EFA

```r
set.seed(42)

library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```r
library(QuantPsyc) # for the multivariate normality test
```

```
## Loading required package: boot
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: purrr
```

```
##
## Attaching package: 'purrr'
```

```
## The following objects are masked from 'package:igraph':
##
##     compose, simplify
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```
library(nFactors) # for the scree plot
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

##
## Attaching package: 'nFactors'

## The following object is masked from 'package:lattice':
##
##     parallel
```

```
library(psych) # for PA FA
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:boot':
##
##     logit
```

```
library(caret) # highly correlated features removal
```

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v readr     2.1.5     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::%--%()       masks igraph::%--%()
## x ggplot2::%+%()          masks psych::%+%()
```

```
## x ggplot2::alpha()        masks psych::alpha()
## x tibble::as_data_frame() masks dplyr::as_data_frame(), igraph::as_data_frame()
## x purrr::compose()        masks igraph::compose()
## x tidyr::crossing()       masks igraph::crossing()
## x dplyr::filter()         masks stats::filter()
## x dplyr::lag()            masks stats::lag()
## x caret::lift()           masks purrr::lift()
## x MASS::select()          masks dplyr::select()
## x purrr::simplify()       masks igraph::simplify()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(paletteer) # color palettes

library(conflicted) # to resolve QuantPsyc x dplyr conflicts
conflict_prefer("select", "dplyr")
```

```
## [conflicted] Will prefer dplyr::select over any other package.
```

```
conflict_prefer("filter", "dplyr")
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

Load and tidy data

```
pretty_names <- read_csv("../feat_name_mapping.csv")
```

```
## Rows: 85 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): name_orig, name_pretty
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data <- read_csv("../measurements/measurements.csv")
```

```
## Rows: 754 Columns: 108
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (20): fpath, KUK_ID, FileName, FileFormat, FolderPath, subcorpus, Source...
## dbl (85): RuleAbstractNouns, RuleAmbiguousRegards, RuleAnaphoricReferences, ...
## lgl  (3): ClarityPursuit, SyllogismBased, Bindingness
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
.firstnonmetacolumn <- 17

data_no_nas <- data %>%
  select(!c(
    fpath,
    # KUK_ID,
    # FileName,
    FolderPath,
    # subcorpus,
    DocumentTitle,
```

```
  ClarityPursuit,
  Readability,
  SyllogismBased,
  SourceDB
)) %>%
# replace -1s in variation coefficients with NAs
mutate(across(c(
  `RuleDoubleAdpos.max_allowable_distance.v`,
  `RuleTooManyNegations.max_negation_frac.v`,
  `RuleTooManyNegations.max_allowable_negations.v`,
  `RuleTooManyNominalConstructions.max_noun_frac.v`,
  `RuleTooManyNominalConstructions.max_allowable_nouns.v`,
  `RuleCaseRepetition.max_repetition_count.v`,
  `RuleCaseRepetition.max_repetition_frac.v`,
  `RulePredSubjDistance.max_distance.v`,
  `RulePredObjDistance.max_distance.v`,
  `RuleInfVerbDistance.max_distance.v`,
  `RuleMultiPartVerbs.max_distance.v`,
  `RuleLongSentences.max_length.v`,
  `RulePredAtClauseBeginning.max_order.v`,
  `mattr.v`,
  `maentropy.v`
), ~ na_if(.x, -1))) %>%
# replace NAs with 0s
replace_na(list(
  RuleGPcoordovs = 0,
  RuleGPdeverbaddr = 0,
  RuleGPpatinstr = 0,
  RuleGPdeverbsubj = 0,
  RuleGPadjective = 0,
  RuleGPpatbenperson = 0,
  RuleGPwordorder = 0,
  RuleDoubleAdpos = 0,
  RuleDoubleAdpos.max_allowable_distance = 0,
  RuleDoubleAdpos.max_allowable_distance.v = 0,
  RuleAmbiguousRegards = 0,
  RuleReflexivePassWithAnimSubj = 0,
  RuleTooManyNegations = 0,
  RuleTooManyNegations.max_negation_frac = 0,
  RuleTooManyNegations.max_negation_frac.v = 0,
  RuleTooManyNegations.max_allowable_negations = 0,
  RuleTooManyNegations.max_allowable_negations.v = 0,
  RuleTooManyNominalConstructions.max_noun_frac.v = 0,
  RuleTooManyNominalConstructions.max_allowable_nouns.v = 0,
  RuleFunctionWordRepetition = 0,
  RuleCaseRepetition.max_repetition_count.v = 0,
  RuleCaseRepetition.max_repetition_frac.v = 0,
  RuleWeakMeaningWords = 0,
  RuleAbstractNouns = 0,
  RuleRelativisticExpressions = 0,
  RuleConfirmationExpressions = 0,
  RuleRedundantExpressions = 0,
  RuleTooLongExpressions = 0,
```

```
    RuleAnaphoricReferences = 0,
    RuleLiteraryStyle = 0,
    RulePassive = 0,
    RulePredSubjDistance = 0,
    RulePredSubjDistance.max_distance = 0,
    RulePredSubjDistance.max_distance.v = 0,
    RulePredObjDistance = 0,
    RulePredObjDistance.max_distance = 0,
    RulePredObjDistance.max_distance.v = 0,
    RuleInfVerbDistance = 0,
    RuleInfVerbDistance.max_distance = 0,
    RuleInfVerbDistance.max_distance.v = 0,
    RuleMultiPartVerbs = 0,
    RuleMultiPartVerbs.max_distance = 0,
    RuleMultiPartVerbs.max_distance.v = 0,
    RuleLongSentences.max_length.v = 0,
    RulePredAtClauseBeginning.max_order.v = 0,
    RuleVerbalNouns = 0,
    RuleDoubleComparison = 0,
    RuleWrongValencyCase = 0,
    RuleWrongVerbonominalCase = 0,
    RuleIncompleteConjunction = 0
  ))

data_clean <- data_no_nas %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    RuleGPcoordovs,
    RuleGPdeverbaddr,
    RuleGPpatinstr,
    RuleGPdeverbsubj,
    RuleGPadjective,
    RuleGPpatbenperson,
    RuleGPwordorder,
    RuleDoubleAdpos,
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleWeakMeaningWords,
    RuleAbstractNouns,
    RuleRelativisticExpressions,
    RuleConfirmationExpressions,
    RuleRedundantExpressions,
    RuleTooLongExpressions,
    RuleAnaphoricReferences,
    RuleLiteraryStyle,
    RulePassive,
    RuleVerbalNouns,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase,
    RuleIncompleteConjunction,
    num_hapax,
    RuleReflexivePassWithAnimSubj,
```

```
    RuleTooManyNominalConstructions,
    RulePredSubjDistance,
    RuleMultiPartVerbs,
    RulePredAtClauseBeginning
  ), ~ .x / word_count)) %>%
  mutate(across(c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredObjDistance,
    RuleInfVerbDistance
  ), ~ .x / sent_count)) %>%
  # remove variables identified as "u counts"
  select(!c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleTooManyNominalConstructions,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredAtClauseBeginning,
    sent_count,
    word_count,
    syllab_count,
    char_count
  )) %>%
  # remove variables identified as unreliable
  select(!c(
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase
  )) %>%
  # remove artificially limited variables
  select(!c(
    RuleCaseRepetition.max_repetition_frac,
    RuleCaseRepetition.max_repetition_frac.v
  )) %>%
  # remove further variables belonging to the 'acceptability' category
  select(!c(RuleIncompleteConjunction)) %>%
  mutate(across(c(class), ~ as.factor(.x)))

# no NAs should be present now
data_clean[!complete.cases(data_clean), ]
```

```
## # A tibble: 754 x 83
##    KUK_ID                FileName FileFormat subcorpus SourceID DocumentVersion
##    <chr>                 <chr>    <chr>      <chr>     <chr>    <chr>
##  1 673b7a37c6537d54ff062~ 002_Kom~ TXT        KUKY      <NA>     Original
##  2 673b7a37c6537d54ff062~ 006_Chc~ TXT        KUKY      <NA>     Redesign
##  3 673b7a37c6537d54ff062~ 004_Nev~ TXT        KUKY      <NA>     Original
##  4 673b7a37c6537d54ff062~ 008_Pol~ TXT        KUKY      <NA>     Original
##  5 673b7a37c6537d54ff062~ 005_Och~ TXT        KUKY      <NA>     Original
```

```
##  6 673b7a37c6537d54ff062~ 016_Obc~ TXT        KUKY      <NA>     Original
##  7 673b7a37c6537d54ff062~ 019_Dět~ TXT        KUKY      <NA>     Redesign
##  8 673b7a37c6537d54ff062~ 007_DŮC~ TXT        KUKY      <NA>     Redesign
##  9 673b7a37c6537d54ff062~ 024_Opa~ TXT        KUKY      <NA>     Original
## 10 673b7a37c6537d54ff062~ 047_Dav~ TXT        KUKY      <NA>     Original
## # i 744 more rows
## # i 77 more variables: ParentDocumentID <chr>, LegalActType <chr>,
## #   Objectivity <chr>, Bindingness <lgl>, AuthorType <chr>,
## #   RecipientType <chr>, RecipientIndividuation <chr>, Anonymized <chr>,
## #   `Recipient Type` <chr>, class <fct>, RuleAbstractNouns <dbl>,
## #   RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>, ...
```

```
data_clean_scaled <- data_clean %>%
  mutate(across(class, ~ .x == "good")) %>%
  mutate(across(.firstnonmetacolumn:length(names(data_clean)), ~ scale(.x)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(.firstnonmetacolumn:length(names(data_clean)),
##   ~scale(.x))`.
## Caused by warning:
## ! Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(.firstnonmetacolumn)
##
##   # Now:
##   data %>% select(all_of(.firstnonmetacolumn))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

Important features identification

```
data_clean_good <- data_clean_scaled %>% filter(class == "good")
data_clean_bad <- data_clean_scaled %>% filter(class == "bad")

feature_importances <- tibble(
  feat_name = character(), p_value = numeric()
)

for (i in .firstnonmetacolumn:ncol(data_clean)) {
  fname <- names(data_clean)[i]

  formula_single <- reformulate(fname, "class")

  glm_model <- glm(formula_single, data_clean, family = "binomial")
  glm_coefficients <- summary(glm_model)$coefficients
  row_index <- which(rownames(glm_coefficients) == fname)
  p_value <- glm_coefficients[row_index, 4]

  feature_importances <- feature_importances %>%
    add_row(feat_name = fname, p_value = p_value)
}
feature_importances
```

```
## # A tibble: 67 x 2
##    feat_name                                   p_value
##    <chr>                                         <dbl>
##  1 RuleAbstractNouns                          0.00187
##  2 RuleAnaphoricReferences                    0.660
##  3 RuleCaseRepetition.max_repetition_count    0.0722
##  4 RuleCaseRepetition.max_repetition_count.v 0.00479
##  5 RuleConfirmationExpressions                0.0985
##  6 RuleDoubleAdpos                            0.312
##  7 RuleDoubleAdpos.max_allowable_distance     0.000154
##  8 RuleDoubleAdpos.max_allowable_distance.v  0.00000356
##  9 RuleGPadjective                            0.380
## 10 RuleGPcoordovs                             0.828
## # i 57 more rows
```

```
selected_features <- feature_importances %>%
  mutate(selected = p_value <= 0.05)
selected_features %>% write_csv("selected_features.csv")
selected_features_names <- selected_features %>%
  filter(selected) %>%
  pull(feat_name)
```

Correlations

See Levshina (2015: 353–54).

