

README for “Adapting to Misspecification”

Timothy B. Armstrong, Patrick Kline and Liyang Sun
MS# 21991; RP# 206 for *Econometrica*, August 10, 2025

Overview

The code in this replication package constructs the adaptive estimator and applies it to the re-analysis of [Angrist and Krueger, 1991], Dehejia and Wahba [1999] and [de Chaisemartin and D’Haultfoeuille, 2020] using MATLAB, R and Stata. Master scripts are provided to run all parts of the workflow: data cleaning, adaptive estimator, and generation of all six tables and seven figures in the main text and online appendix. The replicator should expect the full code to run in approximately 6 hours, half on a standard laptop, and half on a server.

Data Availability and Provenance Statements

We are secondary data users; we did not generate the data. Rather, we accessed the data from the replication package of Angrist and Krueger [1991] from the Angrist Data Archive at <https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive> replication package of Dehejia and Wahba [1999] from the Mostly Harmless Econometrics Archive at <https://economics.mit.edu/people/faculty/josh-angrist/mhe-data-archive>, and replication package of de Chaisemartin and D’Haultfoeuille [2020] from the ICPSR at <https://doi.org/10.3886/E118363V2>. We have also included the data as well as their non-proprietary versions in our replication package.

Statement about Rights

We certify that we have legitimate access to and permission to use the data used in this manuscript.

Summary of Availability

All data **are** publicly available.

Details on each Data Source

Data Name	Data Files	Location	Provided	Citation
Gentzkow et al. (2011)	voting_cnty_clean.dta; voting_cnty_clean.csv	data/Gentzkow et al (2011)/	TRUE	De Chaisemartin & D’Haultfoeuille (2020)

Data Name	Data Files	Location	Provided	Citation
Angrist and Krueger (1991)	NEW7080.dta; NEW7080.csv; cohort3039.csv	data/Angrist and Krueger (1991)/	TRUE	Angrist & Krueger (1991)
Dehejia and Wahba (1999)	nswre74.dta; cps1re74.dta; nswre74.csv; cps1re74.csv	data/Dehejia and Wahba (1999)/	TRUE	Dehejia & Wahba (1999)

All data are obtained from existing replication repositories detailed below, with dictionaries available.

Dataset list

1. `data/Gentzkow et al (2011)/voting_cnty_clean.dta`: Can be downloaded from the ICPSR repository of de Chaisemartin and D’Haultfoeuille [2020] at <https://doi.org/10.3886/E118363V2>. Also provided in this replication package in nonproprietary format as `voting_cnty_clean.csv`. Processed by `Gentzkow_et_al_rho.do` to bootstrap the variance covariance matrix. Serves as input for the figures and tables in Section 2 and Section 5.1 (e.g., Figure 1-3, 5, Tables 1–2).
2. `data/Angrist and Krueger (1991)/NEW7080.dta`: Can be downloaded from the Angrist Data Archive of Angrist and Krueger [1991] at <https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive> (compressed as `NEW7080.rar`). Also provided in this repository in nonproprietary format as `NEW7080.csv`. Cleaned and reformatted by `QOB Table V export.do`, which outputs `cohort3039.csv`. Serves as input for Section 5.2 only and generates Table 3.
3. `data/Dehejia and Wahba (1999)/nswre74.dta` and `cps1re74.dta`: Can be downloaded from the replication archive of Dehejia and Wahba [1999]. Nonproprietary versions are included as `nswre74.csv` and `cps1re74.csv`. Processed by `GetYs3.do`, which computes the simple difference in means and its bootstrap variance–covariance matrix (`Lalonde.log`). Serves as input for Online Appendix E only and generates Tables A1–A3.

Computational requirements

Software Requirements

For the part of analyses on a MacOS laptop

We use Matlab (R2021a). The optimization toolkit `fmincon` (with `sqp` option) and `fminimax` are required.

We use Stata (version 14). The programs `fuzzydid` (distribution date: 20190507) and `moremata` (distribution date: 20170531) are required and can be installed via `ssc install fuzzydid` and `ssc install moremata`. Only the program `fuzzydid` is used, but `fuzzydid` depends on `moremata`. The versions of these two programs do not affect the replication, and a stable version of `fuzzydid`, obtained from the replication repository of de Chaisemartin and D’Haultfoeuille [2020] detailed below in Step 2, is provided in the replication package and used in the replication.

