Fig_3EFG

```r
source("source.R")

path <- "data/dat_all.csv"
path_id <- "data/id.csv"

.pc <- 4

dat_id <-
  path_id %>%
  read_csv()
```

```r
dat_raw <-
  path %>%
  read_csv()

dat_n <-
  dat_raw %>%
  group_nest(type, Name, PM, fm, parents, Stage, calltype) %>%
  mutate(n = map_dbl(data, nrow))

dat_n_wide <-
  dat_n %>%
  select(!data) %>%
  pivot_wider(
    values_from = n,
    values_fill = 0,
    names_from = calltype
  )

.calls <-
  dat_n$calltype %>%
  unique()
```

```r
fit_pca <-
  dat_n_wide %>%
  select(.calls) %>%
  prcomp(scale = FALSE)

fit_pca %>% summary()
```

```
## Importance of components:
##                             PC1      PC2     PC3      PC4     PC5      PC6
## Standard deviation     170.1256 135.9318 32.3655 25.10544 18.9490 13.09556
## Proportion of Variance   0.5803   0.3704  0.0210  0.01264  0.0072  0.00344
## Cumulative Proportion    0.5803   0.9507  0.9717  0.98432  0.9915  0.99496
##                             PC7     PC8     PC9    PC10
## Standard deviation     11.71903 8.86341 5.85305 1.09523
## Proportion of Variance  0.00275 0.00157 0.00069 0.00002
## Cumulative Proportion   0.99771 0.99929 0.99998 1.00000
```

```
dat_ngram_pcs <-
  dat_n_wide %>%
  select(c(type, Name, Stage, PM)) %>%
  bind_cols(fit_pca$x %>% data.frame() %>% select(str_c("PC", 1:.pc)))
```

```
.cols <- str_c("PC", 1:.pc)


dat_ngram_MahaD <-
  dat_ngram_pcs %>%
  group_nest(Stage) %>%
  mutate(base = map(data, \(x){
    x %>%
      filter(type == "UE") %>%
      select(starts_with("PC"))
  })) %>%
  mutate(base = map(data, \(x){
    x %>% filter(type == "UE")
  })) %>%
  mutate(
    x = map(data, \(x){select(x, .cols)}),
    center = map(base, \(x){select(x, .cols) %>% summarise_all(mean) %>% unlist()}),
    cov = map(base, \(x){select(x, .cols) %>% var()})
  )%>%
  mutate(maha = pmap(list(x, center, cov), mahalanobis))
```

```
.threshold <- qchisq(0.95, length(.cols))

dat_mahaD <-
  dat_ngram_MahaD %>%
  select(Stage, data, maha) %>%
  unnest(everything())

write_csv(dat_mahaD, "data/data_MahaD_unigram.csv")
```

```
g_maha <-
  dat_mahaD %>%
  ggplot() +
  aes(PM, log10(sqrt(maha)), color = type) +
  geom_rect(xmin = 0, xmax = 6, ymin = log10(0), ymax = log10(sqrt(.threshold)),
            color = NA, fill = "lightgrey") +
  geom_vline(xintercept = c(2.375, 3.875), color = "white") +
  geom_path(aes(group = Name), alpha = 0.25) +
  geom_point() +
  scale_color_manual(values = c(UE = "black", VPA = "red")) +
  theme(legend.title = element_blank()) +
  labs(x = "PM", y = "log10(Mahalanobis D)")
```

```
dat_judge <-
  dat_mahaD %>%
  mutate(judge = if_else(maha <= .threshold, 1, 0))  %>%
  left_join(dat_id %>% select(Name, Pup), by = "Name") %>%
  mutate(y = factor(Pup) %>% as.numeric()) %>%
#  mutate(y = str_c(type, "_", Name),
#         y = factor(y) %>% as.numeric()) %>%
  group_by(type) %>%
#  mutate(ylab = str_c(type, "_", y - min(y) + 1)) %>%
  ungroup()


.labs <-
  dat_judge %>%
  select(y, Pup) %>%
  distinct() %>%
  arrange(y) %>%
  .$Pup

g_judge <-
  dat_judge %>%
  mutate(judge = if_else(judge == 0, "out", "in")) %>%
  ggplot() +
  aes(PM, y, color = type) +
  geom_vline(xintercept = c(2.375, 3.875), linetype = "dashed", color = "darkgrey") +
  geom_line(aes(group = Name))+
  geom_point(size = 3,aes(fill = judge), shape = 21) +
  scale_fill_manual(values = c("white", "red")) +
  scale_color_manual(values = c(UE = "black", VPA = "red")) +
  scale_y_continuous(breaks = 1:16, labels = .labs) +
  theme(axis.title.y = element_blank(),
        legend.title = element_blank())+
  labs(x = "PM")
```

