Guidelines interpretation GORIC(A) benchmark output

Rebecca M. Kuiper and Leonard Vanbrabant

19 August 2024

Contents

Introduction

To aid qualifying or labeling the height/size of support for the preferred hypothesis, one can inspect case-specific benchmarks for i) the GORIC(A) weights and ii) the ratio of GORIC(A) weights of the preferred hypothesis versus the other hypotheses in the set. This is especially helpful when you found support for one hypothesis, as opposed to support for the overlap or a boundary/border of two or more hypotheses (see 'Guidelines_output_GORIC'). These benchmarks can also help in noticing support for overlap (if you forgot to check the fit values), since the benchmarks will show that there is maximum/bounded support; as will exemplified in the second example below.

The benchmarks are based on (user-specified) sets of population parameter values (as is also used in a power analysis when performing a null hypothesis test). For an ANOVA, the set of population parameter values can be obtained by specifying a population effect size (i.e., Cohen's f) and a specific ratio of population means (which by default is the ratio of sample means, that is, the ratio of means found in the data). For all types of statistical models (including ANOVA), one can specify the set of population parameter values themselves.
We advise on using the null as a reference population, that is, assuming that one or more equalities (instead of the hypothesized inequalities) hold true in the population. Then, you obtain insight in how (un)likely your results are when one or more equalities are true. The extremer your finding, the more support for your informative, theory-based, inequality-constrained hypothesis.
By default, the null population in which the effect size is – or parameter values are – zero ('No-effect') and a population based on the sample effect size – or sample parameter values – ('Observed') is used.

Additionally, the function renders benchmarks for the ratio of log-likelihood (loglik) weights of the preferred hypothesis versus the other hypotheses in the set and for the differences in loglik values of the preferred hypothesis versus the other hypotheses in the set. Note that the log-likelihood benchmarks (under a null population) give insight into the distribution of loglik weight ratios and of the loglik differences in case some or all of the group means or model parameters would be the same or zero (depending in the type of null). When you calculate these loglik benchmarks for a population in which one or more of the hypothesized inequalities is set to an equality, these loglik benchmarks are helpful in determining support for a boundary hypothesis (if

of interest). This will be illustrated in the third example below. Bear in mind that research (e.g., a simulation study) is needed to obtain more insight into whether and how well the log-likelihood benchmarks work.

GORIC(A) weights benchmarks

The GORIC(A) weights benchmarks come from several percentiles of sets of GORIC(A) weights assuming that the specified set of population values is true (for the sample size under consideration). More specifically, the benchmarks are, currently, based on the 5th, 35th, 50th, 65th, and 95th percentiles of the GORIC(A) weights for the preferred hypothesis and of the ratios of the GORIC(A) weights for the preferred hypothesis versus the other hypotheses. Bear in mind that research (e.g., a simulation study) is needed to obtain more insight into how well these choices work.
Notably, you can a-priori decide on what percentiles you believe reflect the different types of support, and **pre-register** that together with your informative hypothesis/-es.

You can compare your GORIC(A) weight and ratios of GORIC(A) weights of the preferred hypothesis to these benchmarks to make a conclusion regarding the strength of support for the hypothesis (given the assumed set of population parameter values and given the sample size). If the benchmarks show a maximum/bounded support (see second example below), there is support for the overlap of two or more hypotheses (which is also signaled by equal log-likelihood values). Otherwise, the GORIC(A) weights (ratios) benchmarks can be used to qualify the height of support (see first example below).

How to use benchmarks

Labelling

We are not in favor of cut-off points (or 'surrounding anchors'), but we need them when we want to label the height of support via the benchmarks, we propose the following:

Benchmark (percentile)	Height support
below 5th	no support
between 5th and 35th	low support
between 35th and 65th	medium support
between 65th and 95th	high (compelling) support
over 95th	very large (tremendous) support

We advise on using some kind of null model as the assumed population. Then, you can see how extreme your finding is (or not) for this null population. The extremer your finding, the more support for your informative, theory-based, inequality-constrained hypothesis.

You can of course use other percentile levels (for which you think the finding is said to be extreme enough etc), but do make sure to define these before seeing the data, and preferably also pre-register them (together with your informative hypothesis/-es).

Use minimum effect

You may want to use a minimum effect. One option, the one we also advise, is to specify your hypothesis such that it inspects, for example, minimum differences between parameters (e.g., $\mu_1 - \mu_2 > 0.2$ instead of $\mu_1 > \mu_2$, that is, $\mu_1 - \mu_2 > 0$).

A second option, although not something we advise on doing, is to investigate benchmarks using a minimum effect size (or looking at multiple ones). In the function, you should specify the population parameter values (reflecting specific effect sizes). In case of an ANOVA model, you can also specify the effect size level(s) (for Cohen's f). Note that the benchmarks differ when you assume different effect sizes or different population parameter estimates.

Sensitivity analysis

If of interest, as a sensitivity analysis, you can calculate the benchmarks for multiple sets of population parameter values (or population effect sizes). Note that this may also complicate drawing conclusions (especially when the assumed sets of population parameter values differ much, like when using multiple effect size heights). We advise on doing this for multiple null populations (setting some to all of the inequality restrictions to equalities).

If your preferred hypothesis does not have the highest fit and you want to inspect benchmarks, we advise on inspecting multiple ratios of population parameter estimates, where some are in agreement with your hypothesis and others in agreement with the data.

Defaults

Once more, we advise on inspecting populations in which some of the inequalities of your hypothesis/-es of interest are set to equalities (especially if the log-likelihood values seem to be close). This way, you can also check for the support of a boundary hypothesis; as discussed in the third example below.

By default, two populations are used: i) a null population in which the population effect size is – or population parameter values are – set to 0 and ii) a population based on the observed effect size – or observed parameter values. One can overrule this by using the `pop_es` or `pop_est` argument, respectively.

Examples

Next, we will discuss two ANOVA examples. More specifically, We will inspect the case-specific benchmarks values (using the ratio of means as in the data). We will look at
- an ANOVA example where we evaluate $H_1 : \mu_1 > \mu_2 > \mu_3$ versus its complement, and H_1 is true;
- an ANOVA example where we evaluate two overlapping hypotheses, namely $H_1 : \mu_1 > \mu_2 > \mu_3$ and $H_2 : \mu_1 > \mu_2,\ \mu_3$, together with the unconstrained, and H_1 is true (and thus the others are as well, but they are not the most parsimonious one).

Later on (in another section), We will also discuss the following example:
- an ANOVA example where we evaluate $H_1 : \mu_1 > \mu_2 > \mu_3$ versus its complement, and the border $\mu_1 = \mu_2 > \mu_3$ is true. Notably, here, both hypotheses are true, H_1 is the most parsimonious one, and we want to conclude that the border is true.

For a description of interpreting GORIC(A) output, see 'Guidelines_output_GORIC' (https://github.com/rebeccakuiper/Tutorials).

Example 1 (ANOVA): H_1 vs its complement

```
# H1 vs complement - H1 is true
H1 <- "D1 > D2 > D3" # mu1 > mu2 > mu3

# Apply GORIC #
set.seed(123)
results_1c <- goric(fit, hypotheses = list(H1), comparison = "complement")
results_1c
```

```
restriktor (0.5-85): generalized order-restricted information criterion:

Results:
        model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1          H1  -155.075    2.833  315.816           0.704            0.697          0.845
2  complement  -155.939    3.667  319.212           0.296            0.303          0.155
---
The order-restricted hypothesis 'H1' has 5.46 times more support than its complement.
```

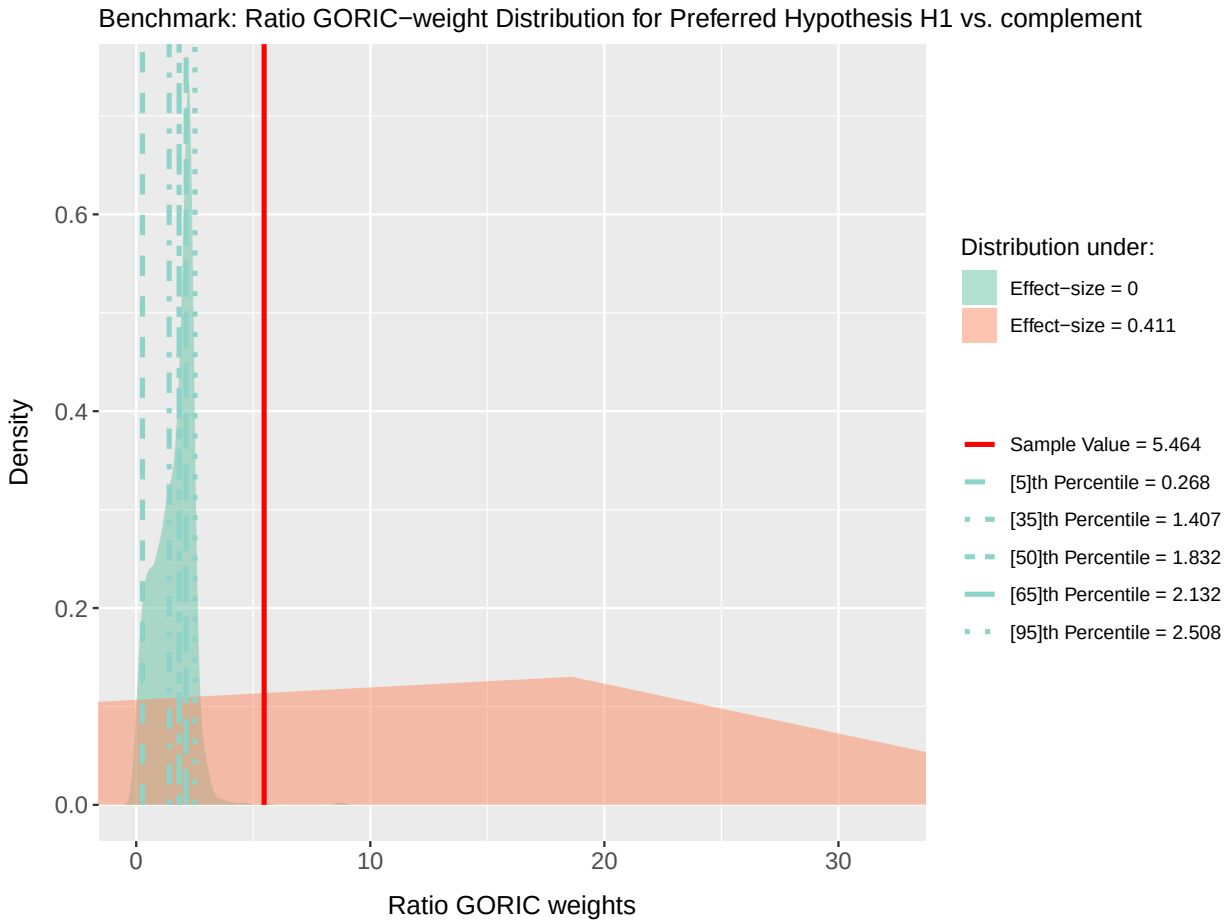
```r
# Benchmarks based on null
benchmarks_1c <- benchmark(results_1c, model_type = "means", ncpus = 8)
```

```
Calculating means benchmark for effect-size = 0
Calculating means benchmark for effect-size = 0.411
```