We use R (4.1.1) and the following libraries

- `latex2exp` (0.9.4)
- `tidyverse` (1.3.1)
- `xtable` (1.8-4)
- `pracma` (2.4.2)
- `akima` (0.6-3.4)
- `R.matlab` (3.7.0)
- `ggplot2` (3.5.1)
- `dplyr` (1.1.4)

For the part of analyses on a Linux server

We use Matlab (R2022a) and the optimization toolkit `CVX` is used for the multivariate adaption in Online Appendix E. Note that `CVX` needs to be installed separately by following instructions on <https://cvxr.com/cvx/doc/install.html>. We only use the free solvers `SeDuMi` and `SDPT3` as shown below

```
>> cvx_version
```

```
-----
CVX: Software for Disciplined Convex Programming      (c)2014 CVX Research
Version 2.2, Build 1148 (62bfcca)                    Tue Jan 28 00:51:35 2020
-----
```

```
Installation info:
```

```
Path: /home/lsun20/cvx
MATLAB version: 9.12 (R2022a)
OS: Linux amd64 version 4.18.0-348.el8.x86_64
Java version: 1.8.0_202
```

```
Verfying CVX directory contents:
```

```
WARNING: The following files/directories are missing:
/home/lsun20/cvx/sedumi/.travis.yml
```

```
These omissions may prevent CVX from operating properly.
```

```
Preferences:
```

```
Path: /home/lsun20/.matlab/cvx_prefs.mat
```

```
License host:
```

```

Username: lsun20
Host ID: 84160c133824 (eno1np0,192.168.1.62)
Installed license:
    No license installed.

```

```

>> cvx_solver

```

Name	Status	Version	Location
SDPT3	selected,default	4.0	{cvx}/sdpt3
SeDuMi		1.3.4	{cvx}/sedumi

Controlled Randomness

Random seed is set at lines 29, 33 of program `code/Matlab/master.m`, lines 96, 138, 159 and 201 of program `code/fig1235tab12/master.R`, lines 199 of program `code/tab3/master.m`, lines 126 of program `code/OAtab1/master.m` and lines 73, 144, and 226 of program `code/OAtab23/master.m`.

Memory and Runtime Requirements

Summary Approximate time needed to reproduce the part of analyses on a server described below: 3 hours.

Approximate time needed to reproduce the part of analyses on a laptop described below: 3 hour.

Approximate storage space needed: 250 MB

Details The part of analyses on a server was last run on a **24-core Linux-based server with Red Hat Enterprise Linux 8.5**.

The part of analyses on a laptop was last run on a **6-core Intel-based 2021 laptop with MacOS version 11.7.3**.

Description of programs

All key intermediate outputs that form the adaptive estimators are saved in `code/Matlab/sim_results/` for reuse across different parts of the manuscript. Figures and tables are saved with intuitive filenames (`tables/table3.tex`, `figures/figure4.png`, etc.) to match the manuscript.

- Programs in `code/Matlab/` construct the adaptive estimators which support our main theoretical results. The file `code/Matlab/master.m` runs them all in sequence as explained in instructions. To speed up the optimization, an initial guess is provided in `sim_results/init_priors.mat`.
- Programs in `data/Gentzkow et al (2011)/` and `code/fig1235tab12/` generate all tables and figures for Section 2 and Section 5.1 of

the main paper. The file `code/fig1235tab12/master.R` reproduces these using estimates precomputed by `data/Gentzkow et al (2011)/Gentzkow_et_al_rho.do` as logged in `Gentzkow_et_al_rho.log`.

