The Sampling Distributions and Covariance Matrix of Phylogenetic Spectra

Peter J. Waddell,* David Penny,* Michael D. Hendy,† and Greg Arnold‡

*Department of Plant Biology, †Department of Mathematics, and ‡Department of Statistics, Massey University

We extend recent advances in computing variance-covariance matrices from genetic distances to a sequence method of phylogenetic analysis. These matrices, together with other statistical properties of corrected sequence spectra, are studied as a foundation for more powerful and testable methods in phylogenetics. We start with $\mathbf{s}$, a vector of the proportion of sites in a sequence of length $c$ showing each of the possible character-state patterns for $t$ taxa. Hadamard conjugations are then used to calculate $\mathbf{\hat{y}}$, a vector of the the support for bipartitions, or splits, in the data, after correcting for all implied multiple changes. These corrections are made independently of any tree and are illustrated with Cavender’s two-character-state model. Each entry in $\mathbf{\hat{y}} (\gamma_0$ excluded) that is not associated with an edge on the tree that generated the data is an invariant (sensu Cavender) with an expected value of 0 as the number of sites $c \rightarrow \infty$. Under an independent identically distributed model (sites are independent and identically distributed), vector $\mathbf{\hat{s}}$ is a random sample from a scaled multinomial distribution. Starting from this point, we illustrate the derivation of $\mathbf{V}[\mathbf{\hat{y}}]$, the variance-covariance matrix of $\mathbf{\hat{y}}$. The bias induced by the delta method, a convenient approximation in deriving $\mathbf{V}[\mathbf{\hat{y}}]$, is evaluated for both population and sample variance-covariance matrices. It is found to be acceptable in the first case and very good in the second. Likewise bias in $\mathbf{\hat{y}}$ due to a logarithmic transform and to short sequences is also acceptable. We infer the marginal distributions of entries in $\mathbf{\hat{y}}$. Simulations with illustrative values of $c$ and $\lambda$ (the rate per site) show how $\mathbf{\hat{y}}$ tends to multivariate normal as $c \rightarrow \infty$. Our results extend naturally to four-color (nucleotide) spectra.

Introduction

A major priority in phylogenetic research is to assess the statistical reliability of trees built from sequence data (Felsenstein 1988; Penny et al. 1992). The variances and covariances of edge lengths in trees estimated from distance data are important in quantifying the sampling errors arising from finite sequence length (see, e.g., Nei and Jin 1989; Bulmer 1991). Here we describe equivalent results for sequence data. For any tree-building method to be consistent (i.e., to converge to the true weighted tree with sufficiently long sequences; Felsenstein 1978), it must compensate for multiple changes.

One new approach to estimating phylogenetic trees is spectral analysis (Hendy and Penny 1993), in which aligned sequences are corrected for all implied multiple changes (hits), prior to choosing an optimal tree(s). With this method the proportions of sites which show each pattern of character states are expressed as a vector $\mathbf{s}$ ($s$ from sequences and $\mathbf{\hat{s}}$ to denote a sample estimate). The Hadamard conjugation (a discrete Fourier transform) applied to $\mathbf{s}$ takes into account all implied multiple changes under the Cavender-Farris model (Farris 1973; Cavender 1978) to give a vector of reweighted sequence pattern values called $\mathbf{\hat{y}}$. Each entry in $\mathbf{\hat{y}} (\gamma_0$ excluded) which is not associated with an edge on the tree which generated the data is an invariant (sensu Cavender 1978) with an expected value of 0 as the number of sites $c \rightarrow \infty$ (Hendy and Penny 1993). From the adjusted weighted bipartitions in $\mathbf{\hat{y}}$, it is possible to choose a set of compatible bipartitions, i.e., a tree, by criteria such as maximum parsimony, compatibility, or closest tree (Hendy 1991). A useful feature of Hadamard conjugations is that they are invertible; that is, given a weighted tree we can use it to calculate the probability of each bipartition pattern (the vector $\mathbf{s}(T)$) under Cavender's model. With real data this can be used to check the goodness of fit between data and model, though here we use the Hadamard conjugation to specify the data from which we will study the effects of sampling error on $\mathbf{\hat{y}}$.

If sites in the sequences are independent, then the patterns observed in a sequence of length $c$ are a random sample from a multinomial distribution (Felsenstein 1981, 1982) with parameters $c$ and $\mathbf{s}(T)$ (the vector of the probabilities of observing each pattern). Knowing
this, we can estimate \( V[\hat{s}] \), the variance-covariance matrix of \( \hat{s} \). The sample error in \( \hat{s} \) is transformed by the Hadamard conjugation into error in \( \hat{y} \). In this paper we show how to calculate \( V[\hat{y}] \), the variance-covariance matrix for the entries in \( \hat{y} \), and from this we derive the size of the correlations of pairs of entries in \( \hat{y} \).

The variances and covariances of \( \hat{y} \) give an approximate measure of the sampling error of an entry in \( \hat{y} \). To precisely know the probability of an error, we need to know the form of the marginal distributions of \( \hat{y} \) (the values \( \hat{y} \) takes when \( \hat{s} \) is a random sample from \( s(T) \)). (For example, we need to know whether the marginal distributions are Poisson, normal, etc.) The bivariate distributions of pairs of entries in \( \hat{y} \) will show whether pairs of entries are linearly correlated. It is already known that \( \hat{y} \) is a consistent estimator (Hendy and Penny 1993).

That is, \( \hat{y} \) estimated from longer and longer sequences, will converge to \( \gamma(T) \), a vector representation of the original weighted tree, \( T \). Since the Hadamard conjugation, like distance corrections, is a nonlinear transformation, we expect a degree of statistical bias (a systematic departure of the mean of \( \hat{y} \) from \( \gamma(T) \) for short sequences) when using it (Stuart and Ord 1987, p. 324). Using simulations, we estimate the bias in \( \hat{y} \) when it is calculated with sequence lengths commonly used in phylogenetic studies.

