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

FRED-MD is a large macroeconomic database designed for the empirical analysis of "big data." The datasets of monthly and quarterly observations mimic the coverage of datasets already used in the literature, but they add three appealing features. They are updated in real-time through the FRED database. They are publicly accessible, facilitating the replication of empirical work.

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: https://research.stlouisfed.org/econ/mccracken/fred-databases/

□ 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: current.csv
☐ 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): Code_Reproduction/Real_Data/Pre-Process/current.csv
☐ Data file(s) is (are) self-describiing (e.g., netCDF files)
☐ Available at the following URL:

Additional information (optional)

Part 2: Code

Abstract

We provide the code for reproduce our simulation and real data application. See workflow and readme.md for more details.

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 4.2.2, RStudio 2022.07.2+576 "Spotted Wakerobin" Release (e7373ef832b49b2a9b88162cfe7eac5f22c40b34, 2022-09-06) for macOS Mozilla/5.0 (Macintosh; Intel Mac OS X 12_5_0) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.10 Chrome/69.0.3497.128 Safari/537.36

Libraries and dependencies used by the code

Glmnetc (4.1-6), SIS (0.8-8), MASS (7.3-58.1), Elastic-Net (1.3), randomForest (4.7-1.1),

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:

License

☒ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other (described here):

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 here:

Workflow details

Location

The workflow is available:
□ As part of the paper's supplementary material
☒ In this Git repository: https://github.com/qooyqpqy123/FARM_Reproduction
□ Other:

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

##Same Content with Readme.md in the github path ##

Reproduction for the methodology of "Are Factor Regression and Sparse Regression Adequate?"

1. **Real Data Analysis**

Data

1.1. Data Description
FRED-MD is a large macroeconomic database designed for the empirical analysis of “big data.” The datasets of monthly and quarterly observations mimic the coverage of datasets already used in the literature, but they add three appealing features. They are updated in real-time through the FRED database. They are publicly accessible, facilitating the replication of empirical work.
1.2. Data Availability
 The original dataset can be found in https://research.stlouisfed.org/econ/mccracken/fred-databases/.
 In our .zip file, the original Dataset is provided in folder ``Code_Reproduction/Real_Data/Pre_process/current.csv."
 The description of the data is provied in "Code_Reproduction/Real_Data/Pre_process/Description_of_Variables.pdf."
 This description describes the meaning of all columns of this datset.

Codes Description and Implementation Details:
 There are in total four folders under the folder: ``Code_Reproduction/Real_Data."
 1. Pre_process,
2. Prediction_Table_3
  3. PCR_Adequate_Table4,
 4. Sparse_Adequate_Table4.
Folder `Code_Reproduction/Real_Data/Pre_process':
 It contains the Codes for real data pre-processing. The document `current.csv' in that folder is our downloaded orignal dataset from the aforementioned website.

 The code **realdata_read.R** contains the codes for pre-processing the data. The line 4 in "realdata_read.R", we have a variable called: "choose_time". If one lets choose_time=1 and run the codes, the output is the processed dataset **realdata_115_126.csv**. If one lets choose_time=2, and run the codes, the output is the processed dataset **realdata_189_126.csv**.

 Run time of **realdata_read.R** is less than 3 min.

 Finally, for document "Description_of_Variables.pdf" inside that document, it contains the detailed description of all variables in the orignal dataset "current.csv".

Folder `Code_Reproduction/Real_Data/Prediction_Table3':
This folder contains the codes for reproducing the prediction results of Table 3 in our paper.

*For code file **realdata_prediction_115.R**, it first reads the data "realdata_115_126.csv" in that folder. On the line 54 of this file, there is a variable "j". If we set j=49, and run the codes, we reproduce the line "HOUSTNE" line during "08.2010-02.2020" in Table 3. If we set j=81 and run the codes, we reproduce the line "GS5" line during "08.2010-02.2020" in Table 3.

*For code file **realdata_prediction_189.R**, it first reads the data "realdata_189_126.csv" in that folder. On the line 54 of this file, there is a variable "j". If we set j=49, and run the codes, we reproduce the line "HOUSTNE" line during "02.1992-10.2007" in Table 3. If we set j=81 and run the codes, we reproduce the line "GS5" line during "02.1992-10.2007" in Table 3.

For both .R files, the output are printed in line 149-161 and 189 for different methods :Lasso (sparse linear regression), PCR (latent factor regression), Ridge (Ridge regression), El-Net (Elastic Net), RF (Random Forest), and FarmSelect (Factor adjusted Lasso), respectively.

Run time of **realdata_prediction_115.R** and **realdata_prediction_189.R** are less than 10 min.

Folder `Code_Reproduction/Real_Data/PCR_Adequate_Table4':
This folder contains the codes for reproducing the hypothesis testing results of column "LA_factor" of Table 4 in our paper.

*For code file **Real_data_PCR_115.R**, it first reads the data "realdata_115_126.csv" in that folder. On the line 17 of this file, there is a variable "j". If we set j=49, and run the codes, we reproduce the line "HOUSTNE", column "LA_factor," during "08.2010-02.2020" in Table 4. If we set j=81 and run the codes, we reproduce the line "GS5" line, column "LA_factor", during "08.2010-02.2020" in Table 4.