- Programs in `code/fig40Afig12/` generate Figure 4 in Section 4.3 and Figure A1 and A2 in Online Appendix C. The file `code/fig40Afig12/master.R` runs the complete workflow.
- Programs in `code/sec44/` replicate the numerical results for Section 4.4 on constrained adaptation. The file `code/sec44/master.m` runs the computation and outputs the result to `figures/sec44.png`.
- Programs in `data/Angrist and Krueger (1991)/` and `code/tab3/` generate Table 3 for Section 5.2. First, `data/Angrist and Krueger (1991)/QOB Table V export.do` cleans and reformats the census data, producing `cohort3039.csv`. Then, `code/tab3/master.m` reads this cleaned dataset and produces `tables/table3.tex`.
- Programs in `data/Dehejia and Wahba (1999)/`, `code/OAtab1/` and `code/OAtab23/` compute the results in Online Appendix E and F. The file `code/OAtab1/master.m` and `code/OAtab23/master.m` reproduce these using estimates precomputed by `data/Dehejia and Wahba (1999)/GetYs3.do` as logged in `data/Dehejia and Wahba (1999)/Lalonde.log`.

Instructions to Replicators

1. Preliminary steps

Begin by navigating to `code/Matlab/`, and create a subdirectory named `sim_results/` to store the output from this step; this output will be used in later replication steps. To speed up optimization, an initial guess is provided in `sim_results/init_priors.mat`. Set `code/Matlab/` as the home directory in Matlab, and run `master.m`, which performs all the computations described below. For completeness, each component of this script is explained in the following text. All directory paths mentioned below are relative to `code/Matlab/`.

1.1 Compute the B -minimax estimators and oracle risk

This part is implemented in `opt_minimax_bounded_normal_mean.m`, which takes 10 minutes on a laptop.

1. `opt_minimax_bounded_normal_mean.m` calls `pmf.m` and uses `fmincon` to solve the least favorable priors via the function `outer_loop_minimax.m`
2. This outputs `sim_results/minimax_rho_B9.csv`, which is the oracle risk $R_{\max}(B)$
3. This also outputs `sim_results/minimax_bounded_normal_mean.mat`, the B -minimax estimator, `sim_results/priors_bounded_normal_mean.mat`,

the least favorable prior, `sim_results/risk_bounded_normal_mean.mat`
the risk functions of the B -minimax estimator.

1.2 Compute the adaptive estimator

This part is implemented in `risk_functions_hausman.m`, which takes 60 minutes on a laptop.

1. Set `Sigma_U0_grid` to be the grid $\tanh((3 : 0.05 : 0.05))^2$ in `risk_functions_hausman.m`, which calls `risk_calc.m` and uses `fmincon` to solve the least favorable priors via the function `outer_loop_minimax.m`.
2. Outputs in `sim_results/` the adaptive prior, `priors.mat`, the adaptive estimator, `policy.mat`, and the risk of the adaptive estimator `risk.mat`

1.3 Compute the adaptive soft-threshold and hard-threshold

This part is implemented in `risk_functions_soft_minimax.m`, which takes 2 minutes on a laptop.

1. Set `Sigma_U0_grid` to be the grid $\tanh((3 : 0.05 : 0.05))^2$ in `risk_functions_soft_minimax.m` which uses `fminimax` to solve the adaptive soft threshold and hard threshold
2. Outputs in `sim_results/` the soft and hard thresholds, `thresholds.mat`, and the risk of the soft and hard thresholding estimator `risk_thresholds.mat`

1.4 Compute the adaptive ERM estimator

This part is implemented in `risk_calc_empirical_MSE.m` which calls `fminimax` and exports the optimal thresholds to `sim_results/emse_corr.mat`. This takes 30 min. on a laptop.

1.5 Compute critical values for FLCI

This part is implemented in `FLCI.m`, which uses simulation to determine the critical values for varying B for the adaptive estimator and for the soft thresholding estimator. This takes 60 min. on a laptop. The outputs are stored in:

- Critical values for the adaptive estimator: `sim_results/flci_adaptive_cv.mat`
- Critical values for the soft-thresholding estimator: `sim_results/flci_adaptive_st_cv.mat`

2. Replicate results in Section 2 and Section 5.1

This section is based on data from Gentzkow et al. [2011], replicated following the replication repository of de Chaisemartin and D'Haultfœuille [2020].

1. First download the replication repository of de Chaisemartin and D'Haultfœuille [2020] from the ICPSR at <https://doi.org/10.3886/>

E118363V2. Copy the files `voting_cnty_clean.dta` to the directory `data/Gentzkow et al (2011)/`. The non-proprietary version of `voting_cnty_clean.dta` is provided in `csv` as `voting_cnty_clean.csv`.

