# EFA

```r
# library(extrafont)
# extrafont::loadfonts(quiet = TRUE)

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
```

```r
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
```

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

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

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

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v readr     2.1.5

## -- 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 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
```

```r
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: 96
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): fpath, KUK_ID, class, FileName, FolderPath, subcorpus, DocumentTit...
## dbl (85): RuleAbstractNouns, RuleAmbiguousRegards, RuleAnaphoricReferences, ...
## lgl  (2): ClarityPursuit, SyllogismBased
##
## 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_clean <- 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,
```

```r
  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
)) %>%
# 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: 0 x 71
## # i 71 variables: KUK_ID <chr>, class <fct>, FileName <chr>, subcorpus <chr>,
## #   RuleAbstractNouns <dbl>, RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>,
## #   RuleCaseRepetition.max_repetition_count.v <dbl>,
## #   RuleConfirmationExpressions <dbl>, RuleDoubleAdpos <dbl>,
## #   RuleDoubleAdpos.max_allowable_distance <dbl>,
## #   RuleDoubleAdpos.max_allowable_distance.v <dbl>, RuleGPadjective <dbl>, ...
```

```
data_clean_scaled <- data_clean %>%
  mutate(across(class, ~ .x == "good")) %>%
  mutate(across(5:length(names(data_clean)), ~ scale(.x)))
```

## 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 5:ncol(data_clean)) {
  fname <- names(data_clean)[i]

  formula_single <- reformulate(fname, "class")
  # print(formula_single)

  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 %>%
  filter(p_value <= 0.05) %>%
  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
  )
}

data_purish <- data_clean %>% select(any_of(selected_features))
```

### High correlations

```
.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: 20 x 4
##    feat1                        feat2                          cor abs_cor
##    <chr>                        <chr>                        <dbl>   <dbl>
##  1 RuleLongSentences.max_length ari                          0.944   0.944
##  2 RuleLongSentences.max_length gf                           0.922   0.922
##  3 ari                          fkgl                         0.984   0.984
##  4 ari                          gf                           0.978   0.978
##  5 ari                          smog                         0.951   0.951
##  6 ari                          RuleLongSentences.max_length 0.944   0.944
##  7 atl                          cli                          0.960   0.960
##  8 cli                          atl                          0.960   0.960
##  9 fkgl                         ari                          0.984   0.984
## 10 fkgl                         gf                           0.967   0.967
## 11 fkgl                         smog                         0.949   0.949
## 12 gf                           smog                         0.987   0.987
## 13 gf                           ari                          0.978   0.978
## 14 gf                           fkgl                         0.967   0.967
## 15 gf                           RuleLongSentences.max_length 0.922   0.922
## 16 maentropy                    mattr                        0.964   0.964
## 17 mattr                        maentropy                    0.964   0.964
## 18 smog                         gf                           0.987   0.987
## 19 smog                         ari                          0.951   0.951
## 20 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

```r
data_pureish_striphigh <- data_purish %>% select(!c(
  ari, gf, maentropy, smog, atl
  # ari, gf, maentropy, smog, atl, fkgl, RuleTooFewVerbs.min_verb_frac
  # ari, gf, maentropy, smog, atl, num_hapax
  # ari, gf, maentropy, smog, atl, num_hapax, fkgl, RuleTooFewVerbs.min_verb_frac
  # ari, gf, maentropy, smog, atl, num_hapax, fkgl, RuleTooFewVerbs.min_verb_frac, RuleTooFewVerbs.min_
  # ari, gf, maentropy, smog, atl, num_hapax, fkgl, RuleTooFewVerbs.min_verb_frac.v
  # ari, gf, maentropy, smog, atl, RuleTooFewVerbs.min_verb_frac, RuleTooFewVerbs.min_verb_frac.v
))

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

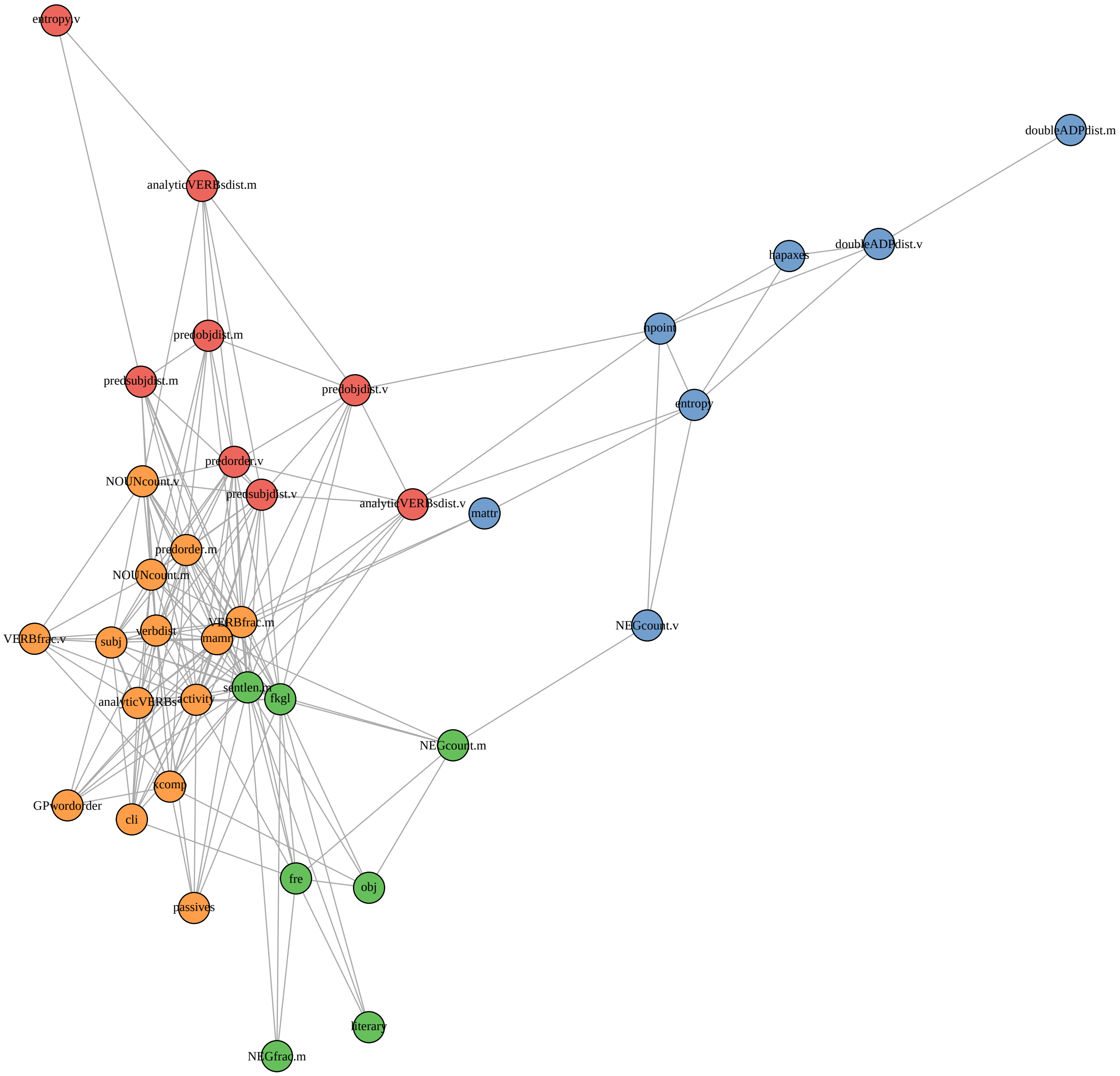
```r
# 0.35 instead of 0.3 otherwise the FA bootstrapping would freeze
.lcorrcutoff <- 0.35

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: 10 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 RuleRedundantExpressions                        0.0104
##  6 RuleRelativisticExpressions                     0.00205
##  7 RuleTooManyNegations.max_negation_frac.v        0.0365
##  8 RuleTooManyNominalConstructions.max_noun_frac.v 0.00000311
##  9 RuleVerbalNouns                                 0.0000748
## 10 RuleWeakMeaningWords                            0.0386
```

```r
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

