Phylogenetic Tree Selection

Hidetoshi Shimodaira

2019/01/14

Phylogenetic Analysis of Mammal Dataset

Load the package *scaleboot*.

```
library(scaleboot)
```

The methods are explained in Shimodaira and Terada (2019). The theory for the selective inference behind the methods is given in Terada and Shimodaira (2017).

Hidetoshi Shimodaira and Yoshikazu Terada. Selective Inference for Testing Trees and Edges in Phylogenetics. 2019.

Yoshikazu Terada and Hidetoshi Shimodaira. Selective inference for the problem of regions via multiscale bootstrap. arXiv:1711.00949, 2017.

Phylogenetic Analysis of 105 trees of 6 taxa

As a working example, we estimate the phylogenetic tree from the same dataset previously analyzed in Shimodaira and Hasegawa (1999), Shimodaira (2001, 2002) using the same model of evolution. The dataset consists of mitochondrial protein sequences of six mammalian species with $n = 3414$ amino acids The taxa are *Homo sapiens* (human), *Phoca vitulina* (seal), *Bos taurus* (cow), *Oryctolagus cuniculus* (rabbit), *Mus musculus* (mouse), and *Didelphis virginiana* (opossum). The software package PAML (Yang 1997) was used to calculate the site-wise log-likelihoods for the trees. The mtREV model (Adachi and Hasgawa 1996) was used for amino acid substitutions, and the site-heterogeneity was modeled by the discrete-gamma distribution (Yang 1996).

We first run a phylogenetic package, such as PAML, to calculate *site-wise log-likelihood* for trees. The tree topology file is mam105.tpl, and the site-wise log-likelhiood file is mam105.mt. The mam105.mt file is converted from mam105.lnf (output from PAML) by seqmt program in CONSEL. We also run treeass in CONSEL to get mam105.ass and mam105.log from mam105.tpl. We use CONSEL only for preparing *mt* and *ass* files. All these files are found in mam15 folder.

Instead of using the program *consel* in CONSEL to compute p-values, we use scaleboot here. First, read the following two files. Then run *relltest* (internally calling *scaleboot* function) to perform multiscale bootstrap resampling.

```
### dont run
nb.rell = 100000
nb.pvclust = 10000
library(parallel)
length(cl <- makeCluster(detectCores()))
mam105.mt <- read.mt("mam15-files/mam105.mt")
mam105.ass <- read.ass("mam15-files/mam105.ass")
sa <- 9^seq(-1,1,length=13) # specify scales for multiscale bootstrap
mam105.relltest <- relltest(mam105.mt,nb=nb.rell,sa=sa,ass=mam105.ass,cluster=cl)
```

We have run the above command in *makedata.R* preveously. To get the results, simply do below, which will also load other objects.

```
data(mam15) # load mam15, mam26, mam105
ls() # look at the objects
```

```
##  [1] "mam105.ass"      "mam105.aux"      "mam105.mt"
##  [4] "mam105.relltest" "mam15.ass"       "mam15.aux"
##  [7] "mam15.mt"        "mam15.relltest"  "mam26.ass"
## [10] "mam26.aux"       "mam26.mt"
```

The output of relltest includes the results of trees and edges. We separate them, and also reorder the trees and edges in decreasing order of likelhiood values below.

```
mam105 <- sbphylo(mam105.relltest, mam105.ass)
```

This includes the multiscale bootstrap probability. The order can be checked as follows. T1, T2, T3, ... are sorted tree (in decreasing order of likelhiood). t1, t2, t3, ... are the original order of trees. E1, E2, E3, ... are sorted edges, and e1, e2, e3, ... are the original order of edges.

```
mam105$order.tree  # sorted tree to original tree
```

```
##   T1   T2   T3   T4   T5   T6   T7   T8   T9  T10  T11  T12  T13  T14  T15
##    4    1    2    8    9    5   10    3    7   11   15   14    6   13   12
##  T16  T17  T18  T19  T20  T21  T22  T23  T24  T25  T26  T27  T28  T29  T30
##   19   17   35   34   41   39   42   29   38   28   18   43   40   24   30
##  T31  T32  T33  T34  T35  T36  T37  T38  T39  T40  T41  T42  T43  T44  T45
##   16   20   25   65   63   33   22   66   68   51   36   50   27   32   26
##  T46  T47  T48  T49  T50  T51  T52  T53  T54  T55  T56  T57  T58  T59  T60
##   37   31   23   21   47   64   56   44   55   57   53   48   69   46   45
##  T61  T62  T63  T64  T65  T66  T67  T68  T69  T70  T71  T72  T73  T74  T75
##   76   60   74   88   49  104   58   75   89   67   73   77  101   85   59
##  T76  T77  T78  T79  T80  T81  T82  T83  T84  T85  T86  T87  T88  T89  T90
```

```
##   52   80   70   84   78   61   71   79   87   72   62   86   82   81   54
##  T91  T92  T93  T94  T95  T96  T97  T98  T99 T100 T101 T102 T103 T104 T105
##   93  102   83  100   92   95  103   96   99   90  105   94   91   98   97
```

```
mam105$invorder.tree  # original tree to sorted tree
```

```
##   t1   t2   t3   t4   t5   t6   t7   t8   t9  t10  t11  t12  t13  t14  t15
##    2    3    8    1    6   13    9    4    5    7   10   15   14   12   11
##  t16  t17  t18  t19  t20  t21  t22  t23  t24  t25  t26  t27  t28  t29  t30
##   31   17   26   16   32   49   37   48   29   33   45   43   25   23   30
##  t31  t32  t33  t34  t35  t36  t37  t38  t39  t40  t41  t42  t43  t44  t45
##   47   44   36   19   18   41   46   24   21   28   20   22   27   53   60
##  t46  t47  t48  t49  t50  t51  t52  t53  t54  t55  t56  t57  t58  t59  t60
##   59   50   57   65   42   40   76   56   90   54   52   55   67   75   62
##  t61  t62  t63  t64  t65  t66  t67  t68  t69  t70  t71  t72  t73  t74  t75
##   81   86   35   51   34   38   70   39   58   78   82   85   71   63   68
##  t76  t77  t78  t79  t80  t81  t82  t83  t84  t85  t86  t87  t88  t89  t90
##   61   72   80   83   77   89   88   93   79   74   87   84   64   69  100
##  t91  t92  t93  t94  t95  t96  t97  t98  t99 t100 t101 t102 t103 t104 t105
##  103   95   91  102   96   98  105  104   99   94   73   92   97   66  101
```

```
mam105$order.edge # sorted edge to original edge
```

```
##  E1  E2  E3  E4  E5  E6  E7  E8  E9 E10 E11 E12 E13 E14 E15 E16 E17 E18
##   2   3  16  24   4  10   7  17   5   9  22   8   1  14  15   6  20  19
## E19 E20 E21 E22 E23 E24 E25
##  12  21  18  13  11  23  25
```

```
mam105$invorder.edge # original edge to sorted edge
```

```
##  e1  e2  e3  e4  e5  e6  e7  e8  e9 e10 e11 e12 e13 e14 e15 e16 e17 e18
##  13   1   2   5   9  16   7  12  10   6  23  19  22  14  15   3   8  21
## e19 e20 e21 e22 e23 e24 e25
##  18  17  20  11  24   4  25
```

