Syst. Biol. 60(5):668–675, 2011

DOI:10.1093/sysbio/syr028
Advance Access publication on April 11, 2011

Improved Least Squares Topology Testing and Estimation

EDWARD SUSKO*

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5;
**Correspondence to be sent to: Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; E-mail: susko@mathstat.dal.ca.*

Received 1 September 2010; reviews returned 11 November 2010; accepted 6 January 2011
Associate Editor: Cecile Ane

Abstract.—Generalized least squares (GLS) methods provide a relatively fast means of constructing a confidence set of topologies. Because they utilize information about the covariances between distances, it is reasonable to expect additional efficiency in estimation and confidence set construction relative to other least squares (LS) methods. Difficulties have been found to arise in a number of practical settings due to estimates of covariance matrices being ill conditioned or even noninvertible. We present here new ways of estimating the covariance matrices for distances that are much more likely to be positive definite, as the actual covariance matrices are. A thorough investigation of performance is also conducted. An alternative to GLS that has been proposed for constructing confidence sets of topologies is weighted least squares (WLS). As currently implemented, this approach is equivalent to the use of GLS but with covariances set to zero rather than being estimated. In effect, this approach assumes normality of the estimated distances and zero covariances. As the results here illustrate, this assumption leads to poor performance. A 95% confidence set is almost certain to contain the true topology but will contain many more topologies than are needed. On the other hand, the results here also indicate that, among LS methods, WLS performs quite well at estimating the correct topology. It turns out to be possible to improve the performance of WLS for confidence set construction through a relatively inexpensive normal parametric bootstrap that utilizes the same variances and covariances of GLS. The resulting procedure is shown to perform at least as well as GLS and thus provides a reasonable alternative in cases where covariance matrices are ill conditioned. [Confidence sets of trees; distance methods; generalized least squares; topology tests; weighted least squares.]

Least squares (LS) estimation and weighted least squares (WLS) estimation have a long history in phylogenetics (Cavalli-Sforza and Edwards 1967; Fitch and Margoliash 1967). Given a set of distances, d_{ij}, and weights, w_{ij}, the WLS statistic for a tree is obtained by choosing the distances δ_{ij} on that tree that minimize

$$\sum_{i<j} w_{ij}(d_{ij} - \delta_{ij})^2. \tag{1}$$

More specifically, δ_{ij} is the sum of the branch lengths along the path from i to j, and it is the branch lengths that are chosen to minimize Equation (1). Because the choice of δ_{ij} is topology dependent, the test statistic calculated in Equation (1) depends on topology. LS methods choose as an estimate the topology that gives the minimum value of Equation (1). For LS estimation, the $w_{ij} = 1$, whereas for conventional WLS estimation, w_{ij} is set to an estimate of the inverse variance for the pair of taxa i and j.

It will be convenient to express Equation (1) in vector notation. Let $\boldsymbol{y}$ denote the vector of estimated distances, $(d_{12},\ldots,d_{1T},d_{23},\ldots,d_{T-1T})$. The distances, δ_{ij} can be expressed as the sum of the edge lengths α_l in the path from i to j. Alternatively,

$$\delta_{ij} = \sum_l x_{ij,l}\alpha_l, \tag{2}$$

where $x_{ij,l} = 1$ or 0 according to whether or not the lth branch is in the path from i to j. Let X be the matrix with $x_{ij,l}$ being the entry in the lth column for the row corresponding to the pair of taxa i and j. Then in matrix notation, $\boldsymbol{\delta} = X\boldsymbol{\alpha}$, and the WLS test statistic is

$$(\boldsymbol{y} - X\alpha)^T W(\,\boldsymbol{y} - X\boldsymbol{\alpha}), \tag{3}$$

where the W matrix is diagonal with the w_{ij} entries along its diagonal. GLS methods are the same as WLS methods except that W is replaced by an estimate of V^{-1}, the inverse of the covariance matrix of the distances.

According to the theory for regression analysis, if the entries of the y vector are normally distributed and uncorrelated, the WLS estimates of α are of uniformly minimum variance among unbiased estimators when the weights in W are taken as the inverse variances of the y vector (cf. Theorem 6.1.4 of Bickel and Doksum 2007 after transformation to so-called canonical form). In the case that the entries are correlated, however, the optimal choice of W is the inverse of the covariance matrix for the $\boldsymbol{y}$ vector. Because of the shared path lengths of pairs, distances are indeed correlated. Bulmer (1991) was the first to investigate the use of GLS and, for distances obtained through maximum likelihood (ML) methods, Susko (2003) provided general formulae for calculation of V and showed that distances have, for large sequence lengths, an approximate normal distribution. Optimality here refers to optimality of edge length estimation for the correct topology that may only be loosely related to good topological estimation. Moreover, the theoretical optimality results assume a known, rather than estimated, covariance matrix is used in GLS calculations.

