Maximum Likelihood Estimates for Binary Random Variables on Trees via Phylogenetic Ideals

Robin Evans

Abstract

In their 2007 paper, E.S. Allman and J.A. Rhodes characterise the phylogenetic ideal of general Markov distributions for binary data on the leaves of a tree. Further, as yet unpublished, results by Zwiernik, Maclagan and Smith enable us to check simple inequality constraints to ensure the joint probabilities are valid for the model. We investigate the use of these results in the finding of maximum likelihood estimates for the joint distribution. We demonstrate a method of finding the MLEs using the ideals for the quartet tree, and make a comparison with the EM algorithm.

1 Introduction

This paper discusses how phylogenetic invariants might be used to fit general Markov models for binary data. Some authors, for example Pachter and Sturmfels [12], have contended that all statistical models are algebraic varieties, meaning that they can be characterised by a set of algebraic constraints on the joint probabilities. For the purposes of most statisticians, one also requires a set of semi-algebraic, or inequality, constraints to ensure, for example, that the probabilities are real and lie on $[0, 1]$.

In the case of binary data following the general Markov model on a tree, these algebraic constraints are available explicitly, due to Allman and Rhodes [2]. Futher, Zwiernik et al. [17] allow us to characterise the additional semi-algebraic constraints required to ensure that we are inside the general Markov model. This paper discusses how the algebraic constraints may be used to fit a model, and compares this approach to a standard EM algorithm as used by, for example, Felsenstein [6].

We begin in section 2 by setting up the model, and mentioning a biological context for phylogenetic trees. Section 3 contains the Allman and Rhodes result, as well as the main contribution of this paper: a method of fitting models directly using this result. Section 4 outlines the fitting the maximum likelihood estimators via the EM algorithm and message passing. We introduce semi-algebraic constraints on the model in section 5, and discuss how these might be applied to the direct fitting. In section 6 we compare experimentally the two approaches to fitting, and discuss the practicalities. Section 7 suggests the direction of future work, whilst section 8 contains some proofs which were not considered informative enough to be in the main body of the paper.

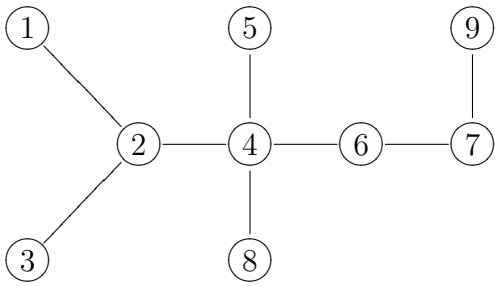
2 Preliminaries

The Global Markov Property

We being by showing, in the case of a tree, an equivalence between models staisfying the global Markov property on directed acyclic graphs (DAGs) and undirected graphs.

Definition 1
A *tree* is a connected acyclic graph. The leaves of a tree are the vertices of order one (i.e. with exactly one incident edge).

The picture below is instructive; for a more formal approach see, for example, Bollobás [4].



The leaves in this example are the nodes 1, 3, 5, 8 and 9. We refer to other vertices as *internal*.

Lemma 1
Let $T = (V, E)$ be a tree, and let $i, j \in V$ be distinct vertices of the tree. Then there exists a unique path $\pi_{i,j}$ from i to j.

Proof. See section 8. □

We will assume that the distribution of the random variables follows the *global Markov property* (GMP); for a discussion see, for example, Jensen [9]. When this holds, we say that the distribution is in the *general Markov model* on the tree. The lemma above shows that the GMP will yield the same independence structure on an undirected tree as a rooted directed one: given an undirected tree $T = (V, E)$, we can pick any vertex $r \in V$ to be the *root* of our tree, and direct all edges in such a way as they point away from r. Then notice that there are no colliders on the directed graph.

In light of this, we can use undirected and (rooted) directed trees interchangeably from now on; in either case, for any two vertices i and j with associated random outcomes Y_i and Y_j, we have $Y_i \perp\!\!\!\perp Y_j | D$, where D is any subset of the vertices on the unique path $\pi_{i,j}$ from i to j. Intuitively then, if we know the outcome of any vertex 'in between' i and j, knowledge of Y_i will tell us nothing further about Y_j, and vice versa.

Bifurcating Trees and Binary Data

From now on we restrict our attention to trees whose vertices are binary random variables, taking states 0 and 1. We denote the random variables corresponding to the leaves of the tree by $X_1, \ldots, X_n$, and those corresponding to internal vertices ('hidden' variables) by $H_1, \ldots, H_{m-n}$. In cases where we do not wish to distinguish between leaves and internal vertices, we use the notation $Y_1, \ldots, Y_m$ instead. We assume that no vertex has order 2; it is not difficult to see that such vertices create identifiability issues.

