Guidelines GORIC(A) benchmarks

Rebecca M. Kuiper

22 November 2023

Contents

Introduction **1**
Example: H_1 vs its complement 2
Example: Overlapping hypotheses 5

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 the ratio of 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 of two or more hypotheses (cf. 'Guidelines_output_GORIC' on https://github.com/rebeccakuiper/Tutorials). Notably, the benchmarks can 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).
One can also use the null as a reference, that is, assuming that the equalities (instead of inequalities) hold true in the population.

The 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 come up with 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), this is another indication that there is support for the overlap of two or more hypotheses; otherwise, the benchmarks can be used to qualify the support (see first example below).
I am not in favor of cut-off points, 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

benchmark (percentile)	Height support
over 95%	very large (tremendous) support

Note that the benchmarks differ when you assume 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).

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

```
# 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
```

Next, I will inspect the case-specific benchmarks values (using the ratio of means as in the data). I will look at
- the example where we evaluated $H_1 : \mu_1 > \mu_2 > \mu_3$ versus its complement; and
- the example where we evaluated 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.

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

Example: H_1 vs its complement

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

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

```
restriktor (0.5-20): 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
```

```
[1] 0.1700569
```

```
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
```

```
# 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
```

```
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
```

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

```
restriktor (0.5-20): 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
pop.es <- c(0, .2, .5)
benchmarks_1c_1000 <- benchmarks_ANOVA(results_1c_1000, pop.es)
#benchmarks_1c_1000$error.prob.pref.hypo
#benchmarks_1c_1000$benchmarks.weight
benchmarks_1c_1000$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
```

```
$`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

$`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
```

As discussed before, 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 17.0056931% (namely, `error.prob.pref.hypo` $\approx$ 0.1700569). 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). The output shows that, when assuming that there is a small effect in the population (i.e., `pop.es` = .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 population:

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

```
      Sample       5%       35%       50%       65%       95%
H1 0.8299431 0.691804 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).

If you do not have prior knowledge regarding the expected population effect size, or if you just want to compare your results with the situation in which the equalities hold true, use a null population (here, an effect size of 0). Then, you obtain insight into how likely the inequalities are.

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

```
      Sample       5%       35%      50%       65%       95%
H1 0.8299431 0.190525 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), 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. 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).

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.

Example: Overlapping hypotheses

```
# H1, H2, and unconstrained (default) - subset 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 (0.5-20): 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
```

As we saw before, both hypotheses are not weak, H_1 is the best, and there is support for the overlap of the hypotheses which is H_1 here (since H_1 is a subset of H_2).

From both types of 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. 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.

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.