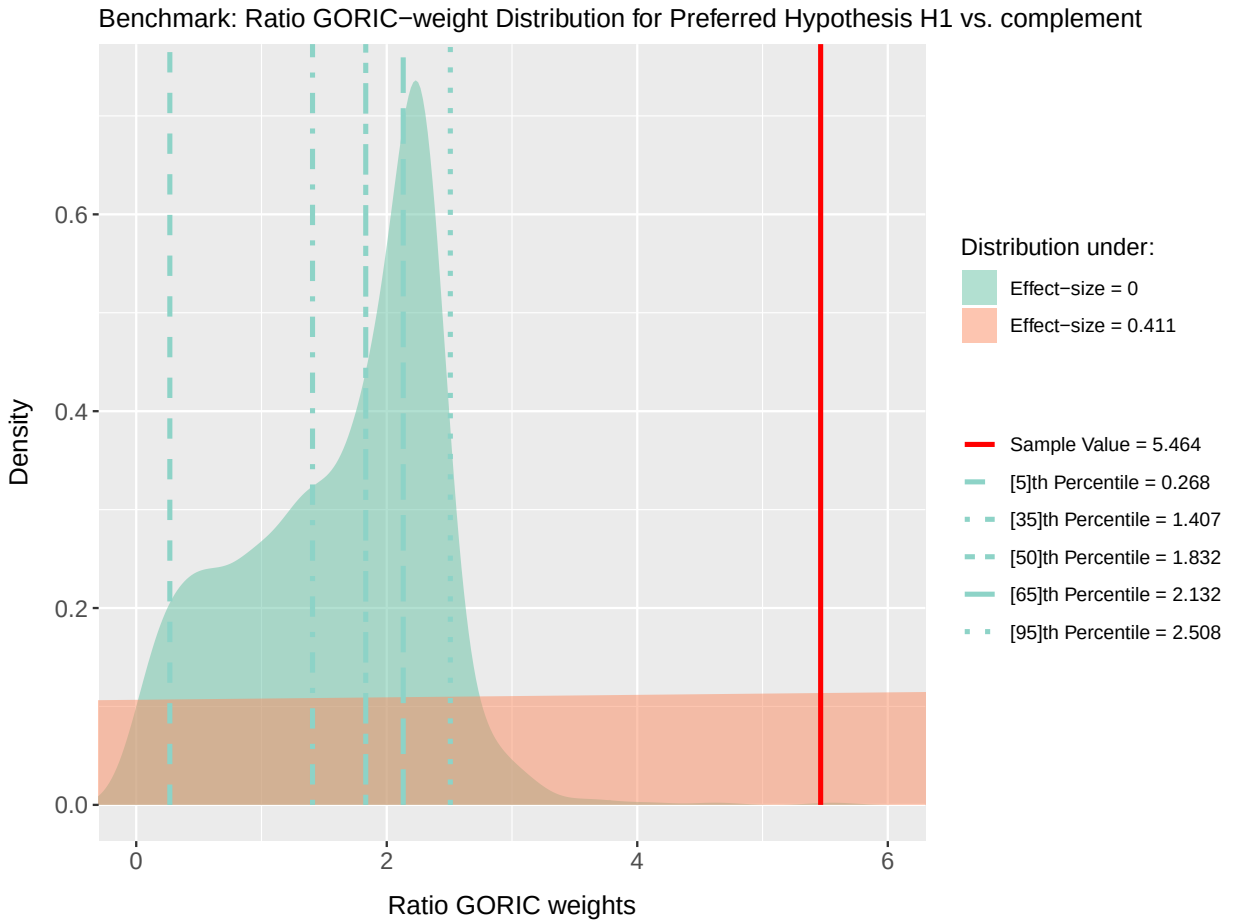
```r
#benchmarks_1c # use in R file
print(benchmarks_1c, color = FALSE) # use in Rmd file, since Rmd cannot deal with colored text
```

```
Benchmark Results
-------------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.155
Number of Groups: 3
Group Sizes: 40, 40, 40
Ratio of Population Means: 4.322, 2.000, 1.000
Population Effect-Sizes (Cohens f): 0.000, 0.411
Observed Effect-Size (Cohens f): 0.411


=============================================================================================
Benchmark: Percentiles of Ratio-of-GORIC-weights for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population effect-size = 0
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 5.464  0.268 1.407 1.832 2.132 2.508

Population effect-size = 0.411
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 5.464  2.130 3.286 4.841 7.586 55.792
```

```r
plot(benchmarks_1c)
```




```
plot(benchmarks_1c, x_lim = c(0, 6))
```



From the `goric()` output, you can conclude that there is support for $H_1 : \mu_1 > \mu_2 > \mu_3$, and H_1 is 0.85 / 0.15 $\approx$ 5.46 times more supported than its complement. The probability that H_1 is not the best is 15.47% (namely, the goric.weight for the complement of H1, which is also given in the benchmark output by `$error_prob_pref_hypo_name`, that is, the `Error probability Preferred Hypothesis vs. complement`). This already gives insight into the (un)certainty and, therefore, helps in qualifying the results. Additionally, the benchmarks can help:

Based on the benchmarks, you can check how plausible your finding is (given the assumed population parameters and given the sample size). If you want to compare your results with the situation in which one or more equalities hold true, use a null population (here, an effect size of 0, indicating that the three means are equal). Then, you obtain insight into how (un)likely / how extreme your finding (based on inequalities) is.

When assuming that there is no effect in the population (i.e., under the null), that is, all three group means are equal, the ratio of GORIC weights of H_1 versus its complement (i.e., 5.46) is larger than the 95th percentile (i.e., 2.51). Hence, our finding is very extreme if the null would be true. Using the table with cut-off values above, this indicates that there is very large (tremendous) support, when assuming no population effect size (given a group size of 40 for each of the 3 groups).

Log-likelihood check:
Before inspecting the height of the support, one may want to establish whether there is support for the overlap or boundary of hypotheses. Since we evaluate an informative hypothesis H_1 versus its complement, we should check whether there is support for a boundary hypothesis (in which one or more inequalities in H_1 is replaced by an equality). For this, one should inspect the log-likelihood / fit values of the hypotheses. When these are close (i.e., the ratio of loglik weights is close to 1 or the difference in loglik values is close to 0), then there is support for (one of their) boundaries. In this case, the loglik values are -155.07 and

-155.94, with corresponding loglik.weights of 0.7 and 0.3 (and thus a difference of approx. 0.86 and a ratio of approx. 2.37). Since it is hard to judge what is close, one can inspect the benchmarks for the ratio of log-likelihood (loglik) weights and/or for the differences in log-likelihood values. This should then be done for a null population in which such a boundary is true. We discuss this in the next section.
For now, we assume that the loglik values are not close.

Population information:
In the data generation, we used a ratio of population means of 3:2:1; implying that H_1 is correct. More specifically, We used population mean values of approximately 0.92, 0.61, and 0.31. This implies that Cohen's f is .25; thus, there is a medium population effect size (which are in the same order as hypothesized). We then sampled 40 observations for each of the three groups, ran an ANOVA (with three groups), and applied the GORIC. Note that Cohen (1992) suggest that a minimum group size of 52 is needed to find a medium effect when doing null hypothesis testing.
Notable, the sample/observed effect size is .411 (with sample means of 1.12, 0.51, and 0.25), which can be seen as a high effect size. This also explains why we, despite the medium population effect size, find tremendous support for our hypothesis.
When we would sample more observations, the GORIC(A) weight for H_1 converges to 1 (denoting full support for H_1). Note that the benchmarks for the GORIC(A) weight for H_1 will remain the same for a null population and will go to 1 for a non-null population. Note that the error probability then goes to 0, and that the ratio of GORIC(A) weights of H_1 versus its complement then goes to infinity.

