

Projection Prediction Inference

Supplementary material

Contents

Data pre-processing	1
Packages & Options	1
Data	1
Covariate selection	3
Reference Model: 2-level Multilevel Model	4
Projection Prediction Inference	5

Data pre-processing

Packages & Options

```
library(tidyverse)
library(brms)
library(projpred)
# library(doParallel)

# retrieve # of cores
ncores <- parallel::detectCores()
# registerDoParallel(cores = ncores)
options(mc.cores = ncores)

# for output clarity
options(scipen = 999)
```

Data

```
df_trials <- read_csv("../../data/laac_data_trial.csv")

# remove variables not relevant to PPI analysis
df_trials <- df_trials %>%
  select(-c(time_waited, birth_date, comment, experimenter, pick, condition, heat))

# data pre-processing

sum_resp <- function(x, ...) {
  # helper function: sum over `correct response` variable (code)
  to_return = tibble(cogn = sum(x$code))
}

resp_sum <- df_trials %>%
  # contains summed code variable for [task, time point, session, subject]
```

```

group_by(time_point, session, subject, task) %>%
group_modify(sum_resp)

df_tmp <- df_trials %>%
  # helper for merging
  select(-c(date, trial_session, trial_time_point, code)) %>%
  unique(by = c("time_point", "session", "subject"))

df_trials_final <- inner_join(df_tmp, resp_sum)

df_trials_final <- df_trials_final %>%
  # grand mean center below variables
  mutate(across(c(sick_severity,
                  le_mean,
                  dist_mean,
                  time_outdoors,
                  age,
                  time_in_leipzig),
              ~scale(., center = T, scale = F))) %>%
  mutate(across(c(session,
                  subject,
                  group,
                  sex,
                  test_day,
                  le_present,
                  dist_present,
                  rearing,
                  observer),
              as_factor)) %>%
  mutate(observer = fct_relevel(observer, "no")) %>%
  mutate(rearing = fct_recode(rearing, "hand" = "unknown")) %>%
  # create new observer variable due to different measurements of observer between phases
  mutate(observer_mod = case_when(
    observer == "yes" ~ "yes",
    observer == "no" ~ "no",
    observer != "no" & observer != "yes" & observer != "NA" ~ "yes",
    TRUE ~ "no"
  ), observer_mod = as_factor(observer_mod))

grp_size <- tibble(
  # number of apes for each species
  a_chimp = 20,
  b_chimp = 6,
  bonobo = 12,
  gorilla = 6,
  orangutan = 6
)

df_trials_final <- df_trials_final %>%
  # relative rank of ape within species (varies between time points)
  group_by(group, time_point) %>%
  mutate(
    rel_rank = case_when(

```

```

    group == "a_chimp" ~ percent_rank(grp_size$a_chimp:1)[rank],
    group == "b_chimp" ~ percent_rank(grp_size$b_chimp:1)[rank],
    group == "bonobo" ~ percent_rank(grp_size$bonobo:1)[rank],
    group == "gorilla" ~ percent_rank(grp_size$gorilla:1)[rank],
    group == "orangutan" ~ percent_rank(grp_size$orangutan:1)[rank]
  )
) %>%
ungroup()

# create complete subsets for each task
t_cau <- filter(df_trials_final, task == "causality")
t_inf <- filter(df_trials_final, task == "inference")
t_quant <- filter(df_trials_final, task == "quantity")
t_gaze <- filter(df_trials_final, task == "gaze_following")
t_grat <- filter(df_trials_final, task == "delay_of_gratification")

t_gaze <- t_gaze %>%
  # create dummy variable indicating if in session 1 or 2
  group_by(time_point, session) %>%
  mutate(tp_mod = cur_group_id()) %>%
  ungroup() %>%
  mutate(day2 = case_when(session == 1 ~ "no",
                          session == 2 ~ "yes"),
         day2 = factor(day2)) %>%
  select(tp_mod, day2, everything())

t_gaze <- t_gaze %>%
  # remove duplicates created by day2
  group_by(subject) %>%
  filter(!duplicated(tp_mod)) %>%
  ungroup()

# filter data to only include time points from phase 1
t_cau_p1 <- filter(t_cau, time_point < 15)
t_inf_p1 <- filter(t_inf, time_point < 15)
t_quant_p1 <- filter(t_quant, time_point < 15)
t_gaze_p1 <- filter(t_gaze, time_point < 15)

# filter data to only include time points from phase 2
t_cau_p2 <- filter(t_cau, time_point >= 15)
t_inf_p2 <- filter(t_inf, time_point >= 15)
t_quant_p2 <- filter(t_quant, time_point >= 15)
t_gaze_p2 <- filter(t_gaze, time_point >= 15)
t_grat_p2 <- filter(t_grat, time_point >= 15)

```

Covariate selection

```

# formula for reference models
fm <- formula(cogn ~ sick_severity +
              test_tp + test_day +
              rel_rank +
              observer_mod +

```

```

      age + time_in_leipzig +
      sex + group +
      rearing +
      le_mean + dist_mean +
      time_outdoors +
      sociality +
      (1|subject))
# formula for gaze reference model
fm_gaze <- formula(cogn ~ sick_severity +
      test_tp + test_day + time_point + day2 +
      rel_rank +
      observer_mod +
      age + time_in_leipzig +
      sex + group +
      rearing +
      le_mean + dist_mean +
      time_outdoors +
      sociality +
      (1 + time_point|subject))

```

Reference Model: 2-level Multilevel Model

```

m_cau_2l_p1 <- brm(fm, data = t_cau_p1,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2021
)

```

Warning: Rows containing NAs were excluded from the model.

```

m_inf_2l_p1 <- brm(fm, data = t_inf_p1,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2021
)

```

Warning: Rows containing NAs were excluded from the model.

```

m_quant_2l_p1 <- brm(fm, data = t_quant_p1,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2021
)

```

Warning: Rows containing NAs were excluded from the model.

```

m_gaze_2l_p1 <- brm(fm_gaze, data = t_gaze_p1,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2021
)

```

Warning: Rows containing NAs were excluded from the model.

```

m_cau_2l_p2 <- brm(fm, data = t_cau_p2,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2022
)

```

Warning: Rows containing NAs were excluded from the model.

```
m_inf_2l_p2 <- brm(fm, data = t_inf_p2,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2022
)
```

Warning: Rows containing NAs were excluded from the model.

```
m_quant_2l_p2 <- brm(fm, data = t_quant_p2,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2022
)
m_gaze_2l_p2 <- brm(fm_gaze, data = t_gaze_p2,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2022
)
```

