

Replication package README

Contents

1. Overview	1
2. Package contents	1
2.1 Folder structure	2
2.2 Key scripts	2
2.3 Data files included in the package	2
3. Data Availability Statement (DAS)	2
3.1 Included data	2
3.2 Provenance and how to obtain the original data	3
3.3 Excluded data	3
4. Computational requirements	3
4.1 Software	3
4.2 R packages	3
4.3 Hardware and OS used for benchmark run	4
4.4 Observed run time (full replication)	4
5. Instructions to run the replication	4
5.1 Quick start (full replication)	4
5.2 First use: 00_* and 99_* scripts	5
5.3 Running components separately	5
5.4 Environment variables (optional)	5
5.5 Controlled randomness (seeds)	5
6. Mapping outputs to tables and figures	6
6.1 Simulation outputs (raw)	6
6.2 Paper tables	6
6.3 Paper figures (from simulations)	7
6.4 Census diagnostic figures	7
6.5 Opportunity Insights (OI) diagnostic figure	7
6.6 Synthetic spectrum diagnostic figure	7
7. Logs and reproducibility records	7
8. References and data citations	8

1. Overview

This replication package reproduces the main simulation results and figures for the associated paper.

The package is organized so that it is **self-contained and portable**: all scripts use relative paths (no `setwd()`), and the universal setup script installs missing R packages, creates output folders, and records the computing environment (R version and package versions).

The main entry point is:

- `scripts/99_run_all.R` — runs the full replication in the intended order.

2. Package contents

2.1 Folder structure

```
project_root/  
  scripts/      # replication scripts (run these)  
  src/          # main code functions used by scripts  
  data/         # input data used by census/OI diagnostics  
  results/      # created by scripts; raw and intermediate outputs (CSV)  
  figures/      # created by scripts; main figures (PNG)  
  figures_alt/  # created by scripts; appendix/alternative figures (PNG)  
  tables/       # created by scripts; paper tables (CSV)  
  info/         # created by scripts; environment logs (R + packages)
```

2.2 Key scripts

- `scripts/00_setup_simulations.R`
Universal setup: installs/loads packages, creates `results/`, `figures/`, `figures_alt/`, `tables/`, `info/`, sets RNG defaults, and writes environment logs to `info/`.
- `scripts/01_run_simulations.R`
Runs Monte Carlo simulation experiments and writes raw results to `results/`.
- `scripts/02_postprocess_simulations.R`
Reads the raw simulation outputs from `results/` and produces “paper-ready” tables in `tables/` and histogram figures in `figures/`.
- `scripts/03_run_census.R`
Uses the census dataset to produce diagnostics and extension figures (including SVD/scree and semi-synthetic noise experiments). Writes outputs to `results/`, `figures/`, and `figures_alt/`.
- `scripts/04_run_oi_svd.R`
Uses Opportunity Insights (OI) data to produce the OI SVD/scree diagnostic figure.
- `scripts/05_run_svd_synthetic.R`
Produces a synthetic SVD spectrum diagnostic figure from a simulated low-rank matrix with Laplace noise.
- `scripts/99_run_all.R`
Main runner that calls scripts (01)–(05) in order.

2.3 Data files included in the package

This package **includes** the required input datasets inside `data/`.

- Census / China Syndrome replication dataset:
 - `data/workfile_china.dta`
- Opportunity Insights dataset:
 - `data/online_table_3.dta`

No confidential or proprietary datasets are used.

3. Data Availability Statement (DAS)

3.1 Included data

All data needed to run the replication are included in this package under `data/` (see Section 2.3).

3.2 Provenance and how to obtain the original data

Even though the data are included here, the original sources are public and can be obtained independently as follows.

