JASA ACS Reproducibility Initiative - Author Contributions Checklist Form

The purpose of the Author Contributions Checklist (ACC) Form is to document the code and data supporting a manuscript, and describe how to reproduce its main results.

As of Sept. 1, 2016, the ACC Form must be included with all new submissions to JASA ACS.

This document is the initial version of the template that will be provided to authors. The JASA Associate Editors for Reproducibility will update this document with more detailed instructions and information about best practices for many of the listed requirements over time.

Table of Contents

Data

Abstract (Mandatory)

The Bangladesh nutrition-environment interaction dataset contains 351 records of 20 variables. The variables include 12 background covariates (age, education, smoking history, sex, birth weight, length, head circumference, birth order and gestational age, HOME score, maternal depression scale, maternal IQ), 3 measurements for *in utero exposure* to environmental metals Arsenic (As), Manganese (Mn) and Lead (Pb), and variables for pregnancy intake for five major nutrient groups (*macronutrient*, *mineral*, *vitamin A*, *vitamin B* and *the other vitamins)*. Please see Section 2 of the manuscript for details.

Availability (Mandatory)

The Principal Investigators of Bangladesh Reproductive Cohort Study (Project Jeebon), who are also co-authors of this publication, have given permission for a de-identified version of the dataset to be made available with this article.

Description (Mandatory if data available)

The data is explained in full in the text of Section 5.1. of the manuscript. It is available in csv format in the supplementary material.

Optional Information (complete as necessary)

N/A

Code

Abstract (Mandatory)
R scripts implementing a suite of routines for estimation and hypothesis testing using CVEK (Cross-validated Ensemble of Kernels). It contains functionalities for defining base models, estimation routines for kernel machine regression and a cross-validated ensemble, and procedures for computing p-values using the proposed variance component test.

Description (Mandatory)

- How delivered: The code is written in R and is included in a zipped file `code.zip`.
  All scripts used to reproduce simulation results/figures are included under the `./code/numeric_study` folder.
  All scripts used to reproduce data analysis results are included under the `./code/data_analysis` folder.

- Licensing information: MIT License. See `./code/LICENSE.md`

- Link to code/repository: Original code written specifically for this publication is included as a zipped file with the submission.

Optional Information (complete as necessary)

Hardware requirements:

Code is developed and tested using R 3.6.1.
The simulation study is conducted under HMS Research Computing's Orchestra cluster (LSF system). Current implementation depends on the following R packages:

- limSolve: version 1.5.6.
- MASS: version 7.3.53.
- Matrix: version 1.2-17
- mvtnorm: version 1.1.1.
- CompQuadForm: version 1.4.3.
- magrittr: version 2.0.1.
- dplyr: version 1.0.2.
- mgcv: version 1.8-33.
- psych: version 2.1.6
- survival: version 2.44-1.1
- survey: version 4.0.
- kernlab: version 0.9-29
- iterators: version 1.0.12
- doParallel: version 1.0.16
- foreach: version 1.4.7

All packages are available through CRAN (https://cran.r-project.org/) and can be installed automatically by running `install.packages(PACKAGE_NAME)` in an R session.

Folder Structure:

./code/numeric_study/ code to reproduce numeric study (Figure 1, C1-C4).

- main.R Wrapper function to run the simulation.
- main_header.R Defines functions to execute individual simulation.
- settings.txt Config file that specifies all simulation settings.
- command_exec*.sh Shell scripts (6 in total) to submit the simulation jobs to cluster.
- plot_power.R R code to generate Figures 1 and C.1-C.4 from simulation.
- README.md The readme file.
- ./plot Empty folder with subfoloders to write Figures 1 and C.1-C.4 to.

./code/data_analysis/ code to reproduce results in data analysis (Table 2).

- main.R Main script for executing data analysis.
- main_header.R Defines utility functions for data analysis.
- README.md The readme file.
- ./data Folder that contains the data file.
- ./result Empty folder where the result table will be written to.

./code/functions/ code for the core functionalities of CVEK models

- ./cvek Functions for CVEK ensemble regression and kernel tests.
- ./baselines Implementations of baseline methods (iSKAT, GKM).

Instructions for Use

Reproducibility (Mandatory)

What is to be reproduced

- Numeric Simulation: Figure 1 and Figure C.1-C.4.
- Data Analysis: Table 2.

How to reproduce analyses

- Numeric Simulations:
  - On an LSF cluster, unzip ./`code/numeric_study` to a directory of choice.
  - Under the directory, run the shell script `command_exec1.sh` to `command_exec6.sh` separately (The script will submit many parallel jobs to run different sub-parts of the simulation. On the LSF cluster, aach job may require up to 240MB memory and takes on average 30 min to complete). Simulation results will be stored as `power_res.txt.`
  - After all jobs are completed, run `./plot_power.R` to generate subfigures of Figure 1 and Figure C.1-C.4. Results will be written to `./plot/.` There will be five folders under ./plot/., and each folder corresponds to a figure, which contains nine subfigures in pdf format. These nine subfigures will be further combined together using LaTeX.

- Data Analysis:
  - Unzip `code` to a directory of choice.
  - Under the directory, either:
    - Start terminal and set the working directory to the `` `./code/data_analysis` `` folder. (e.g., `` `cd ./code/data_analysis` ``). Then, run the analysis by executing the `` `main.R` `` script. e.g. `` `Rscript ./main.R` ``.
    - Alternatively, start R / Rstudio GUI and set the working directory to the `` `data_analysis` `` folder. (e.g., `` `setwd("~/PARENT_DIRECTORY/code/data_analysis/")` ``), and then execute the content of `` `main.R` `` script. (e.g., `` `source("./main.R")` ``)
  - Result table (`` `table.txt` ``) will be written to the `` `./result/` `` directory.

Running time:

- **Numeric Simulations**: For Figure 1, it takes ~5 hours to run all the 12 methods for data generated from Matern with smoothness parameter 3/2, spectral frequency parameter 0.5 and interaction strength 0.5 on a personal computer (Apple M1 chip with 8 core CPU, 8 core GPU, and 16 core Neural Engine, 16GB unified memory).

- **Data Analysis**: On a personal computer, it takes ~50 min - 1 hour to complete the full result for Table 2 (Ubuntu 16.04 computer with a 8-core Intel i7-6700HQ 2.60GHz CPU and 16 Gb memory.)

Notes

Code is available as a zip file with the journal submission.