Warning: Rows containing NAs were excluded from the model.

```
m_grat_2l_p2 <- brm(fm, data = t_grat_p2,
  warmup = 1e3, iter = 4e3, cores = ncores, chains = 4,
  seed = 2022
)
```

Warning: Rows containing NAs were excluded from the model.

```
summary(m_cau_2l_p1)
summary(m_inf_2l_p1)
summary(m_quant_2l_p1)
summary(m_gaze_2l_p1)

summary(m_cau_2l_p2)
summary(m_inf_2l_p2)
summary(m_quant_2l_p2)
summary(m_gaze_2l_p2)
summary(m_grat_2l_p2)

loo_p1 <- lapply(list(m_cau_2l_p1, m_inf_2l_p1, m_quant_2l_p1, m_gaze_2l_p1), loo::loo)
loo_p2 <- lapply(list(m_cau_2l_p2, m_inf_2l_p2, m_quant_2l_p2, m_gaze_2l_p2, m_grat_2l_p2), loo::loo)
```

Projection Prediction Inference

```
all_fixed_effects <- c("sick_severity",
  "test_tp", "test_day",
  "rel_rank",
  "observer_mod",
  "age", "time_in_leipzig",
  "sex", "group",
  "rearing",
  "le_mean", "dist_mean",
  "time_outdoors",
  "sociality")
# delay random intercept to last place so that it doesn't soak up all the variance
s_terms <- c("1", all_fixed_effects,
  paste0(paste(all_fixed_effects, collapse = " + "),
    " + (1 | subject)"))
```

```
# gaze task gets its own search terms vector (including day2 and time_point (fixed/random))
s_terms_gaze <- c("1", c(all_fixed_effects, "day2", "time_point"),
  paste0(paste(all_fixed_effects, collapse = " + "),
    " + (1 | subject)"),
  paste0(paste(all_fixed_effects, collapse = " + "),
    " + (time_point | subject)"))
```

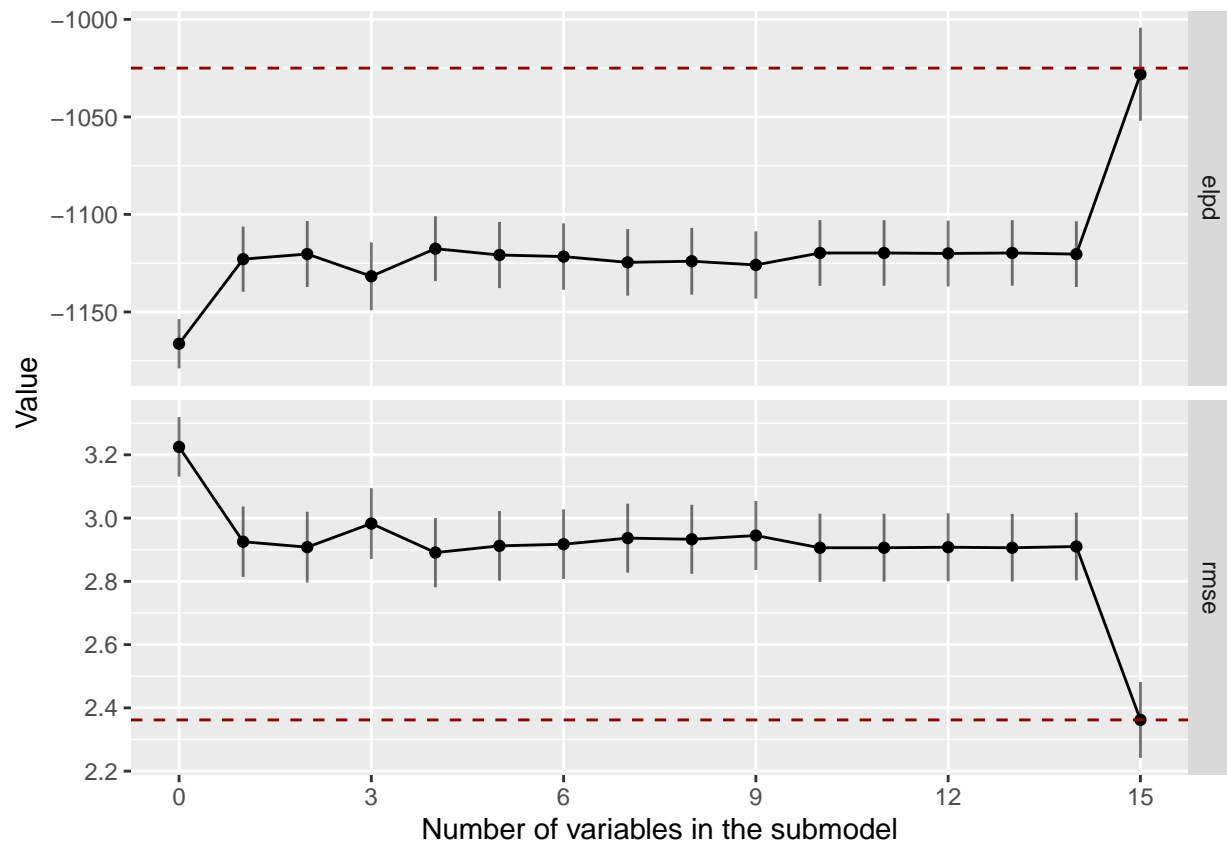
Phase 1

```
cvs_cau_p1 <- cv_varsel(m_cau_2l_p1,
  search_terms = s_terms,
  cv_method = "L00", method = "forward",
  seed = 2020)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
## [1] "Computing L00s..."
## |
```

```
summary(cvs_cau_p1)
```

```
##      size  solution_terms      elpd  elpd.se
## 2      0      <NA> -1166.284 12.72482
## 3      1      group -1122.926 16.80129
## 4      2 time_in_leipzig -1120.298 16.99272
## 5      3      sociality -1131.701 17.49355
## 6      4      dist_mean -1117.568 16.73176
## 7      5      sick_severity -1120.800 17.06588
## 8      6      age -1121.606 17.13295
## 9      7      test_day -1124.578 17.15932
## 10     8      le_mean -1124.001 17.19919
## 11     9      rel_rank -1125.890 17.34207
## 12    10      observer_mod -1119.764 16.95308
## 13    11      time_outdoors -1119.762 16.92067
## 14    12      sex -1120.003 16.96137
## 15    13      rearing -1119.745 16.90651
## 16    14      test_tp -1120.351 16.94601
## 17    15      (1 | subject) -1028.134 24.02604
plot(cvs_cau_p1, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), group

