**JASA ACS Reproducibility Initiative - Author Contributions Checklist Form**

## Data

**Abstract**

We applied our method to multi-omic datasets from the 7 TCGA tumor types, lung

adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), colon

adenocarcinoma (COAD), rectum adenocarcinoma (READ), uterine corpus

endometrial carcinoma (UCEC), ovarian serous cystadenocarcinoma (OV),

and skin cutaneous melanoma (SKCM). For each of the tumor types, we included

multi-platform data, DNA methylation, copy number alteration, mRNA expression data,

and reverse phase protein array (RPPA)-based proteomic data.

**Availability**

We downloaded, assembled, and processed public The Cancer Genome Atlas (TCGA) data (<https://portal.gdc.cancer.gov)>.

**Description**

Using TCGA-assembler R package, we downloaded molecular profiling data, DNA methylation, Copy number variation, mRNA expression, and RPPA-based protein expression data from TCGA Data Coordinating Center for the matched TCGA samples. Then we matched the sample ids across platforms and selected genes/proteins that are included in the key signaling pathways described in pathway information file. The pathway file StringV10\_ppi\_pathway.Rdata in the Data folder includes pathway information in the “pathwaydat” object: the data frame has pathway names (first column), protein names (second column) and the matched gene names (third column). The data folder includes R datasets of the multi-omic data with matched samples. Each of the TCGA\_tumor type\_CNA\_Methylation\_mRNA\_RPPA.rda files include 4 objects, CNA, methylation, mRNA, RPPA that are data frames (genes/proteins in columns and samples are in rows). The rows of those data objects are ordered by the sample ids in the samples object. We also included tab-delimited txt files for all platforms and cancer types, named as platform\_cancer.txt files in the Data folder.

## Code

**Abstract**

For multi-layered Gaussian graphical models from multiple genomic platforms, we propose Bayesian node-wise selection (BANS) framework.

**Description**

The BANS R codes with an example data are at <https://github.com/MinJinHa/BANS>. BANS can be tested through the readme file.

All the R functions for the BANS framework are included in the chaingraph.R file in the Rpackage folder in this Supplementary files. The estimation of a mlGGM with q layers can be achieved from one GGM model for the first layer and q-1 two-layered mlGGMs and each of them is estimated from BANS method. Our BANS framework consists of two parts for graphical structure estimation and structured estimation of signs of the dependencies between and within layers. The main function for structure estimation and the structured estimation of the signs are ch.chaingraph and ch.chaingraph.str functions, respectively.

**Optional Information**

Usage

1. ch.chaingraph

* Title: Node-wise selection for (Two-layered) Gaussian graphical model with symmetric constraint on precision matrix (run jointly for a layer)
* Input

- v.ch: a vector for indices of the target layer

- v.pa: a vector for indices of the parent layer

- Y: nxp matrix that includes genomic measurements for p genes/proteins and n number of samples

- eta.prob: P(eta=1)

- gamma.prob: P(gamma=1)

- lambda and delta: hyper-parameters for variances and the undirected edges

- burnin.S: length of MCMC for burn-in

- inf.S: length of MCMC for inference

* Ouput
  + Gamma: |v.ch| x |v.pa| x inf.S dimensional array for between layer structure
  + eta: |v.ch| x |v.ch| x inf.S dimensional array for within layer structure
  + A: |v.ch| x |v.ch| x inf.S dimensional array for regression coefficients corresponding to edges within layer
  + B: |v.ch| x |v.pa| x inf.S dimensional array for regression coefficients corresponding to edges between layers.
  + kappa: inf.S x |v.ch| matrix for the inverse variance

1. v.chaingraph

* Title: Node-wise selection for (Two-layered) Gaussian graphical model with no symmetric constraint on precision matrix (run for a single node)
* Input
  + v: an integer for a target indice for node-wise regression
  + chlist: list for node numbers for layers
  + palist: list for node number for parent layers

- Y: nxp matrix that includes genomic measurements for p genes/proteins and n number of samples

- eta.prob: P(eta=1)

- gamma.prob: P(gamma=1)