```r
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
)
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.frame.color = "#00000000",
  # vertex.label.family = "Public Sans",
  vertex.label.color = "black",
  vertex.label.cex = 0.7
)
```



## Scaling

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

## Check for normality

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

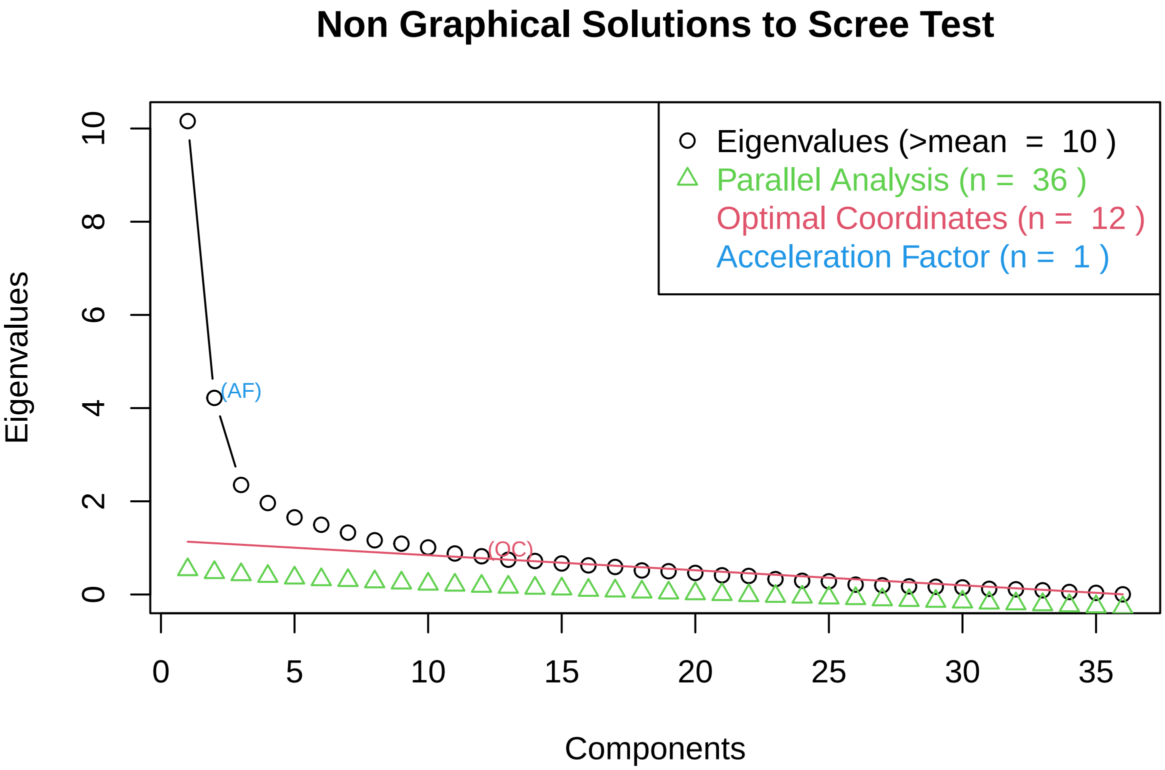
```
##           Beta-hat       kappa p-val
## Skewness  1622.36 203876.6315     0
## Kurtosis  4329.61    438.3355     0
```

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

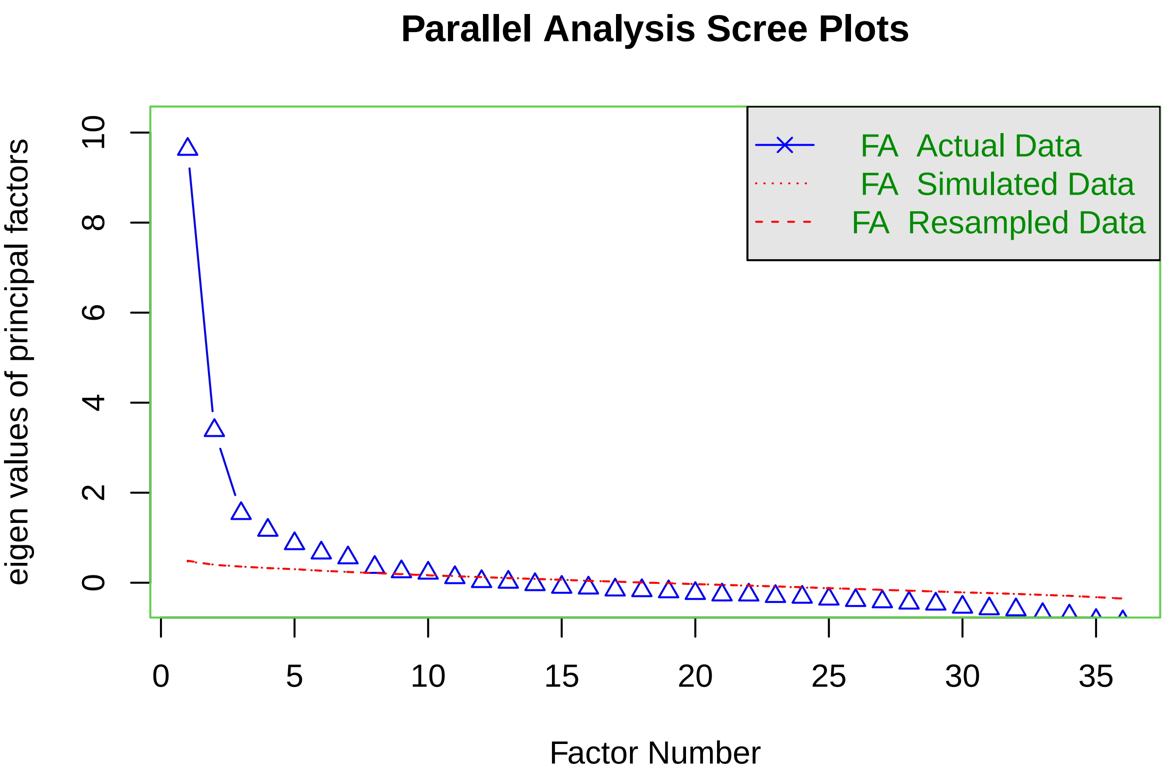
## No. of factors