A different reason for interest in GLS over other LS methods is that, for the correct tree, if W gives a consistent approximation to V^{-1}, the GLS statistic Equation (3) has an approximate chi-squared distribution with

$T(T-1)/2-(2T-3)$ degrees of freedom (Susko 2003). This allows one to construct a 95% confidence set of trees as the set of all trees that have a GLS test statistic value less than the 95th percentile of the chi-squared distribution with $T(T-1)/2-(2T-3)$ degrees of freedom. There is an equivalence to testing topologies here. Any tree that is not in the 95% confidence set can be rejected at the 5% level. Alternatively, a P value for the hypothesis that the given tree is the correct tree, is the probability that a chi-squared random variable is larger than that given tree's GLS statistic. In any case, because of the chi-squared distribution of the GLS statistic, testing rather than estimation was the focus of Susko (2003) and GLS was shown to have reasonable performance in a number of cases involving a relatively small number of taxa. A software implementation was made available at http://www.mathstat.dal.ca/tsusko. A major difficulty with the software implementation turned out to be that, with small numbers of sites or almost identical sequences, the estimated covariance matrix was often ill conditioned or even not invertible. That this can lead to poor performance or even an inability to calculate the GLS statistic has been illustrated in Shi et al. (2005), Sanjuán and Wróbel (2005), and Czarna et al. (2006).

Because of the difficulties with matrix inversion that sometimes arise in computing the GLS statistic, Czarna et al. (2006) proposed using WLS for confidence set construction. Although estimates of the variances need to be inverted to obtain the WLS weights, those variances will only be small for very small distances, a case that can be dealt with through removal of almost identical sequences. The WLS statistic is consequently more stable than the GLS statistic. The difficulty that arises is that the WLS statistic no longer has a known chi-squared distribution, under the null hypothesis that a given topology is correct, that can be used to convert test statistics into P values. Czarna et al. (2006) nevertheless use a chi-squared distribution with $T(T-1)/2-(2T-3)$ degrees of freedom to calculate P values. This would only be truly justified if distances are independent. Distances are not, however, independent, particularly in settings where the covariance matrix is approximately singular as singularity of the covariance matrix can only arise if a distance has zero variance or if one of the distances is linearly related to the others; the latter is a case of strong dependence. Simulations conducted here confirm that using chi-squared approximations for P value calculation gives resulting WLS confidence sets with poor statistical properties relative to GLS methods.

In Susko (2003), GLS was presented mainly as a tool for testing rather than estimation. Although an arbitrary set of trees could be input, tree searching was not conducted and edge lengths were not output. This has changed in the new version of the software that includes routines that allow estimation through subtree prune and regraft (SPR) topology searching. The routines can output all trees encountered during searching that are in a confidence set. Similar simulations as were used for investigating testing performance are used here to compare estimation performance. Surprisingly, given its poor testing performance, WLS performs well in estimation. This suggests that the poor test performance of WLS was due to an inappropriate assumption of independent distances. Moreover, the solid estimation performance suggests that the WLS test statistic might do well at discriminating between different topologies if a correct distribution is used in calculating P values. Although a closed-form distribution for the WLS test statistic is not available, it is known from Susko (2003) that the distances used to construct it are approximately multivariate normally distributed and what their approximate covariances are. This leads to a normal parametric bootstrap approach to calculating WLS P values that is much less computationally intensive than usual bootstrapping.

METHODS

A More Robust Covariance Matrix Estimator

The fundamental difficulty with the covariance matrix estimator given in Susko (2003) is that it might be singular. Consequently, the GLS statistic may not even be defined. What is desired is to obtain an estimator that maintains the property of converging upon the correct covariance matrix with large sequence lengths yet is less likely to be singular. This is accomplished in two ways. First, different formulae are used for covariance matrix construction that are asymptotically equivalent to the formulae in Susko (2003) but guaranteed to be nonnegative definite as is always the case for actual, rather than estimated, covariance matrices. Second, weighting of the contributions to the covariance matrix in Susko (2003), which was by the observed proportions of times patterns arose, is replaced by weighting through the probabilities of patterns under a fitted model. Let x denote a site pattern that will implicitly depend on the taxa under consideration. For instance, with four taxa, $x = AACG$ would be a pattern for Taxa 1–4, whereas $x = CG$ would be the corresponding site pattern for Taxa 3 and 4. The variance of the estimated ML distance given in Susko (2003) can be expressed as

$$\hat{V}_{jj} = \left[\sum_x \hat{p}_x \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_j)}{p(x\,;d_j)} \right\}^2 - \sum_x \hat{p}_x \frac{\frac{\partial^2}{\partial d^2} p(x\,;d_j)}{p(x\,;d_j)} \right]^{-1} \Big/ n, \tag{4}$$

where sums are over all patterns for the pair, $\hat{p}_x$ is the frequency with which the xth pattern arose for the jth pair and $p(x\,;d_j)$ is the probability of pattern x for the jth pair under the model of evolution and evaluated at the jth estimated distance. The covariance between the jth distance and the kth distance can be expressed as

$$\hat{V}_{jk} = n\hat{V}_{jj}\hat{V}_{kk} \sum_x \hat{p}_x \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_j)}{p(x\,;d_j)} \right\} \cdot \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_k)}{p(x\,;d_k)} \right\}, \tag{5}$$

where now the sum is over all patterns for the taxa in the pair j and in the pair k. For instance, if j was the pair of Taxa 1 and 2, and k was 1 and 3, the sum would be over terms like $x = ACG$ giving the character states for 1, 2, and 3. We also use x_j and x_k to denote the patterns for the jth and kth pair; in the above example, $x_j = AC$ and $x_k = AG$. As discussed in Susko (2003), these covariance matrix estimators will be statistically consistent: they will converge upon the true covariance matrix as sequence length gets large.

