X-tree T, the matrix S_T is full rank.*

Proof: We show that for any $e \in E$, the vector $f_e = (0, \dots, 1, \dots, 0)$ of size $|E|$ with a 1 in the e-th position and 0 elsewhere lies in the row span of S. Choose any $i, j, k, l \in X$ such that the paths from i to j and from k to l do not intersect, and the intersection of the paths from i to j and from k to l is exactly the edge e. Note that

$$\frac{1}{2}\sum_{e}(S_{ik,e} + S_{jl,e} - S_{ij,e} - S_{kl,e}) = f_e. \tag{5}$$

Theorem 2.2. (Gauss–Markov Theorem) *Suppose that D is a random dissimilarity map of the form $D = S_T l + \epsilon$ where T is a tree, and ϵ is a vector of random variables satisfying $E(\epsilon) = 0$ and $Var(\epsilon) = V$ where V is an invertible variance-covariance matrix for ϵ.*

Let $M(S_T^t)$ be the linear space generated by the columns of S_T^t and $f \in M(S_T^t)$. Then $f^t\hat{l} = p^t D$ (where $\hat{l}$ given by (4)) has minimum variance among the linear unbiased estimators of $f^t l$.

Proof: Observe that the problem of finding p is equivalent to solving a constrained optimization problem:

$$\min\, p^t V p \text{ subject to } S_T^t p = f. \tag{6}$$

The first condition specifies that the goal is to minimize the variance; the second constraint encodes the requirement that the estimator is unbiased. Using Lagrange multipliers, it is easy to see that the minimum variance unbiased estimator of $f^t l$ is the unique vector p satisfying

$$\begin{aligned} Vp &= S_T\mu \text{ for some } \mu \in \mathbf{R}^{|E|}, & (7)\\ S_T^t p &= f. & (8)\end{aligned}$$

In other words

$$\begin{pmatrix} V & -S_T \\ S_T^t & 0 \end{pmatrix}\begin{pmatrix} p \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix} \tag{9}$$

$$\Rightarrow \begin{pmatrix} p \\ \mu \end{pmatrix} = \begin{pmatrix} V^{-1}S_T U^{-1} S_T^t V^{-1} & (U^{-1}S_T^t V^{-1})^t \\ -U^{-1}S_T^t V^{-1} & U^{-1} \end{pmatrix}\begin{pmatrix} 0 \\ f \end{pmatrix}$$

where $U = S_T^t V^{-1} S_T$.

The Gauss–Markov Theorem can also be proved directly using linear algebra, but the Lagrange multiplier proof has two advantages: First, it provides a description of p different from (4) that is simpler and more informative. Secondly, the technique is general and can be used in many similar settings to find minimum variance unbiased estimators. Hayes and Haslett [15] provide pedagogical arguments in favor of Lagrange multipliers for interpreting least squares coefficients and discuss the origins of this approach in applied statistics [19].

In phylogenetics, Theorem 2.2. (and its proof) are useful because for each edge e, the vector f_e in the standard basis for $M(S_T^t)$ is associated with a vector p such that p^tD is the best linear unbiased estimator for the length of e. Similarly, the tree length is estimated from $f_T = (1, 1, \ldots, 1)$ which is also in $M(S_T^t)$. Condition (7) is particularly interesting because it says that there exists some T-additive map $\Lambda = S_T^t\mu = Vp$, whose (possibly negative) edge lengths are given by the Lagrange multipliers μ.

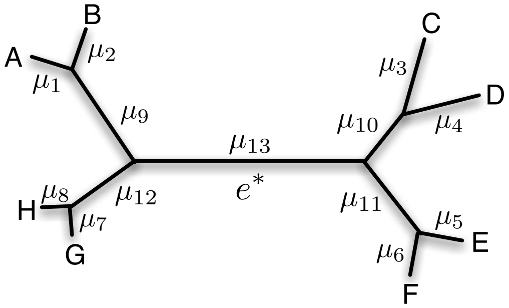

Figure 1: The Lagrange tree Λ for an IIP weighted least squares estimator for the central edge e^* of a complete binary tree with 8 leaves. In Proposition, 3.5. $X = \{A, B, C, D, E, F, G, H\}$, whereas in the proof of Theorem 3.4. the leaf labels represent clades. The IIP property means that the WLS estimate $\hat{l}_{e^*}$ does not depend on D_{AB}, D_{CD}, D_{EF} or D_{GH}.

