1 Overview

This folder contains the data and code needed to replicated the results of "A Novel Class of Unfolding Models for Binary Preference Data".

2 Computational Requirements

This code was run on the standard compute nodes of University of Washington's statistics cluster. This cluster became available to use in 2023 and features 23 compute nodes with 1184 cores and 6.8TB of RAM. The CPU models for the nodes are Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz, Gold 5220R CPU @ 2.10GHz, and Gold 5220R CPU @ 2.20GHz. 37 cores need to be available to fit the models and 42 cores are required to simultaneously compute the various statistics for the plots. While the script only needs one core to makes the figures, that script is the most memory intensive because it requires 70 GB.

3 Data

We analyzed Voteview vote data from the 100th to 117th House of Representative and Martin-Quinn's Supreme Court data from the 1937 to the 2021 term. In this folder, we included the processed version of these files.

The processed version can be derived in the following manner. For Voteview data, votes coded as 1-3 in the Voteview data are re-coded as 1, signifying a yes vote. Votes coded as 4-6 were re-coded as 0, signifying a no vote. Otherwise, the votes were coded as NA or missing. Any legislators with 40% or more missing votes were excluded. Further, any unanimous votes were removed. Meanwhile, we did not change how Martin and Quinn coded the votes. Instead, the processed data files includes a vector with the vote's term. Note that the terms are re-numbered from 1 to 85 instead of 1937 to 2021. It also includes information on which justices' ideological preferences should be positive or negative and for which of the terms they served should this constraint be enforced. In particular, we fixed the sign of the ideal point of Justices Hugo Black for his first term to be negative, William Douglas for the first 29 terms to be negative, Thurgood Marshall for his entire term in office to be negative, and Clarence Thomas for his entire term in office to be negative.

4 Content

- install_packages.R: This script includes all R packages used in the analysis. Here, we used R/4.2.

- *.sh: Once the R packages have all been installed, these scripts executes the code needed to run all the models, process the results, and make the figures. The scripts need to be run in the following order.

    1. run_comparison_mcmc.sh: Executes the MCMC code for IDEAL and BGGUM on the U.S. House of Representative vote data and Martin-Quinn's model on the U.S. Supreme Court vote data from 1937.
    2. run_pum_mcmc.sh: Runs the MCMC code for the probit unfolding model on the U.S. House of Representative vote data and the dynamic probit unfolding model on the U.S. Supreme Court vote data from 1937.
    3. gen_summary_files.sh: Executes the code to process the MCMC draws from the first two steps in order to generate summary information needed for the figure.
    4. plot_figure.sh: Runs the code needed to plot the figures.

- data
    - raw: This folder contains all of the processed data files.
        * processed_house_votes_*.Rdata: Contains the processed Voteview vote data for the corresponding House session of that number.
        * mq_supreme_court_vote_info_2021.Rdata: Contains the Supreme Court vote data from Martin-Quinn's data since the 1937 term.
        * justices_2021.csv: Contains the Martin-Quinn scores for the justices since the 1937 term.
        * H107_votes.ord, H107_*.csv: Contains the raw vote information that is required to find the vote information for votes from the 107th House.
    - analysis: This folder contains the results from the models and intermediate files from processing the results.
        * alg_results: Contains the processed Voteview vote data for the corresponding House session of that number.
            · bggum: Contains BGGUM's MCMC samples for the House with that session number in the file name
            · ideal: Contains IDEAL's MCMC samples for the House with that session number in the file name
            · pum: Contains the probit unfolding model's MCMC samples for the House with that session number in the file name. Note that _alt indicates an alternative prior being used.
            · mq: Contains the Martin-Quinn MCMC samples for the US Supreme Court since the 1937 term. Each number indicates a different chain.
            · dynamic_pum: Contains the Martin-Quinn MCMC samples for the US Supreme Court since the 1937 term. Note that _alt indicates an alternative prior being used.
        * response_curve: Contains information on the response curve for three votes from the 116th House for the probit unfolding model under various priors.

* summary_info: Each file contains the posterior median ranks of the legislators according to the various models from the House of that number. It also has information on the range, standard deviation, and IQR of the Republicans and Democrat ideological preferences.
* waic_info: Contains the WAIC computed for the various models.
  - house_vote_*_block_waic_comp.Rdata: Contains the computed WAIC blocked by legislator for IDEAL, BGGUM, and the probit unfolding model for the House with that session number. Note that _alt indicates an alternative prior being used.
  - house_vote_107_block_vote_waic_comp.Rdata: Contains the WAIC blocked by vote for IDEAL, BGGUM, and the probit unfolding model for the 107th House.
  - mq_2021_dynamic_pum_comp_alt.Rdata, mq_2021_dynamic_pum_comp_alt.Rdata: Holds the computed WAIC blocked by judge and year for MQ and the dynamic probit unfolding model for the US Supreme Court since the 1937 term. Note that _alt indicates an alternative prior being used.