The first correction to avoid badly behaved estimates is to ignore the second term in the sum in Equation (4). This adjustment does not alter the statistical consistency properties because the second term can be shown to be approximately zero with large sequence lengths. At the same time, it produces a covariance matrix estimator that is guaranteed to be nonnegative definite. The reason for this is that the new covariance matrix estimator, $\hat{V}^*$ can now be expressed as

$$\hat{V}^* = \sum_x \hat{p}_x \boldsymbol{a}_x \boldsymbol{a}_x^T, \tag{6}$$

where the sum is now over patterns for all the taxa jointly and $\boldsymbol{a}_x$ is a vector with jth component

$$a_{xj} = \hat{V}_{jj}^* \frac{\frac{\partial}{\partial d} p(x\,;d_j)}{p(x\,;d_j)}.$$

A matrix V is nonnegative definite if $\boldsymbol{b}^T V \boldsymbol{b} \geqslant 0$ for any vector $\boldsymbol{b}$. We have

$$\boldsymbol{b}^T \hat{V}^* \boldsymbol{b} = \sum_x \hat{p}_x [\, \boldsymbol{b}^T \boldsymbol{a}_x]^2 \geqslant 0, \tag{7}$$

so that $\hat{V}^*$ is guaranteed to be nonnegative definite, whereas the original estimator was not.

Equation (7) also suggests a second way of adjusting the estimate of the covariance matrix so that it is more likely to be nonsingular. The estimate $\hat{V}^*$ will be singular only if Equation (7) is zero for some nonzero vector $\boldsymbol{b}$. Although the sum in Equation (7) is over all patterns x, the only patterns that contribute are those with $\hat{p}_x > 0$. The more terms there are with $\hat{p}_x > 0$, the less likely it is that the sum will be zero. With a small number of sites or closely related taxa, it will frequently be the case that many of the $\hat{p}_x$ are zero. This is one of the reasons that singular matrix estimates had arisen with the previous estimate. One way of maintaining the statistical consistency properties while increasing the number of patterns for which $\hat{p}_x > 0$ is to replace $\hat{p}_x$ with the estimated probability of the pattern x under the model. This may, at first sight, seem computationally infeasible as the number of patterns grows exponentially with the number of taxa. However, there are at most four taxa involved in the expressions for each pair in Equation (4) or each set of pairs in Equation (5).

What is needed to implement this second adjustment is an estimated tree. In practice, this is achieved in two ways. The first, which we refer to as single weighting uses the neighbor-joining (Saitou and Nei 1987) topology (NJ topology) with constrained LS edge length estimates. In order to ensure that pattern probabilities are positive for any pattern, zero edge lengths are set to a small but positive constant; this was 1.0×10^{-12} for the results reported here. The second approach, which we refer to as multiple weighting, uses different weights for each hypothesized topology. Pattern probabilities are determined using the hypothesized tree with constrained LS edge length estimates.

In summary, the covariance matrix estimates are obtained through

$$\hat{V}_{jj}^* = \left[\sum_x p_x \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_j)}{p(x\,;d_j)} \right\}^2 \right]^{-1} \Big/ n \tag{8}$$

and

$$\hat{V}_{jk}^* = n \hat{V}_{jj}^* \hat{V}_{kk}^* \sum_x p_x \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_j)}{p(x\,;d_j)} \right\} \cdot \left\{ \frac{\frac{\partial}{\partial d} p(x\,;d_k)}{p(x\,;d_k)} \right\}, \tag{9}$$

where p_x is the estimated probability of pattern x.

That the new form of covariance matrix is less likely to suffer from singularity problems can be illustrated with a four taxa example. Roux (2009) established that the covariance matrix becomes almost singular in the case of a Jukes–Cantor model (Jukes and Cantor 1969) with all small edge lengths. This can be expected to be a problematic setting for GLS because estimated edge lengths will often be close to zero. Table 1 gives the means and variances of the old and new covariance matrix estimates after 1000 simulations from a Jukes–Cantor model and four taxa tree with all equal edge lengths. For comparison, the actual observed covariance matrix entries were also calculated for the 1000 simulated sets of estimated distances. With edge lengths $t = 0.001$ and sequence length $n = 1000$, estimation is

TABLE 1. Means and standard deviations (multiplied by 10^6) of covariance matrix estimates for 1000 data sets simulated from four-taxa trees with all equal edge lengths, t. The simulated column gives the sample covariance matrix entries over the 1000 data sets

