ESTIMATION GRAPHICAL MODEL

24 OKTOBER, 2021

Mona Azadkia, Ahmad Ehyaei

ETH zürich

Contents

1 Introduction **5**

2 Review and background **7**

3 Methods **9**

4 Theoretical properties **11**

5 Simulation **13**

6 Application to TCGA data **15**

6.1 Download Data from TCGA Portal 15

7 Reference **19**

Chapter 1

Introduction

Chapter 2

Review and background

Chapter 3

Methods

Chapter 4

Theoretical properties

Chapter 5

Simulation

Chapter 6

Application to TCGA data

For download the RNA-seq breast cancer data from Cancer Genome Atlas (TCGA) database we use `TCGAbiolinks` package (Silva et al. 2016).

6.1 Download Data from TCGA Portal

TCGAbiolinks' purpose is to make it easier to access GDC data, build preprocessing methods, and provide multiple methods for analysis and visualization. For install package run below commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TCGAbiolinks")
```

Fist we download clinical data related to `TCGA-BRCA` study

```
library(TCGAbiolinks)

clinical_data <- TCGAbiolinks::getLinkedOmicsData(
  project = "TCGA-BRCA",
  dataset = "Clinical"
)
readr::write_rds(clinical_data,"data/clinical_brca.rds")
```

TCGA BRCA Clinical Data Header

attrib_name	TCGA.5L.AAT0	TCGA.5L.AAT1	TCGA.A1.A0SP
years_to_birth	42	63	40
Tumor_purity	0.6501	0.5553	0.6913
pathologic_stage	stageii	stageiv	stageii
pathology_T_stage	t2	t2	t2
pathology_N_stage	n0	n0	n0
pathology_M_stage	m0	m1	m0
histological_type	infiltratinglobularcarcinoma	infiltratinglobularcarcinoma	infiltratingductalcarcinoma
number_of_lymph_nodes	0	0	0
PAM50			Basal
ER.Status			
PR.Status			
HER2.Status			
gender	female	female	female
radiation_therapy	yes	no	
race	white	white	
ethnicity	hispanicorlatino	hispanicorlatino	nothispanicorlatino
Median_overall_survival	0	0	0
overall_survival	1477	1471	584
status	0	0	0
overallsurvival	1477,0	1471,0	584,0

The above table contains clinical data of patients. The name of each column is related to the patient's id.

The GDC mRNA quantification analysis pipeline measures gene level expression in HT-Seq raw read count, Fragments per Kilobase of transcript per Million mapped reads (FPKM), and FPKM-UQ (upper quartile normalization). For see more information see mRNA Analysis Pipeline GDC document.

Definition 6.1.1: HT Seq Normalization

RNA-Seq expression level read counts produced by HT-Seq are normalized using two similar methods: FPKM and FPKM-UQ. Normalized values should be used only within the context of the entire gene set. Users are encouraged to normalize raw read count values if a subset of genes is investigated.

FPKM:

The Fragments per Kilobase of transcript per Million mapped reads (FPKM) calculation normalizes read count by dividing it by the gene length and the total number of reads mapped to protein-coding genes.

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \tag{0.3}$$

Upper Quartile FPKM

The upper quartile FPKM (FPKM-UQ) is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample.

$$FPKM - UQ = \frac{RC_g * 10^9}{RC_{g75} * L}$$

RC_g: Number of reads mapped to the gene
RC_{pc}: Number of reads mapped to all protein-coding genes
RC_{g75}: The 75th percentile read count value for genes in the sample
L: Length of the gene in base pairs; Calculated as the sum of all exons in a gene
The read count is multiplied by a scalar 10^9 during normalization to account for the kilobase and 'million mapped reads' units.
mRNA Expression HT-Seq Normalization

```
measurements = c( "HTSeq - Counts", "HTSeq - FPKM", "HTSeq - FPKM-UQ")
for( m in measurements){
  query <- GDCquery(
    project = "TCGA-BRCA",
    data.category = "Transcriptome Profiling",
    data.type = "Gene Expression Quantification",
    workflow.type = "HTSeq - FPKM-UQ",
    barcode =  colnames(clinical_data)[-1] # List of all patients id.
  )

  GDCdownload(query,files.per.chunk = 100)
   GDCprepare(query = query, summarizedExperiment = FALSE) %>%
  readr::write_rds(sprintf("data/TCGA_BRCA_%s.rds",gsub(" - ","_",m)))
}
```

In the below table, the header of the HT-Seq count data can be found:

```
Sample_HTSeq_Counts = readr::read_rds("data/Sample_TCGA_BRCA_HTSeq_Counts.rds")

Sample_HTSeq_Counts %>%
  gt() %>%
  tab_header(title = "mRNA HT-Seq Count Data") %>%
   fmt_missing(columns = 1:4,missing_text = "")
```

mRNA HT-Seq Count Data

Gene	TCGA-3C-AAAU-01A-11R-A41B-07	TCGA-3C-AALI-01A-11R-A41B-07	TCGA-3C-AALJ-01A-31R-A41
ENSG00000225411.2	1	1	
ENSG00000236662.1	6	6	
ENSG00000244501.3	0	0	
ENSG00000242067.1	0	0	
ENSG00000140319.9	12112	3039	6
ENSG00000200112.1	0	0	
ENSG00000230496.1	0	0	
ENSG00000250261.1	2	0	
ENSG00000135838.12	647	720	
ENSG00000268906.1	0	0	

The `HTSeq-Counts` contains 60488 genes. To replicate (Yang et al. 2021) paper, we consider only genes in RNA-seq whose read number is more than 20 in at least 25% of the samples. We also use `HTSeq-FPKM-UQ`, which is more robust than `HTSeq-Counts` using the total number of reads per sample.

```
# Find Most Frequent Genes
SeqMat = HTSeq_Counts
SeqMat$Gene = NULL
SeqMat  =  as.matrix(SeqMat)
GeneList = HTSeq_Counts$Gene[rowSums(SeqMat>20)/ncol(SeqMat)>0.25]

# Filter HTSeq_FPKM_UQ to Most Frequent Genes
readr::read_rds("data/TCGA_BRCA_HTSeq_FPKM_UQ.rds") %>%
  filter(X1 %in% GeneList) %>%
  rename(Gene = X1 ) %>%
  readr::read_rds("data/RNASeq.rds")
```

Chapter 7

Reference

Silva, Tiago C, Colaprico, Antonio, Olsen, Catharina, D'Angelo, et al. 2016. "TCGA Workflow: Analyze Cancer Genomics and Epigenomics Data Using Bioconductor Packages." F1000Research 5.

Yang, Jenny, Yang Liu, Yufeng Liu, and Wei Sun. 2021. "Model Free Estimation of Graphical Model Using Gene Expression Data." The Annals of Applied Statistics 15 (1): 194–207.