The p-values are calculated by the summary method.

```
mam105.pv <- summary(mam105)
mam105.pv$tree$value[1:5,] # p-values of the best 5 trees
```

```
##        raw        k.1        k.2       sk.1      sk.2      beta0     beta1
## T1 0.57187 0.55889656 0.75185084 0.11779312 0.3719175 -0.4142489 0.2660767
## T2 0.32255 0.30394923 0.46712902 0.60789846 0.7981964  0.2977822 0.2152934
## T3 0.03721 0.03757708 0.12604118 0.07515415 0.2018534  1.4624146 0.3171079
## T4 0.01338 0.01413656 0.08140435 0.02827311 0.1242782  1.7945885 0.3988874
## T5 0.03139 0.03183233 0.12712373 0.06366465 0.1987748  1.4973115 0.3572105
##         stat  shtest
## T1 -2.664116 0.99050
## T2  2.664116 0.92899
## T3  7.397927 0.83972
## T4 17.565794 0.57732
## T5 18.934344 0.54609
```

```
mam105.pv$edge$value[1:5,] # p-values of the best 5 edges
```

```
##        raw       k.1       k.2      sk.1      sk.2      beta0     beta1
## E1 0.99988 0.9998942 0.9999724 0.9997883 0.9999365 -3.8690582 0.1642471
## E2 0.93171 0.9298632 0.9560328 0.8597264 0.9031350 -1.5905870 0.1158142
## E3 0.58535 0.5796771 0.7188055 0.1593543 0.3383682 -0.3901821 0.1891146
## E4 0.32848 0.3179740 0.4352301 0.6359480 0.7749161  0.3182229 0.1551488
## E5 0.03737 0.0368873 0.1235372 0.0737746 0.1980655  1.4727468 0.3152620
```

We also have formatted results.

```
mam105.pv$tree$character[1:5,] # p-values of the best 5 trees
```

```
##    stat     shtest          k.1            k.2
## T1 " -2.66" "0.990 (0.000)" "0.559 (0.001)" "0.752 (0.001)"
## T2 "  2.66" "0.929 (0.001)" "0.304 (0.000)" "0.467 (0.001)"
## T3 "  7.40" "0.840 (0.001)" "0.038 (0.000)" "0.126 (0.002)"
## T4 " 17.57" "0.577 (0.002)" "0.014 (0.000)" "0.081 (0.002)"
## T5 " 18.93" "0.546 (0.002)" "0.032 (0.000)" "0.127 (0.002)"
##    sk.2            beta0          beta1         edge
## T1 "0.372 (0.001)" "-0.41 (0.00)" "0.27 (0.00)" "E1,E2,E3"
## T2 "0.798 (0.001)" " 0.30 (0.00)" "0.22 (0.00)" "E1,E2,E4"
## T3 "0.202 (0.003)" " 1.46 (0.01)" "0.32 (0.00)" "E1,E2,E5"
## T4 "0.124 (0.003)" " 1.79 (0.01)" "0.40 (0.01)" "E1,E3,E6"
## T5 "0.199 (0.003)" " 1.50 (0.01)" "0.36 (0.00)" "E1,E6,E7"
```

```
mam105.pv$edge$character[1:5,] # p-values of the best 5 edges
```

```
##    k.1             k.2             sk.2            beta0
## E1 "1.000 (0.000)" "1.000 (0.000)" "1.000 (0.000)" "-3.87 (0.03)"
## E2 "0.930 (0.000)" "0.956 (0.001)" "0.903 (0.001)" "-1.59 (0.00)"
## E3 "0.580 (0.001)" "0.719 (0.001)" "0.338 (0.001)" "-0.39 (0.00)"
```

```
## E4 "0.318 (0.000)" "0.435 (0.001)" "0.775 (0.001)" " 0.32 (0.00)"
## E5 "0.037 (0.000)" "0.124 (0.002)" "0.198 (0.002)" " 1.47 (0.01)"
##    beta1
## E1 "0.16 (0.01)"
## E2 "0.12 (0.00)"
## E3 "0.19 (0.00)"
## E4 "0.16 (0.00)"
## E5 "0.32 (0.00)"
##    tree
## E1 "T1,T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12,T13,T14,T15"
## E2 "T1,T2,T3,T16,T17,T26,T29,T31,T32,T36,T37,T41,T44,T46,T47"
## E3 "T1,T4,T9,T16,T17,T18,T19,T23,T25"
## E4 "T2,T6,T8,T26,T31,T43,T45,T48,T49"
## E5 "T3,T11,T14,T27,T28,T29,T32,T51,T55,T58,T61,T63,T68,T71,T72"
```

The formatted table can be used for prepare latex table.

```
table2latex <- function(x) {
  rn <- rownames(x)
  cn <- colnames(x); cl <- length(cn)
  cat("\n\\begin{tabular}{",paste(rep("c",cl+1),collapse=""),"}\n",sep="")
  cat("\\hline\n")
  cat("&",paste(cn,collapse=" & "),"\\\\\n")
  for(i in seq(along=rn)) {
    cat(rn[i],"&",paste(x[i,],collapse=" & "),"\\\\\n")
  }
  cat("\\hline\n")
  cat("\\end{tabular}\n")
}
```

