Hierarchical Testing: Evolved Weights

2026-02-20

Load Data and Libraries

```r
suppressWarnings(suppressPackageStartupMessages({
    library(emmeans)
    library(lme4)
    library(brms)
    library(lmerTest)
    library(report)  # bayesian reporting
}))

df = read.csv("all_hv.csv", header = TRUE, stringsAsFactors = FALSE)

df$exp = as.factor(df$exp)
df$exp = relevel(df$exp, ref = "Equal Weights")

df$objective1 = as.factor(ifelse(df$objective1 == 1, "AUROC",
    "Acc"))
df$objective1 = relevel(df$objective1, ref = "AUROC")

df$objective2 = as.factor(ifelse(df$objective2 == 1, "sFNR",
    "DP"))
df$objective2 = relevel(df$objective2, ref = "sFNR")
```

Distribution of hypervolume

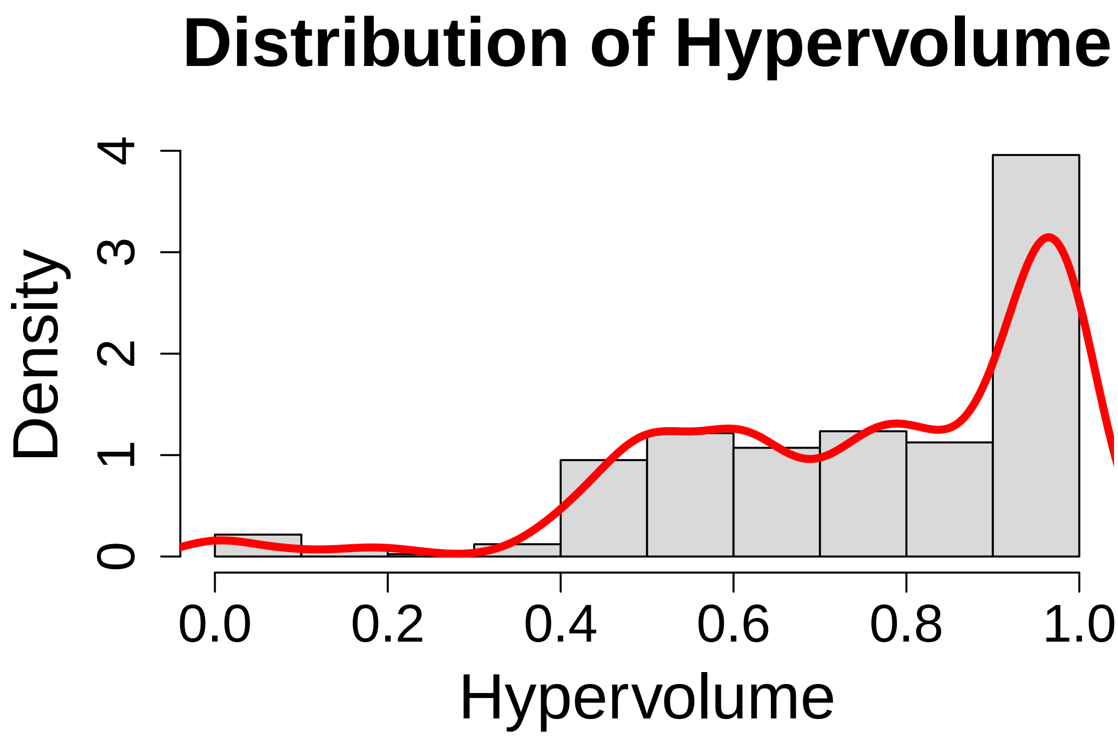
```r
# Pull hv, drop missing/non-finite values
hv <- df$hv
hv <- hv[is.finite(hv)]

# --- Plot: histogram as density + red KDE line ---
op <- par(mar = c(6, 6, 5, 2) + 0.1)  # bigger margins like the example

h <- hist(
  hv,
  breaks = 10,                # adjust (e.g., 8, 12) to match binning
  probability = TRUE,         # y-axis is density ("Probability" in your plot)
  col = "gray85",
  border = "black",
  main = "Distribution of Hypervolume",
  xlab = "Hypervolume",
  ylab = "Density",
  cex.main = 2.2,             # large bold title
  font.main = 2,
```

```r
  cex.lab = 2.0,                # large axis labels
  cex.axis = 1.7
)

# Density curve (red)
lines(
  density(hv, na.rm = TRUE),
  col = "red",
  lwd = 4
)
```




```r
par(op)
```