(A) Census / “China Syndrome” replication data

- File used in this package: `workfile_china.dta`
- Source: Replication package for:
 - Autor, D. H., Dorn, D., and Hanson, G. H. (2013b). Replication data for ‘The China syndrome: Local labor market effects of import competition in the United States’. DOI:10.3886/E112670V1.
- Download location (openICPSR):
<https://www.openicpsr.org/openicpsr/project/112670/version/V1/view>
- Access conditions: publicly available; may require creating a (free) openICPSR account depending on current platform requirements. “Copyright 2013 American Economic Association. Licensed under CC-BY 4.0.”

(B) Opportunity Insights (OI) data

- File used in this package: “Online Data Table 3: Complete CZ-level Dataset: Causal Effects and Co-variates”.
- Source paper:
 - Chetty, R. and Hendren, N. (2018). Replication code for ‘The impacts of neighborhoods on inter-generational mobility I: Childhood exposure effects’. DOI:10.7910/DVN/CEMFTJ.
- Download location (Opportunity Insights data library):
https://opportunityinsights.org/data/?geographic_level=0&topic=0&paper_id=606#resource-listing
- Mirror / archived distribution (Harvard Dataverse, includes CSV/DTA + dictionary):
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EI4WE2>
- Access conditions: publicly available; no special permissions required.

3.3 Excluded data

No additional datasets are required beyond those included in `data/`.

4. Computational requirements

4.1 Software

- Benchmark run environment:
 - R version: **R 4.4.2 (2024-10-31)**
 - Platform: `x86_64-pc-linux-gnu`
 - Running under: Ubuntu 24.04.1 LTS
- The setup script records the exact environment used in:
 - `info/R_version.txt`
 - `info/sessionInfo.txt`
 - `info/package_versions_setup.csv`

4.2 R packages

The universal setup script (`scripts/00_setup_simulations.R`) installs missing packages automatically from CRAN (internet required on first run).

The authors’ package versions (as recorded in `info/package_versions_setup.csv`) are:

package	version
DescTools	0.99.60
dplyr	1.1.4
ExtDist	0.7.4
foreign	0.8.88
ggbiplot	0.6.2
ggplot2	4.0.1
gtools	3.9.5
haven	2.5.4
ivreg	0.6.5
lmtest	0.9.40
Matrix	1.7.2
mvtnorm	1.3.3
quantreg	6.1
Rlab	4
sandwich	3.1.1
tidyr	1.3.1

Note: ggbiplot can be difficult to install on some systems (it is not always available on CRAN). In this package it is treated as optional; it is not required for the core replication outputs listed in Section 6.

4.3 Hardware and OS used for benchmark run

- OS: Linux 4.18.0-553.44.1.el8_10.x86_64 (x86_64)
- CPU cores: 12 logical / 12 physical
- RAM (total): 24.0 GB

4.4 Observed run time (full replication)

The following times were recorded from scripts/99_run_all.R:

step	script	elapsed_sec	elapsed_min
Run simulations	01_run_simulations.R	2686.168	44.7694667
Post-process simulations	02_postprocess_simulations.R	2383.297	39.7216167
Run census diagnostics	03_run_census.R	13.222	0.2203667
Run OI SVD diagnostics	04_run_oi_svd.R	2.031	0.0338500
Run synthetic SVD diagnostics	05_run_svd_synthetic.R	41.308	0.6884667

- Total elapsed: **5126.04 sec (85.43 min)**.
- 01_run_simulations.R is implemented as a serialized workflow (no explicit parallel backend in the script).

5. Instructions to run the replication

5.1 Quick start (full replication)

1. Unzip/copy the replication package to a local directory.
2. Open a terminal in the **project root** folder.

3. Run:

```
Rscript scripts/99_run_all.R
```

On first run, R will install missing packages from CRAN automatically.

