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 MPHIA survey is a new HIV-focused, cross-sectional, household-based, nationally representative survey of adults and adolescents aged 15 years and older as well as children aged 0-14 years. The survey contains the HIV test result and subsequent additional HIV related test results for its participants. The survey also asks demographic questions, such as age group, gender (preference to the gender collected in the adult or adolescent questionnaire), and ethnic group, and HIV related questions, such as access to preventive care and treatment services.

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:

☐ Data are available as part of the paper's supplementary material.

☒ Data are publicly available by request, following the process described here:

- First register at: PHIA Project Document Manager - Register (columbia.edu).
- Then request data at: PHIA Project Document Manager - Datasets (columbia.edu).
- After the request is permitted, data can be downloaded.

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

Non-publicly available data

Description

File format(s)

☒ CSV or other plain text: CSV
☒ Software-specific binary format (.Rda, Python pickle, etc.): .Rda
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (please specify):

Data dictionary

☐ Provided by authors in the following file(s):
☐ Data file(s) is(are) self-describing (e.g., netCDF files)

☒ Available at the following URL: PHIA Project Document Manager - Datasets (columbia.edu).

Additional Information (optional)

Part 2: Code

Abstract

The package MPHIACausal implements the proposed causal structural learning algorithms. It also contains some helper functions for doing analysis in the paper and the pre-processed MPHIA datasets. The README.Rmd provides instructions on how to use the package, and the ResultReproduction.Rmd verifies the causal structural learning results of the proposed algorithm on MPHIA data.

Description

Code format(s)

- ☒ Script files
  - ☒ R
  - ☐ Python
  - ☐ Matlab
  - ☐ Other:
- ☒ Package
  - ☒ R
  - ☐ Python
  - ☐ MATLAB toolbox
  - ☐ Other:
- ☒ Reproducible report
  - ☒ R Markdown
  - ☐ Jupyter notebook
  - ☐ Other:
- ☐ Shell script
- ☐ Other (please specify):

Supporting software requirements

Version of primary software used R version 3.5.2

Libraries and dependencies used by the code

- bnlearn version 4.5,
- dplyr version 1.0.5,
- magrittr version 1.5,
- igraph version 1.2.4.2,
- future version 1.21.0,
- future.apply version 1.7.0,
- MASS version 7.3-51.1,
- nnet version 7.3-12,
- ROCR version 1.0-7

Supporting system/hardware requirements (optional)

Parallelization used

- ☐ No parallel code used
- ☒ Multi-core parallelization on a single machine/node

  - Number of cores used: 10
- ☒ Multi-machine/multi-node parallelization
  - Number of nodes and cores used: 3 nodes, 243 cores

License

- ☐ MIT License (default)
- ☐ BSD
- ☒ GPL v3.0
- ☐ Creative Commons
- ☐ Other: (please specify below)

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 below:

The analysis results on the MPHIA dataset can be reproduced by the Rmd files. The simulation studies take very long time, and need to use high performance computing cluster, so the reproduction scripts for the simulation studies are not included. Instead, we provide the details on how to reproduce the simulation study results in the Rmd files.

Workflow

Location

The workflow is available:

- ☒ As part of the paper's supplementary material.
- ☐ In this Git repository:
- ☐ Other (please specify):

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

Download the R package MPHIACausal_0.1.0.tar.gz, the README.Rmd, and the ResultReproduction.Rmd to the same folder. See README.Rmd to learn how to use the package and reproduce the analysis results, and use ResultReproduction.Rmd to verify the analysis results on MPHIA data.

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:

Additional information (optional)

Notes (optional)