Mixed Effect Modeling

Using lme4 package

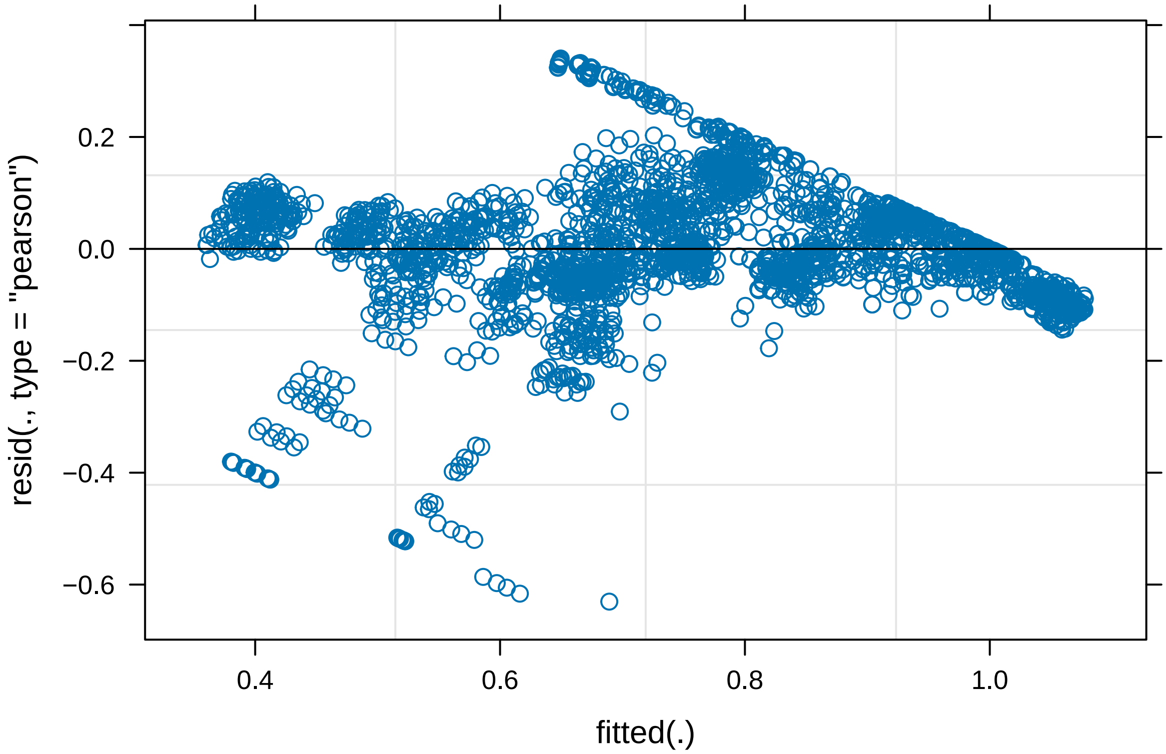
```r
mod_lme4 = lmer(hv ~ exp + objective1 + objective2 + objective1 *
    objective2 + exp * objective1 + exp * objective2 + (1 | dataset/rep),
    data = df)
anova(mod_lme4)  # omnibus
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                      Sum Sq Mean Sq NumDF DenDF   F value    Pr(>F)    
## exp                   2.890   1.445     2  2411   94.1086 < 2.2e-16 ***
## objective1            0.106   0.106     1  2411    6.9332  0.008515 ** 
## objective2           33.360  33.360     1  2411 2172.4653 < 2.2e-16 ***
## objective1:objective2  0.001   0.001     1  2411    0.0919  0.761824    
## exp:objective1        0.032   0.016     2  2411    1.0364  0.354884    
## exp:objective2        1.655   0.828     2  2411   53.8995 < 2.2e-16 ***
```