Definition 2
A *bifurcating* tree is a tree in which all internal vertices have order three. Such trees are sometimes referred to as *binary trees*, but this causes confusion with trees whose vertices correspond to binary random variables, so we avoid this terminology.

We use the notion of a bifurcating tree to find an upperbound on the dimension of the space of distributions over a tree.

Lemma 2
A bifurcating tree with n leaves has exactly $n-2$ internal vertices and $2n-3$ edges.

Proof. See section 8. □

Lemma 3
An m-variate binary distribution P over a tree (not just the leaves) and satisfying the GMP can be characterised by $2m-1$ parameters.

Proof
Let $\boldsymbol{Y}$ be a random variable sampled from P, and suppose it satisfies the tree structure for some tree T of order m. Let $Y_1, \ldots, Y_m$ be the variables of the tree; pick the vertex corressponding to Y_1 as the root, and order the Y_is so that for each i, the parent of i, pa(i) comes before i in the sequence. Then the joint distribution is

$$p(Y_1 = y_1, \ldots, Y_m = y_m) = p(Y_1 = y_1)p(Y_2 = y_2|Y_1 = y_1) \cdots p(Y_m = y_m|Y_1 = y_1, \ldots, Y_{m-1} = y_{m-1}).$$

But since each Y_j is followed by all its descendents, it is separated from all its predecessors in the list by its parent, and thus

$$p(Y_1 = y_1, \ldots, Y_m = y_m) = p(Y_1 = y_1) \prod_{i=2}^{m} p(Y_i = y_i|Y_{\mathrm{pa}(i)} = y_{\mathrm{pa}(i)}). \quad (1)$$

But then $p(Y_1 = y_1)$ is characterised by the single parameter $p(Y_1 = 1)$, and for each i, $p(Y_i = y_i|Y_{\mathrm{pa}(i)} = y_{\mathrm{pa}(i)})$ is characterised by the two parameters $p(Y_i = 1|Y_{\mathrm{pa}(i)} = 1)$ and $p(Y_i = 1|Y_{\mathrm{pa}(i)} = 0)$. Thus we have at most $1 + 2(m-1) = 2m-1$ parameters. □

It is clear that these parameters are free to be chosen as any value between 0 and 1, so the space of binary distributions over the whole tree has dimension $2m-1$. This means that a bifurcating tree

with n leaves (and therefore $m = n + n - 2 = 2n - 2$ vertices in total) is characterised by *at most* $4n - 5$ parameters. Note that the 'at most' here is important, since it is possible that two different distributions over then entire structure of the tree could give the same marginal distribution on the leaves. By way of comparison, the full set of binary distributions over n variables has dimension $2^n - 1$. For even moderate values of n, it becomes clear that we are working over a much smaller space of distributions.

It is easy to extend the bound to non-bifurcating trees, because we can create a general tree (with no vertices of order 2) by collapsing edges from a bifurcating tree. Thus a model over any tree with n leaves is described by at most $4n - 5$ parameters.

It, of course, possible to extract the marginal distibution of the leaves by summing over the possible outcomes of the internal vertices; however, this involves 2^{n-2} terms for a tree with n leaves, and is computationally infeasible.

A Biological Interpretation

The concept of a phylogenetic tree has its origins in evolutionary biology, as the name suggests. In this context, it is a 'tree of life', where leaves represent extant species, and internal nodes are ancestor species at which an evolutionary split occured. Applying the indepedence structure of the tree from the GMP implies that the genome of two species is independent, conditional on the genome of their most recent common ancestor (or indeed any other species on the path in between).

In genetics, the data of interest concerns sequences of the four possible base pairs, usually labelled A, C, G and T in reference to the chemicals involved. In a highly idealised model, we could assume that there are n extant species, where the ith species is defined by a random sequence of these base pairs $(X_{i1}, \ldots, X_{iN})$ with common length N. Then letting $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})$ be the vector made up of the jth base pair for each species, we suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ are independent and identically distributed (i.i.d.) random variables, where the common joint distribution obeys the independence structure of the tree under the GMP.

Of particular interest to evolutionary biologists is the evolutionary 'distance' between two species. As we will see in section 5, this can be related neatly to the correlation between random variables.

In practice of course, base pairs do not change independently—whole sequences of DNA are routinely moved, inverted or removed in the process of evolution, and processes such as horizontal gene transfer will violate the tree structure. However, this model has been observed to work well in practice; see, for example, Hodge and Cope [8].

3 A Method of Direct Fitting

Working in a frequentist context, it is natural to ask which of the possible distributions in the model maximises the likelihood for the data. We will assume from here on that the topology of the tree is known, and that it is known which leaves are which, with respect to the tree structure.

