Estimating Latent Traits from Expert Surveys ReadMe

Kyle L. Marquardt and Daniel Pemstein

May 4, 2021

This dataset contains replication materials for "Estimating latent traits from expert surveys," which this document describes in detail. Note that the simulation analyses involve over 300 latent variable model estimates (three replications of hierarchical and non-hierarchical analyses, with two different underlying distributions, two-three data generating processes, and five different models). Each of these models requires eight cores and between one hour and seven days to run using cluster computing resources at the Swedish National Infrastructure for Computing (OS: CentOS 6 and Ubuntu Xenial [16.04 LTS]).

Running replication code—excluding the run time for latent variable models—takes approximately 1 minute to six hours (OS: Ubuntu 19.10, 64 bit). Analyses require the following R libraries:

- ggplot2
- reshape2
- parallel
- rstan
- truncnorm
- zoo
- coda

The replication data set consists of a main folder with four subfolders. In addition to the subfolders, the main folder also contains the file necessary to both run the bootstrap analyses and compile summary statistics for the latent variable models associated with all analyses of simulated data (*simulation_dataset_script.R*), the resulting dataset (**sim_ds.rds**), and the file which uses this dataset to create all graphics in the paper and appendix (*graphics.R*). Note that *simulation_dataset_script.R* assumes the folder structure described here to run, since it loads files from subfolders.[1]

[1]Due to file size restrictions, we have removed all simulation posterior .rds files from the replication files. Since *simulation_dataset_script.R* requires these files, it will also not run. Simulation posterior .rds files are available from the authors upon request.

The contents of the subfolders are relatively straightforward. The first folder, **Basic_Data**, contains the .rds files that serve as the basis for all analyses in the manuscript (coder-level data for the V–Dem variable "Freedom from political killings"); it also contains analyses of these data. The remaining three subfolders; **R1**, **R2**, and **R3**; represent the three replications of the analyses of simulated data. Accordingly, they have the same structure. Specifically, the main folder for each replication contains the file which creates simulated data (e.g. expert characteristics under different data generating assumptions; *simulation_code.R*) and the resulting simulated dataset (**simdata.rds**).[2] It also contains two subfolders, **Hierarchical** and **NHierarchical**, which contain replication materials for hierarchical and non-hierarchical data generating processes. Both **Hierarchical** and **NHierarchical** each contain two subfolders, **normal** and **uniform**, which contain replication materials for simulated data that have normally- and uniformly-distributed underlying data structures. Both **normal** and **uniform** contain three folders; **am**, **irt**, and **uniform**; which contain replication materials for simulated data that follow A–M, IRT or uniform data generating processes (since there is no simulated hierarchical uniform data generating process, hierarchical replication folders do not contain **uniform**). Finally, **am**, **irt**, and **uniform** all contain the folders **high** and **low**, which refer to the level of simulated error for each data generating process (in the case of the uniform errors, "high" is the truncated-uniform data generating process).

We discuss the contents of the folders in greater detail in turn.

1 Basic Data

This folder contains the .RData files that serve as the basis for all analyses in the manuscript. Specifically, it includes **v2clkill_reduced.rds**, the basis for all analyses in the paper; and **v2clkill_interpolated.rds**, which contains the data we use to create country-date graphics.

The file **v2clkill_reduced.rds** contains coder-level data for the V–Dem v8 variable, "Freedom from Political Killings" (v2clkill). "_reduced" signifies that these data have been reduced from country-date to country-regimes, which we define as periods in which no experts change either their self-reported confidence in their coding or their actual codings. The object includes six objects of relevance to the manuscript:

wdata : Country-regime × coder data on regime trait (political killings); ordinal 1-5. Codes of -1 represent missing data.

N : Number of observations (country-regimes).

J : Number of coders.

K : Number of ordinal values.

cdata : Identifier of main-country coded for each coder (for use in hierarchical clustering of thresholds).

[2] *simulation_code.R* requires **v2clkill_reduced.rds** located in **Basic_Data**.

The file **v2clkill_interpolated.rds** contains the non-reduced data for "Freedom from political killings," which we use to interpolate the data.

Finally, there is the file *analysis.R*, which creates the dataset **zds.RData**. The file *basic_analysis.R* uses **zds.RData** to produce Figures 6-7 and the figures in Appendix B. The remaining files in the folder are related to different models.

1.1 Models

We mainly model the country-regime latent trait—in both the actual (non-simulated) and simulated data— using five different latent variable models.

