Estimators for Topic Sampling Designs

Reproduction tl;dr

The code below (or `00-do-all.R`) reproduces all analysis in the paper. The entire analysis takes about 17 hours to complete.

The file `install-packages.R` (written dynamically below) will install all the packages I use for this project in the appropriate version. Please run if you want to use the same package versions as I did.

```
# remove all generated files
unlink("output/*")
unlink("output/simulations/*")

# run each component of the analysis
# note: scripts 03a, 03b, and 03c, print reported quantities
#       so you probably want to run those more interactively,
#       but they are very fast
source("01-sdc-clean-data.R")
source("02-sdc-do-sims.R")
source("03a-eval-delta-bar.R", echo = TRUE)
source("03b-eval-sigma.R", echo = TRUE)
source("03c-eval-delta-j.R", echo = TRUE)
```

Output

The code above computes several quantities of interest reported in the paper and produces two figures

- `figs/fig1-sdc-sd.png`, which corresponds to Figure 1 in the paper.
- `figs/fig2-sdc-part.png`, which corresponds to Figure 1 in the paper.

Session Information

```
  sessioninfo::session_info(info = "platform")
```

```
- Session info -------------------------------------------------------------
 setting  value
 version  R version 4.3.2 (2023-10-31)
 os       macOS Ventura 13.4
 system   aarch64, darwin20
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/New_York
 date     2024-02-23
 pandoc   3.1.6.1 @ /usr/local/bin/ (via rmarkdown)

-------------------------------------------------------------------------------
```

System Information

```
  # os info
  Sys.info() |>
    tibble::as_tibble(rownames = "par") |>
    dplyr::filter(par != "nodename")
```

```
# A tibble: 7 x 2
  par            value
  <chr>          <chr>
1 sysname        Darwin
2 release        22.5.0
3 version        Darwin Kernel Version 22.5.0: Mon Apr 24 20:53:19 PDT 2023; ro~
4 machine        arm64
5 login          root
6 user           carlislerainey
7 effective_user carlislerainey
```

```
# cpu info
benchmarkme::get_cpu()
```

```
$vendor_id
character(0)

$model_name
[1] "Apple M2 Max"

$no_of_cores
[1] 12
```

```
# ram info
ram_info <- system("sysctl hw.memsize", intern = TRUE)
print(ram_info)
```

```
[1] "hw.memsize: 34359738368"
```

R Packages Used

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
# Function to get package version
get_version <- function(package_names){
  sapply(package_names, function(pkg) as.character(packageVersion(pkg)))
}
```

```r
dep <- renv::dependencies() %>%
  bind_rows(renv::dependencies("README.qmd")) %>%
  select(Package) %>%
  distinct() %>%
  arrange(Package) %>%
  mutate(Version = get_version(Package))
```

```
Finding R package dependencies ... Done!
Finding R package dependencies ... Done!
```

```r
dep
```

```
       Package  Version
1          arm   1.13.1
2  benchmarkme    1.0.8
3   doParallel   1.0.17
4        dplyr    1.1.4
5      foreach    1.5.2
6   kableExtra    1.4.0
7         lme4 1.1.35.1
8    patchwork    1.2.0
9       readxl    1.4.3
10     remotes  2.4.2.1
11        renv    1.0.3
12   rmarkdown     2.25
13    rstanarm   2.32.1
14      scales    1.3.0
15 sessioninfo    1.2.2
16      tibble    3.2.1
17   tidybayes    3.0.6
18   tidyverse    2.0.0
```

Write Script to Install Packages and Versions

```r
# create a new script file
script_file <- file("install-packages.R", "w")

# write install.packages() commands for each package to the file
```

```r
for (i in 1:nrow(dep)) {
  install_cmd <- sprintf("remotes::install_version('%s', version = '%s', repos = 'http://c
  cat(install_cmd, file = script_file, append = TRUE)
}

# close the file
close(script_file)
```