Week 8: Testing trees, Bootstraps, jackknifes, gene frequencies

Genome 570

February, 2014
Normal distribution: curvature of log of height

\[
\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}
\]

Taking the logarithm of the height of the density curve of a normal distribution whose variance is \(\sigma^2\), we see that it is a quadratic curve whose curvature is \(-1/\sigma^2\).
The likelihood curve is nearly a normal distribution for large amounts of data. If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data.

If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of datawhen we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data.

If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data. If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data.

If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data.

If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
The likelihood curve is nearly a normal distribution for large amounts of data.

\[ \hat{\theta} \] from \( t(x) \), the "sufficient statistic"

If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.
Curvatures and covariances of ML estimates

ML estimates have covariances computable from curvatures of the expected log-likelihood:

$$\text{Var} \left[ \hat{\theta} \right] \simeq -1/ \left( \frac{d^2 E(\log(L))}{d\theta^2} \right)$$

The same is true when there are multiple parameters:

$$\text{Var} \left[ \hat{\theta} \right] \simeq \mathbf{V} \simeq -\mathbf{C}^{-1}$$

where

$$C_{ij} = E \left( \frac{\partial^2 \log(L)}{\partial \theta_i \partial \theta_j} \right)$$
With large amounts of data, asymptotically

When the true value of \( \theta \) is \( \theta_0 \),

\[
\frac{\hat{\theta} - \theta_0}{\sqrt{v}} \sim \mathcal{N}(0, 1)
\]

Since \( 1/v \) is the negative of the curvature of the log-likelihood:

\[
\ln L(\theta_0) = \ln L(\hat{\theta}) - \frac{1}{2} \left( \theta_0 - \hat{\theta} \right)^2
\]

so that twice the difference of log-likelihoods is the square of a normal:

\[
2 \left( \ln L(\hat{\theta}) - \ln L(\theta_0) \right) \sim \chi^2_1
\]
Corresponding results for multiple parameters

\[
\ln L(\theta_0) \simeq \ln L(\theta_0) - \frac{1}{2} (\theta_0 - \theta)^T C (\theta_0 - \theta)
\]

\[
(\theta - \theta_0)^T C (\theta - \theta_0) \sim \chi^2_p
\]

so that the log-likelihood difference is:

\[
2 \left( \ln L(\hat{\theta}) - \ln L(\theta_0) \right) \sim \chi^2_p
\]

When in the (true) null hypothesis \( \theta_0 \) we have \( q \) of the \( p \) parameters constrained:

\[
2 \left( \ln L(\hat{\theta}) - \ln L(\theta_0) \right) \sim \chi^2_q
\]
A log-likelihood curve

A Likelihood curve in one parameter
Its maximum likelihood estimate

A Likelihood curve in one parameter and the maximum likelihood estimate

\[ \text{Ln (Likelihood)} \]

length of a branch in the tree

maximum likelihood estimate (MLE)
The (approximate, asymptotic) confidence interval

A Likelihood curve in one parameter and the maximum likelihood estimate and confidence interval derived from it.

\[ \ln (\text{Likelihood}) \]

- 1/2 the value of a chi-square with 1 d.f. significant at 95%
- 95% confidence interval
- Maximum likelihood estimate (MLE)
Contours of a log-likelihood surface in two dimensions
Contours of a log-likelihood surface in two dimensions
Log-likelihood-based confidence set for two variables

- shaded area is the joint confidence interval

The height of this contour is less than at the peak by an amount equal to 1/2 the chi-square value with two degrees of freedom which is significant at 95% level.
Confidence interval for one variable

height of this contour is less than at the peak by an amount equal to 1/2 the chi-square value with one degree of freedom which is significant at 95% level.
Confidence interval for the other variable

height of this contour is less than at the peak by an amount equal to 1/2 the chi-square value with one degree of freedom which is significant at 95% level
Inferring the transition/transversion ratio for an F84 model with the 14-species primate mitochondria data set.
LRT of a molecular clock – how many parameters?