Two factors affecting the form of the multivariate distribution of \( \hat{y} \) are \( c \) (the sequence length) and \( \lambda \) (the rate of change per site). Parameter \( c \) will influence the magnitude of the error in \( \hat{s} \) (relative to entries in \( s(T) \)), while \( \lambda \) will determine how many multiple changes are expected. We infer the effects of \( c \) and \( \lambda \) on the distribution of \( \hat{y} \) by following the intermediate steps in the Hadamard conjugation from \( s \) to \( \hat{y} \). The changes in the multivariate distribution of \( \hat{y} \) are illustrated with simulations at selected values for \( c \) and \( \lambda \).

Last, the delta method (Stuart and Ord 1987, p. 324; Bulmer 1991; Rzhetsky and Nei 1992), a convenient first-order approximation used to estimate the variances and covariances of variables after a nonlinear transformation, is expected to cause bias in \( V[\hat{y}] \) (Stuart and Ord 1987, p. 324). We examine the magnitude and pattern of this bias when calculating population and sample variance-covariance matrices of \( \hat{y} \).

**Methods**

**Our Illustrative Model**

Following Hendy and Penny (1993), let us describe \( T \), our weighted tree of \( t \) taxa, by the vector of weighted edge lengths \( \gamma(T) \). The tree \( T \) that we use to illustrate our methods (fig. 1a) is similar to Felsenstein’s (1978) example with only two different edge weights, which are chosen so that parsimony applied to uncorrected sequence data will converge to the wrong tree. An edge’s length (weight) is the total expected number of changes (counting multiplicity of changes) per nucleotide site on that edge. These are represented by \( \gamma(T) \) of figure 1b. The mechanism of character-state change used here is that of Cavender (1978) and Hendy and Penny (1989, 1993), the two-state analogue of the Jukes-Cantor equation. Changes at sites occur independently of each other, and the probability of changing is identical at all sites, so the mechanism of change is independent and identically distributed (i.i.d.).

Following Hendy and Penny (1993), we apply the Hadamard conjugation to \( \gamma(T) \) to obtain \( s(T) \), the vector of the probabilities for each of the possible \( 2^{t-1} \) site patterns (bipartitions); that is,

\[
s(T) = H^{-1}(\exp(H\gamma(T))) .
\]

In this formula \( H \) is a symmetric Hadamard matrix of \( 2^{t-1} \) rows, \( H^{-1} = 2^{(t-1)}H \), and the exponent function is applied component-wise. The Hadamard matrix used is that specified in Hendy and Penny (1993), which for four taxa is

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\
\end{bmatrix}
\]
Table 1
s(T) and V[s]

<table>
<thead>
<tr>
<th>INDEX</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipartition</td>
<td>{0/1234}</td>
<td>{1/234}</td>
<td>{2/134}</td>
<td>{12/34}</td>
<td>{13/24}</td>
<td>{23/14}</td>
<td>{123/4}</td>
<td></td>
</tr>
<tr>
<td>s(T)</td>
<td>0.6497</td>
<td>0.1283</td>
<td>0.0200</td>
<td>0.0226</td>
<td>0.1283</td>
<td>0.0258</td>
<td>0.0070</td>
<td>0.0200</td>
</tr>
<tr>
<td>V[s]a</td>
<td>Index:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.2281</td>
<td>-0.0831</td>
<td>-0.0130</td>
<td>-0.0146</td>
<td>-0.0831</td>
<td>-0.0167</td>
<td>-0.0046</td>
<td>-0.0130</td>
</tr>
<tr>
<td>1</td>
<td>-0.0831</td>
<td>0.1119</td>
<td>-0.0026</td>
<td>-0.0029</td>
<td>-0.0165</td>
<td>-0.0033</td>
<td>-0.0009</td>
<td>-0.0026</td>
</tr>
<tr>
<td>2</td>
<td>-0.0130</td>
<td>-0.0026</td>
<td>0.0196</td>
<td>-0.0005</td>
<td>-0.0026</td>
<td>-0.0005</td>
<td>-0.0001</td>
<td>-0.0004</td>
</tr>
<tr>
<td>3</td>
<td>-0.0146</td>
<td>-0.0029</td>
<td>-0.0005</td>
<td>0.0221</td>
<td>-0.0029</td>
<td>-0.0006</td>
<td>-0.0002</td>
<td>-0.0005</td>
</tr>
<tr>
<td>4</td>
<td>-0.0831</td>
<td>-0.0165</td>
<td>-0.0026</td>
<td>-0.0029</td>
<td>0.1119</td>
<td>-0.0033</td>
<td>-0.0009</td>
<td>-0.0026</td>
</tr>
<tr>
<td>5</td>
<td>-0.0167</td>
<td>-0.0033</td>
<td>-0.0005</td>
<td>-0.0006</td>
<td>-0.0033</td>
<td>0.0251</td>
<td>-0.0002</td>
<td>-0.0005</td>
</tr>
<tr>
<td>6</td>
<td>-0.0046</td>
<td>-0.0009</td>
<td>-0.0001</td>
<td>-0.0002</td>
<td>-0.0009</td>
<td>0.0070</td>
<td>-0.0001</td>
<td>-0.0001</td>
</tr>
<tr>
<td>7</td>
<td>-0.0130</td>
<td>-0.0026</td>
<td>-0.0004</td>
<td>-0.0005</td>
<td>-0.0005</td>
<td>-0.0001</td>
<td>0.0196</td>
<td></td>
</tr>
</tbody>
</table>

a Indexing is that of Hendy and Penny (1993).
b Generated with the vector γ(T) in fig. 1b, by eq. (1).

For convenience, the values in this and the subsequent covariance matrices (tables 2, 3, and 4) are calculated independent of a sequence length, so, to calculate their entries for a sequence of length c, simply divide each entry by c.

The calculations of equation (1) are fully invertible. Below (eq. [4]) we give the inverse of equation (1), starting with s(T) to obtain γ(T). We then break the calculations up into a series of intermediate steps and give the results of a numerical example, starting with s(T), at the top of tables 1–4.

The vector s(T) calculated from γ(T) in fig. 1b, by equation (1), is shown in table 1 and in graphical form as the gray columns of fig. 2. The frequencies of the bipartitions in a sequence of length c are, under this model, a random observation from a multinomial distribution (Felsenstein 1981) with parameters c and s(T). Therefore $\hat{s}$ (a vector $f/c$, where $f_i$ is the observed frequency of the $i$th bipartition), our maximum-likelihood estimate of s(T), has the distribution of the above multinomial scaled by 1/c. The vector s(T) is our population mean vector, and $\hat{s}$ is our sample mean vector.