```
cv_inf_p1 <- cv_varsel(m_inf_2l_p1,
  search_terms = s_terms,
  cv_method = "L00", method = "forward",
  seed = 2020)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
```

```
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for details.
```

```
## [1] "Computing L00s..."
```

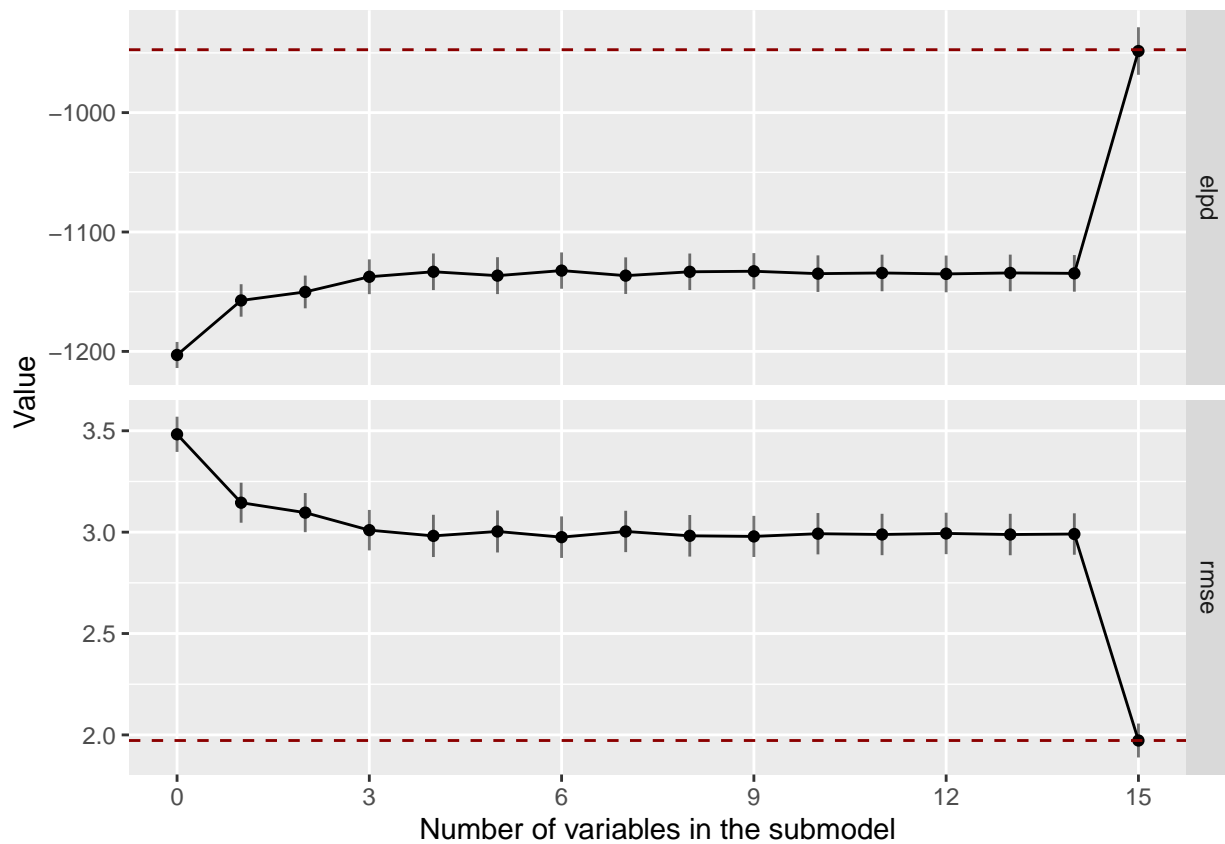
```
##      |
```

```
summary(cv_inf_p1)
```

```
##      size solution_terms      elpd elpd.se
```

```
## 2      0      <NA> -1202.9839 10.97475
## 3      1 time_in_leipzig -1157.3107 13.67198
## 4      2      group -1150.1792 13.80130
## 5      3      age -1137.4899 14.64495
## 6      4 sick_severity -1133.2909 15.41999
## 7      5      test_day -1136.5384 15.52882
## 8      6 time_outdoors -1132.3066 15.28879
## 9      7 observer_mod -1136.5442 15.41194
## 10     8      sociality -1133.3114 15.38975
## 11     9      rearing -1132.8525 15.25838
## 12    10      rel_rank -1134.8773 15.41398
## 13    11      dist_mean -1134.2995 15.41905
## 14    12      test_tp -1135.0920 15.47459
## 15    13      le_mean -1134.2714 15.43826
## 16    14      sex -1134.6379 15.44841
## 17    15 (1 | subject) -948.4261 20.02142
```