			Old		New	
Setting	Entry	Simulated	Mean	SD	Mean	SD
$n = 1000$	Var(d_{12})	2.10	−351.94	969.06	2.06	1.46
$t = 0.001$	Var(d_{13})	2.85	−153.04	667.11	2.96	1.70
	Cov(d_{12}, d_{13})	1.03	181.01	1260.48	1.01	1.00
	Cov(d_{12}, d_{34})	0.03	144.00	1129.84	0.00	0.02
	Cov(d_{12}, d_{14})	0.96	181.01	1260.48	1.01	1.00
	Cov(d_{13}, d_{24})	1.01	81.98	850.29	0.98	0.95
$n = 1000$	Var(d_{12})	20.57	20.36	4.73	20.36	4.73
$t = 0.01$	Var(d_{13})	31.24	30.82	5.94	30.83	5.94
	Cov(d_{12}, d_{13})	9.98	10.01	3.18	10.03	3.16
	Cov(d_{12}, d_{34})	−0.16	−0.02	0.63	0.00	0.00
	Cov(d_{12}, d_{14})	9.26	10.03	3.17	10.03	3.16
	Cov(d_{13}, d_{24})	8.92	10.08	3.25	10.10	3.17
$n = 10000$	Var(d_{12})	0.21	0.20	0.05	0.20	0.05
$t = 0.001$	Var(d_{13})	0.30	0.30	0.06	0.30	0.06
	Cov(d_{12}, d_{13})	0.11	0.10	0.03	0.10	0.03
	Cov(d_{12}, d_{34})	0.00	0.00	0.00	0.00	0.00
	Cov(d_{12}, d_{14})	0.11	0.10	0.03	0.10	0.03
	Cov(d_{13}, d_{24})	0.10	0.10	0.03	0.10	0.03

improved dramatically with the new covariances. The reasons for this are clear because of the simplicity of the setting. Neighboring taxa expect substitutions only once in 1000 sites. It is thus not uncommon that two or more sequences will be identical in which case the old form gives a singular matrix and even gives a negative variance estimate. The setting with 1000 sites and edge lengths of 0.001 push the limits of what is achievable for variance approximation. The other settings listed in Table 1 illustrate that the two variance approximations are comparable in settings where the variance matrices are more stable. However, even in the setting with $t = 0.001$ and $n = 10,000$, where estimation was good, the estimated covariance matrix with the old approximation was not positive definite 26.8% of the time. In the settings $t = 0.001$, $n = 1000$, and $t = 0.01$, $n = 1000$, the estimated matrix was not positive definite 36.3% and 5.1% of the time. In contrast, the estimated covariance matrix using the new approximations was always positive definite.

WLS Testing

Sanjuán and Wróbel (2005) and Czarna et al. (2006) considered the use of WLS rather than GLS for confidence set construction. There are reasons to suspect that WLS may perform better than GLS in cases where the covariance matrix is close to singular. This can most easily be seen by restating GLS as a form of WLS. Let $\boldsymbol{u}_i$ denote the ith eigenvector of the covariance matrix, having associated eigenvalue λ_i. Then, letting $d_i^* = \sum_i u_{ij} d_j$ be the distances linearly transformed by the ith eigenvector, the eigenvalue λ_i can be interpreted as the estimated variance of d_i^*. GLS estimation is equivalently WLS estimation with the d_i^* as variables and the $1/\lambda_i$ as weights. WLS, by comparison, used the distances d_i directly as variables and weights by the inverse of their estimated variances. In both cases, small errors in the estimated variances may result in large errors in test statistics because these invert the variances. In cases where the covariance matrix is almost singular, eigenvalues will be close to zero, whereas the variance of distances need not be, thus making the WLS statistic more stable with respect to small errors in variance estimates. To illustrate, in the case of a four-taxon tree with equal edge lengths of 0.001 and sequence length 1000, the smallest variance for a distance is approximately 2×10^{-5}, whereas the smallest variance for an eigen transformed distance is approximately 2×10^{-7}, 100 times as small.

Although WLS statistics can be expected to be better behaved, they do not have a known distribution under the null hypothesis that a given tree is correct. Accordingly, it is difficult to determine cutoffs for P values for tests. Czarna et al. (2006) deal with this by assuming independence of the distances. If true, this assumption would lead to the same approximate large-sequence length distribution that applies to GLS. In some cases, the assumption of independence is reasonable. For instance, the observed correlation of d_{12} and d_{34} in Table 1 are close to zero. In fact, for the Jukes–Cantor model, this correlation is exactly zero. (This follows from, for instance, the results in Steel et al. 2000; the covariances reported in Table 1 vary from zero because they are based on a finite number of simulated data sets.) However, in many more cases, it is a poor approximation. For instance, the observed correlation was 0.40 for the distances d_{12} and d_{13} over 1000 simulations with $t = 0.01$. The correlations for d_{12}, d_{14}, d_{13}, d_{23}, and other pairs of distances were similarly close to 0.40. As pointed out in Czarna et al. (2006), the end effect on the properties of the WLS test that assumes independence is to make the test conservative: a 95% confidence set of trees has probability larger than 95% of containing the true tree and consequently contains more trees than if the actual (albeit unknown) distribution for the WLS statistic is used.

