Replication Package for: Estimation based on nearest neighbor matching: from density ratio to average treatment effect

Overview

This replication package accompanies Zhexiao Lin, Peng Ding, and Fang Han. (forthcoming). "Estimation based on nearest neighbor matching: from density ratio to average treatment effect". Econometrica.

Data Availability and Provenance Statements

- □ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

(Optional, but recommended) License for Data

See LICENSE.txt for details.

Summary of Availability

- □ All data **are** publicly available.
- ☑ Some data **cannot be made** publicly available.
- □ **No data can be made** publicly available.

Details on each Data Source

Data.Name	Data.Files	Location	Provided	Citation
LaLonde data	exp_generated.feather	data/	FALSE	Athey et al (2021)
Shadish et al. data	shadish.txt	data/	TRUE	Shadish et al (2008)

where the `Data.Name` column is then expanded in the subsequent paragraphs.

LaLonde data

> Simulated LaLonde data is generated by Athey et al (2021), and can be directly downloaded from https://drive.google.com/drive/folders/1CCU1zuibrPOZHK6NAAAVrX5TKQZZOMTF.

Data file: `data/exp_generated.feather`

Shadish et al. data

> Shadish et al. data is from Shadish et al (2008) and is not redistributed. Original data can be requested from Shadish. Save the file in the folder `data`.

Data file: `data/shadish.txt`

Dataset list

Data file	Source	Notes	Provided
`data/exp_generated.feather`	Athey		Yes
`data/shadish.txt`	Shadish	Confidential	No

Computational requirements

Software Requirements

- Python 3.8.12
  - `numpy` 1.22.1
  - `torch` 1.13.1
  - `pandas` 1.2.4
  - `matplotlib` 3.6.3
  - `pyarrow` 9.0.0
  - `POT` 0.9.0
  - `wgan` 0.2
  - `hypergrad` 0.1
  - the file "`requirements.txt`" lists these dependencies; please run "`pip install -r requirements.txt`" at the first step. See https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability for further instructions on creating and using the "`requirements.txt`" file.
- R 4.3.0
  - `dplyr` (1.1.2)
  - `FNN` (1.1.3.2)
  - `xtable` (1.8-4)
  - `optparse` (1.7.3)
  - `future.apply` (1.10.0)
  - `feather` (0.3.5)
  - `tictoc` (1.2)
  - `future` (1.32.0)

  - `arrow` (12.0.1)
  - the file "`check_dependency.R`" will install all dependencies (latest versions), and should be run prior to running other programs.

Portions of the code use bash scripting and they require Linux.

Controlled Randomness

- ☑ Random seed is set at line **78** of the program **“comparison.R”**; line **77** of the program **“comparison_shadish.R”**; line **77** of the program **“se.R”**; line **76** of the program **“se_shadish.R”**; lines **6-7** of the program **“gan_shadish_baseline.py”**.

We found that different operating systems and machines could render different results for the same random seed. This observation is supported by the PyTorch document in https://pytorch.org/docs/stable/notes/randomness.html, which states that "*[c]ompletely reproducible results are not guaranteed across PyTorch releases, individual commits, or different platforms. Furthermore, results may not be reproducible between CPU and GPU executions, even when using identical seeds*".

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☐ 2-8 hours
- ☑ 8-24 hours
- ☐ 1-3 days
- ☐ 3-14 days
- ☐
  > 14 days
- ☐ Not feasible to run on a desktop machine, as described below.

Details

The codes were last run on an **Intel(R) Xeon(R) CPU E5-2643 v2 @ 3.50GHz, 24 cores with 128 GB RAM, Ubuntu 22.04.1 LTS**. Computation took 12 hours. Results of Table II may slightly differ on different systems; see the section "Controlled Randomness" for further details.

Description of programs/code

All commands are the bash scripts.

- Install the python requirements. Run

```
pip install -r requirements.txt
```

- Install the R requirements. Run

```
Rscript --save check_dependency.R
```

- `shadish.R` preprocesses the Shadish et al. data. Output file is saved as `data/shadish.csv`. Run

```
Rscript --save shadish.R
```

- `gan_estimation/gan_shadish_baseline.py` generates simulated Shadish et al. data. Simulated Shadish et al. data is saved as `data/shadish_generated.feather`. Related figures are also saved in the folder `data`. Run

```
python3 gan_estimation/gan_shadish_baseline.py
```

- `comparison.R` and `comparison_shadish.R` compare the estimators. Results are saved as `result/out.feather` and `result/out_shadish.feather`. Run

```
Rscript --save comparison.R
Rscript --save comparison_shadish.R
```

- `se.R` and `se_shadish.R` approximate the semiparametric efficiency lower bound. Run

```
Rscript --save se.R
Rscript --save se_shadish.R
```

- `produce_tables.R` and `produce_tables_shadish.R` produce the tables in the manuscript. Results in the main tables are saved as `result/table.txt` and `result/table_shadish.txt` for different M and `result/table_alpha.txt` and `result/table_shadish_alpha.txt` for different α. Results in the parenthesis are saved as `result/tablese.txt` and `result/tablese_shadish.txt` for different M and `result/tablese_alpha.txt` and `result/tablese_shadish_alpha.txt` for different α. Run

```
Rscript --save produce_tables.R
Rscript --save produce_tables_shadish.R
```

(Optional, but recommended) License for Code

The code is licensed under a MIT license. See LICENSE.txt for details.

Instructions to Replicators

- Delete the feather files in `result` folder.
- Run `main.sh` on Linux shell.

List of tables and programs

The provided code reproduces:

- ☑ All numbers provided in text in the paper
- ☑ All tables and figures in the paper
- □ Selected tables and figures in the paper, as explained and justified below.

Figure/Table #	Program	Line Number	Output file	Note
Table 1 (M)	produce_tables.R		table.txt	
Table 1 (M, parentheses)	produce_tables.R		tablese.txt	
Table 1 (α)	produce_tables.R		table_alpha.txt	
Table 1 (α, parentheses)	produce_tables.R		tablese_alpha.txt	
Table 2 (M)	produce_tables_shadish.R		table_shadish.txt	
Table 2 (M, parentheses)	produce_tables_shadish.R		tablese_shadish.txt	
Table 2 (α)	produce_tables_shadish.R		table_shadish_alpha.txt	
Table 2 (α, parentheses)	produce_tables_shadish.R		tablese_shadish_alpha.txt	

References

- Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2021). Using wasserstein generative adversarial networks for the design of monte carlo simulations. Journal of Econometrics.
- Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2021). Simulated LaLonde data Datafile Version: https://drive.google.com/drive/folders/1CCU1zuibrPOZHK6NAAAVrX5TKQZZOMTF.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. Journal of the American Statistical Association, 103(484), 1334-1343.

Acknowledgements

The README used the template from https://social-science-data-editors.github.io/template_README/.