```
plot(cvs_inf_p1, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), time_in_leipzig, group, age

```
cvs_quant_p1 <- cv_varsel(m_quant_2l_p1,
  search_terms = s_terms,
  cv_method = "L00", method = "forward",
  seed = 2020)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
```

```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

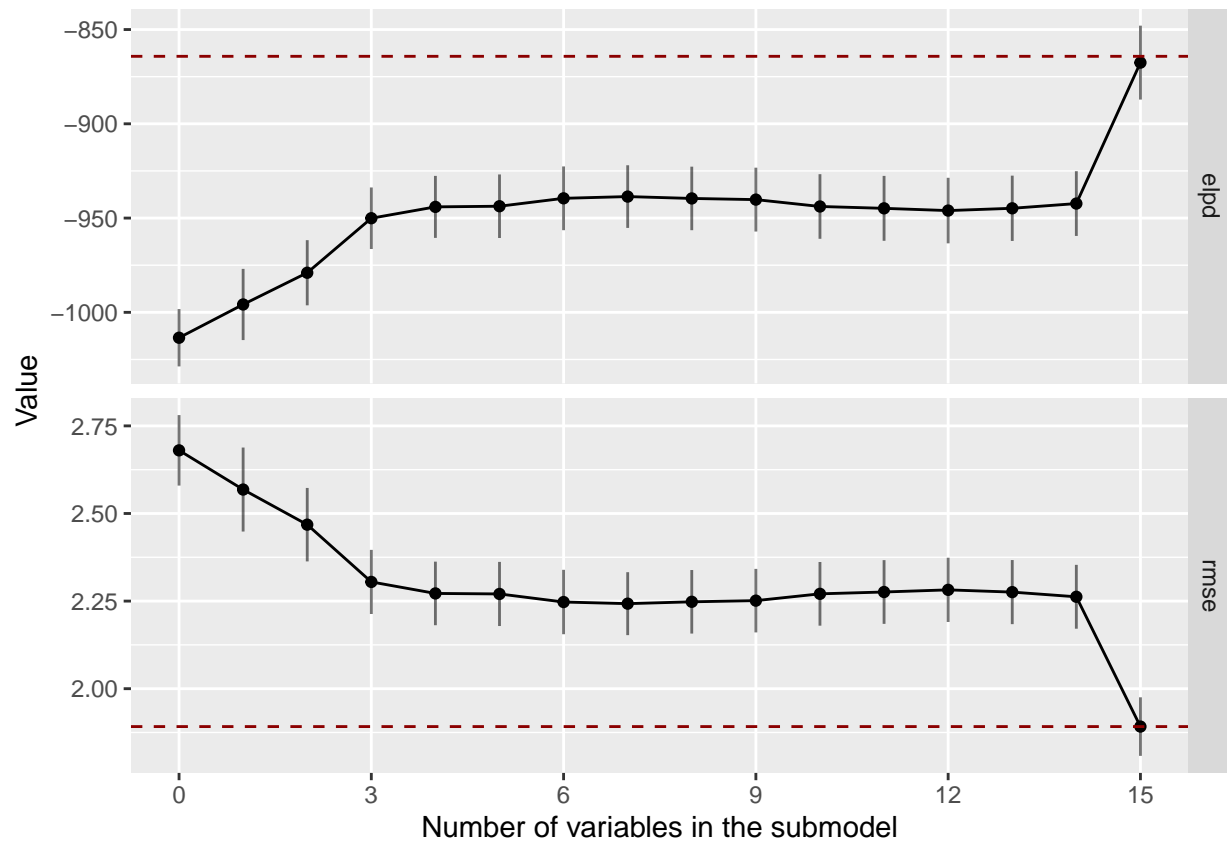


```
## [1] "Computing LOOs..."
## |
```

```
summary(cvs_quant_p1)
```

##	size	solution_terms	elpd	elpd.se
## 2	0	<NA>	-1013.4742	15.29171
## 3	1	time_in_leipzig	-995.8257	18.98727
## 4	2	rearing	-979.0000	17.37981
## 5	3	group	-950.0854	16.41186
## 6	4	observer_mod	-944.0649	16.53589
## 7	5	sex	-943.7152	16.95251
## 8	6	rel_rank	-939.5206	17.01434
## 9	7	age	-938.6060	16.71206
## 10	8	sick_severity	-939.5859	16.94362
## 11	9	test_tp	-940.2080	17.02886
## 12	10	test_day	-943.8538	17.23026
## 13	11	time_outdoors	-944.8374	17.30622
## 14	12	dist_mean	-946.0202	17.48274
## 15	13	sociality	-944.8157	17.41873
## 16	14	le_mean	-942.3036	17.26396
## 17	15	(1 subject)	-867.5398	19.71658

```
plot(cvs_quant_p1, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), time_in_leipzig, rearing, group

```
cvs_gaze_p1 <- cv_varsel(m_gaze_2l_p1,
  search_terms = s_terms_gaze,
  cv_method = "L00", method = "forward",
  seed = 2020)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
```

```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

```
## [1] "Computing L00s..."
```

```
##      |
```

```
summary(cvs_gaze_p1)
```

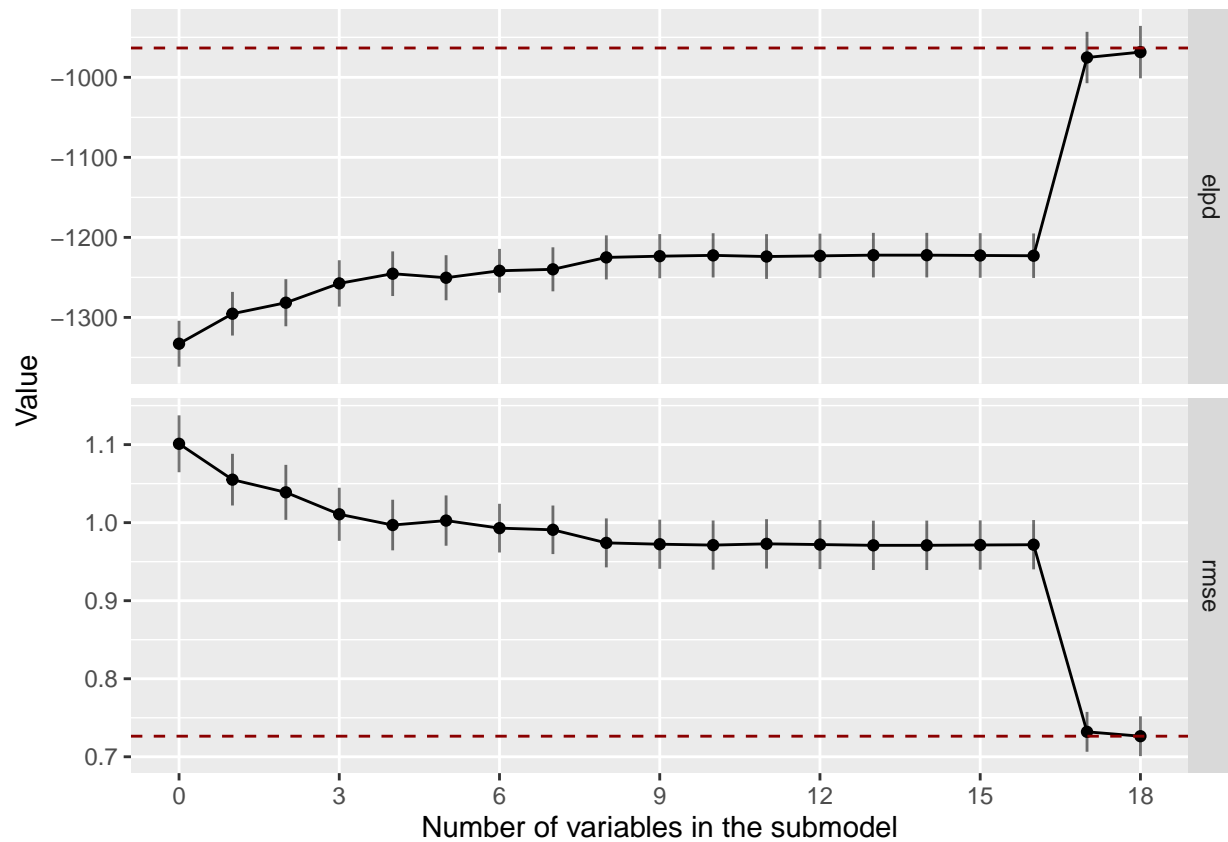
```
##      size      solution_terms      elpd  elpd.se
```

