MULTISCALE BOOTSTRAP USING SCALEBOOT PACKAGE

HIDETOSHI SHIMODAIRA

1. Introduction

scaleboot is an add-on package for R. It is for calculating approximately unbiased (AU) p-values for a general problem from a set of multiscale bootstrap probabilities (BPs). Scaling is equivalent to changing the sample size of a data in bootstrap resampling. We compute BPs at several scales, from which a very accurate p-value is calculated (Shimodaira 2002). This multiscale bootstrap method has been implemented in CONSEL (Shimodaira and Hasegawa 2001) for phylogenetic inference and as the R add-on package pvclust (Suzuki and Shimodaira 2006) for hierarchical clustering. The point of the scaleboot package is to calculate an improved version of the AU p-value that is justified even for hypotheses with nonsmooth boundaries (Shimodaira 2008).

The basic usage of this package is illustrated in a simple example below. Then real applications in hierarchical clustering and phylogenetic inference are shown later.

For the use of scaleboot, Shimodaira (2008) may be referenced.

New update in 2018: Selective inference is now implemented. The selective inference version of the approximately unbiased p-value is denoted as SI in scaleboot and pvclust. The original approximately unbiased p-value, which is denoted as AU, is not designed for selective inference. The bootstrap probability (BP) is also not designed for selective inference. The theory and example of selective inference is described in Terada and Shimodaira (2017).

2. Install

scaleboot is easily installed from CRAN online. Windows users can install the package by choosing “scaleboot” from the pull-down menu. Otherwise, run R on your computer and type

```R
> install.packages("scaleboot")
```

You can also download the package file from the URL below and install it manually.

http://stat.sys.i.kyoto-u.ac.jp/prog/scaleboot/

3. Simple Example

3.1. Simulation Data. We first generate a simulation data.

```R
> simdata <- function(n,y,sd) {
+   m <- length(y)
+   x <- matrix(rnorm(m*n,0,sd),m,n)
+   t(x + (y - apply(x,1,mean)))
+ } > X <- simdata(100,c(0,1,1,1,1,1,1,1,1,1),10)
> round(X[1:3,],3)
```

This document is a part of the scaleboot package (Version 0.4-0 or newer). The source file is usesb.Rnw. I thank Paul A. Sheridan for his comments to improve the earlier version of the manuscript.
The matrix $X = (x_{ij})$ above is of size $n \times m$ with $n = 100$, $m = 10$. We consider $X$ as a data of sample size $n$, and rows $x_i = (x_{i1}, \ldots, x_{im})$, $i = 1, \ldots, n$, are observations of a random vector of $m$ dimensions.

3.2. Null Hypothesis. Let $\mu$ be the unknown population mean of the row vectors. An estimate of $\mu$ is the sample average of the rows defined as $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Let $f(\mu)$ be a 0/1-valued (or false/true valued) function of $\mu$. The null hypothesis we are going to test is represented as $f(\mu) = 1$. For example, $f(\mu) = 1$ if $\mu_1$ is the largest among $\mu_1, \ldots, \mu_m$, and $f(\mu) = 0$ otherwise. This $f(\mu)$ is implemented as `mc1(mu)` below.

```r
> mc1 <- function(x) all(x[1] >= x[-1])
> mc1(y)
[1] FALSE
```

Although $f(y) = 0$ gives a rough idea whether $f(\mu) = 1$, we want to calculate a real number ranging between 0 and 1 which indicates the possibility of $f(\mu) = 1$. This is what `scaleboot` calculates as $p$-values.

3.3. Bootstrap Probabilities. A naive way to calculate a $p$-value is by bootstrap resampling. Let $X^* = (x^*_{ij})$ be a bootstrap sample of $X$; each row $x^*_i$ is obtained by resampling with replacement from the rows $x_1, \ldots, x_n$. Let $n'$ be the size of the resampling so that $X^*$ is a matrix of size $n' \times m$. The bootstrap replicate is $y^* = \bar{x}^* = \frac{1}{n'} \sum_{i=1}^{n'} x^*_i$. The following code generates an $X^*$ with $n' = n$, and calculates $f(y^*)$. The resampling is made via a weight vector $w$; $w_i$ is the number of times that $x_i$ is resampled in $X^*$.

```r
> countw <- function(x,w,fn) {
+   y <- apply(w*x,2,sum)/sum(w)
+   fn(y)
+ }
> w <- as.vector(rmultinom(1,100,rep(1,100)))
> countw(X,w,mc1)
[1] FALSE
```

Let $B$ be the number of bootstrap samples we will generate, and $y^*_1, \ldots, y^*_B$ be the bootstrap replicates. Typically, $B = 10,000$. The BP is computed as $\sum_{i=1}^{B} f(y^*_i)/B$, where the ordinary BP uses $n' = n$. Since first introduced by Felsenstein (1985), it has been widely used as a $p$-value, but the bias is in fact rather large.