In the tree table below, we omitted *stat* (log-likelihood difference), *shtest* (Shimodaira-Hasegawa test p-value). The other values are: *k.1* (BP, bootstrap probability), *k.2* (AU, approximately unbiased p-value), *sk.2* (SI, selective inference p-value), *beta0* (β_0, signed distance), *beta1* (β_1, mean curvature), *edge* (the associated edges).

```
table2latex(mam105.pv$tree$character[1:20,-(1:2)]) # the best 20 trees
```

```
##
## \begin{tabular}{ccccccc}
## \hline
## & k.1 & k.2 & sk.2 & beta0 & beta1 & edge \\
## T1 & 0.559 (0.001) & 0.752 (0.001) & 0.372 (0.001) & -0.41 (0.00) & 0.27 (0.00) & E1,E2,E3 \\
## T2 & 0.304 (0.000) & 0.467 (0.001) & 0.798 (0.001) &  0.30 (0.00) & 0.22 (0.00) & E1,E2,E4 \\
## T3 & 0.038 (0.000) & 0.126 (0.002) & 0.202 (0.003) &  1.46 (0.01) & 0.32 (0.00) & E1,E2,E5 \\
## T4 & 0.014 (0.000) & 0.081 (0.002) & 0.124 (0.003) &  1.79 (0.01) & 0.40 (0.01) & E1,E3,E6 \\
## T5 & 0.032 (0.000) & 0.127 (0.002) & 0.199 (0.003) &  1.50 (0.01) & 0.36 (0.00) & E1,E6,E7 \\
## T6 & 0.005 (0.000) & 0.032 (0.002) & 0.050 (0.002) &  2.21 (0.02) & 0.35 (0.01) & E1,E4,E7 \\
## T7 & 0.015 (0.000) & 0.100 (0.003) & 0.150 (0.003) &  1.72 (0.01) & 0.44 (0.01) & E1,E6,E8 \\
## T8 & 0.001 (0.000) & 0.011 (0.001) & 0.016 (0.002) &  2.74 (0.03) & 0.43 (0.02) & E1,E4,E9 \\
## T9 & 0.000 (0.000) & 0.001 (0.000) & 0.001 (0.000) &  3.67 (0.09) & 0.46 (0.04) & E1,E3,E10 \\
## T10 & 0.002 (0.000) & 0.022 (0.002) & 0.033 (0.002) &  2.43 (0.02) & 0.42 (0.01) & E1,E8,E9 \\
## T11 & 0.000 (0.000) & 0.004 (0.001) & 0.006 (0.002) &  3.14 (0.07) & 0.51 (0.03) & E1,E5,E8 \\
## T12 & 0.000 (0.000) & 0.000 (0.000) & 0.001 (0.000) &  3.78 (0.09) & 0.41 (0.04) & E1,E9,E10 \\
## T13 & 0.000 (0.000) & 0.000 (0.000) & 0.001 (0.001) &  3.96 (0.19) & 0.54 (0.09) & E1,E7,E11 \\
## T14 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.66 (0.31) & 0.65 (0.12) & E1,E5,E11 \\
## T15 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  5.28 (0.34) & 0.43 (0.11) & E1,E10,E11 \\
## T16 & 0.000 (0.000) & 0.000 (0.000) & 0.001 (0.000) &  3.63 (0.04) & 0.23 (0.01) & E2,E3,E12 \\
## T17 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  3.81 (0.04) & 0.22 (0.01) & E2,E3,E13 \\
## T18 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.33 (0.10) & 0.34 (0.03) & E3,E6,E12 \\
## T19 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.36 (0.11) & 0.32 (0.04) & E3,E6,E13 \\
## T20 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  3.90 (0.12) & 0.44 (0.05) & E6,E8,E14 \\
## \hline
## \end{tabular}
```

In the edge table below, we omitted *tree* (associated trees). The other values are: *k.1* (BP, bootstrap probability), *k.2* (AU, approximately unbiased p-value), *sk.2* (SI, selective inference p-value), *beta0* (β_0, signed distance), *beta1* (β_1, mean curvature).

```
table2latex(mam105.pv$edge$character[,-6]) # all the 25 edges
```

```
##
## \begin{tabular}{ccccc}
## \hline
## & k.1 & k.2 & sk.2 & beta0 & beta1 \\
## E1 & 1.000 (0.000) & 1.000 (0.000) & 1.000 (0.000) & -3.87 (0.03) & 0.16 (0.01) \\
## E2 & 0.930 (0.000) & 0.956 (0.001) & 0.903 (0.001) & -1.59 (0.00) & 0.12 (0.00) \\
## E3 & 0.580 (0.001) & 0.719 (0.001) & 0.338 (0.001) & -0.39 (0.00) & 0.19 (0.00) \\
## E4 & 0.318 (0.000) & 0.435 (0.001) & 0.775 (0.001) &  0.32 (0.00) & 0.16 (0.00) \\
## E5 & 0.037 (0.000) & 0.124 (0.002) & 0.198 (0.002) &  1.47 (0.01) & 0.32 (0.00) \\
## E6 & 0.060 (0.000) & 0.074 (0.001) & 0.141 (0.002) &  1.50 (0.00) & 0.05 (0.00) \\
## E7 & 0.038 (0.000) & 0.091 (0.002) & 0.154 (0.002) &  1.56 (0.01) & 0.22 (0.00) \\
## E8 & 0.018 (0.000) & 0.068 (0.002) & 0.110 (0.003) &  1.80 (0.01) & 0.31 (0.01) \\
## E9 & 0.003 (0.000) & 0.014 (0.001) & 0.023 (0.002) &  2.48 (0.02) & 0.27 (0.02) \\
## E10 & 0.000 (0.000) & 0.000 (0.000) & 0.001 (0.000) &  3.72 (0.07) & 0.29 (0.03) \\
## E11 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.31 (0.10) & 0.35 (0.03) \\
```

```
## E12 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  3.68 (0.05) & 0.17 (0.02) \\
## E13 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  3.90 (0.04) & 0.15 (0.02) \\
## E14 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.03 (0.09) & 0.30 (0.04) \\
## E15 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.03 (0.13) & 0.38 (0.06) \\
## E16 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.44 (0.05) & 0.12 (0.01) \\
## E17 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.70 (0.07) & 0.19 (0.02) \\
## E18 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  3.94 (0.09) & 0.26 (0.04) \\
## E19 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  5.23 (0.43) & 0.57 (0.13) \\
## E20 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  5.66 (0.29) & 0.28 (0.09) \\
## E21 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  6.38 (0.33) & 0.24 (0.08) \\
## E22 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  5.62 (0.21) & 0.17 (0.07) \\
## E23 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  4.86 (0.43) & 0.70 (0.13) \\
## E24 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  5.61 (0.17) & 0.23 (0.04) \\
## E25 & 0.000 (0.000) & 0.000 (0.000) & 0.000 (0.000) &  6.32 (0.71) & 0.52 (0.20) \\
## \hline
## \end{tabular}
```

We have auxiliary information in *mam105.aux*. The topologies are in the order of mam105.tpl (the same order as mam105.mt). The edges are in the order of mam105.cld (extracted from mam105.log, which is the log file of treeass).

```
names(mam105.aux)
```

```
## [1] "tpl" "cld" "tax"
```

```
mam105.aux$tpl[1:3] # topologies (the first three trees, in the order of mam105.tpl file)
```

```
##                                          t1
## "(Homsa,((Phovi,Bosta),Orycu),(Musmu,Didvi));"
##                                          t2
## "((Homsa,Orycu),(Phovi,Bosta),(Didvi,Musmu));"
##                                          t3
## "((Homsa,Musmu),((Phovi,Bosta),Orycu),Didvi);"
```

```
mam105.aux$cld[1:3] # edges  (the first three edges, in the order of  mam105.cld file)
```

```
##       e1       e2       e3
## "++----" "-++---" "++++--"
```

```
mam105.aux$tax # taxa, the order corresponds to the positions of + and - in the clade pattern.
```

```
## [1] "Homsa" "Phovi" "Bosta" "Orycu" "Musmu" "Didvi"
```

We can specify these auxiliary information in *sbphylo*.

```
mam105 <- sbphylo(mam105.relltest, mam105.ass, treename=mam105.aux$tpl,edgename=mam105.aux$cld,taxaname=mam105.aux$tax)
```