```

## 2    0                <NA> -1332.8640 28.82966
## 3    1                group -1295.3877 27.42496
## 4    2                rearing -1281.6340 29.58314
## 5    3            time_outdoors -1257.5132 29.12716
## 6    4                 age -1245.3894 28.05369
## 7    5            sociality -1250.4262 28.32989
## 8    6                 sex -1241.7161 27.47326
## 9    7            sick_severity -1239.9077 27.67060
## 10   8            observer_mod -1225.0218 27.67382
## 11   9        time_in_leipzig -1223.4863 27.78601
## 12  10                 day2 -1222.4733 27.83285
## 13  11             dist_mean -1223.9155 28.11508
## 14  12             time_point -1223.0766 27.91434
## 15  13             test_day -1222.2077 28.02984
## 16  14             test_tp -1222.2278 28.04756
## 17  15             rel_rank -1222.5950 28.00839
## 18  16             le_mean -1222.9229 28.02147
## 19  17      (1 | subject)  -975.2275 32.28426
## 20  18 (time_point | subject) -968.4631 32.93261

```

```
plot(cvs_gaze_p1, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), group, rearing, time_outdoors, age, sociality, sex, sick_severity, observer_mod

Phase 2

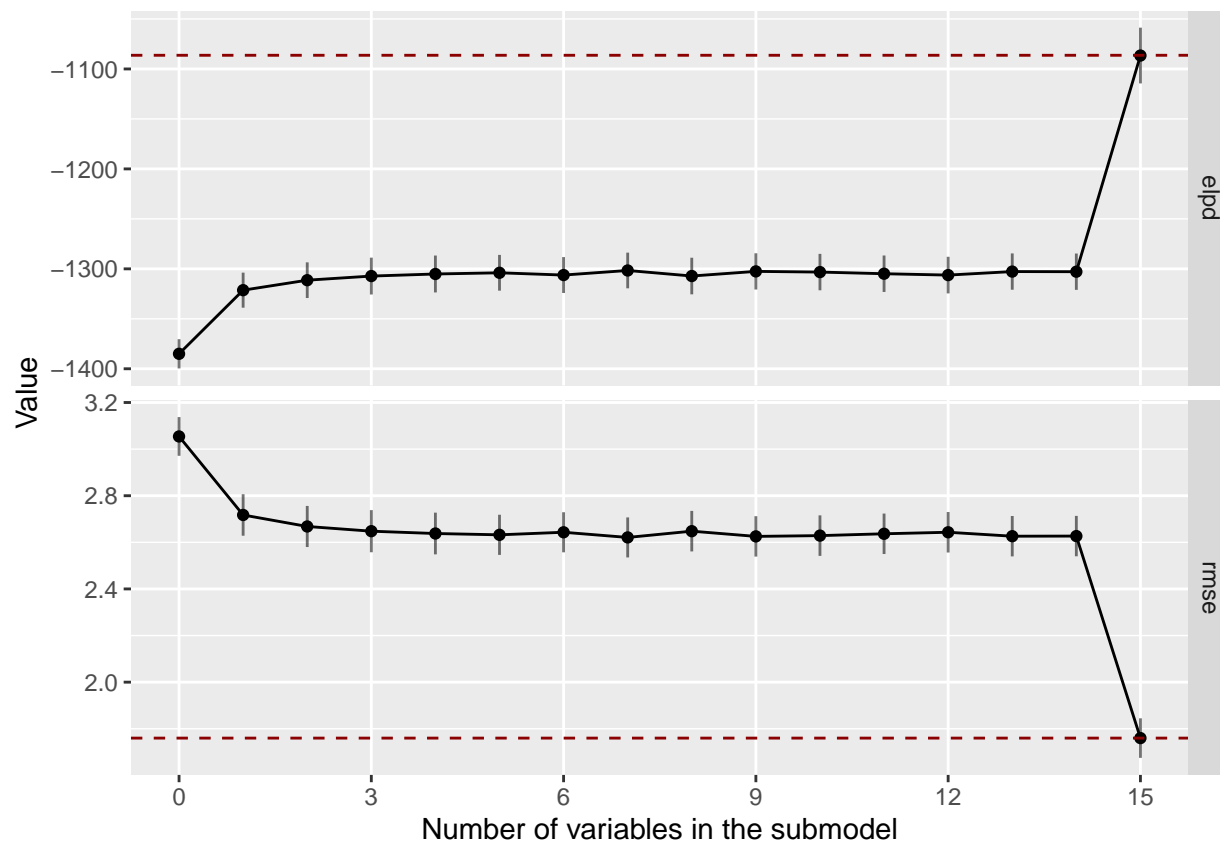
```
cvs_cau_p2 <- cv_varsel(m_cau_2l_p2,
                        search_terms = s_terms,
                        cv_method = "L00", method = "forward",
                        seed = 2022)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
## [1] "Computing L00s..."
##    |
```

```
summary(cvs_cau_p2)
```

```
##    size solution_terms      elpd elpd.se
## 2     0          <NA> -1385.016 14.69496
## 3     1           group -1321.352 17.64576
## 4     2  time_outdoors -1311.360 17.94981
## 5     3 time_in_leipzig -1307.235 18.54256
## 6     4            age -1305.158 18.56643
## 7     5          rearing -1303.951 17.98294
## 8     6  observer_mod -1306.220 18.07087
## 9     7  sick_severity -1301.649 17.94775
## 10    8    sociality -1307.202 18.40965
## 11    9           sex -1302.558 18.23648
## 12   10      test_day -1303.249 18.32480
## 13   11      rel_rank -1304.890 18.34704
## 14   12      dist_mean -1306.265 18.43430
## 15   13      test_tp -1302.741 18.27451
## 16   14      le_mean -1302.859 18.27852
## 17   15  (1 | subject) -1086.583 28.06979
```

```
plot(cvs_cau_p2, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), group, time_outdoors, (time_in_leipzig?)

