# Guidelines GORIC(A) benchmarks

Rebecca M. Kuiper

06 December 2023

## Contents



## Introduction

To aid qualifying/labeling the height of support for the preferred hypothesis, one can inspect case-specific benchmarks for the GORIC(A) weights (and for 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 (cf. 'Guidelines_output_GORIC' on https://github.com/rebeccakuiper/Tutorials). Notably, 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 populations values (as is also used in a power analysis when performing a null hypothesis test). For an ANOVA, the set of populations values can be obtained by specifying population effect size and a specific ratio of population means (which can be the ratio of sample means, that is, the ratio of means found in the data), using the 'benchmarks_ANOVA' function. For all types of statistical models (including ANOVA), one can specify the set of population parameter values themselves, using the 'benchmarks' function.
I 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.

Additionally, the functions render benchmarks for the ratio of log-likelihood (loglik) weights of the preferred hypothesis versus the other hypotheses in the set and for the (absolute) differences in loglik values of the preferred hypothesis versus the other hypotheses in the set. Note that the log-likelihood benchmarks (under a mull population) give insight into the distribution of loglik weight ratios and of the (absolute) loglik differences in case some or all of the group means would be the same. 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.

To use the benchmark function, first install and load it:

```r
# Install and load benchmark function
if (!require("devtools")) install.packages("devtools")
library(devtools)
install_github("rebeccakuiper/benchmarks")
library(benchmarks)
# For more information regarding the included functions:
#?benchmarks
#?benchmarks_ANOVA
```

# 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 populations values is true (for the sample size under consideration). More specifically, the benchmarks are based on the 5%, 35%, 50%, 65%, and 95% percentiles of the GORIC(A) weights for the preferred hypothesis and of the ratios of the GORIC(A) weights for the preferred hypothesis with the other hypotheses.

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 estimates 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 support (see first example below).

Additionally, one can check the log-likelihood benchmarks. This can give insight into whether there is support for the boundary of hypotheses (see the third example below, part of the third section).

## Labelling

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

| benchmark (percentile) | Height support |
|---|---|
| below 5% | no support |
| between 5% and 35% | low support |
| between 35% and 65% | medium support |
| between 65% and 95% | high (compelling) support |
| over 95% | very large (tremendous) support |

I 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.

Alternatively, you may want to use a minimum effect. One option 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 is to use the GORIC(A) weights benchmark for a null model, using a minimum percentile level (for which you think the finding is said to be extreme enough). A third option is to investigate benchmarks using a minimum effect size (or looking at multiple ones). When using the 'benchmarks_ANOVA()' function, you can specify the effect size level(s). When using the 'benchmarks()' function, you should specify the population parameter values (reflecting specific effect sizes).
Note that the benchmarks differ when you assume different effect sizes or different population parameter estimates (additionally, the benchmarks change with sample size, as do the GORIC(A) weights themselves).

If of interest (e.g., as a sensitivity analysis), you can do this for multiple sets of population parameter estimates; but this may also complicate drawing conclusions (especially when the assumed sets of population parameter estimates differ much).

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

Once more, I would inspect 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.

## Examples

Next, I will discuss two ANOVA examples. More specifically, I will inspect the case-specific benchmarks values (using the ratio of means as in the data). I 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), I 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' on 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 (0.5-30): generalized order-restricted information criterion:

Results:
        model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1          H1  -123.384    2.832  252.431           0.679            0.698          0.830
2  complement  -124.133    3.668  255.601           0.321            0.302          0.170
---
The order-restricted hypothesis 'H1' has 4.880 times more support than its complement.
```

```
# Benchmarks based on null
#pop.es <- c(0)
#benchmarks_1c <- benchmarks_ANOVA(results_1c, pop.es)
#benchmarks_1c$error.prob.pref.hypo
#benchmarks_1c$benchmarks.weight
#benchmarks_1c$benchmarks.ratios

# Benchmarks based on 3 effect sizes (null/no, small, medium)
pop.es <- c(0, .2, .5)
```

```
benchmarks_1c <- benchmarks_ANOVA(results_1c, pop.es)
benchmarks_1c$error.prob.pref.hypo
```

```
[1] 0.1700569
```

From the goric output, you can conclude that there is support for $H_1 : \mu_1 > \mu_2 > \mu_3$, and $H_1$ is 0.83 / 0.17 $\approx$ 4.88 times more supported than its complement.
The probability that $H_1$ is not the best is 17.01% (namely, the goric.weight for the complement of H1, which is also given by `error.prob.pref.hypo` $\approx$ 0.17). 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 parameter estimates 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 with respect to the inequalities is:

```
benchmarks_1c$benchmarks.weight$`pop.es = 0`
```

```
      Sample        5%       35%       50%       65%       95%
H1 0.8299431 0.1905254 0.5761572 0.644275 0.6819728 0.7141905
```

```
benchmarks_1c$benchmarks.ratios$`pop.es = 0`
```

```
                    Sample        5%      35%      50%      65%      95%
H1 vs. H1         1.000000 1.0000000 1.000000 1.000000 1.000000 1.000000
H1 vs. complement 4.880384 0.2353701 1.359366 1.811168 2.144385 2.498835
```

When assuming that there is no effect in the population (i.e., `pop.es` = 0), that is, all three group means are equal, the GORIC weight of $H_1$ (i.e., 0.83) is higher than the 95% percentile. The same holds true for the ratio of GORIC weights of $H_1$ versus its complement. Hence, our finding is a very extreme one 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 total sample size of 100).

Perhaps you want to inspect a minimum effect. You can either create the hypotheses in such a way that they reflect this, or you can inspect benchmarks for specific effect sizes (or specific population parameter values). Here, we will look at the benchmarks for an effect size of 0.2, that is, a small effect size:

```
benchmarks_1c$benchmarks.weight$`pop.es = 0.2`
```

```
      Sample        5%       35%       50%       65%       95%
H1 0.8299431 0.5471092 0.6976378 0.7064523 0.7336471 0.8699218
```

```
benchmarks_1c$benchmarks.ratios$`pop.es = 0.2`
```

```
                    Sample       5%      35%      50%      65%      95%
H1 vs. H1         1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
H1 vs. complement 4.880384 1.208038 2.307291 2.406602 2.754418 6.687681
```

The output shows that, when assuming that there is a small effect in the population (i.e., `pop.es` = 0.2) and that the ratio of population means is the same as in the data, the GORIC weight of $H_1$ (i.e., 0.83) lies in between the 65% and 95% percentiles and is very close to the 95% (cf. `benchmarks.weight`). The same holds true for the ratio of GORIC weights of $H_1$ versus its complement (cf. `benchmarks.ratios`). Using the table with cut-off values above, this would indicate that there is high (to very high) support, when assuming a small population effect size and a population mean ratio as equal to the sample mean ratio (given a total sample size of 100).

Next, I show that the benchmarks depend on your assumed effect size in the population:

```
benchmarks_1c$benchmarks.weight$`pop.es = 0.5`
```

```
      Sample        5%       35%       50%       65%       95%
H1 0.8299431 0.6918041 0.8077821 0.8662529 0.9153995 0.9882138
```

```
benchmarks_1c$benchmarks.ratios$`pop.es = 0.5`
```

```
                    Sample      5%      35%      50%     65%      95%
H1 vs. H1         1.000000 1.00000 1.000000 1.000000  1.0000  1.00000
H1 vs. complement 4.880384 2.24469 4.202446 6.476797 10.8203 83.84509
```

When assuming that there a medium effect in the population (i.e., `pop.es` = .5) and that the ratio of population means is the same as in the data, the GORIC weight of $H_1$ (i.e., 0.83) lies in between the 35% and 50% and thus 65% percentiles. The same holds true for the ratio of GORIC weights of $H_1$ versus its complement. Using the table with cut-off values above, this indicates that there is medium support, when assuming a medium population effect size and a population mean ratio as equal to the sample mean ratio (given a total sample size of 100).

Log-likelihood check:
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 absolute difference in loglik values is close to 0), then there is support for (one of their) boundaries. In this case, the loglik values are -123.384 and -124.133, with corresponding loglik.weihts of 0.679 and 0.321 (and thus a difference of approx. 0.75 and a ratio of approx. 2.12).
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 (absolute) differences in log-likelihood values. This should then be done for a population in which such a boundary is true. We discuss this in the next section.
For now, we assume that the loglik values or not close.

Population information:
In the data generation, I used a ratio of population means of 3:2:1; implying that $H_1$ is correct. More specifically, I used population mean values of approximately 0.98, 0.65, and 0.33. This implies that Cohen's $d$ is approximately .27; thus, there is a small to medium population effect size (which are in the same order as hypothesized). I then sampled 100 observations, ran an ANOVA (with three groups), and applied the GORIC.
When I 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 the 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 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 (0.5-30): generalized order-restricted information criterion:
```

```
Results:
           model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1             H1  -123.384    2.832  252.431           0.333            0.548          0.548
2             H2  -123.384    3.500  253.767           0.333            0.281          0.281
3  unconstrained  -123.384    4.000  254.767           0.333            0.171          0.171
```