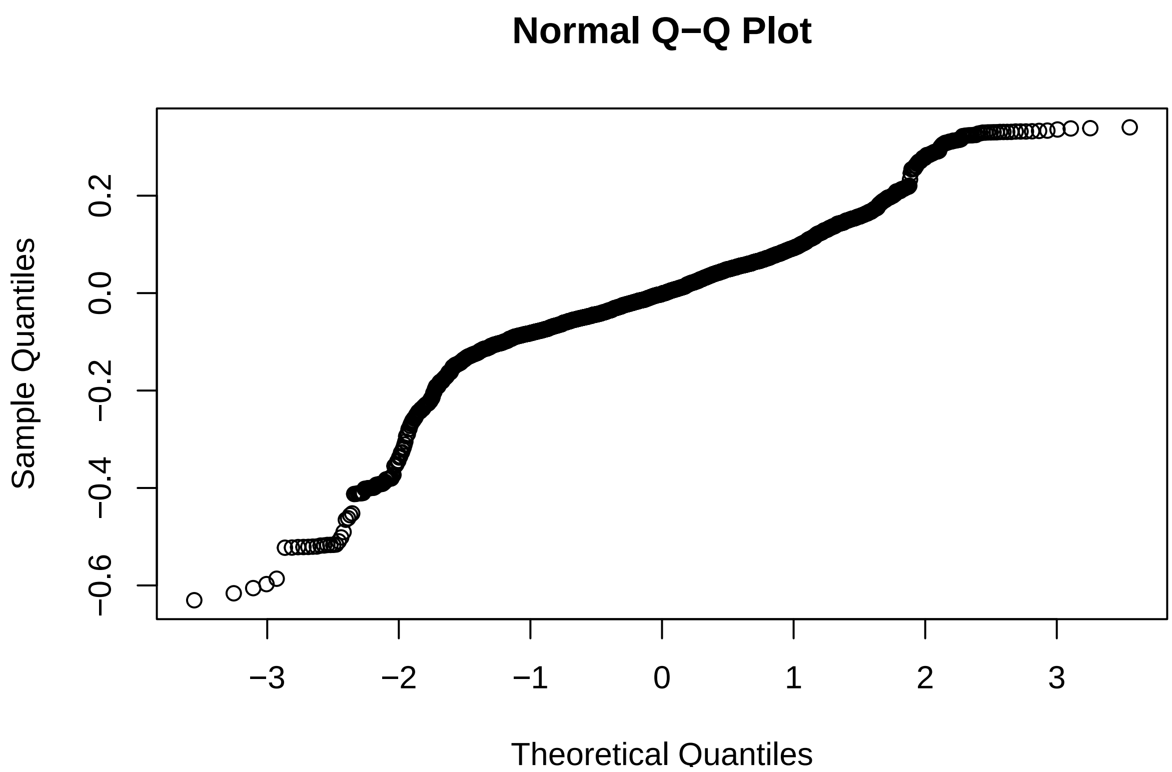
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Data does not meet model assumptions for lme4.