```
analyze_correlation <- function(data) {
  cor_matrix <- cor(data)

  cor_tibble_long <- cor_matrix %>%
    as_tibble() %>%
    mutate(feat1 = rownames(cor_matrix)) %>%
    pivot_longer(!feat1, names_to = "feat2", values_to = "cor") %>%
    mutate(abs_cor = abs(cor))

  cor_matrix_upper <- cor_matrix
  cor_matrix_upper[lower.tri(cor_matrix_upper)] <- 0

  cor_tibble_long_upper <- cor_matrix_upper %>%
    as_tibble() %>%
    mutate(feat1 = rownames(cor_matrix)) %>%
    pivot_longer(!feat1, names_to = "feat2", values_to = "cor") %>%
    mutate(abs_cor = abs(cor)) %>%
    filter(feat1 != feat2 & abs_cor > 0)

  list(
    cor_matrix = cor_matrix,
    cor_matrix_upper = cor_matrix_upper,
    cor_tibble_long = cor_tibble_long,
    cor_tibble_long_upper = cor_tibble_long_upper
  )
}
```

```r
data_purish <- data_clean %>% select(any_of(selected_features_names)) %>%
  # remove features expected to have low communalities
  select(!c(
    RuleDoubleAdpos.max_allowable_distance,
    RuleDoubleAdpos.max_allowable_distance.v,
    RuleGPwordorder,
    RuleLiteraryStyle,
    maentropy.v,
    RuleTooManyNegations.max_negation_frac,
    RulePredSubjDistance.max_distance,
    RuleTooManyNegations.max_allowable_negations,
    RuleTooManyNegations.max_allowable_negations.v,
    RuleTooManyNominalConstructions.max_allowable_nouns.v,
    RuleTooFewVerbs.min_verb_frac.v,
    RulePredObjDistance.max_distance.v,
    RulePredObjDistance.max_distance,
    RulePredAtClauseBeginning.max_order.v,
    RuleInfVerbDistance
  )) %>%
  # remove features expected to have low loadings
  select(!c(
    RuleMultiPartVerbs.max_distance.v,
    RulePredSubjDistance.max_distance.v,
    RuleLongSentences.max_length
  ))
```

Extremely non-normal data

```r
# # remove where median == 0?
# keep <- character()
# for (i in seq_along(colnames(data_purish))) {
#   cname <- colnames(data_purish)[i]
#   q <- quantile(data_purish[, i][[1]], probs = 0.10)[[1]]
#   if (q > 0) {
#     keep <- c(keep, cname)
#     cat("keep", cname, "\n")
#   } else {
#     cat("throw out", cname, "\n")
#   }
# }
# data_purish <- data_purish %>% select(any_of(keep))
```

High correlations

```r
.hcorrcutoff <- 0.9

analyze_correlation(data_purish)$cor_tibble_long %>%
  filter(feat1 != feat2 & abs_cor > .hcorrcutoff) %>%
  arrange(feat1, -abs_cor) %>%
  print(n = 100)
```

```
## # A tibble: 16 x 4
##    feat1     feat2       cor abs_cor
```

```
##    <chr>     <chr>     <dbl>   <dbl>
##  1 ari       fkgl      0.984   0.984
##  2 ari       gf        0.978   0.978
##  3 ari       smog      0.951   0.951
##  4 atl       cli       0.960   0.960
##  5 cli       atl       0.960   0.960
##  6 fkgl      ari       0.984   0.984
##  7 fkgl      gf        0.967   0.967
##  8 fkgl      smog      0.949   0.949
##  9 gf        smog      0.987   0.987
## 10 gf        ari       0.978   0.978
## 11 gf        fkgl      0.967   0.967
## 12 maentropy mattr     0.964   0.964
## 13 mattr     maentropy 0.964   0.964
## 14 smog      gf        0.987   0.987
## 15 smog      ari       0.951   0.951
## 16 smog      fkgl      0.949   0.949
```

exclude:

- **ari:** corr. w/ RuleLongSentences.max_length > 0.94; sentence length seems more universal, let's make it a substitute
- **gf:** corr. w/ RuleLongSentences.max_length > 0.92; sentence length seems more universal, let's make it a substitute
- **maentropy:** corr. w/ mattr > 0.96, but mattr is implemented in QuitaUp. besides, the interesting thing about maentropy is its variation
- **smog:** corr. w/ fkgl almost 0.95, but fkgl coefficients adjusted for Czech are available
- **atl:** corr. w/ cli around 0.96; unlike cli, atl is not a readability metric

```
high_correlations <- findCorrelation(cor(data_purish), verbose = TRUE)
```

```
## Compare row 20  and column  26 with corr  0.978
##   Means:  0.399 vs 0.207 so flagging column 20
## Compare row 26  and column  32 with corr  0.987
##   Means:  0.378 vs 0.195 so flagging column 26
## Compare row 32  and column  24 with corr  0.949
##   Means:  0.35 vs 0.184 so flagging column 32
## Compare row 21  and column  22 with corr  0.96
##   Means:  0.27 vs 0.176 so flagging column 21
## Compare row 28  and column  30 with corr  0.964
##   Means:  0.194 vs 0.171 so flagging column 28
## All correlations <= 0.9
```

```
names(data_purish)[high_correlations]
```

```
## [1] "ari"       "gf"        "smog"      "atl"       "maentropy"
```

```
data_pureish_striphigh <- data_purish %>% select(!all_of(high_correlations))

analyze_correlation(data_pureish_striphigh)$cor_tibble_long %>%
  filter(feat1 != feat2 & abs_cor > .hcorrcutoff) %>%
  arrange(feat1, -abs_cor) %>%
  print(n = 100)
```

```
## # A tibble: 0 x 4
## # i 4 variables: feat1 <chr>, feat2 <chr>, cor <dbl>, abs_cor <dbl>
```

Low correlations

```
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

low_correlating_features <- analyze_correlation(data_pureish_striphigh)$
  cor_tibble_long %>%
  filter(feat1 != feat2) %>%
  group_by(feat1) %>%
  summarize(max_cor = max(abs_cor)) %>%
  filter(max_cor < .lcorrcutoff) %>%
  pull(feat1)

feature_importances %>% filter(feat_name %in% low_correlating_features)
```

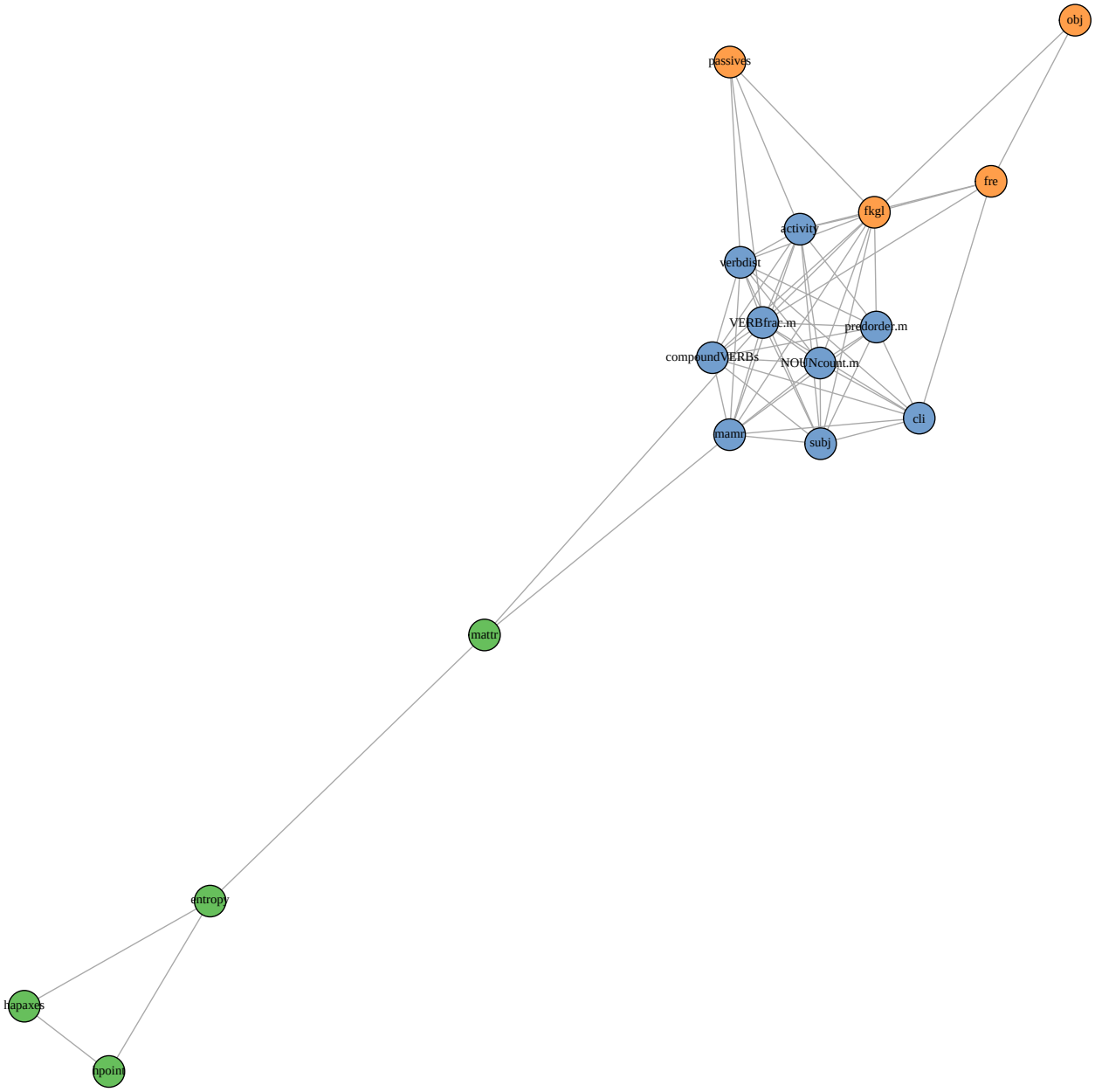
```
## # A tibble: 11 x 2
##    feat_name                                          p_value
##    <chr>                                                <dbl>
##  1 RuleAbstractNouns                                 0.00187
##  2 RuleCaseRepetition.max_repetition_count.v         0.00479
##  3 RuleGPdeverbaddr                                  0.0112
##  4 RuleGPdeverbsubj                                  0.0133
##  5 RuleMultiPartVerbs.max_distance                   0.00320
##  6 RuleRedundantExpressions                          0.0104
##  7 RuleRelativisticExpressions                       0.00205
##  8 RuleTooManyNegations.max_negation_frac.v          0.0365
##  9 RuleTooManyNominalConstructions.max_noun_frac.v 0.00000311
## 10 RuleVerbalNouns                                   0.0000748
## 11 RuleWeakMeaningWords                              0.0386
```

```
data_pure <- data_pureish_striphigh %>%
  select(!any_of(low_correlating_features))

cnames <- map(
  colnames(data_pure),
  function(x) {
    pull(pretty_names %>%
      filter(name_orig == x), name_pretty)
  }
) %>% unlist()

colnames(data_pure) <- cnames
```

Visualisation

```
my_colors <- paletteer::paletteer_d("ggthemes::Classic_10_Medium")

network_edges <- analyze_correlation(data_pure)$cor_tibble_long_upper %>%
  filter(abs_cor > 0.3)

network <- graph_from_data_frame(
  network_edges,
  directed = FALSE
)
```

```r
E(network)$weight <- network_edges$abs_cor
network_communities <- cluster_optimal(network)

network_membership <- membership(network_communities)

plot(
  network,
  layout = layout.fruchterman.reingold,
  vertex.color = map(
    network_communities$membership,
    function(x) my_colors[x]
  ) %>% unlist(use.names = FALSE),
  vertex.size = 6,
  vertex.label.color = "black",
  vertex.label.cex = 0.7
)
```



Scaling