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

```
              vs. H1 vs. H2 vs. unconstrained
H1             1.000  1.950             3.216
H2             0.513  1.000             1.649
unconstrained  0.311  0.607             1.000
```

```
# Benchmarks
pop.es <- c(0, .2, .5)
benchmarks_12u <- benchmarks_ANOVA(results_12u, pop.es)
benchmarks_12u$error.prob.pref.hypo
```

```
[1] 0.1700569
```

```
benchmarks_12u$benchmarks.weight
```

```
$`pop.es = 0`
     Sample       5%       35%       50%       65%      95%
H1 0.548338 0.129804 0.4557295 0.5061509 0.5356164 0.548338

$`pop.es = 0.2`
     Sample        5%       35%      50%      65%      95%
H1 0.548338 0.4013116 0.5483277 0.548338 0.548338 0.548338

$`pop.es = 0.5`
     Sample        5%      35%      50%      65%      95%
H1 0.548338 0.5481983 0.548338 0.548338 0.548338 0.548338
```

```
benchmarks_12u$benchmarks.ratios
```

```
$`pop.es = 0`
                       Sample        5%      35%      50%      65%      95%
H1 vs. H1            1.000000 1.0000000 1.000000 1.000000 1.000000 1.000000
H1 vs. H2            1.950401 0.3332983 1.686983 1.920856 1.950401 1.950401
H1 vs. unconstrained 3.215667 0.2670816 1.861925 2.500761 2.958712 3.215667

$`pop.es = 0.2`
                       Sample       5%      35%      50%      65%      95%
H1 vs. H1            1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
H1 vs. H2            1.950401 1.113707 1.950401 1.950401 1.950401 1.950401
H1 vs. unconstrained 3.215667 1.629242 3.215503 3.215667 3.215667 3.215667

$`pop.es = 0.5`
                       Sample       5%      35%      50%      65%      95%
H1 vs. H1            1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
H1 vs. H2            1.950401 1.949300 1.950401 1.950401 1.950401 1.950401
H1 vs. unconstrained 3.215667 3.213853 3.215667 3.215667 3.215667 3.215667
```

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, we would find again that there is support for the overlap (here, $H_1$) and we should not use the benchmarks: From both types of GORIC(A) weights benchmarks (i.e., `benchmarks.weight` and `benchmarks.ratios`) and when the effect size is higher than 0, we can see that there is maximum support, and our support resembles that. Hence, we conclude that there is support for the overlap. Since there is a maximum support, it is not possible to label this support (notably, one should thus also not compare the results to the benchmarks for a 0 effect size). When you would evaluate the overlap (or the preferred hypothesis) versus its complement, you can inspect the height of the support.

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 absolute 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 mull population) give insight into the distribution of loglik weight ratios and of the (absolute) loglik differences in case some or all of the group means would be the same.

Population information:
In the data generation, I used a ratio of population means of 3:2:1; implying that $H_1$ is correct. More specifically, I used population mean values of approximately 0.98, 0.65, and 0.33. This implies that Cohen's $d$ is approximately .27; thus, there is a small to medium population effect size (which are in the same order as hypothesized). I then sampled 100 observations, ran an ANOVA (with three groups), and applied the GORIC.
When I 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 bounds it can take on. The benchmarks for the GORIC(A) weights will also attain the maximum value (here, 0.548) as will the ratio of weights (e.g., 1.950 for $H_1$ vs $H_2$); and it will for each positive population effect size. For a null population, the GORIC(A) weight sbenchmarks will remian teh 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 absolute 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 (absolute) differences in log-likelihood values. This should then be done for a population in which such a boundary is true.

Next, I will inspect two examples: one in which the border is true (Example 3) and one where it is not (Example 1 continued). I will first use a total sample size of 100, like in the example above; afterwards, I will inspect a higher sample size to obtain insight into the asymptotic properties of the loglik weights.

