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 transNOAH breast cancer trial dataset (GSE50948), as described in Prat et al. (2014). This data consists of gene expression profiling measurements from 156 patients. For our analysis, we use expres sion measurements of the gene probe 204531 s at, associated with BRCA1, as the response variable Y .

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:

☐ 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.
☐ Software-specific binary format (.Rda, Python pickle, etc.): pkcle
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☒ Other (please specify): The dataset is saved as Rdata format.

Data dictionary

☒ Provided by authors in the following file(s): transNOAH.Rdata in the folder "Real Data Analysis"
☐ Data file(s) is(are) self-describing (e.g., netCDF files)
☒ Available at the following URL: Using the following command:
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("GEOquery")
Sys.setenv("VROOM_CONNECTION_SIZE" = 500000)
readr::local_edition(1)
transNOAH = GEOquery::getGEO("GSE50948")
save(transNOAH, file = "transNOAH.RData")

Additional Information (optional)

Part 2: Code

Abstract

All the codes of the simulation study and real data analysis are included in different folders based on their tasks. Details can refer to the illutration file in the folders.

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 4.2.1

Libraries and dependencies used by the code Biobase version 2.58.0; BiocGenerics version 0.440; data.table version 1.14.8; GEOquery version 2.66.0; glmnet version 4.1-7; ncvreg version 3.13.0; MASS version 7.3-58.1; SIHR version 2.0.1; stringr version 1.5.0; tidyr version 1.3.0.

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: 300 nodes and 300 cores

License

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

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:

Workflow

Location

The workflow is available:

- ☒ As part of the paper's supplementary material.
- ☒ In this Git repository: https://github.com/anonymousci42/HdimLinearInference
- ☐ 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

Step 1: load the required R packages.

Step 2: Run the "NPE1.R" script to generate the results of the proposed method without power enhancement term JPE. We only provide a script with fixed parameters, the whole tasks cannot be done in a single local computer.

Step 3: Run the "PPE1.R" and "SIHR.R" to generate the results of two comparative methods.

Step 4: Run the "PE1.R" in the folder PE_FULL, to reproduce the results of the proposed method with power enhancement term JPE.

Step 5: Run the "Data_Analysis.R" to summarize all the output into different tables.

Step 6: Run the "Real_Data_Analysis.R" in the folder "Real Data Analysis" to reproduce the result of the transNOAH data.

Implementation details and illustration of outputs are provided in the README.md.

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)

Due to the large computational cost, We use the clusters of the University to finish all the tasks.

Notes (optional)