- lambda and delta: hyper-parameters for variances and the undirected edges

- burnin.S: length of MCMC for burn-in

- inf.S: length of MCMC for inference

* Ouput
  + Gamma: |v.ch| x |v.pa| x inf.S dimensional array for between layer structure
  + eta: inf.S x |v.ch| dimensional matrix for undirected edges connected to v
  + A: inf.S x |v.ch| dimensional matrix for regression coefficients corresponding to undirected edges connected to v
  + B: |v.ch| x |v.pa| x inf.S dimensional array for regression coefficients corresponding to edges between layers.
  + kappa: inf.S x |v.ch| matrix for the inverse variance

1. ch.chaingraph.str

* Title: Structured estimation for (Two-layered) Gaussian graphical model
* Input

- v.ch: a vector for indices of the target layer

- v.pa: a vector for indices of the parent layer

- Y: nxp matrix that includes genomic measurements for p genes/proteins and n number of samples

- G: Adjacency matrix for the mlGGM across p variables.

- lambda and delta: hyper-parameters for variances and the undirected edges

- burnin.S: length of MCMC for burn-in

- inf.S: length of MCMC for inference

* Ouput
  + A: |v.ch| x |v.ch| x inf.S dimensional array for regression coefficients corresponding to edges within layer
  + B: |v.ch| x |v.pa| x inf.S dimensional array for regression coefficients corresponding to edges between layers.
  + kappa: inf.S x |v.ch| matrix for the inverse variance

Dependencies

The BANS R packages and other R codes requires several R packages under

R version 3.6.0:

library(Matrix) (Matrix\_1.2-17)

library(MASS) (MASS\_7.3-51.4 )

library(rms) (rms\_5.1-3.1)

library(Rcpp) (Rcpp\_1.0.1)

library(RcppArmadillo) (RcppArmadillo\_0.9.500.2.0)

library(RcppEigen) (RcppEigen\_0.3.3.5.0)

## Instructions for Use

**Reproducibility**

1. Simulation

* Table 1 and Figure 3

We compared the performance of BANS method with other joint estimation approaches, MRCE and CAPME. The simulation results in Table 1 and Figure 3 are generated from simul\_BANS.R for BANS, simul\_p200q10\_mrce.R for MRCE, and simul\_p200q10\_capme.R for CAPME in the Rcode/Simulation folder of this Supplementary materials.

* Figure 4

We also numerically evaluate the sign consistency of the estimated partial correlations for undirected edges within layers, and estimated coefficients for directed edges between layers using our structured sampling approach. The ROC curves in Figure 4 are based on the results from simul\_BANS.R for structure estimation, and simul\_BANS\_str.R for the structured estimations of signs in the Rcode/Simulation folder of this Supplementary materials.

1. Application

In the real data application, we applied BANS to 7 cancer types (see Data description section). All R codes that were used to make inference on the signed graph structures are included in Rcode/Application folder: fitting\_tumor type.R and fitting\_tumor type\_str.R for tumor type=COAD, LUAD, LUSC, OV, READ, SKCM, and UCEC.

* Figure 6 and Figure 7

UpSet plots are produced for each of the platform-level relations. Those UpSet plots are produced based on the R code, Rcode/Application/UpSetR.R by calling the estimated edge set information in results/edgeattr\_tumor type.txt files for tumor type= COAD, LUAD, LUSC, OV, READ, SKCM, and UCEC. The edge attribute files for each cancer includes nodes in the first two columns, edge type in the third column, posterior inclusion probabilities in the fourth column, signs of the edges in the sixth column, and the pathway memberships in the last column.

* Figure 8

The connectivity scores and the standard deviations for variability are computed based on Rcode/Application/ConnectivityScore.R. This also includes the R codes for generating heatmap and the barplots.

* Figure 9

The p53 networks across cancer types are generated from Cytoscape. The Cytoscape file is results/TP53.cys.

**Replication**

The BANS R codes with an example data are at <https://github.com/MinJinHa/BANS>. BANS can be tested through the readme file under the github page.

**Computation Time**

Using Linux server with 2.93 GHz Intel processor and 48 GB RAM, the figure below shows the computation time as the number of nodes increases for the two-layered Gaussian Graphical Models estimation using 5000 MCMC iterations. BANS-parallel took around 15 minutes to perform a node-wise regression for number of nodes from 20 to 100 across two-layers, while BANS took 37 hours to learn the p=100 network jointly. Thus, BANS-parallel is useful considering the gain in computational efficiency.

figure/BANS_time.pdf