```
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 =  10  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 =
##     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   PA4   PA2   PA3   PA5   PA6   PA7   h2    u2 com
## doubleADPdist.m      -0.28  0.00  0.20  0.08 -0.11  0.06 -0.02 0.14 0.864 2.5
## doubleADPdist.v      -0.11  0.06  0.56  0.01 -0.11  0.07  0.09 0.33 0.672 1.3
## GPwordorder          0.26  0.04  0.00 -0.02 -0.14  0.16 -0.11 0.18 0.823 2.8
## xcomp                0.50  0.26  0.01  0.04 -0.04  0.52  0.00 0.60 0.403 2.5
## literary             0.01  0.24 -0.02  0.08  0.24 -0.11  0.03 0.22 0.776 2.7
## sentlen.m           -0.63  0.43 -0.06  0.04  0.14  0.00 -0.09 0.93 0.072 2.0
## analyticVERBs        1.06  0.00 -0.12  0.39  0.02 -0.20  0.05 0.71 0.293 1.4
## analyticVERBsdist.m  0.21 -0.04 -0.04  0.82 -0.04 -0.04 -0.01 0.47 0.530 1.2
## analyticVERBsdist.v -0.03  0.10  0.25  0.28  0.07 -0.13 -0.03 0.33 0.672 2.9
## passives             0.10  0.22 -0.03  0.06  0.13 -0.60 -0.09 0.46 0.537 1.5
## predorder.m         -0.59  0.23 -0.15  0.13 -0.03  0.01 -0.10 0.60 0.404 1.6
## predorder.v         -0.12  0.11 -0.03  0.55  0.14  0.07  0.08 0.52 0.485 1.4
## obj                 -0.01  0.50 -0.04 -0.02  0.30  0.50 -0.13 0.66 0.337 2.8
## predobjdist.m        0.00  0.01 -0.15  0.65 -0.09 -0.05 -0.03 0.38 0.623 1.2
## predobjdist.v        0.04  0.10  0.12  0.47  0.11  0.01  0.05 0.36 0.641 1.4
## subj                 0.69  0.00  0.14 -0.09  0.02 -0.10 -0.23 0.55 0.451 1.4
## predsubjdist.m      -0.22  0.11 -0.09  0.36 -0.13 -0.01 -0.19 0.33 0.670 3.0
## predsubjdist.v      -0.18  0.08  0.06  0.40  0.17  0.02  0.00 0.45 0.551 1.9
## VERBfrac.m           0.78 -0.23 -0.06  0.20 -0.01  0.36  0.02 0.90 0.097 1.8
## VERBfrac.v          -0.55 -0.22 -0.04  0.10 -0.09  0.05  0.15 0.33 0.673 1.7
## NEGcount.m          -0.06 -0.14 -0.03 -0.09  0.97 -0.03  0.07 0.81 0.192 1.1
## NEGcount.v           0.21 -0.14  0.11 -0.03  0.81 -0.07  0.09 0.58 0.417 1.3
## NEGfrac.m           -0.10 -0.59 -0.07 -0.08  0.24  0.13 -0.06 0.33 0.672 1.6
## NOUNcount.m         -0.85  0.15  0.01 -0.02 -0.19 -0.11  0.04 0.80 0.201 1.2
## NOUNcount.v         -0.29 -0.03 -0.01  0.39  0.02  0.08  0.14 0.34 0.662 2.3
## activity             0.63 -0.27 -0.04  0.12  0.10  0.52 -0.04 0.92 0.083 2.5
## cli                  0.52  0.44  0.04 -0.10 -0.24  0.03  0.29 0.49 0.508 3.1
## entropy              0.10  0.00  0.80  0.01  0.13  0.04  0.45 0.92 0.081 1.7
## fkgl                -0.37  0.80 -0.01 -0.05  0.07  0.01  0.02 0.98 0.018 1.4
## fre                 -0.04 -1.02 -0.01  0.09  0.08 -0.04 -0.16 0.91 0.089 1.1
## hpoint               0.10 -0.02  0.90 -0.07  0.13 -0.02 -0.08 0.85 0.145 1.1
## entropy.v            0.06 -0.18  0.11  0.43 -0.08  0.03 -0.26 0.26 0.742 2.3
## mamr                 0.81  0.05 -0.09 -0.02 -0.08 -0.04 -0.25 0.75 0.254 1.3
## mattr               -0.25  0.12  0.00 -0.08  0.11  0.05  0.72 0.62 0.383 1.4
## hapaxes             -0.07 -0.03 -0.81  0.13 -0.10  0.08  0.33 0.77 0.225 1.5
## verbdist            -0.74  0.02 -0.07  0.02 -0.11 -0.34 -0.10 0.79 0.209 1.5
##
##                        PA1  PA4  PA2  PA3  PA5  PA6  PA7
## SS loadings           6.97 3.33 2.72 2.38 2.08 1.77 1.30
## Proportion Var        0.19 0.09 0.08 0.07 0.06 0.05 0.04
## Cumulative Var        0.19 0.29 0.36 0.43 0.49 0.53 0.57
## Proportion Explained  0.34 0.16 0.13 0.12 0.10 0.09 0.06
## Cumulative Proportion 0.34 0.50 0.63 0.75 0.85 0.94 1.00
##
##  With factor correlations of
##       PA1   PA4   PA2   PA3   PA5   PA6   PA7
## PA1  1.00 -0.33 -0.08 -0.61 -0.31  0.29  0.01
## PA4 -0.33  1.00  0.32  0.39  0.50 -0.15 -0.03
## PA2 -0.08  0.32  1.00  0.32  0.34  0.00  0.00
## PA3 -0.61  0.39  0.32  1.00  0.33 -0.15 -0.15
## PA5 -0.31  0.50  0.34  0.33  1.00 -0.07 -0.02
## PA6  0.29 -0.15  0.00 -0.15 -0.07  1.00 -0.21
## PA7  0.01 -0.03  0.00 -0.15 -0.02 -0.21  1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 7 factors are sufficient.
```