The fomatted tables are now accampanied by tree topology and clade pattern.

```
mam105.pv <- summary(mam105)
mam105.pv$tree$character[1:5,] # p-values of the best 5 trees
```

```
##    stat     shtest          k.1             k.2
## T1 " -2.66" "0.990 (0.000)" "0.559 (0.001)" "0.752 (0.001)"
## T2 "  2.66" "0.929 (0.001)" "0.304 (0.000)" "0.467 (0.001)"
## T3 "  7.40" "0.840 (0.001)" "0.038 (0.000)" "0.126 (0.002)"
## T4 " 17.57" "0.577 (0.002)" "0.014 (0.000)" "0.081 (0.002)"
## T5 " 18.93" "0.546 (0.002)" "0.032 (0.000)" "0.127 (0.002)"
##    sk.2            beta0          beta1
## T1 "0.372 (0.001)" "-0.41 (0.00)" "0.27 (0.00)"
## T2 "0.798 (0.001)" " 0.30 (0.00)" "0.22 (0.00)"
## T3 "0.202 (0.003)" " 1.46 (0.01)" "0.32 (0.00)"
## T4 "0.124 (0.003)" " 1.79 (0.01)" "0.40 (0.01)"
## T5 "0.199 (0.003)" " 1.50 (0.01)" "0.36 (0.00)"
##    tree                                          edge
## T1 "(Homsa,(Phovi,Bosta),((Didvi,Musmu),Orycu));" "E1,E2,E3"
## T2 "(Homsa,((Phovi,Bosta),Orycu),(Musmu,Didvi));" "E1,E2,E4"
## T3 "((Homsa,Orycu),(Phovi,Bosta),(Didvi,Musmu));" "E1,E2,E5"
## T4 "(Homsa,(Phovi,Bosta),(Didvi,(Orycu,Musmu)));" "E1,E3,E6"
## T5 "(Homsa,((Phovi,Bosta),(Orycu,Musmu)),Didvi);" "E1,E6,E7"
```

```
mam105.pv$edge$character[1:5,] # p-values of the best 5 edges
```

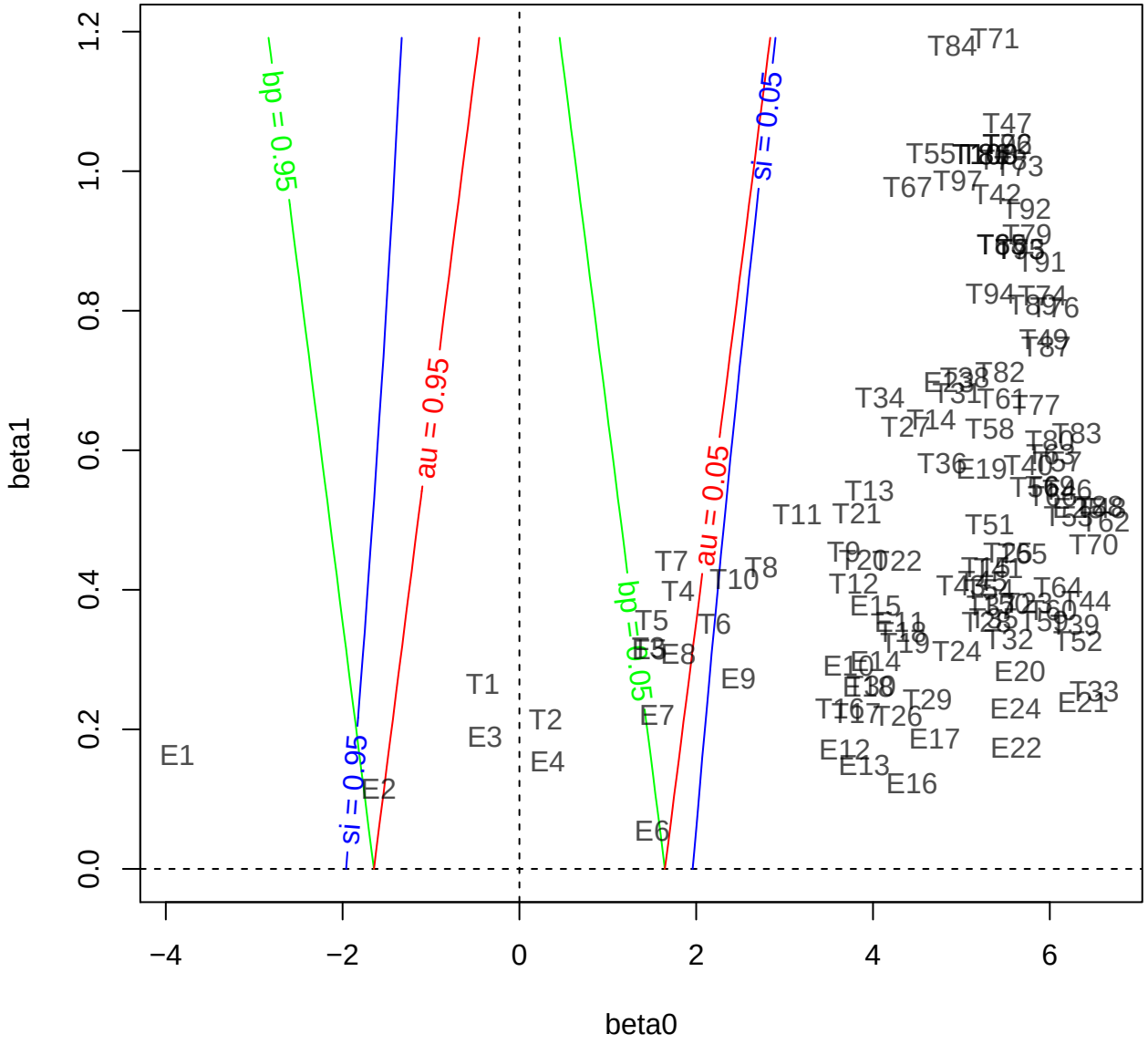
```
##    k.1             k.2             sk.2            beta0
## E1 "1.000 (0.000)" "1.000 (0.000)" "1.000 (0.000)" "-3.87 (0.03)"
## E2 "0.930 (0.000)" "0.956 (0.001)" "0.903 (0.001)" "-1.59 (0.00)"
## E3 "0.580 (0.001)" "0.719 (0.001)" "0.338 (0.001)" "-0.39 (0.00)"
## E4 "0.318 (0.000)" "0.435 (0.001)" "0.775 (0.001)" " 0.32 (0.00)"
## E5 "0.037 (0.000)" "0.124 (0.002)" "0.198 (0.002)" " 1.47 (0.01)"
##    beta1          edge
## E1 "0.16 (0.01)" "-++---"
## E2 "0.12 (0.00)" "++++--"
## E3 "0.19 (0.00)" "+++---"
## E4 "0.16 (0.00)" "-+++--"
## E5 "0.32 (0.00)" "+--+--"
##    tree
```

```
## E1 "T1,T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12,T13,T14,T15"
## E2 "T1,T2,T3,T16,T17,T26,T29,T31,T32,T36,T37,T41,T44,T46,T47"
## E3 "T1,T4,T9,T16,T17,T18,T19,T23,T25"
## E4 "T2,T6,T8,T26,T31,T43,T45,T48,T49"
## E5 "T3,T11,T14,T27,T28,T29,T32,T51,T55,T58,T61,T63,T68,T71,T72"
```

Geometric Quantities

The two geometric quantities play important roles in our theory of multiscale bootstrap. They are signed distance (β_0) and mean curvature (β_1). We look at estimated values of (β_0, β_1) for trees and edges.

```
a1 <- attr(summary(mam105$trees,k=2),"table") # extract (beta0,beta1) for trees
a2 <- attr(summary(mam105$edges,k=2),"table") # extract (beta0,beta1) for edges
beta <- rbind(a1$value,a2$value)[,c("beta0","beta1")]
sbplotbeta(beta,col=rgb(0,0,0,alpha=0.7))
```