```
data_scaled <- data_pure %>%
  mutate(across(seq_along(data_pure), ~ scale(.x)[, 1]))

final_collist <- data_scaled %>% colnames()
```

Check for normality

```
mult.norm(data_scaled %>% as.data.frame())$mult.test
```

```
##          Beta-hat      kappa p-val
```

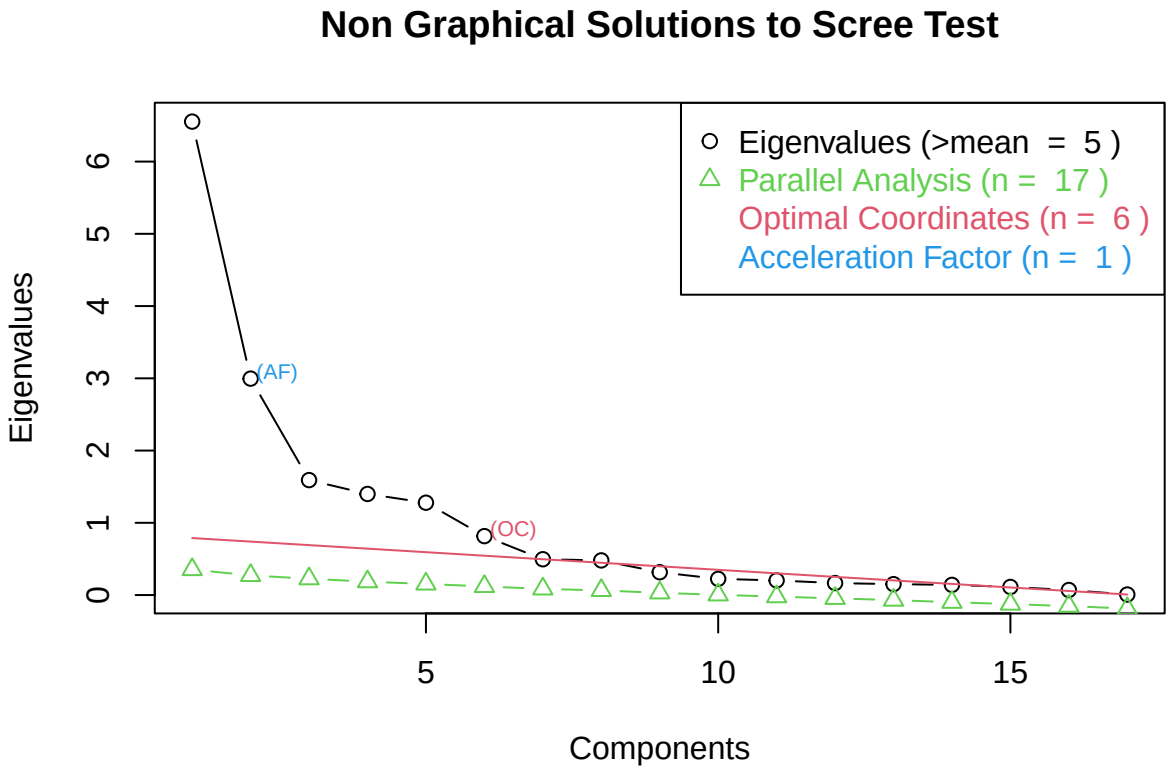
```
## Skewness 351.5182 44174.1153     0
## Kurtosis 858.5678   289.3036     0
```

Low (null) p-values show that we can reject the hypothesis that the data would be in a multivariate normal distribution. I.e. the distribution isn't multivariate normal.

FA

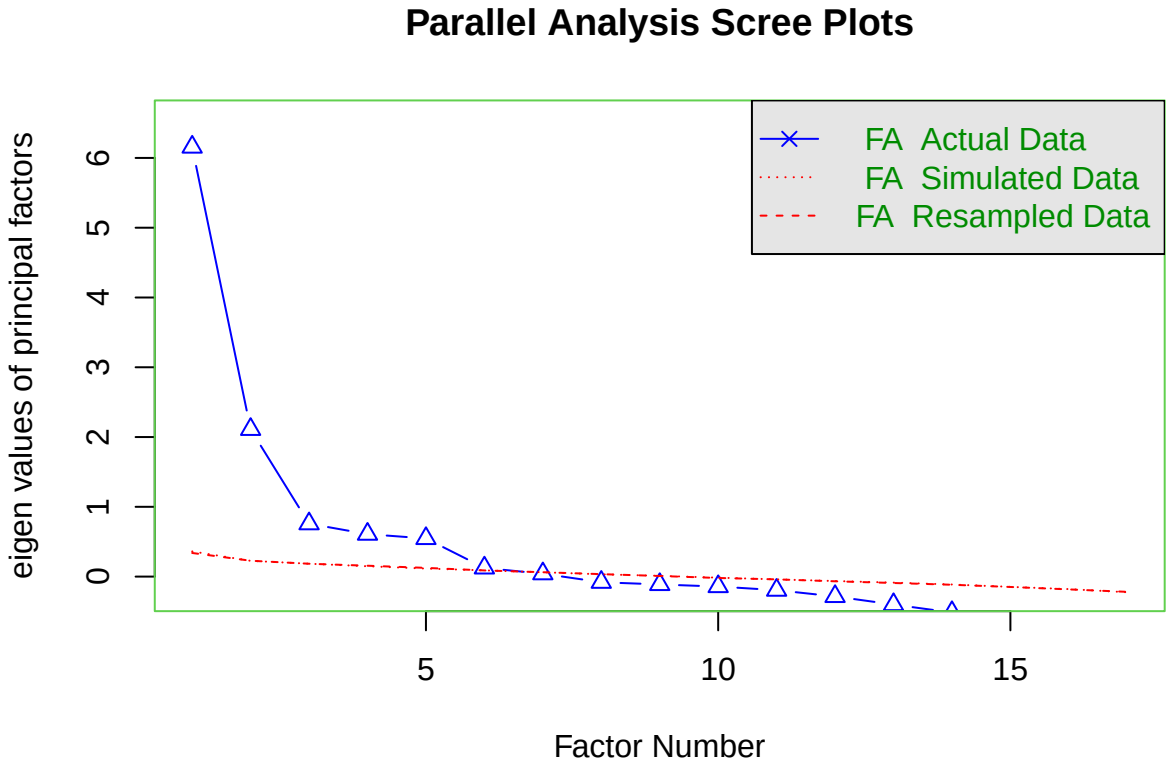
No. of factors

```
eigen <- eigen(cor(data_scaled))
par <- nFactors::parallel(
  subject = nrow(data_scaled),
  var = ncol(data_scaled),
  rep = 100,
  quantile = .95,
  model = "factors"
)
scree <- nScree(x = eigen$values, aparallel = par$eigen$qevpea)
plotnScree(scree)
```




```
fa.parallel(data_scaled, fm = "pa", fa = "fa", n.iter = 20)
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```




```
## Parallel analysis suggests that the number of factors =  6  and the number of components =  NA
```

