Bayesian Hierarchical Clustering

Rich Savage

November 1, 2022

This is a simple Sweave test of the Bayesian Hierarchical Clustering method, as implemented in the R package *BHC*. It runs the example code as given in the R help file and generates a resulting dendrogram plot, to show the sort of results one can expect.

```
> require(BHC)
> ##BUILD SAMPLE DATA AND LABELS
> data         <- matrix(0,15,10)
> itemLabels   <- vector("character",15)
> data[1:5,]   <- 1 ; itemLabels[1:5]   <- "a"
> data[6:10,]  <- 2 ; itemLabels[6:10]  <- "b"
> data[11:15,] <- 3 ; itemLabels[11:15] <- "c"
> timePoints   <- 1:10 # for the time-course case
> ##DATA DIMENSIONS
> nDataItems <- nrow(data)
> nFeatures  <- ncol(data)
> ##RUN MULTINOMIAL CLUSTERING
> hc1 <- bhc(data,itemLabels,verbose=TRUE)

[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]   0.7642338 -88.7830929
[1]   1.236254 -99.514068
[1]   0.4309005 -79.1391819
[1]   0.4309005 -79.1391819
[1]   0.4309005 -79.1391819
[1] Hyperparameter: 0.430900452187475
[1] Lower bound on overall LogEvidence: -7.9139e+01
[1] *******************

> ##RUN TIME-COURSE CLUSTERING
> hc2 <- bhc(data, itemLabels, 0, timePoints, "time-course",
+           numReps=1, noiseMode=0, numThreads=1, verbose=TRUE)

[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: time-course"
[1]   0.0000 242.4406
[1] Hyperparameter: 0
[1] Lower bound on overall LogEvidence: 2.4244e+02
[1] *******************

> ##OUTPUT CLUSTER LABELS TO FILE
> WriteOutClusterLabels(hc1, "labels.txt", verbose=TRUE)
```

```
[1] ---CLUSTER 1 ---
[1] c
[1] c
[1] c
[1] c
[1] c
[1] ---CLUSTER 2 ---
[1] a
[1] a
[1] a
[1] a
[1] a
[1] ---CLUSTER 3 ---
[1] b
[1] b
[1] b
[1] b
[1] b
```

```
> ##FOR THE MULTINOMIAL CASE, THE DATA CAN BE DISCRETISED
> newData      <- data[] + rnorm(150, 0, 0.1);
> percentiles  <- FindOptimalBinning(newData, itemLabels, transposeData=TRUE, verbose=TRUE)
```

```
DATA DISCRETISATION
-------------------
Percentiles: 0.1 0.8 0.1
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 271.207188900625
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -100.3550
[1] 1265.9246 -100.3267
[1] 1564.651 -100.318
[1] 1806.2707 -100.3131
[1] 1898.6031 -100.3115
[1] 1955.6676 -100.3106
[1] 1990.9355 -100.3101
[1] 2012.7322 -100.3098
[1] 2026.2033 -100.3096
[1] 2034.5289 -100.3095
[1] 2039.6744 -100.3094
[1] 2042.8545 -100.3093
[1] 2044.8199 -100.3093
[1] 2046.0346 -100.3093
[1] 2046.7853 -100.3093
[1] 2047.2493 -100.3093
[1] 2047.5826 -100.3093
```

```
[1] 2047.5826 -100.3093
[1] 2047.5826 -100.3093
[1] Hyperparameter: 2047.58264449935
[1] Lower bound on overall LogEvidence: -1.0031e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.15 0.7 0.15
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 333.875058211373
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -143.8307
[1] 1265.9246 -143.7785
[1] 1564.6508 -143.7624
[1] 1808.0122 -143.7532
[1] 1715.0564 -143.7564
[1] 1899.6794 -143.7503
[1] 1956.3328 -143.7487
[1] 1991.3466 -143.7477
[1] 2012.9863 -143.7471
[1] 2026.3603 -143.7468
[1] 2034.6259 -143.7466
[1] 2039.7344 -143.7464
[1] 2042.8916 -143.7463
[1] 2044.8428 -143.7463
[1] 2046.0488 -143.7463
[1] 2046.7941 -143.7462
[1] 2047.2547 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] Hyperparameter: 2047.58805303045
[1] Lower bound on overall LogEvidence: -1.4375e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.2 0.6 0.2
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 333.875058211373
(Need to add this to the model logEvidence)
```

```
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -143.8307
[1] 1265.9246 -143.7785
[1] 1564.6508 -143.7624
[1] 1808.0122 -143.7532
[1] 1715.0564 -143.7564
[1] 1899.6794 -143.7503
[1] 1956.3328 -143.7487
[1] 1991.3466 -143.7477
[1] 2012.9863 -143.7471
[1] 2026.3603 -143.7468
[1] 2034.6259 -143.7466
[1] 2039.7344 -143.7464
[1] 2042.8916 -143.7463
[1] 2044.8428 -143.7463
[1] 2046.0488 -143.7463
[1] 2046.7941 -143.7462
[1] 2047.2547 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] Hyperparameter: 2047.58805303045
[1] Lower bound on overall LogEvidence: -1.4375e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.25 0.5 0.25
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 333.875058211373
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -143.8307
[1] 1265.9246 -143.7785
[1] 1564.6508 -143.7624
[1] 1808.0122 -143.7532
[1] 1715.0564 -143.7564
[1] 1899.6794 -143.7503
[1] 1956.3328 -143.7487
[1] 1991.3466 -143.7477
[1] 2012.9863 -143.7471
[1] 2026.3603 -143.7468
[1] 2034.6259 -143.7466
[1] 2039.7344 -143.7464
```

```
[1] 2042.8916 -143.7463
[1] 2044.8428 -143.7463
[1] 2046.0488 -143.7463
[1] 2046.7941 -143.7462
[1] 2047.2547 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] 2047.5881 -143.7462
[1] Hyperparameter: 2047.58805303045
[1] Lower bound on overall LogEvidence: -1.4375e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.3 0.4 0.3
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.35 0.3 0.35
We have the following parameters for the data array:
```

```
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 351.378004511408
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -158.8043
[1] 1265.9246 -158.7461
[1] 1564.6508 -158.7281
[1] 1808.2877 -158.7178
[1] 1715.2267 -158.7214
[1] 1899.8497 -158.7146
[1] 1956.4381 -158.7128
[1] 1991.4116 -158.7117
[1] 2013.0265 -158.7111
[1] 2026.3852 -158.7107
[1] 2034.6413 -158.7104
[1] 2039.7439 -158.7103
[1] 2042.8974 -158.7102
[1] 2044.8464 -158.7101
[1] 2046.0510 -158.7101
[1] 2046.7954 -158.7101
[1] 2047.2555 -158.7101
[1] 2047.5889 -158.7101
[1] 2047.5889 -158.7101
[1] 2047.5889 -158.7101
[1] Hyperparameter: 2047.58890859597
[1] Lower bound on overall LogEvidence: -1.5871e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.26 0.48 0.26
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
```

```
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.27 0.46 0.27
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
```

```
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.28 0.44 0.28
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.29 0.42 0.29
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
```

