Maximum likelihood estimation and natural pairwise estimating equations are identical for three sequences and a symmetric 2-state substitution model

Hobolth, Asger; Wiuf, Carsten

Published in:
Theoretical Population Biology

DOI:
10.1016/j.tpb.2023.12.004

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Maximum likelihood estimation and natural pairwise estimating equations are identical for three sequences and a symmetric 2-state substitution model

Asger Hobolth<sup>a,∗</sup>, Carsten Wiuf<sup>b</sup>

<sup>a</sup>Department of Mathematics, Aarhus University, Ny Munkegade 118, Building 1530, DK-8000 Aarhus C, Denmark
<sup>b</sup>Department of Mathematical Sciences, University of Copenhagen, Denmark

**A R T I C L E I N F O**

Keywords:
- Maximum likelihood estimation
- Pairwise comparisons
- Phylogenetic trees

**A B S T R A C T**

Consider the problem of estimating the branch lengths in a symmetric 2-state substitution model with a known topology and a general, clock-like or star-shaped tree with three leaves. We show that the maximum likelihood estimates are analytically tractable and can be obtained from pairwise sequence comparisons. Furthermore, we demonstrate that this property does not generalize to larger state spaces, more complex models or larger trees. Our arguments are based on an enumeration of the free parameters of the model and the dimension of the minimal sufficient data vector. Our interest in this problem arose from discussions with our former colleague Freddy Bugge Christiansen.

1. **Motivation: A tribute to our former colleague Freddy**

We were colleagues with Freddy Bugge Christiansen for many years at the Bioinformatics Research Center (BiRC) at Aarhus University. During the early years of the center, Freddy was the only senior academic at BiRC, which was located in a beautiful old building that used to house Danish army officers. This unique setting made us feel special. Freddy was surrounded by a large group of early career researchers, and he always motivated us to pursue our academic dreams and ambitions. He was eager to engage in scientific discussions and provide academic and personal advice to his younger colleagues. Freddy's influence played a significant role in shaping our careers, as we both developed a strong interest in statistical genetics.

All three of us had previously been students at different times at the Department of Theoretical Statistics at Aarhus University. This shared background fostered a common interest in fundamental statistical problems, and we often approached problems from a statistical perspective. The present paper is a tribute to Freddy and falls within this vein of statistical inquiry.

During alternate fall terms, I (Asger) taught a course titled “Statistical Methods in Bioinformatics” based on the book by Warren Ewens and Gregory Grant (Ewens and Grant, 2005). As part of the course, the students were given a 48-hour take-home exam. Before each exam, I frequently sought Freddy's input to the proposed exercises, and to assist with grading the students' answers. Freddy always took the time to ensure that the exercises were clear and that the students received a fair assessment.

One topic in the course was the EM-algorithm (Dempster et al., 1977) and its application in molecular evolution (Holmes and Rubin, 2002; Hobolth and Jensen, 2005). As a preparation for an exam exercise, I simulated a general tree with three leaves using a symmetric 2-state substitution model. I then generated appropriate initial values for the EM-algorithm based on simple estimating equations derived from pairwise sequence comparisons, and performed the iterative procedure. Surprisingly, the parameter estimates did not change, and I could not find any bugs in my code. I was puzzled by this outcome and sought to understand the underlying reason.

In this paper, we demonstrate that the solutions to the pairwise estimating equations coincide with the maximum likelihood estimates (MLEs) for three sequences and a symmetric 2-state substitution model. Consequently, the parameters in the EM algorithm should not change. Therefore, my implementation was correct, although unnecessary. Furthermore, in this paper, we also address a broader question: when are the maximum likelihood estimates for branch lengths and substitution rates identical to those obtained from pairwise comparisons?