Although the actual limiting distribution of the WLS statistic is not available in closed form, the availability of the covariance estimates (8) and (9) make it possible to obtain appropriate thresholds and P values through a simple parametric bootstrap. To see this, it is valuable to note that the WLS statistic can be expressed as

$$\boldsymbol{y}^T[W - WX(X^TWX)^{-1}X^tW]\boldsymbol{y}. \tag{10}$$

It follows from the results in Susko (2003) that $\boldsymbol{y}$ has an approximate multivariate normal distribution with mean vector $X\delta$ and covariance matrix V that can be estimated through Equations (8) and (9). Because

$$[W - WX(X^TWX)^{-1}X^TW]X\delta = 0$$

the same WLS statistic is obtained in Equation (10) if $\boldsymbol{y}$ is replaced by $\boldsymbol{y}^* = \boldsymbol{y} - X\delta$. The advantage with using $\boldsymbol{y}^*$ is that it has an approximate multivariate normal distribution with means zero and covariance matrix V; the only unknown in the distribution is V that can be estimated through Equations (8) and (9). Thus, the following parametric bootstrap can be used to approximate the distribution of the WLS statistic.

1. Generate $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_B$ from a multivariate normal distribution with means zero and covariance matrix $\hat{V}^*$ given by Equations (8) and (9).
2. Substitute $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_B$ in Equation (10) to obtain B generated WLS statistics, $WLS_1, \ldots, WLS_B$.
3. If WLS_o is the observed WLS statistic for the data, an approximate P value is the proportion of $WLS_i \geqslant WLS_o$.

Note that the bootstrapping here is much less computationally intensive than conventional bootstrapping that requires repeated generation of alignments, repeated ML estimation of all distances, and repeated WLS test statistic calculation for these distances.

We will refer to the WLS test that uses 1–3 as the WLSN test; the N indicates normal simulation is being used. Because of the differing choices for estimation, there are several versions that will be considered. First, in calculating observed WLS statistics, estimation can either constrain edge lengths to be positive or not. Second, single or multiple covariance matrices can be used. Single matrices are constructed from Equations (8) and (9) using pattern probabilities calculated for an NJ tree with LS edge lengths. Multiple matrices use separate pattern probabilities for each hypothesized tree under consideration. Together, these options give rise to four different versions of the test. Similarly, there are four versions of the GLS test. The weights in the observed WLS statistics are the inverse variance estimates (8), where the p_x are the probabilities of pairwise patterns calculated using the estimated distances for the pair under consideration.

Results

Simulation Settings

Simulations were conducted with the five through eight taxa topologies given in Figure 1 as well as with the 10, 15, and 20 taxa trees that are given there. For the five through eight taxa simulations, terminal edge lengths for each simulation were generated from a $U(0.01, 0.1)$ distribution and internal edge lengths were set to 0.01. The substitution model used was the F84 model (Felsenstein (2005), DNAML program since 1984, PHYLIP Version 2.6) with all equal frequencies and transition–transversion ratio set to 2. The 10, 15, and 20 taxon trees are the same trees that were used in Shi et al. (2005); these are listed in the appendix of that reference. The same substitution model was used for simulation as in Shi et al. (2005): an HKY model (Hasegawa et al. 1985) with transition–transversion ratio set to 2.93 and frequencies of A, C, G, and T equal to 0.37, 0.24, 0.12, and 0.27.

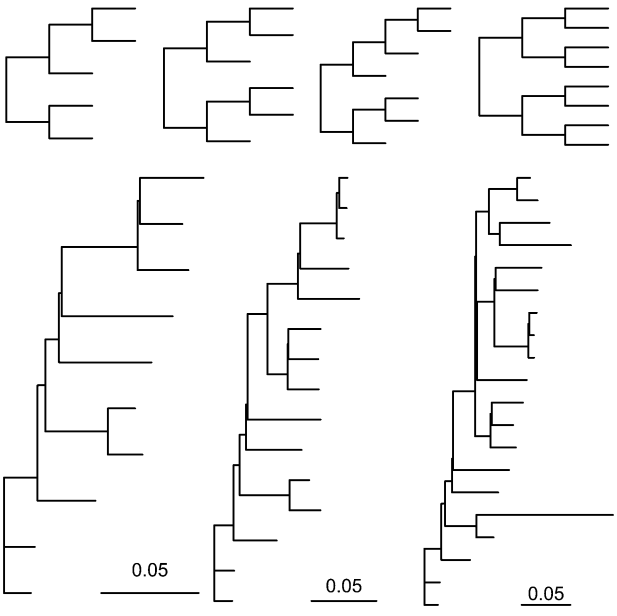

FIGURE 1. The trees used in simulation. The four topologies on the first row are for the examples involving five through eight taxa; internal edge lengths were set to 0.01 and terminal edge lengths were generated uniformly in the range 0.01 through 0.1. The bottom 3 trees were used for 10, 15, and 20 taxa.