3.4. $P$-value Calculation. `scaleboot` calculates corrected $p$-values for improving BPs. First load the package by

```r
> library(scaleboot)
```
Below, \( sa \) specifies the scales, and \( nb \) specifies \( B \) for each scale, so that \( 10,000 \times 13 = 130,000 \) bootstrap samples are generated internally. It takes a few minutes on a pc.

\[
\begin{align*}
&> \text{sa} <- 9^\text{seq(-1,1,length=13)} \\
&> \text{nb} <- 10000 \\
&> \text{X.sb} <- \text{scaleboot(X, nb, sa, countw, mc1)}
\end{align*}
\]

The main result (for \( k = 3 \)) is shown by

\[
> \text{summary(X.sb)} \quad \# k = 3 \quad \text{(default)}
\]

Raw Bootstrap Probability (scale=1) : 0.93 (0.10)

Hypothesis: null

Corrected P-values for Models (percent,Frequentist):

\[
\begin{array}{cccccc}
\text{Model} & \text{k.3} & \text{sk.3} & \text{beta0} & \text{beta1} & \text{aic} & \text{weight} \\
\hline
\text{sing.3} & 37.68 (2.94) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\text{poly.3} & 17.23 (1.26) & 22.84 (1.47) & 1.67 (0.03) & 0.59 (0.02) & 15.02 \\
\text{poly.2} & 6.34 (0.34) & 9.70 (0.49) & 1.92 (0.02) & 0.39 (0.01) & 177.96 \\
\text{poly.1} & 0.07 (0.00) & 0.13 (0.01) & 3.21 (0.02) & 0.00 (0.00) & 3534.85 \\
\end{array}
\]

Best Model: sing.3

Corrected P-values by the Best Model and by Akaike Weights Averaging:

\[
\begin{array}{cccccc}
\text{Model} & \text{k.3} & \text{sk.3} & \text{beta0} & \text{beta1} & \text{aic} & \text{weight} \\
\hline
\text{best} & 37.68 (2.94) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\text{average} & 37.68 (2.94) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\end{array}
\]

The selective inference \( p \)-values (SI) are denoted as \( \text{sk.3} \), and the non-selective AU \( p \)-values are denoted as \( \text{k.3} \). We can also see additional results by specifying \( k \),

\[
> \text{summary(X.sb, k=1:3)} \quad \# k = 1, 2, 3
\]

Raw Bootstrap Probability (scale=1) : 0.93 (0.10)

Hypothesis: null

Corrected P-values for Models (percent,Frequentist):

\[
\begin{array}{cccccc}
\text{Model} & \text{k.1} & \text{k.2} & \text{k.3} & \text{sk.1} & \text{sk.2} & \text{sk.3} & \text{beta0} & \text{beta1} & \text{aic} & \text{weight} \\
\hline
\text{sing.3} & 1.04 (0.04) & 16.89 (0.82) & 37.68 (2.94) & 2.08 (0.09) & 22.50 (0.97) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\text{poly.3} & 1.20 (0.05) & 14.18 (0.93) & 17.23 (1.26) & 2.39 (0.10) & 19.60 (1.15) & 22.84 (1.47) & 1.67 (0.03) & 0.59 (0.02) & 15.02 \\
\text{poly.2} & 1.03 (0.05) & 6.34 (0.34) & 6.34 (0.34) & 2.06 (0.09) & 9.70 (0.49) & 9.70 (0.49) & 1.92 (0.02) & 0.39 (0.01) & 177.96 \\
\text{poly.1} & 0.07 (0.00) & 0.07 (0.00) & 0.07 (0.00) & 0.13 (0.01) & 0.13 (0.01) & 0.13 (0.01) & 3.21 (0.02) & 0.00 (0.00) & 3534.85 \\
\end{array}
\]