Model

https://www.rdocumentation.org/packages/psych/versions/2.5.3/topics/fa

```
# appears to be the happiest when nfactors = 6 or 7
# throws the The estimated weights for the factor scores are probably incorrect.
# Try a different factor score estimation method. warning otherwise
fa_res <- fa(
  data_scaled,
  nfactors = 7,
  fm = "pa",
  rotate = "promax",
  oblique.scores = TRUE,
  scores = "tenBerge",
  n.iter = 100
)
```

```
## Loading required namespace: GPArotation
```

```
fa_res
```

```
## Factor Analysis with confidence intervals using method = fa(r = data_scaled, nfactors = 7, n.iter = 1
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Factor Analysis using method =  pa
## Call: fa(r = data_scaled, nfactors = 7, n.iter = 100, rotate = "promax",
##     scores = "tenBerge", fm = "pa", oblique.scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                  PA1   PA2   PA6   PA3   PA4   PA5   PA7   h2     u2 com
## compoundVERBs  0.79  0.00 -0.17 -0.01  0.04  0.24  0.51 0.78 0.2219 2.0
## passives       0.10 -0.03  0.00 -0.17  0.03  0.82  0.12 0.59 0.4143 1.2
```

```
## predorder.m   -0.72 -0.02  0.13 -0.07 -0.11 -0.05  0.02 0.55 0.4487 1.1
## obj            0.24 -0.01  0.96 -0.15 -0.02 -0.05 -0.12 0.71 0.2887 1.2
## subj           0.72  0.09  0.02  0.09 -0.15  0.18 -0.01 0.53 0.4684 1.3
## VERBfrac.m     0.75 -0.01 -0.03 -0.01 -0.03 -0.31  0.24 0.91 0.0932 1.6
## NOUNcount.m   -1.06  0.06 -0.16  0.11 -0.07 -0.13 -0.07 0.86 0.1443 1.1
## activity       0.72  0.01  0.15 -0.15 -0.02 -0.43  0.18 0.93 0.0695 2.0
## cli            0.20 -0.04 -0.14  0.95  0.08 -0.23 -0.03 0.91 0.0852 1.3
## entropy        0.12  0.74 -0.05  0.04  0.54  0.00  0.03 0.95 0.0461 1.9
## fkgl          -0.40  0.04  0.60  0.07  0.04  0.19  0.08 1.00 0.0013 2.1
## fre            0.17 -0.03 -0.53 -0.53 -0.04 -0.07 -0.11 0.99 0.0127 2.4
## hpoint         0.01  0.94  0.01 -0.02 -0.01 -0.02 -0.02 0.87 0.1348 1.0
## mamr           0.68 -0.06 -0.04  0.22 -0.32  0.04  0.06 0.75 0.2506 1.7
## mattr         -0.06 -0.12 -0.01  0.08  0.83  0.04  0.02 0.72 0.2818 1.1
## hapaxes        0.09 -0.94 -0.02  0.03  0.29  0.04 -0.01 0.86 0.1397 1.2
## verbdist      -0.94  0.02 -0.27 -0.09 -0.10  0.08  0.07 0.82 0.1803 1.2
##
##                        PA1  PA2  PA6  PA3  PA4  PA5  PA7
## SS loadings           5.52 2.34 1.71 1.31 1.26 1.22 0.35
## Proportion Var        0.32 0.14 0.10 0.08 0.07 0.07 0.02
## Cumulative Var        0.32 0.46 0.56 0.64 0.71 0.79 0.81
## Proportion Explained  0.40 0.17 0.13 0.10 0.09 0.09 0.03
## Cumulative Proportion 0.40 0.57 0.70 0.79 0.89 0.97 1.00
##
##  With factor correlations of
##       PA1   PA2   PA6  PA3   PA4   PA5   PA7
## PA1  1.00  0.02 -0.38 0.04 -0.27 -0.49 -0.09
## PA2  0.02  1.00  0.30 0.17  0.17  0.19 -0.01
## PA6 -0.38  0.30  1.00 0.27  0.13  0.22  0.28
## PA3  0.04  0.17  0.27 1.00  0.10  0.29  0.33
## PA4 -0.27  0.17  0.13 0.10  1.00  0.11 -0.05
## PA5 -0.49  0.19  0.22 0.29  0.11  1.00  0.22
## PA7 -0.09 -0.01  0.28 0.33 -0.05  0.22  1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 7 factors are sufficient.
##
## df null model =  136  with the objective function =  17.23 with Chi Square =  12859.06
## df of  the model are 38  and the objective function was  0.41
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.02
##
## The harmonic n.obs is  754 with the empirical chi square  19.64  with prob <  0.99
## The total n.obs was  754  with Likelihood Chi Square =  305.8  with prob <  1.5e-43
##
## Tucker Lewis Index of factoring reliability =  0.924
## RMSEA index =  0.097  and the 90 % confidence intervals are  0.087 0.107
## BIC =  54.04
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 PA1  PA2  PA6  PA3  PA4  PA5
## Correlation of (regression) scores with factors   0.99 0.98 0.99 0.99 0.95 0.93
## Multiple R square of scores with factors          0.97 0.96 0.98 0.98 0.90 0.86
## Minimum correlation of possible factor scores     0.95 0.91 0.96 0.96 0.80 0.73
```

```
##                                                 PA7
## Correlation of (regression) scores with factors   0.85
## Multiple R square of scores with factors          0.72
## Minimum correlation of possible factor scores     0.44
##
##  Coefficients and bootstrapped confidence intervals
##                 low   PA1 upper   low   PA2 upper   low   PA6 upper   low   PA3
## compoundVERBs  0.73  0.79  0.86 -0.05  0.00  0.03 -0.20 -0.17 -0.09 -0.07 -0.01
## passives       0.04  0.10  0.15 -0.06 -0.03  0.01 -0.05  0.00  0.04 -0.23 -0.17
## predorder.m   -0.80 -0.72 -0.61 -0.07 -0.02  0.04  0.06  0.13  0.20 -0.21 -0.07
## obj            0.15  0.24  0.29 -0.03 -0.01  0.03  0.87  0.96  1.02 -0.20 -0.15
## subj           0.58  0.72  0.81  0.03  0.09  0.17 -0.06  0.02  0.09  0.03  0.09
## VERBfrac.m     0.67  0.75  0.82 -0.05 -0.01  0.02 -0.06 -0.03  0.02 -0.06 -0.01
## NOUNcount.m   -1.14 -1.06 -0.90  0.01  0.06  0.09 -0.19 -0.16 -0.10  0.07  0.11
## activity       0.64  0.72  0.79 -0.03  0.01  0.03  0.12  0.15  0.20 -0.18 -0.15
## cli            0.15  0.20  0.24 -0.05 -0.04 -0.01 -0.18 -0.14 -0.10  0.91  0.95
## entropy        0.08  0.12  0.15  0.70  0.74  0.78 -0.07 -0.05 -0.02  0.02  0.04
## fkgl          -0.46 -0.40 -0.33  0.01  0.04  0.06  0.54  0.60  0.67  0.05  0.07
## fre            0.12  0.17  0.20 -0.04 -0.03 -0.01 -0.60 -0.53 -0.48 -0.60 -0.53
## hpoint        -0.03  0.01  0.05  0.91  0.94  0.96 -0.02  0.01  0.04 -0.05 -0.02
## mamr           0.57  0.68  0.77 -0.11 -0.06  0.00 -0.11 -0.04  0.01  0.16  0.22
## mattr         -0.10 -0.06 -0.01 -0.16 -0.12 -0.08 -0.05 -0.01  0.03  0.04  0.08
## hapaxes        0.04  0.09  0.12 -0.97 -0.94 -0.90 -0.05 -0.02  0.02  0.01  0.03
## verbdist      -1.00 -0.94 -0.82 -0.02  0.02  0.06 -0.35 -0.27 -0.18 -0.13 -0.09
##                upper   low   PA4 upper   low   PA5 upper   low   PA7 upper
## compoundVERBs  0.05  0.01  0.04  0.07  0.18  0.24  0.37  0.39  0.51  0.75
## passives      -0.12  0.00  0.03  0.06  0.72  0.82  0.98 -0.05  0.12  0.44
## predorder.m    0.02 -0.16 -0.11 -0.04 -0.17 -0.05  0.10 -0.18  0.02  0.22
## obj           -0.12 -0.05 -0.02  0.00 -0.14 -0.05 -0.01 -0.23 -0.12 -0.02
## subj           0.15 -0.22 -0.15 -0.08  0.07  0.18  0.29 -0.24 -0.01  0.31
## VERBfrac.m     0.02 -0.06 -0.03  0.00 -0.37 -0.31 -0.24  0.17  0.24  0.39
## NOUNcount.m    0.16 -0.11 -0.07 -0.03 -0.18 -0.13 -0.06 -0.36 -0.07  0.08
## activity      -0.12 -0.05 -0.02  0.01 -0.50 -0.43 -0.36  0.12  0.18  0.24
## cli            1.01  0.06  0.08  0.11 -0.30 -0.23 -0.17 -0.12 -0.03  0.08
## entropy        0.07  0.49  0.54  0.57 -0.04  0.00  0.04 -0.01  0.03  0.10
## fkgl           0.11  0.02  0.04  0.05  0.14  0.19  0.28 -0.02  0.08  0.17
## fre           -0.47 -0.06 -0.04 -0.02 -0.13 -0.07 -0.04 -0.18 -0.11 -0.06
## hpoint         0.00 -0.04 -0.01  0.02 -0.06 -0.02  0.02 -0.06 -0.02  0.02
## mamr           0.27 -0.37 -0.32 -0.26 -0.07  0.04  0.13 -0.15  0.06  0.39
## mattr          0.11  0.78  0.83  0.88 -0.01  0.04  0.10 -0.09  0.02  0.16
## hapaxes        0.06  0.25  0.29  0.32 -0.02  0.04  0.08 -0.06 -0.01  0.08
## verbdist      -0.05 -0.13 -0.10 -0.07 -0.03  0.08  0.24 -0.10  0.07  0.23
##
##  Interfactor correlations and bootstrapped confidence intervals
##           lower estimate  upper
## PA1-PA2 -0.0714    0.017  0.106
## PA1-PA6 -0.5273   -0.381 -0.171
## PA1-PA3 -0.5739    0.044  0.350
## PA1-PA4 -0.6641   -0.270  0.297
## PA1-PA5 -0.6635   -0.486  0.061
## PA1-PA7 -0.5572   -0.093  0.344
## PA2-PA6  0.2043    0.297  0.389
## PA2-PA3  0.0676    0.170  0.278
## PA2-PA4 -0.0032    0.168  0.298
```

```
## PA2-PA5 -0.0031    0.193  0.321
## PA2-PA7 -0.2905   -0.013  0.204
## PA6-PA3  0.1100    0.269  0.355
## PA6-PA4 -0.0055    0.134  0.396
## PA6-PA5 -0.0373    0.219  0.352
## PA6-PA7 -0.2027    0.285  0.594
## PA3-PA4 -0.1132    0.100  0.493
## PA3-PA5 -0.0740    0.288  0.405
## PA3-PA7 -0.2092    0.328  0.556
## PA4-PA5 -0.1376    0.113  0.318
## PA4-PA7 -0.3189   -0.053  0.573
## PA5-PA7 -0.4608    0.225  0.504
```

Loadings

```
fa_res$loadings
```

```
##
## Loadings:
##                PA1    PA2    PA6    PA3    PA4    PA5    PA7
## compoundVERBs   0.794        -0.165                  0.239  0.505
## passives        0.101               -0.169          0.817  0.123
## predorder.m    -0.719         0.129        -0.107
## obj             0.236         0.965 -0.155                -0.124
## subj            0.721                       -0.154  0.183
## VERBfrac.m      0.745                               -0.312  0.240
## NOUNcount.m    -1.060        -0.158  0.109        -0.131
## activity        0.716         0.151 -0.147        -0.432  0.176
## cli             0.203        -0.138  0.955        -0.229
## entropy         0.121  0.742                0.541
## fkgl           -0.404         0.605                0.194
## fre             0.171        -0.535 -0.527               -0.113
## hpoint                 0.936
## mamr            0.685                0.222 -0.318
## mattr                -0.122                0.826
## hapaxes              -0.942                0.294
## verbdist       -0.937        -0.272        -0.105
##
##                  PA1   PA2   PA6   PA3   PA4   PA5   PA7
## SS loadings    5.533 2.351 1.774 1.360 1.227 1.171 0.411
## Proportion Var 0.325 0.138 0.104 0.080 0.072 0.069 0.024
## Cumulative Var 0.325 0.464 0.568 0.648 0.720 0.789 0.813
```

```
for (i in 1:fa_res$factors) {
  cat("\n-----", colnames(fa_res$loadings)[i], "-----\n")

  loadings <- fa_res$loadings[, i]
  load_df <- data.frame(loading = loadings)

  load_df_filtered <- load_df %>%
    mutate(abs_l = abs(loading)) %>%
    mutate(str = case_when(
      abs_l > 0.7 ~ "***",
      abs_l <= 0.7 & abs_l > 0.5 ~ "** ",
```

```
      abs_l <= 0.5 & abs_l > 0.3 ~ "*  ",
      abs_l <= 0.3 & abs_l > 0.1 ~ ".  ",
      .default = ""
    )) %>%
    arrange(-abs_l) %>%
    filter(abs_l > 0.1)

  load_df_filtered %>%
    mutate(across(c(loading, abs_l), ~ round(.x, 3))) %>%
    print()

  cat("\n")
}
```

```
##
## ----- PA1 -----
##               loading abs_l str
## NOUNcount.m    -1.060 1.060 ***
## verbdist       -0.937 0.937 ***
## compoundVERBs   0.794 0.794 ***
## VERBfrac.m      0.745 0.745 ***
## subj            0.721 0.721 ***
## predorder.m    -0.719 0.719 ***
## activity        0.716 0.716 ***
## mamr            0.685 0.685 **
## fkgl           -0.404 0.404 *
## obj             0.236 0.236 .
## cli             0.203 0.203 .
## fre             0.171 0.171 .
## entropy         0.121 0.121 .
## passives        0.101 0.101 .
##
##
## ----- PA2 -----
##         loading abs_l str
## hapaxes  -0.942 0.942 ***
## hpoint    0.936 0.936 ***
## entropy   0.742 0.742 ***
## mattr    -0.122 0.122 .
##
##
## ----- PA6 -----
##               loading abs_l str
## obj             0.965 0.965 ***
## fkgl            0.605 0.605 **
## fre            -0.535 0.535 **
## verbdist       -0.272 0.272 .
## compoundVERBs  -0.165 0.165 .
## NOUNcount.m    -0.158 0.158 .
## activity        0.151 0.151 .
## cli            -0.138 0.138 .
## predorder.m     0.129 0.129 .
##
##
```

```
## ----- PA3 -----
##             loading abs_l str
## cli           0.955 0.955 ***
## fre          -0.527 0.527 **
## mamr          0.222 0.222 .
## passives     -0.169 0.169 .
## obj          -0.155 0.155 .
## activity     -0.147 0.147 .
## NOUNcount.m   0.109 0.109 .
##
##
## ----- PA4 -----
##             loading abs_l str
## mattr         0.826 0.826 ***
## entropy       0.541 0.541 **
## mamr         -0.318 0.318 *
## hapaxes       0.294 0.294 .
## subj         -0.154 0.154 .
## predorder.m  -0.107 0.107 .
## verbdist     -0.105 0.105 .
##
##
## ----- PA5 -----
##               loading abs_l str
## passives        0.817 0.817 ***
## activity       -0.432 0.432 *
## VERBfrac.m     -0.312 0.312 *
## compoundVERBs   0.239 0.239 .
## cli            -0.229 0.229 .
## fkgl            0.194 0.194 .
## subj            0.183 0.183 .
## NOUNcount.m    -0.131 0.131 .
##
##
## ----- PA7 -----
##               loading abs_l str
## compoundVERBs   0.505 0.505 **
## VERBfrac.m      0.240 0.240 .
## activity        0.176 0.176 .
## obj            -0.124 0.124 .
## passives        0.123 0.123 .
## fre            -0.113 0.113 .
```

hypotheses:

- **PA1:** register – narrativity, richness of expression; shorter clauses (-technical / +narrative)
  - narrativity? (1st and 2nd persons etc.)
- **PA2:** text length (-short / +long)
  - hapaxes load negatively, because I normed them over word count
- **PA6:** sentence complexity (more clauses) (-simple / +complex)
  - slightly longer nominal constructions / more objects, more years of education necessary, predicates slightly further in the clause, slightly more verbs
  - `fkgl` in strong correlation with `sentlen.m`
- **PA3:** word length (-short / +long)
  - cli highly correlates with atl, meaning the factor likely expresses mostly token lengths

  - slightly more passives, slightly more objects, slightly less verbal overall / slightly longer nom. constructions, slightly morphologically richer, many years of education necessary
  - more enumerations? but one would expect higher `activity` differences to occur if that was the case
- **PA4:** lexical richness (-poor / +rich)
- **PA5:** passivity (-active / +passive)
  - compound verbs, because that's what passives are in Czech
  - smaller activity, because passive participles count as `ADJ` in UD.
- **PA7:** compound verbs (-less / +more)

strong correlations:

- **PA1–PA6:** (-0.38) narrativity leads to simple clauses
- **PA2–PA6:** (+0.30) longer texts include more complex sentences
- **PA1–PA5:** (-0.49, topconf = +0.09) narrative texts more active

  NOTE: variables with low communalities are excluded from the analysis, yet still likely play a role in legal writing readability. this includes both those selected for the analysis and the excluded ones.