```
cv_inf_p2 <- cv_varsel(m_inf_2l_p2,
  search_terms = s_terms,
  cv_method = "L00", method = "forward",
  seed = 2022)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 600000)
```

```
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for details.
```

```
## [1] "Computing L00s..."
```

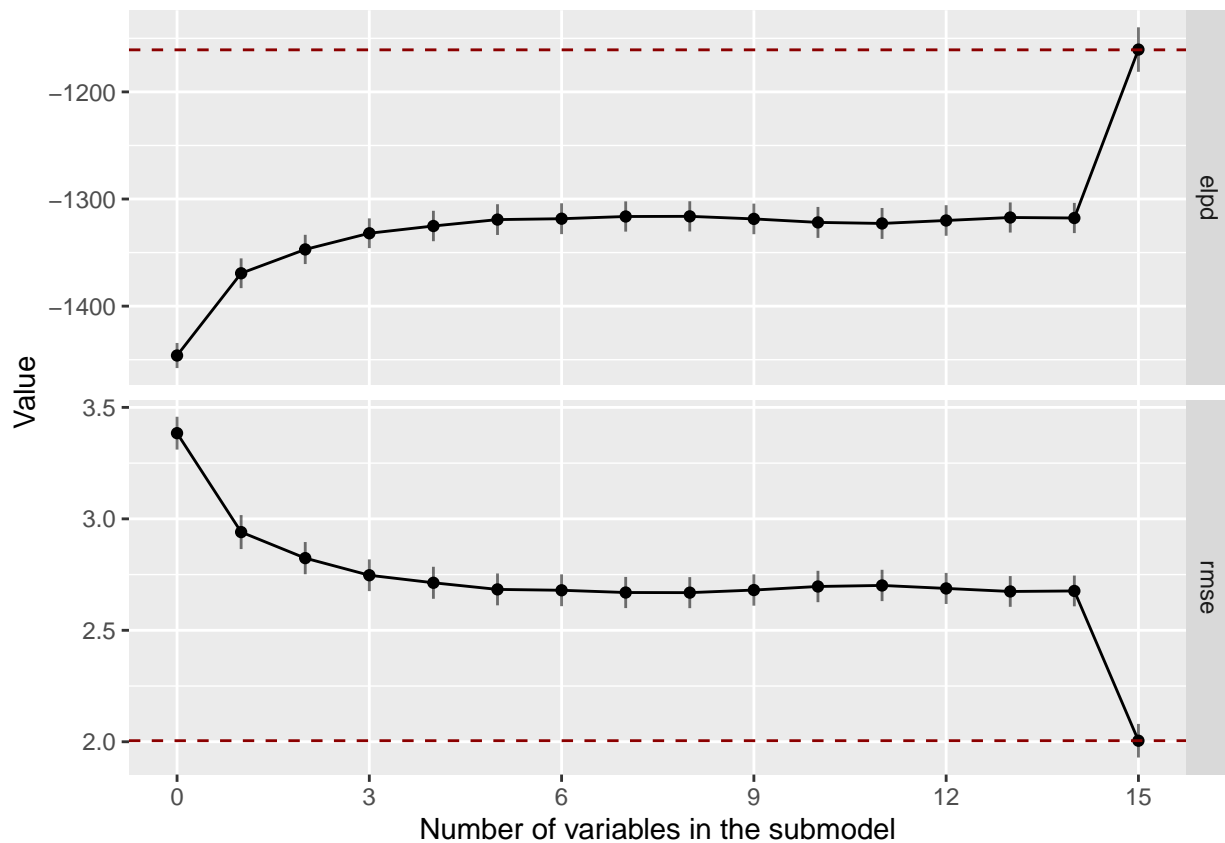
```
## |
```

```
|
```

```
summary(cvs_inf_p2)
```

```
##      size solution_terms      elpd elpd.se
## 2      0              <NA> -1446.101 11.77939
## 3      1 time_in_leipzig -1369.305 13.97695
## 4      2              group -1347.068 13.73156
## 5      3              age -1331.969 13.97187
## 6      4 time_outdoors -1325.193 14.28571
## 7      5              sex -1319.204 14.43684
## 8      6      sociality -1318.384 14.48859
## 9      7 sick_severity -1316.323 14.19400
## 10     8      rearing -1316.190 14.17482
## 11     9      dist_mean -1318.566 14.36981
## 12    10      le_mean -1321.842 14.52652
## 13    11      rel_rank -1322.805 14.52641
## 14    12      test_day -1320.042 14.34843
## 15    13 observer_mod -1317.258 14.14640
## 16    14      test_tp -1317.728 14.16571
## 17    15 (1 | subject) -1160.477 20.89182
```

```
plot(cvs_inf_p2, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), time_in_leipzig, group, age, time_outdoors