Diagnostics of multiscale bootstrap

In *scaleboot*, p-values are computed by multiscale bootstrap. We compute bootstrap probabilities at several scales, and fit models of scaling-law to them. We look at the model fitting for diagnostics.

tree T1

Look at the model fitting of tree T1. Candidate models are used for fitting, and sorted by AIC values. Model parameters $(\beta_0, \beta_1, \beta_2)$ are estimated by the maximum likelihood method. Models are sorted by AIC. We also plot $\psi(\sigma^2)$ function. It is defined as

$$\psi(\sigma^2) = \Phi^{-1}(1 - \mathrm{BP}(\sigma^2)), \quad \sigma^2 = \frac{n}{n'}$$

for the sample size of dataset n, and that of bootstrap replicates n'. We compute bootstrap probabilities (BP) for several $n' = n/\sigma^2$ values. Then fitting parametric models to $\psi(\sigma^2)$. The most standard model is *poly.2*

$$\mathrm{poly}.2(\sigma^2) = \beta_0 + \beta_1 \sigma^2,$$

and its generalization

$$\mathrm{poly}.k(\sigma^2) = \sum_{i=0}^{k-1} \beta_i \sigma^{2i},$$

for $k = 1, 2, 3$. Also considered is the singular model

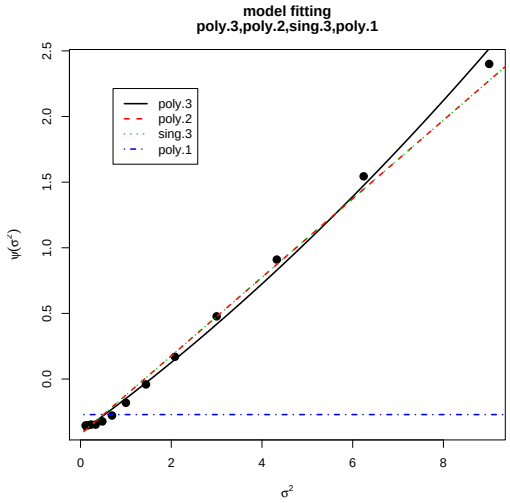
$$\mathrm{sing}.3 = \beta_0 + \frac{\beta_1 \sigma^2}{1 + \beta_2(\sigma - 1)}.$$

The result is as follows. The best fitting model is *poly.3*.

```
(f <- mam105$trees$T1) # the list of fitted models (MLE and AIC)
```

```
##
## Multiscale Bootstrap Probabilities (percent):
## 1     2     3     4     5     6     7     8     9     10    11    12    13
## 85.57 81.07 76.49 72.62 67.93 63.08 57.19 51.35 45.32 39.14 33.09 26.82 21.19
##
## Numbers of Bootstrap Replicates:
## 1     2     3     4     5     6     7     8     9     10    11    12    13
## 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05
##
## Scales (Sigma Squared):
## 1      2      3      4      5      6      7 8      9    10 11    12    13
## 0.1111 0.1603 0.2311 0.3333 0.4808 0.6933 1 1.442 2.08 3  4.327 6.241 9.008
##
## Coefficients:
##         beta0             beta1            beta2
## poly.3 -0.4059 (0.0011) 0.2494 (0.0021) 0.0083 (0.0003)
## poly.2 -0.4218 (0.0009) 0.2992 (0.0009)
## sing.3 -0.4218 (0.0009) 0.2992 (0.0009) 0.0000 (0.0000)
## poly.1 -0.2708 (0.0008)
##
## Model Fitting:
##         rss       df pfit   aic
## poly.3   1148.40 10 0.0000   1128.40
## poly.2   1857.34 11 0.0000   1835.34
## sing.3   1857.34 10 0.0000   1837.34
## poly.1 122971.02 12 0.0000 122947.02
##
## Best Model:  poly.3
```

```
plot(f,legend="topleft",pch=16,cex=1.5,lwd=2) # fitting curves
```

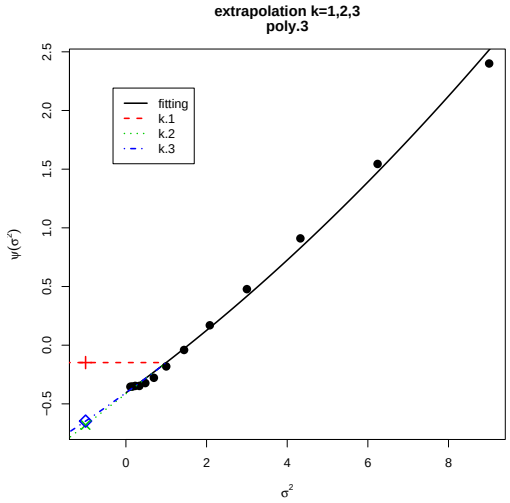


p-values are computed using the fitted models. We extrapolate $\psi(\sigma^2)$ to $\sigma^2 = 0$ and $\sigma^2 = -1$, and these values are used in AU and SI. On the other hand BP is computed as $1 - \Phi(\psi(1))$; this improves the raw value of BP(1) in terms of standard error. For each model, we extrapoloate $\psi(\sigma^2)$. We consider the Taylor expansion of $\psi(\sigma^2)$ at $\sigma^2 = 1$, and extrapolate $\psi(\sigma^2)$ by polynomial of degree $k-1$. In the below, we use $k = 1, 2, 3$ for the Taylor expansion. $k = 1$ is used for BP. $k = 2, 3$ can be used for AU and SI. The default value of k in *sbphylo* is $k = 2$.

```
(g <- summary(mam105$trees$T1,k=1:3))
```

```
##
## Raw Bootstrap Probability (scale=1) : 57.19 (0.16)
##
## Hypothesis: alternative
##
## Corrected P-values for Models (percent,Frequentist):
##        k.1          k.2          k.3          sk.1         sk.2         sk.3         beta0        beta1        aic       weight
## poly.3 55.89 (0.05) 75.19 (0.07) 74.12 (0.10) 11.78 (0.11) 37.19 (0.10) 36.06 (0.12) -0.41 (0.00)  0.27 (0.00)   1128.40 100.00
## poly.2 54.88 (0.04) 76.45 (0.05) 76.45 (0.05)  9.75 (0.07) 38.42 (0.09) 38.42 (0.09) -0.42 (0.00)  0.30 (0.00)   1835.34
## sing.3 54.88 (0.04) 76.45 (0.05) 76.45 (0.05)  9.75 (0.07) 38.42 (0.09) 38.42 (0.09) -0.42 (0.00)  0.30 (0.00)   1837.34
## poly.1 60.67 (0.03) 60.67 (0.03) 60.67 (0.03) 21.34 (0.06) 21.34 (0.06) 21.34 (0.06) -0.27 (0.00)  0.00 (0.00) 122947.02
##
## Best Model:  poly.3
##
## Corrected P-values by the Best Model and by Akaike Weights Averaging:
##         k.1          k.2          k.3          sk.1         sk.2         sk.3         beta0        beta1
## best    55.89 (0.05) 75.19 (0.07) 74.12 (0.10) 11.78 (0.11) 37.19 (0.10) 36.06 (0.12) -0.41 (0.00)  0.27 (0.00)
## average 55.89 (0.05) 75.19 (0.07) 74.12 (0.10) 11.78 (0.11) 37.19 (0.10) 36.06 (0.12) -0.41 (0.00)  0.27 (0.00)
```

```
plot(g,legend="topleft",pch=16,cex=1.5,lwd=2)
```



In the table, *k.1* is BP. *k.2* or *k.3* is used for AU. *sk.2* or *sk.3* is used for SI. beta0 and beta1 are estimated values of β_0 and β_1, obtained as the tangent line at $\sigma^2 = 1$. Thus these β_0 and β_1 correspond to the Talor expansion with $k = 2$.