  NOTE: some high-correlating variables were excluded from the FA.

Healthiness diagnostics

```
fa_res$loadings[] %>%
  as_tibble() %>%
  mutate(feat = cnames) %>%
  select(feat, everything()) %>%
  pivot_longer(!feat) %>%
  mutate(value = abs(value)) %>%
  group_by(feat) %>%
  summarize(maxload = max(value)) %>%
  arrange(maxload)
```

```
## # A tibble: 17 x 2
##    feat          maxload
##    <chr>           <dbl>
##  1 fre             0.535
##  2 fkgl            0.605
##  3 mamr            0.685
##  4 activity        0.716
##  5 predorder.m     0.719
##  6 subj            0.721
##  7 entropy         0.742
##  8 VERBfrac.m      0.745
##  9 compoundVERBs   0.794
## 10 passives        0.817
## 11 mattr           0.826
## 12 hpoint          0.936
## 13 verbdist        0.937
## 14 hapaxes         0.942
## 15 cli             0.955
## 16 obj             0.965
## 17 NOUNcount.m     1.06
```

```
fa_res$communality %>% sort()
```

```
##          subj   predorder.m      passives           obj         mattr
##     0.5315955     0.5512988     0.5856923     0.7113441     0.7181813
##          mamr compoundVERBs      verbdist   NOUNcount.m       hapaxes
##     0.7493974     0.7781089     0.8196810     0.8557424     0.8602815
##        hpoint    VERBfrac.m           cli      activity       entropy
##     0.8652002     0.9067914     0.9147991     0.9305464     0.9539003
##           fre          fkgl
##     0.9872669     0.9987222
```

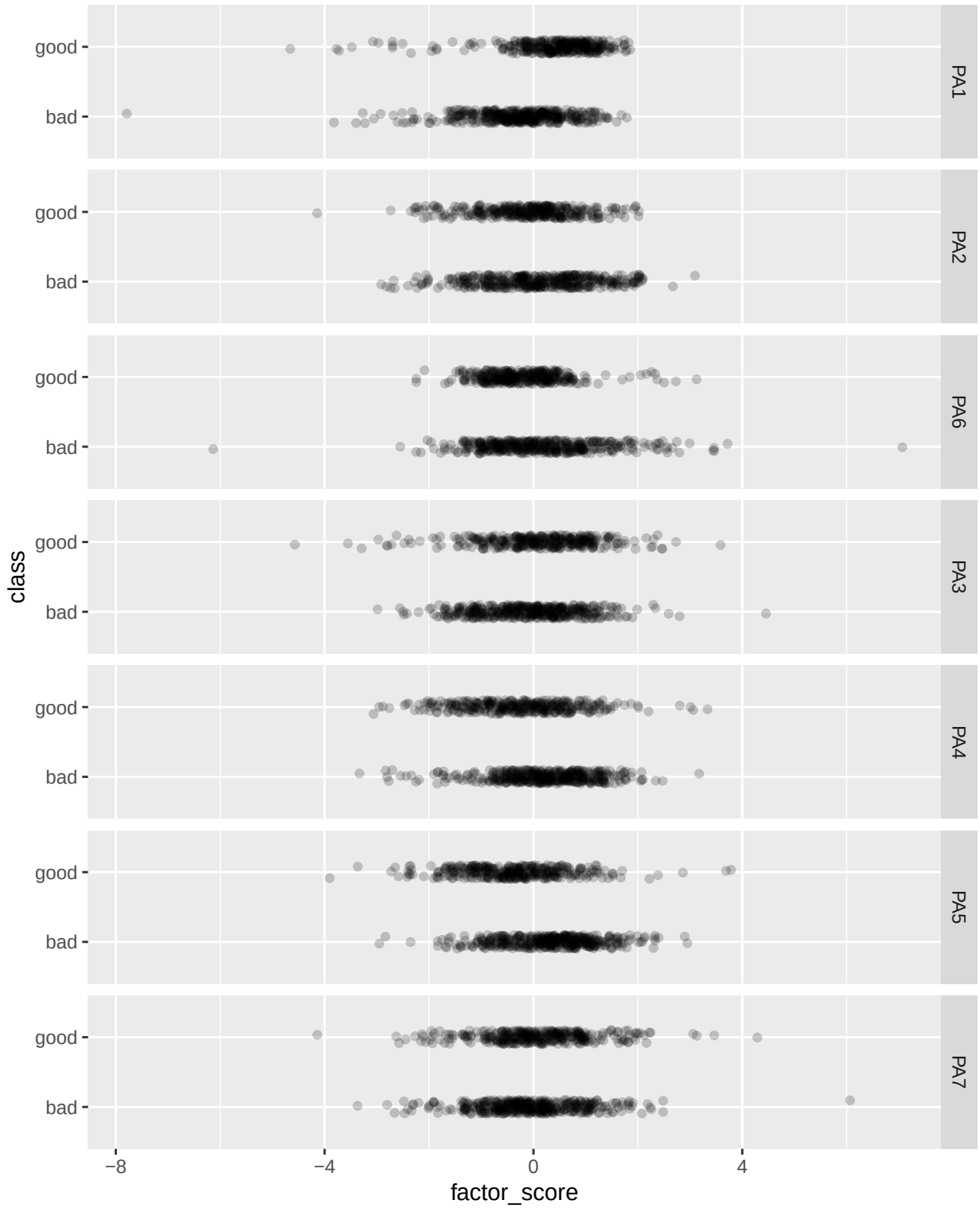
Uniquenesses

```
fa_res$uniquenesses %>% round(3)
```

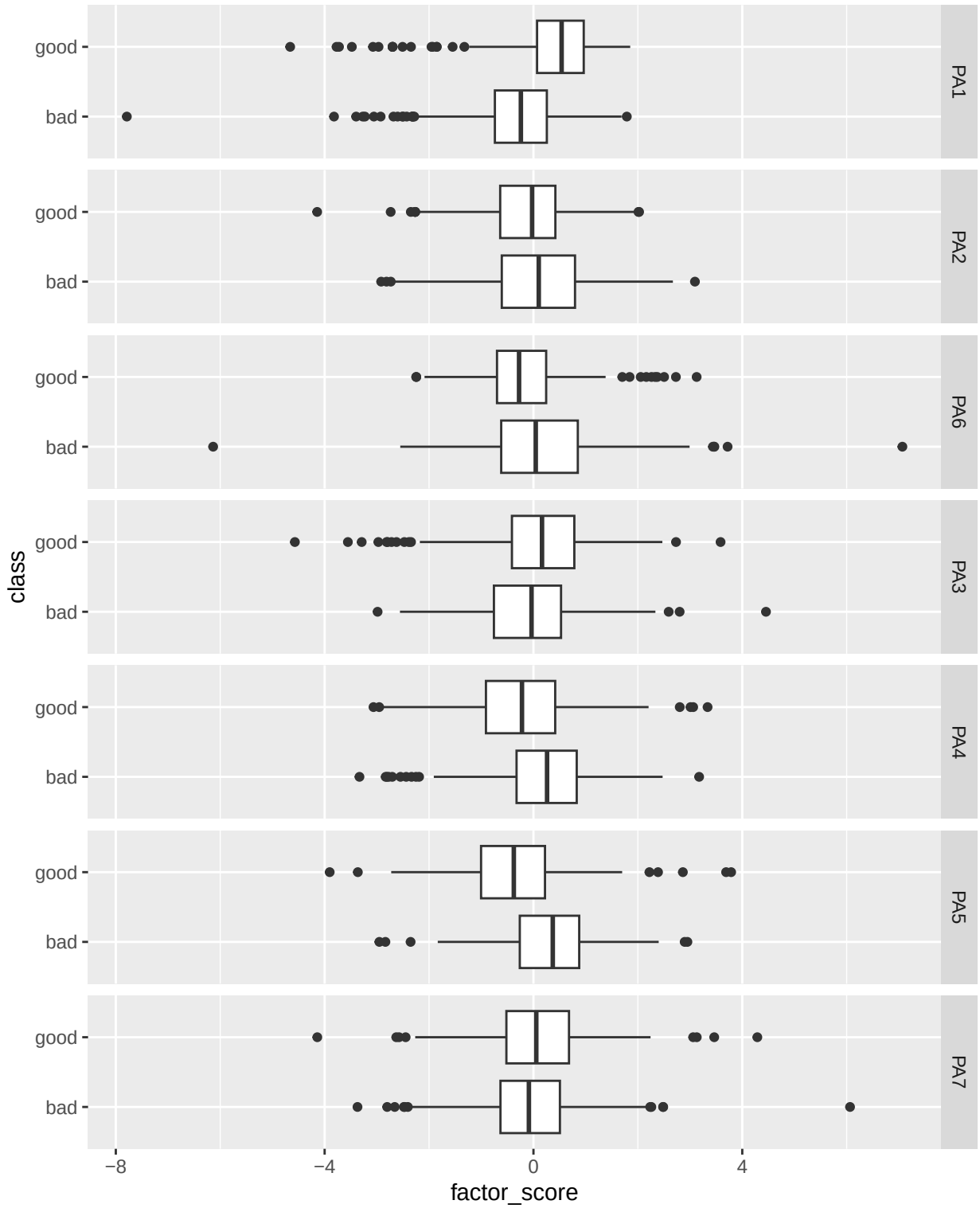
```
## compoundVERBs      passives   predorder.m           obj          subj
##         0.222         0.414         0.449         0.289         0.468
##    VERBfrac.m   NOUNcount.m      activity           cli       entropy
##         0.093         0.144         0.069         0.085         0.046
##          fkgl           fre        hpoint          mamr         mattr
##         0.001         0.013         0.135         0.251         0.282
##       hapaxes      verbdist
##         0.140         0.180
```