Calculating the Variance-Covariance Matrix of Phylogenetic Spectra

The Variance-Covariance Matrix $V[\hat{s}]$ of the Sequence Spectrum $\hat{s}$

The entries in the multinomial population variance-covariance matrix (referred to henceforth as the “covariance matrix”) $V[s]$ of $\hat{s}$, estimated from c sequence sites, are

Table 2
r(T) and V[r]

<table>
<thead>
<tr>
<th>INDEX</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path set</td>
<td>{0}</td>
<td>{1,4}</td>
<td>{2,4}</td>
<td>{1,2}</td>
<td>{3,4}</td>
<td>{1,3}</td>
<td>{2,3}</td>
<td>{1,2,3,4}</td>
</tr>
<tr>
<td>r(T)</td>
<td>1.000</td>
<td>0.6065</td>
<td>0.8607</td>
<td>0.6376</td>
<td>0.6376</td>
<td>0.4274</td>
<td>0.6065</td>
<td>0.4066</td>
</tr>
<tr>
<td>V[r]a</td>
<td>Index:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.6321</td>
<td>0.1156</td>
<td>0.4740</td>
<td>0.0407</td>
<td>0.3784</td>
<td>0.0387</td>
<td>0.3599</td>
</tr>
<tr>
<td>2</td>
<td>0.0000</td>
<td>0.1156</td>
<td>0.2592</td>
<td>0.0577</td>
<td>0.0577</td>
<td>0.0387</td>
<td>0.1156</td>
<td>0.0775</td>
</tr>
<tr>
<td>3</td>
<td>0.0000</td>
<td>0.4740</td>
<td>0.0577</td>
<td>0.5934</td>
<td>0.0000</td>
<td>0.3340</td>
<td>0.0407</td>
<td>0.3784</td>
</tr>
<tr>
<td>4</td>
<td>0.0000</td>
<td>0.0407</td>
<td>0.0577</td>
<td>0.0000</td>
<td>0.5934</td>
<td>0.3340</td>
<td>0.4740</td>
<td>0.3784</td>
</tr>
<tr>
<td>5</td>
<td>0.0000</td>
<td>0.3784</td>
<td>0.0387</td>
<td>0.3340</td>
<td>0.3340</td>
<td>0.8173</td>
<td>0.3784</td>
<td>0.6869</td>
</tr>
<tr>
<td>6</td>
<td>0.0000</td>
<td>0.0387</td>
<td>0.1156</td>
<td>0.0407</td>
<td>0.4740</td>
<td>0.3784</td>
<td>0.6321</td>
<td>0.3599</td>
</tr>
<tr>
<td>7</td>
<td>0.0000</td>
<td>0.3599</td>
<td>0.0775</td>
<td>0.3784</td>
<td>0.3784</td>
<td>0.6869</td>
<td>0.3599</td>
<td>0.8347</td>
</tr>
</tbody>
</table>

a The first row and column of this matrix will always be 0, since $r_i$ is the sum of all entries in $\hat{s}$. For convenience of presentation, all entries in this matrix have yet to be divided by a specific sequence length c.
Table 3

\( V'(b) \) Estimated Using the Delta Method

<table>
<thead>
<tr>
<th>INDEX</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path set</td>
<td>{0}</td>
<td>{1,4}</td>
<td>{2,4}</td>
<td>{1,2}</td>
<td>{3,4}</td>
<td>{1,3}</td>
<td>{2,3}</td>
<td>{1,2; 3,4}</td>
</tr>
<tr>
<td>( \rho(T) )</td>
<td>0.00</td>
<td>-0.50</td>
<td>-0.15</td>
<td>-0.45</td>
<td>-0.45</td>
<td>-0.85</td>
<td>-0.50</td>
<td>-0.90</td>
</tr>
<tr>
<td>( V'[b] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The variances of the longest sets of paths (underlined) \( \rho_3 \) (negative twice the estimated distance between taxa 1 and 3) and \( \rho_6 \) (negative twice the estimated length of the nonintersecting path set between taxa 1, 2, 3, and 4 [i.e., edges 1, 2, 4, and 7]) have increased considerably, because of the uncertainty created by multiple changes at a site. Since these two entries estimate the weights on very similar sets of tree edges (because they are counting mostly identical sequence bipartitions), they have a large covariance and a correlation of approximately 0.85.

\[ V_i = s(T)(1-s(T))/c, \tag{2} \]

and

\[ V_{ij} = -s(T)_i s(T)_j/c, \tag{3} \]

where \( V_i \) is the variance of \( \hat{s}_i \) and \( V_{ij} (i \neq j) \) is the covariance of \( \hat{s}_i \) with \( \hat{s}_j \). \( V[\hat{s}] \) is symmetric with \( (2^{r-1})^2 = 4^{r-1} \)

entries. Table 1 shows the covariance matrix of our example. For ease of presentation, the values that appear in all covariance matrices in this paper are not divided by sequence length. With sampled data, we replace \( s(T) \) with its maximum-likelihood estimate \( \hat{s} \), obtaining the sample covariance matrix \( \hat{V}[\hat{s}] \).

The marginal distributions of entries in \( \hat{s} \) are binomial distributions (with parameters \( c \) and \( s_i \)) scaled