Best Model: sing.3

Corrected P-values by the Best Model and by Akaike Weights Averaging:

\[
\begin{array}{cccccc}
\text{Model} & \text{k.1} & \text{k.2} & \text{k.3} & \text{sk.1} & \text{sk.2} & \text{sk.3} & \text{beta0} & \text{beta1} & \text{aic} & \text{weight} \\
\hline
\text{best} & 1.04 (0.04) & 16.89 (0.82) & 37.68 (2.94) & 2.08 (0.09) & 22.50 (0.97) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\text{average} & 1.04 (0.04) & 16.89 (0.82) & 37.68 (2.94) & 2.08 (0.09) & 22.50 (0.97) & 42.97 (2.83) & 1.64 (0.02) & 0.68 (0.01) & -12.35 & 100.00 \\
\end{array}
\]

A class of AU \( p \)-values indexed by \( k = 1, 2, 3 \), are calculated, and they are labelled as \( k.1, k.2, \) and \( k.3 \). The \( p \)-values are shown in percent, and the standard errors are given in parentheses. We should look at the row of \text{average} (the bottom line), and we can ignore the other rows. \( p_1 \approx 1\% \) corresponds to the ordinary \( p \)-value, and \( p_2 \approx 18\% \) corresponds to the AU \( p \)-value of Shimodaira (2002). What we recommend to use here is \( p_3 \approx 40\% \); this is the AU \( p \)-value of Shimodaira (2008). For this particular example, the common practice for calculating a \( p \)-value is to use the multiple comparisons method. If it is applied to \( y \), the \( p \)-value is \( p = 66\% \), which is rather close to \( p_3 \) in our example, whereas \( p_1 \) is obviously too small.

Internally, as explained in the next section, several models are fitted to the observed bootstrap probabilities. They are sorted in increasing order of AIC, and the best model is \text{sing.3}. The row of \text{sing.3} is duplicated in the row \text{best}, two lines from the bottom. The Akaike weights \( \propto \exp(-\text{AIC}/2) \) are also computed for
models and the p-values are averaged using the weights so that we get the row of average, the bottom line. Typically, the two bottom lines are almost identical, and average is regarded as a smoothed version of best for small changes in data. I recommend to use average instead of best.

3.5. **Internal Steps.** We consider the following three steps (i)-(iii). Internally, the `scaleboot` function (i) performs the multiscale bootstrap, and (ii) estimates coefficients for candidate models. Then the `summary` method (iii) calculates the corrected p-values. These steps are explained below.

The results of steps (i) and (ii) are shown here.

```
> X.sb
Multiscale Bootstrap Probabilities (percent):
1 2 3 4 5 6 7 8 9 10 11 12 13
0.00 0.01 0.05 0.12 0.29 0.68 0.93 1.57 2.21 2.77 3.40 3.94 4.69

Numbers of Bootstrap Replicates:
1 2 3 4 5 6 7 8 9 10 11 12 13
10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000

Scales (Sigma Squared):
1 2 3 4 5 6 7 8 9 10 11 12 13
0.1111 0.1603 0.2309 0.3333 0.4808 0.6944 1.0000 1.4490 2.0833 3.0303 4.3484 6.2500 9.0909

Coefficients:
<table>
<thead>
<tr>
<th></th>
<th>beta0</th>
<th>beta1</th>
<th>beta2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sing.3</td>
<td>1.1518 (0.1347)</td>
<td>1.1601 (0.1401)</td>
<td>0.8332 (0.1221)</td>
</tr>
<tr>
<td>poly.3</td>
<td>1.6337 (0.0284)</td>
<td>0.6569 (0.0210)</td>
<td>-0.0318 (0.0024)</td>
</tr>
<tr>
<td>poly.2</td>
<td>1.9212 (0.0219)</td>
<td>0.3943 (0.0069)</td>
<td></td>
</tr>
<tr>
<td>poly.1</td>
<td>3.2056 (0.0182)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model Fitting:

<table>
<thead>
<tr>
<th></th>
<th>rss</th>
<th>df</th>
<th>pfit</th>
<th>aic</th>
</tr>
</thead>
<tbody>
<tr>
<td>sing.3</td>
<td>7.65</td>
<td>10</td>
<td>0.6629</td>
<td>-12.35</td>
</tr>
<tr>
<td>poly.3</td>
<td>35.02</td>
<td>10</td>
<td>0.0001</td>
<td>15.02</td>
</tr>
<tr>
<td>poly.2</td>
<td>199.96</td>
<td>11</td>
<td>0.0000</td>
<td>177.96</td>
</tr>
<tr>
<td>poly.1</td>
<td>3558.85</td>
<td>12</td>
<td>0.0000</td>
<td>3534.85</td>
</tr>
</tbody>
</table>

Best Model: sing.3
```