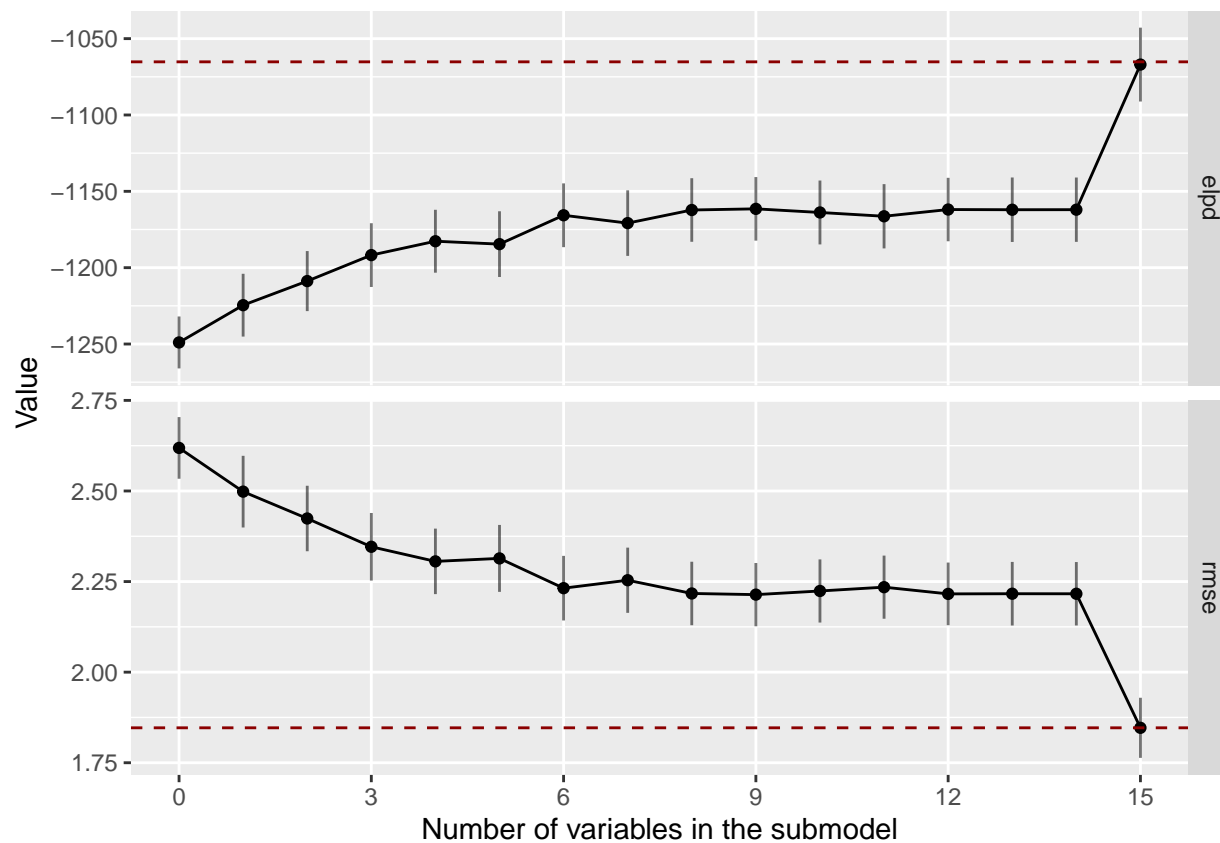
```
cvs_quant_p2 <- cv_varsel(m_quant_21_p2,
  search_terms = s_terms,
  cv_method = "L00", method = "forward",
  seed = 2022)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 600000)
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for deta
## [1] "Computing LOOs..."
##      |
```

```
summary(cvs_quant_p2)
```

```
##      size  solution_terms      elpd  elpd.se
## 2      0      <NA> -1248.971 17.10132
## 3      1      rel_rank -1224.566 20.68574
## 4      2      rearing -1208.788 19.77086
## 5      3 time_in_leipzig -1191.770 20.98866
## 6      4      group -1182.687 20.72264
## 7      5  observer_mod -1184.582 21.63684
## 8      6  time_outdoors -1165.682 21.00602
## 9      7      test_tp -1170.812 21.58873
## 10     8      dist_mean -1162.232 20.93727
## 11     9      test_day -1161.480 20.93512
## 12    10      le_mean -1163.851 21.03599
## 13    11  sick_severity -1166.339 21.17300
## 14    12      age -1161.928 20.89748
## 15    13      sex -1162.056 21.23606
## 16    14  sociality -1162.023 21.19257
## 17    15  (1 | subject) -1066.990 24.32482
```

```
plot(cvs_quant_p2, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), rel_rank, rearing, time_in_leipzig, group, time_outdoors, test_tp

```
cv_s_gaze_p2 <- cv_varsel(m_gaze_21_p2,
  search_terms = s_terms_gaze,
  cv_method = "L00", method = "forward",
  seed = 2022)
```

```
## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 600000)
```

```
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for details.
```

```
## [1] "Computing L00s..."
```

```
## |
```

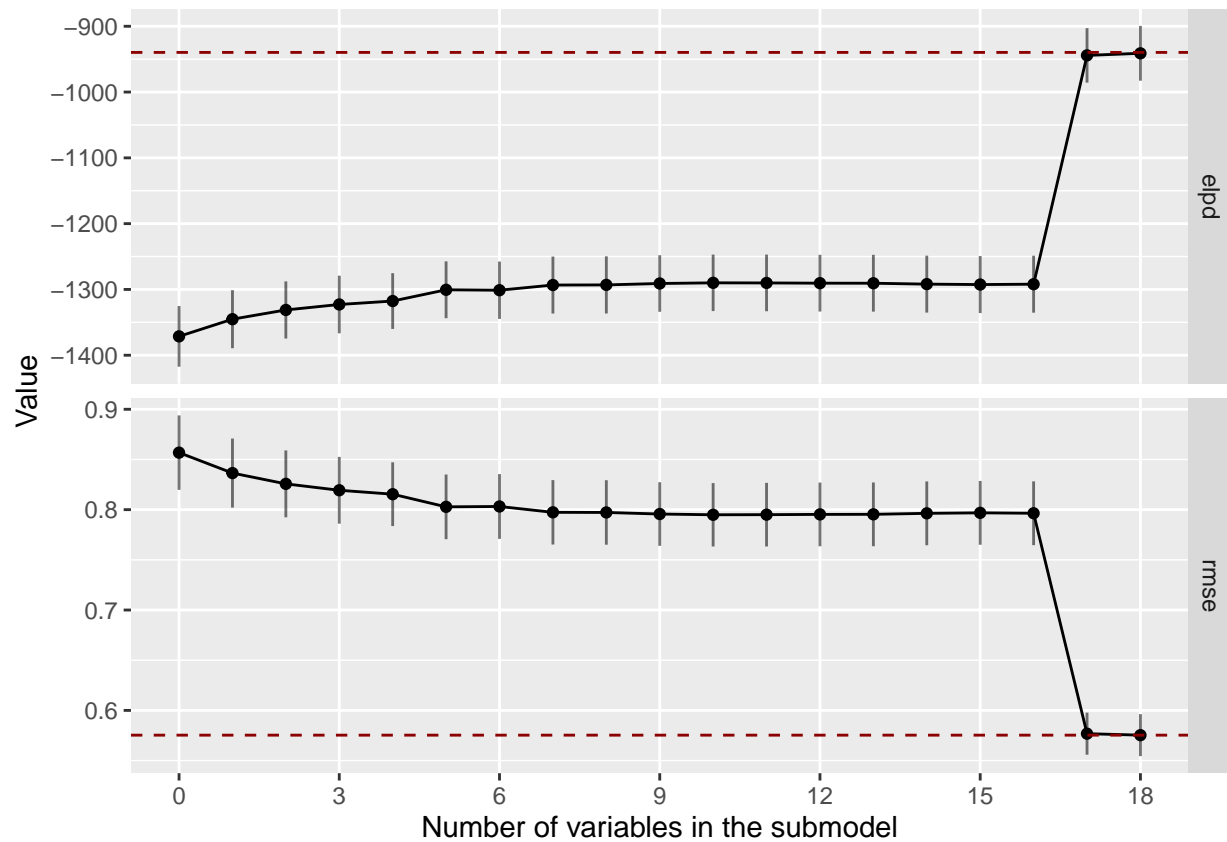
```
|
```



```
summary(cvs_gaze_p2)
```

##	size	solution_terms	elpd	elpd.se
## 2	0	<NA>	-1371.3280	46.37679
## 3	1	group	-1345.2073	44.37327
## 4	2	sex	-1331.1870	43.67917
## 5	3	observer_mod	-1322.8534	44.06713
## 6	4	age	-1317.6484	42.56238
## 7	5	rearing	-1300.5452	43.45194
## 8	6	sick_severity	-1301.1635	43.83051
## 9	7	sociality	-1293.3058	43.57111
## 10	8	time_in_leipzig	-1293.1034	43.69256
## 11	9	rel_rank	-1290.9634	43.25502
## 12	10	test_day	-1289.9420	43.15980
## 13	11	day2	-1290.1149	43.27658
## 14	12	time_point	-1290.4706	43.39314
## 15	13	le_mean	-1290.5677	43.41433
## 16	14	dist_mean	-1291.9200	43.52289
## 17	15	test_tp	-1292.5802	43.56191
## 18	16	time_outdoors	-1292.0216	43.54598
## 19	17	(1 subject)	-944.2455	41.60639
## 20	18	(time_point subject)	-941.1030	41.83415