```
##
## df null model =  630  with the objective function =  30.2 with Chi Square =  22351.26
## df of  the model are 399  and the objective function was  6.51
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  754 with the empirical chi square  1285.27  with prob <  1e-93
## The total n.obs was  754  with Likelihood Chi Square =  4788.13  with prob <  0
##
## Tucker Lewis Index of factoring reliability =  0.679
## RMSEA index =  0.121  and the 90 % confidence intervals are  0.118 0.124
## BIC =  2144.59
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                   PA1  PA4  PA2  PA3  PA5  PA6
## Correlation of (regression) scores with factors     1 0.99 0.97 0.93 0.98 0.94
## Multiple R square of scores with factors            1 0.98 0.94 0.86 0.95 0.88
## Minimum correlation of possible factor scores       1 0.97 0.88 0.73 0.90 0.77
##                                                    PA7
## Correlation of (regression) scores with factors   0.96
## Multiple R square of scores with factors          0.93
## Minimum correlation of possible factor scores     0.86
##
##  Coefficients and bootstrapped confidence intervals
##                       low   PA1 upper   low   PA4 upper   low   PA2 upper   low
## doubleADPdist.m     -1.21 -0.28  0.52 -0.30  0.00  0.28 -0.50  0.20  1.06 -1.26
## doubleADPdist.v     -0.58 -0.11  0.31 -0.18  0.06  0.30 -1.36  0.56  2.84 -0.34
## GPwordorder         -0.64  0.26  1.37 -0.06  0.04  0.13 -0.13  0.00  0.12 -0.73
## xcomp               -0.91  0.50  2.23 -0.37  0.26  0.93 -0.08  0.01  0.10 -0.90
## literary            -0.27  0.01  0.22 -0.73  0.24  1.34 -0.13 -0.02  0.10 -0.42
## sentlen.m           -2.38 -0.63  0.90 -0.60  0.43  1.58 -0.43 -0.06  0.25 -1.31
## analyticVERBs       -1.05  1.06  3.33 -0.33  0.00  0.41 -0.24 -0.12  0.12 -0.05
## analyticVERBsdist.m -0.02  0.21  0.33 -0.51 -0.04  0.53 -0.34 -0.04  0.35 -3.42
## analyticVERBsdist.v -0.37 -0.03  0.23 -0.18  0.10  0.42 -0.82  0.25  1.53 -1.23
## passives            -0.12  0.10  0.26 -0.48  0.22  1.05 -0.18 -0.03  0.15 -0.29
## predorder.m         -2.60 -0.59  1.12 -0.06  0.23  0.51 -0.75 -0.15  0.35 -1.33
## predorder.v         -1.18 -0.12  0.73 -0.47  0.11  0.79 -0.17 -0.03  0.15 -2.01
## obj                 -0.43 -0.01  0.56 -0.84  0.50  1.98 -0.42 -0.04  0.26 -1.35
## predobjdist.m       -0.49  0.00  0.40 -0.46  0.01  0.54 -0.52 -0.15  0.18 -3.54
## predobjdist.v       -0.37  0.04  0.36 -0.56  0.10  0.87 -0.36  0.12  0.72 -2.58
## subj                -1.28  0.69  3.02 -0.30  0.00  0.28 -0.27  0.14  0.60 -0.60
## predsubjdist.m      -0.96 -0.22  0.34 -0.23  0.11  0.54 -0.43 -0.09  0.23 -3.08
## predsubjdist.v      -1.05 -0.18  0.52 -0.34  0.08  0.57 -0.41  0.06  0.62 -1.92
## VERBfrac.m          -1.05  0.78  2.88 -0.73 -0.23  0.22 -0.10 -0.06  0.01 -0.24
## VERBfrac.v          -2.64 -0.55  1.21 -0.68 -0.22  0.22 -0.15 -0.04  0.07 -0.22
## NEGcount.m          -0.70 -0.06  0.48 -0.21 -0.14  0.01 -0.20 -0.03  0.19 -2.72
## NEGcount.v          -0.08  0.21  0.48 -0.22 -0.14  0.00 -0.45  0.11  0.81 -1.93
## NEGfrac.m           -0.41 -0.10  0.22 -2.66 -0.59  1.26 -0.23 -0.07  0.06 -1.12
## NOUNcount.m         -3.33 -0.85  1.30 -0.38  0.15  0.73 -0.23  0.01  0.20 -1.20
## NOUNcount.v         -1.68 -0.29  0.84 -0.43 -0.03  0.44 -0.26 -0.01  0.29 -1.74
## activity            -1.01  0.63  2.56 -1.06 -0.27  0.43 -0.15 -0.04  0.06 -0.56
## cli                 -0.76  0.52  1.98 -1.28  0.44  2.39 -0.27  0.04  0.42 -2.39
## entropy             -0.03  0.10  0.19 -0.29  0.00  0.35 -1.94  0.80  4.02 -1.93
## fkgl                -1.24 -0.37  0.39 -1.61  0.80  3.46 -0.24 -0.01  0.19 -0.84
## fre                 -0.41 -0.04  0.28 -4.84 -1.02  2.41 -0.11 -0.01  0.06 -0.41
## hpoint              -0.58  0.10  0.96 -0.33 -0.02  0.25 -1.76  0.90  3.99 -1.23
## entropy.v           -0.18  0.06  0.24 -0.79 -0.18  0.38 -0.25  0.11  0.59 -2.69
## mamr                -1.58  0.81  3.61 -0.03  0.05  0.13 -0.42 -0.09  0.20 -0.40
## mattr               -1.97 -0.25  1.16 -0.62  0.12  0.97 -0.38  0.00  0.50 -4.63
## hapaxes             -1.25 -0.07  0.87 -0.24 -0.03  0.23 -3.27 -0.81  1.33 -2.36
## verbdist            -2.96 -0.74  1.15 -0.08  0.02  0.11 -0.39 -0.07  0.20 -1.26
##                       PA3 upper   low   PA5 upper   low   PA6 upper   low   PA7
## doubleADPdist.m      0.08  1.65 -0.55 -0.11  0.24 -0.54  0.06  0.76 -1.17 -0.02
## doubleADPdist.v      0.01  0.44 -0.52 -0.11  0.21 -0.41  0.07  0.63 -2.68  0.09
## GPwordorder         -0.02  0.82 -0.53 -0.14  0.22 -0.43  0.16  0.85 -0.97 -0.11
## xcomp                0.04  1.16 -0.10 -0.04  0.07 -1.55  0.52  3.02 -0.34  0.00
## literary             0.08  0.63 -0.33  0.24  0.89 -0.49 -0.11  0.20 -0.32  0.03
## sentlen.m            0.04  1.70 -0.59  0.14  0.98 -0.16  0.00  0.13 -0.34 -0.09
## analyticVERBs        0.39  0.64 -0.53  0.02  0.48 -0.75 -0.20  0.32 -2.48  0.05
## analyticVERBsdist.m  0.82  5.72 -0.64 -0.04  0.48 -0.31 -0.04  0.25 -4.05 -0.01
## analyticVERBsdist.v  0.28  2.05 -0.08  0.07  0.25 -0.73 -0.13  0.35 -0.76 -0.03
## passives             0.06  0.31 -0.29  0.13  0.58 -3.16 -0.60  1.51 -2.15 -0.09
## predorder.m          0.13  1.91 -0.51 -0.03  0.51 -0.18  0.01  0.20 -0.66 -0.10
## predorder.v          0.55  3.56 -0.22  0.14  0.54 -0.34  0.07  0.59 -1.12  0.08
## obj                 -0.02  1.69 -1.11  0.30  1.92 -1.06  0.50  2.41 -0.64 -0.13
## predobjdist.m        0.65  5.60 -1.04 -0.09  0.75 -0.37 -0.05  0.21 -4.10 -0.03
## predobjdist.v        0.47  3.98 -0.04  0.11  0.26 -0.15  0.01  0.20 -1.67  0.05
## subj                -0.09  0.39 -0.34  0.02  0.41 -0.56 -0.10  0.30 -2.87 -0.23
## predsubjdist.m       0.36  4.36 -0.57 -0.13  0.26 -0.25 -0.01  0.23 -3.35 -0.19
## predsubjdist.v       0.40  3.10 -0.37  0.17  0.78 -0.20  0.02  0.29 -1.30  0.00
## VERBfrac.m           0.20  0.63 -0.38 -0.01  0.31 -1.12  0.36  2.18 -1.18  0.02
## VERBfrac.v           0.10  0.40 -0.81 -0.09  0.52 -0.32  0.05  0.44 -1.88  0.15
## NEGcount.m          -0.09  2.23 -2.01  0.97  4.28 -0.21 -0.03  0.19 -0.61  0.07
## NEGcount.v          -0.03  1.59 -1.62  0.81  3.52 -0.23 -0.07  0.11 -0.73  0.09
## NEGfrac.m           -0.08  0.89 -0.41  0.24  0.97 -0.38  0.13  0.76 -1.09 -0.06
## NOUNcount.m         -0.02  1.41 -0.99 -0.19  0.53 -0.55 -0.11  0.20 -1.73  0.04
## NOUNcount.v          0.39  2.79 -0.48  0.02  0.48 -0.51  0.08  0.79 -0.57  0.14
## activity             0.12  0.92 -0.07  0.10  0.32 -1.42  0.52  2.91 -1.57 -0.04
## cli                 -0.10  1.75 -1.55 -0.24  0.90 -0.20  0.03  0.29 -3.32  0.29
## entropy              0.01  1.59 -0.28  0.13  0.65 -0.14  0.04  0.21 -3.91  0.45
## fkgl                -0.05  0.90 -0.39  0.07  0.63 -0.10  0.01  0.09 -1.75  0.02
## fre                  0.09  0.65 -0.15  0.08  0.32 -0.31 -0.04  0.21 -4.26 -0.16
## hpoint              -0.07  1.39 -0.55  0.13  0.96 -0.35 -0.02  0.26 -0.52 -0.08
## entropy.v            0.43  4.11 -0.39 -0.08  0.17 -0.14  0.03  0.22 -3.95 -0.26
## mamr                -0.02  0.42 -0.33 -0.08  0.12 -0.25 -0.04  0.21 -3.93 -0.25
## mattr               -0.08  3.67 -0.03  0.11  0.29 -0.18  0.05  0.24 -5.47  0.72
## hapaxes              0.13  2.12 -0.89 -0.10  0.56 -0.33  0.08  0.56 -1.57  0.33
## verbdist             0.02  1.57 -0.43 -0.11  0.17 -1.98 -0.34  0.99 -0.95 -0.10
##                     upper
## doubleADPdist.m      1.27
## doubleADPdist.v      3.33
## GPwordorder          0.59
## xcomp                0.34
## literary             0.34
## sentlen.m            0.18
## analyticVERBs        2.11
## analyticVERBsdist.m  3.34
## analyticVERBsdist.v  0.53
## passives             1.64
## predorder.m          0.38
## predorder.v          1.07
## obj                  0.33
## predobjdist.m        3.38
## predobjdist.v        1.50
## subj                 1.95
## predsubjdist.m       2.37
## predsubjdist.v       1.11
## VERBfrac.m           0.97
## VERBfrac.v           2.50
## NEGcount.m           0.92
## NEGcount.v           0.95
## NEGfrac.m            0.90
## NOUNcount.m          2.16
## NOUNcount.v          0.93
## activity             1.21
## cli                  4.52
## entropy              5.65
## fkgl                 2.10
## fre                  3.35
## hpoint               0.41
## entropy.v            2.68
## mamr                 2.84
## mattr                8.07
## hapaxes              2.50
## verbdist             0.63
##
##  Interfactor correlations and bootstrapped confidence intervals
##          lower estimate upper
## PA1-PA4 -0.803  -0.3318  0.57
## PA1-PA2 -0.562  -0.0773  0.35
## PA1-PA3 -0.979  -0.6090  0.58
## PA1-PA5 -0.814  -0.3127  0.51
## PA1-PA6 -0.435   0.2863  0.64
## PA1-PA7 -0.366   0.0137  0.34
## PA4-PA2  0.033   0.3215  0.58
## PA4-PA3 -0.254   0.3943  0.78
## PA4-PA5 -0.094   0.4992  0.77
## PA4-PA6 -0.419  -0.1513  0.55
## PA4-PA7 -0.316  -0.0257  0.31
## PA2-PA3 -0.097   0.3194  0.57
## PA2-PA5 -0.039   0.3399  0.57
## PA2-PA6 -0.283  -0.0013  0.44
## PA2-PA7 -0.250  -0.0046  0.28
## PA3-PA5 -0.251   0.3335  0.66
## PA3-PA6 -0.429  -0.1536  0.38
## PA3-PA7 -0.351  -0.1472  0.31
## PA5-PA6 -0.318  -0.0709  0.35
## PA5-PA7 -0.312  -0.0210  0.29
## PA6-PA7 -0.399  -0.2059  0.27
```

### Loadings

