Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech

Matthew Gentzkow
Jesse Shapiro
Matt Taddy

March 2019

Replication Code and Data

A Archive Overview

The directories listed in Table 1 produce all plots, tables and calculations reported in the paper and online appendix.

Table 1: Archive Overview

Directory Name	Purpose
environment	Initializes environment for analysis on local machines and computing clusters
lib	Houses abstracted scripts called on by multiple directories in `/source/analysis`
output	Houses output and intermediate data from directories in `/source/analysis`
data	Contains all data necessary for the replication of tables and figures in the main paper and online appendix
source/analysis	Estimates of all reported models and constructs of all figures
source/paper/	Houses .lyx files for main paper and online appendix
source/raw/	Contains several small raw data files used in `/source/analysis`

B Directory Structure

Each subfolder of `/source/analysis` contains a `make.py` script which executes all scripts in the required order. You will not be able to run this script. Treat it as documentation on the correct order of the other scripts.

C Point Estimates for Maximum Likelihood, Leave-out, and Penalized Estimators

Table 2 gives the locations of the `make.py` scripts which produce point estimates of the "real" series for the maximum likelihood, leave-out, and penalized estimators.

Table 2: Paths to Estimators

Estimator	Plot	Path
Maximum Likelihood	Figure 1.A	`source/analysis/descriptive_measures/plugin_partisanship`
Leave-out	Figure 2.A	`source/analysis/descriptive_measures/loo_expected_posterior`
Penalized	Figure 2.B	`source/analysis/mn_model_estimation/mn_nosmoothing`

D Proprietary Data

The data underlying Online Appendix Figure 21 are proprietary and available by subscription from GfK Mediamark Research & Intelligence. We have provided code to clean and analyze the raw data. This code can be used by a researcher who subscribes to these data.

E System Requirements

All analysis code was originally run on Mac OS X machines as well as the University of Chicago Midway and Stanford University Sherlock research computing clusters. A research computing cluster is necessary for the estimation of most specifications.