A very hard problem is to choose the best topology and leaf labelling in the first place; see, for example, Strimmer and von Haeseler [16] for some ideas.

A direct approach would try to characterise the set of distributions on n-variables which can be induced by a tree structure. From an algebraic perspective, we are interested in finding *phylogenetic invariants*, polynomials in the joint probabilities which are constrained to be zero by the model structure. As a trivial example, we know that

$$\sum_{i_1=0}^{1} \cdots \sum_{i_n=0}^{1} p(X_1 = i_1, \ldots, X_n = i_n) - 1 = 0 \tag{2}$$

for any distribution. Allman and Rhodes [2] show that the set of phylogenetic invariants is fully characterised by edge-flattenings of the tree structure. Essentially we are interested in the tensor

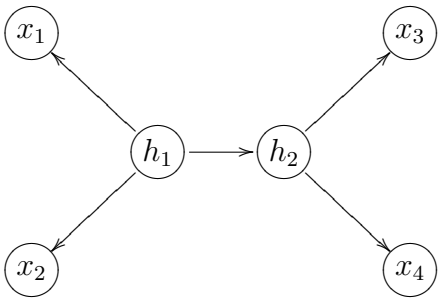
$$p_{i_1 \cdots i_n} = p(X_1 = i_1, \ldots, X_n = i_n)$$

where $i_1, \ldots, i_n \in \{0, 1\}$. The complete set of equations such as (2) is called the *phylogenetic variety*; together with some semi-algebraic (inequality) constraints, the variety determines the model completely.

Definition 3
Let $e = \{h_1, h_2\} \in E$ be an internal edge of the tree T, meaning that both h_1 and h_2 are internal vertices. Let a_1 and a_2 be the two subsets of leaves in T which are separated by e. Then the *edge-flattening* of T with respect to e is created by considering the joint distribution of a_1, h_1 and a_2 as a simplified tree structure (note that we can choose h_2 instead of h_1, this is arbitrary).

The matrix P_e associated with the edge-flattening is given by writing the entries of the tensor $p_{i_1 \cdots i_n}$ in such a way that each row represents a particular state of a_1, and each column a particular state of a_2.

Example 1
Consider the quartet tree shown below.



The only internal edge is $e = \{h_1, h_2\}$, and the associated flattening is given by

$$(x_1, x_2) \text{ —— } h_1 \text{ —— } (x_3, x_4)$$

The flattening matrix is given by

$$P_e = \left[p_{(a_1)(a_2)}\right] = \begin{pmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{pmatrix}.$$

Note that the entire first row, for example, corresponds to $(X_1, X_2) = (0, 0)$. A further example for the tree with 5 leaves is found in section 7.

Now, we can characterise the new distribution on this simplified tree by writing

$$p(a_1, a_2, h_1) = p(h_1)p(a_1|h_1)p(a_2|h_1)$$
$$p(a_1, a_2) = \sum_{i=0}^{1} p(h_1 = i)p(a_1|h_1 = i)p(a_2|h_1 = i),$$

and in fact

$$P_e = M_1^T \begin{pmatrix} p(h_1 = 0) & 0 \\ 0 & p(h_1 = 1) \end{pmatrix} M_2$$

where

$$M_1 = \begin{pmatrix} p(x_1 = 0, x_2 = 0|h_1 = 0) & \cdots & p(x_1 = 1, x_2 = 1|h_1 = 0) \\ p(x_1 = 0, x_2 = 0|h_1 = 1) & \cdots & p(x_1 = 1, x_2 = 1|h_1 = 1) \end{pmatrix}$$

and M_2 is defined similarly for X_3 and X_4. And in the general case of an edge-flattening where a_i contains k_i leaves, M_i will be a 2×2^{k_i} matrix. Thus the rank of P_e is at most 2. This is the key observation, since elementary linear algebra tells us that every 3×3 minor of P_e must therefore be zero. This defines our phylogenetic variety. The following result is Theorem 4 of Allman and Rhodes [2].

Theorem 4
For a bifurcating tree $T = (V, E)$ whose vertices take two states and follow the GMP, the phylogenetic ideal is generated by the 3×3 minors of all edge-flattenings of the tensor $p_{i_1 \cdots i_n}$.

We have showed the necessity of these minors being zero; the theorem shows their sufficiency in determining the model, up to inequality constraints. This motivates the question of whether the variety can be used to fit the tree model to data.