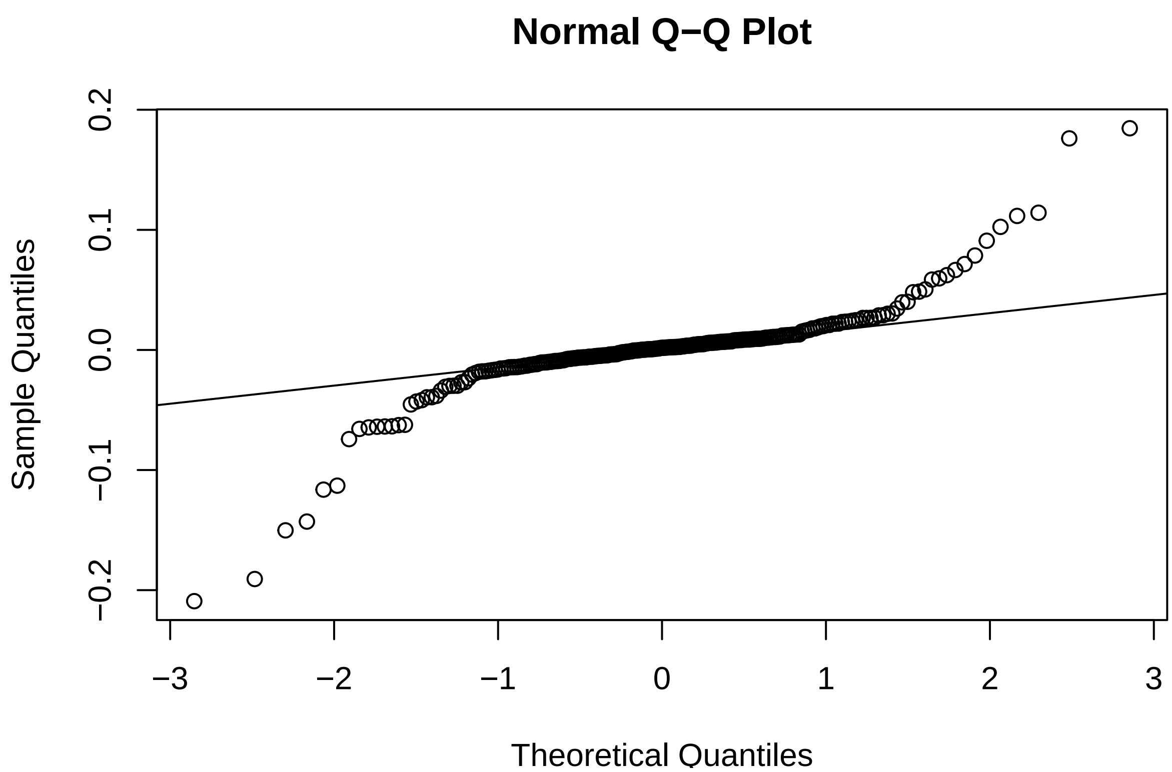
```r
# check homogeneity of variance: failed
plot(mod_lme4)
```




```r
# check normality of variance: failed
qqnorm(resid(mod_lme4))
```




```r
# check normality of random effects: failed
qqnorm(unlist(ranef(mod_lme4)))
qqline(unlist(ranef(mod_lme4)))
```



Using brms package

Fit Bayesian generalized (non-)linear multivariate multilevel models using 'Stan' for full Bayesian inference.

```
model = brm(hv ~ exp * objective1 * objective2 + (1 | dataset/rep),
    family = brms::zero_one_inflated_beta(), chains = 2, data = df)
```

Test of Effects

Method of inference: Sivula, Magnusson and Vehtari (2020)
Resources: https://arxiv.org/pdf/2008.10296, https://easystats.github.io/report/reference/report.compare.lo o.html

- Significant main effect of weighting method (exp).
- Significant main effect of performance objective (objective 1).
- Significant main effect of fairness objective (objective 2).
- Significant interaction between weighting method and fairness objective (objective 2).
- Interpretation: The effect of weighting method depends on fairness objective, regardless of choice of performance objective.

```
##        term    elpd_diff   se_diff         z             p
## 1       exp -143.9703192 18.766223  7.671779  1.696269e-14
## 2      obj1  -22.8675099  7.168822  3.189856  1.423437e-03
## 3      obj2 -740.7891790 28.385510 26.097441 3.897947e-150
## 4  expXobj1   -2.0940007  2.046388  1.023267  3.061817e-01
## 5  expXobj2  -39.8681181  9.740546  4.093006  4.258159e-05
## 6 obj1Xobj2   -3.9715006  3.905747  1.016835  3.092318e-01
## 7      3way   -0.7338148  2.067999  0.354843  7.227072e-01
```

