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

In Section 6, we considered the lymphoma data set available from R package **spls**. It consists of 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 samples of follicular lymphoma (FL), and 11 samples of chronic lymphocytic leukemia (CLL), coded as $Y = 0$, 1, and 2, respectively. And $p = 4026$ gene expression measurements are recorded as the predictor X.

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: The data is available from R package **spls**.

☐ 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.): .Rda
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☒ Other (please specify): The data is a part of the R package **spls**.

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: https://cran.r-project.org/web/packages/spls/spls.pdf

Additional Information (optional)

- Lymphoma data set: For more details of the data set, please refer to R package **spls** and the paper:
  - Chung, D., Chun, H., and Keles, S. (2019). spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.2-3. https://CRAN.R-project.org/package=spls

– Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1).

Part 2: Code

Abstract

Each folder reproduces the results in simulation studies (Section 5) or real data analysis (Section 6).

Description

All statistical analyses were performed using R version 4.0.0. Each folder is described as follows:

1. *simulation_p1000*, *simulation_p3000*: reproduce results in Table 1 with $p = 1000$ and $p = 3000$. The output is saved to the folder *output*. Please refer to Section **Workflow** for details.

- *simulationX_dat.R*: simulates data in model X.
- *simulationX_LassoSIR.R*: implements LassoSIR in model X.
- *simulationX.R*: implements SEAS-SIR, SEAS-Intra, and SEAS-PFC in models 1-4, and implements Lasso in models 1-2.

2. *real_data_analysis*: reproduces the results in the real data analysis.

- *lymphoma_est.R*: reproduces the Estimation part (columns 2–5) in Table 2.
- *lymphoma_pred.R*: reproduces the Classification error part (columns 6–9) in Table 2.
- *plot_lymphoma.R*: reproduces Figure 1.

The common R files contained in all these folders are described as follows:

- *seas.R*: SEAS algorithm (Algorithm 1)
- *utility.R*: auxiliary functions.
- *LassoSIR_revised.R*: We revise the 'LassoSIR' function from R package 'LassoSIR'. The revised function accepts the user-specified cross-validation folds index and the user-specified tuning parameter sequence and records the computation time.

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

R

Version of primary software used R 4.0.0

Libraries and dependencies used by the code

- energy_1.7-8
- e1071_1.7-9
- ggplot2_3.3.2
- glmnet_4.1-2
- LassoSIR_0.1.1
- latex2exp_0.4.0
- MASS_7.3-54
- Matrix_1.2-18
- msda_1.0.2
- nnet_7.3-13
- pbmcapply_1.5.0
- randomForest_4.6-14
- spls_2.2-3

Supporting system/hardware requirements (optional)

Platform: x86_64-redhat-linux-gnu (64-bit)

Running under: CentOS Linux 8

Parallelization used

☒ No parallel code used: *plot_lymphoma.R* in the folder *real_data_analysis*
☒ Multi-core parallelization on a single machine/node
  – Number of cores used: 16 cores on one node are used for all the other codes.
☐ Multi-machine/multi-node parallelization
  – Number of nodes and cores used:

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:

- *simulation_p1000*: Reproduces results in Table 1 with $p = 1000$.

- *simulation_p3000*: Reproduces results in Table 1 with $p = 3000$.
- *real_data_analysis*:
  - *lymphoma_est.R*: Reproduces the Estimation part (columns 2–5) in Table 2.
  - *lymphoma_pred.R*: Reproduces the Classification error part (columns 6–9) in Table 2.
  - *plot_lymphoma.R*: Reproduces Figure 1.

Workflow

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

- The folders *simulation_p1000* and *simulation_p3000* reproduce the results in Table 1. Since their workflows are similar, we use *simulation_p1000* as an example. We show how to generate the results in model (M1) with $p = 1000$. First, we run file *simulation1_dat.R* to randomly generate the basis matrix β (saved to the folder *beta*), the cross-validation folds index (saved to the folder *dat*), and the data replicates (saved to the folder *dat*). Then we run files *simulation1_LassoSIR.R* and *simulation1.R* to generate all comparison criteria for each method. The results are saved to the folder *output*.
- For the folder *real_data_analysis*, run *lymphoma_pred.R* and *lymphoma_est.R* to reproduce the results in Table 2. Run *plot_lymphoma.R* to reproduce Figures 1.

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)

In our numerical studies, we include natural SSIR and refined SSIR (Tan et al., 2020) as competitors. However, since the authors' codes are not open to the public, we do not include the implementation of these two competitors in our reproducible materials. The codes may be requested from the authors.

Reference: Tan, K., Shi, L., Yu, Z., et al. (2020). Sparse sir: Optimal rates and adaptive estimation. *The Annals of Statistics*, 48(1):64-85.

Notes (optional)