The results of (i) are the BPs for the 13 scales shown at first. Let $\alpha_{\sigma^2}$ denote the BP at scale $\sigma^2$. Each BP is calculated from 10,000 bootstrap samples of size $n'$ as the frequency of observing $f(y^*) = 1$. In `scaleboot`, $n'$ is `round(n/sa[i])`, for $i = 1, \ldots, 13$. Then, the scale is recalculated as $\sigma^2 = n/n'$ for taking account of the discreteness.

Step (ii) is performed by the `sbfit` function called internally from the `scaleboot` function for fitting parametric models to observed $\alpha_{\sigma^2}$’s. By default, four models are considered as candidates: `poly.1`, `poly.2`, `poly.3`, and `sing.3`. Each of these models is denoted as $\psi(\sigma^2|\beta)$. Let $z_{\sigma^2} = \Phi^{-1}(1 - \alpha_{\sigma^2})$ be the bootstrap $z$-value at scale $\sigma^2$, where $\Phi^{-1}(p) = \text{qnorm}(p)$. We work on $\sigma z_{\sigma^2}(y)$, which may be called a normalized bootstrap $z$-value. Considering $\sigma z_{\sigma^2}$ as a function of $\sigma^2$, the coefficient vector $\beta$ is estimated by fitting $\sigma z_{\sigma^2} = \psi(\sigma^2|\beta)$. Let $\hat{\beta}$ denote the estimated
The goodness of fit is measured by the difference of AIC values between the specified model and an unconstrained binomial model:

$$AIC = (-2\hat{\ell} + 2k) - (-2\ell + 2S),$$

where $\ell = \sum_{i=1}^{S}(C_i \log(C_i/B_i) + (B_i - C_i) \log(1 - C_i/B_i))$. 

The details of model fitting are as follows. Let $B_i$ and $C_i$ be the number of replicates and the observed number of times that $f(y^*) = 1$, respectively, for the bootstrap resampling of scale $\sigma_i^2$, $i = 1, \ldots, S$. Since each $C_i$ is binomially distributed, the log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{S} \left\{ C_i \log \Phi(-\psi(\sigma_i^2|\beta)/\sigma_i) + (B_i - C_i) \log \Phi(\psi(\sigma_i^2|\beta)/\sigma_i) \right\},$$

where $\Phi(q) = \text{pnorm}(q)$. The estimate $\hat{\beta}$ is obtained by maximizing $\ell(\beta)$ numerically.
Figure 2. Model Fitting ($x = \log \sigma, y = \alpha \sigma^2$)

Step (iii) is performed by the the summary method as already mentioned. The first line shows the “raw” BP $\alpha_1$ (the BP obtained from the ordinary bootstrap resampling). The main results are the corrected $p$-values, which follow next. For each model, we calculate $q_k$, $k = 1, 2, 3$, by

$$q_k = \sum_{j=0}^{k-1} \left( -1 - \sigma_0^2 \right)^j \frac{j!}{j!} \frac{\partial^j \psi(\sigma^2|\hat{\beta})}{\partial (\sigma^2)^j} \bigg|_{\sigma_0^2}.$$  

Then the corrected $p$-values are calculated by $p_k = 1 - \Phi(q_k)$. By default $\sigma_0^2 = 1$. The calculation of $q_k$ is interpreted as extrapolation of $\sigma z_\sigma^2$ to $\sigma^2 = -1$ by using the first $k$ terms of the Taylor series. According to the theory of Shimodaira (2008), the unbiased $p$-value is, if it exists, obtained by taking the limit $k \to \infty$. The extrapolated curves are shown (Fig. 3) by

```r
> plot(summary(X.sb),legend="topleft")
```

4. Hierarchical Clustering

4.1. Pvclust Package. The scaleboot package includes an interface for the pvclust package (Suzuki and Shimodaira 2006). We use pvclust to calculate multiscale BPs for clusters by bootstrapping hierarchical clustering, from which we calculate an improved version of AU $p$-values using scaleboot. See `help(lung73)` for further details of the following example.