The following theorem provides a combinatorial characterization of tree-additive maps, and hence of the *Lagrange tree* Λ:

Definition 2.3. (Weak four point condition) *A dissimilarity map D satisfies the weak four point condition if for any $i, j, k, l \in X$, two of the following three linear forms are equal:*

$$D_{ij} + D_{kl}, \quad D_{ik} + D_{jl}, \quad D_{il} + D_{jk}. \tag{10}$$

Theorem 2.4. (Weak fundamental theorem of phylogenetics) *A dissimilarity map D is tree-additive if and only if it satisfies the weak four point condition.*

Theorem 2.4. was first proved in [21]. For a recent exposition see Corollary 7.6.8 of [26] where it is derived using the theory of group-valued dissimilarity maps. We note that the pair of equal quantities in the four point condition define the topology of a quartet. Furthermore the topology of the tree is defined uniquely by the topologies of all its quartets. We again refer the reader to [26] for details.

The *Lagrange equations* (7) and (8) together with Theorem 2.4. form the mathematical basis for our results:

Remark 2.5. *Condition (7) specifies that Vp must be a T-additive map. It follows that Vp satisfies the weak four point condition. In other words, (7) amounts to a combinatorial characterization of Vp, and hence p. Condition (8) imposes a normalization requirement on p. Together these conditions are useful for finding p, and also for understanding its combinatorial properties.*

The structure of the Lagrange tree in the case of OLS is the middle quartet of the tree shown in Figure 1. It immediately reveals interesting properties of the estimator. For example the fact that it is a tree on four taxa implies the IIP property. The content of [5, Appendix 2] is that for tree length estimation under the balanced minimum evolution model, the Lagrange tree is the star tree. In fact, we will see that most of the known combinatorial results about least squares estimates of edge and tree lengths can be explained by Remark 2.5. and interpreted in terms of the structure of the Lagrange tree.

3 Main Theorem

Our main result is a characterization of IIP WLS estimators. In the sections that follow we will see that the IIP property for WLS is not only biologically desirable, but also statistically motivated and algorithmically convenient. We begin by introducing some notation and concepts that are necessary for stating our main theorem.

Definition 3.1. (Clade) *A clade of a phylogenetic X-tree T is a subset $A \subset X$ such that there exists an edge in T whose removal induces the partition $\{A, X \setminus A\}$. We also use clade to mean the induced topology $T|_A$.*

Given a dissimilarity map D and a variance matrix V, we set

$$\begin{aligned} D_{AB} &:= \sum_{a\in A, b\in B} V_{ab}^{-1} D_{ab}, \text{ and} \\ Z_{AB} &:= \sum_{a\in A, b\in B} V_{ab}^{-1}. \end{aligned}$$

where A, B are disjoint clades. If $e_1, \ldots, e_k \in E(T)$ form a path with ends determining clades A and B, then by the notation $D_{e_1 \cdots e_k}$ and $Z_{e_1 \cdots e_k}$ we mean D_{AB} and Z_{AB} respectively. Note that if e is an edge in a tree T then (7,8) imply that the Lagrange tree for any WLS estimate of e satisfies $\Lambda_e = f_e$.

Definition 3.2. (Semi-multiplicative map) *A dissimilarity map D is semi-multiplicative with respect to disjoint clades A, B if for any $a_1, a_2 \in A$ and $b_1, b_2 \in B$*

$$D_{a_1 b_1} D_{a_2 b_2} = D_{a_1 b_2} D_{a_2 b_1}. \tag{11}$$

We say that D is semi-multiplicative with respect to T if for any pair of disjoint clades A, B, not defined by the same edge of T, (11) holds.

Lemma 3.3. *D is semi-multiplicative if and only if every clade A of T has the property that for any $A' \subset A$, and any clade B disjoint from A and induced by a different edge, for all $x \in B$,*

$$Z_{\{x\}A'} / Z_{\{x\}A} = \xi^B_{A'A}, \tag{12}$$

where $\xi^B_{A'A}$ does not depend on x.

It is an easy exercise to prove that A satisfies (12) for all relevant B if and only if (12) holds for the the two clades disjoint from A and defined by the two edges adjacent to the edge defining A.

The semi-multiplicative condition is slightly weaker than $\log D$ being tree-additive. Indeed, removing the requirement that the clades A, B are defined by different edges of T leaves one one with a multiplicative analog of the four-point condition. By Theorem 2.4., this is equivalent to $D_{ij} = \prod_{e \in \overline{i,j}} w(e)^{-1}$ for some $w : E(T) \to \mathbf{R}_+$ [13].

