Author Contributions Checklist Form

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

This data set contains 120 twelve-week old male rats and 18976 different probes (genes) from eye tissue were measured. The intensity values were normalized using the RMA (robust multi-chip averaging method to obtain summary expression values for each probe.

## 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 dat*a 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

Discussion of lack of publicly available data:

## Description

### File format(s)

CSV or other plain text: .txt

Software-specific binary format (.Rda, Python pickle, etc.):

Standardized binary format (e.g., netCDF, HDF5, etc.):

Other (described here):

### Data dictionary

Provided by the authors in the following file(s): RatEyeExpression.txt in the folder Real Data/Data

Data file(s) is (are) self-describiing (e.g., netCDF files)

Available at the following URL:

The data is also made publicly available at the GitHub repository

https://github.com/multizhang288/Model-Free-Inference

### Additional information (optional)

# Part 2: Code

## Abstract

-Tables\_Figures.Rmd: Generate all the tables and figures based on the saved intermediate results.

-Folder Study 1 contains the simulation codes and results of 5.1.1 and 5.1.2

-Folder Study 2 contains the simulation codes and results of Study 2, S.4.2, S.4.3, S.4.4 and S.4.6

-Folder FDR contains the simulation codes and results of 5.1.3: FDR control procedures of the proposed method.

-Folder Real Data is about the subsection 5.3 Real Data Analysis and S.4.5

All codes are made publicly available at the GitHub repository

https://github.com/multizhang288/Model-Free-Inference

## 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 (described here):

### **Supporting software requirements**

Version of primary software used

R 4.0.3

Libraries and dependencies used by the code

dplyr version 1.0.10 stringr version 1.5.0

ggplot2 version 3.4.1 splines version 4.2.2

gridExtra version 2.3 tidyr version 1.2.1

mvtnorm version 1.1-3 hdi version 0.1-9

ncvreg version 3.13.0

patchwork version 1.1.2

pracma version 2.4.2

sn version 2.1.0

### 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 300 cores

### License

MIT License (default)

BSD

GPL v3.0

Creative Commons

Other (described here):

### Additional information (optional)

Click or tap here to enter text.

# 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 here:

Figures 1, S.1 and S.3, Tables 3-5, S.1-S.5 are reproduced in Tables\_Figures.Rmd.

Figure S.2 can be reproduced by Real Data/Plot.R

In Table 2, we reproduce the Wn and TNL. TZZ is based on the existing package hdi with command lasso.proj.

In Table 3, we reproduce the FDR control results of the proposed method. The results of knockoff method is directly from the exisitng package knockoff under R 3.6.0.

## Workflow details

### Location

The workflow is available:

As part of the paper’s supplementary material

In this Git repository: https://github.com/multizhang288/Model-Free-Inference

Other:

Some readme-style files that document workflow are provided.

Since data, intermediate results, and code will take over 200MB, we can’t upload all intermediate results along with this revision submission. We have put all code, real data, and intermediate results in the GitHub repository

https://github.com/multizhang288/Model-Free-Inference

### 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 following R-packages:

dplyr version 1.0.10 stringr version 1.5.0

ggplot2 version 3.4.1 splines version 4.2.2

gridExtra version 2.3 tidyr version 1.2.1

mvtnorm version 1.1-3

ncvreg version 3.13.0

patchwork version 1.1.2

pracma version 2.4.2

sn version 2.1.0

The following steps can be implemented independently.

Step 2: Run the simulations of section 5.1.1 (The choice of h) and section 5.1.2 (Comparison with existing methods). Codes are in Folder “Study 1”

1. Check the path of this folder and set it as a working directory before you run the codes.
2. Run the codes in the folders of “Impact of h” and “Results\_h\_5”. Codes are saved in the corresponding subfolders “Codes”.
3. It might take more than 200 hours if you go through all the settings in the paper. The script only contains one setting for the purpose of the demo.
4. We store the intermediate results of all simulations in each “Intermediate Data”.
5. “Results” contains the final results that can be directly used to reproduce the tables and figures.

Step 3: Run the simulations of section Study 2, S.4.2, S.4.3, S.4.4 and S.4.6. Codes are in Folder “Study 2”

1. Check the path of this folder and set it as a working directory before you run the codes.
2. Run the codes in the folders of “Distribution”, “Extreme\_d”, “Semi-synthetic”, “Sparsity”, “Transformation” and “Adversarial\_Example”.
3. It might take more than 150 hours if you go through all the settings in the paper. The script only contains one setting for the purpose of the demo.
4. We store the intermediate results of all simulations in each “Intermediate Data”.
5. “Results” contains the final results that can be directly used to reproduce the tables and figures.

Step 4: Run the real data analysis, 5.3 and S.4.5. Codes are in Folder “Real Data”

1. Check the path of this folder and set it as a working directory before you run the codes.
2. Run the codes “Real\_Analysis.R” and “Plot.R”. Codes are saved in the corresponding subfolders “Codes”.
3. It might take 2-3 hours.
4. We store the intermediate results of all simulations in each “Intermediate Data”.
5. “Results” contains the final results that can be directly used to reproduce the tables and figures.

Step 5: Run the simulations of section 5.1.3. Codes are in Folder “FDR”.

1. Check the path of this folder and set it as a working directory before you run the codes.
2. Run the “Simualtion\_Pvalues” to generate the p-values and statistics of all variables of 1000 replications. It is not a feasible task on a local desktop.
3. We store all the statistics data in the Folder “Data”.
4. Run the “Estimation\_FDR.R”. It takes about 7-8 hours.
5. We store all the results of FDR and Power in the Folder “Results”.

Step 6: Run the Tables\_Figures.Rmd to generate the Tables and Figures. Note that they are not latex codes. The final results can be found in Tables\_Figures.html.

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:

We use the clusters of the University to finish all the tasks.

### Additional documentation (optional)

# Notes (optional)

It may take very long time to run all simulations. For your convenience, we have also saved intermediate results of our simulations, and the intermediate results can be found at the link of the GitHub repository

<https://github.com/multizhang288/Model-Free-Inference>

Since the intermediate results take over more than 200MB, we can’t upload all intermediate results along with the submission.