2. **Overview**

This short paper collects some observations on parameter estimation in phylogenetic models with few leaves. Specifically, we show that for a symmetric 2-state model with three leaves and known topology, the maximum likelihood estimates (MLEs) of the branch lengths can be obtained in analytical form from pairwise comparisons of the sequences. Furthermore, we show that in general the MLEs of branch
lengths and substitution rates cannot be expressed solely in terms of pairwise comparisons between the leaves for asymmetric models with three leaves or symmetric models with four or more leaves.

The classical procedure for finding the MLE for a set of sequences related by a known topology is by calculating the likelihood using Felsenstein’s pruning algorithm (see Chapter 16 in Felsenstein, 2004), and then use a numerical optimizer for parameter estimation. Another strategy is to use the EM-algorithm (e.g. Holmes and Rubin, 2002 or Hobolth and Jensen, 2005), but this procedure also requires Felsenstein’s pruning algorithm. Felsenstein’s pruning algorithm efficiently sums all possible states of the inner nodes, but pairwise comparisons are much simpler. We show that pairwise comparisons are only sufficient for maximum likelihood estimation in the very simple case of a symmetric 2-state model with three leaves.

Neyman (1971) is perhaps the first to suggest the maximum likelihood method for estimating phylogenetic parameters. Explicit analytical expressions of the MLEs are naturally of interest as it simplifies analysis and facilitates interpretations. Yang (2000) finds an analytical expression for the unique MLEs by differentiation of the likelihood function allows for a simple analytical solution of the equations for a 2-state symmetric model for three leaves with a molecular clock. We essentially revisit these results, and show that the MLEs are much simpler. We show that pairwise comparisons are only sufficient for maximum likelihood estimation in the very simple case of a symmetric 2-state model with three leaves.

Table 1 provides an overview of our observations. We consider two-, three-, and four-leaves phylogenetic trees with $K$ different alleles and a symmetric or asymmetric substitution model. We assume that the topology is known. The parameters in the phylogenetic tree estimation problem with three leaves can be expressed in analytical form from pairwise comparisons of the sequences if the substitution model is symmetric and has two states. If the model is asymmetric, or the number of alleles $K$ is larger than two, or the number of leaves is larger than three, then the MLE can no longer be expressed in terms of pairwise differences between the leaves.

From Table 1, we see that the number of free parameters of the model agrees with the dimension of the minimal sufficient data vector for two leaves and for three leaves and a 2-state symmetric model. Using properties of the multinomial distribution, it is therefore possible to obtain simple estimating equations for the parameters in these cases by equating empirical frequencies with expected probabilities of data patterns. Further, and interestingly, precisely for (a) two leaves, or (b) the symmetric 2-state substitution model with three leaves, the pairwise comparisons between the sequences form a minimal sufficient statistics. The problem of determining the MLE therefore reduces to a problem of estimating parameters from pairwise comparisons. For three leaves and more complex substitution models or more than three leaves, the number of free parameters is smaller than the dimension of the minimal sufficient data vector, and the MLE is not available in a simple analytical closed form. These statements are shown in detail in the next sections.

3. Two sequences

We start by comparing two sequences in a 2-state model with asymmetric transition rates between the states. The transition probabilities of different endpoints, $0 \rightarrow 1$ and $1 \rightarrow 0$, respectively, over time $t > 0$, are given by

\[ q_{01}(t) = \frac{\alpha}{(1+\alpha)} (1 - e^{-t(1+\alpha)}) \quad \text{and} \quad q_{10}(t) = \frac{1}{(1+\alpha)} (1 - e^{-t(1+\alpha)}) \]

Here, the transition rate from $0 \rightarrow 1$ is $\alpha > 0$ and from $1 \rightarrow 0$ is 1. The stationary distribution of the continuous time Markov chain is $x = (x_0, x_1) = (1, \alpha)/(1+\alpha)$. There are four different site patterns $(0,0), (0,1), (1,0), (1,1)$. The probabilities of the patterns are

\[ p_{01}(t) = \frac{1}{1+\alpha} - p_{00}(t) = x_0 q_{01}(t) = \frac{\alpha}{(1+\alpha)^2} (1 - e^{-t(1+\alpha)}) \]

and

\[ p_{10}(t) = \frac{1}{1+\alpha} - p_{11}(t) = x_1 q_{10}(t) = \frac{\alpha}{(1+\alpha)^2} (1 - e^{-t(1+\alpha)}) \]

