**JASA A&CS 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 A&CS.

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.

## Data

**Abstract (Mandatory)**

The triage data, provided by the Alberta Medical Center, is collected during the triage phase of a hospital visit in the Emergency Department. The data set contains the demographic information, medical complaints, vital readings, and documented symptoms of the patients. These documented symptoms, or triage notes, are manually typed into the system by triage nurses and is the focus of the analysis in this paper.

**Availability (Mandatory)**

The authors provide de-identified dataset, i.e. **‘Altered Level of Consciousness’** to the public for reproducibility and further research purposes. The dataset can be downloaded from “cralo31/SONMF/Dataset/” on Github.

**Description (Mandatory if data available)**

The ‘Altered Level of Consciousness’ dataset has 5220 observations (patient records) and 12 columns. The response variable used in this analysis is ‘DISPOSITION’, which is the final disposition of a patient after a doctor’s decision has been made. This is a binary variable, with a balanced ratio between **A**dmittance (51.15%) vs. **D**ischarge (48.85%).

The ‘TriageNote’ column, which consists of the manually typed nurse notes, is the focus of this study. Additional triage information of the patients (i.e. temperature, heart rate, blood pressure, etc.) are also provided in this dataset, but they are not considered in this study.

**Optional Information (complete as necessary)**

All patient records are de-identified for privacy reasons.

## Code

**Abstract (Mandatory)**

The code to reproduce the results in the JASA A&CS article “Topic modeling on triage notes with semi-orthogonal non-negative matrix factorization”.

**Description (Mandatory)**

The code to reproduce the results in the manuscript (and appendix) can be found both at ‘cralo31/SONMF/code’ on Github or in the ‘/reproducibility/code/’ folder in the supplementary materials. The code has been divided into multiple ‘.txt’ subfiles, according to the corresponding sections in the manuscript. Each code subfile is independent from the rest except for the “code\_Setup.txt” file.

The first step is to run the “code\_Setup.txt” file to install all the relevant packages to ensure the code runs properly. This file also provides some examples for the ‘MatrixFact’ package, which is an R package developed by the authors for the proposed method, along with other NMF methods compared in the manuscript. The ‘MatrixFact’ package can be downloaded at ‘cralo31/MatrixFact’ on Github.

Package: MatrixFact

Type: Package

Title: Algorithms for Non-negative Matrix Factorization

Version: 1.0.0

Date: 2019-06-11

Author: Yutong Li, Ruoqing Zhu, Annie Qu

Maintainer: Yutong Li <yutongli92@gmail.com>

Description: This package performs 6 variations of the Non-negative Matrix Factorization method for matrices with either continuous or binary entries, such as the original NMF by Lee & Seung (1999) <DOI:10.1038/44565> and SONMF by Li et. al(2018) <arxiv.org/abs/1805.02306>.

Encoding: UTF-8

License: GPL (>= 2)

RoxygenNote: 6.1.0

Imports: Rcpp (>= 0.12.18), irlba, pracma

LinkingTo: Rcpp, RcppArmadillo, irlba, pracma

**Optional Information (complete as necessary)**

The R version used for this study is 3.6.2. Multi-thread CPUs are recommended since most of the analyses in the manuscript are performed using parallel computing with multiple cores due to intensive computational demand. GPU is not utilized in this study.

The required R packages and corresponding versions to fully reproduce the results are listed below.

|  |  |  |
| --- | --- | --- |
| 1 | pheatmap | 1.0.12 |
| 2 | class | 7.3-17 |
| 3 | randomForest | 4.6-14 |
| 4 | doRNG | 1.8.2 |
| 5 | rngtools | 1.5 |
| 6 | caret | 6.0-86 |
| 7 | ggplot2 | 3.3.2 |
| 8 | lattice | 0.20-38 |
| 9 | pracma | 2.2.9 |
| 10 | rlang | 0.4.8 |
| 11 | RcppArmadillo | 0.10.1.2.0 |
| 12 | Rcpp | 1.0.05 |
| 13 | raster | 3.4-5 |
| 14 | sp | 1.4-2 |
| 15 | ranger | 0.12.1 |
| 16 | rorpcor | 1.6.9 |
| 17 | e1071 | 1.7-4 |
| 18 | doParallel | 1.0.16 |
| 19 | iterators | 1.0.12 |
| 20 | dplyr | 1.0.2 |
| 21 | dlmnet | 4.0-2 |
| 22 | irlba | 2.3.3 |
| 23 | matrix | 1.2-18 |
| 24 | tm | 0.7-8 |
| 25 | nLP | 0.2-1 |
| 26 | foreach | 1.5.1 |
| 27 | readr | 1.4.0 |
| 28 | stringr | 1.4.0 |
| 29 | MatrixFact | 1.0.0 |

## Instructions for Use

**Reproducibility (Mandatory)**

All results in the simulation sections are seeded and reproducible.