```
fa_res$loadings
```

```
##
## Loadings:
##                     PA1    PA4    PA2    PA3    PA5    PA6    PA7
## doubleADPdist.m     -0.278         0.201        -0.105
## doubleADPdist.v     -0.114         0.559        -0.114
## GPwordorder          0.264                      -0.135  0.157 -0.115
## xcomp                0.503  0.264                       0.518
## literary                    0.241                0.238 -0.113
## sentlen.m           -0.628  0.431                0.138
## analyticVERBs        1.059        -0.116  0.393        -0.203
## analyticVERBsdist.m  0.214                0.824
## analyticVERBsdist.v                0.253  0.281        -0.135
## passives             0.100  0.222                0.135 -0.598
## predorder.m         -0.590  0.230 -0.152  0.131               -0.103
## predorder.v         -0.120  0.112         0.548  0.136
## obj                         0.503                0.295  0.500 -0.128
## predobjdist.m                     -0.146  0.649
## predobjdist.v               0.102  0.124  0.468  0.106
## subj                 0.689         0.136                      -0.228
## predsubjdist.m      -0.223  0.112         0.355 -0.126        -0.186
## predsubjdist.v      -0.181                0.398  0.165
## VERBfrac.m           0.778 -0.231         0.201         0.360
## VERBfrac.v          -0.546 -0.221                              0.155
## NEGcount.m                 -0.137                0.969
## NEGcount.v           0.208 -0.136  0.114         0.807
## NEGfrac.m                  -0.585                0.237  0.133
## NOUNcount.m         -0.853  0.155               -0.190 -0.110
## NOUNcount.v         -0.286                0.386                0.141
## activity             0.625 -0.266         0.122         0.524
## cli                  0.519  0.444               -0.242         0.288
## entropy                            0.803         0.131         0.450
## fkgl                -0.369  0.796
## fre                        -1.025                             -0.159
## hpoint               0.101         0.897         0.134
## entropy.v                  -0.176  0.109  0.431               -0.258
## mamr                 0.808                                    -0.250
## mattr               -0.253  0.124                0.109         0.722
## hapaxes                           -0.812  0.127                0.331
## verbdist            -0.742                      -0.108 -0.336
##
##                  PA1   PA4   PA2   PA3   PA5   PA6   PA7
## SS loadings    6.770 3.235 2.691 2.646 2.169 1.577 1.316
## Proportion Var 0.188 0.090 0.075 0.073 0.060 0.044 0.037
## Cumulative Var 0.188 0.278 0.353 0.426 0.486 0.530 0.567
```

```
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)) %>%
    arrange(-abs_l) %>%
    filter(abs_l > 0.3)

  load_df_filtered %>%
    round(3) %>%
    print()