How does each equation constrain the branch lengths in the unrooted tree? What about the red equation?

Week 8: Testing trees, Bootstraps, jackknifes, gene frequencies – p.23/69
## Likelihood Ratio Test for a molecular clock

Using the 7-species mitochondrial DNA data set (the great apes plus Bovine and Mouse), we get with Ts/Tn = 30 and an F84 model:

<table>
<thead>
<tr>
<th>Tree</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>No clock</td>
<td>$-1372.77620$</td>
</tr>
<tr>
<td>Clock</td>
<td>$-1414.45053$</td>
</tr>
<tr>
<td>Difference</td>
<td>$41.67473$</td>
</tr>
</tbody>
</table>

Chi-square statistic: $2 \times 41.675 = 83.35$, with $n - 2 = 5$ degrees of freedom – highly significant.
Model selection using the LRT

Parameters

29

F84, T estimated

28

F81

F84, T=2

27

K2P, T estimated

26

Jukes–Cantor

K2P, T=2

25

The problem with using likelihood ratio tests is the multiplicity of tests and the multiple routes to the same hypotheses.
### The Akaike Information Criterion

Compare between hypotheses $-2 \ln L + 2p$ (the same as reducing the log-likelihood by the number of parameters)

<table>
<thead>
<tr>
<th>Model</th>
<th>$\ln L$</th>
<th>Number of parameters</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jukes-Cantor</td>
<td>$-3068.29186$</td>
<td>25</td>
<td>6186.58</td>
</tr>
<tr>
<td>K2P, $R = 2.0$</td>
<td>$-2953.15830$</td>
<td>25</td>
<td>5956.32</td>
</tr>
<tr>
<td>K2P, $\hat{R} = 1.889$</td>
<td>$-2952.94264$</td>
<td>26</td>
<td>5957.89</td>
</tr>
<tr>
<td>F81</td>
<td>$-2935.25430$</td>
<td>28</td>
<td>5926.51</td>
</tr>
<tr>
<td>F84, $R = 2.0$</td>
<td>$-2680.32982$</td>
<td>28</td>
<td>5416.66</td>
</tr>
<tr>
<td>F84, $\hat{R} = 28.95$</td>
<td>$-2616.3981$</td>
<td>29</td>
<td>5290.80</td>
</tr>
</tbody>
</table>
Can we test trees using the LRT?

If so, how many degrees of freedom for the comparison of the two peaks? These are three-species clocklike trees (shown here plotted in a “profile log-likelihood plot” plotting the highest likelihood for each value of the interior branch length).
The bootstrap

estimate of $\theta$  
(unknown) true value of $\theta$

empirical distribution of sample

150 data points

(unknown) true distribution

Bootstrap replicates  (each 150 draws)

Distribution of estimates of parameters

An example with mixed normal distributions. Draw from the empirical distribution 150 times if there are 150 data points. With replacement!
The bootstrap for phylogenies

Drawing columns of the data matrix, with replacement.
A partition defined by a branch in the first tree

How many times each partition of species is found:

- AE | BCDF
- ACE | BDF
- ACEF | BD  1
- AC | BDEF
- AEF | BCD
- ADEF | BC
- ABDF | EC
- ABCE | DF
Another partition from the first tree

Trees:

```
E   B
A   D
C   F

E   B
A   C
D   F
```

How many times each partition of species is found:

- AE | BCDF
- ACE | BDF 1
- ACEF | BD 1
- AC | BDEF
- AEF | BCD
- ADEF | BC
- ABDF | EC
- ABCE | DF
The third partition from that tree

Trees:

How many times each partition of species is found:

- **AE | BCDF**: 1
- **ACE | BDF**: 1
- **ACEF | BD**: 1
- **AC | BDEF**: 1
- **AEF | BCD**: 1
- **ADEF | BC**: 1
- **ABDF | EC**: 1
- **ABCE | DF**: 1
Partitions from the second tree

Trees:

- 

- 

- 

- 

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>BCDF</td>
</tr>
<tr>
<td>ACE</td>
<td>BDF</td>
</tr>
<tr>
<td>ACEF</td>
<td>BD</td>
</tr>
<tr>
<td>AC</td>
<td>BDEF</td>
</tr>
<tr>
<td>AEF</td>
<td>BCD</td>
</tr>
<tr>
<td>ADEF</td>
<td>BC</td>
</tr>
<tr>
<td>ABDF</td>
<td>EC</td>
</tr>
<tr>
<td>ABCE</td>
<td>DF</td>
</tr>
</tbody>
</table>
Partitions from the third tree

Trees:

How many times each partition of species is found:

- AE | BCDF  2
- ACE | BDF  2
- ACEF | BD  1
- AC | BDEF  1
- AEF | BCD  1
- ADEF | BC  1
- ABDF | EC  1
- ABCE | DF  1
Partitions from the fourth tree

Trees:

How many times each partition of species is found:

- **AE | BCDF**: 3
- **ACE | BDF**: 2
- **ACEF | BD**: 1
- **AC | BDEF**: 1
- **AEF | BCD**: 1
- **ADEF | BC**: 2
- **ABDF | EC**: 1
- **ABCE | DF**: 2
Partitions from the fifth tree

Trees:

```
  E
 /|
/  |
A---B
  |
  D
  |
  C
```

```
  E
 /|
/  |
C---F
  |
  D
```

How many times each partition of species is found:

- **AE | BCDF**: 3
- **ACE | BDF**: 3
- **ACEF | BD**: 1
- **AC | BDEF**: 1
- **AEF | BCD**: 1
- **ADEF | BC**: 2
- **ABDF | EC**: 1
- **ABCE | DF**: 3
The table of partitions from all trees

Trees:

- \( \text{AE} | \text{BCDF} \)
- \( \text{ACE} | \text{BDF} \)
- \( \text{ACEF} | \text{BD} \)
- \( \text{AC} | \text{BDEF} \)
- \( \text{AEF} | \text{BCD} \)
- \( \text{ADEF} | \text{BC} \)
- \( \text{ABDF} | \text{EC} \)
- \( \text{ABCE} | \text{DF} \)

How many times each partition of species is found:

- \( \text{AE} | \text{BCDF} \) 3
- \( \text{ACE} | \text{BDF} \) 3
- \( \text{ACEF} | \text{BD} \) 1
- \( \text{AC} | \text{BDEF} \) 1
- \( \text{AEF} | \text{BCD} \) 1
- \( \text{ADEF} | \text{BC} \) 2
- \( \text{ABDF} | \text{EC} \) 1
- \( \text{ABCE} | \text{DF} \) 3
The majority-rule consensus tree

Trees:

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>BCDF</td>
</tr>
<tr>
<td>ACE</td>
<td>BDF</td>
</tr>
<tr>
<td>ACEF</td>
<td>BD</td>
</tr>
<tr>
<td>AC</td>
<td>BDEF</td>
</tr>
<tr>
<td>AEF</td>
<td>BCD</td>
</tr>
<tr>
<td>ADEF</td>
<td>BC</td>
</tr>
<tr>
<td>ABDF</td>
<td>EC</td>
</tr>
<tr>
<td>ABCE</td>
<td>DF</td>
</tr>
</tbody>
</table>
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these “characters” that occur in more than 50% of the trees are all pairwise compatible.
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these “characters” that occur in more then 50% of the trees are all pairwise compatible.
- By the Pairwise Compatibility Theorem (remember that?) they must then be jointly compatible
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these “characters” that occur in more then 50% of the trees are all pairwise compatible.
- By the Pairwise Compatibility Theorem (remember that?) they must then be jointly compatible
- So there must be a tree that contains them all.
The MR tree with 14-species primate mtDNA data
Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a "multiple-tests" problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods
Other resampling methods

