CPT estimation with R and stan

Alina Ferecatu

Last compiled on 2021-12-19

Contents

Overview

This notebook accopanies the code in the *CPT_estimation_Rstan* Github repository. We estimate the cumulative prospect theory model with a power value function and the Gonzalez and Wu (1999) probability weighting function.

The most recent version of this package can be found at: **https://github.com/alinafere/CPT_estimation_Rstan**.

The estimation runs in R, using the *Rstan* package: **https://mc-stan.org/users/interfaces/rstan**. In addition to R, you need to install *rstan*, *loo*, and *tidyverse* on your machine.

The estimation is based on two types of data: 1. data gathered using the laddering technique (see Wu and Gonzalez (1996)) and 2. data gathered using two-outcome gamble pairs (see Rieskamp (2008)).

The analyses were conducted on a Macbook Pro with an Intel Core i7 processor, speed 3.5 GHz, and memory 16GB RAM.

This README-file contains some details on how to run the analyses in R. For each analysis, I describe the following:

- **Data**: The input data and variable names,
- **Output**: Describes the output files and variables,
- **Code Structure**: How to run the codes.

Structure of each R file

- **Prerequisites**: The necessary packages are listed in the beginning of each file.
- **Set paths**: Set local paths necessary to load the data and the save the outputs. The paths are PATH_FUNCTION, PATH_DATA, PATH_RESULTS, PATH_PLOTS.
- **Source functions**: source the R file *COT_functions.R*, that includes functions relevant for all analyses. The file containing the functions is loaded in each R file.
- **Data**: Loads the input data for each analysis. I describe below the data and variable names per R file.
- **Output**: Save the output for each analysis. I describe below the type of output for each R file.

Source functions used in all R files

R code *CPT_functions.R*

Data

None

Output

None

Code Structure and Instructions

- Code structure:
  - Basic functions.
  - Estimation functions written in *stan* two CPT models, for each type of data.
  - Model comparison functions, that computes in-sample fit measures.
- Instructions:
  - Source the functions file in all other R files.

CPT estimation for laddering technique (see Wu and Gonzalez 1996)

R code *CPT_estimation_for_laddering_technique.R*

The r code *CPT_estimation_for_laddering_technique.R* estimates the CPT model for the Wu and Gonzalez (1996) data set gathered using the laddering technique.

Data

1. *wu_gonzalez_1996_dataR.csv*, the dataset published adn explained in Wu and Gonzalez (1996). It contains the following variables:

- *Survey*: Each survey contains a mix of choices from a subset of rungs in each ladder: when using ladders 1, 2, 4, takes either first or last 4 rungs; when using ladders 3 and 5, takes either even or odd rungs.
- *Choice*: 1 risky, 0 safe, turned into the variable *decision* in the estimation: 1- risky; 2-safe.
- *Date*: when the surveys took place: ladders 1-2-4 in April, ladders 3-5 in May.
- *r1*: probability and payoff of the first R gamble, set within each ladder.
- *r2*: probability and payoff of the second R gamble, changing with the common consequence in each ladder.
- *s1*: probability and payoff of the S gamble, changing with the common consequence in each ladder.
- *ladder*: ladder index, 1 through 5.
- *rung*: rung id in ladder, 1 through 8. There are 8 per ladder, each answered by 105 subjects.
- *laddername*: ladder name using the first (R, S) gamble pair of each ladder.
- *Subjectid*: subject unique id (using the survey id). Total of 420 subjects. Each subject answered 8 or 12 questions.
- *ps*: probability of the S gamble (includes probability 1).

Output

The following output types are reported:

1. Two *stanfit* objects, which contain the estimation results of the CPT individual and aggregate level models. We used two HMC chains, with 5000 iterations per chain. The first 3000 iterations are used for warmup. The two chains run in parallel. The estimation of the aggreagte level model takes 20 minutes to run. The estimation of the individual level model takes about 2 days and 6 hours to run. This is because we have a large number of individuals each with their unique CPT parameter estimates.
2. Parameter estimates, model comparison results and posterior predictive checks (correlation between observed and predicted safer choice proportions).
3. Figures comparing the aggregate and the individual level estimations.
4. Estimated value and probability weighting function plots.

Code Structure

- Code structure:
  - Import data and functions.
  - Summary stats.
  - Stan estimation.
  - Parameter estimates and model comparions.
  - Comparison plots and function plots.

CPT estimation for two-outcome gamble pairs (see Rieskamp (2008) Study 2)

R code *CPT_estimation_for_two_outcome_gamble_pairs.R*

The r code *CPT_estimation_for_two_outcome_gamble_pairs.R* estimates the CPT model for the Rieskamp (2008) data set gathered using two outcome gamble pairs. Subjects choose between option A with gambles (p1, x1; q1, y1) and option B with gambles (p2, x2; q2, y2). The data set is publicly available at https://www.researchgate.net/publication/258513197_DATA_Study2_Rieskamp_2008 [retrieved on Dec. 16, 2021].

Data

1. *rieskamp08_gambles.csv*, records the gamble pairs. It contains the following variables:

- *choicepair*: gamble pair id, from 1 to 180 (1:60 - gains, 61:120 - losses, 121:180 - mixed gambles)
- *A*: first option
- *B*: second option
- *A1*: first outcome
- *A2*: second outcome
- *B1*: first outcome
- *B2*: second outcome
- *prob*: probability
- *payoff*: monetary consequence

2. *rieskamp08_choices.csv*, records subjects' choices. It contains the following variables:

- *choicepair*: same as in *rieskamp08_gambles.csv*.
- *subjects*: columns 1 to 30, indexing the subject id.

Output

The following output types are reported:

1. Two *stanfit* objects, which contain the estimation results of the CPT individual and aggregate level models. We used two HMC chains, with 5000 iterations per chain. The first 3000 iterations are used for warmup. The two chains run in parallel. The estimation of the aggregate level model takes about 30 minutes to run. The estimation of the individual level model takes about 2.5 hours to runs.
2. Parameter estimates, model comparison results and posterior predictive checks (correlation between observed and predicted).
3. Figures comparing the aggregate and the individual level estimations.
4. Estimated value and probability weighting function plots.

Code Structure

- Code structure:
  - Import data and functions.
  - Summary stats.
  - Stan estimation.
  - Parameter estimates and model comparions.
  - Comparison plots and function plots.

References

Gonzalez R, Wu G (1999) On the Shape of the Probability Weighting Function. Cognitive Psychology 38(1):129-166.

Rieskamp J (2008) The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition 34(6):1446.

Wu G, Gonzalez R (1996) Curvature of the Probability Weighting Function. Management Science, 42(12):1676-1690.