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

We use two real datasets in the analysis. (1) The supermarket dataset with each record corresponds to a daily observation collected from a major supermarket located in northern China. The response of interest is the number of customers in one particular day. Each predictor corresponds to one particular product's sale volume on that day. (2) Arcene dataset is to distinguish cancer versus normal patterns from mass-spectrometric data. This is a two-class classification problem with continuous input variables.

Availability

- ☒ Data **are** publicly available.
- ☒ Data **cannot be made** publicly available.

The Arcene dataset is publicly available online. The supermarket dataset cannot be made publicly available due to data sharing restrictions.

Publicly available data

- ☒ Data are available online at: https://archive.ics.uci.edu/dataset/167/arcene
- ☐ 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

Unfortunately, public sharing of the supermarket dataset is not possible due to data use agreements and confidentiality constraints imposed by the data provider. The dataset contains proprietary business information that could potentially reveal sensitive commercial patterns or affect competitive interests.

We believe the scientific contribution of this paper outweighs the lack of reproducibility. The goal of our work is to develop methodology for identifying important predictors of a response variable—the number of customers in a supermarket. The specific results from the dataset are only of direct interest to the supermarket owner, rather than the broader scientific community.

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):

Data dictionary

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

Additional Information (optional)

Part 2: Code

Abstract

The code implements the two simulations described in the main text as well as additional results presented in the supplementary material. For the real data analysis, it implements the computational methods for identifying important predictors in the supermarket and Arcene datasets, supporting the analyses reported in the paper.

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 4.4.1

Libraries and dependencies used by the code geigen 2.3 VGAM 1.1.11 Matrix 1.7.0 foreach 1.5.2 parallel 4.4.1 doParallel 1.0.17 MASS 7.3.60.2 expm 1.0.0

Supporting system/hardware requirements (optional)

platform aarch64-apple-darwin20
arch aarch64
os darwin20
system aarch64, darwin20
status
major 4
minor 4.1
year 2024
month 06

day 14
svn rev 86737
language R
version.string R version 4.4.1 (2024-06-14)

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:

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:
- ☐ 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

The workflow is documented in a README file included in the supplementary material. Follow the instructions there to reproduce the analyses. The code uses R scripts to process the Arcene dataset (publicly available) and the supermarket dataset (non-public). Ensure all listed R packages are installed and use R version 4.4.1. Parallelization requires a machine with at least 10 cores.

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)