Theorem 3.4. (Characterization of IIP WLS estimators) *A WLS edge length estimator for an edge in a tree T has the IIP property if and only if the variance matrix is semi-multiplicative with respect to T.*

The proof of the theorem reduces to the WLS solution for the length of an edge in a tree with at most eight leaves (edge e^* in Figure 1):

Proposition 3.5. *Let T be the phylogenetic X-tree shown in Figure 1. The Lagrange tree $\Lambda = S_T\mu$ for the WLS problem of estimating the length of the edge e^* satisfies the property that $\mu_1 = -\mu_2$, $\mu_3 = -\mu_4$, $\mu_5 = -\mu_6$ and $\mu_7 = -\mu_8$. Furthermore, these Lagrange multipliers and the remaining ones $\mu_9, \ldots, \mu_{13}$ can be computed by solving $\mu = (S_T^t V^{-1} S_T)^{-1} f_{e^*}$.*

Proof: Using the notation of Figure 1, with the convention that the edge labeled by μ_i is e_i, it follows from (8) that $\Lambda_{e_i} = 0$ for $i = 1, 2, 9$. But $\Lambda_{e_i} = \Lambda_{e_i e_j} + \Lambda_{e_i e_k}$ for $\{i, j, k\} = \{1, 2, 9\}$, which implies that $\Lambda_{e_i e_j} = 0\, \forall i, j \in \{1, 2, 9\}$. Therefore $V_{AB}^{-1}\Lambda_{AB} = V_{AB}^{-1}(\mu_1 + \mu_2) = 0$ and the result follows. The arguments for e_3, e_4, e_5, e_6 and e_7, e_8 are identical. The complete solution for the μ for a given V is given by $\mu = (S_T^t V^{-1} S)^{-1} f_{e^*}$, which reduces to the inversion of a 13×13 matrix.

Note that the proof only uses the fact that e_1, e_2 are adjacent leaf edges not adjacent to e^*. The conclusion $\mu_{e_1} = -\mu_{e_2}$ will hold identically in any tree for a pair of edges of this type.

Proof of Theorem 3.4.: We begin by showing that if V is semi-multiplicative then the WLS edge length estimators have the IIP property. This calculation involves showing that for any phylogenetic X-tree T and edge $e^* \in T$, the Lagrange tree for e^* is the tree in Figure 1, where A,B,C, D, E, F, G, H are clades with the property that their intra-clade Lagrange multipliers are zero.

Let $e_1, \ldots, e_k$, with $k \leq 8$, be the edges of T such that either $d(e^*, e_i) = 2$ or $d(e^*, e_i) < 2$ and e_i is a leaf edge. For $i \in \{1, \ldots, k\}$, let C_i be the clade defined by e_i such that $e^* \notin C_i$. Let $T^{/e^*}$ to be the phylogenetic $X^{/e^*}$-tree, where $X^{/e^*} = \{C_1, \ldots, C_k\}$, with topology induced by T in the natural way (see Figure 1). Set $V^{/e^*}$ be the diagonal variance matrix on pairs of nodes in $X^{/e^*}$ given by $V^{/e^*}_{C_i C_j} = Z^{-1}_{C_i C_j}$.

If $\mu^{/e^*}$ are the Lagrange multipliers and $\Lambda^{/e^*}$ is the Lagrange tree given by estimating $\hat{l}_{e^*}$ for topology $T^{/e^*}$ and variance $V^{/e^*}$ then the T-additive map given by $\Lambda = S_T^t \mu$ with

$$\mu_e = \begin{cases} \mu_e^{/e^*} \text{ if } e \in E(T^{/e^*}), \\ 0 \text{ otherwise.} \end{cases} \tag{13}$$

satisfies the Lagrange equations for T. Thus μ are the Lagrange multipliers for $\hat{l}_{e^*}$ and $\hat{l}_{e^*} = \Lambda^t V^{-1} D$.

We let $\Lambda^{/e*}$, $Z^{/e*}$ denote the natural correspondents of Λ and Z for the problem of estimating $\hat{l}_{e^*}$ from and $V^{/e^*}$ and $T^{/e^*}$. It is an easy exercise to check that for all $e \in E(T^{/e^*})$, we have $Z_e^{/e*} = Z_e$ and $\Lambda_e^{/e*} = \Lambda_e$. This implies that $\Lambda_e = f_e$ for all $e \in E(T^{/e^*})$, i.e. the Lagrange equation (8) is satisfied for $e \in E(T^{/e^*})$.

Now consider edge $e \in C_1$. We need to verify that $\Lambda_e = 0$. Since $\Lambda_{ij} = 0$ for all $i, j \in C_1$, $\Lambda_e = \Lambda_{e \cdots e_2} + \Lambda_{e \cdots e_9}$. Now for all $i \in C_1$ and $j \in C_2$, $\Lambda_{ij} = \mu_1 + \mu_2 = 0$, so $\Lambda_{e \cdots e_2} = 0$. Finally let $A' \subset A$ be the clade defined by e and let A'' be the clade defined by e_9 which does not intersect A. The fact that V is semi-multiplicative implies that for any taxon $x \in A''$

$$Z_{\{x\}A'}/Z_{\{x\}A} = \xi^{C_1}_{A'A} \tag{14}$$

where $\xi_{A'A}$ does not depend on the taxon x. This implies $\Lambda_{e \cdots e_9} = \xi^{C_1}_{A'A} \Lambda_{e_1 \cdots e_9} = 0$ by the proof of Proposition 3.5..