*For code file **Real_data_PCR_189.R**, it first reads the data "realdata_189_126.csv" in that folder. On the line 17 of this file, there is a variable "j". If we set j=49, and run the codes, we reproduce the line "HOUSTNE", column "LA_factor," during "02.1992-10.2007" in Table 4. If we set j=81 and run the codes, we reproduce the line "GS5" line, column "LA_factor", during "02.1992-10.2007" in Table 4.

For both .R files, the output are printed p-values of the test on whether PCR is adequate.

Run time of **Real_data_PCR_115.R** and **Real_data_PCR_189.R** are less than 10 min.

Folder `Code_Reproduction/Real_Data/Sparse_Adequate_Table4':
This folder contains the codes for reproducing the hypothesis testing results of column "SP_Linear" of Table 4 in our paper.

*For code file **Real_data_Sparse_115.R**, it first reads the data "realdata_115_126.csv" in that folder. On the line 94 of this file, there is a variable "j". If we set j=49, and run the codes, we reproduce the line "HOUSTNE", column "SP_Linear," during "08.2010-02.2020" in Table 4. If we set j=81 and run the codes, we reproduce the line "GS5" line, column "SP_Linear", during "08.2010-02.2020" in Table 4.

*For code file **Real_data_Sparse_189.R**, it first reads the data "realdata_189_126.csv" in that folder. On the line 94 of this file, there is a variable "j". If we set j=49, and run the codes, we

reproduce the line "HOUSTNE", column "SP_Linear," during "02.1992-10.2007" in Table 4. If we set j=81 and run the codes, we reproduce the line "GS5" line, column "SP_Linear", during "02.1992-10.2007" in Table 4.

For both .R files, the output are printed p-values of the test on whether Sparse Regression is adequate.

Run time of **Real_data_Sparse_115.R** and **Real_data_Sparse_189.R** are less than 10 min.

2. **Simulation Studies**

Codes Description and Implementation Details:

Folder `Code_Reproduction/simulation/Consitency_Figure1/':
There are three sub-folders inside this folder, namely, Gaussian, Uniform, Heavy-tail.

For folder Gaussian and Uniform, they contain codes **Gauss_consistency.R** and **Uniform_consistency.R**, respectively. By running these codes directly, we reproduce the methodology for generating results for Figure 1 (a) and (b).

For folder Heavy-tail: it contains codes: **huber_estimation.R** and **T-consistency.R**. By running these codes directly, we reproduce the methodology for generating the red line and blue line of (c) of Figure 1.

Run time of **Gauss_consistency.R**, **Uniform_consistency.R**, **huber_estimation.R** and **T-consistency.R** are around 5 hrs, respectively.

Folder `Code_Reproduction/simulation/Simulation_Table1/':
There is one file **pcr_adequate.R** under this folder. It reproduces the methodology of Table 1.
There are several parameters in this file need to be pre-specified. They correspond to different simulation settings of the paper.

1. First, on line 7, if we let the parameter "choose_dim=1", it outputs the results with dimension p=200. If we let the parameter "choose_dim=2", it outputs the results with dimension p=500.

2. Second, on line 10, if we let the parameter "choose=1", it outputs the results under the setting where factors F and idiosyncratic components are generated via setting 1 in section 5.2 of this paper. If we let the parameter "choose=2",it outputs the results under the setting where factors F and idiosyncratic components are generated via setting 2 in section 5.2 of this paper .

3. Third, on line 16, if we let the parameter "choose_noise=1", it outputs the results when the noise is generated based on Gaussian distribution. If we let the parameter "choose_noise=2",it outputs the results under the setting where the noise follows uniform distribution.

Run time of **pcr_adequate.R** for any given setting described above is around 12 hrs.

Folder `Code_Reproduction/simulation/Simulation_Table2/':

There is one file **inference.R** under this folder. It reproduces the methodology of Table 2.
There are several parameters in this file need to be pre-specified. They correspond to different simulation settings of the paper.

1. First, on line 93, if we let the parameter "choose_p=1", it outputs the results with dimension p=250. If we let the parameter "choose_p=2", it outputs the results with dimension p=600.

2. Second, on line 94, if we let the parameter "choose_noise=1", it outputs the results where the noise is generated based on Gaussian distribution. If we let the parameter "choose_noise=2",it outputs the results under the setting where the noise follows uniform distribution.

3. Third, on line 95, if we let the parameter "choose_mix=1", it outputs the results under the setting where the covariates are generated i.i.d. If we let the parameter "choose_mix=2",it outputs the results under the setting where the covariates are strong mixing.

Run time of **inference.R** for any given setting described above is around 5 hrs.

Folder `Code_Reproduction/simulation/Prediction_Section_B.1/':

There are four files inside this folder. **Prediction_Table1.R**, **Prediction_Table2.R**, **Prediction_Table3.R**, and **Prediction_Table4.R**
They reproduces the Table 1-4 in Appendix, respectively.

Run time of **Prediction_Table1.R**, **Prediction_Table2.R**, **Prediction_Table3.R**, and **Prediction_Table4.R** for any given setting described above is less than 30 min.

Folder `Code_Reproduction/simulation/Figure7_appendix/':

There is one file **qqplot_figure7.R** in the folder. It reproduces the methodology to generate Figure 7 in Appendix. The implementation is to run codes in 1-160 lines to generate "sure.csv". After that, one needs to read in this "sure.csv" to plot the qq-plot (corresponds to lines 164-171).

Run time of **qqplot_figure7.R** for any given setting described above is around 5 hrs.

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 documentation (optional)

Notes (optional)