For now, I will use that the loglik values are said to be the same if the loglik differences are in between the 5% and 95% 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 four types of loglik benchmarks:
* the loglik ratios (i.e., ratio of loglik weights), which can take on values between 0 and infinity and where 1 means that the loglik values are the same;
* the loglik ratios transformed such that the values lie between 1 and infinity, where 1 means that the loglik values are the same;

* the loglik differences, which can take on values between minus infinity and infinity and where 0 means that the loglik values are the same;
* the absolute loglik differences, which can take on values between 0 and infinity and where 0 means that the loglik values are the same;
Note that the first and the third contain the same information, as do the second and the fourth.
It seems like the difference between two log-likelihoods works best and is closely related to the likelihood ratio test. Hence, I will only show the output for this one. - One could also do a likelihood ratio test, using a Chi-square distribution (using the difference in penalty values as the degrees of freedom). Notably, one should actually use a Chi-bar-square distribution. This uses Chi-bar-square weights, that is, level probabilities, and these can be retrieved from the goric object. Further research is needed to find out how this should be done and to see how well it works.

## 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 (0.5-30): generalized order-restricted information criterion:

Results:
        model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1          H1  -121.591    2.832  248.847           0.534            0.698          0.725
2  complement  -121.727    3.668  250.790           0.466            0.302          0.275
---
The order-restricted hypothesis 'H1' has 2.642 times more support than its complement.
```

Before we inspect the height of support for the preferred hypotheses, we can check whether there is support for the boundary/border of the two (non-overlapping) hypotheses. By eyeballing, I believe the log-likelihood values are close (-121.591 vs -121.727).
To obtain better evidence for the closeness of the loglik values, I will use the log-likelihood benchmarks functions for several specifications of null populations:

```
## Loglik benchmarks based on null / no effect sizes
## That is, setting all three means equal in the population,
## using the benchmarks_ANOVA function:
#pop.es <- c(0)
#benchmarks_1c_border <- benchmarks_ANOVA(results_1c_border, pop.es)
##benchmarks_1c_border$benchmarks.LLratios
##benchmarks_1c_border$benchmarks.LLratios_ge1
#benchmarks_1c_border$benchmarks.difLL
##benchmarks_1c_border$benchmarks.absdifLL

# To obtain more insight into closeness log-likelihood values:
# Loglik benchmarks based on using all possible equalities,
# using the benchmarks function:
est <- coef(fit_border)
pop.est <- matrix(c(
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  est[1], mean(est[2:3]), mean(est[2:3]),
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3])
```

```
                  ),
                  byrow = TRUE, ncol = length(est))
benchmarks_1c_border_allpos <- benchmarks(results_1c_border, pop.est)
#benchmarks_1c_border_allpos$benchmarks.LLratios
#benchmarks_1c_border_allpos$benchmarks.LLratios_ge1
benchmarks_1c_border_allpos$benchmarks.difLL
```

```
$`pop.est.nr. = 1`
                     Sample        5%         35%           50%        65%      95%
H1 vs. H1         0.0000000  0.000000  0.00000000  0.000000e+00 0.00000000 0.000000
H1 vs. complement 0.1353646 -1.421586 -0.08768385 -3.684246e-06 0.06151908 1.023305

$`pop.est.nr. = 2`
                     Sample        5%         35%           50%        65%      95%
H1 vs. H1         0.0000000  0.000000  0.00000000  0.0000000000 0.00000000 0.000000
H1 vs. complement 0.1353646 -1.296229 -0.09539898 -0.0009850665 0.03951493 0.810325

$`pop.est.nr. = 3`
                     Sample        5%        35%        50%         65%        95%
H1 vs. H1         0.0000000  0.000000  0.0000000  0.0000000  0.00000000 0.00000000
H1 vs. complement 0.1353646 -2.419411 -0.4900731 -0.2336335 -0.07146743 0.08778951
```

```
#benchmarks_1c_border_allpos$benchmarks.absdifLL
```

In this example, the difference in log-likelihood values (.14) lies between the 65% and the 95% percentile of the loglik benchmarks (so, in between the 5% and 95%) of the first two null populations; and it is higher than the 95% benchmark of the third null population (in which all means are set equal). Based on this, I 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 (absolute) difference is close to 0); and that there is support for a boundary hypothesis in which two means are the same.

One could also do a likelihood ratio test, using a Chi-square distribution. One should actually use a Chi-bar-square distribution, but additional research is needed.