The five through eight taxa simulations have the advantage that all trees can be searched for estimation, and all trees can be tested for inclusion in a confidence set; the total number of possible eight taxa trees is 10,395. For the 10, 15, and 20 taxa simulations that consider estimation, an NJ tree was used as a starting tree and SPR searching was conducted. Given a current tree, the test statistic (either GLS or WLS) was calculated for SPR trees, in order of distance from the current tree, until a tree was found giving a better test statistic value, at which point a new SPR search started from this new tree; if no such tree could be found, the SPR search was stopped. For simulations considering confidence set construction performance, all trees within one SPR of the generating tree were considered for inclusion in the confidence set. For the WLSN test, $B = 1000$ simulations were used in Steps 1 and 2.

Simulation Results

We first consider estimation performance. Table 2 gives the numbers of times a given Robinson–Foulds distance was obtained for the five through eight taxa simulations. The first thing that is apparent is that, for any of the LS methods, improvements in topological estimation are attainable by constraining edge lengths to be nonnegative. Because the computational cost of doing so is not substantial, we only consider LS methods with constrained edge length estimation in what follows. None of the LS methods give consistently better performance than the baseline method, NJ. This is perhaps not that surprising as the probabilities of correct reconstruction are relatively high and all the methods are comparable in performance. As Table 3 indicates, in the 10, 15, and 20 taxa simulations, where estimation is more difficult, performance is still comparable to NJ.

Next, we consider testing performance. Table 4 gives results for the five through eight taxa simulations. The first observation of note is that the performance of WEIGHTLESS program version 2.93 (Sanjuán and Wróbel 2005) and WLS are almost identical; a minor bug in the WEIGHTLESS implementation had the WLS statistics being twice as large as they should be which was corrected for. The difference between WLS and WEIGHTLESS is only in the estimation of the variances used for weighting. For WEIGHTLESS, these are constructed using bootstrapping, whereas for WLS, they are calculated using Equation (8). Because bootstrapping is much more computationally intensive, the similarity of the results suggest that it is better to use Equation (8); these are consequently the only choices considered in the 10 through 20 taxa simulations.

WLSN refers to the confidence set constructed using the WLS statistic but with normal simulations to

TABLE 2. The numbers of times a given Robinson–Foulds (RF) distance was obtained between the generating and estimated tree over 1000 simulations with and without nonnegativity constraints on edge lengths

	RF Distance														
	Five Taxa			Six Taxa				Seven Taxa				Eight Taxa			
	0	2	4	0	2	4	6	0	2	4	⩾ 6	0	2	4	⩾ 6
No Constraint															
LS	832	161	7	776	214	9	1	728	245	27	0	735	234	21	4
WLS	842	151	7	790	200	10	0	740	235	25	0	764	211	23	2
GLS(single)	728	245	27	644	306	47	3	543	342	98	14	569	298	92	42
GLS(multiple)	834	159	7	782	206	11	1	734	234	31	1	772	201	23	4
Nonnegative															
LS	970	30	0	971	29	0	0	957	43	0	0	944	50	5	1
WLS	973	26	1	973	27	0	0	958	42	0	0	950	45	5	0
GLS(single)	956	41	3	942	57	1	0	924	67	9	0	918	70	10	2
GLS(multiple)	973	26	1	975	25	0	0	959	41	0	0	956	42	2	0
NJ	970	30	0	971	29	0	0	958	42	0	0	964	32	3	1

determine critical values. The test statistics for WLS and WLSN were exactly the same for each of the simulated data sets. These two methods differ only in the way in which cutoffs for inclusion in the confidence set are determined. The reason for the poor performance of WLS is thus clear and indicated by its coverage. Although the WLSN confidence sets usually have approximate 95% coverage, the WLS sets always contain the generating tree. Thus, for WLS, thresholds for inclusion in the confidence set have been set much larger than is required for a 95% confidence set, and consequently, the confidence set tends to include many more trees than it needs to. The two methods WLSN and GLS give coverages that are reasonably comparable to 95% over 1000 simulations. They also exhibit similar confidence set sizes in the five through eight taxa simulations. WLSN, however, has a substantially higher proportion of cases where the generating tree is the only tree in the confidence set.

Because of the similarity of the results between WLS and WEIGHTLESS, the less computationally intensive WLS routine was used for the 10, 15, and 20 taxa simulations reported in Table 5. As with the five through eight taxa simulations, the thresholds used by WLS cause many more trees to be included in the confidence sets than is necessary. In the 10, 15, and 20 taxa simulations, the WLS and GLS tests tend to undercover. WLSN tends to give smaller regions with GLS including one to three more trees on average. The results reported in Tables 4 and 5 were for the GLS and WLSN methods separately estimating covariance matrices for each topology tested. As Table 6 indicates, it is always better to use multiple matrices when testing. However, the performance of GLS degrades much more than performance of WLS when single covariance matrices are used. The results with larger numbers of taxa, in particular, suggest that GLS should usually be used with multiple covariance matrices.