Example 2 (ANOVA): Overlapping hypotheses

```
# H1, H2, and unconstrained (default) - subset/overlap, that is, H1 is true
H1 <- "D1 > D2 > D3"                # mu1 > mu2 > mu3
H2 <- "D1 > D2" # H2: D1 > D2, D3  # mu1 > mu2,  mu3

# Apply GORIC #
set.seed(123)
results_12u <- goric(fit, hypotheses = list(H1, H2))
results_12u
```

```
restriktor (0.5-85): generalized order-restricted information criterion:

Results:
           model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1             H1  -155.075    2.833  315.816           0.333            0.548          0.548
2             H2  -155.075    3.500  317.149           0.333            0.281          0.281
3  unconstrained  -155.075    4.000  318.149           0.333            0.171          0.171
```

```
round(results_12u$ratio.gw, 3)
```

```
              vs. H1 vs. H2 vs. unconstrained
H1             1.000  1.948             3.211
H2             0.513  1.000             1.649
unconstrained  0.311  0.607             1.000
```

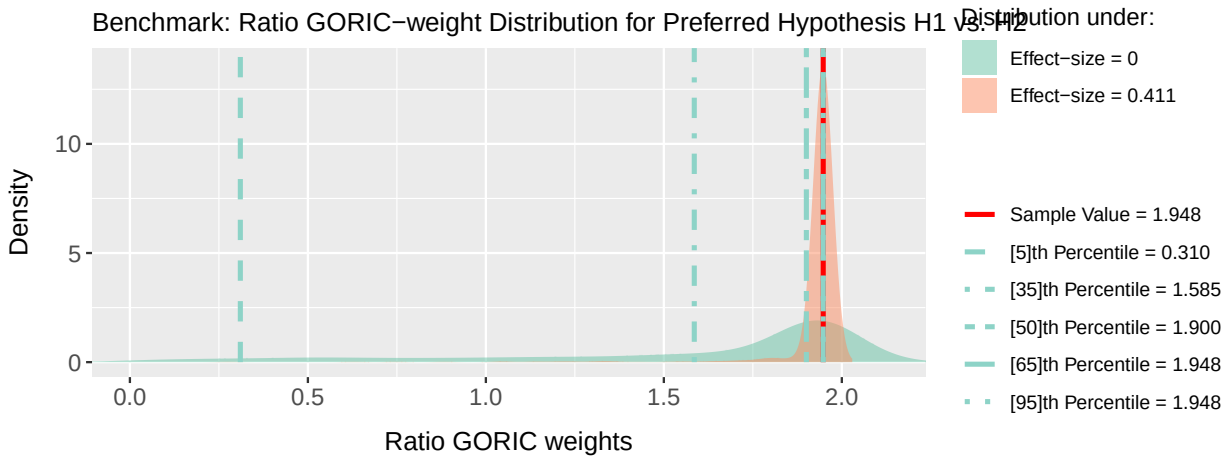
```
# Benchmarks
benchmarks_12u <- benchmark(results_12u, model_type = "means", ncpus = 8)
```

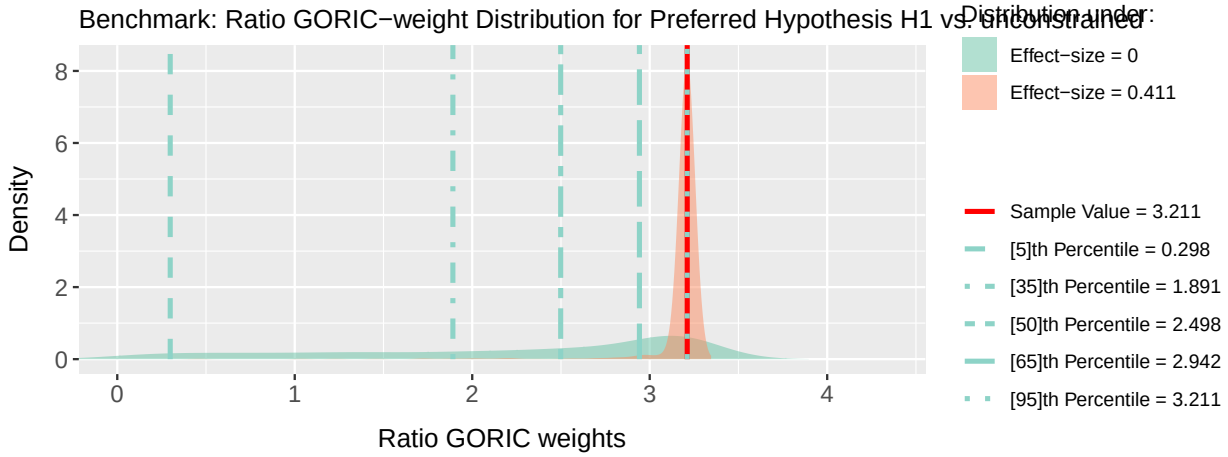
```
Calculating means benchmark for effect-size = 0
Calculating means benchmark for effect-size = 0.411
```

```
#benchmarks_12u # R file
print(benchmarks_12u, color = FALSE) # Rmd file
```

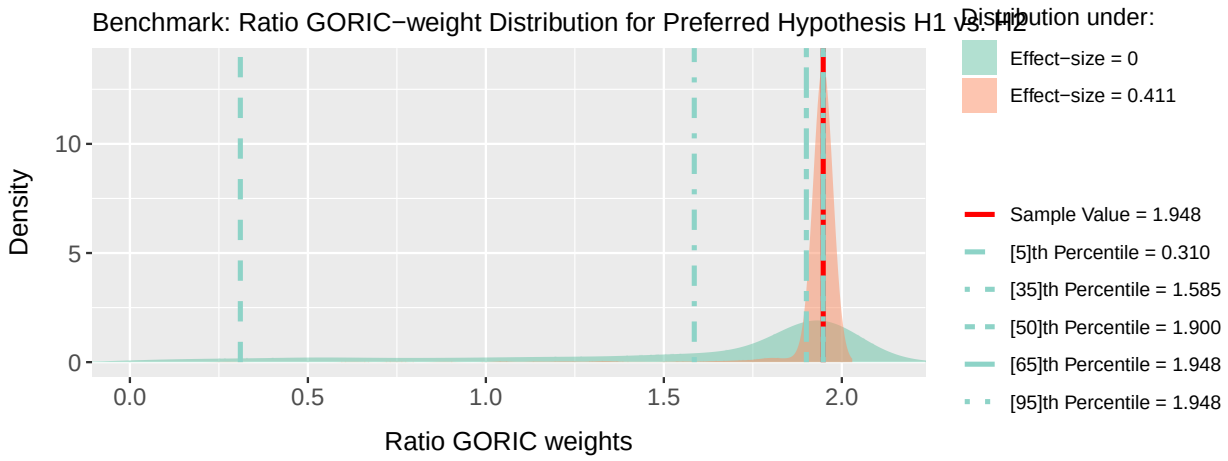
```
Benchmark Results
-----------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.155
Number of Groups: 3
Group Sizes: 40, 40, 40
Ratio of Population Means: 4.322, 2.000, 1.000
Population Effect-Sizes (Cohens f): 0.000, 0.411
Observed Effect-Size (Cohens f): 0.411


=========================================================================================
Benchmark: Percentiles of Ratio-of-GORIC-weights for the Preferred Hypothesis 'H1'
-----------------------------------------------------------------------------------------
Population effect-size = 0
                     Sample 5%    35%   50%   65%   95%
H1 vs. H2            1.948  0.310 1.585 1.900 1.948 1.948
H1 vs. unconstrained 3.211  0.298 1.891 2.498 2.942 3.211

Population effect-size = 0.411
                     Sample 5%    35%   50%   65%   95%
H1 vs. H2            1.948  1.877 1.948 1.948 1.948 1.948
H1 vs. unconstrained 3.211  3.091 3.211 3.211 3.211 3.211
```

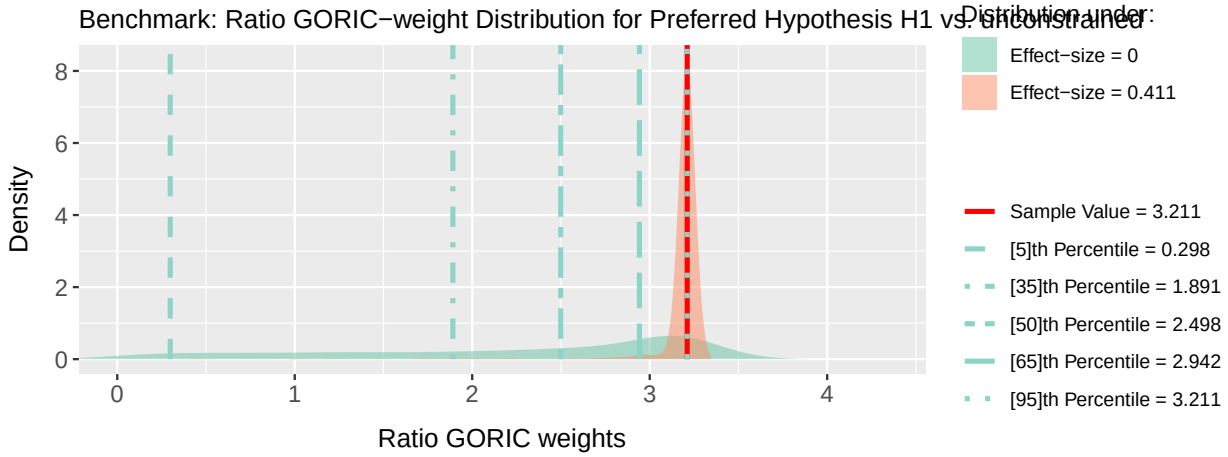
```
plot(benchmarks_12u, output_type = "rgw")
```



