Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ 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).

✅ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The external dataset (i.e., the data from the Work, Family, and Health Study) used in the data application section of our manuscript is publicly available at https://www.icpsr.umich.edu/web/DSDR/studies/36158/summary.The Work, Family, and Health Study is a large-scale cluster randomized experiment for improving the well-being and work-family balance of employees.

The relevant variable information can be found in **Appendix E: More Details on the Real Data Application** of the supplementary materials. More detailed information can be located directly within the codebook in the following sections: **SCWM_TIMEALL**: Page 42, **SCEM_DIST**: Page 75, **SCEM_STRS**: Page 79, **SCWM_FTWC**: Page 37, **RMZFN**: Page 9, **EMPLOYEE**: Page 4, **SCWM_CWH**: Page 15, **CONDITION**, Page 10.

Availability

✅ Data **are** publicly available
☐ Data **cannot be made** publicly available

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

✅ Data are available online at:
https://www.icpsr.umich.edu/web/DSDR/studies/36158/summary
☐ Data are available as part of the paper's supplementary material.
☐ Data are publicly available by request, following the process described here:

☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

Discussion of lack of publicly available data:

Description

File format(s)

✅ CSV or other plain text: tsv(tab-separated values) file
☐ Software-specific binary format (.Rda, Python pickle, etc.):
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (described here):

Data dictionary

✅ Provided by the authors in the following file(s): 36158-Codebook.pdf (downloaded from the publicly available website of the Work, Family, and Health Study)
☐ Data file(s) is (are) self-describiing (e.g., netCDF files)
☐ Available at the following URL:

Additional information (optional)

Part 2: Code

Abstract

The code/ directory includes the Python code for the real data application and simulation, each detailed in their README files. The data/ directory contains real data preprocessing scripts, described in its README file. The manuscript/ directory contains the figures for the manuscript. The output/ directory includes the results and computational outputs, with explanations in the README file.

Description

Code format(s)

☐ Script files
- ☐ R ✅ Python ☐ Matlab
- ☐ Other:

☐ Package
- ☐ R ✅ Python ☐ MATLAB toolbox
- ☐ Other:

☐ Reproducible report
- ☐ R Markdown ☐ Jupyter notebook
- ✅ Other: Markdown

☐ Shell script

☐ Other (described here)

Both the Python and R packages have been developed for the implementation of the methods proposed in our manuscript. For blinded review, we will provide the publicly available links to our Python and R packages after the peer review process.

Supporting software requirements

Version of primary software used

python-iArt 1.1.3

Libraries and dependencies used by the code

Python packages: numpy; pandas; scikit-learn; xgboost; statsmodels; lightgbm.

Supporting system/hardware requirements (optional)

Parallelization used

☐ No parallel code used

☐ Multi-core parallelization on a single machine/node

Number of cores used:

✅Multi-machine/multi-node parallelization

Number of nodes and cores used: 2000 nodes, 2000 cores

License

✅ MIT License (default)

☐ BSD

☐ GPL v3.0

☐ Creative Commons

☐ Other (described here):

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:
✅ Any numbers provided in text in the paper
✅ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))
✅ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified here:

Workflow details

Location

The workflow is available:
✅ As part of the paper's supplementary material
☐ In this Git repository:
☐ Other:

Format(s)

☐ Single master code file
☐ Wrapper (shell) script(s)
☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
✅ Text file (e.g., a readme-style file) that documents workflow
☐ Makefile
☐ Other (more detail in 'Instructions' below)

Instructions

All workflow instructions can be found in the corresponding folders in the code folder. Specifically, the real data application instructions are located in the application folder, and the simulation instructions are in the simulation folder, both of which are inside the code folder.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ <1 minute

☐ 1-10 minutes

☐ 10-60 minutes

☐ 1-8 hours

☐ >8 hours

✅ Not feasible to run on a desktop machine, as described here:

Simulation studies and real data applications in our manuscript involve extensive computation. Fortunately, the nature of our proposed methods and algorithms allows multi-node parallelization. The simulation studies in our manuscript (main text) run for over 8 hours on a 2000-node high-performance computing cluster. The real data application in our manuscript runs for over 8 hours on 1000 nodes.

Additional documentation (optional)

Notes (optional)