5.2 First use: 00_* and 99_* scripts

- `scripts/99_run_all.R` is the top-level “full replication” runner. It first sources `scripts/00_setup_simulations.R`, then calls `scripts/01_run_simulations.R` through `scripts/05_run_svd_synthetic.R` in order.
- Each component script (01 to 05) also sources `scripts/00_setup_simulations.R` internally, so stand-alone runs still perform setup automatically.
- On first use, package installation and environment logging happen through `00_setup` during normal execution.
- Users normally do **not** need to run `00_setup` manually.

5.3 Running components separately

You can run each component script directly, e.g.:

```
Rscript scripts/01_run_simulations.R
Rscript scripts/02_postprocess_simulations.R
Rscript scripts/03_run_census.R
Rscript scripts/04_run_oi_svd.R
Rscript scripts/05_run_svd_synthetic.R
```

5.4 Environment variables (optional)

Some scripts allow overriding defaults without editing code:

- Census data path:
 - `CENSUS_DTA` — full path to `workfile_china.dta`
- OI data options:
 - `OI_DTA` — full path to `online_table_<k>.dta`
 - `OI_TABLE` — which online table to use (default: 3)
 - `OI_FILL_NA` — NA handling option passed to `extract_X()` (default: 0)
- Synthetic spectrum options:
 - `SPEC_N`, `SPEC_R`, `SPEC_SEED`, `SPEC_VARIANCES`

Example (Mac/Linux):

```
CENSUS_DTA="/path/to/workfile_china.dta" Rscript scripts/03_run_census.R
```

Example (Windows PowerShell):

```
$env:CENSUS_DTA="C:\path\to\workfile_china.dta"
Rscript scripts\03_run_census.R
```

5.5 Controlled randomness (seeds)

The package sets seeds in several places. To manipulate controlled randomness, edit the seed values below (or set `SPEC_SEED`) and re-run the relevant script(s).

- Global RNG defaults in `scripts/00_setup_simulations.R`:
 - `RNGkind(kind = "Mersenne-Twister", normal.kind = "Inversion", sample.kind = "Rejection")` at line 79 (or legacy fallback line 81 on older R versions)
 - `SEED_GLOBAL <- 1` at line 83
 - `set.seed(SEED_GLOBAL)` at line 84

- This setup script is sourced by `scripts/01_run_simulations.R` through `scripts/05_run_svd_synthetic.R`.
- Simulation seeds in `scripts/01_run_simulations.R`:
 - Long-table blocks:
 - * `SEED_LONG_NOISY <- 1` at line 166
 - * `SEED_LONG_DISCRETE <- 1` at line 167
 - * `SEED_LONG_PRIVATE <- 1` at line 168
 - Short/histogram blocks:
 - * `SEED_SHORT_NOISY <- 100` at line 170
 - * `SEED_SHORT_DISCRETE <- 10000` at line 171
 - * `SEED_SHORT_PRIVATE <- 100000` at line 172
- Census seeds in `scripts/03_run_census.R`:
 - `set.seed(1)` at line 252
 - `seed <- 1` at line 263
 - Additional seed applications in calibration paths: `set.seed(seed)` at lines 274, 360, 363, 394, and 397.
- Synthetic SVD seed in `scripts/05_run_svd_synthetic.R`:
 - `seed <- as.integer(Sys.getenv("SPEC_SEED", unset = "1"))` at line 87
 - `set.seed(seed)` at line 114
 - Override example (Mac/Linux):

```
SPEC_SEED=123 Rscript scripts/05_run_svd_synthetic.R
```

- Override example (Windows PowerShell):

```
$env:SPEC_SEED="123"
Rscript scripts\05_run_svd_synthetic.R
```

How to change randomness:

1. Edit seed constants in `scripts/01_run_simulations.R` and/or `scripts/03_run_census.R`, or set `SPEC_SEED` for `scripts/05_run_svd_synthetic.R`.
2. Re-run the affected script(s), or run `Rscript scripts/99_run_all.R` for a full refreshed replication.

6. Mapping outputs to tables and figures

This section lists where outputs are written and how they map to the paper exhibits.