- Delete-half jackknife. Sample a random 50% of the sites, without replacement.
- Delete-1/e jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).
- Reweighting characters by choosing weights from an exponential distribution.
- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1
- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth
- (to correct for correlation among adjacent sites) (Künsch, 1989)
- Block-bootstrapping – sample \(\frac{n}{b}\) blocks of \(b\) adjacent sites.
With the delete-half jackknife
Bootstrap versus jackknife in a simple case

Exact computation of the effects of deletion fraction for the jackknife

(suppose 1 and 2 are conflicting groups)

We can compute for various n’s the probabilities of getting more evidence for group 1 than for group 2

A typical result is for $n_1 = 10$, $n_2 = 8$, $n = 100$ :

<table>
<thead>
<tr>
<th></th>
<th>Bootstrap</th>
<th>$\delta = 1/2$</th>
<th>$\delta = 1/e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Prob}(m_1 &gt; m_2)$</td>
<td>0.6384</td>
<td>0.5923</td>
<td>0.6441</td>
</tr>
<tr>
<td>$\text{Prob}(m_1 &gt; m_2)$</td>
<td>0.7230</td>
<td>0.7587</td>
<td>0.8040</td>
</tr>
<tr>
<td>$\text{Prob}(m_1 = m_2)$ + $\frac{1}{2} \text{Prob}(m_1 = m_2)$</td>
<td>0.6807</td>
<td>0.6755</td>
<td>0.7240</td>
</tr>
</tbody>
</table>
## Probability of a character being omitted from a bootstrap

\[
(1 - \frac{1}{N})^N
\]

<table>
<thead>
<tr>
<th>$N$</th>
<th>$(1 - 1/N)^N$</th>
<th>11</th>
<th>25</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.35049</td>
<td>0.36040</td>
<td>0.36603</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.35200</td>
<td>0.36166</td>
<td>0.36665</td>
</tr>
<tr>
<td>3</td>
<td>0.29630</td>
<td>0.35326</td>
<td>0.36256</td>
<td>0.36696</td>
</tr>
<tr>
<td>4</td>
<td>0.31641</td>
<td>0.35434</td>
<td>0.36323</td>
<td>0.36714</td>
</tr>
<tr>
<td>5</td>
<td>0.32768</td>
<td>0.35526</td>
<td>0.36375</td>
<td>0.36727</td>
</tr>
<tr>
<td>6</td>
<td>0.33490</td>
<td>0.35607</td>
<td>0.36417</td>
<td>0.36751</td>
</tr>
<tr>
<td>7</td>
<td>0.33992</td>
<td>0.35679</td>
<td>0.36479</td>
<td>0.36770</td>
</tr>
<tr>
<td>8</td>
<td>0.34361</td>
<td>0.35742</td>
<td>0.36524</td>
<td>$\infty$</td>
</tr>
<tr>
<td>9</td>
<td>0.34644</td>
<td>0.35798</td>
<td>0.36557</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.34868</td>
<td>0.35849</td>
<td>0.36583</td>
<td></td>
</tr>
</tbody>
</table>
A toy example to examine bias of $P$ values

Assuming a normal distribution, trying to infer whether the mean is above 0, when the mean is unknown and the variance known to be 1.
Bias in the P values

note that the true P is more extreme than the average of the P's

estimate of the "phylogeny"

topology II  0  topology I

the true mean
How much bias in the P values?
Bias in the P values with different priors

![Graph showing the probability of correct topology against P for expectation of μ for different values of $n \sigma^2 = 0.1, 1.0, 2.0$.](image)
The parametric bootstrap
The parametric bootstrap with the primates data

- Bovine
- Lemur
  - Tarsier
  - Squirrel Monkey
  - Mouse
  - Jp Macaque
    - Barbary Mac
    - Crab–Eating Mac
    - Rhesus Mac
  - Gorilla
  - Chimp
  - Human
  - Orang
  - Gibbon
Goldman’s test using simulation

(related to the "parametric bootstrap")

tree

no clock

log–likelihood

\[ T \]

clock

\[ T_c \]

simulating data sets

\[ 2(l - l_c) \]

estimating clocklike and nonclocklike trees from each data set ...

\[ 2(l - l_c) \]

\[ 2(l - l_c) \]

\[ 2(l - l_c) \]

\[ \cdots \]

\[ 2(l - l_c) \]
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
Brownian motion along a tree
Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances $v_3$ and $v_1$) so it is normally distributed with mean 0 and variance $v_3 + v_1$. 
Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances $v_3$ and $v_1$) so it is normally distributed with mean 0 and variance $v_3 + v_1$.