4.2. Using Pvclust. This example uses the lung dataset (Garber et al. 2001) included in pvclust. It is a DNA microarray data of 73 lung tissues (arrays) with 916 observations of genes. To draw dendrograms in terms of the arrays, we
resample genes in our analysis; this may be interpreted as assessing the uncertainty due to the variability of genes. The function \texttt{pvclust} first obtains a dendrogram by a hierarchical clustering method, and then calculates the multiscale BPs for each cluster of the dendrogram.

```r
> library(pvclust)
> data(lung)
> sa <- 9^seq(-1,1,length=13)
> nb <- 10000
> lung73.pvclust <- pvclust(lung,r=1/sa,nboot=nb)
```

The above code may take a day, so it would be a good idea to run with \texttt{nb=1000} so that it would run 10 times faster. However, \texttt{nb=1000} should be used just for checking the program, and \texttt{nb=10,000} (at least) is recommended for publishing the results.

4.3. Model Fitting. We next apply the \texttt{sbfit} function of \texttt{scaleboot} to the multiscale BPs. For each cluster of the dendrogram, parametric models are fitted to the BPs.

```r
> library(scaleboot)
> lung73.sb <- sbfit(lung73.pvclust)
```

4.4. Lung73 Dataset. The results of the previous two sections (\texttt{lung73.pvclust} and \texttt{lung73.sb}) are in fact stored in the \texttt{lung73} dataset of \texttt{scaleboot}. For users who want to try the examples, just type as follows.

```r
> library(scaleboot)
> data(lung73)
```
Figure 4. Dendrogram of lung73 dataset (k = 3)

We have used a cluster computer of 40 cpus for parallel computing using the snow package. The following code may run in under an hour.

```r
> library(snow)
> cl <- makeCluster(40)
> library(pvclust)
> data(lung)
> sa <- 9^seq(-1,1,length=13)
> nb <- 10000
> lung73.pvclust <- parPvclust(cl,lung,r=1/sa,nboot=nb)
> library(scaleboot)
> lung73.sb <- sbfit(lung73.pvclust,cluster=cl)
```

4.5. $P$-value Calculation. To calculate AU $p$-values ($p_{AU}$) from `lung73.sb` and write them back to `lung73.pvclust`, we do

```r
> lung73.k3 <- sbpvclust(lung73.pvclust,lung73.sb)
```

To see the results, we simply plot the dendrogram (Fig. 4) by

```r
> library(pvclust)
> plot(lung73.k3, cex=0.5, cex.pv=0.7)
> pvrect(lung73.k3, pv="si") # find clusters with $p>0.95$. Now use SI instead of AU.
```

There three types of $p$-values are printed at each edge of the cluster. SI is the selective inference version of approximately unbiased $p$-value, which is newly introduced in Terada and Shimodaira (2017). AU is the non-selective inference version of approximately unbiased $p$-value, which has been used since Shimodaira (2002). BP is the original bootstrap probability. SI, AU, BP are denoted as $sk.3$, $k.3$, $bp$.
**Figure 5.** Model fitting for node 67

k.1, respectively, in tables generated by scaleboot for \( k = 3 \). When you know the cluster before looking at the data (such as cluster id=31, the left most cluster in Fig. 4; it consists of normal cell and fetal cell, so no lung cancer at all), you may use AU. However, you should use SI for most of the clusters when they are identified by looking at the tree.