Plots

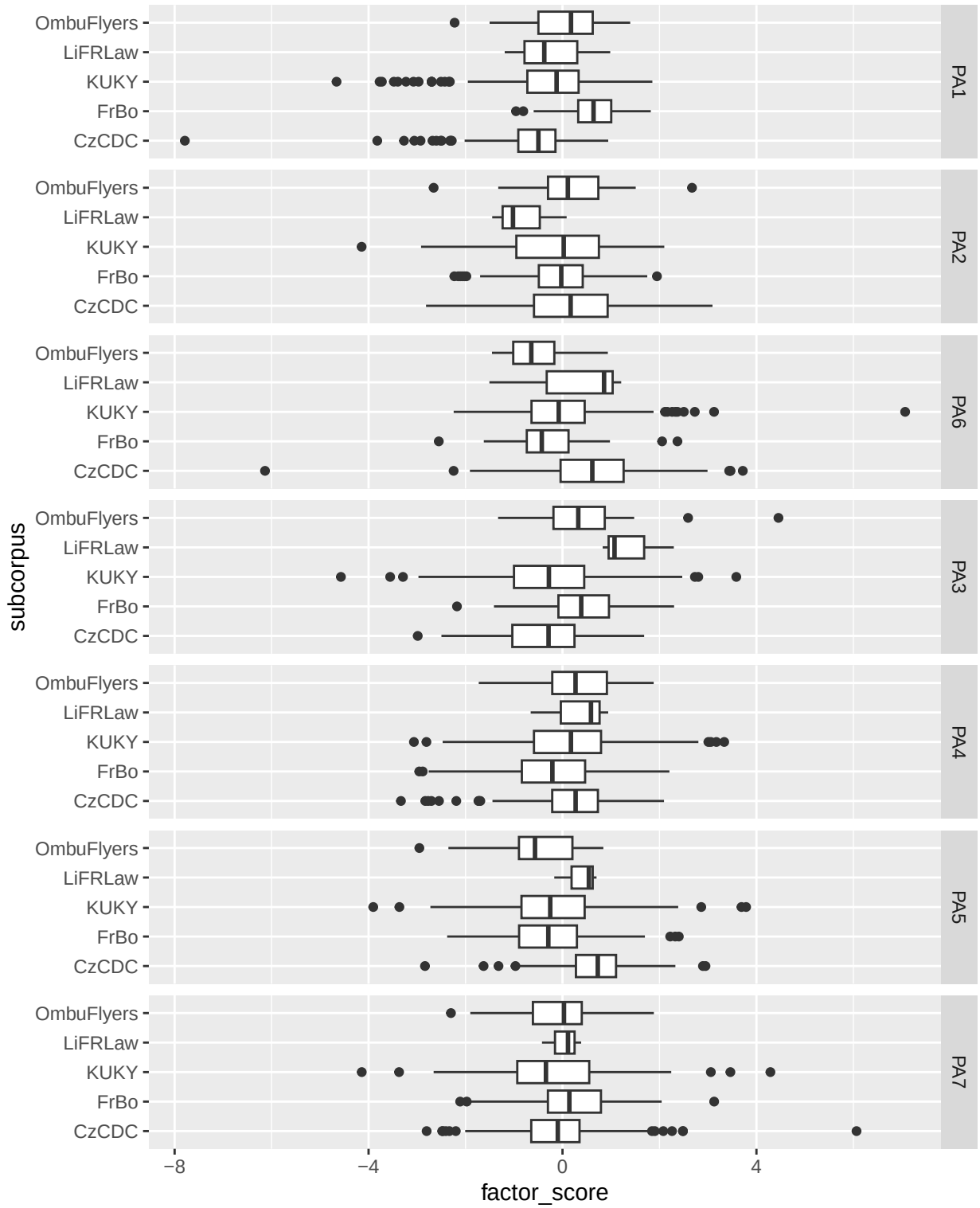
```
data_factors <- bind_cols(data_clean, fa_res$scores %>% as.data.frame())
cnames <- map(
  colnames(data_factors),
  function(x) {
    name <- pull(pretty_names %>%
      filter(name_orig == x), name_pretty)
    if (length(name) == 1) {
      return(name)
    } else {
      return(x)
    }
  }
) %>% unlist()
colnames(data_factors) <- cnames

data_factors_long <- data_factors %>%
  pivot_longer(PA1:PA7, names_to = "factor", values_to = "factor_score") %>%
  mutate(across(
    factor,
    ~ factor(.x, levels = c("PA1", "PA2", "PA6", "PA3", "PA4", "PA5", "PA7"))
  ))

data_factors_long %>%
  ggplot(aes(x = factor_score, y = class)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_jitter(width = 0, height = 0.1, alpha = 0.2)
```



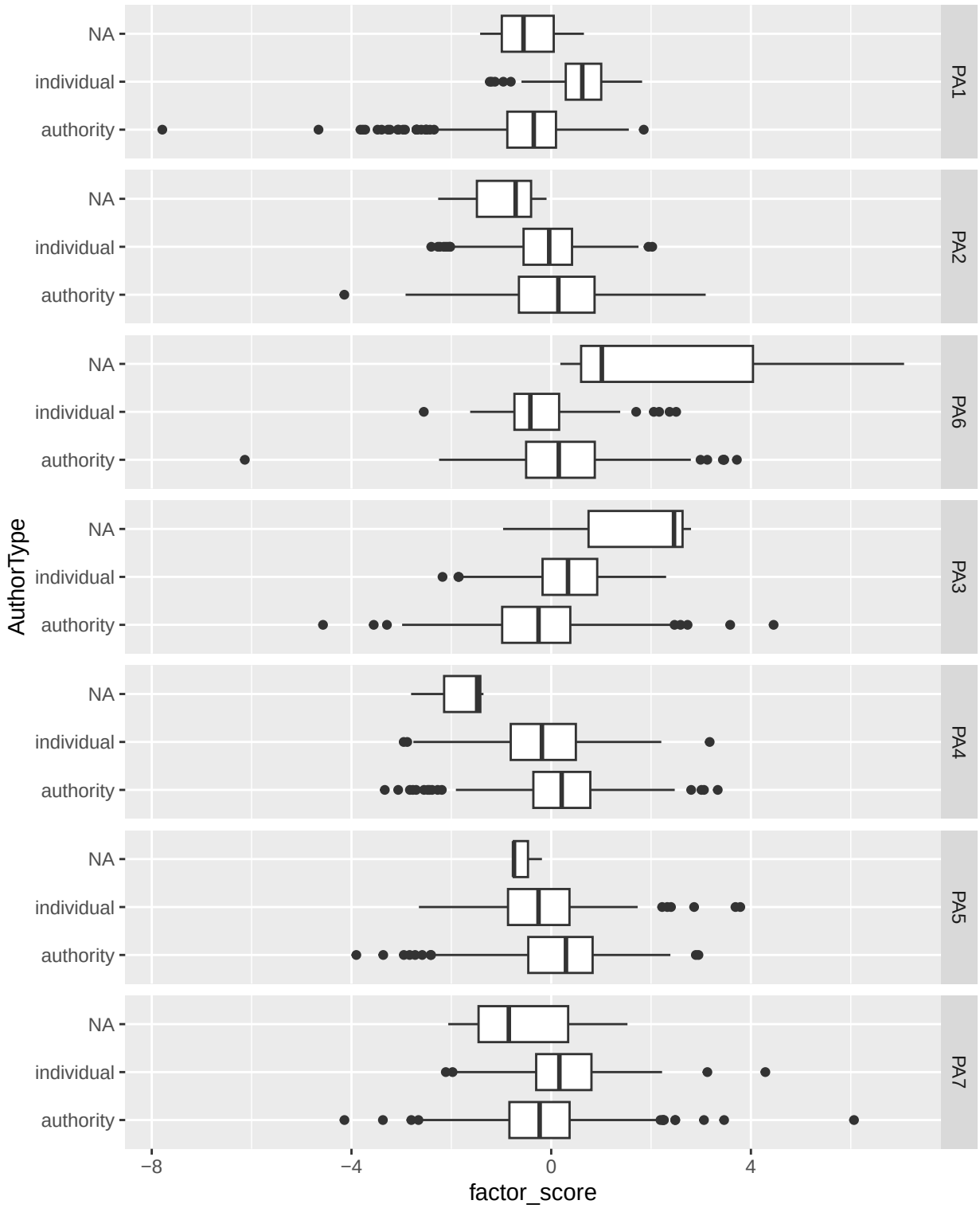

```
data_factors_long %>% ggplot(aes(x = factor_score, y = class)) +
  geom_boxplot() +
  facet_grid(factor ~ .)
```



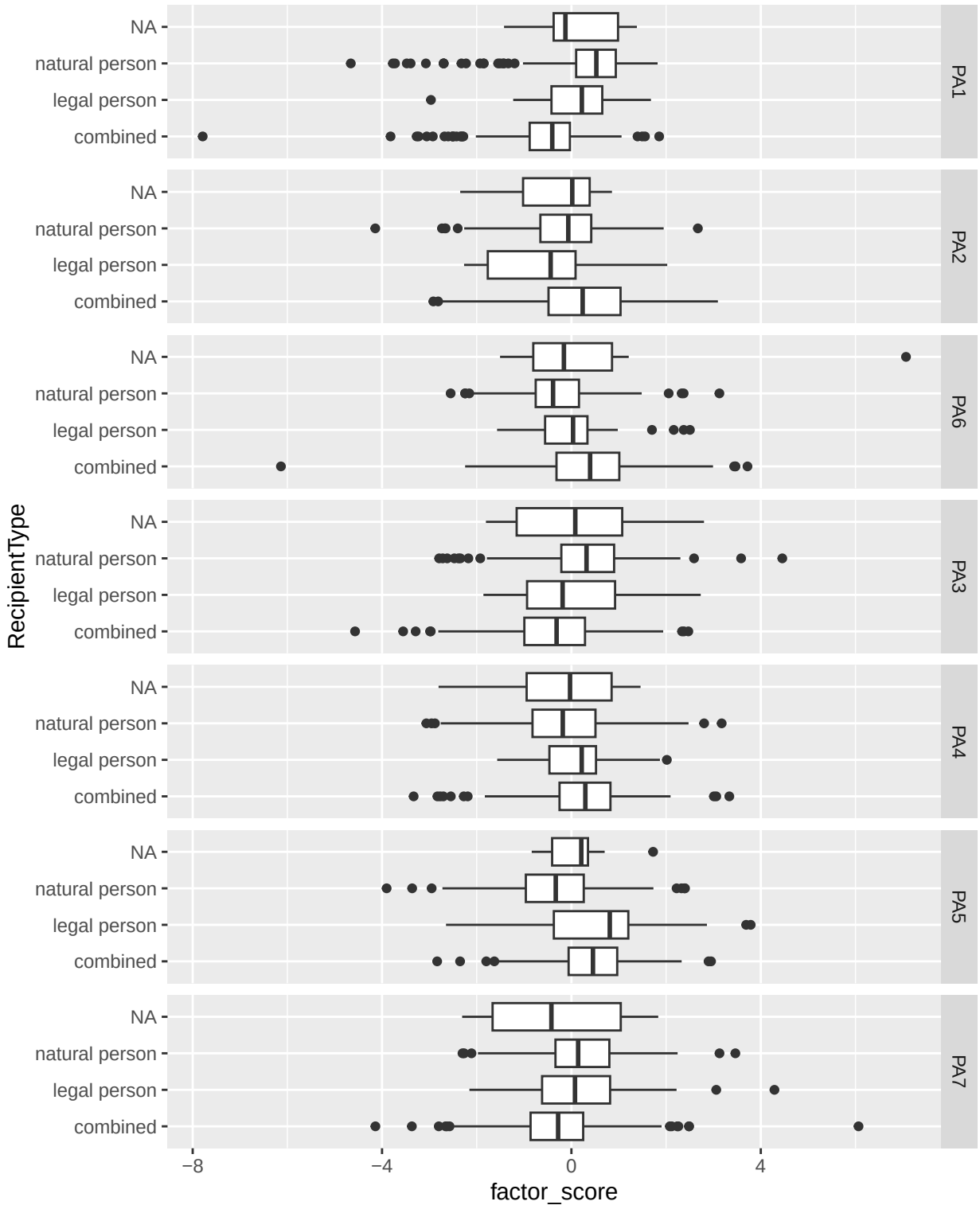

```
data_factors_long %>% ggplot(aes(x = factor_score, y = subcorpus)) +
  geom_boxplot() +
  facet_grid(factor ~ .)
```




```
data_factors_long %>%
  ggplot(aes(x = factor_score, y = AuthorType)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_boxplot()
```