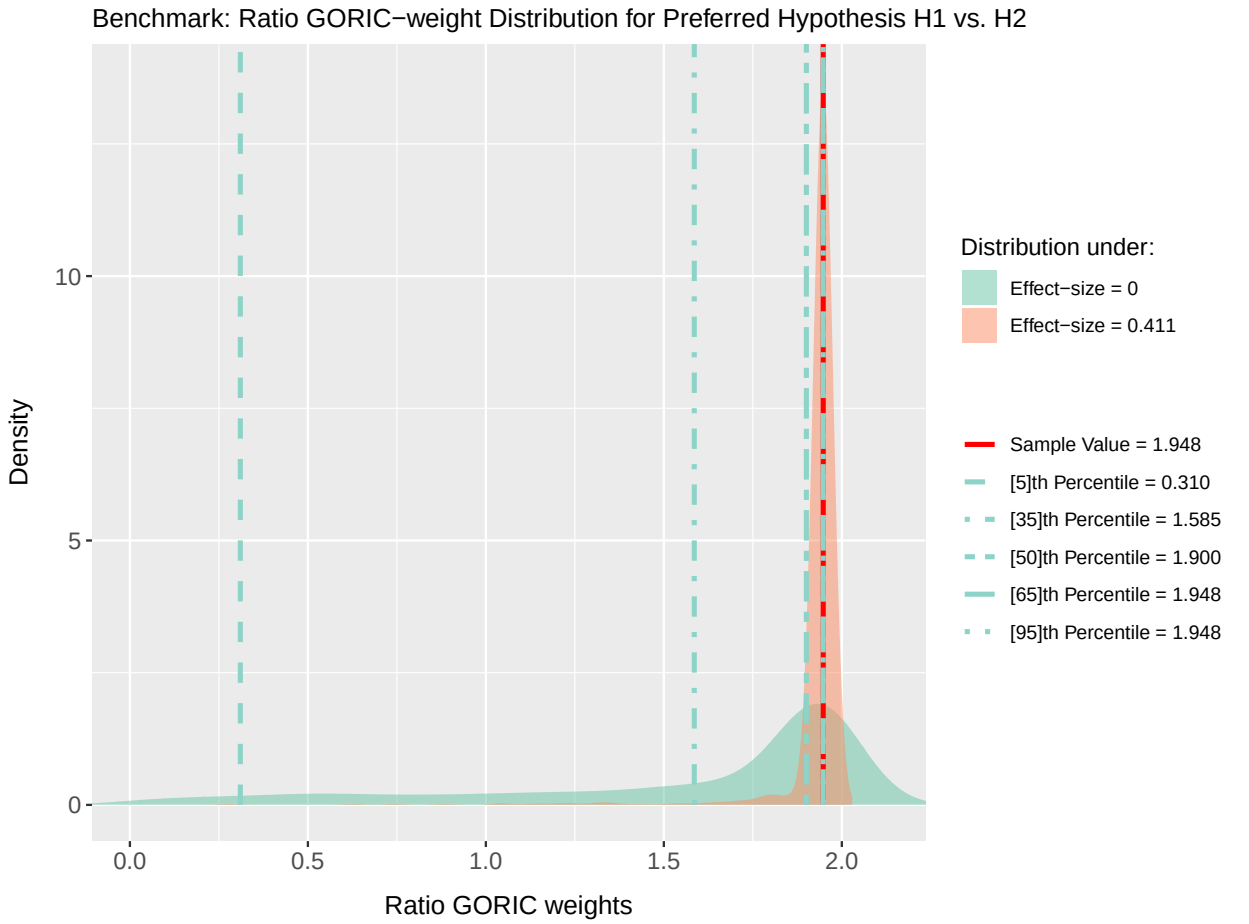



```
plot_out <- plot(benchmarks_12u) # save all plots in object plot_out
```






```
plot(plot_out$grobs$`H1 vs. H2`) # call separate plots
```



From the GORIC output, we can see that both hypotheses are not weak and that H_1 is the best. Additionally, we can see that the log-likelihood values are exactly the same, and thus there is support for the overlap of the hypotheses which is H_1 here (since H_1 is a subset of H_2).

If we would now check the GORIC(A) weights benchmarks (especially for non-nulls), we would find again that there is support for the overlap (here, H_1): Namely, the benchmarks for multiple percentiles (here, even all percentiles) have the same value as our finding, that is, the sample value.

Notably, if there is support for the overlap or boundary, it is not meaningful to use the benchmarks to label the height of support for H_1, since The ratio of support reached its maximum. Hence, we do not proceed with labeling the height of support (by comparing the sample value to the benchmarks under the null population). Instead, we conclude that there is support for the overlap.
If of interest, one can next inspect the support for the overlap, here H_1, versus its complement. When that does not result in support for a boundary, then the height of the support for H_1 versus its complement can be labelled using the benchmarks. This is then helpful information for future research: Generating a new theory and having another value to compare future results to.

Log-likelihood check:
Here, one clearly finds that the loglik / fit values are exactly the same. Hence, the ratio of loglik weights is exactly 1 and the difference in loglik values is exactly 0. Consequently, there is support for the overlap. In this case, we do not need to check the log-likelihood benchmarks.
Note that the log-likelihood benchmarks (under a null population) give insight into the distribution of loglik weight ratios and of the loglik differences in case some or all of the group means would be the same.

Population information:
In the data generation, We used a ratio of population means of 3:2:1; implying that H_1 is correct. More

specifically, We used population mean values of approximately 0.92, 0.61, and 0.31. This implies that Cohen's f is .25; thus, there is a medium population effect size (which are in the same order as hypothesized). We then sampled 40 observations for each of the three groups, ran an ANOVA (with three groups), and applied the GORIC. Note that Cohen (1992) suggest that a minimum group size of 52 is needed to find a medium effect when doing null hypothesis testing.
When We would sample more observations, it does not (really) affect the GORIC(A) weights for H_1, H_2, and the unconstrained: It will converge to the bounds (i.e., the maximum support) it can take on. The benchmarks for the GORIC(A) weights will also attain the maximum value as will the ratio of weights; and it will for each positive population effect size. For a null population, the GORIC(A) weight benchmarks will remain the same.

Log-likelihood benchmarks

Before inspecting the height of the support, one may want to establish whether there is support for the overlap / boundary of hypotheses. Since we evaluate an informative hypothesis H_1 versus its complement, we should check whether there is support for a boundary hypothesis (in which one or more inequalities in H_1 is replaced by an equality). For this, one should inspect the log-likelihood / fit values of the hypotheses. When these are close (i.e., the ratio of loglik weights is close to 1 or the difference in loglik values is close to 0), then there is support for (one of their) boundaries.

Since it is hard to judge what is close, one can inspect the benchmarks for the ratio of log-likelihood (loglik) weights and for the differences in log-likelihood values. This should then be done for a population in which such a boundary is true.

Next, We will inspect two examples: one in which the border is true (Example 3) and one where it is not (Example 1 continued). We will first use a group size of 40, like in the example above; afterwards, We will inspect a higher sample size to obtain insight into the asymptotic properties of the loglik weights.

For now, We will use that the loglik values are said to be the same if the loglik differences are in between the 5th and 95th percentiles of the loglik benchmarks for a null population. You can of course use a narrower range to be more strict.

Remarks:
- The loglik benchmarks need to be more properly investigated.
- The benchmark function contains two types of loglik benchmarks:
* the loglik ratios (i.e., ratio of loglik weights; `output_type = "rlw"`), which can take on values between 0 and infinity, where 1 means that the loglik values are the same;
* the loglik differences (`output_type = "ld"`), which can take on values between minus infinity and infinity and where 0 means that the loglik values are the same;
- One could think about doing a likelihood ratio test (LRT), using a Chi-square distribution or, because of the inequality restrictions, a Chi-bar-square distribution. However, the LRT test is appropriate for nested models, while we are interested in models that are non-nested and even non-overlapping (except for the boundary).

Example 3 (ANOVA): Border is true

```
# H1 vs complement - border (nl., mu1 = mu2 > mu3) is true
H1 <- "D1 > D2 > D3" # mu1 > mu2 > mu3

# Apply GORIC #
set.seed(123)
results_1c_border <- goric(fit_border, hypotheses = list(H1), comparison = "complement")
results_1c_border
```

```
restriktor (0.5-85): generalized order-restricted information criterion:
```

```
Results:
        model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1          H1  -150.549    2.833  306.765           0.482            0.697          0.682
2  complement  -150.477    3.667  308.288           0.518            0.303          0.318
---
The order-restricted hypothesis 'H1' has 2.14 times more support than its complement.
```

Before we inspect the height of support for the preferred hypotheses, we should check whether there is support for the boundary/border of the two (non-overlapping) hypotheses. By eyeballing, We believe the log-likelihood values are close. To obtain better evidence for the closeness of the loglik values, We will use the log-likelihood benchmarks functions for several specifications of null populations.

```
# Default null (when using `model_type = "means"`)
## Loglik benchmarks based on default null / no effect sizes, that is,
## setting all three means equal in the population
#benchmarks_1c_border <- benchmark(results_1c_border, model_type = "means", ncpus = 8)
## loglik diff
##print(benchmarks_1c_border, output_type = "ld") # in R file
#print(benchmarks_1c_border, output_type = "ld", color = FALSE) # in Rmd file
#plot(benchmarks_1c_border, output_type = "ld")
## ratio loglik weights
##print(benchmarks_1c_border, output_type = "rlw") # in R file
#print(benchmarks_1c_border, output_type = "rlw", color = FALSE) # in Rmd file
#plot(benchmarks_1c_border, output_type = "rlw")
```

```
# Specifying multiple null populations, that is,
# using all possibilities of setting inequalities to equalities.
# Here, we will use the default `model_type` (i.e., "asymp") which takes population parameter values (i
est <- coef(fit_border)
pop_est <- matrix(c(
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3]),
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  mean(est[1:2]), est[2], mean(est[1:2]),
                  est[1], mean(est[2:3]), mean(est[2:3]),

                  est[1], est[2], est[3]
                  ),
                  byrow = TRUE, ncol = length(est))
rownames(pop_est) <- c("PE_123eq", "PE_12eq", "PE_13eq", "PE_23eq", "Observed")
benchmarks_1c_border_allpos <- benchmark(results_1c_border, pop_est = pop_est, ncpus = 8)
```