Table 4

\( V'[\gamma] \) Estimated Using the Delta Method

<table>
<thead>
<tr>
<th>INDEX</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipartition</td>
<td>{0/1234}</td>
<td>{1/234}</td>
<td>{2/134}</td>
<td>{12/34}</td>
<td>{3/124}</td>
<td>{13/24}</td>
<td>{23/14}</td>
<td>{123/4}</td>
</tr>
<tr>
<td>( \gamma(T) )</td>
<td>-0.475</td>
<td>0.200</td>
<td>0.025</td>
<td>0.025</td>
<td>0.200</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
</tr>
<tr>
<td>( V'[\gamma] )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* For ease of presentation, values in this matrix have yet to be divided by \( c \). The largest covariance (excluding \( \gamma_6 \), which cannot be an edge in any tree) is the negative covariance (underlined) between \( \gamma_1 \) or \( \gamma_4 \), which estimate the long pendant edges 1 and 4 of the model tree (fig. 1), and \( \gamma_5 \), which groups the long edges together. Bipartition \( \gamma_1 \) causes parsimony applied to \( s(T) \) to converge to the wrong tree, while \( \gamma_5 \) (its corrected estimated value) has a relatively large variance. This large variance makes \( \gamma_5 \) more likely than \( \gamma_6 \) to be greater than \( \gamma_1 \) (which estimates the internal edge in our model tree) due to sampling error. When applied to this \( \gamma_1 \), many optimality criteria, such as parsimony, compatibility, or closest tree, will simply choose the largest of the three previous entries as the internal edge in the optimal tree.
by 1/c. When \( cs_j \) and \( c(1-s_j) \) are both >5, then the normal approximation to the binomial is reasonable (Freund and Walpole 1987, p. 229). A better approximation to these binomials, when \( c \) is large and \( cs_j < 5 \), is the Poisson distribution. When all \( cs_j > 5 \), then \( \hat{s} \) is distributed approximately multivariate normally. As \( c \to \infty \), \( \hat{s} \) tends to a multivariate normal distribution, with variance tending to zero.

**Intermediate Steps in the Hadamard Conjugation from \( \hat{s} \) to \( \hat{\gamma} \)**

Given our specified mechanism of character change, we can derive a consistent estimator \( \hat{\gamma} \) of \( \gamma(T) \) by inverting equation (1) and replacing \( s(T) \) with \( \hat{s} \). Thus, using the inverse of the Hadamard conjugation, we obtain

\[
\hat{\gamma} = H^{-1}(\ln(H\hat{s})),
\]

where \( \ln \) (the natural logarithm) is applied to each component of \( H\hat{s} \). Following Hendy and Penny (1993), we break the conjugation into three steps

\[
\hat{\gamma} = H^{-1}(\ln(H\hat{s})) = \hat{\rho} = \ln \hat{\gamma} = H^{-1}\hat{\rho} = H^{-1}(\ln(H\hat{s})),
\]

where \( \rho \) is the covariance matrix of \( \hat{s} \) (Krzeanowski p. 205) and the transpose \( H' = H \), since \( H \) is symmetric (See the Appendix for a more efficient implementation of these operations.) Table 2 shows the results of these operations on \( V[\hat{\gamma}] \).

The entries in \( \hat{\gamma} \) (like the entries in \( \hat{s} \)) have marginal distributions which are scaled binomial though their covariance structure is not that of a binomial. We now describe the marginal distribution of entries in \( \hat{\gamma} \). For any subset \( J \subseteq \{0, \ldots, n\} \) of indices, let \( \hat{s}_J = \sum_{j \in J} \hat{s}_j \) and \( \hat{s}_J = \sum_{j \in J} s_j \) (where \( s_j \) is the observed number of changes). Recall that \( \hat{\gamma} \) is distributed as a binomial with parameters \( c \) and \( \rho = \hat{\gamma} \) is distributed as a scaled binomial; that is, \( \hat{s}_J \sim (B(c, \rho)) \) scal
c). Let \( I = \{ j : h_{ij} = -1 \} \) (\( h_{ij} \) from the Hadamard matrix), then \( \bar{r}_i = 1 - 2\bar{s}_i \) is distributed as \( 1 - 2(B(c, s_j) \) scaled by \( 1/c \). The variance of \( \bar{r}_i \) is therefore \( 4 \cdot s_j(1 - s_j)/c = (1 - \bar{r}_i^2)/c \). In biological data sets, entries in \( \bar{r} \) will often be estimated from an expected five or more observable sequence changes; that is, \( E[\bar{r}_i] > 5 \), where \( E \) denotes expected value. Correspondingly, the marginal distributions of \( \bar{r}_i \) will often be well approximated by a normal distribution with mean \( r_i \) and variance as given above.

Since \( \bar{r} \) is a linear transformation of \( \hat{s} \), both \( \hat{s} \) and \( \bar{r} \) tend to multivariate normal as \( c \to \infty \), with variance tending to zero. This is a generalization of the pairwise-distance matrix (a subset of \( \bar{r} \)) tending to multivariate normal (see, e.g., Bulmer 1991).

**The Covariance Matrix** \( \mathbf{V}[\hat{\rho}] \) **of the Estimated Path Lengths** \( \hat{\rho} \)

The total implied number of changes per site on the \( i \)th even-sized set of edge paths is \( -\rho_i/2 \), where \( \rho_i = \ln \bar{r}_i \). This value is also the component-wise maximum-likelihood estimate of \( \rho_i \). These estimates are written in this form because multiplication of \( \hat{\rho} \) by \( \mathbf{H}^{-1} \) will give the implied total number of changes on each bipartition in \( \hat{\gamma} \), the corrected sequence vector.

The entries in \( \mathbf{V}[\bar{r}] \) must be adjusted for this nonlinear transformation, to give \( \mathbf{V}[\hat{\rho}] \). We may estimate the variance of \( \hat{\rho}_i \) exactly as \( \text{Var}[\hat{\rho}_i] = \text{Var}[\ln(\bar{r}_i/c)] \), which involves summing many binomial terms. This computation is time consuming, so we estimate the variance of \( \rho_i \) by using the first-order approximation known as the delta method (Stuart and Ord 1987, p. 324). The delta method uses the result that any transformation changes the variance by a factor approximately equal to the square of the gradient of the transformation curve. We denote covariance and correlation matrices estimated using the delta method with a prime, e.g., \( \mathbf{V}^\prime[\bar{r}] \).

The gradient of the transformation \( \rho_i = \ln(r_i) \), with respect to \( r_i \), is \( 1/r_i \). So letting \( V_{ij} \) be the \( ij \)th entry in \( \mathbf{V}[\bar{r}] \), we have

\[
V[\hat{\rho}]_{ii} = (1/r_i)^2 V_{ii}; \quad (9)
\]

a similar result for covariances gives

\[
V[\hat{\rho}]_{ij} = 1/r_i \cdot 1/r_j \cdot V_{ij}, \quad (10)
\]

where \( r_i \) and \( r_j \) are, respectively, the \( i \)th and \( j \)th entries in the vector \( r \) and \( i \neq j \). Table 3 shows the results of these calculations. Later we investigate the size and direction of bias resulting from the delta method.