Each of the five models has a set of four files: *x.R*, the R file for running the analysis; *x.stan*, the Stan model file; *x.sh*, the job file; and **v2clkill.i.x.RData**, *i* posterior draws from *x* model run on the actual V–Dem data. Each *.RData* also has an associated *.out* file; the *.out* file that is temporally closest to a *.RData* file is the log file for that file. The names of the files designate the specific model:

beta : IRT model that accounts for DIF with coder-specific hierarchical thresholds clustered about both the main country an expert codes and universal thresholds.

irt : IRT model that accounts for DIF with coder-specific hierarchical thresholds clustered about only universal thresholds.

irt2 : IRT model that accounts for DIF with coder-specific hierarchical thresholds without clustering.

ham : A–M model that accounts for DIF with coder-specific intercept and slope parameters clustered about both the main country an expert codes and universal intercepts and slopes.

bam : A–M model that accounts for DIF with coder-specific intercept and slope parameters without clustering.

Note that the R files use **v2clkill_reduced.rds** as their input data.

1.2 zds.RData

This file summarizes the analyses of the actual V–Dem data, "Freedom from political killings." Rows represent country-date observations. Columns are as follows:

wmed : Zero-centered median across expert scores

z.avg : Normalized mean across expert scores

z.bavg : Mean across bootstrapped normalized mean draws.

z.bmed : Median across bootstrapped zero-centered median draws.

z.am : Posterior median from A–M model.

z.ham : Posterior median from hierarchical A–M model.

z.irt : Posterior median from IRT model.

z.hirt : Posterior median from hierarchical IRT model (one level of clustering).

z.hirt2 : Posterior median from hierarchical IRT model (two levels of clustering).

reduced : Indicator for whether or not an observation is the a reduced country-date.

cid : Country ID.

date : Date.

"lower" and "upper" represent the upper and lower bounds for 95% confidence intervals/credible regions about different model point estimates.

2 Simulation Analysis

All lowest level simulation analysis folders contain the analyses of a single simulated dataset with a specific data generating process for a given replication (e.g. **R1/NHierarchical/uniform/uniform/low** contains analyses of a simulated dataset in the first replication with a non-hierarchical structure, uniformly-distributed true-values, and a data-generating process that uses uniform distributions with low error). The files included in each folder are accordingly the same, and there is substantial overlap with the files in the folder **Basic_Data**. Specifically, we model the country-regime latent trait—in both the actual (non-simulated) and simulated data— using the same five different latent variable models.

Again, each of the five models has a set of four files: *x.R*, the R file for running the analysis; *x.stan*, the Stan model file; *x.sh*, the job file; and **v2clkill.i.x.rds**, i posterior draws from x model run on the simulated data (**v2clkill.rds**).[3] Each *.rds* also has an associated *.out* file; the *.out* file that is temporally closest to a *.rds* file is the log file for that file. Model nomenclature is the same as in the folder **Basic_Data**.

In addition to model files, each folder contains the file *sim.R*, which converts the simulated forms of error from the root directory for a given replication into the simulated dataset, **v2clkill.rds**. Note that this file assumes the nested file structure of the dataset.

3 sim_ds.rds

This file is the output from *simulation_dataset_script.R*; *graphics.R* uses it to create most of the graphics in the article (Figures 2-5, Appendix Figures E-G). *simulation_dataset_script.R* produces Figure 1. Object rows summarize the output of analyses of each model used to aggregate simulated data sets. Columns are as follows:

[3]We removed the simulation posterior .rds files from the replication dataset due to file size limitations. They are available from the authors upon request.

Model : Model for aggregating data. Model nomenclature is the same as elsewhere, except that "beta" becomes "HIRT2" and "irt2" become "HIRT."

Variation : Level of simulated expert error ("Low" or "High").

DGP : Simulated DGP ("A–M," "IRT" and "Uniform").

Distribution : Distribution of true values ("Normal" and "Uniform"),

Hierarchical : Whether or not simulated data have hierarchical structure ("Y" and "No").

Replication : Simulated replication ("1," "2" and "3").

Converged : Proportion of latent variable estimates with $\hat{r} > 1.1$.

MSE : Mean square error of estimated values from true values.

Pearson : Pearson correlation correlation of estimated values with true values.

Kendall : Kendall correlation correlation of estimated values with true values.

HPD : Credible region coverage.