The Quartet Tree

A plausible procedure for fitting valid joint probabilities in the case of the quartet tree is as follows. Suppose we pick any twelve positive numbers to fill the first two rows and columns of P_e:

$$P_e = \begin{pmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & & \\ p_{1100} & p_{1101} & & \end{pmatrix}.$$

Let $M_{(abc)(def)}$ be the minor generated by the rows a, b, c and columns d, e, f. Then (ignoring singularities for the moment), we can fix $(P_e)_{3,3} = p_{1010} = q$ so that the minor $M_{(123)(123)}$ is zero:

$$M_{(123)(123)} = \begin{vmatrix} p_{0000} & p_{0001} & p_{0010} \\ p_{0100} & p_{0101} & p_{0110} \\ p_{1000} & p_{1001} & q \end{vmatrix} = 0.$$

This amounts to simply solving a linear equation in one variable. Similarly, we can fix $(P_e)_{3,4}$ so that $M_{(123)(124)} = 0$, and $(P_e)_{4,3}$ and $(P_e)_{4,4}$ so that $M_{(124)(123)} = 0$ and $M_{(124)(124)} = 0$ respectively. At this point we have filled each of the entries of P_e, but we have only set 4 of the 16 possible 3×3 minors to be zero. The following results shows that we are in fact done. We use $P_{(i_1 \cdots i_k)(j_1 \cdots j_l)}$ to denote the submatrix of P formed by taking entries in the rows $i_1, \ldots, i_k$ and the columns $j_1, \ldots, j_l$. (Thus $M_{(abc)(def)} = |P_{(abc)(def)}|$)

Lemma 5
Let P be an $m \times n$ matrix with $m, n > 2$, and pick any values P_{ij} for the top two rows and columns ($i \leq 2$ or $j \leq 2$) of P such that none of the 2×2 minors specified by the top two rows and columns

$$M_{(12)(j_1 j_2)}, \quad M_{(i_1 i_2)(12)} \qquad i_1, i_2 = 1, \ldots, m,\ j_1, j_2 = 1, \ldots, n$$

are zero. Then we can choose the remaining entries of P to ensure that every 3×3 minor of P is zero.

Proof
We can easily pick values as above to ensure that $M_{(12i)(12j)} = 0$ for each i, j. To see this

$$M_{(12i)(12j)} = \begin{vmatrix} P_{11} & P_{12} & P_{1j} \\ P_{21} & P_{22} & P_{2j} \\ P_{i1} & P_{i2} & q \end{vmatrix}$$

so

$$q = \begin{pmatrix} P_{i1} & P_{i2} \end{pmatrix} \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}^{-1} \begin{pmatrix} P_{1j} \\ P_{2j} \end{pmatrix}$$

where the inverse exists by the assumption concerning 2×2 minors. Now note that $M_{(12i)(12j)} = 0$ implies that there is a non-trivial linear multiple of the columns $P_{(12i)(1)}, P_{(12i)(2)}, P_{(12i)(j)}$ of the minor which vanishes;

$$\lambda P_{(12i)(1)} + \mu P_{(12i)(2)} + \nu P_{(12i)(j)} = 0$$

for some constants λ, μ, ν not all zero. But the condition on the minors $M_{(12)(12)}, M_{(12)(1j)}, M_{(12)(2j)}$ means that all three constants must be non-zero, and hence we can re-write $P_{(12i)(j)}$ as a linear multiple of the other two columns.

But then we can write $P_{(12i)(j)}$ as a linear multiple of $P_{(12i)(1)}, P_{(12i)(2)}$ for all j including, trivially, $j = 1, 2$. Thus any 3×3 submatrix $P_{(12i)(jkl)}$ has the three columns as linear multiples of only two vectors, and hence the minor $M_{(12i)(jkl)}$ is zero.

Then for a general submatrix $P_{(i_1i_2i_3)(j_1j_2j_3)}$, each column is a linear multiple of $P_{(i_1i_2i_3)(1)}$ and $P_{(i_1i_2i_3)(2)}$, so the same result applies. □

Thus in the quartet model, we could choose 12 values of the joint distribution tensor to put in the top two rows and columns, and then fill in the rest to satisfy the phylogenetic ideal. In order to then ensure that the joint probabilities summed to 1, we could divide by the sum of the entries to normalise, preserving the property of the 3×3 minors.

Note that this procedure does not guarantee that the resulting values will be the joint probabilities of a model satisfying the general Markov property on a tree; indeed, they could be negative. In practice, however, it appears to work well.

4 The EM Algorithm

The EM algorithm, or Expectation Maximisation algorithm, is a method of finding the maximum likelihood estimator of a set of parameters where data is unobserved or missing; often the likelihood of the observed data is difficult to maximise. In some cases this missing data may have a physical or intuitive significance in terms of the problem, in others it may simply be chosen for convenience.