All values in \( \hat{\rho} \) will have greater variances and covariances than the corresponding entries in \( \bar{r} \), because the gradient of the log curve with respect to the \( r \) axis is 1 at \( r_i = 1 \), increasing without limit as \( r_i \to 0 \). As \( c \to \infty \), \( \bar{r} \) tends to multivariate normality with variance \( \to 0 \), so \( \hat{\rho} \) also tends to multivariate normality with variance \( \to 0 \). The logarithmic function introduces positive skewness into the marginal distributions of \( \hat{\rho} \) relative to \( \bar{r} \). This skewness (which causes bias, since \( E[\ln(\bar{r}_i)] \) is always \( > \ln(E[\bar{r}_i]) \) tends to zero as \( c \to \infty \) (and variance \( c \to 0 \)). Skewness also tends to zero as \( \lambda \to 0 \) (i.e., the paths require almost no correction).

If path sets \( i \) and \( j \) have no common edges in the tree, then their correlation will be 0. Otherwise the correlation of two additive path sets of edges will be the sum of the lengths of their common edges divided by the sum of the lengths of all the edges in the two path sets (additive implies counting all changes that occur). So as \( \hat{\rho}_i \) and \( \hat{\rho}_j \) count events on increasingly equivalent path sets, their correlation tends to one. This interpretation is a generalization of the covariance structure of pairwise distances on trees (Nei and Jin 1989; Bulmer 1991).

**The Covariance and Correlation Matrix of** \( \gamma \): **The Corrected Data**

The conjugate spectrum \( \gamma \) of transformed bipartitions is calculated from \( \hat{\rho} \) by equation (5). Its covariance matrix \( \mathbf{V}[\gamma] \) is equal (Krzanowski 1988, p. 205) to

\[
\mathbf{V}[\gamma] = \mathbf{H}^{-1} \mathbf{V}[\hat{\rho}] \mathbf{H}^{-1}. \quad (11)
\]

These operations are computationally almost identical to those involved in going from \( \mathbf{V}[\hat{s}] \) to \( \mathbf{V}[\bar{r}] \), since \( \mathbf{H}^{-1} \) is \( 2^{-c-1} \mathbf{H} \). We still have the same \( 2^{c-1} - 1 \) degrees of freedom in \( \gamma \) that we started with in \( \hat{s} \). Table 5 shows \( \mathbf{V}[\gamma] \), while figure 3 shows the relative sizes of errors on entries in \( \hat{s} \), as implied by \( \mathbf{V}[\gamma] \) with \( c = 1,000 \). Hendy and Penny (1989) observed that long edges cause parsimony on uncorrected data to be inconsistent (which they paraphrased as "long edges attract"). Here we observe that with a consistent method of tree building, long edges cause increased variance of corrected bipartitions which are not in the tree that generated the data. In this case (and we expect in general with appropriate path corrections) there is little increase in the variance of the entry in \( \gamma \), relating to the correct internal edge(s) but a large increase in the variance of entries in \( \gamma \) grouping these long edges together. This will greatly (as a multiple of the probability without long edges) increase the probability of such entries being in the tree selected by any optimality criteria, i.e., choose the wrong tree. In general, resolving the branching order around long edges will be more error prone and will require more data for the same degree of precision than if all edges were the same length.

Let \( \mathbf{C}[\gamma] \) be the correlation matrix of \( \gamma \) so that

\[
\mathbf{C}[\gamma] = \mathbf{W} \mathbf{V}[\gamma] \mathbf{W}, \quad (12)
\]
where $W$ is a diagonal matrix with entries $W_{ii} = 1/\sqrt{V_{ii}}$ (the inverses of the standard deviations of $\hat{\gamma}$).

The correlations of entries in $\hat{\gamma}$ (excluding those with $\hat{\gamma}_0$), for the tree in figure 1a and with $c = 1,000$, are apparent as the ellipsoid bivariate distributions shown in figure 6b. We consider now the causes of the correlation structure of the weighted transformed bipartitions ($\hat{\gamma}_i$, $i \neq 0$). All entries in $\hat{\gamma}$ are correlated. Excluding $\hat{\gamma}_0$, these correlations range from $-0.48$ to $+0.23$. This is due to the nonindependence of cells in the multinomially distributed $s$ and the transformation for multiple changes that alter the relative sizes of these correlations. For example, the cells in $s$, excluding the larger cells $s_0$, $s_1$, and $s_4$, have correlations between 0 and $-0.07$, but the corresponding small entries in $\hat{\gamma}$ may have large correlations; for example, $\gamma_2$ and $\gamma_6$ have a correlation of $0.46$.

The differences between corresponding entries in $C[y]$ and $C[S]$ (the multinomial correlation matrix of $s$) are the result of multiple changes at sites in sequences evolving on the weighted tree in figure 1. For example, in table 5 the largest correlation is the negative correlation ($-0.48$) of $\hat{\gamma}_5$ with both $\hat{\gamma}_1$ and $\hat{\gamma}_4$. This correlation is predominantly due to parallel changes along the edges $\gamma(T)_1$ and $\gamma(T)_4$. In the majority of such parallel changes, patterns $s_1$ or $s_4$ becomes pattern $s_5$, which adds support to $\hat{\gamma}_5$ at the expense of $\hat{\gamma}_1$ and $\hat{\gamma}_4$. The only other double change that would add support to $s_5$ would be parallel changes on the short edges $e_2$ and $e_7$. This is much less likely, as are sets of three or more changes giving the pattern $s_5$. As a result, the small negative correlations between $\gamma_5$ and either $\gamma_2$ or $\gamma_6$ are little changed from the corresponding entries in $C[S]$. The impact of multiple substitutions in increasingly differentiating the correlation structure of $\hat{\gamma}$ from the multinomial correlations of $s$ is illustrated in figure 4. In this plot, the only difference between axes is that the $\lambda$ on the $X$-axis is only one-fourth of that on the $Y$-axis (that is, the correlations of the white with the gray columns in fig. 2). As $\lambda \to 0$, the correlations of $\gamma$ ($\gamma_i$, $i \neq 0$) tend to those of $s$, which also tend to zero.

