Bayesian Hierarchical Varying-sparsity Regression Models with Application to Cancer Proteogenomics

by Yang Ni, Francesco C. Stingo, Min Jin Ha, Rehan Akbani, & Veerabhadran Baladandayuthapani

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 dataset contains gene expressions, protein expressions and patients' survival times across 4 cancers (kidney renal clear cell carcinoma, ovarian serous cystadenocarcinoma, skin cutaneous melanoma and head and neck squamous cell carcinoma. The genes/proteins are key members of 12 core pathways in those 4 cancers.

Availability (Mandatory)
We retrieve the genomic, proteomic and clinical data from TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/?) using TCGA-Assembler (Zhu et al., 2014)

Description (Mandatory if data available)
The TCGA data portal has been moved to https://gdc.cancer.gov/access-data
TCGA-Assembler (http://www.compgenome.org/TCGA-Assembler/) or other similar software can be used to download the data.

Optional Information (complete as necessary)

Code

Abstract (Mandatory)
The code implements the MCMC algorithm of BEHAVIOR model described in Section 2.
It allows for drawing posterior samples from the model and making prediction in survival times for testing dataset.

Description (Mandatory)
The Matlab compiled executable is submitted in a zip file with the manuscript, which can be run on any platform with or without Matlab license.

Optional Information (complete as necessary)
Free MATLAB Runtime (v9.1) can be downloaded and installed from (http://www.mathworks.com/products/compiler/mcr/).

Reproducibility (Mandatory)
Table 1 and Figures 2-6

Replication (Optional)

Notes

The main function is BEHAVIOR which takes 6 inputs and returns 10 outputs which are necessary to reproduce our results in simulations and case studies.

-Input files (put in the same directory of BEHAVIOR)
“parameter.csv”: in the format of [N,s] where N is the number of MCMC iterations and s is the seed used by random number generator.
“y.csv”: n by 2 survival variable with 1st column being survival/censoring times and 2nd column being censoring indicator (1=death,0=censored)
“P.csv”: n by p proteins
“G.csv”: n by p genes
“Pt.csv”: n by p proteins for test data
“Gt.csv”: n by p genes for test data

-Output variables are stored in “BEHAVIOR.mat":
“reg_coef”: protein effect
“reg_rate”: inclusion probability of protein
“lambda”: threshold
“lin_coef”: linear gene effect
“nlin_coef”: nonlinear gene effect
“const_coef”: constant gene effect
“lin_rate”: inclusion probability of linear gene effect
“nlin_rate”: inclusion probability of nonlinear gene effect
“const_rate”: inclusion probability of constant gene effect
“ypred”: predictive values of y for training data Pt and Gt.