All results regarding the ‘Altered Level of Consciousness’ dataset are seeded and reproducible.

Note that since the remaining triage note datasets are not available publicly, part of the manuscript is not fully reproducible. However, all the remaining datasets utilize the same analysis pipeline as the “Altered Level of Consciousness” dataset. Therefore, the provided code and dataset is sufficient for general reproducibility of our manuscript.

Below are the approximate runtimes for each major code chunk.

1. Section 5.1 (continuous simulation): **Figure 2,3** and the results in **Table 1,2** takes about 45 minutes to produce. **Figure 4** and **Table 3** takes about 25 minutes. Both estimates are from running the 100 simulations in parallel with 8 CPU cores.
2. Section 5.2 (binary simulation): **Figure 5** and **Table 4** takes about 90 minutes to run 100 simulations in parallel with 8 CPU cores.
3. The results in Section 6.1 are very computationally intensive to produce. It is recommended to use HPC clusters for these analyses. In the manuscript, 20 CPU cores were utilized in parallel to produce the results in **Table 5** and **Table 6**, and consequently, **Table 7**.
   1. Section 6.1 (triage bag-of-words): It takes up to 25 hours to run 20 trials (5-fold CV per trial) in parallel using 20 CPU cores to run the four supervised learning models on the ‘Altered Level of Consciousness’ bag-of-words matrix **(Table 5).** In particular, the random forest model, along with the considered tuning parameters, takes about 18 hours to run to run 20 trials in parallel using 20 CPU cores. The other three models take a much shorter time, with 2 hours for penalized linear regression, 3 hours for KNN, and 1.5 hours for SVM respectively. On the other hand, it can take more than 50 hours to produce the results for the larger datasets, i.e. ‘Cough’ and ‘Lower Extremity Injury’, using the same setup.
   2. Section 6.1 (triage SONMF): Depending on the NMF method and factorization rank, it takes anywhere between 15 minutes to 4 hours to produce one line of classification result in **Table 6**, given that the 20 trials were ran in parallel with 20 cores. The larger the factorization rank, the longer it takes to run the matrix factorization method, and hence the longer it takes to produce a result.
   3. Section 6.1 (triage factorization cost and sparsity): Depending on the NMF method and factorization rank, it takes anywhere between 3 minutes to 40 minutes to produce one line of result cost and sparsity result in **Table 6**, given that the 20 trials were ran in parallel with 20 cores. The larger the factorization rank, the longer it takes to run the matrix factorization method, and hence the longer it takes to produce a result.
4. Section 6.2 (triage topics): It takes around 5 minutes to run the functions to generate the word-topics **(Figure 6, Table 8,9)** and document-topics **(Figure 7, Table 10)** for the ‘Altered Level of Consciousness’ dataset. Note that the two datasets used in the manuscript, i.e. “Lower Extremity Injury” and “Symptoms of Stroke” are not available to the public.
5. Section A.2 (initialization study): It takes around 90 minutes – 120 minutes to produce each of the three plots **(Figure 8)** using one CPU core. Note that rather than using the wrapper ‘nmf.main()’ function from the ‘MatrixFact’ package, the raw ‘SONMF()’ function was called directly from the ‘src’ folder in the ‘MatrixFact’ package. Parallel computing is not available for this section.
6. Section A.3 (rank-deficient): It takes around 3 hours to run 100 trials and generate all three plots **(Figure 9)** with 8 cores running in parallel.
7. Section A.4 (binary SONMF step size): It takes around 450 minutes to run 100 trials for the 5 different step sizes (~90 minutes each) across 3 different rank settings with 8 cores running in parallel. The results of the 5 different step sizes are then combined to plot **Figure 10**.
8. Section A.5
   1. Section A.5 (triage additional): The cluster membership **(Figure 11)** plot and the elbow plot **(Figure 12)** takes about 3 minutes to produce.
   2. Section A.5 (triage all data): **Table 10 – 15** are computationally intensive to produce and the time required depends on the size of the dataset, the NMF method and the factorization rank. For example, the largest two datasets, ‘Cough’ and ‘Lower Extremity Injury’, takes anywhere between 45 minutes to 20 hours to run 20 trials in parallel with 20 cores for the four models on one NMF method + rank setting.