Since $\mu_e = 0$ for all $e \notin T^{/e^*}$, it is enough to show that $\Lambda^{/e^*}$ satisfies the IIP property. This follows from Proposition 3.5.. Therefore, V has the IIP property with respect to T, i.e. $\Lambda_{ij} = 0$ for all $i, j \in X$ such that $\overline{i, j}$ does not intersect e^*.

This concludes the proof for the ”if” part of Theorem 3.4.. For the ”only if” direction, we will prove by induction that (12) is satisfied by all clades A of T, and thus the variance V is semi-multiplicative with respect to T. The base case is provided by clades formed by a single leaf, for which (12) holds vacuously.

For the induction step, suppose clades A and B both satisfy (12), and that they are defined by adjacent edges e_A and e_B (see Figure 2). Let e_C be the other edge adjacent to e_A and e_B and let $C = X \setminus (A \cup B)$ be the clade it defines. We would like to prove that the clade $(A \cup B)$ also satisfies (12). If $|C| = 1$, this holds vacuously. We may therefore assume that there exist two more edges e_1, e_2 incident with e_C. Let $C_i \subset C$ be the clade defined by e_i, for $i = 1, 2$. It suffices to prove that $(A \cup B)$ satisfies (12) with respect to C_1 and C_2. Notice that A and B already satisfy (12) with respect to C_1 and C_2. Therefore it is enough to show that

$$\frac{Z_{\{x\}A}}{Z_{\{x\}(A \cup B)}} = \xi^{C_1}_{A(A \cup B)} \tag{15}$$

is the same for all $x \in C_1$, and similarly for all $x \in C_2$.

Now consider the problem of estimating $\hat{l}_{e_A}$. Let μ be the corresponding Lagrange multipliers and $\Lambda = S_T \mu$ be the Lagrange tree they define. By the IIP property, Λ defines an identically zero tree additive map on the clade

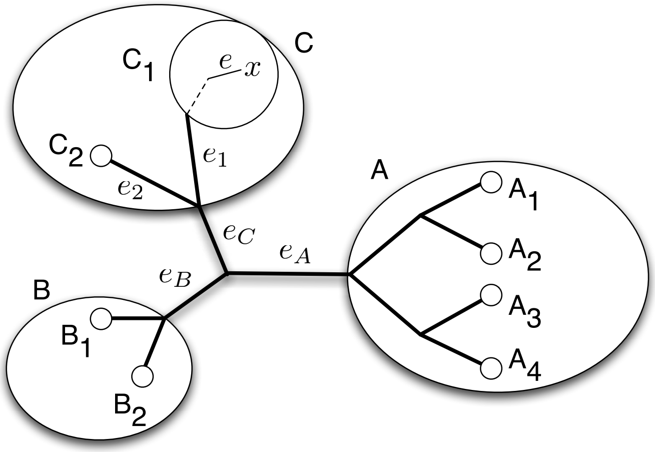

Figure 2: Configuration of the induction in the proof that IIP WLS models are semi-multiplicative.

C. Therefore the edge lengths corresponding to this map are all zero. This implies $\mu_e = 0$ for all $e \in E(C), e \neq e_1, e_2$, and also $\mu_{e_1} + \mu_{e_2} = 0$.

Let $A_1, \ldots, A_k$, with $k \leq 4$ and $B_1, \ldots, B_t$, with $t \leq 2$, be the subclades of A, respectively B, corresponding to nodes of $T^{/e_A}$. Then for any $x \in C_1$ and $y \in A_i$, and $z \in B_j$, $\Lambda_{xy} = \Lambda^{/e_A}_{C_1 A_i}$ does not depend on x, y and $\Lambda_{xz} = \Lambda^{/e_A}_{C_1 B_j}$ does not depend on x, z.

Now pick $x \in C_1$ and let e be the leaf edge adjacent to it. Then $\Lambda_e = 0$. Since all Lagrange multipliers are 0 inside the clade C_1, $\Lambda_e = \Lambda_{e \ldots e_1} = \Lambda_{e \ldots e_2} + \Lambda_{e \ldots e_c}$. Since $\mu_{e_1} + \mu_{e_2} = 0$, $\Lambda_{e \ldots e_2} = 0$. Thus $\Lambda_{e \ldots e_C} = \Lambda_{\{x\}A} + \Lambda_{\{x\}B} = 0$. Equivalently,

$$\sum_{i=1}^{k} Z_{\{x\},A_i} \Lambda^{/e_A}_{C_1 A_i} + \sum_{j=1}^{t} Z_{\{x\},B_j} \Lambda^{/e_A}_{C_1 B_j} = 0 \Leftrightarrow$$

$$Z_{\{x\},A} \sum_{i=1}^{k} \xi^{C_1}_{A_i A} \Lambda^{/e_A}_{C_1 A_i} + Z_{\{x\},B} \sum_{j=1}^{t} \xi^{C_1}_{B_j B} \Lambda^{/e_A}_{C_1 B_j} = 0 \tag{16}$$