```
alpha <- .10 # This corresponds to using the 5% and 95% percentiles
df_diff = 3.668-2.832 # These are the penalty (df) values for H1 and 'Hc'
LR <- -2*-121.727 + 2*-121.591 # = 2*'benchmarks.difLL'
#
# Critical value
qchisq(alpha, df = df_diff, lower.tail=FALSE)
```

```
[1] 2.342696
```

```
#2.343
# LR from sample, to be compared to the critical value
LR
```

```
[1] 0.272
```

```
# 0.272
```

Based on this test, I would also conclude that the loglik values do not differ significantly (since the value from the sample is smaller that the critical value).

Population information:
In the data generation, I 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, I used population mean values of approximately 0.77, 0.77, and 0.31. This implies that

Cohen's $d$ is approximately .22; thus, there is a small to medium population effect size. I then sampled 100 observations, ran an ANOVA (with three groups), and applied the GORIC.
When I 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, total sample size is 1000 (instead of 100)

# 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_1000 <- goric(fit_border_1000, hypotheses = list(H1), comparison = "complement")
results_1c_border_1000
```

```
restriktor (0.5-30): generalized order-restricted information criterion:

Results:
        model     loglik  penalty     goric  loglik.weights  penalty.weights  goric.weights
1          H1  -1284.884    2.833  2575.434           0.540            0.697          0.730
2  complement  -1285.046    3.667  2577.425           0.460            0.303          0.270
---
The order-restricted hypothesis 'H1' has 2.706 times more support than its complement.
```

We will also now check whether there is support for the border of the two (non-overlapping) hypotheses. By eyeballing, I believe the log-likelihood values are close (-1284.884 vs -1285.046), but - to obtain more evidence - I will use the log-likelihood benchmarks functions for several specifications of null populations:

```
## Loglik benchmarks based on null / no effect sizes
## That is, setting all three means equal in the population,
## using the benchmarks_ANOVA function:
#pop.es <- c(0)
#benchmarks_1c_border_1000 <- benchmarks_ANOVA(results_1c_border_1000, pop.es)
##benchmarks_1c_border_1000$benchmarks.LLratios
##benchmarks_1c_border_1000$benchmarks.LLratios_ge1
#benchmarks_1c_border_1000$benchmarks.difLL
##benchmarks_1c_border_1000$benchmarks.absdifLL

# To obtain more insight into closeness log-likelihood values:
# Loglik benchmarks based on using all possible equalities (and none),
# using the benchmarks function:
est <- coef(fit_border_1000)
pop.est <- matrix(c(
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  est[1], mean(est[2:3]), mean(est[2:3]),
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3])
                  ),
                  byrow = TRUE, ncol = length(est))
benchmarks_1c_border_1000_allpos <- benchmarks(results_1c_border_1000, pop.est)
#benchmarks_1c_border_1000_allpos$benchmarks.LLratios
#benchmarks_1c_border_1000_allpos$benchmarks.LLratios_ge1
```

```
benchmarks_1c_border_1000_allpos$benchmarks.difLL
```

```
$`pop.est.nr. = 1`
                     Sample       5%         35%           50%        65%      95%
H1 vs. H1         0.0000000  0.00000  0.00000000  0.000000e+00 0.00000000 0.000000
H1 vs. complement 0.1617548 -1.42333 -0.08873944 -6.073469e-06 0.06184266 1.134347

$`pop.est.nr. = 2`
                     Sample        5%         35%           50%        65%      95%
H1 vs. H1         0.0000000  0.000000  0.00000000 0.000000e+00 0.00000000 0.000000
H1 vs. complement 0.1617548 -1.296229 -0.08355899 4.590777e-05 0.07467684 1.344459

$`pop.est.nr. = 3`
                     Sample        5%        35%        50%       65%        95%
H1 vs. H1         0.0000000  0.000000  0.0000000  0.0000000  0.000000 0.00000000
H1 vs. complement 0.1617548 -2.419487 -0.4853368 -0.2316165 -0.070985 0.09029137
```

```
#benchmarks_1c_border_1000_allpos$benchmarks.absdifLL
```

Also in this example with a much hger sample size, the difference in log-likelihood values (.16) lies between the 65% and the 95% percentile of the loglik benchmarks (so, in between the 5% and 95%) of the first two null populations; and it is higher than the 95% benchmark of the third null population (in which all means are set equal). Based on this, I 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 (absolute) difference is close to 0); and that there is support for a boundary hypothesis in which two means are the same.

