Overview

The code in this replication package reconstructs the analysis and exhibits in “Robust Machine Learning Algorithms for Text Analysis” by Ke, Montiel Olea and Nesbit. The replicator should expect the code to run for about 18 hours.

Data Availability and Provenance Statements

The empirical sections (Section 6 in the main text and Section F in the Online Appendix) conducts analysis using the FOMC transcript data. The data is made publicly available by the Federal Reserve Board, which can be accessed at https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm. We manually downloaded the 149 raw pdf files for the periods covered in our analysis (Aug 1987 to Jan 2006) and store them in the following folder:

  - `Preprocessing/FOMC_pdf`

Computational requirements

Software Requirements

The program is run using the following programs:

- Python 3.8.16
  - the file “`requirements.txt`” lists these dependencies, please run “`pip install -r requirements.txt`” as the first step. See https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability for further instructions on creating and using the “`requirements.txt`” file.
- Matlab (code was run with Matlab Release 2022b)

Controlled Randomness

- Random seed is set at line 88 of program `preprocessing/estimation_and_nmf.py`.

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard 2023 machine is about 18 hours.

Details

The code is run on 4-coure Intel-based laptop with Windows 10. The simulation part of the code takes less than 10 minutes. The preprocessing part of the code takes about 17 hours. The generation of plots takes about 30 minutes.

To store the Nonnegative Matrix Factorization results, it’s required to have a storage space of >150 GB.

Description of programs/code

Simulation

The replication codes for simulation (Section 6.1 of main text and Section E of online appendix) are in `simulation` folder. They are self-contained scripts that could be run individually. Below describes each folder and which plot the script replicates for the simulation section :

- 1Sensitivity: Run `Sensitivity.m` for Figure 3 in main text.
- 2Range: Run `Range_main.m` for Figure 4, 5 and 6 in main text.
- 3MonteCarlo: Run `Main_MonteCarlo.m` for Figure 7 in main text and Figure 4 in online appendix.
- 4Approximation: Run `Main_approx.m` for Figure 3 in online appendix.
- 5AnchorWord: Run `Main_AnchorWord.m` for Figure 5 and 6 in online appendix.

Empirical Exercise

The code for the empirical section consists of two parts in two folders. The `preprocessing` folder contains the code to preprocess raw pdf files, clean up the corpus, generate term-frequency matrices, plot word clouds, and run nonnegative matrix factorizations. The empirical_analysis folder contains code that generates the plots in the paper.

Instructions to Replicators

- Edit `preprocessing/Constant.py` to adjust the current working directory and the folder path for NMF_draws_folder, which should have at least 150 GB of free storage.
- Edit line 5 in `empirical_analysis/Main_empirical.m` to point to the folder path for NMF_draws_folder as defined in the previous step.
- Run the code in subfolders in the `simulation` folder, which will generate all the simulation results.
- Run `preprocessing/Main_preprocess.py` to preprocess pdf files.
- Run `preprocessing/Main_generate_NMF_draws.py` to generate and store NMF draws. This step will take about 16 hours and requires a storage space of >150GB.
- Run `empirical_analysis/Main_empirical.m` to generate all the plots in the empirical section of the paper.

List of tables and programs

The provided code reproduces:

Figure	Program	Output file
Figure 1	n.a. (no data)	
Figure 2	n.a. (no data)	
Figure 3	simulation/1Sensitivity/Sensitivity.m	Sensitivity_N10.eps; Sensitivity_N100.eps
Figure 4	simulation/2Range/Range_main.m	Range_N10.eps; Range_N100.eps
Figure 5	simulation/2Range/Range_main.m	CredibleSet90_Range_N10.eps; CredibleSet90_Range_N100.eps
Figure 6	simulation/2Range/Range_main.m	Algo2Range_N10.eps; Algo2Range_N100.eps
Figure 7	simulation/3MonteCarlo/Main_MonteCarlo.m	Freq_MC.eps; Robust_MC.eps
Figure 8	preprocessing/Main_preprocess.py	preprocessing plots/ WordCloud_FOMC1_onlyTF.png

Figure 9	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/ posterior_alpha_1.25_beta_0.025 _percent_diff.eps;
		empirical_analysis/Figures/ prior_alpha_1.25_beta_0.025 _percent_diff.eps;

Fig 1 OA	n.a. (no data)	
Fig 2 OA	n.a. (no data)	
Fig 3 OA	simulation/4Approximation/Main_approx.m	Approximation_Range_N10.eps; Approximation_Range_N100.eps
Fig 4 OA	simulation/3MonteCarlo/Main_MonteCarlo.m	diff_MC.eps
Fig 5 OA	simulation/5AnchorWord/Main_AnchorWord.m	AnchorWord_vs_posterior1.eps
Fig 6 OA	simulation/5AnchorWord/Main_AnchorWord.m	AnchorWord_vs_posterior2.eps

Fig 7 OA	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/ posterior_alpha_1.25_beta_0.025 _regression.eps;
		empirical_analysis/Figures/ prior_alpha_1.25_beta_0.025 _regression.eps;

Fig 8 OA	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/ posterior_alpha_1_beta_1 _percent_diff.eps;

		empirical_analysis/Figures/prior_alpha_1_beta_1_percent_diff.eps;

Fig 9 OA	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/posterior_alpha_1_beta_1_regression.eps;
		empirical_analysis/Figures/prior_alpha_1_beta_1_regression.eps;

Fig 10 OA	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/posterior_alpha_1.25_beta_0.025_percent_diff_FOMC2.eps;
		empirical_analysis/Figures/prior_alpha_1.25_beta_0.025_percent_diff_FOMC2.eps;

Fig 11 OA	empirical_analysis/Main_empirical.m	empirical_analysis/Figures/posterior_alpha_1.25_beta_0.025_ regression_FOMC2.eps;
		empirical_analysis/Figures/prior_alpha_1.25_beta_0.025_regression_FOMC2.eps;

Acknowledgements

The code in `preprocessing/preprocess_pdf` uses part of the `topicmodels` package developed by Stephan Hansen and is used in Hansen, McMahon, Prat (2018). The original package is now retired but archives of the code can be accessed at https://github.com/sekhansen/text-mining-tutorial.

References

Hansen, Stephen, Michael McMahon, and Andrea Prat. "Transparency and deliberation within the FOMC: A computational linguistics approach." *The Quarterly Journal of Economics* 133, no. 2 (2018): 801-870.