TABLE 3. Estimation results over 1000 simulations from each of the 10,15, and 20 taxa trees. Given are the proportion of times the generating topology was estimated and other summary statistics for the RF distances between estimated and generating trees over the 1000 simulations. Except for NJ, nonnegativity constraints were imposed on edge lengths.

Number			RF Distance			
of taxa	Method	Proportion	Mean	Median	SD	Max
10	LS	0.278	1.966	2.000	1.538	6.000
	WLS	0.272	1.978	2.000	1.533	6.000
	GLS(single)	0.240	2.206	2.000	1.677	8.000
	GLS(multiple)	0.271	2.000	2.000	1.545	6.000
	NJ	0.262	2.042	2.000	1.556	8.000
15	LS	0.121	3.270	4.000	2.029	10.000
	WLS	0.132	3.232	4.000	2.037	10.000
	GLS(single)	0.095	3.872	4.000	2.325	12.000
	GLS(multiple)	0.115	3.428	4.000	2.062	10.000
	NJ	0.112	3.418	4.000	2.076	10.000
20	LS	0.024	5.756	6.000	2.640	14.000
	WLS	0.029	5.600	6.000	2.606	14.000
	GLS(single)	0.031	5.940	6.000	2.863	18.000
	GLS(multiple)	0.032	5.538	6.000	2.642	14.000
	NJ	0.024	5.622	6.000	2.581	14.000

TABLE 4. Properties of 95% confidence sets of trees. Given are the proportions of time the generating tree was the only tree in the set, the mean number of trees in the set and the proportions of time the generating tree was in the set (Coverage). Results were obtained for the WLS with variances estimated through bootstrapping and chi-squared critical values (WEIGHTLESS), WLS with variance formulae (WLS), WLS with nonnegativity restriction on edge lengths and normal simulation to obtain critical values (WLSN) and GLS; covariance matrices were calculated with probability weighting done separately for each tree tested

Taxa	Method	Proportion	Mean size	Coverage
5	WEIGHTLESS	0.000	13.666	1.000
	WLS	0.000	13.676	1.000
	WLSN	0.724	1.468	0.943
	GLS	0.685	1.540	0.937
6	WEIGHTLESS	0.000	78.316	1.000
	WLS	0.000	78.556	1.000
	WLSN	0.574	1.889	0.942
	GLS	0.511	2.077	0.935
7	WEIGHTLESS	0.000	436.131	1.000
	WLS	0.000	438.490	1.000
	WLSN	0.508	2.767	0.951
	GLS	0.416	3.163	0.945
8	WEIGHTLESS	0.000	6523.393	1.000
	WLS	0.000	6557.942	1.000
	WLSN	0.412	4.054	0.952
	GLS	0.360	4.504	0.948

TABLE 5. Properties of 95% confidence sets of trees for simulations with 10, 15, and 20 taxa. Given are the proportions of time the generating tree was the only tree in the set, the mean number of trees in the set and the proportion of times the generating tree was in the set (Coverage). Results were obtained for the WLS which chi-squared critical values, WLS with nonnegativity restriction on edge lengths and matrices were calculated with probability weighting done separately for each tree tested

Taxa	Method	Proportion	Mean size	Coverage
10	WLS	0.000	60.242	1.000
	WLSN	0.000	9.871	0.945
	GLS	0.006	10.833	0.925
15	WLS	0.000	241.090	1.000
	WLSN	0.000	14.757	0.924
	GLS	0.002	19.503	0.889
20	WLS	0.000	711.350	1.000
	WLSN	0.000	35.372	0.957
	GLS	0.000	38.668	0.921

DISCUSSION

We have presented here modifications of the variance formulae of Susko (2003) that are asymptotically equivalent to the previous formulae but much less likely to have the difficulties of the previous GLS test statistic that were due to ill behaved, almost singular covariance matrix estimates. These matrices can be used in GLS statistic calculation or to determine cutoffs for inclusion in confidence sets for the WLS statistics. What was a bit surprising in the estimation and testing simulations was that WLSN had a consistently better performance than GLS; although often the improvement was marginal. Both methods performed much better than the WLS approach suggested by Czarna et al. (2006) that effectively applies the GLS test with covariances set to zero. The simulations here make clear that the resulting chi-squared thresholds for inclusion in a confidence set are much too large and cause many more topologies to be included in the confidence region than is needed.

Although the WLSN test was the best performer, this performance improvement comes with a computational cost. To illustrate that, CPU times were obtained on the same machine and for the same 7 and 20 taxon data sets. In the seven taxa case, WLSN and GLS P values were obtained for 945 trees, and in the 20 taxa case, P values were obtained for 100 trees. The times required for GLS were 13.2 s and 2 min 52 s for 7 and 20 taxa cases, whereas for WLSN, they were 55.5 s and 7 min 42 s. The additional time required is primarily due to the normal simulation required for P value determination. For estimation alone, WLS can be expected to be faster.

