PSIVCM-Paper2021

Ng Hoi Min (hoi-min.ng@connect.polyu.hk)

Introduction

This directory contains the codes to perform all the analyses and reproduce all the figures and tables presented in the paper: Ng HM, Jiang BY, Wong KY. Penalized estimation of a class of single-index varying-coefficient models for integrative genomic analysis. 2021.

Before running the codes, install the R-package **psivcm** which can be found in the home directory. The programs also require the R-packages **grpreg**, **splines2**, **survival**, and **SurvC1**. A list of configurations is shown at the end of this document.

Simulation Settings

We introduce the simulation settings considered in the paper:

- Setting 1 (`model = 1`): the main study.
- Setting 2 (`model = 2`): the additional simulation study concerning a main effect model.
- Setting 3 (`model = 3`): the additional simulation study concerning a model with non-linear main effects.
- Setting 4 (`model = 4`): the additional simulation study concerning an interaction model.

Simulation Data Sets

The simulation data sets used in the paper are stored in the directory `Simulation/SimulationData`. The zip files in the directory contain the file `Simulation-beta0.csv`, 303 files with names in the form of `SimulationData-p[number of covariates in X]-[replication number].csv`, and 909 files with names in the form of `AdditionalSimulationData-setting[setting number]-[number of covariates in X]-[replication number].csv`. For details, see the following data documentation:

- The file `Simulation-beta0.csv` stores the initial values of single-index parameter that are used for all simulation replicates. Each row in a file contains a set of initial values. This file consists of rows in the following form:

```
head(as.matrix(read.csv("./Simulation/SimulationData/Simulation-beta0.csv", header = FALSE)))
```

```
##              V1          V2         V3          V4
## [1,] -0.3270345  0.09586933 -0.4362323  0.83280182
## [2,] -0.4164886  0.08583603  0.7373273 -0.52489798
## [3,] -0.6202452 -0.18861771  0.1668641 -0.74288331
## [4,]  0.1775360 -0.44433592  0.7299041  0.48814599
## [5,] -0.4100676  0.67512308 -0.6122771  0.03420716
```

- Each of the 303 files contains the simulation data for 500 subjects (5000 subjects for `replication number = 0`) for a specific simulation replicate under setting 1. Each row in a file contains data for a subject. The first element on each row is the continuous outcome, the second element and the third element are the right-censored outcome corresponding to the observed time and the event indicator (that equals 1 or 0 if the event is observed or right-censored), respectively. The remaining elements are the observed values of $\boldsymbol{X}$ and $\boldsymbol{U}$. Each file consists of rows in the following form:

```
head(as.matrix(read.csv("./Simulation/SimulationData/SimulationData-p20-0.csv")))
```

```
##               Y      time status         X1         X2          X3         X4
## [1,] -2.2168735 2.4191907      1  1.2629543  0.3925581 -1.21955126 -0.4087576
## [2,]  2.0932791 0.2879223      1 -0.3262334  0.4582467 -1.20146018 -1.0759654
## [3,]  0.3108285 0.3650713      1  1.3297993 -1.2196454 -0.49604255  0.8524326
## [4,] -3.4811298 2.3271571      1  1.2724293 -1.1220983  0.06693112  1.1176122
## [5,] -0.5815491 1.1647004      1  0.4146414  0.9933303 -0.05694914  0.9210962
## [6,]  4.8036485 0.2369540      1 -1.5399500 -1.8332470  0.25558017  0.1765438
##              X5          X6         X7         X8         X9        X10
## [1,] -0.5288658  0.31351432 -1.6106983 -0.7343096  0.1129743 -1.9107605
## [2,]  0.1329492  0.52623495  0.3783567  0.7000116 -0.5159348  0.3370675
## [3,] -0.2724691 -1.17989691 -1.2570990  1.0060950  0.7626851  1.0427772
## [4,]  0.6429857 -1.62834804 -0.2540740  1.0474575  0.2103232 -1.1408760
## [5,]  0.9979110 -0.09839114  1.4356872 -1.0881598 -1.8987746  1.4644663
## [6,] -1.0999127  0.91341797  1.4030621  1.1347706 -0.5999515  1.2671652
##             X11        X12         X13        X14        X15        X16
## [1,] -1.1595593 -0.5143815  0.58447183  0.4801717 -1.0985986  0.7141512
## [2,]  0.5294831 -0.5141191 -0.07573724  2.2161453  0.9145290  0.3376984
## [3,]  0.1316142 -1.0854109 -0.67335942 -0.2323169 -0.9487341  1.8974608
## [4,]  0.4708434  0.7170363  0.96146897  1.9005009 -0.2414130  1.7364285
## [5,]  0.6656041  0.3690821 -0.70523555  0.8444283 -0.8924001 -0.2626786
## [6,]  0.2364149  0.4012748 -0.06542054  0.8801384  0.1802821  0.3154389
##              X17        X18        X19        X20          U1         U2
## [1,]  1.50972999  1.1260095 -1.4552316 -1.2579881  1.94778492  1.1272995
## [2,]  0.88602138  0.1113577  0.7698239 -0.5207599  0.06377977  0.5762405
## [3,] -0.45867129 -1.4442747 -1.4679826 -0.6897890  0.69070570 -0.4855107
## [4,] -1.84599119  1.0016545  0.8515475 -0.3146988 -0.37134317  0.8483763
## [5,] -0.06687141 -1.1242585  0.2075026 -0.6963116  0.40274504 -1.7698105
## [6,]  0.69728718  2.4427992  0.5318263  1.2938959  0.26475219 -0.2971756
##               U3         U4
## [1,] -0.23746516  0.6777674
## [2,]  0.36338160  0.2375314
## [3,]  0.05900119  0.4663721
## [4,] -0.48466276  0.3328498
## [5,]  0.01474475 -0.7225432
## [6,]  0.12840204 -0.3299229
```