Application of the EM algorithm to fitting graphical models is discussed in detail in Lauritzen [10]. The EM algorithm, as formalised by Dempster et al. [5], is aimed at precisely the kind of problem we find here: the likelihood over the leaves is very difficult to characterise, and therefore maximising is very difficult; however the likelihood of the whole tree is very simple, as shown in (1).

In our case, we have observed data over the leaves of our tree, and wish to find the tree structure which maximises the likelihood. However, the likelihood for the leaves can only be obtained by summing over the 2^{n-2} possible values of the binary random variables assumed to exist at the internal vertices, which quickly becomes computationally difficult.

Conversely however, the *complete* likelihood of the entire tree has the simple form given in (1), and the joint distirubtion could be estimated very easily. Thus we will impute the 'missing' data at the internal vertices, in the form of counts.

Method

1. Pick some starting values of the unknown parameters of interest, θ_0; $n = 0$.
2. **E-step**: given the observed data X and current parameter estimate θ_n, impute the missing data Y by $\mathbb{E}_{\theta_n}[Y|X]$.
3. **M-step**: maximise the 'complete' likelihood of the observed and missing data with respect to θ, and call the MLE θ_{n+1}.
4. $n \leftarrow n + 1$; go to step 2 and repeat until some specified convergence has been reached.

In our case, the observed data are counts of the values at the leaves of the tree, and the missing data is the expected division of these counts into the 'complete counts' over the entire tree.

The EM algorithm has some good convergence properties; in particular the likelihood is guaranteed not to decrease after an iteration, which in complete likelihoods without local maxima is usually enough for convergence to the MLE. It is well known that the EM algorithm can converge very slowly, and indeed the number of iterations required was observed to increase significantly in n. Further, the E-step in our example is not at all trivial, and requires us to use a message-passing algorithm to find the conditional probabilities for each of the 2^n possible outcomes observed at the leaves; consequently the computational difficulty of each iteration increases exponentially with the number of nodes.

5 Inequality Constraints

Whilst we have shown that all models in the space must satisfy the equality constraints imposed by the edge-flattenings, and Allman and Rhodes have shown that no other algebraic constraints are required, these conditions are not sufficient to guarantee that we are in a valid tree model. The following theorem due to unpublished work by Zwiernik et al. [17] completes the characterisation of the model.

Theorem 6
Let T be a tree under the binary general Markov model. Let P be the joint distribution on the leaves $X_1, \ldots, X_n$, and ρ_{ij} be the correlation between outcomes at leaves i and j under P. If the joint distribution of P satisfies the edge-flattening equality constraints of Theorem 4, then P is a member of the binary general Markov model if and only if:

(i) for all (not necessarily distinct) leaves $i, j, k, l \in \{1, \ldots, n\}$,

$$|\rho_{ij}\rho_{kl}| \geq \min\left\{|\rho_{ik}\rho_{jl}|, |\rho_{il}\rho_{jk}|\right\};$$

(ii) for all distinct $i, j, k \in \{1, \ldots, n\}$,

$$\rho_{ij}\rho_{jk}\rho_{ki} > 0.$$

Zwiernik et al. [17] also mention an important parallel between this result and a tree metric. First we need a lemma.

Lemma 7
Let T be a tree under the binary general Markov model, and let ρ_{ij} denote the correlation between the random variables at vertices i and j. Suppose further that k lies on the unique path between i and j. Then

$$\rho_{ij} = \rho_{ik}\rho_{kj}.$$

Proof. Can be found in slides of Zwiernik et al. [17]. □

Now, let the distance d_{ij} between two vertices i and j be defined by $d_{ij} = -\log|\rho_{ij}|$; it is easy to verify that this is a metric using the lemma. Thus we can think of the 'distance' between two vertices simply as a measure of how uncorrelated they are.

6 Results

All programming was implemented in `R`, making use of the `bindata` package to simulate data.

The direct approach was implemented for the quartet model as outlined above. The `nlm` function in `R` was then used to maximise the likelihood over the 12 'free' parameters; the likelihood was given a near infinite negative value if any joint probabilities were fitted as negative by the procedure. Starting values were chosen by perturbing the uniform distribution on the leaves—the perturbation made zero 2×2 minors unlikely.

`nlm` is something of a 'black box' function, and it frequently fails to find the global maximum (presumably by getting stuck in a local maximum). More work is needed to produce an algorithm which does not fail so easily. This should not be too difficult; using a variety of starting points and comparing answers should work well. Choosing a different starting value for the same data always resulted in it finding the correct answer.

In most cases the solution arrived at was identical to that given by the EM algorithm, but in a significant minority some local maximum was reached instead. In a few cases with small sample sizes, the direct method was observed to give a higher likelihood than the EM; checking by the conditions of Theorem 6 revealed these solutions to be outside the model.