Workflow summary:

```
scripts/ ---> results/ ---> figures/ + figures_alt/
                        \
                        ---> tables/
```

6.1 Simulation outputs (raw)

Produced by `scripts/01_run_simulations.R` in `results/`:

- `results/noisy.csv`
- `results/discrete.csv`
- `results/private.csv`
- `results/compare_noise.csv`, `results/compare_noise_ate.csv`, `results/compare_noise_se.csv`
- `results/compare_discrete.csv`, `results/compare_discrete_ate.csv`, `results/compare_discrete_se.csv`
- `results/compare_private.csv`, `results/compare_private_ate.csv`, `results/compare_private_se.csv`

6.2 Paper tables

Produced by `scripts/02_postprocess_simulations.R` in `tables/`:

- tables/noisy_tuning.csv
- tables/noisy_levels.csv
- tables/discrete_tuning.csv
- tables/private_tuning.csv
- tables/private_levels.csv

These CSVs correspond to the tuning/levels panels used to construct **Table 2** in the paper.

6.3 Paper figures (from simulations)

Produced by scripts/02_postprocess_simulations.R in figures/:

- figures/noisy_PCR_hist2.png
- figures/discrete_PCR_hist2.png
- figures/private_PCR_hist2.png

These figures correspond to the simulation histograms in **Figure 2** (panels for measurement error, discretization, and differential privacy).

6.4 Census diagnostic figures

Produced by scripts/03_run_census.R:

Main-text outputs in figures/ (unscaled pipeline):

- figures/scree_real2.png (*SVD/scree diagnostic; Figure 1a*)
- figures/noisy_CI_semi2.png (*Estimation effects; Figure 3a*)
- figures/discrete_CI_semi2.png (*Estimation effects; Figure 3b*)
- figures/private_calibration2.png (*Estimation effects; Figure 3c*)

Appendix/alternative outputs in figures_alt/ (scaled pipeline):

- figures_alt/noisy_CI_semi2.png (*Figure I.1a*)
- figures_alt/discrete_CI_semi2.png (*Figure I.1b*)
- figures_alt/private_calibration2.png (*Figure I.1c*)

Figures are written to figures/ (main text) and figures_alt/ (appendix). Tables are written to tables/. These final outputs are produced from intermediate CSV files in results/.

6.5 Opportunity Insights (OI) diagnostic figure

Produced by scripts/04_run_oi_svd.R in figures/:

- figures/scree_oi_<OI_TABLE>_<OI_FILL_NA>.png
Default: figures/scree_oi_3_0.png
(*Figure 1b*)

6.6 Synthetic spectrum diagnostic figure

Produced by scripts/05_run_svd_synthetic.R in figures/:

- figures/spectrum.png (*Figure 1c*)

7. Logs and reproducibility records

Each time the setup script runs, it writes/overwrites environment records in info/:

- info/sessionInfo.txt
- info/R_version.txt
- info/package_versions_setup.csv

These files document the exact software environment used to run the replication.

8. References and data citations

Autor, D. H., Dorn, D., and Hanson, G. H. (2013b). Replication data for ‘The China syndrome: Local labor market effects of import competition in the United States’. DOI:10.3886/E112670V1.

Chetty, R. and Hendren, N. (2018). Replication code for ‘The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects’. DOI:10.7910/DVN/CEMFTJ.

openICPSR (Replication package). “The China Syndrome: Local Labor Market Effects of Import Competition in the United States” (Project 112670, Version V1). <https://www.openicpsr.org/openicpsr/project/112670/version/V1/v>

Opportunity Insights Data Library (Replication data). “Online Data Table 3: Complete CZ-level Dataset: Causal Effects and Covariates.” https://opportunityinsights.org/data/?geographic_level=0&topic=0&paper_id=606#resource_listing

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (Opportunity Insights dataset archive). Harvard Dataverse, DOI:10.7910/DVN/EI4WE2. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EI4WE2>