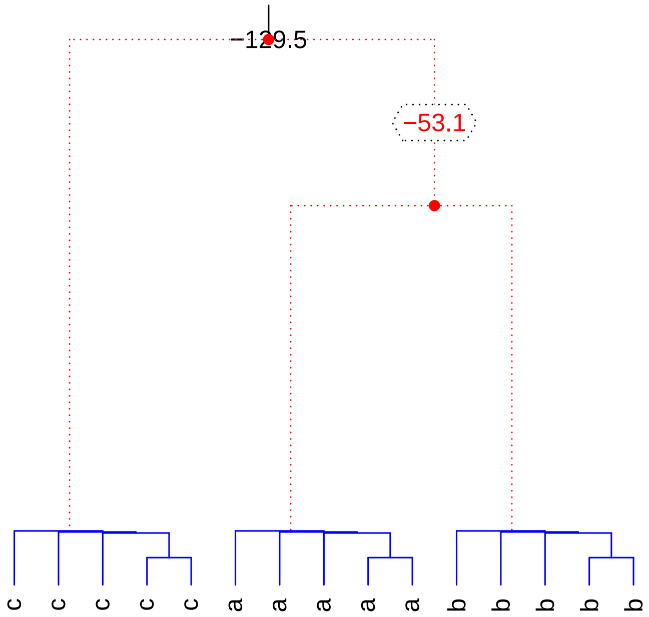
```
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.3 0.4 0.3
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
```

```
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.31 0.38 0.31
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.32 0.36 0.32
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***
```

```
Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.33 0.34 0.33
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
```

```
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

DATA DISCRETISATION
-------------------
Percentiles: 0.34 0.32 0.34
We have the following parameters for the data array:
nGenes:       15
nExperiments: 10
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 358.011081430942
(Need to add this to the model logEvidence)
-------------------
[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]  782.5754 -163.5695
[1] 1265.9246 -163.5086
[1] 1564.6508 -163.4898
[1] 1808.379 -163.479
[1] 1715.2829 -163.4827
[1] 1899.9058 -163.4756
[1] 1956.4728 -163.4737
[1] 1991.4331 -163.4726
[1] 2013.0397 -163.4719
[1] 2026.3934 -163.4715
[1] 2034.6464 -163.4713
[1] 2039.7470 -163.4711
[1] 2042.899 -163.471
[1] 2044.848 -163.471
[1] 2046.0517 -163.4709
[1] 2046.7959 -163.4709
[1] 2047.2558 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] 2047.5892 -163.4709
[1] Hyperparameter: 2047.58919090975
[1] Lower bound on overall LogEvidence: -1.6347e+02
[1] *******************

OPTIMISED DISCRETISATION
------------------------
```

```
Percentiles: 0.3 0.4 0.3
LogEvidence: 194.5402

> discreteData <- DiscretiseData(t(newData), percentiles=percentiles)

DATA DISCRETISATION
-------------------
Percentiles: 0.3 0.4 0.3
We have the following parameters for the data array:
nGenes:       10
nExperiments: 15
***Please check that these are the right way round! (it affects the discretisation)***

Discretisation logEvidence: 98.7600287840702
(Need to add this to the model logEvidence)
-------------------

> discreteData <- t(discreteData)
> hc3          <- bhc(discreteData, itemLabels, verbose=TRUE)

[1] Running Bayesian Hierarchical Clustering....
[1] "DataType: multinomial"
[1] Optimising global hyperparameter...
[1]    0.8411863 -115.5350443
[1]    1.283814 -121.371320
[1]    0.5078529 -111.1485633
[1]    0.5078529 -111.1485633
[1]    0.5078529 -111.1485633
[1] Hyperparameter: 0.507852925225962
[1] Lower bound on overall LogEvidence: -1.1115e+02
[1] *******************

>
```



The plot shows a simple example dendrogram. (note that the structure is quite distinctive; this may the be the result of discretising and analysing a small data-set).