This imposes a linear equation on $Z_{\{x\}A}$ and $Z_{\{x\}B}$ whose coefficients do not depend on x. Thus the following also does not depend on x:

$$\xi^{C_1}_{A(A \cup B)} = \frac{Z_{\{x\}A}}{Z_{\{x\}(A \cup B)}} = \frac{Z_{\{x\}A}}{Z_{\{x\}A} + Z_{\{x\}B}}. \tag{17}$$

4 An optimal algorithm for WLS edge lengths

Theorem 4.1. (Computing WLS edge lengths) *Let D be a dissimilarity map and V an IIP variance matrix. The set of all WLS edge lengths estimates for a tree T can be computed in $O(n^2)$ where n is the number of leaves in T.*

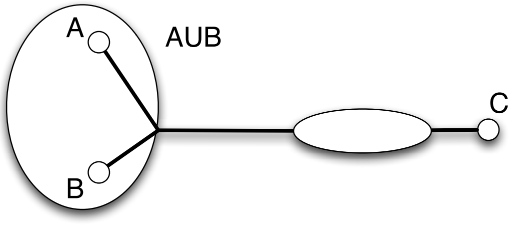

Figure 3: Configuration of the dynamic programming recursion for computing WLS edge lengths. A, B and $A \cup B$ are clades, and C is a clade disjoint from $A \cup B$.The oval in the middle represents the rest of the tree.

Proof: It is apparent from the proof of Theorem 3.4. that all one needs in order to compute the WLS edge lengths are the values of D_{AB} and Z_{AB}, where A and B are disjoint clades of T. We define the height of a tree to be the distance between its root and its farthest leaf, where the root is taken to be the closest endpoint of the edge defining the clade. Thus the height of a clade formed by just one leaf is 0.

Now consider the configuration in Figure 3. The clades A, B, C are all pairwise disjoint and A and B are adjacent. It is easy to see that $A \cup B$ form a clade for which

$$\begin{aligned} Z_{A\cup B,C} &= Z_{AC} + Z_{BC}, & (18) \\ D_{A\cup B,C} &= (D_{AC}Z_{AC} + D_{BC}Z_{BC})/Z_{A\cup B,C}. & (19) \end{aligned}$$

Therefore one needs only constant time to compute $D_{A\cup B,C}$ and $Z_{A\cup B,C}$ if D_{AC},Z_{AC},D_{CB} and Z_{CB} are known. Clearly, there are $O(n)$ clades since there are $O(n)$ edges, and thus there are $O(n^2)$ pairs of disjoint clades. We can compute D_{AB} and Z_{AB} for all pairs AB through a simple dynamic program. We start with pairs of trees of height 0, for which the values of D and Z are trivially given by δ and V^{-1}. After round $2t$ of the algorithm we will know D_{AB} and Z_{AB} for all disjoint pairs A, B of height at most t and after round $2t+1$ we know D_{AB} and Z_{AB} for all disjoint pairs A, B of height $t+1$ and t respectively. The algorithm clearly requires constant time per clade pair. Subsequently, all $O(n)$ edge lengths can be computed in constant time per edge: the calculation of each edge length involves only a constant number of multiplications and one matrix inversion (of size at most 13×13). Thus the algorithm is optimal since its running time is proportional to the size of the input.

We note that many algorithms have been proposed for computing WLS edge lengths for certain specific models (these are discussed in the next section). Existing approaches rely on different recursive schemes that lead to markedly different algorithms. Some attempt to reduce the size of the problem by agglomerating leaves ([4]); others start with a star topology and gradually extend it by refining internal nodes ([27]). In fact, all these methods implicitly compute Lagrange multipliers in a recursive way, and dealing directly with Lagrange multipliers may in many cases clarify the exposition and suggest simplified implementations. As we can see from the above theorem however, once one has the closed form expressions for the edge lengths, these inductive arguments can be easily replaced by our dynamic program.

5 The multiplicative model and other corollaries

In this section we begin by giving formulas for the WLS edge lengths assuming a a *tree-multiplicative* variance matrix, i.e. $V_{ij} = \prod_{e \in \overline{i,j}} w_e^{-1}$ for some $w : E(T) \to \mathbf{R}_+$. Throughout the section, $e^* \in E(T)$ denotes the edge for which the WLS length is being computed. If e^* is an internal edge then A, B, C, D are the adjacent clades. In the case that e^* is adjacent to a leaf, that leaf is labeled i and the adjacent clades A, B.