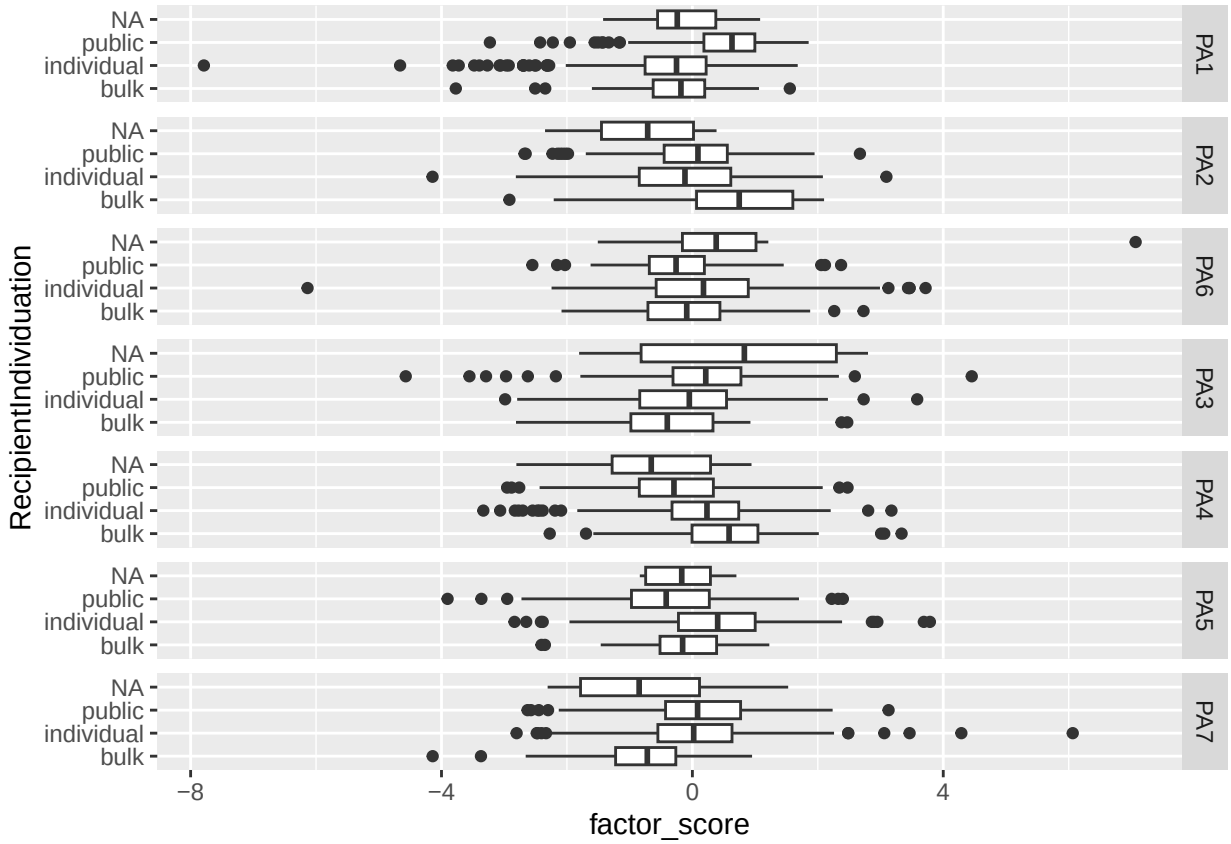

```
data_factors_long %>%
  ggplot(aes(x = factor_score, y = RecipientType)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_boxplot()
```



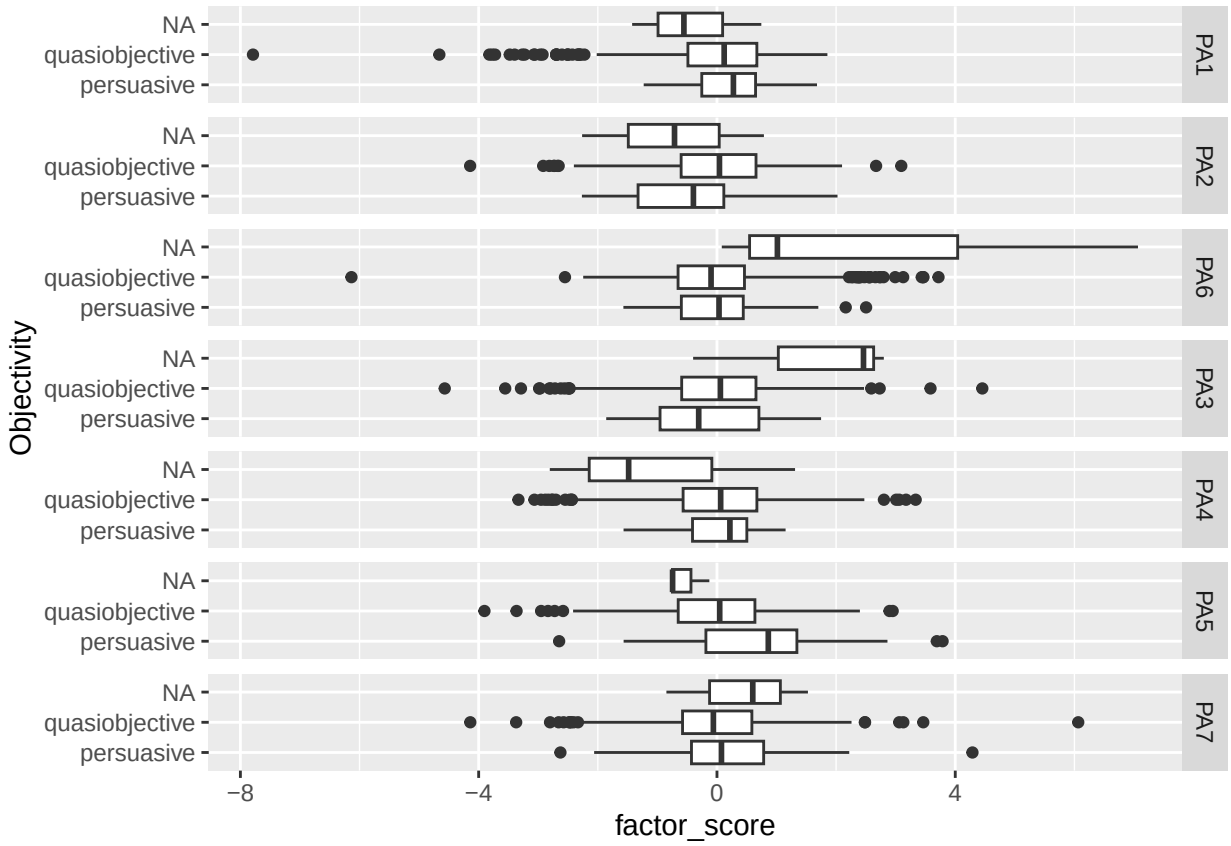
court decisions often `combined`.

```
data_factors_long %>%
  ggplot(aes(x = factor_score, y = RecipientIndividuation)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
```

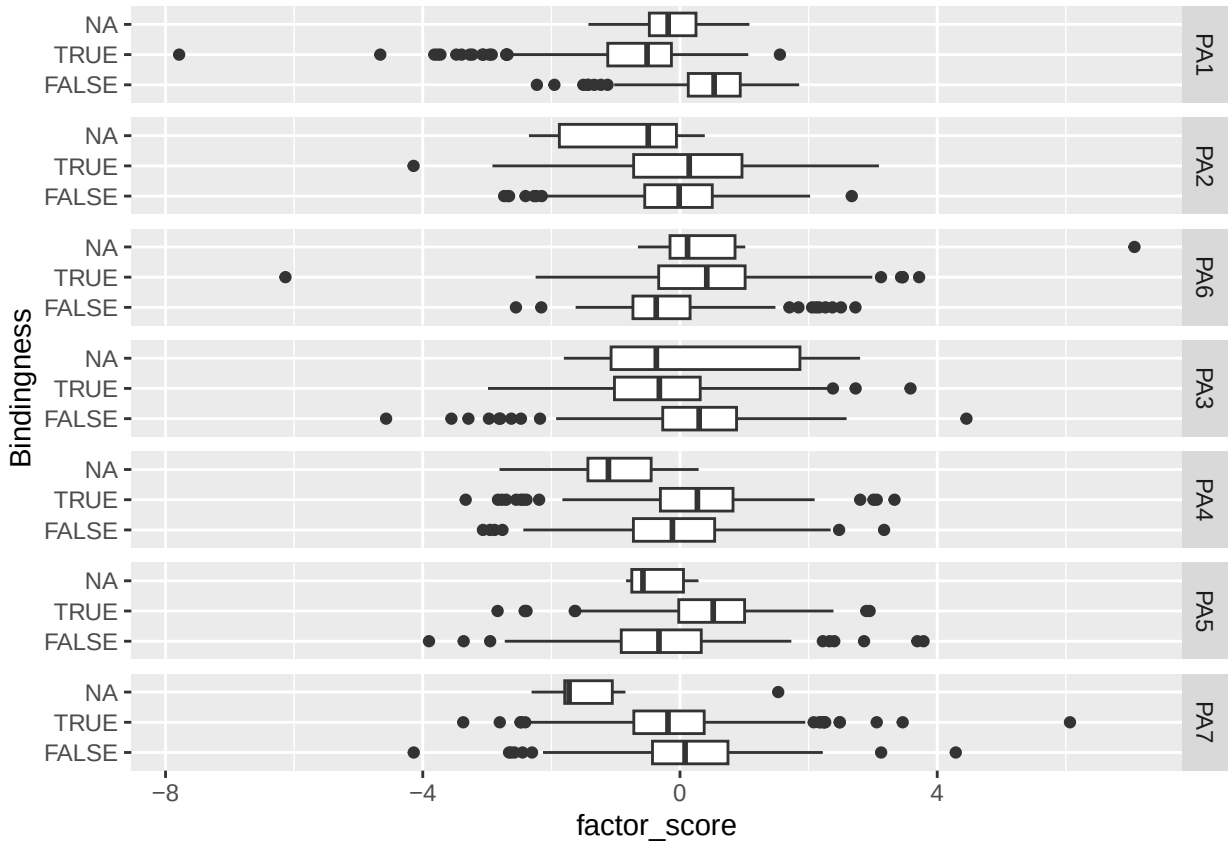
```
  geom_boxplot()
```




```
data_factors_long %>%
  ggplot(aes(x = factor_score, y = Objectivity)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_boxplot()
```




```
data_factors_long %>%
  ggplot(aes(x = factor_score, y = Bindingness)) +
  facet_grid(factor ~ .) +
  theme(legend.position = "bottom") +
  geom_boxplot()
```