We ran a trial comparison of 50 sets of data, simulated from a randomly chosen[1] general Markov model on the tree. The EM algorithm was run until the log-likelihood changed by less than 0.001; in general this resulted in the EM reaching a maximum approximately 0.02 less than that found by `nlm`. In 13 cases `nlm` failed to find the best answer, so these were discarded. However, in the remaining 37 cases where both algorithms converged to the same solution, the direct approach with `nlm` was generally faster. Figure 1 shows actual run times for each of these trials, and EM does significantly worse overall.

We also ran an experiment to demonstrate the problems with the EM algorithm. Figure 2 shows a log-log plot of actual time taken by a computer to run EM for various values of n. The algorithm was run until the log-likelihood changed by less than 0.01. Whilst the actual times taken for $n = 10$ were impractically large—around 7 minutes—more worrying is that the slope of increase appeared to be around 6.8, indicating a polynomial relationship of approximately order 7. Interestingly the relationship did not appear to be exponential, which is surprising given that we have to run the message passing for each of 2^n possible datasets—perhaps a study with larger values of n would reveal that the relationship was, in fact, non-polynomial. It may be that the criterion for convergence was flawed, since EM moves very slowly with each iteration for large n, even when

[1] Here we use random in the sense that the distances from section 5 were generated with a gamma distribution.

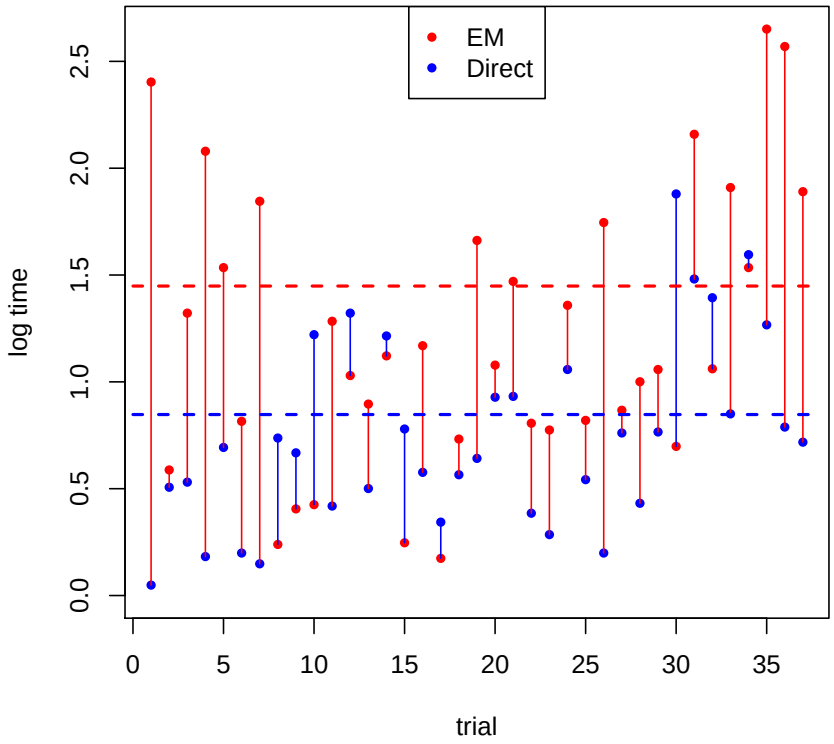


Figure 1: log run time (in seconds) for both methods over 37 different trials. The abundance of red lines indicates that EM was usually the slower method. Dashed lines are mean run times.