In *sbphylo*, you can replace $k = 2$ by $k = 3$ (or you could specify $k = 4$) as follows. This may improve the accuracy of AU and SI when $\psi(\sigma^2)$ deviates from the linear model *poly.2*. There is a trade-off between the accuracy and stability, so $k = 2$ or $k = 3$ would be a good choice, instead of using larger values such as $k = 4$.

```
mam105.pv3 <- summary(mam105,k=2:3) # simply specify k=3 is also fine
mam105.pv3$tree$value[1:5,] # p-values of the best 5 trees
```

```
##        raw        k.1        k.2        k.3       sk.1      sk.2      sk.3
## T1 0.57187 0.55889656 0.75185084 0.74118983 0.11779312 0.3719175 0.3605526
## T2 0.32255 0.30394923 0.46712902 0.44831498 0.60789846 0.7981964 0.7847087
## T3 0.03721 0.03757708 0.12604118 0.13753611 0.07515415 0.2018534 0.2150208
## T4 0.01338 0.01413656 0.08140435 0.08896401 0.02827311 0.1242782 0.1331120
## T5 0.03139 0.03183233 0.12712373 0.13341131 0.06366465 0.1987748 0.2059271
##         beta0     beta1      stat  shtest
## T1 -0.4142489 0.2660767 -2.664116 0.99050
## T2  0.2977822 0.2152934  2.664116 0.92899
## T3  1.4624146 0.3171079  7.397927 0.83972
## T4  1.7945885 0.3988874 17.565794 0.57732
## T5  1.4973115 0.3572105 18.934344 0.54609
```

```
mam105.pv3$edge$value[1:5,] # p-values of the best 5 edges
```

```
##        raw       k.1       k.2       k.3       sk.1      sk.2      sk.3
## E1 0.99988 0.9998942 0.9999724 0.9999728 0.9997883 0.9999365 0.9999374
## E2 0.93171 0.9298632 0.9560328 0.9561256 0.8597264 0.9031350 0.9032764
## E3 0.58535 0.5796771 0.7188055 0.7208303 0.1593543 0.3383682 0.3403908
## E4 0.32848 0.3179740 0.4352301 0.4307262 0.6359480 0.7749161 0.7715507
## E5 0.03737 0.0368873 0.1235372 0.1661949 0.0737746 0.1980655 0.2458701
```

```
##         beta0     beta1
## E1 -3.8690582 0.1642471
## E2 -1.5905870 0.1158142
## E3 -0.3901821 0.1891146
## E4  0.3182229 0.1551488
## E5  1.4727468 0.3152620
```