- Each of the 909 files contains the responses for other simulation settings (setting 2 to 4). Each row in a file contains data for a subject. The first element on each row is the continuous outcome, the second element and the third element are the right-censored outcome corresponding to the observed time and the event indicator (that equals 1 or 0 if the event is observed or right-censored), respectively. Each file consists of rows in the following form:

```
head(as.matrix(read.csv("./Simulation/SimulationData/AdditionalSimulationData-setting2-p20-0.csv")))
```

```
##               Y      time status
## [1,] -0.40720396 0.9788257      1
## [2,]  1.86695412 0.3224194      1
## [3,]  0.03114291 0.4198659      1
## [4,] -3.07537013 1.8998360      1
## [5,] -1.77551362 2.1158278      1
## [6,]  6.42168940 0.1055142      1
```

Users can instead simulate the random data by running the program `SimulateData.R` in the directory

`Simulation`. It will generate the initial values of single-index parameter and 101 simulation data sets (100 training data sets and a validation data set) for each simulation setting to the directory `Simulation/SimulationData`.

Simulation Studies

Run the program `SimulationAnalysis.R` in the directory `Simulation`. The program performs analysis on 100 simulation data sets and validates on a validation data set (with `replication number = 0`). For each simulation setting and estimation method (and initial values of single-index parameter for the proposed method), the program generates 2 output files in the directory `Simulation/SimulationResults`. Each row of the output file contains the simulation results for a replication. Files with prefix `SimulationResults-` in the name contain a summary of the model performance, including the sensitivity, FDR, cardinality, MSE, C-index, and lambda value(s). Files with prefix `Estimates-` in the name contain the estimated parameters of the model with smallest modified BIC value.

To reproduce all the analyses in the paper, users should follow the comments stated in lines 12-19 of the program `SimulationAnalysis.R` and change the simulation setting accordingly. Since some of the programs may take long time to run, we have included all the simulation results in `SimulationResults.zip` in the directory `./Simulation/SimulationResults`. Users should copy the intermediate results to `./Simulation/SimulationResults`. The program `SummerizeResults.R` in the directory `Simulation` summarize the simulation results for each setting over 100 replicates.

Analysis of TCGA data

We provided two real data examples for the application of the proposed method. The raw data downloaded from UCSC Xena data hubs can be found in the directory `RealDataAnalysis/RealDataAnalysisData`. Run the program `DataProcessing.R` to extract relevant data for the analyses. The processed data will be written in the same directory. Run the program `RealDataAnalysis.R` in the directory `RealDataAnalysis`. This program generate the files `RealData-NSCLC-beta0.csv` and `RealData-LGG-beta0.csv` that store 50 initial values of single-index parameter to the directory `RealDataAnalysisData`. Each row in a file contains a set of initial values of single-index parameter. The program then performs analyses on the NSCLC and LGG data sets. For each analysis, the program generates an output files in the directory `RealDataAnalysis/RealDataAnalysisResults` and each row in the output file contains the estimated regression parameters under each initial value of the single-index parameter.

Generation of Figures

Upon completion of all analyses, run the programs `plotFigure1andS1andS2andS3andS4andS5.R` in the directory `Simulation` and `plotFigure2and3.R` in the directory `RealDataAnalysis`, which generate the figures for the estimated coefficient. These programs generate Figures 1 and S1-S5 to the directory `Simulation/SimulationResults` and Figures 2-3 to the directory `RealDataAnalysis/RealDataAnalysisResults`.

List of configurations

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] survC1_1.0-3    survival_3.2-13 splines2_0.4.4  psivcm_0.1.0
## [5] grpreg_3.4.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7      lattice_0.20-44 digest_0.6.28   grid_4.0.2
##  [5] magrittr_2.0.1  evaluate_0.14   rlang_0.4.12    stringi_1.7.5
##  [9] Matrix_1.3-4    rmarkdown_2.11  splines_4.0.2   tools_4.0.2
## [13] stringr_1.4.0   xfun_0.27       yaml_2.2.1      fastmap_1.1.0
## [17] compiler_4.0.2  htmltools_0.5.2 knitr_1.36
```