The Marginal Distributions of Entries in $\hat{\gamma}$

In order to know the probability that, due to sampling error alone, an entry in $\hat{\gamma}$ exceeds a certain value,
we need to know not only its mean and variance but also the form of its marginal distribution. Here we explore the marginal distributions of entries in $\hat{y}$ by varying (1) sequence lengths (using $c = 100$ and 1,000) and (2) the rate of change per site ($h$) without altering the relative weights of edges on the tree (we denote the relatively high rate in our model tree [fig. 1a] as $\lambda_1$, then reduce this rate by a factor of two ($\lambda_1/2$) and four ($\lambda_1/4$) as shown in fig. 2). We will restrict our attention to the entries in $\hat{y}$, which may be used as edge-length estimates in a tree, and will ignore $y_0$, which is a function of the other seven entries. Because of this tree's symmetry, the Monte Carlo distributions of $\hat{y}_1$ are equivalent to those of $\hat{y}_4$, and likewise for for the pair $\hat{y}_2$ and $\hat{y}_7$. All combinations of $c$ and $\lambda$ were checked with high resolution univariate, bivariate, and normal probability plots, but only illustrative examples are shown here.

The generalities that emerge are:

1. An arbitrarily small rate of change implies few multiple changes, so the the conjugation causes little difference between the values in $\hat{s}$ and $\hat{y}$. As a consequence, $\hat{y}$ has a distribution close to a multinomial. Figure 5a and b illustrate the binomial-like marginal distributions of $\hat{y}_1$ and $\hat{y}_3$ with $\lambda_1/4$ and $c = 100$. Increasing $c$ increases the expected values in $\hat{s}$, causing the marginal distributions to become more normal in shape.

2. When $c$ is small and $\hat{s}$ is distributed like a multinomial, there may be slightly nonlinear relationships between entries in $\hat{y}$ irrespective of the rate of change (for example, between $\hat{y}_1$ and $\hat{y}_2$ in fig. 6a, $c = 100$ and $\lambda_1/4$). High rates of change can exacerbate such effects. These relationships between $\hat{y}$ become linear as $c$ increases. Entries in $\hat{y}$ with $c = 1,000$ and $\lambda_1$ (fig. 6b) show that the distribution has converged close to multivariate normality (apart from a slight amount of skewness).

3. Smoothing due to correction for implied multiple changes can introduce either positive or negative skewness to the marginal distributions of $\hat{y}$. For example, the bipartition $\hat{y}_1$ in figure 5d shows smoothing and skewness. Smoothing has erased the discreteness present in the marginal distribution of $\hat{s}_3$ ($\sim B(100, 3.6)$). Smoothing is pronounced on this variable, because the two longest sets of paths, $\beta_3$ and $\beta_7$, which have the highest skewness, variance, and also a high correlation, both have the same sign in the summation from $p$ to give $\gamma_3$. That is, the largest potential errors, after taking into account multiple substitutions, are additive when estimating $\gamma_3$. The entries in $\gamma$ relating to the two longest tree edges ($\gamma_1$ and $\gamma_4$) have skewness and smoothing for the same reason, but their skewness is positive because of a reversal of the signs on $\beta_3$ and $\beta_7$ in their estimation from $p$ (fig. 5c).

4. High $\lambda$ and small $c$ both increase the skewness resulting from correction for multiple changes, which in turn results from bias. Figure 5c and d shows the skewness on $\gamma_1$ and $\gamma_3$, with $c = 100$ and $\lambda_1$. Increasing $c$ to 1,000, as shown in fig. 6b, reduces the skewness greatly.

5. If other factors are kept constant, as $c$ increases the amount of smoothing decreases around individual peaks in the marginal distributions of $\hat{y}$. However this decrease is not as rapid as the decrease in the size of the gaps between peaks, so that the marginal distributions of $\hat{y}$ become effectively continuous as $c$ becomes large.

6. In general, if the variance of $\gamma_1$ is $>5/c$, then the marginal distribution approximates a normal distribution in overall shape. Figure 5a illustrates a worst-case example where $\text{Var}[\hat{y}_1]$ just meets this criterion, and the marginal distribution retains much binomial character. When the variance of $\gamma_1$ is $<5/c$, $\hat{y}_1$ tends to show the characteristics of a binomial distribution with an expected value $<5$. An example of this is seen in figure 5b. We make these comments in light of the commonly used rule for approximating a binomial distribution with a normal distribution when $cs_i$ and $c(1-s_i)$ are $>5$ (Freund and Walpole 1987, p. 229). In work not presented, we describe hypothesis testing of entries in $\hat{y}$, elaborating on these findings.

In our simulations, the largest bias occurred on variable $\hat{y}_1$ with $c = 100$ and $\lambda_1$. This bias was 0.0065.
Fig. 5.—Marginal distributions of $\hat{\gamma}$ from a simulation of size $n = 10,000$. a and b, $\hat{\gamma}_1$ and $\hat{\gamma}_3$ with $\lambda_{14}$ and $c = 100$. c and d, $\hat{\gamma}_1$ and $\hat{\gamma}_5$ with $c = 100$ and $\lambda_4$. Implied number of changes per site.
FIG. 6.—Bivariate plots of entries in $\hat{\gamma}$ (excluding $\hat{\gamma}_6$) from a simulation of size $n = 10,000$. $a$, Parameters $c = 100$ and $\lambda_{14}$. $b$, Parameters $c = 1,000$ and $\lambda_{14}$.

The other major effect on the distribution of $\hat{\gamma}$ is that, as the number of taxa ($t$) increases, the number of possible sequence patterns increases exponentially. In addition, most of these patterns will be incompatible with any given tree, which typically results in any one of them having a low expected frequency of occurrence. So for even modest $t$, the majority of entries in $\hat{s}$ will have expected frequencies $<1$ for any biologically realistic sequence length (even for sequences of the order of $10^6$). As a consequence, most entries in $\hat{s}$ will have marginal distributions like those of a Poisson variable with an expected value $<1$. This poses a major problem for statistical testing, especially for overall tests of fit of data to model (including the tree), where the most tractable approaches using multivariate normal statistics (c.e.,
Bulmer 1991) will often be invalid. We are presently evaluating alternative strategies.