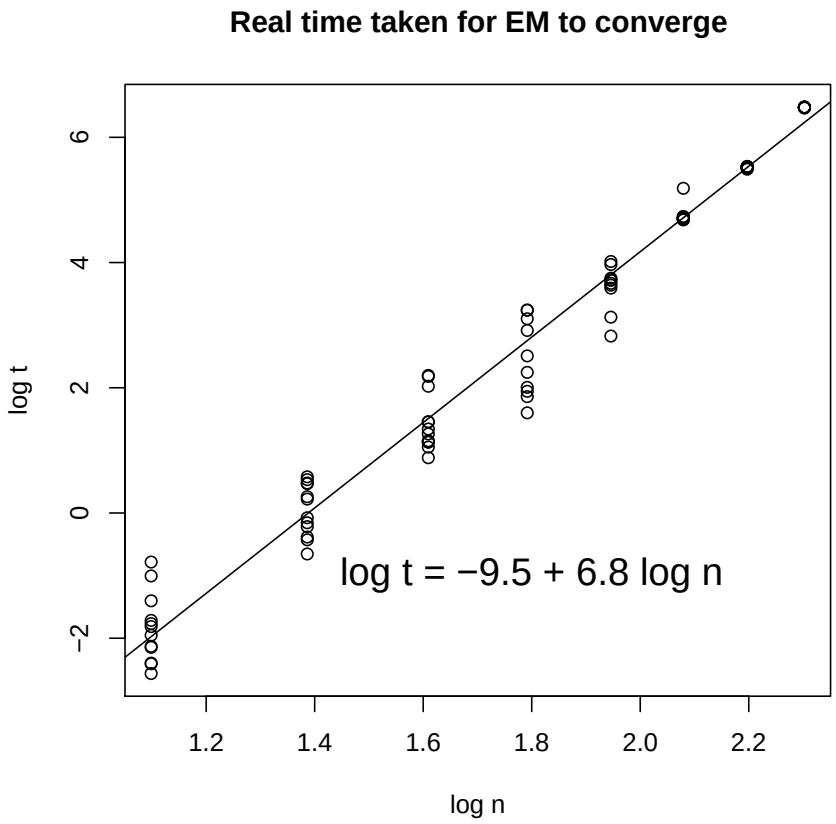


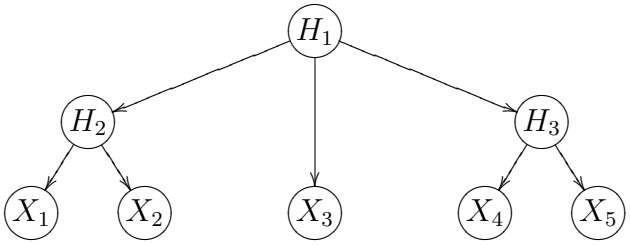
Figure 2: log of actual time taken against $\log n$ for various runs of the EM algorithm. The actual values of n are from 3 to 10, and times range from about 0.1 seconds to 7 minutes.

some distance from the final answer. In any case, EM appears to be impractically slow.

7 Extensions

The 5-Taxon Tree

Our programming so far only extends to the quartet tree, which is highly restrictive. An important objective should be to fit models directly for any size tree. The 5 leaved tree is shown below.



Here there are two edges to flatten $\{h_1, h_2\}$ and $\{h_1, h_3\}$, and hence we have two edge-flattening matrices; these are given below.

$$\begin{pmatrix} p_{00000} & p_{00100} & p_{01000} & p_{01100} & p_{10000} & p_{10100} & p_{11000} & p_{11100} \\ p_{00001} & p_{00101} & p_{01001} & p_{01101} & p_{10001} & p_{10101} & p_{11001} & p_{11101} \\ p_{00010} & p_{00110} & p_{01010} & p_{01110} & p_{10010} & p_{10110} & p_{11010} & p_{11110} \\ p_{00011} & p_{00111} & p_{01011} & p_{01111} & p_{10011} & p_{10111} & p_{11011} & p_{11111} \end{pmatrix}$$

$$\begin{pmatrix} p_{00000} & p_{01000} & p_{10000} & p_{11000} \\ p_{00001} & p_{01001} & p_{10001} & p_{11001} \\ p_{00010} & p_{01010} & p_{10010} & p_{11010} \\ p_{00011} & p_{01011} & p_{10011} & p_{11011} \\ p_{00100} & p_{01100} & p_{10100} & p_{11100} \\ p_{00101} & p_{01101} & p_{10101} & p_{11101} \\ p_{00110} & p_{01110} & p_{10110} & p_{11110} \\ p_{00111} & p_{01111} & p_{10111} & p_{11111} \end{pmatrix}$$

The following is an idea of how we might extend the method used for the quartet tree.

We specify the dark entries above, and choose the greyed out entries to ensure that the 3×3 minor conditions hold. First choose $p_{01010}, p_{01011}, p_{01110}$, $p_{01111}, p_{10010}, p_{10011}$ using the first matrix, in the same manner as was used for the quartet tree; then add these entries to the appropriate place in the second matrix. But now the top two rows and left-most two columns of the second matrix have been completed, and so we can use it to find all the remaining values to force the minors to be zero.

It is not too difficult to see, from the 'stacking' relationship between the columns of the two matrices that this means that all 3×3 minors of both matrices are now zero. The nice feature of this is

that we seem to have specified the 'correct' number of parameters: recall that by Lemma 3, an upper bound on the dimension of this model is 15. The method above allows 16 'free' parameters, but we normalise at the end to ensure the sum is 1; hence in some sense we are maximising in a 15 dimensional space.