Probe Effects

- Interpretation: Overall, evolved weights are most beneficial when optimizing on DP, regardless of what the other objective is.
  Note that the point estimates displayed are medians (not means), and the intervals represent 95% credible intervals (Bayesian).

```
## objective1 = AUROC, objective2 = sFNR:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights    0.00932    -0.121    0.1624
##  Equal Weights - Evolved Weights         -0.05259    -0.193    0.1023
##  Deterministic Weights - Evolved Weights -0.06134    -0.208    0.0852
##
## objective1 = Acc, objective2 = sFNR:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights    0.00204    -0.147    0.1418
##  Equal Weights - Evolved Weights         -0.19728    -0.343   -0.0467
##  Deterministic Weights - Evolved Weights -0.19997    -0.352   -0.0548
##
## objective1 = AUROC, objective2 = DP:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights   -0.04615    -0.173    0.0836
##  Equal Weights - Evolved Weights         -0.74535    -0.875   -0.6136
```

```
##  Deterministic Weights - Evolved Weights -0.70025    -0.818   -0.5535
##
## objective1 = Acc, objective2 = DP:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights   -0.04463    -0.169    0.0800
##  Equal Weights - Evolved Weights         -0.64860    -0.774   -0.5261
##  Deterministic Weights - Evolved Weights -0.60240    -0.721   -0.4749
##
## Note: contrasts are still on the logit scale. Consider using
##       regrid() if you want contrasts of back-transformed estimates.
## Point estimate displayed: median
## HPD interval probability: 0.95
```

```
## objective2 = sFNR:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights    0.00651    -0.094    0.1026
##  Equal Weights - Evolved Weights         -0.12514    -0.226   -0.0210
##  Deterministic Weights - Evolved Weights -0.12992    -0.238   -0.0321
##
## objective2 = DP:
##  contrast                                estimate lower.HPD upper.HPD
##  Equal Weights - Deterministic Weights   -0.04518    -0.136    0.0431
##  Equal Weights - Evolved Weights         -0.69811    -0.785   -0.6036
##  Deterministic Weights - Evolved Weights -0.64969    -0.739   -0.5598
##
## Results are averaged over the levels of: objective1
## Note: contrasts are still on the logit scale. Consider using
##       regrid() if you want contrasts of back-transformed estimates.
## Point estimate displayed: median
## HPD interval probability: 0.95
```

Test of Contrasts

- No difference between equal and deterministic weights.
- Evolved weights are significantly better than equal weights.
- Evolved weights are significantly better than deterministic weights.
- Evolved weights are significantly better than equal & deterministic weights (together).

Please note that the overall effects here are not directly interpretable since the effect of weighting method depends on choice of fairness objective.

```r
emm = emmeans(model, specs = "exp", type = "response")

contrast(emm, list(`Equal vs Deterministic` = c(1, -1, 0), `Equal vs Evolved` = c(1,
    0, -1), `Deterministic vs Evolved` = c(0, 1, -1), `Equal+Deterministic vs Evolved` = c(0.5,
    0.5, -1)), adjust = "tukey", ratio = FALSE)
```

```
##  contrast                       estimate lower.HPD upper.HPD
##  Equal vs Deterministic          -0.0201    -0.087     0.044
##  Equal vs Evolved                -0.4113    -0.478    -0.344
##  Deterministic vs Evolved        -0.3892    -0.462    -0.327
##  Equal+Deterministic vs Evolved  -0.4007    -0.457    -0.340
```

```
##
## Results are averaged over the levels of: objective1, objective2
## Note: contrasts are still on the logit scale. Consider using
##       regrid() if you want contrasts of back-transformed estimates.
## Point estimate displayed: median
## HPD interval probability: 0.95
```