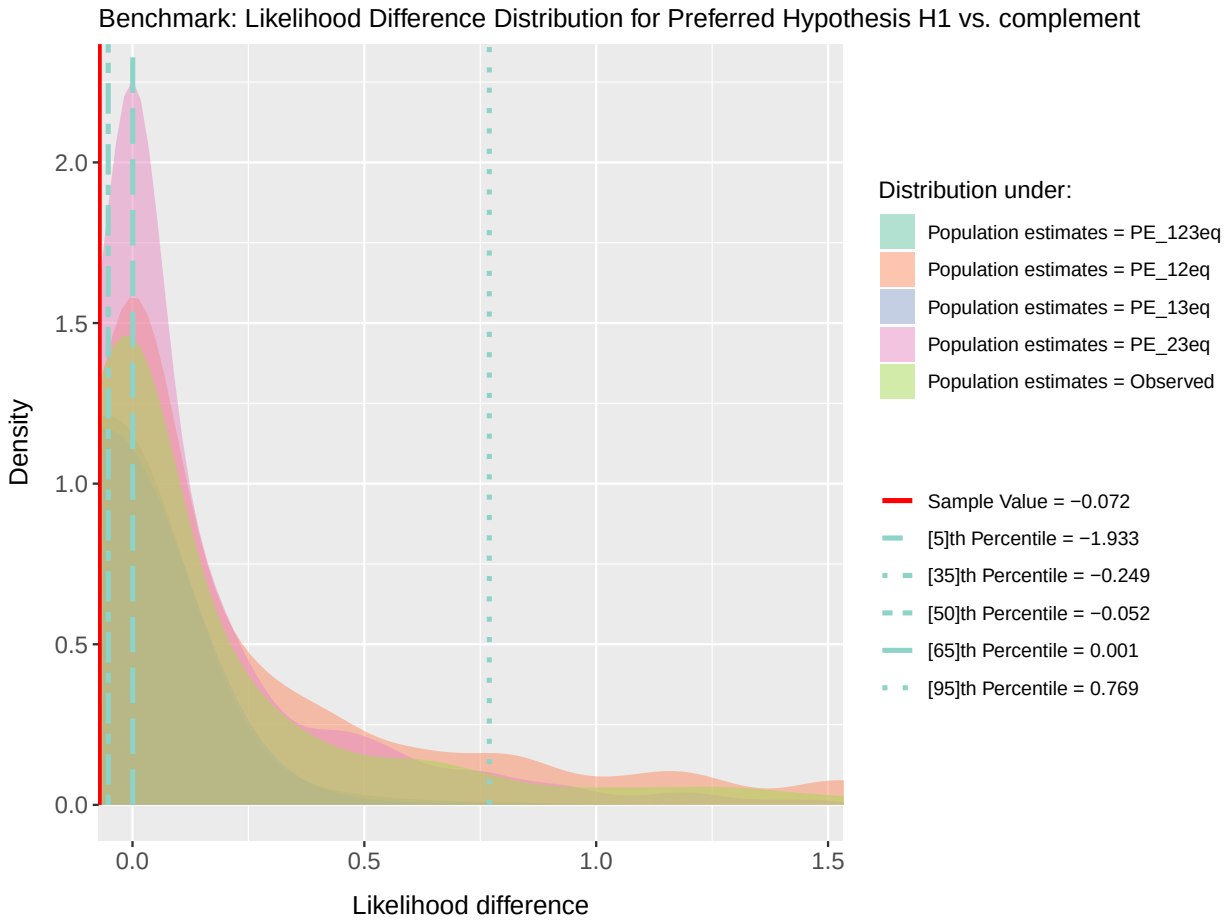
```
Calculating asymptotic benchmark for population estimates = PE_123eq
Calculating asymptotic benchmark for population estimates = PE_12eq
Calculating asymptotic benchmark for population estimates = PE_13eq
Calculating asymptotic benchmark for population estimates = PE_23eq
Calculating asymptotic benchmark for population estimates = Observed
```

```
# loglik difference
#print(benchmarks_1c_border_allpos, output_type = "ld") # R file
print(benchmarks_1c_border_allpos, output_type = "ld", color = FALSE) # Rmd file
```

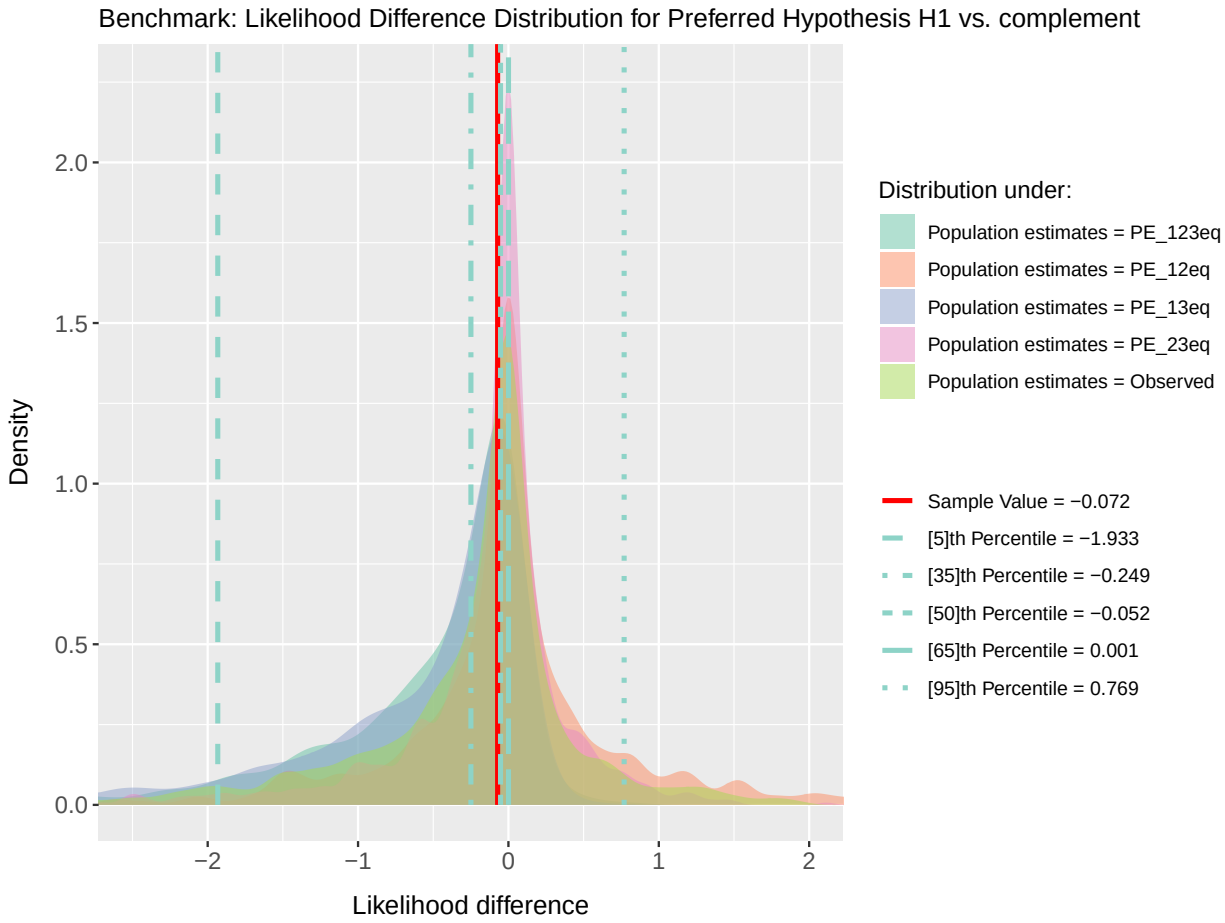
```
Benchmark Results
-----------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.318
```

```
Sample Size: 120
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.696  0.696  0.696
PE_12eq   0.913  0.913  0.264
PE_13eq   0.913  0.949  0.913
PE_23eq   0.877  0.606  0.606
Observed  0.877  0.949  0.264


=============================================================================================
Benchmark: Percentiles of Differences in Likelihood Values for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement -0.072 -2.278 -0.547 -0.265 -0.090 0.067

Population estimates = PE_12eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement -0.072 -1.412 -0.087 -2.062e-04 0.074 1.272

Population estimates = PE_13eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement -0.072 -2.475 -0.569 -0.261 -0.094 0.062

Population estimates = PE_23eq
                  Sample 5%     35%    50%    65%   95%
H1 vs. complement -0.072 -1.326 -0.121 -0.012 0.017 0.590

Population estimates = Observed
                  Sample 5%     35%    50%    65%       95%
H1 vs. complement -0.072 -1.933 -0.249 -0.052 5.572e-04 0.769
```

```
plot(benchmarks_1c_border_allpos, output_type = "ld")
```




```
plot(benchmarks_1c_border_allpos, output_type = "ld", x_lim = c(-2.5,2))
```