One could also do a likelihood ratio test, using a Chi-square distribution. One should actually use a Chi-bar-square distribution, but additional research is needed.

```
alpha <- .10 # This corresponds to using the 5% and 95% percentiles
df_diff = 3.667-2.833 # These are the penalty (df) values for H1 and 'Hc'
LR <- -2*-1285.046 + 2*-1284.884 # = 2*'benchmarks.difLL'
#
# Critical value
qchisq(alpha, df = df_diff, lower.tail=FALSE)
```

```
[1] 2.338119
```

```
#2.338
# LR from sample, to be compared to the critical value
LR
```

```
[1] 0.324
```

```
# 0.324
```

Based on this test, I would also conclude that the loglik values do not differ significantly (since the value from the sample is smaller that the critical value).

## 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 -123.384 and -124.133, with corresponding loglik.weights of 0.679 and 0.321 (and thus a difference of approx. 0.75 and a ratio of approx. 2.12). 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 (absolute)

differences in log-likelihood values. This should then be done for a population in which such a boundary is true.

```
# To obtain more insight into closeness log-likelihood values:
# Benchmarks based on using all possible equalities (and none),
# using the benchmarks function:
est <- coef(fit)
pop.est <- matrix(c(
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  est[1], mean(est[2:3]), mean(est[2:3]),
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3])
                  ),
                  byrow = TRUE, ncol = length(est))
benchmarks_1c_allpos <- benchmarks(results_1c, pop.est)
#benchmarks_1c_allpos$benchmarks.LLratios
benchmarks_1c_allpos$benchmarks.LLratios_ge1
```

```
$`pop.est.nr. = 1`
                   Sample       5%      35%      50%      65%      95%
H1 vs. H1         1.00000 1.000000 1.000000 1.000000 1.000000 1.000000
H1 vs. complement 2.11521 1.001626 1.102072 1.232758 1.529933 4.672766

$`pop.est.nr. = 2`
                   Sample      5%      35%      50%      65%      95%
H1 vs. H1         1.00000 1.00000 1.000000 1.000000 1.000000 1.000000
H1 vs. complement 2.11521 1.00216 1.115449 1.267364 1.593347 5.867315

$`pop.est.nr. = 3`
                   Sample       5%      35%      50%      65%      95%
H1 vs. H1         1.00000 1.000000 1.000000 1.000000 1.000000  1.00000
H1 vs. complement 2.11521 1.002044 1.115849 1.298028 1.636077 11.23987
```

```
#benchmarks_1c_allpos$benchmarks.difLL
#benchmarks_1c_allpos$benchmarks.absdifLL
```

In this example, the difference in log-likelihood values (.75) lies between the 65% and the 95% percentile of the loglik benchmarks (so, in between the 5% and 95%) of the first two null populations; and it is higher than the 95% benchmark of the third null population (in which all means are set equal). Based on this, I 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 (absolute) difference is close to 0); and that there is support for a boundary hypothesis in which two means are the same.
I know that in the population (since I sampled the data) not a boundary hypothesis but H1 itself is true. Hence, in thus case the loglik benchmarks do not help us. Moreover, based on eyeballing, I would have said that the loglik values differ. Clearly, more research is needed for the loglik benchmarks.

One could also do a likelihood ratio test, using a Chi-square distribution. One should actually use a Chi-bar-square distribution, but additional research is needed.

```
alpha <- .10 # This corresponds to using the 5% and 95% percentiles
df_diff = 3.668-2.832 # These are the penalty (df) values for H1 and 'Hc'
LR <- -2*-124.133 + 2*-123.384 # = 2*'benchmarks.difLL'
#
# Critical value
qchisq(alpha, df = df_diff, lower.tail=FALSE)
```

```
[1] 2.342696
```

```
#2.343
# LR from sample, to be compared to the critical value
LR
```

```
[1] 1.498
```

```
# 1.498
```

Based on this test, I would also conclude that the loglik values do not differ significantly (since the value from the sample is smaller that the critical value).

Population information:
In the data generation, I used a ratio of population means of 3:2:1; implying that $H_1$ is correct. More specifically, I used population mean values of approximately 0.98, 0.65, and 0.33. This implies that Cohen's $d$ is approximately .27; thus, there is a small to medium population effect size (which are in the same order as hypothesized). I then sampled 100 observations, ran an ANOVA (with three groups), and applied the GORIC.
When I would sample more observations, the GORIC(A) weight for $H_1$ converges to 1 (denoting full support for $H_1$), as will the benchmarks for the GORIC(A) weight 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.