- code: This folder contains the C++ and R code needed to run the models and derive the intermediate files. The version of R used is 4.2.2. The packages used include *tidyverse* version 2.0.0, *mvtnorm* version 1.1-3, *msm* version 1.7.1, *latex2exp* version 0.9.6, *ggrepel* version 0.9.3, *bggum* version 1.0.2, *MCMCpack* version 1.6-3, *Rcpp* version 1.0.10, *RcppDist* version 0.1.1, *RcppArmadillo* version 0.12.0.1.0, *RColorBrewer* version 1.1-3, *wnominate* version 1.4, and *reshape2* version 1.4.4.
  - C++/: three_utility_probit_helper_functions.cpp contains the C++ code needed to run the probit unfolding model, dynamic probit unfolding model, and compute the WAIC. rtn1.* are helper functions to generate draws from the truncated normal distribution.
  - R/alg_code: Contains the code to fit the models
    * run_bggum.R (∼16 hours, 34 minutes to 30 hours, 30 minutes): Runs R code to fit BGGUM to the data as implemented in the package, *bggum*, to Voteview data.
    * run_probit.R (∼10 to 18 minutes): Runs R code to fit the IDEAL model as implemented in the package, *MCMCpack*, to Voteview data.
    * sample_probit_bggum.R: Contains the R code needed to run the probit unfolding model. It uses the R package, *Rcpp*, to interface with C++ and use functions from three_utility_probit_helper_functions.cpp.
    * run_sample_probit_bggum.R (∼12 hours, 39 minutes to 21 hours, 22 minutes): Contains the R code used to fit the probit unfolding model to Voteview data. This script uses functions from sample_probit_bggum.R. Note that _alt fits the probit unfolding model with an alternative prior.
    * run_mq_irt.R (∼9 hours, 47 minutes): Runs R code to fit the Martin-Quinn model as implemented in the package, *MCMCpack*, to Martin-Quinn's dataset.
    * sample_dynamic_probit_bggum.R: Contains the R code needed to run the dynamic probit unfolding model. It uses the R package, *Rcpp*, to interface with C++ and use functions from three_utility_probit_helper_functions.cpp.

    - run_sample_dynamic_probit_bggum.R (∼20 hours): Contains the R code used to fit the dynamic probit unfolding model to Martin-Quinn's dataset. This script uses functions from sample_probit_bggum.R. Note that _alt fits the dynamic probit unfolding model with an alternative prior.
  - R/info_code: Contains code needed to generate the intermediate files from the results.
    - calc_waic.R: Contains functions in R to compute the WAIC for various models. It uses the R package, *Rcpp*, to interface with C++ and use functions from three_utility_probit_helper_functions.cpp.
    - run_calc_waic_block.R (∼1 hour, 23 minutes to 2 hours, 34 minutes): Runs R code to compute the WAIC blocked by legislator for a given House session based on the Voteview data and the IDEAL, BGGUM, and the probit unfolding models. This script calls functions from calc_waic.R.
    - run_calc_waic_block_vote.R (∼1 hour, 17 minutes): Runs R code to compute the WAIC blocked by vote for a given House session based on the Voteview data and the IDEAL, BGGUM, and the probit unfolding models. This script calls functions from calc_waic.R.
    - run_calc_waic_dynamic.R (∼7 hours, 40 minutes): Runs R code to compute the WAIC blocked by legislator for Martin-Quinn's data and the Martin-Quinn model and the dynamic probit unfolding models. This script calls functions from calc_waic.R.
    - calc_response_curve.R (∼4 hours, 55 minutes): Runs R code to compute the response curve for three votes from the 116th House.
    - run_summarize_results_stats.R (∼3 minute): Runs R code to compute the posterior median ranks from various models for a given House sessions. It also calculates the IQR, standard deviation, and range for the Democratic and Republican ideological preferences.
  - R/alg_code: Contains code from the Annals of Applied Statistics paper, AOAS1454, needed to process H107_votes.ord to get the vote information.

- plot_figure.R (∼17 minutes): This script contains the R code used to make the figures in the paper. This script should only be run after all algorithms have finished running and the results have been processed.

- results: This folder contains the plots created by plot_figure.R. Main text figures are named figure_*.pdf and supplementary material figures are named figure_supplementary_*.pdf. In both cases, the number corresponds to the figure number in the main text.

Note: Because of Markov chain Monte Carlo error, the posterior draws may vary depending on the machine used. This may lead to plots that are different than those in the paper, such as Figure 5 in the main text and Figure 5a, 5b, 6a, and 6b in the supplementary material.

References

Duck-Mayr, J. & Montgomery, J. (2022). Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons. *Political Analysis* pp. 1–20.

Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2023). Voteview: Congressional Roll-Call Votes Database.

Martin, A. D. & Quinn, K. M. (2002). Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* **10**, 134–153.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software* **42**, 22.

Yu, X. & Rodriguez, A. (2021). Spatial voting models in circular spaces: A case study of the us house of representatives. *The Annals of Applied Statistics* **15**, 1897–1922.