Properties of Covariance Matrices of \( \hat{\gamma} \) Calculated by the Delta Method

We now look at properties of covariance and correlation matrices estimated using the delta method. Here the population covariance matrix \( \mathbf{V}[\gamma] \) was estimated from 10,000 sample \( \hat{\gamma} \) (each originating as an independent sample of size \( c \) from \( s(T) \) for \( T \), with a specified \( \lambda \)). \( \mathbf{C}[\gamma] \) was then calculated from \( \mathbf{V}[\gamma] \). As noted earlier, \( \mathbf{V}[\gamma] \) is an approximation of \( \mathbf{V}[\mathbf{s}] \), the population covariance matrix of \( \hat{\gamma} \). Let \( \mathbf{V}[\mathbf{f}] \) be an estimate of the population covariance matrix calculated using the delta method, but replacing \( s(T) \) with \( \mathbf{s} \). Here we examine the bias of entries in \( \mathbf{V}[\mathbf{f}] \), \( \mathbf{C}[\mathbf{f}] \), \( \hat{\mathbf{V}}[\mathbf{f}] \), and \( \hat{\mathbf{C}}[\mathbf{f}] \), for combinations of \( c \) and \( \lambda \).

To examine the extent and pattern of bias introduced by the delta method when calculating the population covariance matrix, we plotted the elements of \( \mathbf{V}[\mathbf{f}] \) against \( \mathbf{V}[\gamma] \). With \( c = 1 \), all these plots were linear with a gradient indistinguishable from one. This was also the case with \( \lambda_{1/4} \) and \( c = 100 \). As a consequence, correlation matrices, \( \mathbf{C}[\gamma] \), estimated from \( \mathbf{V}[\gamma] \), in these instances were also indistinguishable from \( \mathbf{C}[\mathbf{f}] \). These results indicate that the delta method had converged to give accurate estimates of \( \mathbf{V}[\mathbf{f}] \). The gradient of the plots with \( c = 100 \) and \( \lambda_{1/2} \) and then \( \lambda_1 \) were 1.1 and 1.15, respectively, indicating underestimates of the entries in \( \mathbf{V}[\mathbf{f}] \) by approximately 10% and 15%, respectively. Closer examination of the size of the underestimate of each entry in \( \mathbf{V}[\mathbf{f}] \), in percentage terms, showed that those near zero did not all have the same percentage of bias (plots not shown). This showed up in the corresponding plots of \( \mathbf{C}[\mathbf{f}] \) versus \( \mathbf{C}[\gamma] \) as a slight S bend in what would otherwise have been a straight line constrained to go through 1, 0, and -1. Maximum departures of these correlation matrix entries with \( \lambda_{1/2} \) and \( \lambda_1 \) (.04 and .02, respectively) are considered insignificant.

Error and Bias when Estimating \( \mathbf{V}[\gamma] \) from Random Samples, \( \mathbf{s} \), Using the Delta Method

The marginal distributions of entries in \( \hat{\mathbf{V}}[\gamma] \) show features similar to those of \( \hat{\gamma} \), i.e., the possibility of some discreteness, smoothing, and skewness. This may be better understood by realizing that the calculation of these covariance matrices can also be achieved by applying a Hadamard conjugation to a vector representation of \( \mathbf{V}[\mathbf{s}] \) (see the Appendix). As expected, these marginal distributions show more smoothing and skewness than do those of \( \hat{\gamma} \) (distributions not shown). This is because \( \mathbf{V}[\mathbf{s}] \) has values more fine grained than those of \( \mathbf{s} \) and because the covariance matrix of \( \rho \) is a quadratic function of the covariance matrix of \( r \).

In contrast to the underestimation of entries in \( \hat{\mathbf{V}}[\gamma] \) by \( \mathbf{V}[\gamma] \), \( \hat{\mathbf{V}}[\gamma] \) appears considerably closer to being an unbiased estimator of \( \mathbf{V}[\gamma] \). Whereas entries in \( \hat{\mathbf{V}}[\gamma] \) are biased downward in absolute magnitude, \( \hat{\mathbf{V}}[\gamma] \) gives, on average, slight overestimates of the absolute magnitude of all entries in \( \mathbf{V}[\gamma] \). This is understandable because the delta method, which involves squares and products of pairs of estimated gradients, shows positive bias overall when applied to estimates of the gradients which are already skewed because of the natural log function. In effect, these two sources of bias cancel each other to a large degree, to give an overall small amount of bias.

The sampling variance of \( \hat{\mathbf{V}}[\mathbf{s}] \) is a larger contributor to the MSE of entries in \( \hat{\mathbf{V}}[\gamma] \) than is the bias. In all cases, including the most extreme, with \( c = 100 \) and \( \lambda_1 \), the bias contributed to < 1.5% of the MSE of any entry in \( \hat{\mathbf{V}}[\gamma] \). We could discern no correlation between the bias (as a proportion of the MSE) and either the magnitude of entries in \( \mathbf{V}[\gamma] \) or the observed variance of entries in \( \hat{\mathbf{V}}[\gamma] \) (plots not shown). Thus the influence of bias appears to be quite evenly distributed throughout entries in \( \hat{\mathbf{V}}[\gamma] \).

Converting variances to standard deviations (a square root transformation) removes the skewness in the marginals of these covariance matrix entries, except when \( c = 100 \). In this case there remains skewness due to the natural logarithm transformation from \( r \) to \( \rho \). Entries in \( \mathbf{C}[\gamma] \) are generally closer to normal in their marginal distributions than are entries in \( \mathbf{V}[\gamma] \) (except when distorted by the bounds of 1 or -1), yet they still have large variances. For example, the correlation entry of \( \gamma_5 \) and \( \gamma_6 \) in \( \mathbf{C}[\gamma] \) had 2.5% and 97.5% quantiles of approximately .5 to .6 and mean .05, when estimated with \( c = 100 \) and \( \gamma_1 \).