Proposition 5.1. *If V is a tree-multiplicative variance matrix then the WLS edge length of an internal edge is*

$$\begin{aligned} 2\hat{l}_{e^*} &= \frac{Z_{AD}+Z_{CB}}{Z_{A\cup B,C\cup D}}(D_{AC}+D_{BD}) \\ &+ \frac{Z_{AC}+Z_{DB}}{Z_{A\cup B,C\cup D}}(D_{AD}+D_{BC}) \\ &- D_{AB}-D_{CD}. \end{aligned} \tag{20}$$

If e^* is adjacent to a leaf then the WLS length is

$$2\hat{l}_{e^*} = D_{Ai} + D_{Bi} - D_{AB}. \tag{21}$$

At first glance these formulas may seem surprising, but the derivation is straightforward after solving for the Lagrange multipliers.

Proof: By the results of the previous section, it is enough to verify that the Lagrange equations hold. By Proposition 3.5. this is equivalent to verifying that the Lagrange equations hold for $T^{/e^*}$ and $V^{/e^*}$, which is a simple exercise left to the reader.

We now present a number of previous results about least squares that can be interpreted (and in some cases completed) using Theorems 3.4., 4.1., and Lemma 5.1.. All the models we discuss are special cases of the multiplicative variance model and all of our statements can be easily proven by substituting the appropriate form of V into (20,21).

Ordinary least squares.

This is the first model considered for least squares phylogenetics, and is the most studied model for edge and tree length estimation. It corresponds to the variance matrix equal to the identity matrix.

Corollary 5.2. (Rzhetsky [24]) *The ordinary least squares estimate* $p^tD = f_e^t(S_T^tS_T)^{-1}S_T^tD$ *for the length of edge e is given by*

$$\begin{aligned} 2\hat{l}_{e^*} &= \frac{n_An_D+n_Bn_C}{(n_A+n_B)(n_C+n_D)}(D_{AC}+D_{BD}) \\ &+ \frac{n_An_C+n_Bn_D}{(n_A+n_B)(n_C+n_D)}(D_{AD}+D_{BC}) \\ &- D_{AB}-D_{CD}, \end{aligned} \tag{22}$$

where n_A, n_B, n_C *and* n_D *are the number of leaves in the clades* A, B, C *and* D*, and* $D_{AC} = \sum_{a\in A, c\in C} D_{ac}$. *If* e^* *is a leaf edge,* $\hat{l}_e$ *is given by:*

$$2\hat{l}_{e^*} = D_{Ai}+D_{Bi}-D_{AB}. \tag{23}$$

Our algorithm for computing edge lengths (Theorem 4.1.) reduces, in the case of OLS, to that of [6]. It has the same optimal running time as the algorithms in [1, 10, 27].

Balanced minimum evolution.

The **B**alanced **M**inimum **E**volution model was introduced by Pauplin in [22]. The motivation was that in the computation of $\hat{l}_{e^*}$ in the OLS model, the distances D_{ac} and D_{bd} can receive different weights than D_{ad} and D_{bc} where $a \in A, b \in B, c \in C$ and $d \in D$. Pauplin therefore suggested an alternative model where all clades are weighted equally.

Corollary 5.3. (Pauplin's edge formula) *The WLS edge lengths with variance model* $V_{ij} \propto 2^{|\overline{i,j}|}$ *are given by* $\hat{l}_{e^*} = \frac{1}{4}(D_{AC} + D_{BD} + D_{AD} + D_{BC}) - \frac{1}{2}(D_{AB} - D_{CD})$ *for internal edges and* $\hat{l}_{e^*} = \frac{1}{2}(D_{Ai} + D_{Bi}) - \frac{1}{2}(D_{AB})$ *for edges adjacent to leaves.*

Proof: This corresponds to the multiplicative variance model with $w_e = 0.5$ for all edges e. One can easily show that in this case $Z_{AB} \propto 2^{-|\overline{A,B}|}$ and the result follows trivially from Theorem 3.4..

As far as we are aware, this is the first proof that the formulas given by Pauplin for edge lengths are in fact the WLS edge weights under the variance model described above. This implies:

Remark 5.4. *The edge weights of the neighbor-joining tree obtained from the standard reduction formula are equal to the weighted least squares edge length estimates under the BME model.*

This result is a companion to the the connection between Pauplin's tree length formula and WLS tree length under the BME model that was established by Desper and Gasquel in [5]. They proved the following:

Corollary 5.5. (Desper and Gascuel [5]) *The tree length estimator given by* $\hat{l} = \sum_{ab} D_{ab} 2^{1-p_{ij}}$ *is the minimum variance tree length estimator for the BME model. It is also identical to the one given by the coefficients* $p^t = f^t(S_T^t V^{-1} S_T)^{-1} S_T^t V^{-1}$.