  cat("\n")
}
```

```
##
## ----- PA1 -----
##               loading abs_l
## analyticVERBs   1.059 1.059
## NOUNcount.m    -0.853 0.853
## mamr            0.808 0.808
## VERBfrac.m      0.778 0.778
## verbdist       -0.742 0.742
## subj            0.689 0.689
## sentlen.m      -0.628 0.628
## activity        0.625 0.625
## predorder.m    -0.590 0.590
## VERBfrac.v     -0.546 0.546
## cli             0.519 0.519
## xcomp           0.503 0.503
## fkgl           -0.369 0.369
##
##
## ----- PA4 -----
##           loading abs_l
## fre        -1.025 1.025
## fkgl        0.796 0.796
## NEGfrac.m  -0.585 0.585
## obj         0.503 0.503
## cli         0.444 0.444
## sentlen.m   0.431 0.431
##
##
## ----- PA2 -----
##                 loading abs_l
## hpoint            0.897 0.897
## hapaxes          -0.812 0.812
## entropy           0.803 0.803
## doubleADPdist.v   0.559 0.559
##
##
## ----- PA3 -----
##                     loading abs_l
## analyticVERBsdist.m   0.824 0.824
## predobjdist.m         0.649 0.649
## predorder.v           0.548 0.548
## predobjdist.v         0.468 0.468
## entropy.v             0.431 0.431
## predsubjdist.v        0.398 0.398
## analyticVERBs         0.393 0.393
## NOUNcount.v           0.386 0.386
## predsubjdist.m        0.355 0.355
##
##
## ----- PA5 -----
##            loading abs_l
## NEGcount.m   0.969 0.969
## NEGcount.v   0.807 0.807
##
##
## ----- PA6 -----
##            loading abs_l
## passives    -0.598 0.598
## activity     0.524 0.524
## xcomp        0.518 0.518
## obj          0.500 0.500
## VERBfrac.m   0.360 0.360
## verbdist    -0.336 0.336
##
##
## ----- PA7 -----
##         loading abs_l
## mattr     0.722 0.722
## entropy   0.450 0.450
## hapaxes   0.331 0.331
```

hypotheses:

- **PA1:** written, formal register (complex) vs. more spoken-like register
  - long, severely complex, nominalized sentences / shorter, more verbal sentences
- **PA4:** structure size? elaboratedness of expression? advancement (in years of age)?
  - short words, short sentences, more negations / long words, long sentences, more objects
  - cli: word complexity - sentence easiness
  - the negations might be because of the varying sentence length
- **PA2:** text length & enumerations
- **PA3:** intra-text (syntactic, possibly content-related) variation
  - note that the loadings of `VERBfrac.v` and `NEGcount.v` are negligible
  - however, the loading of `entropy.v` is significant
- **PA5:** negation
- **PA6:** passive / active
- **PA7:** unique words

  **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.

Strong correlations:

- **PA1–PA3:** possible register switching
- **PA4–PA5:** expression sophisticatedness

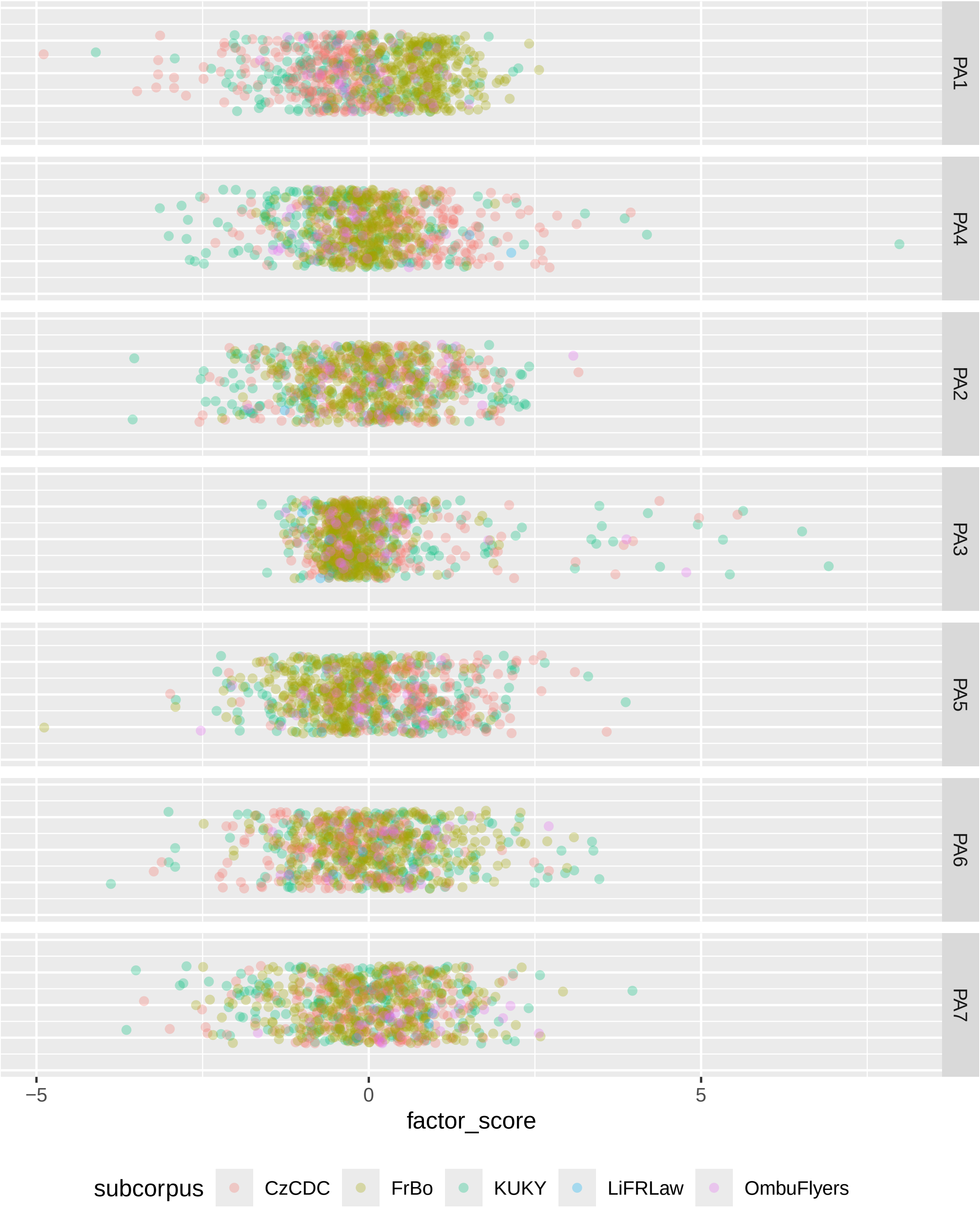
### Uniquenesses

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

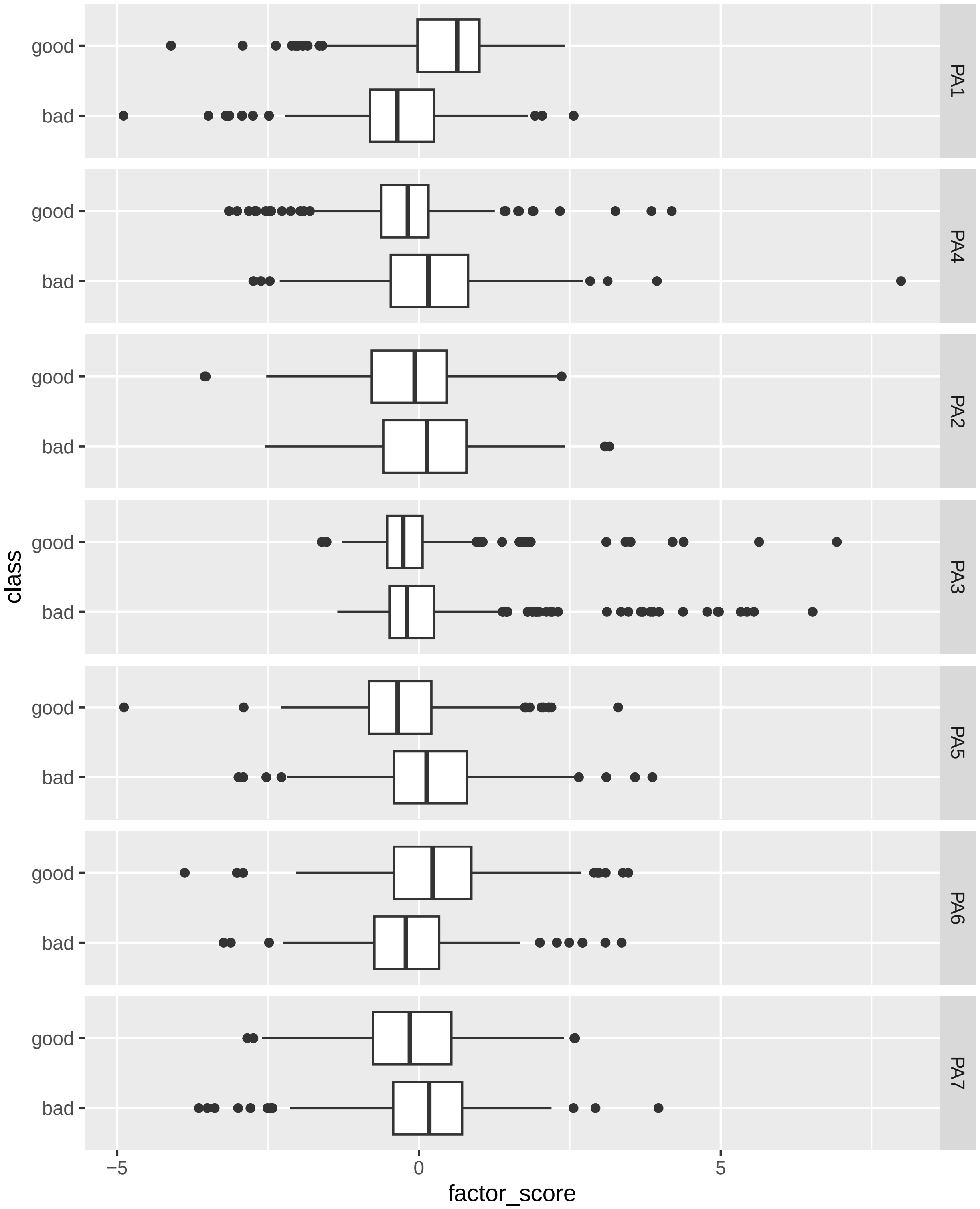
```
##     doubleADPdist.m     doubleADPdist.v         GPwordorder               xcomp
##               0.864               0.672               0.823               0.403
##            literary           sentlen.m       analyticVERBs analyticVERBsdist.m
##               0.776               0.072               0.293               0.530
## analyticVERBsdist.v            passives         predorder.m         predorder.v
##               0.672               0.537               0.404               0.485
##                 obj       predobjdist.m       predobjdist.v                subj
##               0.337               0.623               0.641               0.451
##      predsubjdist.m      predsubjdist.v          VERBfrac.m          VERBfrac.v
##               0.670               0.551               0.097               0.673
##          NEGcount.m          NEGcount.v           NEGfrac.m         NOUNcount.m
##               0.192               0.417               0.672               0.201
##         NOUNcount.v            activity                 cli             entropy
##               0.662               0.083               0.508               0.081
##                fkgl                 fre              hpoint           entropy.v
##               0.018               0.089               0.145               0.742
##                mamr               mattr             hapaxes            verbdist
##               0.254               0.383               0.225               0.209
```

## Plots

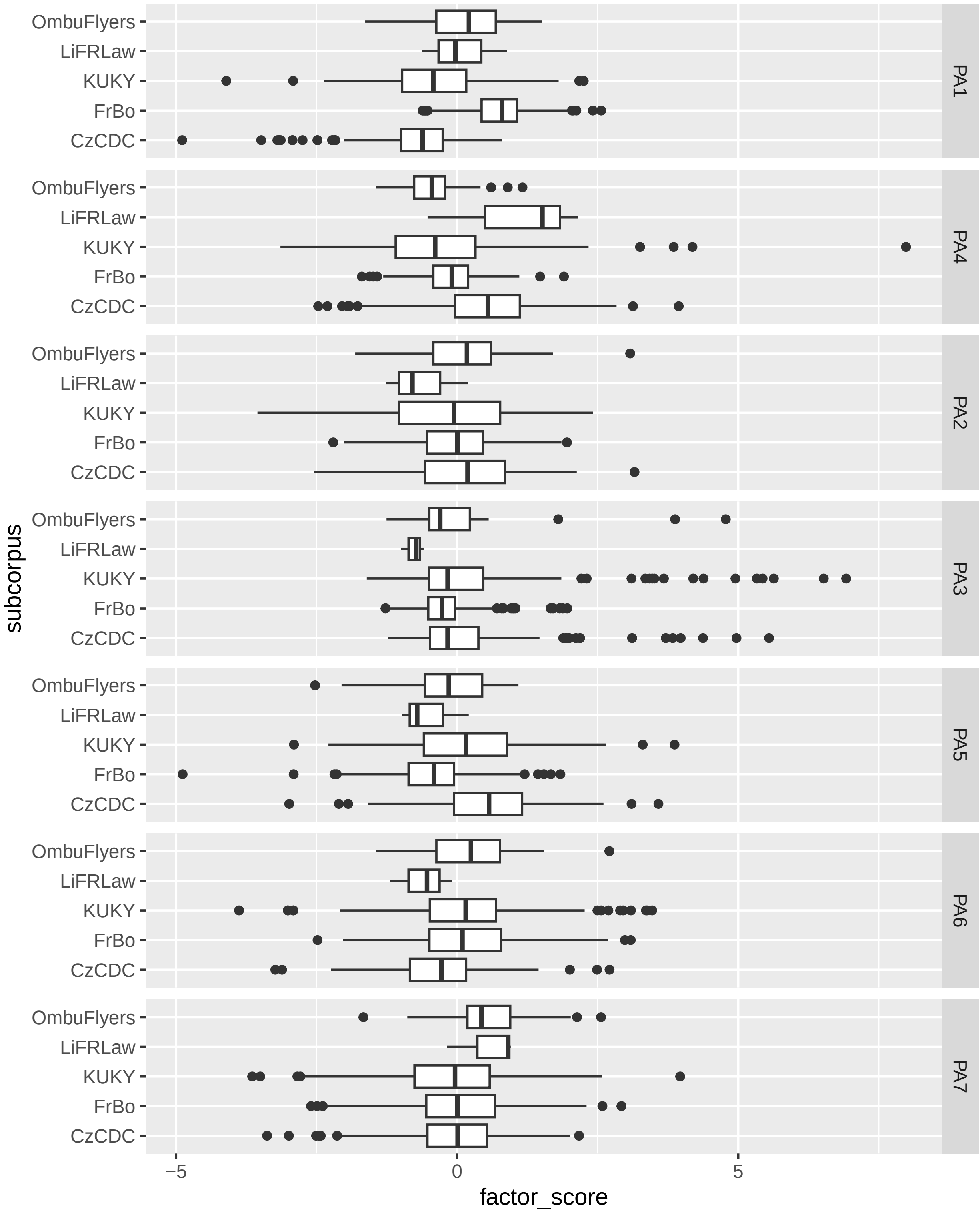
```
data_factors <- bind_cols(data_clean, fa_res$scores %>% as.data.frame())
data_factors_long <- data_factors %>%
  pivot_longer(PA1:PA7, names_to = "factor", values_to = "factor_score") %>%
  mutate(across(
    factor,
    ~ factor(.x, levels = c("PA1", "PA4", "PA2", "PA3", "PA5", "PA6", "PA7"))
  ))

data_factors_long %>% ggplot(aes(x = factor_score, y = 0, color = subcorpus)) +
  facet_grid(factor ~ .) +
  ylim(-0.5, 0.5) +
  theme(
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    legend.position = "bottom"
  ) +
  geom_jitter(width = 0, height = 0.3, alpha = 0.3)
```



```
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 ~ .)
```


OmbuFlyers
LiFRLaw
KUKY
FrBo
CzCDC
PA1
PA4
PA2
PA3
PA5
PA6
PA7
subcorpus
−5
0
5
factor_score