TABLE 6. The proportions of time the generating tree was the only tree in the set and the mean number of trees in the set when using single or multiple (one for every hypothesized tree) covariance matrix estimates

			Number of Taxa		
Weighting	Method		5	6	7
Single	WLSN	Proportion	0.589	0.419	0.362
		Mean size	1.874	2.667	3.425
	GLS	Proportion	0.335	0.114	0.062
		Mean size	2.798	7.352	18.221
Multiple	WLSN	Proportion	0.710	0.591	0.519
		Mean size	1.530	1.931	2.421
	GLS	Proportion	0.686	0.530	0.417
		Mean size	1.578	2.127	2.932

The CPU times reported are for multiple covariance matrices and nonnegativity constraints. Estimation can constrain edge lengths or not and covariance matrices can be calculated for a single pilot tree estimate or separately for each tree tested. In terms of performance, the results here indicate that it is better to constrain edge length estimates. This comes with only a slightly higher computational cost. In fact, CPU times with nonnegativity constraints were sometimes shorter than without. This is likely due to an inefficient implementation of the unconstrained approach by comparison with the nonnegative LS routine that was used and due to Lawson and Hanson (1974). Nevertheless, the similarity of the times suggest that constraints are worth the computational cost. Leaving edge lengths unconstrained does not usually greatly affect the GLS or WLS test score of the true tree but the added flexibility of negative edge lengths allows poor trees to fit distances better and thus deflates GLS or WLS scores for these trees. Although using separate covariance matrices also gave better test performance, it comes with a more substantial computational cost than nonnegativity constraints. For the 7 and 20 taxon examples where CPU times were obtained, the CPU times for GLS with a single covariance matrix were 0.1 s and 3 s for the 7 and 20 taxa examples by comparison with the 13.2 s and 2min 52 s times obtained with multiple covariance matrices. For WLSN, the times were 14 s and 3 min 33 s by comparison with the 55.5 s and 7 min 42 s obtained with multiple covariance matrices.

For the GLS test, results with a larger number of taxa suggest that, for final results, multiple covariance matrices should be used in spite of the additional computational cost. For WLSN, however, the performance improvements were not as substantial and the additional computational cost might reasonably be avoided in cases where a larger number of trees are tested for inclusion in the confidence set. An alternative practical approach to dealing with larger data sets is to use a single matrix, and the GLS test on a large initial set of hypothesized trees and then rerun with multiple covariance matrices on a reduced set. It should be emphasized that by large data sets, we mean that the number of taxa are large. Because site patterns for between 2 and 4 taxa are being summed over in covariance calculations and distance estimation, increasing the number of sites does not appreciably increase computational cost. At the same time, as is illustrated by contrasting the Table 1 results for $t = 0.001$ and $n = 100$ with $n = 10,000$, the behavior of covariance approximations can be expected to improve.

Software for the methods is available at http://www.mathstat.dal.ca/ tsusko. Separate programs are used for GLS and WLSN, and for each of these there are separate programs available for estimation and for testing. All the programs use a PAML-style (Yang 1997, 2007) control file. The testing routines require a set of trees as input and output the trees with P values. The estimation routines use SPR searches starting from an NJ or user input tree and output all trees encountered with P values larger than a threshold. If that threshold is set as 95%, for instance, the output is the set of trees in the 95% confidence set, among those encountered. By running the routine with several starting points, one can approximate the 95% confidence set in cases where it is impossible to list all trees.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Bickel P.J., Doksum K.A. 2007. Mathematical statistics: basic ideas and selected topics. (NJ): Pearson.

Bulmer M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8:868–883.

Czarna A., Sanjuán R., González-Candelas F., Wróbel B. 2006. Topology testing of phylogenies using least-squares methods. BMC Evol. Biol. 6:105.

Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution. 21:550–570.

Felsenstein, J. 2005. Phylip: phylogenetic inference program. Version 3.6. Seattle (WA): University of Washington.

Fitch W.M., Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279–284.

Hasegawa M., Kishino H., Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In Munro H.N., editor. Mammalian protein metabolism. New York (NY): Academic Press. p. 121–123.

Lawson C.L., Hanson R.J. 1974. Solving least squares problems. Eaglewood Cliffs (NJ): Prentice-Hall.

Roux C.Z. 2009. The applicability of ordinary least squares to consistently short distances between taxa in phylogenetic tree construction and the normal distribution test consequences. Bull. Math. Biol. 71:771–780.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sanjuán R., Wróbel B. 2005. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance methods. Syst. Biol. 54:218–229.

Shi X., Gu H., Susko E., Field C. 2005. The comparison of confidence regions in phylogeny. Mol. Biol. Evol. 22:2285–2296.

Steel M. Huson D., Lockhart P.J. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. 49:225–232.

Susko E. 2003. Confidence regions and hypothesis tests for topologies using generalized least squares. Mol. Biol. Evol. 20:862–868.

Yang Z. 2007. PAML 4: a program for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 1997. PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.