In a pairwise alignment with counts $(n_{00}, n_{01}, n_{10}, n_{11})$, the likelihood is

\[ L(t) \propto p_{00}(t)^{n_{00}} p_{01}(t)^{n_{01}} p_{10}(t)^{n_{10}} p_{11}(t)^{n_{11}} \]

(5)

where the multinomial coefficient has been omitted. We note that the pairwise comparison $(n_{00}, n_{01}, n_{10}, n_{11})$ is a minimal sufficient statistic.

The likelihood function allows for a simple analytical solution of the MLEs of rate $\alpha$ and branch length $\beta$, namely

\[ \hat{\alpha} = \frac{2p_{11} + \hat{p}_{01} + \hat{p}_{10}}{2p_{00} + \hat{p}_{01} + \hat{p}_{10}} \quad \text{and} \quad \hat{\beta} = \frac{1}{1+\hat{\alpha}} \log \left( \frac{1 - (1+\hat{\alpha})^2(\hat{p}_{00} + \hat{p}_{11})}{2\hat{\alpha}} \right), \]

provided that the term in the logarithm is positive (if this is not the case then either $\hat{\beta} = 0$ or $\infty$). Here $\hat{p}_{00}, \hat{p}_{01}, \hat{p}_{10}, \hat{p}_{11}$ are the empirical frequencies of the corresponding probabilities, for example, $\hat{p}_{00} = n_{00}/n$, where $n$ is the total number of observed sites. The expressions can be obtained by noting that $2p_{00}(t) + p_{01}(t) + p_{10}(t) = 2/(1+\alpha)$. This result, though simple, does not appear to be widely known.

If $\alpha = 1$ (that is, the model is a symmetric 2-state model), then the MLE of $\beta$ is given as in (5) by replacing $\hat{\alpha}$ with 1,

\[ \hat{\beta} = -\frac{1}{2} \log \left( 1 - 2(\hat{p}_{00} + \hat{p}_{11}) \right). \]

References

with similar equations for \( \hat{t}_y \). In the model implies that the frequencies are smaller than 0.5 (otherwise, one or more branch lengths are merged, the pairs with \((a, b, c) = (0, 0, 0)\), \((1, 0, 0)\), and \((1, 1, 0)\) are merged, and so on, see Fig. 1. Note that the probabilities only depend on the branch lengths in terms of pairwise estimating equations because

\[
\hat{p}_{000} = \frac{1}{8} \left( 1 + e^{-2\hat{t}_y} \right) \left( 1 + e^{-2\hat{t}_x} \right) \left( 1 + e^{-2\hat{t}_z} \right),
\]

and similar calculations give

\[
\hat{p}_{010} = \frac{1}{4} \left( 1 - e^{-2\hat{t}_x} + e^{-2\hat{t}_y} + e^{-2\hat{t}_z} - e^{-2\hat{t}_u} \right).
\]

We conclude that an analytical MLE based on pairwise comparisons are not sufficient. Hence, our result from the previous section does not generalize beyond \( n = 3 \) alleles, a symmetric substitution model, and a star-shaped, clock-like or general tree. These findings are summarized in column 4–6 in Table 1.

### 5. Beyond three sequences and a symmetric 2-state model

For three sequences and an asymmetric model, three sequences and more than two states, or more than four sequences, the counts of the pairwise comparisons are not sufficient. Hence, our result from the previous section does not generalize beyond \( n = 3 \) and a symmetric
Consider first the 3-state symmetric general three leaves tree (the column \( n = 3, K = 3 \), symmetric model, general tree in Table 1). The data consists of the counts of five configurations, namely \( (n_{000}, n_{011}, n_{010}, n_{101}, n_{001}) \), corresponding to alignment columns where all the states are the same, two are the same or they are all different. The pairwise differences on the \( ab \) branch is \( n_{100} + n_{011} + n_{102} \), on the \( ac \) branch \( n_{010} + n_{011} + n_{012} \), and on the \( bc \) branch \( n_{001} + n_{011} + n_{012} \). If we add or subtract the vector \((-1, 1, -1, 1, -1, 1)\) to the original count data, then the likelihood changes, but the pairwise differences stay the same. This shows that the pairwise differences are not a sufficient statistics for the 3-state symmetric general three leaves tree.