To calculate \( p_2 \) instead of \( p_3 \), specify \( k=2 \),
\[
> \text{lung73.k2 <- sbpvclust(lung73.pvclust,lung73.sb, k=2)}
\]

**4.6. Diagnostics of Fitting.** The fitted curves are drawn by the plot method. For node 67, say, a plot with legend is obtained (Fig. 5) by
\[
> \text{plot(lung73.sb[[67]],legend="topleft")}
\]
All the calculated \( p \)-values for node 67 are given by
\[
> \text{summary(lung73.sb[[67]])}
\]
Raw Bootstrap Probability (scale=1) : 3.78 (0.19)

Hypothesis: null

**Corrected P-values for Models (percent,Frequentist):**

<table>
<thead>
<tr>
<th>Model</th>
<th>percent</th>
<th>Sk.3</th>
<th>beta0</th>
<th>beta1</th>
<th>aic</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>sing.3</td>
<td>95.00</td>
<td>(0.18)</td>
<td>97.45</td>
<td>(0.09)</td>
<td>0.53</td>
<td>(0.01)</td>
</tr>
<tr>
<td>poly.3</td>
<td>92.75</td>
<td>(0.28)</td>
<td>97.09</td>
<td>(0.14)</td>
<td>0.33</td>
<td>(0.00)</td>
</tr>
<tr>
<td>poly.2</td>
<td>83.64</td>
<td>(0.32)</td>
<td>92.78</td>
<td>(0.19)</td>
<td>0.31</td>
<td>(0.00)</td>
</tr>
<tr>
<td>poly.1</td>
<td>18.53</td>
<td>(0.10)</td>
<td>37.05</td>
<td>(0.21)</td>
<td>0.90</td>
<td>(0.00)</td>
</tr>
</tbody>
</table>
Figure 6. Extrapolation for node 67

Best Model: sing.3

Corrected $P$-values by the Best Model and by Akaike Weights Averaging:

<table>
<thead>
<tr>
<th></th>
<th>k.3</th>
<th>sk.3</th>
<th>beta0</th>
<th>beta1</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>95.00 (0.18)</td>
<td>97.45 (0.09)</td>
<td>0.53 (0.01)</td>
<td>1.29 (0.01)</td>
</tr>
<tr>
<td>average</td>
<td>95.00 (0.18)</td>
<td>97.45 (0.09)</td>
<td>0.53 (0.01)</td>
<td>1.29 (0.01)</td>
</tr>
</tbody>
</table>

The extrapolation using the best models (averaged by the Akaike weights) is shown (Fig. 6) by

> plot(summary(lung73.sb[[67]]),legend="topleft")

For a set of nodes, $p$-values are given by

> summary(lung73.sb[c(62,67,69,71)])

Corrected $P$-values by Akaike Weights Averaging (percent,Frequentist):

<table>
<thead>
<tr>
<th></th>
<th>k.3</th>
<th>sk.3</th>
<th>beta0</th>
<th>beta1</th>
<th>hypothesis</th>
<th>model</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>95.81 (0.20)</td>
<td>98.78 (0.07)</td>
<td>96.96 (0.16)</td>
<td>-2.00 (0.02)</td>
<td>0.25 (0.01)</td>
<td>alternative</td>
<td>poly.2</td>
</tr>
<tr>
<td>67</td>
<td>3.78 (0.19)</td>
<td>96.00 (0.18)</td>
<td>97.45 (0.09)</td>
<td>0.53 (0.01)</td>
<td>1.29 (0.01)</td>
<td>null</td>
<td>sing.3</td>
</tr>
<tr>
<td>69</td>
<td>30.51 (0.46)</td>
<td>75.04 (0.21)</td>
<td>9.00 (0.34)</td>
<td>-0.08 (0.00)</td>
<td>0.60 (0.00)</td>
<td>alternative</td>
<td>poly.2</td>
</tr>
<tr>
<td>71</td>
<td>25.63 (0.44)</td>
<td>86.02 (0.27)</td>
<td>26.18 (0.47)</td>
<td>-0.18 (0.00)</td>
<td>0.83 (0.01)</td>
<td>alternative</td>
<td>poly.3</td>
</tr>
</tbody>
</table>

Also plots are shown (Fig. 7) by

> plot(lung73.sb[c(62,67,69,71)])

5. Phylogenetic Inference

5.1. CONSEL Software. scaleboot has a front end for phylogenetic inference, and it may eventually replace the CONSEL software (Shimodaira and Hasegawa 2001)
for testing phylogenetic trees. Currently, scaleboot does not have a method for converting files obtained from other commonly used phylogenetic software packages, and so we must use CONSEL for this purpose before applying scaleboot to calculate an improved version of AU $p$-values for trees and edges. See help(mam15) for further details of the following example.