### Higher sample size

```
# Now, total sample size is 300 (instead of 100)

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

# Apply GORIC #
set.seed(123)
results_1c_300 <- goric(fit_300, hypotheses = list(H1), comparison = "complement")
results_1c_300
```

### Total sample size of 300

```
restriktor (0.5-30): generalized order-restricted information criterion:

Results:
        model    loglik  penalty    goric  loglik.weights  penalty.weights  goric.weights
1          H1  -376.425    2.833  758.517           0.954            0.697          0.980
2  complement  -379.462    3.667  766.258           0.046            0.303          0.020
---
The order-restricted hypothesis 'H1' has 47.965 times more support than its complement.
```

Also here, we will check whether there is support for the border of the two (non-overlapping) hypotheses. By eyeballing, I believe the log-likelihood values are not close (-376.425 vs -379.462, leading to loglik.weights of 0.954 and 0.046, respectively). To obtain more evidence, I will use the log-likelihood benchmarks functions for several specifications of null populations:

```
## Loglik benchmarks based on null / no effect sizes
## That is, setting all three means equal in the population,
## using the benchmarks_ANOVA function:
#pop.es <- c(0)
#benchmarks_1c_300 <- benchmarks_ANOVA(results_1c_300, pop.es)
##benchmarks_1c_300$benchmarks.LLratios
##benchmarks_1c_300$benchmarks.LLratios_ge1
```

```
#benchmarks_1c_300$benchmarks.difLL
##benchmarks_1c_300$benchmarks.absdifLL

# To obtain more insight into closeness log-likelihood values:
# Loglik benchmarks based on using all possible equalities (and none),
# using the benchmarks function:
est <- coef(fit_300)
pop.est <- matrix(c(
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  est[1], mean(est[2:3]), mean(est[2:3]),
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3])
                  ),
                  byrow = TRUE, ncol = length(est))
benchmarks_1c_300_allpos <- benchmarks(results_1c_300, pop.est)
#benchmarks_1c_300_allpos$benchmarks.LLratios
#benchmarks_1c_300_allpos$benchmarks.LLratios_ge1
benchmarks_1c_300_allpos$benchmarks.difLL
```

```
$`pop.est.nr. = 1`
                    Sample        5%         35%           50%        65%      95%
H1 vs. H1         0.000000  0.000000  0.00000000  0.000000e+00 0.00000000 0.000000
H1 vs. complement 3.037129 -1.422054 -0.08842792 -8.319458e-06 0.06165588 1.104856

$`pop.est.nr. = 2`
                    Sample        5%         35%          50%        65%     95%
H1 vs. H1         0.000000  0.000000  0.00000000 0.000000e+00 0.00000000 0.00000
H1 vs. complement 3.037129 -1.296229 -0.08355899 4.590777e-05 0.07518101 1.40009

$`pop.est.nr. = 3`
                    Sample        5%        35%        50%         65%        95%
H1 vs. H1         0.000000  0.000000  0.0000000  0.0000000  0.00000000 0.00000000
H1 vs. complement 3.037129 -2.419489 -0.4854035 -0.2315981 -0.07077023 0.09057281
```

```
#benchmarks_1c_300_allpos$benchmarks.absdifLL
```

In this example with a higher sample size, the difference in log-likelihood values (3.04) is higher than the 95% percentile of the loglik benchmarks for all three null populations. Based on this, I 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 $H_1$.

One could also do a likelihood ratio test, using a Chi-square distribution. One should actually use a Chi-bar-square distribution, but additional research is needed.

```
alpha <- .10 # This corresponds to using the 5% and 95% percentiles
df_diff = 3.667-2.833 # These are the penalty (df) values for H1 and 'Hc'
LR <- -2*-379.462 + 2*-376.425 # = 2*'benchmarks.difLL'
#
# Critical value
qchisq(alpha, df = df_diff, lower.tail=FALSE)
```

```
[1] 2.338119
```

```
#2.338
# LR from sample, to be compared to the critical value
LR
```

```
[1] 6.074
```

```
# 6.074
```

Based on this test, I would also conclude that the loglik values do differ significantly (since the value from the sample is larger that the critical value).

```
# Now, total sample size is 1000 (instead of 100)

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

# Apply GORIC #
set.seed(123)
results_1c_1000 <- goric(fit_1000, hypotheses = list(H1), comparison = "complement")
results_1c_1000
```