Secondly, consider the asymmetric general three leaves tree determined by the 8 data points from (8), i.e. the column with \( n = 3, K = 2 \), asymmetric model, general tree in Table 1. The pairwise comparisons of sequence \( a \) and \( b \) and \( c \), and \( b \) and \( c \), are given by the minimal statistics

\[
(n_{000} + n_{010} + n_{100} + n_{101} + n_{110} + n_{111}).
\]

\[
(n_{000} + n_{010} + n_{100} + n_{101} + n_{110} + n_{111}).
\]

\[
(n_{010} + n_{011} + n_{010} + n_{110} + n_{110} + n_{111}).
\]

These three pairwise statistics remain the same if we add or subtract the vector \((-1, 1, -1, 1, -1, 1)\) from the original data points (8). However, the likelihood changes with the addition or subtraction of the vector.

Finally, consider a general tree with four leaves and a 2-state asymmetric model, general tree in Table 1. The pairwise comparisons

\[
(n_{000} + n_{010} + n_{100} + n_{101} + n_{010} + n_{101} + n_{110} + n_{111}).
\]

As for three leaves we can define pairwise differences. There are six pairwise differences. An argument that MLE and estimation based on these statistics

\[
(R_{ab}, R_{ac}, R_{bc}) = (n_{ab}, n_{ac}, n_{bc}) = (m_{010} + m_{101}, m_{001} + m_{011}, m_{010} + m_{001} + m_{011} + m_{100})
\]

is proportional to

\[
(p_{001} p_{101} - p_{010}) D_{ab} (p_{100} p_{010} - p_{011}) D_{ac} (p_{011} p_{100} - p_{101}) D_{bc},
\]

which follows by making the appropriate substitutions in the likelihood of the data. Hence in particular the empirical pairwise differences form a sufficient statistic. It is intriguing that this likelihood is not a multinomial distribution despite the fact that the marginal distributions of \( R_{ab}, R_{ac}, R_{bc} \) are binomials. Neither does it seem transparent how one would get the simple form of the maximum likelihood estimates from the likelihood above. In Yang (2000), the MLE solution is found by differentiation of the log-likelihood function, which hides the simple connection to the pairwise estimation equations.

In Chor et al. (2003), a closed-form expression is found for the MLE for \( n = 4 \) in a 2-state symmetric model with a molecular clock, assuming the four leaves are related by a fork (two subtrees with two leaves each). The expression cannot be cast in terms of pairwise comparisons.

We have described situations where the MLEs agree with the solution obtained by pairwise sequence comparisons. In these cases, pairwise sequence comparison is the same as calculating a distance between sequences using Eqs. (5)–(7). There are many distance based tree reconstruction methods (see e.g. Chapter 11 and Chapter 12 in Felsenstein, 2004), and generally one should not expect a distance based method to provide the same solution as the MLE solution, even in the cases considered here. For example, in the case of neighbor-joining (Saitou and Nei, 1987), the update rules conserve the total amount of evolution (the sum of all distances) in each step. This feature is in contrast to MLE where the sum of all distances calculated from the estimated tree might differ from the distances used as input to neighbor-joining. Only in the case of two and three sequences will these approaches trivially agree.

CRediT authorship contribution statement

Asger Hobolth: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. Carsten Wiuf: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgments

We are grateful to the editor Marcy Uyenoyama and the reviewer for comments and suggestions that substantially improved the presentation.

References