Similarly for tip 2 (variance is $v_3 + v_2$).
Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances $v_3$ and $v_1$) so it is normally distributed with mean 0 and variance $v_3 + v_1$.

Similarly for tip 2 (variance is $v_3 + v_2$).

They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.
Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances $v_3$ and $v_1$) so it is normally distributed with mean 0 and variance $v_3 + v_1$.

Similarly for tip 2 (variance is $v_3 + v_2$).

They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.

Variance is the expectation of the square (of deviation from the mean), and covariance is the expectation of the product of those deviations, for the two variables.
Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances $v_3$ and $v_1$) so it is normally distributed with mean 0 and variance $v_3 + v_1$.

Similarly for tip 2 (variance is $v_3 + v_2$).

They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.

Variance is the expectation of the square (of deviation from the mean), and covariance is the expectation of the product of those deviations, for the two variables.

In fact the covariance of the values at tip 1 and tip 2 is the variance of the shared term that is the same in both of them, so it is $v_3$. 
### Covariances of species on the tree

\[
\begin{pmatrix}
\nu_1 + \nu_8 + \nu_9 & \nu_8 + \nu_9 & \nu_9 & 0 & 0 & 0 & 0 \\
\nu_8 + \nu_9 & \nu_2 + \nu_8 + \nu_9 & \nu_9 & 0 & 0 & 0 & 0 \\
\nu_9 & \nu_9 & \nu_3 + \nu_9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \nu_4 + \nu_{12} & \nu_{12} & \nu_{12} & \nu_{12} \\
0 & 0 & 0 & \nu_{12} & \nu_{5} + \nu_{11} + \nu_{12} & \nu_{11} + \nu_{12} & \nu_{11} + \nu_{12} \\
0 & 0 & 0 & \nu_{12} & \nu_{11} + \nu_{12} & \nu_{6} + \nu_{10} + \nu_{11} + \nu_{12} & \nu_{10} + \nu_{11} + \nu_{12} \\
0 & 0 & 0 & \nu_{12} & \nu_{11} + \nu_{12} & \nu_{10} + \nu_{11} + \nu_{12} & \nu_{7} + \nu_{10} + \nu_{11} + \nu_{12}
\end{pmatrix}
\]
Covariances are of form

$$
\begin{bmatrix}
    a & b & c & 0 & 0 & 0 & 0 \\
    b & d & c & 0 & 0 & 0 & 0 \\
    c & c & e & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & f & g & g & g \\
    0 & 0 & 0 & g & h & i & i \\
    0 & 0 & 0 & g & i & j & k \\
    0 & 0 & 0 & g & i & k & l \\
\end{bmatrix}
$$
Likelihood under Brownian motion with two species

\[ f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) \]

\[ L = \prod_{i=1}^{p} \frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{v_1v_2}} \exp \left( -\frac{1}{2} \left[ \frac{(x_{1i} - x_{0i})^2}{v_1} + \frac{(x_{2i} - x_{0i})^2}{v_2} \right] \right) \]
What happens if we estimate means and branch lengths?

Do we get the right answer if we estimate for each coordinate (each character) the value at the root and the branch lengths \( v_1 \) and \( v_2 \)? Actually no.

Below, we will do this by finding values of these that maximize the likelihood, and show that the likelihood becomes infinite if either \( v_1 \) or \( v_2 \) approaches zero.

Even if we constrain there to be a clock, so \( v_1 = v_2 \) and look only at their sum \( v_1 + v_2 \) this turns out to be half as big as the truth, even with an infinite number of characters.