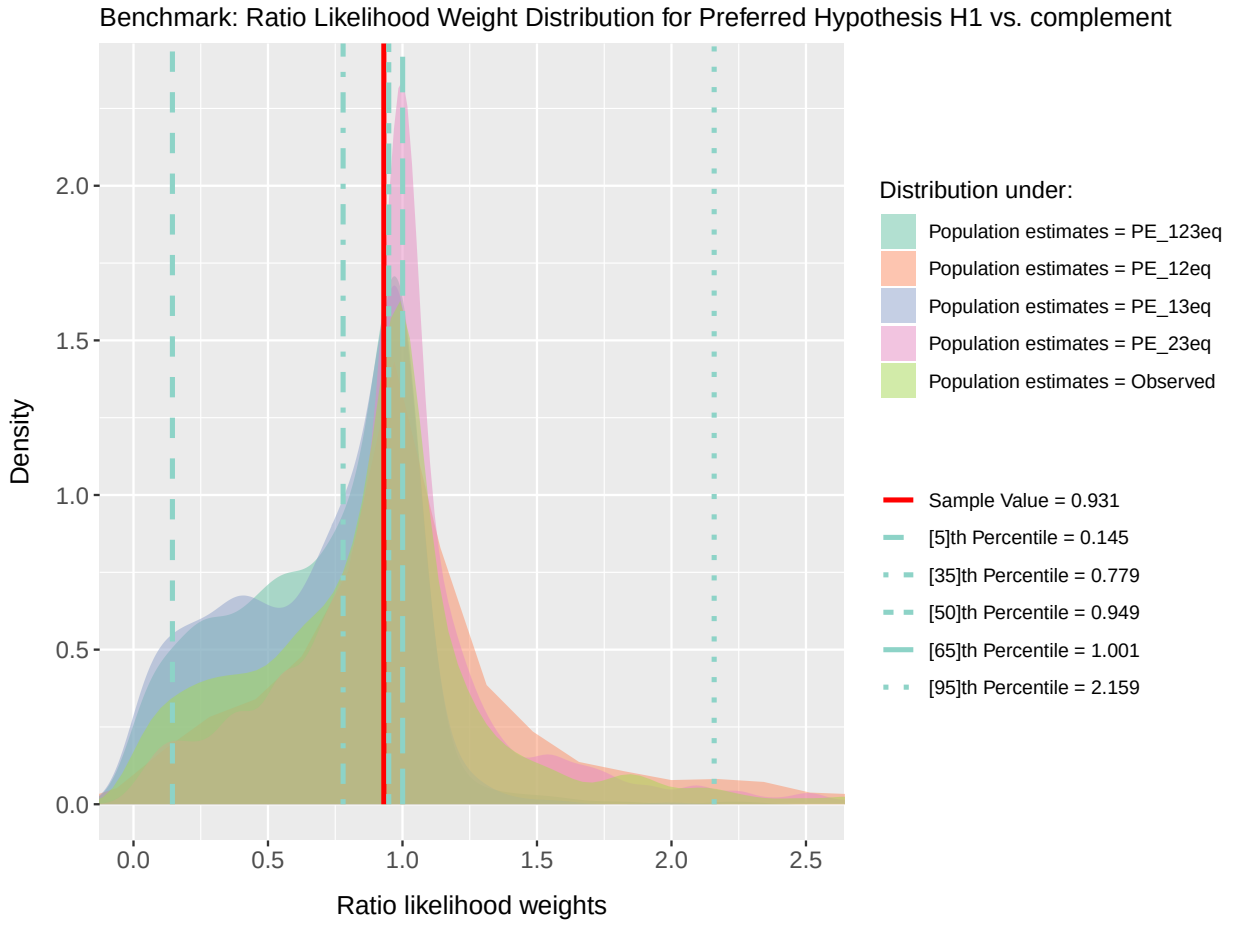



```
# ratio of loglik weights
#print(benchmarks_1c_border_allpos, output_type = "rlw") # R file
print(benchmarks_1c_border_allpos, output_type = "rlw", color = FALSE) # Rmd file
```

```
Benchmark Results
-------------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.318
Sample Size: 120
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.696  0.696  0.696
PE_12eq   0.913  0.913  0.264
PE_13eq   0.913  0.949  0.913
PE_23eq   0.877  0.606  0.606
Observed  0.877  0.949  0.264


=========================================================================================
Benchmark: Percentiles of Ratio-of-likelihood-weights for the Preferred Hypothesis 'H1'
-----------------------------------------------------------------------------------------
Population estimates = PE_123eq
```

```
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 0.931  0.102 0.578 0.767 0.914 1.069

Population estimates = PE_12eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 0.931  0.244 0.916 1.000 1.077 3.567

Population estimates = PE_13eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 0.931  0.084 0.566 0.770 0.911 1.063

Population estimates = PE_23eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 0.931  0.265 0.886 0.988 1.017 1.804

Population estimates = Observed
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 0.931  0.145 0.779 0.949 1.001 2.159
```

```
plot(benchmarks_1c_border_allpos, output_type = "rlw")
```



In this example, the difference in log-likelihood values and the ratio of log-likelihood weights lies between the 5th and the 95th percentile of the corresponding benchmarks (for all null populations). Based on this, we conclude that the log-likelihood values of H_1 and its complement are close (i.e., their loglik weight ratio is close to 1 and their difference is close to 0). Possibly, one can also conclude that there is a bit more support

for a boundary hypothesis in which either the first two or the last means are the same (because the sample value is less extreme for those orderings).

Population information:
In the data generation, We used a ratio of population means of 2.5:2.5:1; implying that the boundary of H_1 and its complement is correct (and that H_1 is preferred over its complement, since it is more parsimonious). More specifically, We used population mean values of approximately 0.88, 0.88, and 0.35. This implies that Cohen's f is .25; thus, there is a medium population effect size. We then sampled 40 observations per group, ran an ANOVA (with three groups), and applied the GORIC.
When We would sample more observations, the GORIC(A) weight for H_1 converges to 1 (denoting full support for H_1). Note that the error probability then goes to 0 and the the ratio of GORIC(A) weights of H_1 versus its complement then goes to infinity. Nevertheless, this is not of interest now, now we are interested in the closeness of log-likelihood values.

Higher sample size

```
# Now, group size is 200 (instead of 40)

# H1 vs complement - border (nl., mu1 = mu2 > mu3) is true
H1 <- "D1 > D2 > D3" # mu1 > mu2 > mu3

# Apply GORIC #
set.seed(123)
results_1c_border_200 <- goric(fit_border_200, hypotheses = list(H1), comparison = "complement")
results_1c_border_200
```

```
restriktor (0.5-85): generalized order-restricted information criterion:

Results:
        model    loglik  penalty     goric  loglik.weights  penalty.weights  goric.weights
1          H1  -829.928    2.833  1665.522           0.509            0.697          0.704
2  complement  -829.962    3.667  1667.258           0.491            0.303          0.296
---
The order-restricted hypothesis 'H1' has 2.38 times more support than its complement.
```

We will also now check whether there is support for the border of the two (non-overlapping) hypotheses:

```
# Specifying multiple null populations, that is,
# using all possibilities of setting inequalities to equalities.
# Here, we will use the default `model_type` (i.e., "asymp") which takes population parameter values (i
est <- coef(fit_border_200)
pop_est <- matrix(c(
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3]),
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  mean(est[1:2]), est[2], mean(est[1:2]),
                  est[1], mean(est[2:3]), mean(est[2:3]),

                  est[1], est[2], est[3]
                  ),
                  byrow = TRUE, ncol = length(est))
rownames(pop_est) <- c("PE_123eq", "PE_12eq", "PE_13eq", "PE_23eq", "Observed")
benchmarks_1c_border_allpos_200 <- benchmark(results_1c_border_200, pop_est = pop_est, ncpus = 8)
```

```
Calculating asymptotic benchmark for population estimates = PE_123eq
Calculating asymptotic benchmark for population estimates = PE_12eq
Calculating asymptotic benchmark for population estimates = PE_13eq
```

```
Calculating asymptotic benchmark for population estimates = PE_23eq
Calculating asymptotic benchmark for population estimates = Observed
```

```
# loglik difference
#print(benchmarks_1c_border_allpos_200, output_type = "ld") # R file
print(benchmarks_1c_border_allpos_200, output_type = "ld", color = FALSE) # Rmd file
```

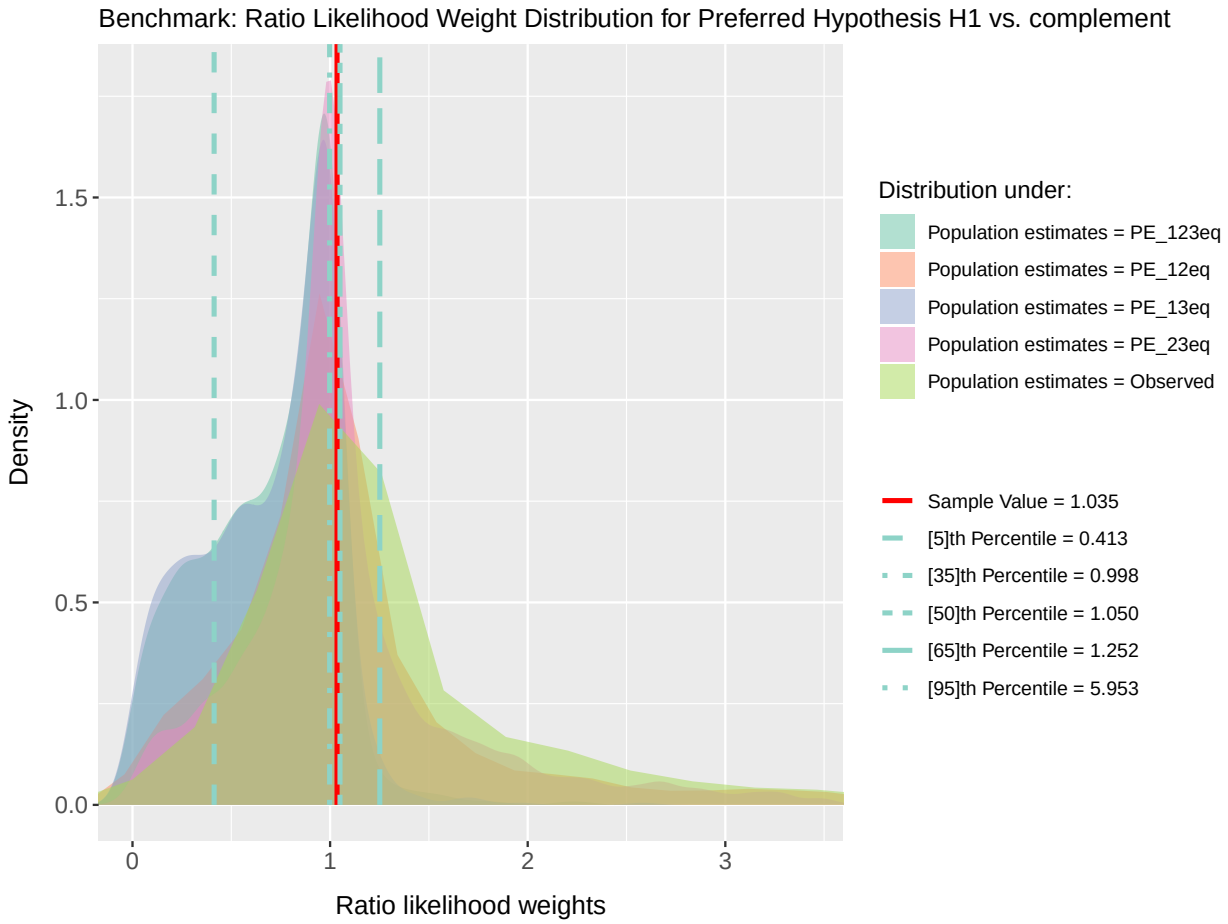
```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.296
Sample Size: 600
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.729  0.729  0.729
PE_12eq   0.874  0.874  0.440
PE_13eq   0.874  0.861  0.874
PE_23eq   0.886  0.650  0.650
Observed  0.886  0.861  0.440


=============================================================================================
Benchmark: Percentiles of Differences in Likelihood Values for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 0.035  -2.278 -0.547 -0.265 -0.090 0.067

Population estimates = PE_12eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement 0.035  -1.412 -0.087 -1.574e-04 0.078 1.411

Population estimates = PE_13eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 0.035  -2.397 -0.580 -0.268 -0.097 0.075

Population estimates = PE_23eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement 0.035  -1.231 -0.084 -2.958e-04 0.061 0.971

Population estimates = Observed
                  Sample 5%     35%    50%   65%   95%
H1 vs. complement 0.035  -0.884 -0.002 0.049 0.224 1.784
```

```
#plot(benchmarks_1c_border_allpos_200, output_type = "ld")
# ratio of loglik weights
#print(benchmarks_1c_border_allpos_200, output_type = "rlw") # R file
print(benchmarks_1c_border_allpos_200, output_type = "rlw", color = FALSE) # Rmd file
```

```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
```

```
Error probability Preferred Hypothesis vs. complement: 0.296
Sample Size: 600
Number of Parameters: 3
Population Estimates (PE):
            D1     D2     D3
PE_123eq  0.729  0.729  0.729
PE_12eq   0.874  0.874  0.440
PE_13eq   0.874  0.861  0.874
PE_23eq   0.886  0.650  0.650
Observed  0.886  0.861  0.440


================================================================================================
Benchmark: Percentiles of Ratio-of-likelihood-weights for the Preferred Hypothesis 'H1'
------------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 1.035  0.102 0.578 0.767 0.914 1.069

Population estimates = PE_12eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 1.035  0.244 0.916 1.000 1.081 4.102

Population estimates = PE_13eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 1.035  0.091 0.560 0.765 0.908 1.078

Population estimates = PE_23eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 1.035  0.292 0.920 1.000 1.063 2.640

Population estimates = Observed
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 1.035  0.413 0.998 1.050 1.252 5.953
```

```
plot(benchmarks_1c_border_allpos_200, output_type = "rlw")
```