5.2. Mammal Dataset. We work on an example of phylogenetic analysis of six mammal species: Homo sapiens (human), Phoca vitulina (harbor seal), Bos taurus (cow), Oryctolagus cuniculus (rabbit), Mus musculus (mouse), Didelphis virginiana (opossum). The dataset was originally used in Shimodaira and Hasegawa (1999).

For Unix users, download mam15-files.tgz, and for Windows users download mam15-files.zip. The details of dataset files are as follows. mam15.aa: amino acid sequences ($n = 3414$) of mtDNA for the six mammals. mam15.ass: association vectors for edges and trees. mam15.lnf: site-wise log-likelihood values (output from PAML). mam15.log: detailed information for the associations. mam15.mt: site-wise log-likelihood values (output from seqmt). mam15.tpl: 15 tree topologies.

5.3. Likelihood Calculation of Trees. The main body of the dataset is the amino acid sequences (mam15.aa). We consider $m = 15$ tree topologies of the six mammals (mam15.tpl):

```
(('Homsa',('Phovi','Bosta')),'Orycu',('Musmu','Didvi')); t1
('Homsa','Orycu',(('Phovi','Bosta'),('Musmu','Didvi'))); t2
('Homsa',(('Phovi','Bosta'),'Orycu'),('Musmu','Didvi')); t3
('Homsa','Orycu','Musmu',(('Phovi','Bosta'),'Didvi')); t4
(('Homsa',('Phovi','Bosta')),'Orycu','Musmu','Didvi'); t5
```
The maximum likelihood estimates for these trees are calculated by PAML (Yang 1997). Let $x_{ij}$ be the site-wise log-likelihood for sites $i = 1, \ldots, n$, and trees $j = 1, \ldots, m$. The log-likelihood of tree-$j$ is $\sum_{i=1}^{n} x_{ij}$. A large $n$ justifies the central limit theorem for $y = \bar{x}$, and allows us to resample $x_{ij}$ directly without recalculation of the maximum likelihood estimates. The matrix $X = (x_{ij})$ is produced by PAML and stored in mam15.lnf. It is converted by CONSEL to a simpler format and stored in mam15.mt. The command is

```
seqmt --paml mam15.lnf
```

5.4. $P$-value Calculation for Trees. The AU $p$-values for trees are calculated simply by

```
> library(scaleboot)
> mam15.mt <- read.mt("mam15.mt")
> mam15.trees <- relltest(mam15.mt)
> summary(mam15.trees)
```

The relltest function above may take a half hour. The next section can be skipped if only tree selection is of interest.

5.5. $P$-value Calculation for Clusters. We can also calculate AU $p$-values for clusters (edges) of trees. We have to know, for each cluster, in which of the 15 trees it is included. The file mam15.ass has this information, which was generated using CONSEL by the command

```
treeass --outgroup 6 mam15.tpl > mam15.log
```

It also produces mam15.log for human readable information. A part of mam15.log is as follows.

```
# leaves: 6
6
 1 Homsa
 2 Phovi
 3 Bosta
 4 Orycu
 5 Musmu
 6 Didvi

# base edges: 10
10 6

123456
 1 +++--- ;
 2 ++++- ;
 3 ---++- ;
 4 -+++-- ;
 5 --+-++ ;
```
The clusters (edges) defined above are named e1,...,e10. For example, e1 = +++-- = (Homsa, Phovi, Bosta).

The AU-p-values for clusters as well as trees are calculated simply by

```r
> library(scaleboot)
> mam15.mt <- read.mt("mam15.mt")
> mam15.ass <- read.ass("mam15.ass")
> mam15.relltest <- relltest(mam15.mt,ass=mam15.ass)
> summary(mam15.relltest)
```

5.6. Mam15 Dataset. The results of the previous sections (mam15.mt, mam15.ass, and mam15.relltest) are in fact stored in mam15 dataset of scaleboot. For users who want to try the examples, just type as follows.

```r
> library(scaleboot)
> data(mam15)
```

The results for trees are extracted by

```r
> mam15.trees <- mam15.relltest[1:15]
```

We have used a cluster computer of 40 cpus for parallel computing using the snow package. The following code may take only 10 minutes, although we have used the number of resamples 10 times larger than the default value.

```r
> library(snow)
> cl <- makeCluster(40)
> library(scaleboot)
> mam15.mt <- read.mt("mam15.mt")
> mam15.ass <- read.ass("mam15.ass")
> mam15.relltest <- relltest(mam15.mt,nb=100000,ass=mam15.ass)
```