**Total sample size of 1000**

```
restriktor (0.5-30): generalized order-restricted information criterion:

Results:
        model     loglik  penalty     goric  loglik.weights  penalty.weights  goric.weights
1          H1  -1284.884    2.833  2575.434           1.000            0.697          1.000
2  complement  -1292.998    3.667  2593.329           0.000            0.303          0.000
---
The order-restricted hypothesis 'H1' has 7689.236 times more support than its complement.
```

Also here, we will check whether there is support for the border of the two (non-overlapping) hypotheses. By eyeballing, I believe the log-likelihood values are not close (-1284.884 vs -1292.998, leading to loglik.weights of 1 and 0, respectively). To obtain more evidence, I will use the log-likelihood benchmarks functions for several specifications of null populations:

```
## Loglik benchmarks based on null / no effect sizes
## That is, setting all three means equal in the population,
## using the benchmarks_ANOVA function:
#pop.es <- c(0)
#benchmarks_1c_1000 <- benchmarks_ANOVA(results_1c_1000, pop.es)
##benchmarks_1c_1000$benchmarks.LLratios
##benchmarks_1c_1000$benchmarks.LLratios_ge1
#benchmarks_1c_1000$benchmarks.difLL
##benchmarks_1c_1000$benchmarks.absdifLL

# To obtain more insight into closeness log-likelihood values:
# Loglik benchmarks based on using all possible equalities (and none),
# using the benchmarks function:
est <- coef(fit_1000)
pop.est <- matrix(c(
                  est[1], est[2], est[3],
                  mean(est[1:2]), mean(est[1:2]), est[3],
                  est[1], mean(est[2:3]), mean(est[2:3]),
                  mean(est[1:3]), mean(est[1:3]), mean(est[1:3])
                  ),
                  byrow = TRUE, ncol = length(est))
benchmarks_1c_1000_allpos <- benchmarks(results_1c_1000, pop.est)
#benchmarks_1c_1000_allpos$benchmarks.LLratios
```

```
#benchmarks_1c_1000_allpos$benchmarks.LLratios_ge1
benchmarks_1c_1000_allpos$benchmarks.difLL
```

```
$`pop.est.nr. = 1`
                    Sample       5%      35%     50%      65%      95%
H1 vs. H1         0.000000 0.000000 0.000000 0.00000 0.000000  0.00000
H1 vs. complement 8.113968 3.055875 6.656764 7.66645 8.681848 12.45022

$`pop.est.nr. = 2`
                    Sample        5%         35%           50%        65%      95%
H1 vs. H1         0.000000  0.000000  0.00000000  0.000000e+00 0.00000000 0.000000
H1 vs. complement 8.113968 -1.371695 -0.08262674 -5.937951e-06 0.05301166 1.370121

$`pop.est.nr. = 3`
                    Sample        5%        35%          50%        65%      95%
H1 vs. H1         0.000000  0.000000  0.0000000 0.0000000000 0.00000000 0.000000
H1 vs. complement 8.113968 -1.391727 -0.0599013 0.0005269092 0.07020758 1.352651

$`pop.est.nr. = 4`
                    Sample        5%        35%        50%         65%        95%
H1 vs. H1         0.000000  0.000000  0.0000000  0.0000000  0.00000000 0.00000000
H1 vs. complement 8.113968 -2.416984 -0.5748742 -0.2858309 -0.08080468 0.08519334
```

```
#benchmarks_1c_1000_allpos$benchmarks.absdifLL
```

In this example with a much higher sample size, the difference in log-likelihood values (3341) is much higher than the 95% percentile of the loglik benchmarks for all three null populations. Based on this, I 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 $H_1$.

One could also do a likelihood ratio test, using a Chi-square distribution. One should actually use a Chi-bar-square distribution, but additional research is needed.

```
alpha <- .10 # This corresponds to using the 5% and 95% percentiles
df_diff = 3.667-2.833 # These are the penalty (df) values for H1 and 'Hc'
LR <- -2*-1292.998 + 2*-1284.884 # = 2*'benchmarks.difLL'
#
# Critical value
qchisq(alpha, df = df_diff, lower.tail=FALSE)
```

```
[1] 2.338119
```

```
#2.338
# LR from sample, to be compared to the critical value
LR
```

```
[1] 16.228
```

```
# 16.228
```

Based on this test, I would also conclude that the loglik values do differ significantly (since the value from the sample is larger that the critical value).