Also in this example, with a much higher sample size, the difference in log-likelihood values and the ratio of log-likelihood weights lies between the 5th and the 95th percentiles of the corresponding benchmarks (for all null populations). Based on this, We conclude that the log-likelihood values of H_1 and its complement are close (i.e., their loglik weight ratio is close to 1 and their difference is close to 0). Additionally, one can argue that there is more support for a boundary hypothesis in which group means 1 and 2 are the same, since the sample value is the closest to the 50th percentile of that null distribution.

Example 1 (ANOVA) Ctd.

In this subsection, we look at Example 1 again, where we evaluate H1 ← "D1 > D2 > D3" versus its complement, like in Example 3. In this example, H_1 is true in the population, while in Example 3 the truth is on the border.

In this example, the loglik values are -155.07 and -155.94, with corresponding loglik.weights of 0.7 and 0.3 (and thus a difference of approx. 0.86 and a ratio of approx. 2.37). Since it is hard to judge what is close, one can inspect the benchmarks for the ratio of log-likelihood (loglik) weights and for the differences in log-likelihood values:

```
# Specifying multiple null populations, that is,
# using all possibilities of setting inequalities to equalities.
# Here, we will use the default `model_type` (i.e., "asymp") which takes population parameter values (i
est <- coef(fit)
pop_est <- matrix(c(
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3]),
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  mean(est[1:2]), est[2], mean(est[1:2]),
```

```
                  est[1], mean(est[2:3]), mean(est[2:3]),

                  est[1], est[2], est[3]
                  ),
                  byrow = TRUE, ncol = length(est))
rownames(pop_est) <- c("PE_123eq", "PE_12eq", "PE_13eq", "PE_23eq", "Observed")
benchmarks_1c_allpos <- benchmark(results_1c, pop_est = pop_est, ncpus = 8)
```

```
Calculating asymptotic benchmark for population estimates = PE_123eq
Calculating asymptotic benchmark for population estimates = PE_12eq
Calculating asymptotic benchmark for population estimates = PE_13eq
Calculating asymptotic benchmark for population estimates = PE_23eq
Calculating asymptotic benchmark for population estimates = Observed
```

```
# loglik difference
#print(benchmarks_1c_allpos, output_type = "ld") # R file
print(benchmarks_1c_allpos, output_type = "ld", color = FALSE) # Rmd file
```

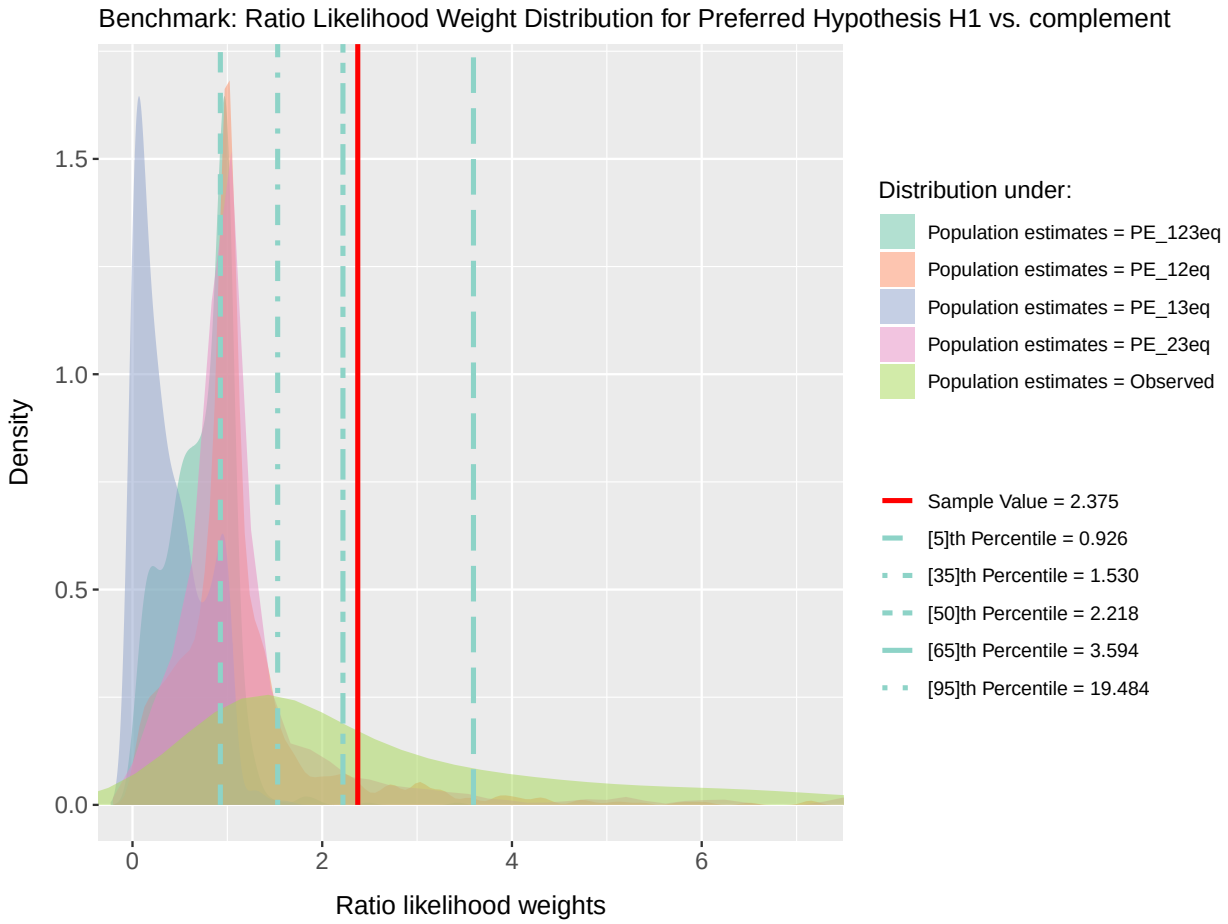
```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.155
Sample Size: 120
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.628  0.628  0.628
PE_12eq   0.815  0.815  0.253
PE_13eq   0.815  0.513  0.815
PE_23eq   1.117  0.383  0.383
Observed  1.117  0.513  0.253


=============================================================================================
Benchmark: Percentiles of Differences in Likelihood Values for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 0.865  -2.059 -0.523 -0.264 -0.080 0.102

Population estimates = PE_12eq
                  Sample 5%     35%    50%       65%   95%
H1 vs. complement 0.865  -1.435 -0.068 1.929e-04 0.071 1.094

Population estimates = PE_13eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 0.865  -4.966 -1.847 -1.223 -0.723 -2.716e-04

Population estimates = PE_23eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement 0.865  -1.187 -0.083 -6.327e-04 0.052 1.210

Population estimates = Observed
```

```
                  Sample 5%     35%   50%   65%   95%
H1 vs. complement 0.865  -0.077 0.426 0.797 1.279 2.970
```

```
#plot(benchmarks_1c_allpos, output_type = "ld")
# ratio of loglik weights
#print(benchmarks_1c_allpos, output_type = "rlw") # R file
print(benchmarks_1c_allpos, output_type = "rlw", color = FALSE) # Rmd file
```

```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.155
Sample Size: 120
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.628  0.628  0.628
PE_12eq   0.815  0.815  0.253
PE_13eq   0.815  0.513  0.815
PE_23eq   1.117  0.383  0.383
Observed  1.117  0.513  0.253


===================================================================================================
Benchmark: Percentiles of Ratio-of-likelihood-weights for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.375  0.128 0.592 0.768 0.923 1.108

Population estimates = PE_12eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.375  0.238 0.934 1.000 1.073 2.987

Population estimates = PE_13eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.375  0.007 0.158 0.294 0.485 1.000

Population estimates = PE_23eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.375  0.305 0.921 0.999 1.054 3.352

Population estimates = Observed
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.375  0.926 1.530 2.218 3.594 19.484
```

```
plot(benchmarks_1c_allpos, output_type = "rlw")
```



In this example, the difference in log-likelihood values and the ratio of log-likelihood weights are larger than the 95th percentile of the corresponding benchmarks for almost all null populations. Only for the null population where the means of groups 2 and 3 are set equal, the sample value is in between the 65th and 95th percentile. When looking at the distributions, the sample value seems more plausible to come from the non-null population ('Observed'), but it could also fit in the null distribution 'PE_23eq'. Hence, we cannot rule out that the log-likelihood values are close, indicating support for the means of groups 2 and 3 being the same.