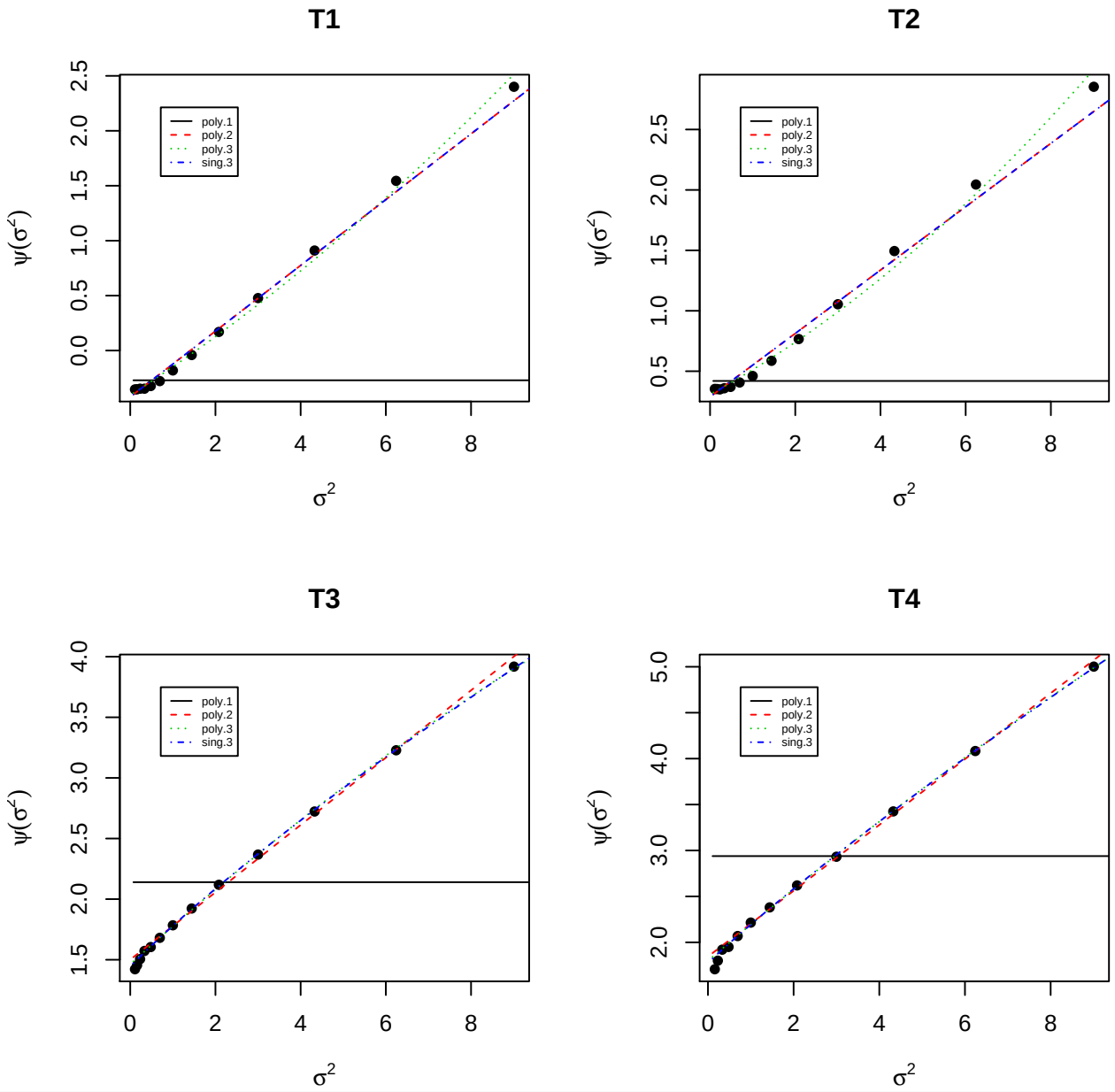
trees T1, T2, T3, T4

The fitting and p-values can be seen for several trees at the same time. Look at the results for the best 4 trees.

```
(f <- mam105$trees[1:4])
```

```
##
## Test Statistic, and Shimodaira-Hasegawa test
##    stat  shtest
## t4 -2.66 99.05 (0.03)
## t1  2.66 92.90 (0.08)
## t2  7.40 83.97 (0.12)
## t8 17.57 57.73 (0.16)
##
## Multiscale Bootstrap Probabilities (percent):
##    1  2  3  4  5  6  7  8  9  10 11 12 13
## t4 86 81 76 73 68 63 57 51 45 39 33 27 21
## t1 14 19 23 27 30 31 32 31 30 27 24 21 17
## t2  0  0  0  0  1  2  4  5  7  9 10 10 10
## t8  0  0  0  0  0  1  1  2  3  5  5  5  5
##
## Numbers of Bootstrap Replicates:
## 1     2     3     4     5     6     7     8     9     10    11    12    13
## 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05
##
## Scales (Sigma Squared):
## 1      2      3      4      5      6      7 8     9    10 11    12    13
## 0.1111 0.1603 0.2311 0.3333 0.4808 0.6933 1 1.442 2.08 3  4.327 6.241 9.008
##
## AIC values of Model Fitting:
##    poly.1    poly.2  poly.3  sing.3
## T1 122947.02 1835.34 1128.40 1837.34
## T2  88533.62 2499.41 1178.97 2501.41
## T3  35229.15   93.54    3.02   -7.89
## T4  32433.71   47.48    6.91    0.96
```

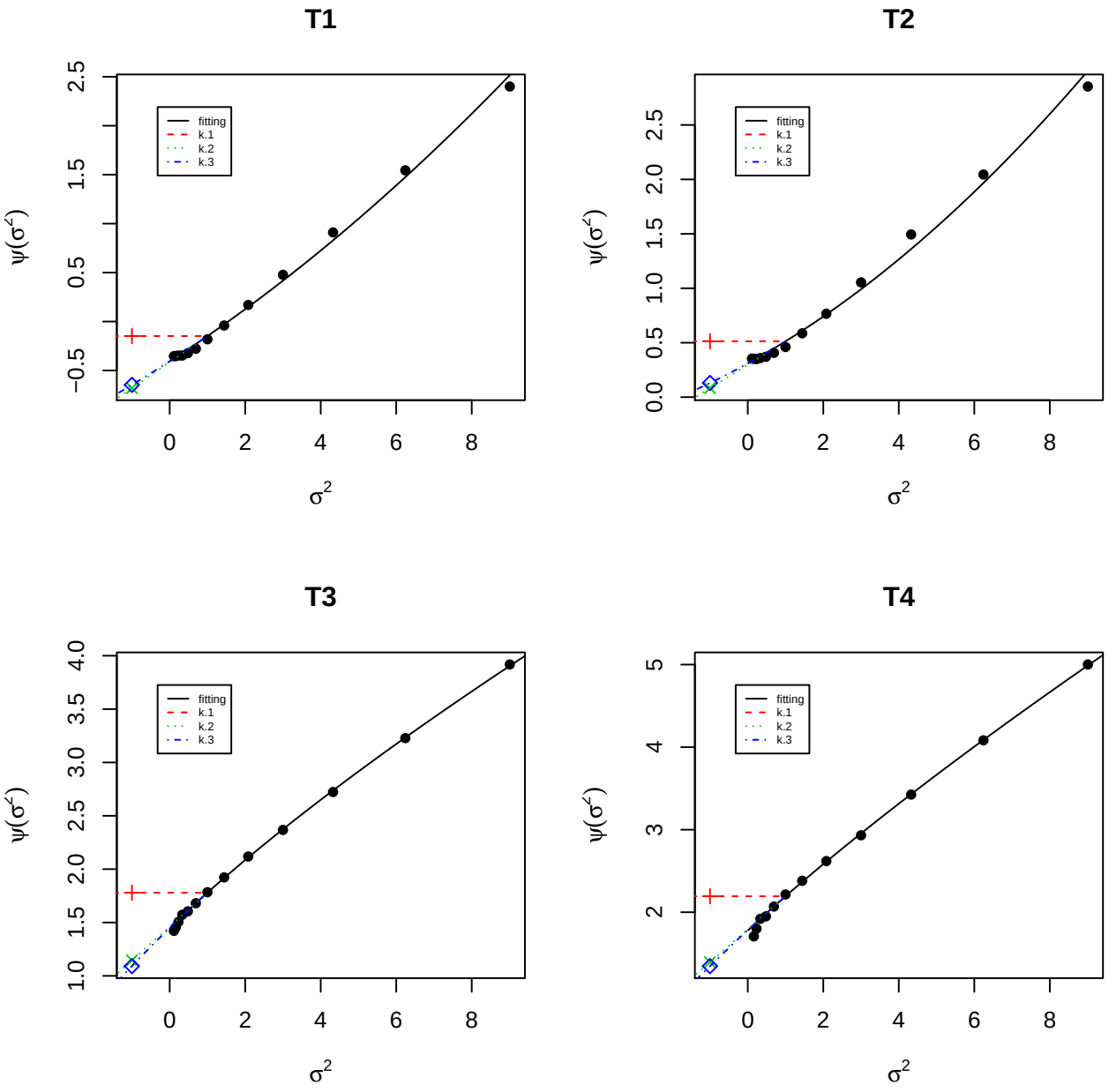
```
plot(f,legend="topleft",pch=16,cex=1,lwd=1,cex.legend=0.5) # fitting curves
```




```
(g <- summary(mam105$trees[1:4],k=1:3))
```

```
##
## Corrected P-values by Akaike Weights Averaging (percent,Frequentist):
##    raw          k.1          k.2          k.3          sk.1         sk.2         sk.3         beta0        beta1       hypothesis  model  weight
## T1 57.19 (0.16) 55.89 (0.05) 75.19 (0.07) 74.12 (0.10) 11.78 (0.11) 37.19 (0.10) 36.06 (0.12) -0.41 (0.00) 0.27 (0.00) alternative poly.3 100.00
## T2 32.26 (0.15) 30.39 (0.05) 46.71 (0.09) 44.83 (0.13) 60.79 (0.10) 79.82 (0.08) 78.47 (0.10)  0.30 (0.00) 0.22 (0.00) null        poly.3 100.00
## T3  3.72 (0.06)  3.76 (0.03) 12.60 (0.19) 13.75 (0.33)  7.52 (0.05) 20.19 (0.26) 21.50 (0.41)  1.46 (0.01) 0.32 (0.00) null        sing.3  99.57
## T4  1.34 (0.04)  1.41 (0.02)  8.14 (0.23)  8.90 (0.37)  2.83 (0.04) 12.43 (0.31) 13.31 (0.47)  1.79 (0.01) 0.40 (0.01) null        sing.3  95.14
```

```
plot(g,legend="topleft",pch=16,cex=1,lwd=1,cex.legend=0.5) # extrapolation
```