2. Create a subdirectory `data/Gentzkow et al (2011)/ado/`. From the replication repository of de Chaisemartin and D'Haultfœuille [2020], copy the following files

```
fuzzydid.ado
fuzzydid.sthlp
build_cond_expectation_1.ado
build_cond_expectation_2.ado
estim_wrapper.ado
five_fold_cv.ado
legendrisation.ado
lqte_estim_nox.ado
late_estim_nox.ado
late_estim_x.ado
special_cases.ado
createyvalues.mo
my_sort.mo
qq_transfo.mo
```

to the directory `data/Gentzkow et al (2011)/ado/`. This is a stable version of the `fuzzydid` that will be used in the next step.

3. Set `data/Gentzkow et al (2011)/` as home directory. Run `Gentzkow_et_al_rho.do`, which slightly modifies the original replication file `Gentzkow_et_al.do` of de Chaisemartin and D'Haultfœuille [2020] to replicate point estimates and calculate the variance covariance matrix using 100 bootstrapped samples. This takes about one hour and half on a laptop. To save time, the 100 bootstrapped estimates that are used to compute the variance covariance matrix are saved in `did_m_fd_bs.csv`. The outputs are saved in `Gentzkow_et_al_rho.log`. To ensure the enclosed version of `fuzzydid` is used and to prevent overwriting any existing `fuzzydid` and `moremata` installation, `Gentzkow_et_al.do` temporarily adds `data/Gentzkow et al (2011)/ado/` to the top of `adopath`, so that `data/Gentzkow et al (2011)/ado/fuzzydid` is used. At the end of `Gentzkow_et_al.do`, `data/Gentzkow et al (2011)/ado/` is removed from `adopath`, restoring the user's original `adopath` configuration after execution.
4. To replicate the results in Section 2, Figure 5 and Section 5.1, set `code/fig1235tab12/` to be the home directory. Run `master.R`. The point estimates, standard errors and correlation coefficient in the beginning of `master.R` are directly copied from `Gentzkow_et_al_rho.log`. This should take less than 5 minutes on a laptop.

3. Replicate results in Section 4.3 and Online Appendix C

Run `fig40Afig12/master.R`, which outputs Figure 4 and surrounding text explanation, Online Appendix Figures 1 and 2. This should take less than 5 minutes on a laptop.

4. Replicate results in Section 4.4

The last paragraph of Section 4.4 discusses how constraining the worst case risk still achieves the property that the maximal risk decreases relative to the risk of the unbiased estimator is larger than the maximal risk increase relative to the unbiased estimator. To replicate this numerical result, set `code/sec44/` to be the home directory. Run `master.m`. The result can be read off from the output, which is in `figures/sec44.png/`. This should take less than 5 minutes on a laptop.

5. Replicate results in Section 5.2

This section is based on the 1970 and 1980 census extract from Angrist and Krueger [1991], which was downloaded from the Angrist Data Archive at <https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive>.

1. After extrating the compressed file `NEW7080.rar`, save `NEW7080.dta` to `data/Angrist and Krueger (1991)/` and set it as home directory. A non-proprietary version of `NEW7080.dta` is provided as `NEW7080.csv`.
2. To obtain the original IV and OLS estimates and their variance covariance matrix, We adapt replication code for Table V in Angrist and Krueger (1991), which is provided in Angrist Data Archive as well. The modified script `QOB Table V export.do` cleans the raw data and save the cleaned data `cohort3039.csv` as part of this replication package. This cleaned dataset is then imported into MATLAB to reproduce the original IV and OLS estimates. This script `QOB Table V export.do` outputs a log file `AK91 replication.log`. This should take about one minute on a laptop.
3. Set `code/tab3/` as home directory. Run `master.m`, which reads `data/Angrist and Krueger (1991)/cohort3039.csv` and outputs Table 3 to `tables/table3.tex/`. This should take less 5 minutes on a laptop. Note that the results for the unconstrained optimal adaptive estimator are based on solution to an optimization problem solved using Matlab's `fmincon`. Due to the numerical nature of the optimizer, small differences in the optimization path arising from different Matlab versions and computing environments can lead to slight variation in the final result. In this case, the difference can affect the “Max Risk” and “Max Regret” of the unconstrained optimal adaptive estimator at the second decimal place, but does not affect the arguments in the surrounding text of the paper.