Each time we add a taxon (leaf) to the tree, we get another edge-flattening, and hence another matrix whose 3×3 minors must be zero. The method above of 'shuffling' between the matrices to find solutions does not obviously generalise to larger trees, and hence it is not clear how to automate it. A computationally simple solution to this problem would be a key objective of further research.

Topology and Model Selection

A question we have so far largely ignored in this paper is how to choose the best tree topology to use. Ideally, one would want to maximise the likelihood over the set of all possible trees, both by changing the topology of the tree, and by changing the labelling of the leaves. One suggestion made by Allman and Rhodes [2] is to first maximise the likelihood naïvely over the space of all distributions. The joint probabilities yielded by this process could then be plugged into the phylogenetic invariants for a particular topology, and assuming that the model is correct, the solutions should be close to zero.

It is not currently understood how these empirical phylogenetic invariants behave under the 'null' distribution, and results in this area would be extremely useful, possibly allowing traditional hypothesis testing. Similarly, it would be useful to understand their behaviour under misspecification, so we could measure the 'power' of such a test.

Concluding Remarks

The main contribution of this paper has been an explicit attempt to fit this set of models to data directly, using the Allman and Rhodes result. Further work should focus on making this technique reliable, particularly in using a better algorithm for maximising the likelihood. Extending the method to trees with an arbitrary number of taxa would be a prerequisite for this method to become useful in practice.

8 Proofs

Proof of Lemma 1
Suppose not for contradiction. Then there are two distinct paths $\pi = (\pi_0, \pi_1, \ldots, \pi_m)$ and $\pi' = (\pi'_0, \pi'_1, \ldots, \pi'_n)$, with $\pi_0 = \pi'_0 = i$ and $\pi_m = \pi'_n = j$, which connect i and j. Let $p = \inf\{k \geq 1 : \pi_{k+1} \neq \pi'_{k+1}\}$, the last vertex before the paths first diverge, which must exist since the paths are non-equal. Similarly, let $q = \inf\{k \geq p+1 : \pi_k = \pi'_l \text{ for some } l \geq p+1\}$, the next point after π_p at which the paths meet, which again must exist since the paths certainly meet again at j. But

then the path given by

$$\pi_p \to \pi_{p+1} \to \cdots \to \pi_q = \pi'_l \to \pi'_{l-1} \to \cdots \to \pi'_p$$

is a cycle; this is a contradiction. □

Proof of Lemma 2
We proceed by induction: for $n = 2$ this is trivial. Suppose the result if true for $n = k - 1$ and let $T = (V, E)$ be a tree with k leaves. Pick a leaf $x \in V$; then x has only one incident edge $e_1 = \{x, y\}$, and y must be an internal node (else x and y are the only vertices in the tree). Denote the other edges incident to y by e_2, e_3. Then by removing x, y, and the three mentioned edges, and joining y's two remaining neighbours by a new edge e^*, we have created a new tree $T' = (V \setminus \{x, y\}, (E \cup \{e^*\}) \setminus \{e_1, e_2, e_3\})$ with $k - 1$ leaves. Thus T' has $k - 3$ internal vertices and $2k - 5$ edges by the induction hypothesis, and so T has $k - 2$ internal vertices and $2k - 3$ edges. □

References

[1] E.S. Allman and J.A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186:113–144, 2003.

[2] E.S. Allman and J.A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40(2):127–148, 2007.

[3] A.H. Andersen. Multidimensional contingency tables. *Scan. J. of Stat.*, 1:115–27, 1974.

[4] B. Bollobás. *Graph Theory: An Introductory Course*. Springer-Verlag, 1979.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39(1):1–38, 1977.

[6] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Ev.*, 17:368–376, 1981.

[7] I.B. Hallgrímsdóttir, R.A. Milowski, and J. Yu. *The EM Algorithm for Hidden Markov Models*, pages 250–262. 2008.

[8] T. Hodge and M.J. Cope. A myosin family tree. *J. of Cell Sci.*, 113(19):3353–3354, 2000.

[9] F.V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[10] S.L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1992.

[11] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. B*, 50(2):157–224, 1988.

[12] L. Pachter and B. Sturmfels. Tropical geometry of statistical models. Technical report, Department of Mathematics, University of California, Berkeley, 2008.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[14] T.S. Richardson. Analysis of the binary instrumental variable model. Talk given January 7th and February 18th, University of Washington, 2009.

[15] D.J. Spiegelhalter. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

[16] K. Strimmer and A. von Haeseler. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *J. Mol. Evol.*, 13(7):964–969, 1996.

[17] P. Zwiernik, D. Maclagan, and J.Q. Smith. Geometry of phylogenetic tree models for binary data. Video of talk and accompanying slides can be found at www2.warwick.ac.uk/fac/sci/statistics/staff/research_students/zwiernik/.