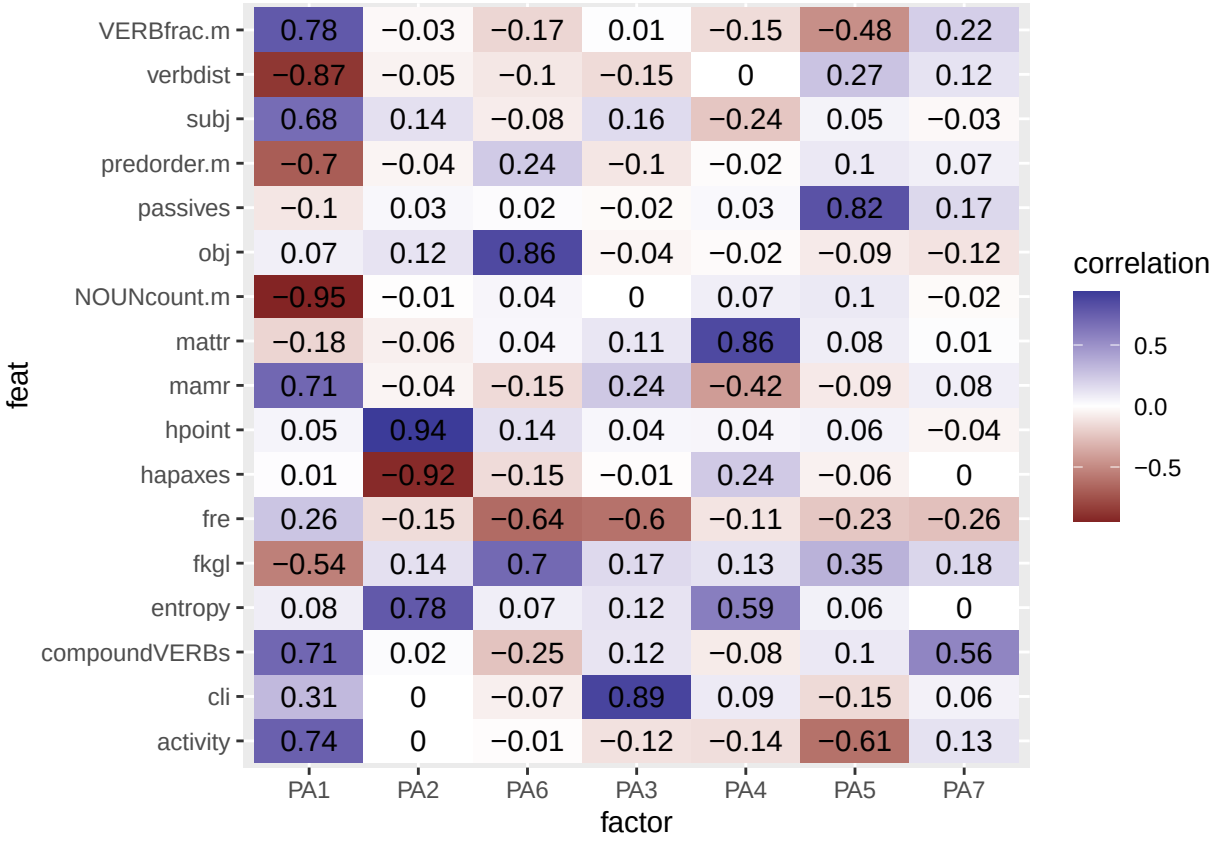


Feature-factor correlations

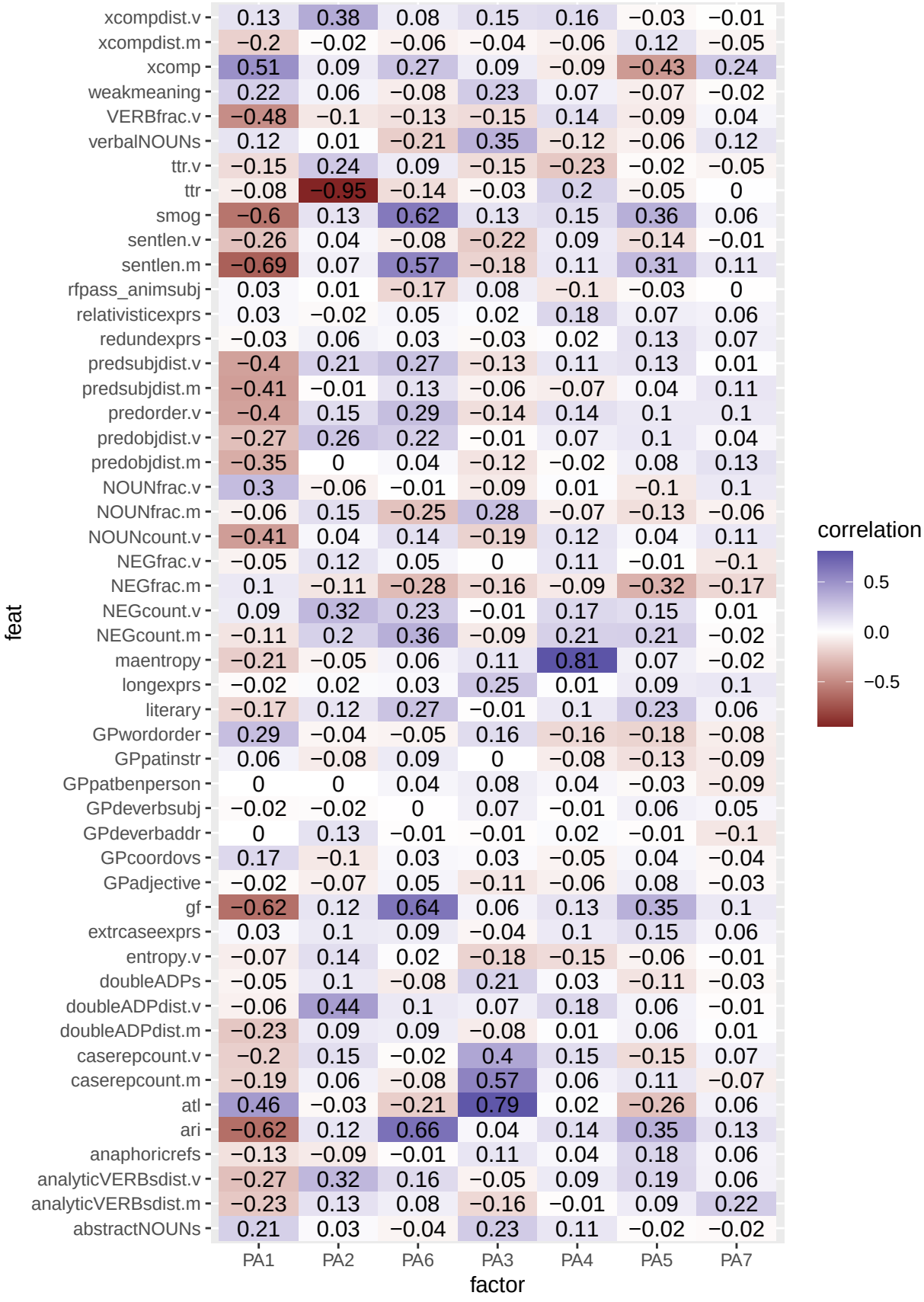
```
data_factors_longer <- data_factors_long %>%
  pivot_longer(
    abstractNOUNs:verbdist,
    names_to = "feat", values_to = "feat_value"
  )

data_factors_correlations <- data_factors_longer %>%
  group_by(feat, factor) %>%
  summarize(correlation = cor(feat_value, factor_score))
```

```
## `summarise()` has grouped output by 'feat'. You can override using the
## `.groups` argument.
```

```
data_factors_correlations %>%
  filter(feat %in% final_collist) %>%
  ggplot(aes(
    x = factor,
    y = feat,
    fill = correlation,
    label = round(correlation, 2)
  )) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient2()
```




```
data_factors_correlations %>%
  filter(!(feat %in% final_collist)) %>%
  ggplot(aes(
    x = factor,
    y = feat,
    fill = correlation,
    label = round(correlation, 2)
  )) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient2()
```


feat
xcompdist.v 0.13 0.38 0.08 0.15 0.16 −0.03 −0.01
xcompdist.m −0.2 −0.02 −0.06 −0.04 −0.06 0.12 −0.05
xcomp 0.51 0.09 0.27 0.09 −0.09 −0.43 0.24
weakmeaning 0.22 0.06 −0.08 0.23 0.07 −0.07 −0.02
VERBfrac.v −0.48 −0.1 −0.13 −0.15 0.14 −0.09 0.04
verbalNOUNs 0.12 0.01 −0.21 0.35 −0.12 −0.06 0.12
ttr.v −0.15 0.24 0.09 −0.15 −0.23 −0.02 −0.05
ttr −0.08 −0.95 −0.14 −0.03 0.2 −0.05 0
smog −0.6 0.13 0.62 0.13 0.15 0.36 0.06
sentlen.v −0.26 0.04 −0.08 −0.22 0.09 −0.14 −0.01
sentlen.m −0.69 0.07 0.57 −0.18 0.11 0.31 0.11
rfpass_animsubj 0.03 0.01 −0.17 0.08 −0.1 −0.03 0
relativisticexprs 0.03 −0.02 0.05 0.02 0.18 0.07 0.06
redundexprs −0.03 0.06 0.03 −0.03 0.02 0.13 0.07
predsubjdist.v −0.4 0.21 0.27 −0.13 0.11 0.13 0.01
predsubjdist.m −0.41 −0.01 0.13 −0.06 −0.07 0.04 0.11
predorder.v −0.4 0.15 0.29 −0.14 0.14 0.1 0.1
predobjdist.v −0.27 0.26 0.22 −0.01 0.07 0.1 0.04
predobjdist.m −0.35 0 0.04 −0.12 −0.02 0.08 0.13
NOUNfrac.v 0.3 −0.06 −0.01 −0.09 0.01 −0.1 0.1
NOUNfrac.m −0.06 0.15 −0.25 0.28 −0.07 −0.13 −0.06
NOUNcount.v −0.41 0.04 0.14 −0.19 0.12 0.04 0.11
NEGfrac.v −0.05 0.12 0.05 0 0.11 −0.01 −0.1
NEGfrac.m 0.1 −0.11 −0.28 −0.16 −0.09 −0.32 −0.17
NEGcount.v 0.09 0.32 0.23 −0.01 0.17 0.15 0.01
NEGcount.m −0.11 0.2 0.36 −0.09 0.21 0.21 −0.02
maentropy −0.21 −0.05 0.06 0.11 0.81 0.07 −0.02
longexprs −0.02 0.02 0.03 0.25 0.01 0.09 0.1
literary −0.17 0.12 0.27 −0.01 0.1 0.23 0.06
GPwordorder 0.29 −0.04 −0.05 0.16 −0.16 −0.18 −0.08
GPpatinstr 0.06 −0.08 0.09 0 −0.08 −0.13 −0.09
GPpatbenperson 0 0 0.04 0.08 0.04 −0.03 −0.09
GPdeverbsubj −0.02 −0.02 0 0.07 −0.01 0.06 0.05
GPdeverbaddr 0 0.13 −0.01 −0.01 0.02 −0.01 −0.1
GPcoordovs 0.17 −0.1 0.03 0.03 −0.05 0.04 −0.04
GPadjective −0.02 −0.07 0.05 −0.11 −0.06 0.08 −0.03
gf −0.62 0.12 0.64 0.06 0.13 0.35 0.1
extrcaseexprs 0.03 0.1 0.09 −0.04 0.1 0.15 0.06
entropy.v −0.07 0.14 0.02 −0.18 −0.15 −0.06 −0.01
doubleADPs −0.05 0.1 −0.08 0.21 0.03 −0.11 −0.03
doubleADPdist.v −0.06 0.44 0.1 0.07 0.18 0.06 −0.01
doubleADPdist.m −0.23 0.09 0.09 −0.08 0.01 0.06 0.01
caserepcount.v −0.2 0.15 −0.02 0.4 0.15 −0.15 0.07
caserepcount.m −0.19 0.06 −0.08 0.57 0.06 0.11 −0.07
atl 0.46 −0.03 −0.21 0.79 0.02 −0.26 0.06
ari −0.62 0.12 0.66 0.04 0.14 0.35 0.13
anaphoricrefs −0.13 −0.09 −0.01 0.11 0.04 0.18 0.06
analyticVERBsdist.v −0.27 0.32 0.16 −0.05 0.09 0.19 0.06
analyticVERBsdist.m −0.23 0.13 0.08 −0.16 −0.01 0.09 0.22
abstractNOUNs 0.21 0.03 −0.04 0.23 0.11 −0.02 −0.02
PA1 PA2 PA6 PA3 PA4 PA5 PA7
factor
correlation
0.5
0.0
−0.5