Proof: The second part of the corollary follows trivially from Theorem 2.2.. The first part follows from a simple combinatorial argument by adding up the WLS edge lengths. Alternatively, one can notice directly that since $p_{ab} = 2^{1-p_{ij}}$, it follows that $p_{ab}V_{ab}$ is the uniform vector, and thus defines a T-additive map, corresponding to the star topology (equal-length leaf edges and zero-length internal edges). Finally, $\sum_{i,j} S_{ij,e} p = 1$ follows from an easy counting argument. Further elaboration on Remark 5.4. is beyond the scope of this paper.

The taxon-weighted variance model.

Another well known WLS model was introduced by Denis and Gascuel in [4]. Under this model we set $V_{ij} = t_i t_j$ for some $t_1, \ldots, t_n \in \mathbf{R}_+$. In the tree-multiplicative model, this corresponds to setting $w_e = 1$ for internal edges and $w_e = t_i$ when e is the leaf edge adjacent to leaf i. The paper [4] gives a beautiful proof for the statistical consistency of this model (which implies statistical consistency of OLS), and also provides an $O(n^2)$ algorithm for computing the WLS edge lengths. However, the algorithm is based on a recursive agglomeration scheme and an explicit formula for the edge lengths based on the values of D is not given. Such a formula follows from Theorem 3.4.:

Corollary 5.6. *For e an internal edge of T, the WLS edge length $\hat{l}_{e^*}$ is given by*

$$\begin{aligned} 2\hat{l}_{e^*} &= \frac{T_A T_D + T_C T_B}{(T_A + T_B)(T_C T_D)}(D_{AC} + D_{AC}) \\ &+ \frac{T_A T_C + T_D T_B}{(T_A + T_B)(T_C T_D)}(D_{AD} + D_{BC}) \\ &- (D_{AB} + D_{CD}) \end{aligned} \tag{24}$$

where $T_X = \sum_{x \in X} t_x$ and $D_{XY} = \sum_{x \in X, y \in Y} \frac{t_x t_y}{T_X T_Y} D_{xy}$. If e^* is adjacent to a leaf,

$$2\hat{l}_{e^*} = D_{Ai} + D_{Bi} - D_{AB}. \tag{25}$$

6 Final remarks

An important question is whether the variance matrices required for the IIP property to hold are realistic for problems where branch lengths are estimated using standard evolutionary models. In fact, semi-multiplicative matrices do not exactly capture the desired form of the variance, but they are good approximations. We illustrate this for the Jukes–Cantor model [17]:

Proposition 6.1. (Variance of distance estimates [2, 20]) *Let the random variable Y be the fraction of different nucleotides between two sequences of length n that are generated from the Jukes–Cantor process with branch length δ. Then the expected value of the empirical distance $D = -\frac{3}{4}\log\left(1 - \frac{4}{3}Y\right)$ is δ and its variance is*

$$Var(D) \approx \frac{3}{16n}\left(3e^{\frac{8}{3}\delta} + 2e^{\frac{4}{3}\delta} - 3\right). \tag{26}$$

This result can be extended to more general models. Since the branch lengths for an evolutionary model are tree-additive, this shows that for many regimes of the parameter δ, a tree-multiplicative model for variances is very reasonable. For a discussion on the statistics rationale behind least squares see [8].

Unfortunately, the Fitch–Margoliash assumption that the variance $V_{ij} = \mathrm{Var}(D_{ij}) \propto D_{ij}^2$ is inaccurate in light of (26), nor does it lead to IIP estimates since V is not semi-multiplicative. This means that for generic dissimilarity maps, the Fitch–Margoliash least squares estimates of edge lengths will depend on irrelevant distance estimates.

Another point that is important is that although it follows from Theorem 2.2. that for any V and f there is a unique BLUE p for $f^t l$, the converse of this statement is not true. For example, if p is BLUE for $f^t l$ with variance matrix V, then p is BLUE for $f^t l$ with variance matrix kV where $k \geq 0$. This is obvious because $S_T^t p$ remains the same, and kVp is a T-additive map if Vp is a T-additive map. However this point has more subtle (and serious) consequences:

Proposition 6.2. (Non-uniqueness of tree length) *The WLS estimated tree length with* $V = (c_1 + c_2(|\overline{i,j}| - 1))2^{|\overline{i,j}|}$ *does not depend on the constants* c_1 *and* c_2.