Why? The problem seems to be that we are estimating too many parameters. There is one parameter (the root value) for each character. So the ratio of data to parameters does not rise to infinity as we increase the number of parameters. In circumstances like this, likelihood methods can misbehave.
Minimizing for each character $i$

\[ Q = \frac{(x_{1i} - x_{0i})^2}{v_1} + \frac{(x_{2i} - x_{0i})^2}{v_2} \]

so:

\[ \frac{dQ}{dx_{0i}} = -2\frac{(x_{1i} - x_{0i})}{v_1} - 2\frac{(x_{2i} - x_{0i})}{v_2} = 0 \]

and then:

\[ \hat{x}_{0i} = \frac{\frac{1}{v_1}x_{1i} + \frac{1}{v_2}x_{2i}}{\frac{1}{v_1} + \frac{1}{v_2}} \]

So that we have a maximum likelihood estimate of the starting value $x_{0i}$ for each character.

The result is that

\[ Q = \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2} \]
Likelihood after estimating initial coordinates

Substituting in our estimates of $x_{0i}$, we end up with

$$L = \frac{1}{(2\pi)^p (v_1 v_2)^{\frac{1}{2}p}} \exp \left( -\frac{1}{2} \sum_{i=1}^{p} \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2} \right)$$

and this finally turns into:

$$\ln L = -p \ln(2\pi) - \frac{1}{2} p \ln (v_1 v_2) - \frac{1}{2} \sum_{i=1}^{p} \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2}$$

This actually goes to infinity as either $v_1$ or $v_2$ goes to zero! This is related to the problem that Edwards and Cavalli-Sforza had with their maximum likelihood method in 1964.
If there is a clock ...

If instead we constrain \( v_1 = v_2 \) because assume a clock:

\[
\ln L = K' - p \ln(v_1 + v_2) - \frac{1}{2} \frac{D^2}{(v_1 + v_2)}
\]

which leads to

\[
\hat{v}_1 = \hat{v}_2 = \frac{D^2}{4p}
\]

(which is half as big as it should be!)

The number of parameters being estimated is \( p + 1 \), which rises as we consider more characters. The fact that the ratio of data to parameters does not rise without limit is the reason why likelihood misbehaves in this case.
The difference between ML and REML

Information we use for ML inference:

species 2  

 species 1  species 4  species 3

1.0  2.0  3.0  4.0

Information we use for REML inference:

species 2  

 species 1  species 4  species 3

1.0+x  2.0+x  3.0+x  4.0+x

Does it matter that we don’t know $x$? It makes it unnecessary to estimate the starting value $x_0$, and that eliminates $p$ parameters. It means that the ratio of data to parameters does then rise as we add characters.
Using only differences between populations (REML)

We assume that we have observed only the differences $x_{1i} - x_{2i}$, and not the actual locations on the phenotype scale. Then

$$ L = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi \sqrt{v_1 + v_2}}} \exp \left( -\frac{1}{2} \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2} \right) $$

$$ \ln L = K - \frac{p}{2} \ln (v_1 + v_2) + \frac{1}{2 (v_1 + v_2)} \sum_{i=1}^{n} (x_{i1} - x_{i2})^2 $$
Likelihood with two species using REML

\[ \ln L = K - \frac{p}{2} \ln (v_1 + v_2) + \frac{D^2}{2(v_1 + v_2)} \]

\[ \ln L = K - \frac{p}{2} \ln (v_T) + \frac{D^2}{2v_T} \]

\[ \hat{v}_T = \frac{D^2}{p} \]

The number of parameters being estimated is 1 (it is the sum \( v_1 + v_2 \)). The number of parameters does not rise as we consider more characters.
“Pruning” a tree in the Brownian motion case

δ = \( \frac{v_1 v_2}{v_1 + v_2} \)

\( x_{12} = \frac{v_2 x_1 + v_1 x_2}{v_1 + v_2} \)

The likelihood for the tree is the product of the likelihoods for these two trees. By repeatedly applying this we can decompose the tree into \( n - 1 \) independent two-species trees. Getting their likelihoods is easy.