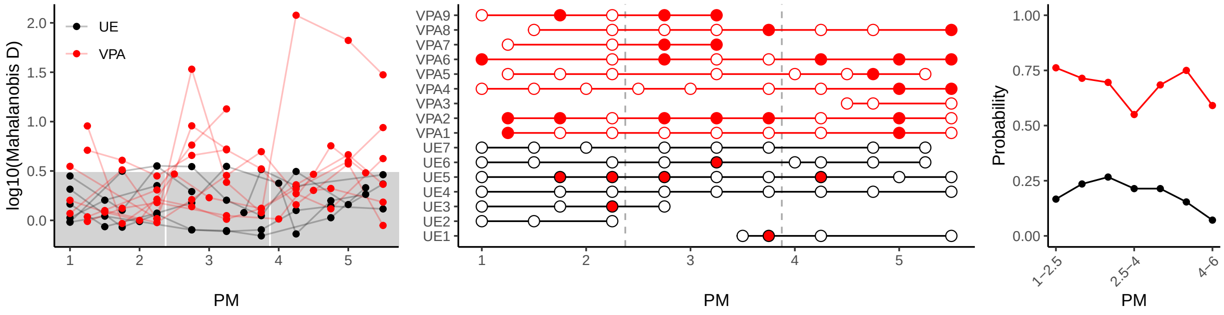
```
dat_r <-
  dat_judge %>%
  filter(type == "VPA") %>%
  group_nest() %>%
  mutate(section = list(seq(1, 4, by = 0.5))) %>%
  unnest(section) %>%
  mutate(end = section + 1.5) %>%
  mutate(end = if_else(end > 5, 6, end)) %>%
  mutate(data = map2(data, section, ~.x %>% filter(PM >= .y))) %>%
  mutate(data = map2(data, end, ~.x %>% filter(PM < .y))) %>%
  mutate(n = map_dbl(data, nrow),
         FN = map_dbl(data, ~sum(.$judge))) %>%
  mutate(r = FN / n) %>%
  mutate(xlab = str_c(section, "-", end))

dat_r2 <-
  dat_judge %>%
  filter(type == "UE") %>%
  group_nest() %>%
  mutate(section = list(seq(1, 4, by = 0.5))) %>%
  unnest(section) %>%
  mutate(end = section + 1.5) %>%
  mutate(end = if_else(end > 5, 6, end)) %>%
  mutate(data = map2(data, section, ~.x %>% filter(PM >= .y))) %>%
  mutate(data = map2(data, end, ~.x %>% filter(PM < .y))) %>%
  mutate(n = map_dbl(data, nrow),
         FN = map_dbl(data, ~sum(.$judge))) %>%
  mutate(r = 1 - FN / n) %>%
  mutate(xlab = str_c(section, "-", end))

g_FN <-
  dat_r %>%
  ggplot() +
  aes(section, r) +
  geom_path(color = "red") +
  geom_point(color = "red") +
  geom_path(data = dat_r2, color = "black") +
  geom_point(data = dat_r2, color = "black") +
  scale_y_continuous(limits = c(0, 1),
                     breaks = seq(0, 1, by = 0.25)) +
  scale_x_continuous(breaks = c(1, 2.5, 4),
                     labels = c(dat_r$xlab %>% .[c(1, 4, 7)])) +
  labs(x = "PM", y = "Probability") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
g <-
  patchwork::wrap_plots(
  g_maha +
    theme(legend.position = c(0, 1),
          legend.justification = c(0, 1),
          legend.background = element_rect(fill = NA)),
  g_judge +
    theme(legend.position = "none"),
  g_FN,
  widths = c(2, 3, 1)
)

g
```



```r
ggsave("fig/gram4_pca_mahaD_unigram.png", g,
       width = 11, height = 4)

ggsave("fig/gram4_pca_mahaD_unigram.svg", g,
       width = 11, height = 3)
```

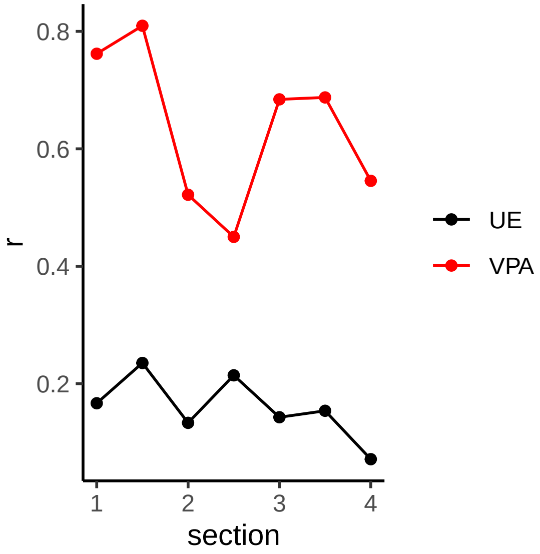
```r
.cols <- str_c("PC", 1:.pc)


dat_ngram_MahaD2 <-
  dat_ngram_pcs %>%
  group_nest() %>%
  mutate(section = list(seq(1, 4, by = 0.5))) %>%
  unnest(section) %>%
  mutate(end = section + 1.5) %>%
  mutate(end = if_else(end > 5, 6, end)) %>%
  mutate(data = map2(data, section, ~.x %>% filter(PM >= .y))) %>%
  mutate(data = map2(data, end, ~.x %>% filter(PM < .y))) %>%
  mutate(base = map(data, \(x){
    x %>%
      filter(type == "UE") %>%
      select(starts_with("PC"))
  })) %>%
  mutate(base = map(data, \(x){
    x %>% filter(type == "UE")
  })) %>%
  mutate(
    x = map(data, \(x){select(x, .cols)}),
    center = map(base, \(x){select(x, .cols) %>% summarise_all(mean) %>% unlist()}),
    cov = map(base, \(x){select(x, .cols) %>% var()})
  )%>%
  mutate(maha = pmap(list(x, center, cov), mahalanobis))


.threshold <- qchisq(0.95, length(.cols))

dat_judge2 <-
  dat_ngram_MahaD2 %>%
  select(section, end, data, maha) %>%
  unnest(everything()) %>%
  mutate(judge = if_else(maha <= .threshold, 1, 0)) %>%
  group_nest(section, end, type) %>%
  mutate(IN = map_dbl(data, ~sum(.$judge)),
         n = map_dbl(data, nrow),
         r = IN / n,
         r = if_else(type == "UE", 1 - r, r))

dat_judge2 %>%
  ggplot() +
  aes(section, r, color = type) +
  geom_path() +
  geom_point() +
  scale_color_manual(values = c(UE = "black", VPA = "red"))
```


0.8
0.6
0.4
0.2
r
1
2
3
4
section
UE
VPA