Online Appendix E - Pooling controls

This section is based on the NSW experimental data and CPS data analysed in Dehejia and Wahba [1999]. Set the folder `Dehejia and Wahba (1999)` as the home path. Download `nswre74.dta` and `cps1re74.dta` from the replication repository of Dehejia and Wahba [1999]. Non-proprietary versions of both are provided in csv format.

Preliminaries Run program `GetYs3.do` to obtain the original Dehejia-Wahba analysis of the NSW estimate: simple difference in means, Y_{R1} , Y_{R2} and their variance covariance matrix, using 1,000 bootstraps. The runtime is about five minutes on a laptop. The variance covariance matrix and estimates are recorded in the log file `Lalonde.log` and copied as input to the following `master.m` analysis files.

6. Replicate Table A1

To compute the trivariate adaptation as in Table A1 and surrounding text, replicate in the following steps. This takes about 2 hours on a server.

1. Navigate to `code/OAtab1/` and set it as the home directory and create a `sim_results/` subdirectory.
2. Run `master.m` which uses `pmf2.m` that is saved in the same directory. It uses CVX to calculate the least favorable prior. Refer to **Software Requirements** for installation instructions for CVX. The optimization takes approximately two hours on a server, so the results are saved to `sim_results/dim2_adaptive_mu_minmax_nsw_diff_means_B9b_095_y_025_log_cvx.csv` to enable faster execution in the subsequent step that generates the output.
3. All results are outputted to `tables/tableA1.tex` including the p-values in the text below Table A1.

7. Replicate Table A2 and A3

To compute the bivariate adaptation as in Table A2 and the composite adaptation as in Table A3, replicate in the following steps. This takes less than one minute on a laptop.

1. Navigate to `code/OAtab23/` and set it as the home directory. Run `master.m` which calls `../Matlab/adaptive_estimate.m` and computes the adaptive estimates.
2. All results are outputted to `tables/tableA2a.tex`, `tables/tableA2b.tex` and `tables/tableA3.tex`

List of tables and programs

The provided code reproduces all original tables and figures in the paper.

Figures 1, 2, 3 and 5, Tables 1 and 2. Program: `/code/fig1235tab12/master.R`.
Output file: `figures/figure1.png`, `figures/figure2.png`, `figures/figure3.png`,
`figures/figure5.png` `tables/table1.tex`, `tables/table2a.tex`, `tables/table2b.tex`.

Figure 4, log approximation to Figure 4, Online Appendix Figures 1 and 2.
Program: `/code/fig40Afig12/master.R`. Output files: `figures/figure4.png`,
`figures/figure4text.txt`, `figures/figureA1.png`, `figures/figureA2.png`

Section 4.4, Last paragraph, Program: `/code/sec44/master.m`. Output file:
`figures/sec44.png`.

Table 3. Program: `/code/tab3/master.m`. Output file: `tables/table3.tex`.

Online Appendix Table 1. Program: `/code/0Atab1/master.m`. Output file:
`tables/tableA1.tex`.

Online Appendix Tables 2 and 3. Program: `/code/0Atab23/master.m`. Output
file: `tables/tableA2a.tex`, `tables/tableA2b.tex` and `tables/tableA2.tex`.

References

- J.D. Angrist and A.B. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106, 1991. Data deposited at <https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive>.
- Clément de Chaisemartin and Xavier D’Haultfoeuille. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996, September 2020. ISSN 0002-8282. Data deposited at <https://doi.org/10.3886/E118363V2>.
- Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999. Data deposited at <https://economics.mit.edu/people/faculty/josh-angrist/mhe-data-archive>.
- Matthew Gentzkow, Jesse M. Shapiro, and Michael Sinkinson. The Effect of Newspaper Entry and Exit on Electoral Politics. *American Economic Review*, 101(7):2980–3018, December 2011. ISSN 0002-8282. doi: 10.1257/aer.101.7.2980. URL <https://www.aeaweb.org/articles?id=10.1257/aer.101.7.2980>.