Proposition 6.2. has significance for the interpretation of the neighbor-joining algorithm. Based on [5], in [12] it is shown that neighbor-joining minimizes the balanced evolution criterion at each step. The criterion is argued to be statistically relevant by virtue of the fact that it is the BLUE for the tree length under the assumption that $V_{ij} \propto 2^{|\overline{i,j}|}$. Proposition 6.2. shows that there are many (significantly) different variance assumptions that yield the same tree length estimate. In fact, for some tree topologies, it is even possible that the OLS tree length is equal to the BME WLS tree length (for example for 5 taxa trees). This means that by minimizing the tree length some information about the variance is being discarded, and from this point of view the fact that the balanced minimum evolution criterion is equal to the BLUE tree length for multiple variance assumptions can be seen as a weakness of balanced minimum evolution methods, not a strength.

There are other issues that are important in least squares applications in phylogenetics that we have not mentioned in this paper. One obvious difficulty with applying WLS methods to tree length estimation is that the resulting estimators are tree-additive, and not necessarily tree-metrics. That is, there may be edge length estimates that are negative. A number of strategies for solving the non-negative WLS problem have been proposed [7, 11, 16, 18].

Our optimal algorithm for weighted least squares edge length estimates for multiplicative matrices is similar in spirit to a some of the algorithms in [1]. In fact, we believe that all the fast algorithms for WLS edge lengths can be understood within a single framework. The unifying concept is the observation that they all essentially estimate the Lagrange tree, either via a top-down, or bottom-up approach. We defer a detailed discussion of this to another paper. Finally, a key issue is that of consistency for specific forms of variance matrices assigned to all trees [4, 28]. An obvious question is what classes of semi-multiplicative variance matrices result in consistent tree estimates. A full discussion of this topic is also beyond the scope of this paper.

7 Acknowledgments

Radu Mihaescu was supported by a National Science Foundation Graduate Fellowship and partially by the Fannie and John Hertz foundation. Lior Pachter was supported in part by NSF grant CCF-0347992.

References

[1] D Bryant and P Waddell. Rapid evaluation of least squares and minimum evolution criteria on phylogenetic trees. *Mol. Biol. Evol.*, 15(10):1346 – 1359, 1998.

[2] D Bulmer. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8(6):868–883, 1991.

[3] L Cavalli-Sforza and A Edwards. Phylogenetic analysis models and estimation procedures. *Evolution*, 32:550–570, 1967.

[4] O Denis, F Gascuel. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127:63–77, 2003.

[5] R Desper and O Gascuel. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3):587–98, 2004.

[6] R Desper and M Vingron. Tree fitting: topological recognition from ordinary least-squares edge length estimates. *Journal of Classification*, 19:87–112, 2002.

[7] J Felsenstein. An alternating least-squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46:101–111, 1997.

[8] J Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Inc., 2003.

[9] WM Fitch and E Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[10] O Gascuel. Concerning the NJ algorithm and its unweghted version, UNJ. In *Mathematical Hierarchies and Biology*, volume V, pages 149–170. American Mathematical Society, 1997.

[11] O Gascuel and D Levy. A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *Journal of Classification*, 13:129–155, 1996.

[12] O Gascuel and M Steel. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, 2006.

[13] J Gill, S Linusson, V Moulton, and M Steel. A regular decomposition of the edge-product space of phylogenetic trees. *Advances in Applied Mathematics, in press*, 2008.

[14] JA Hartigan. Representation of similarity matrices by trees. *Journal of the American Statistical Association*, 62:1140–1158, 1967.

[15] K Hayes and J Haslett. Simplifying general least squares. *The American Statistician*, 53(4):376–381, 1999.

[16] LJ Hubert and P Arabie. Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology*, 48:281–317, 1995.

[17] TH Jukes and C Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press, 1969.

[18] V Makarenov and B Leclerc. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification*, 16:3–26, 1999.

[19] G Matheron. *Les Variables Regionalisés et Leur Estimation.* Paris: Mason, 1962.

[20] M Nei and L Jin. Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution*, 6:290–300, 1989.

[21] AN Patrinos and SL Hakimi. The distance matrix of a graph and its tree realization. *Quarterly Journal of Applied Mathematics*, 30:255–269, 1972.

[22] Y Pauplin. Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.*, 51:41–47, 2000.

[23] P Ray. Independence of irrelevant alternatives. *Econometrica*, 41:987–991, 1973.

[24] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.*, 10:1073–1095, 1993.

[25] C Semple and M Steel. Cyclic permutations and evolutionary trees. *Applied Mathematics*, 32:669–680, 2003.

[26] C Semple and M Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[27] W Vach. *Least squares approximation of additive trees in Conceptual and Numerical Analysis of Data, O. Opitz (ed)*, pages 230–238. Springer, Heidelberg, 1989.

[28] SJ Willson. Consistent formulas for estimating the total lengths of trees. *Discrete Applied Mathematics*, 148:214–239, 2005.