Next, we will inspect the case where we have a higher sample size.

Higher sample size

```
# Now, group size is 200 (instead of 40)

# H1 vs complement
H1 <- "D1 > D2 > D3" # mu1 > mu2 > mu3

# Apply GORIC #
set.seed(123)
results_1c_200 <- goric(fit_200, hypotheses = list(H1), comparison = "complement")
results_1c_200
```

```
restriktor (0.5-85): generalized order-restricted information criterion:

Results:
```

```
        model    loglik  penalty     goric  loglik.weights  penalty.weights  goric.weights
1          H1  -829.928    2.833  1665.522           0.889            0.697          0.949
2  complement  -832.009    3.667  1671.351           0.111            0.303          0.051
---
The order-restricted hypothesis 'H1' has 18.43 times more support than its complement.
```

Also here, we will check whether there is support for the border of the two (non-overlapping) hypotheses:

```
# Specifying multiple null populations, that is,
# using all possibilities of setting inequalities to equalities.
# Here, we will use the default `model_type` (i.e., "asymp") which takes population parameter values (i
est <- coef(fit_200)
pop_est <- matrix(c(
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3]),
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  mean(est[1:2]), est[2], mean(est[1:2]),
                  est[1], mean(est[2:3]), mean(est[2:3]),

                  est[1], est[2], est[3]
                  ),
                  byrow = TRUE, ncol = length(est))
rownames(pop_est) <- c("PE_123eq", "PE_12eq", "PE_13eq", "PE_23eq", "Observed")
benchmarks_1c_allpos_200 <- benchmark(results_1c_200, pop_est = pop_est, ncpus = 8)
```

```
Calculating asymptotic benchmark for population estimates = PE_123eq
Calculating asymptotic benchmark for population estimates = PE_12eq
Calculating asymptotic benchmark for population estimates = PE_13eq
Calculating asymptotic benchmark for population estimates = PE_23eq
Calculating asymptotic benchmark for population estimates = Observed
```

```
# loglik difference
#print(benchmarks_1c_allpos, output_type = "ld") # R file
print(benchmarks_1c_allpos_200, output_type = "ld", color = FALSE) # Rmd file
```

```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.051
Sample Size: 600
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.634  0.634  0.634
PE_12eq   0.755  0.755  0.392
PE_13eq   0.755  0.589  0.755
PE_23eq   0.921  0.491  0.491
Observed  0.921  0.589  0.392


=============================================================================================
Benchmark: Percentiles of Differences in Likelihood Values for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 2.081  -2.278 -0.547 -0.265 -0.090 0.067
```

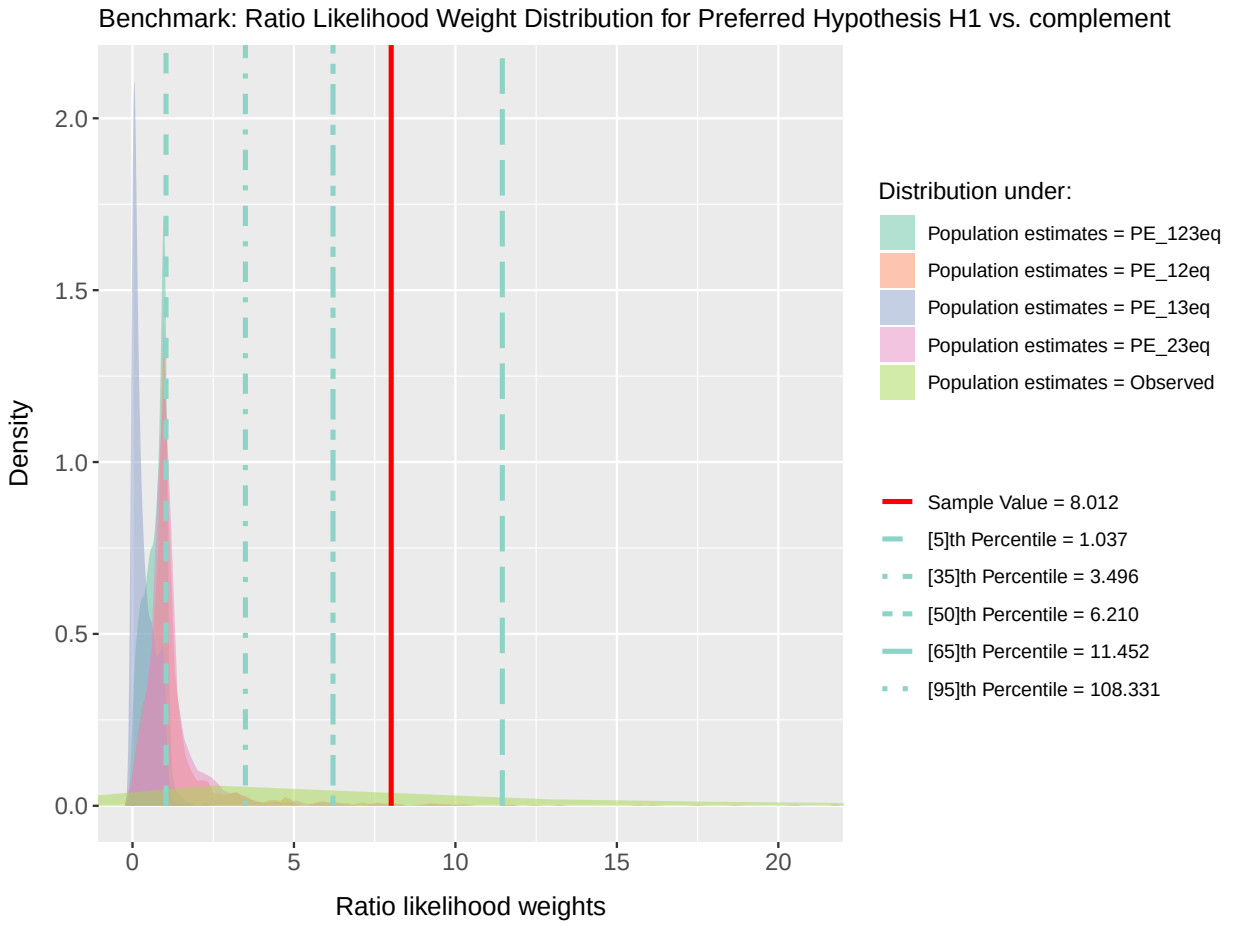
```
Population estimates = PE_12eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement 2.081  -1.412 -0.087 -1.574e-04 0.077 1.356

Population estimates = PE_13eq
                  Sample 5%     35%    50%    65%    95%
H1 vs. complement 2.081  -5.640 -2.312 -1.575 -0.984 -0.028

Population estimates = PE_23eq
                  Sample 5%     35%    50%        65%   95%
H1 vs. complement 2.081  -1.231 -0.081 -3.876e-05 0.087 1.199

Population estimates = Observed
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 2.081  0.037 1.252 1.826 2.438 4.685
```

```r
#plot(benchmarks_1c_allpos_200, output_type = "ld")
# ratio of loglik weights
#print(benchmarks_1c_allpos_200, output_type = "rlw") # R file
print(benchmarks_1c_allpos_200, output_type = "rlw", color = FALSE) # Rmd file
```

```
Benchmark Results
-------------------------------------------------------------------
Preferred Hypothesis: H1
Error probability Preferred Hypothesis vs. complement: 0.051
Sample Size: 600
Number of Parameters: 3
Population Estimates (PE):
             D1     D2     D3
PE_123eq  0.634  0.634  0.634
PE_12eq   0.755  0.755  0.392
PE_13eq   0.755  0.589  0.755
PE_23eq   0.921  0.491  0.491
Observed  0.921  0.589  0.392


=============================================================================================
Benchmark: Percentiles of Ratio-of-likelihood-weights for the Preferred Hypothesis 'H1'
---------------------------------------------------------------------------------------------
Population estimates = PE_123eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 8.012  0.102 0.578 0.767 0.914 1.069

Population estimates = PE_12eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 8.012  0.244 0.916 1.000 1.080 3.882

Population estimates = PE_13eq
                  Sample 5%    35%   50%   65%   95%
H1 vs. complement 8.012  0.004 0.099 0.207 0.374 0.972

Population estimates = PE_23eq
                  Sample 5%    35%   50%   65%   95%
```

```
H1 vs. complement 8.012  0.292 0.922 1.000 1.091 3.316

Population estimates = Observed
                  Sample 5%    35%   50%   65%    95%
H1 vs. complement 8.012  1.037 3.496 6.210 11.452 108.331
```

```
plot(benchmarks_1c_allpos_200, output_type = "rlw")
```



In this example with a higher sample size, the difference in log-likelihood values and the ratio of log-likelihood weights are higher than the 95th percentile of the corresponding benchmarks (for all null populations). Based on this, We conclude that the log-likelihood values of H_1 and its complement are not close. Thus, there is no support for a boundary hypothesis, only for H_1. In that case, we can inspect the height of the support via the GORIC(A) benchmarks (as was done in a previous section).