```
plot(cvs_gaze_p2, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), group, sex, observer_mod, age, (rearing?)

```

cvs_grat_p2 <- cv_varsel(m_grat_2l_p2,
                        search_terms = s_terms,
                        cv_method = "L00", method = "forward",
                        seed = 2022)

## Warning in cv_varsel.refmodel(refmodel, ...): K provided, but cv_method is L00.
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for deta
## [1] "Computing L00s..."
##      |

```

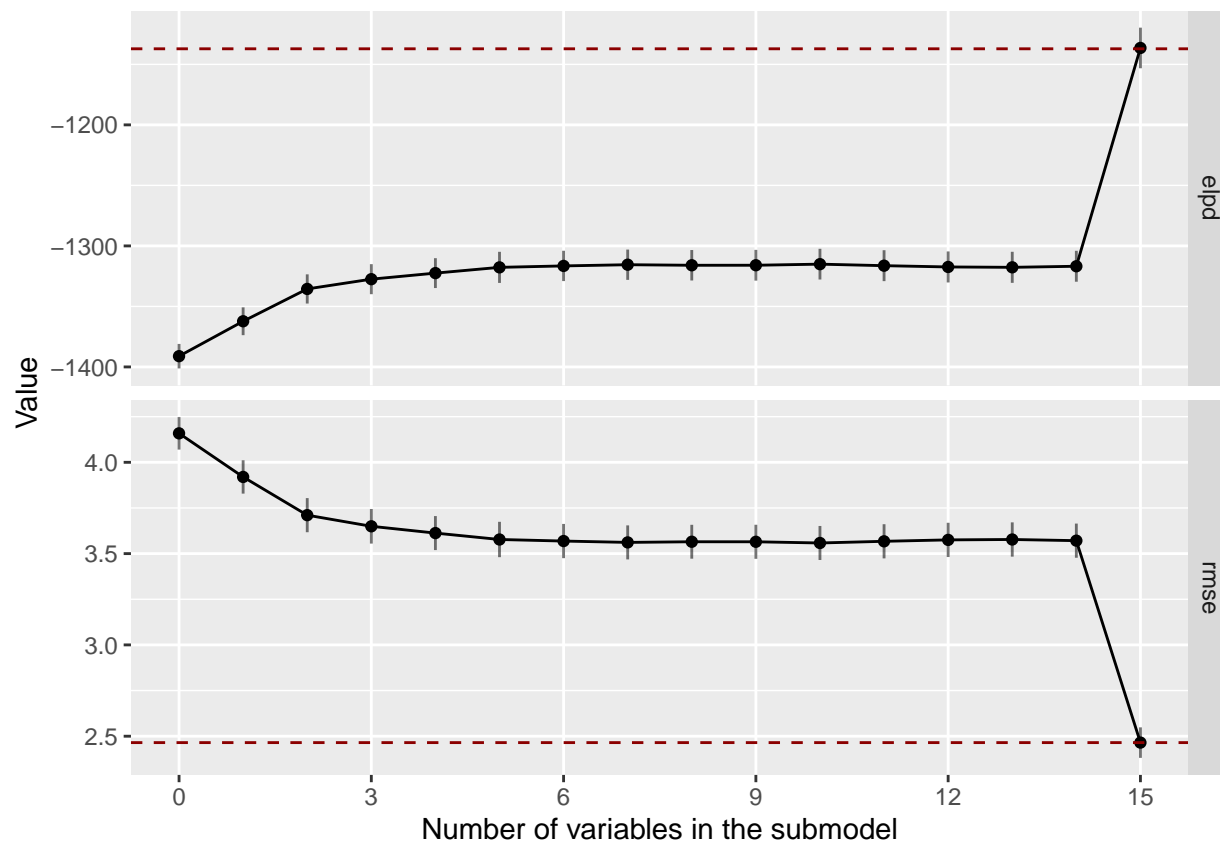
```
summary(cvs_grat_p2)
```

```

##      size solution_terms      elpd elpd.se
## 2      0      <NA> -1391.109 10.17323
## 3      1 time_in_leipzig -1362.270 11.55423
## 4      2  observer_mod -1335.562 12.12460
## 5      3      sex -1327.493 12.48185
## 6      4      rel_rank -1322.476 12.40804
## 7      5 sick_severity -1317.736 12.92059
## 8      6 time_outdoors -1316.543 12.61068
## 9      7      test_tp -1315.540 12.58188
## 10     8      group -1315.999 12.66718
## 11     9      sociality -1315.976 12.78212
## 12    10      test_day -1315.085 12.75679
## 13    11      age -1316.340 12.84615
## 14    12      rearing -1317.399 12.89820
## 15    13      le_mean -1317.696 12.90063
## 16    14      dist_mean -1316.820 12.85426
## 17    15  (1 | subject) -1136.435 16.87488

```

```
plot(cvs_grat_p2, stats = c('elpd', 'rmse'))
```



selected covariates: (1 | subject), time_in_leipzig, observer_mod, sex, rel_rank, sick_severity

```
saveRDS(list(m_cau_2l_p1, m_inf_2l_p1, m_quant_2l_p1, m_gaze_2l_p1,
            m_cau_2l_p2, m_inf_2l_p2, m_quant_2l_p2, m_gaze_2l_p2, m_grat_2l_p2,
            loo_p1, loo_p2),
        file = "results/ref_models.rds")
```

```
saveRDS(list(cvs_cau_p1, cvs_inf_p1, cvs_quant_p1, cvs_gaze_p1,
            cvs_cau_p2, cvs_inf_p2, cvs_quant_p2, cvs_gaze_p2, cvs_grat_p2),
        file = "results/projection_results.rds")
```