5.7. Interpreting the Results. First we sort the results in increasing order of log-likelihood difference,

```r
> stat <- attr(mam15.trees,"stat")
> o <- order(stat)
> mam15.trees <- mam15.trees[o]
> summary(mam15.trees, k=1:3)
```

Corrected P-values by Akaike Weights Averaging (percent,Frequentist):

<table>
<thead>
<tr>
<th></th>
<th>k.1 (%)</th>
<th>k.2 (%)</th>
<th>k.3 (%)</th>
<th>sk.1 (%)</th>
<th>sk.2 (%)</th>
<th>beta0 (%)</th>
<th>beta1 (%)</th>
<th>hypothesis</th>
<th>model</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>57.49 (0.16)</td>
<td>56.10 (0.04)</td>
<td>74.55 (0.05)</td>
<td>74.55 (0.06)</td>
<td>36.37 (0.09)</td>
<td>36.37 (0.09)</td>
<td>-0.41 (0.00)</td>
<td>0.25 (0.00)</td>
<td>alternative</td>
<td>poly.2</td>
</tr>
<tr>
<td>t2</td>
<td>3.67 (0.06)</td>
<td>3.65 (0.03)</td>
<td>13.10 (0.19)</td>
<td>16.60 (0.46)</td>
<td>7.31 (0.05)</td>
<td>20.75 (0.26)</td>
<td>24.66 (0.53)</td>
<td>1.46 (0.01)</td>
<td>0.36 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t3</td>
<td>3.21 (0.06)</td>
<td>3.16 (0.02)</td>
<td>12.44 (0.19)</td>
<td>13.60 (0.27)</td>
<td>6.32 (0.05)</td>
<td>19.51 (0.35)</td>
<td>20.62 (0.34)</td>
<td>1.51 (0.01)</td>
<td>0.35 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t4</td>
<td>1.54 (0.04)</td>
<td>1.49 (0.02)</td>
<td>10.28 (0.27)</td>
<td>14.59 (0.66)</td>
<td>2.97 (0.03)</td>
<td>15.23 (0.35)</td>
<td>19.98 (0.75)</td>
<td>1.72 (0.01)</td>
<td>0.45 (0.01)</td>
<td>null</td>
</tr>
<tr>
<td>t5</td>
<td>0.85 (0.02)</td>
<td>0.70 (0.01)</td>
<td>0.96 (0.10)</td>
<td>1.17 (0.15)</td>
<td>0.15 (0.01)</td>
<td>1.46 (0.12)</td>
<td>1.71 (0.20)</td>
<td>2.76 (0.03)</td>
<td>0.42 (0.01)</td>
<td>null</td>
</tr>
<tr>
<td>t6</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t7</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t8</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t9</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t10</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t11</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
<tr>
<td>t12</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>null</td>
</tr>
</tbody>
</table>

Next we look at the p-values. We confirm that p1 (the second column, indicated as k.1) is almost the same as the raw BP (the first column); this should be so if the model fitting is good. Only two trees, i.e., t1 and t3, have p1 > 0.05. It is known that the bias of p1 is large so that often leads to false positives for tree
When we apply hypothesis testing to all the 10 clusters, we may use AU p-values. Let p be k.2. Then e8, e9 and e10 have p < 0.05, and e5, e6, e7 also have p < 0.10. They may be rejected at either α = 0.05 or 0.10. However, people often want to look at only the clusters appearing in the top tree. In this table, only e1 and e2 are included in t1. We should use SI p-values for testing each of these two clusters by taking it as the alternative hypothesis. Let p be k.k. Then e2 has p > 0.90, indicating that the null hypothesis (e2 is not true) is rejected as 1 − p < 0.10 at α = 0.10.

Finally we examine model fitting. According to the AIC values, the fitting is good overall except for the top two trees; however note that the AIC values should be about 10 times smaller if the default value of nb=10,000 was used. The fitting curves for the top four trees are shown (Fig. 8) by

> plot(mam15.trees[1:4])

According to the plots, the fitting is rather good even for the top two trees.

References

Figure 8. Model fitting for the top four trees


Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

Email address: shimo@i.kyoto-u.ac.jp