edges E1, E2, E3, E4

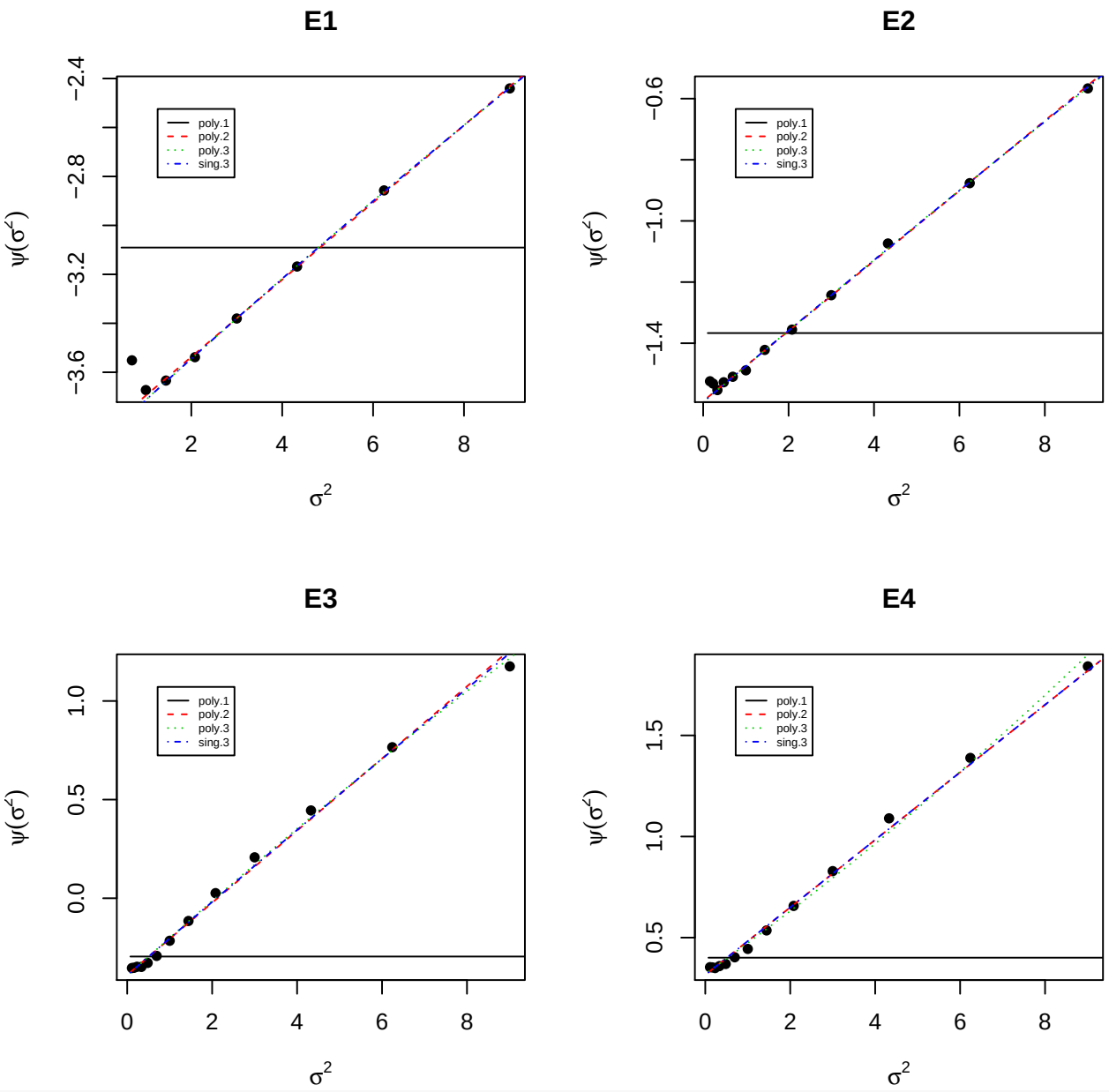
Look at the results for the best 4 edges.

```
(f <- mam105$edges[1:4])
```

```
##
## Test Statistic, and Shimodaira-Hasegawa test
##     stat   shtest
## e2  -48.52 100.00 (0.00)
## e3  -17.57  99.83 (0.01)
## e16  -2.66  96.59 (0.06)
## e24   2.66  90.02 (0.09)
##
## Multiscale Bootstrap Probabilities (percent):
##     1   2   3   4   5   6   7   8   9  10 11 12 13
## e2  100 100 100 100 100 100 100 100 99 97 94 87 79
## e3  100 100 100 100 100  99  97  93  88 83 76 70 64 57
## e16  86  81  76  73  68  64  59  54 49 45 42 38 35
## e24  14  19  23  27  30  31  33  33 32 32 30 29 27
##
## Numbers of Bootstrap Replicates:
```

```
## 1     2     3     4     5     6     7     8     9     10    11    12    13
## 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05 1e+05
##
## Scales (Sigma Squared):
## 1      2      3      4      5      6      7 8     9    10 11    12    13
## 0.1111 0.1603 0.2311 0.3333 0.4808 0.6933 1 1.442 2.08 3  4.327 6.241 9.008
##
## AIC values of Model Fitting:
##    poly.1   poly.2 poly.3 sing.3
## E1  4243.10 -16.28 -15.34 -15.18
## E2 10331.22  -3.81  -2.73  -2.30
## E3 47975.33 478.93 455.67 475.44
## E4 37830.63 493.48 409.76 495.48
```

```
plot(f,legend="topleft",pch=16,cex=1,lwd=1,cex.legend=0.5) # fitting curves
```




```
(g <- summary(mam105$edges[1:4],k=1:3))
```

```
##
## Corrected P-values by Akaike Weights Averaging (percent,Frequentist):
##    raw          k.1          k.2          k.3          sk.1         sk.2         sk.3         beta0        beta1       hypothesis  model  weight
## E1 99.99 (0.00) 99.99 (0.00) 100.00 (0.00) 100.00 (0.00) 99.98 (0.00) 99.99 (0.00) 99.99 (0.00) -3.87 (0.03) 0.16 (0.01) alternative poly.2  45.45
```

```
## E2 93.17 (0.08) 92.99 (0.04)  95.60 (0.05)  95.61 (0.06) 85.97 (0.07) 90.31 (0.10) 90.33 (0.12) -1.59 (0.00) 0.12 (0.00) alternative poly.2  48.71
## E3 58.54 (0.16) 57.97 (0.05)  71.88 (0.07)  72.08 (0.11) 15.94 (0.10) 33.84 (0.09) 34.04 (0.12) -0.39 (0.00) 0.19 (0.00) alternative poly.3 100.00
## E4 32.85 (0.15) 31.80 (0.05)  43.52 (0.09)  43.07 (0.13) 63.59 (0.10) 77.49 (0.08) 77.16 (0.11)  0.32 (0.00) 0.16 (0.00) null        poly.3 100.00
plot(g,legend="topleft",pch=16,cex=1,lwd=1,cex.legend=0.5) # extrapolation
```