In a given sample, the variance entries in the sample covariance matrices \( \hat{\mathbf{V}}[\gamma] \) showed a high degree (all >.85) of correlation with corresponding entries in \( \mathbf{s} \). This indicates that the raw count for a character bipartition is the dominant variable in determining what the variance of the same bipartition in \( \hat{\gamma} \) will be. This is not unexpected in light of the relatively small correlations between entries within \( \hat{\gamma} \). As \( \lambda \to 0 \), the variance of \( \hat{\gamma}_i \), tends to the variance of \( \hat{s}_i \). The correlation of the \( i \)th variance in \( \hat{\mathbf{V}}[\gamma] \) with \( \hat{\gamma}_i \), in these simulations, was always distinctly less than its correlation with \( \hat{s}_i \). Again this is expected because \( \hat{\gamma} \) aims to represent the expected values of bipartitions, after taking into account multiple changes, but does not aim to represent their variances, whereas \( \hat{s}_i \), by its multinomial nature, reflects both in its magnitude. Most entries in \( \hat{\mathbf{V}}[\gamma] \) also show high correlations with many other entries in the same matrix. This is understandable, in that we are estimating the \( (2^c - 1)^2 \) entries in the covariance matrix from the \( 2^{c-1} \) values in \( \mathbf{s} \). In future, we will explore the implications of these findings.
for tree-selection methods that use information from a sample covariance matrix.

**Discussion**

The aim of this study has been to survey the statistical properties of Hadamard conjugations, as the foundation for putting this promising approach into a well-understood statistical framework. In this we have been successful, deriving the covariance matrix of \( \hat{\gamma} \) and studying the form of its multivariate distribution. This allows the development of appropriate statistical tests of parameters such as edge lengths and their relative rates for trees selected from \( \hat{\gamma} \) (P. J. Waddell, unpublished data). The observation that the distribution of \( \hat{\gamma} \) seems to incur a low amount of bias, even when relatively short sequences are analyzed, adds confidence to the estimates it gives and makes the construction of statistical tests on \( \hat{\gamma} \) more straightforward.

Another useful and pleasing finding is that the delta method introduces only a small bias when it is used to estimate the population covariance matrix from a random sample of sites. We recommend using the delta method with real data. As an alternative, when doing simulations with a known \( s(T) \), we would suggest caution in using \( V'[\hat{\gamma}] \) to estimate the population covariance matrix when \( n \) is small and/or \( \lambda \) high. The ideal would be to compute the exact variance of \( \rho_i \) (noting that \( \hat{\gamma}_i \) varies as a function of a binomial distribution), yet this becomes computationally expensive as path sets sum more events in the sequences. In such cases we anticipate that approximating \( \text{Var}[\rho_i] \) with the variance of a log-normal distribution will yield more accurate approximations than the delta method. That is, let the estimate of \( \text{Var}[\rho_i] \) be equal to the second moment of the distribution \( \ln(N \sim (\mu, \sigma)) \), where \( \mu = r_i \) and \( \sigma \) the variance of \( r_i \). One can be quite conservative in applying this improved approximation, since the largest increase in accuracy, as well as the greatest saving computationally, will be in estimating the variance of the longest sets of paths. We would then calculate the covariances of \( \rho_i \), with other entries replacing the delta method’s \( 1/r_i \) with the square root of \( \text{Var}[\rho_i]/\text{Var}[r_i] \).

Steel et al. (1992) have recently extended Hadamard conjugations to four character states under Kimura’s three-parameter model. The four-color \( \mathcal{S} \) vectors contain \( 4^{t-1} \) entries, but have the form of the Hadamard conjugation; that is, \( \hat{\gamma} = H^{-1}(\ln(HS)) \). Therefore the steps in computing the covariance matrix of \( \hat{\gamma} \) are the same, and the distributions of \( \mathcal{S}, \bar{S}, \hat{\rho}, \) and \( \hat{\gamma} \) are analogous to the two-color case illustrated here. The major difference when analyzing sequence data is that, with four colors, the entries in \( \mathcal{S} \) become sparse with increasing \( t \) much more quickly than in the two-color case. As a result, the vast majority of entries in \( \hat{\gamma} \) can have predominantly binomial forms (with expected values close to zero), an important consideration for statistical testing.

There remains the problem that, as the number of taxa increases, the number of entries in the covariance matrix of \( \hat{\gamma} \) which is \((2^{t-1})^2\) (or, worse still, \((4^{t-1})^2\) with four colors) soon exceeds the storage capacity of all computers. We are presently assessing methods (including approximations) that allow us to calculate just the variances (or covariances) of specific entries in \( \hat{\gamma} \), as an alternative approach to this problem.

**APPENDIX**

**The Calculation of HVH**

Since \( H \) and \( V \) are square \((2^{t-1} \times 2^{t-1})\) matrices, by ordinary multiplication we require a total of \( 2 \times (2^{t-1} \times 2^{t-1} \times 2^{t-1}) = 2^{3t-2} \) operations to calculate \( HVH \). The Fast Hadamard transform (a quicker way of multiplying a vector by \( H \); Tolimieri et al. 1989; Hendy and Penny 1993) reduces the total number of operations to \((t-1)2^{2t-1}\). This is still computationally expensive but possible, for up to 11 taxa, in less than 1 min on a 486 computer with sufficient RAM to store a matrix of \( 2^{2t-2} \) entries.

If we form the vectors \( v \) and \( \hat{v} \) each of \( 4^{t-1} \) components by concatenating the columns of \( V \) and \( HVH \), respectively, then, because of the structure of \( H \), \( v' = vH' \), where \( H' \) is the Kronecker product of \( H \) with itself, a Hadamard matrix of \( 4^{t-1} \) rows and columns (Tolimieri et al. 1989). Now applying the Fast Hadamard transform, the computation \( vH' \) also requires \((t-1)2^{2t-1} \) operations.

Estimating \( V'[\rho] \) from \( V[r] \) by the delta method uses the multiplication \( DV[r]D \), where \( D \) is a diagonal matrix with diagonal entries \( \delta_i \) corresponding to the gradient at the point of transformation of \( r_i \) to \( \rho_i \). This can be efficiently computed by multiplying the rows and then the columns in \( V[r] \) by the components of the vector \( \delta = (\delta_i) \).

**LITERATURE CITED**


**Felsenstein, J.** 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401–410.


MASATOSHI NEI, reviewing editor

Received April 16, 1993

Accepted October 27, 1993