Sparse CCA

Introduction

Let X denote the n*p matrix, represents mOTU profile in this report, and let Y denote the n * q matrix, represents PANSS. Canonical-correlation analysis (CCA) can find two vectors, u and v that maximize the correlation between Xu and Yv. On the one hand, we usually want to do feature selection in metagenomics analyze, which CCA can’t work on. On the other hand, CCA can’t get unique vectors if p or q exceeds n. Therefore, we choose sparse CCA instead of CCA to select variables(1).

In this report, I use R package PMA, which contains two sparse methods, lasso and fused lasso. I use lasso because the variables in Y is not ordered.

Data information

1) Sample information

171 samples collected from community and Research center for mental disorders, includes 81 healthy controls (HC) and 90 schizophrenes (SZ). In SZ, 49 patients are first-episode, and 41 patients are relapse. 360 mOTUs profile which represents bacteria abundance was calculated by analyzing the sequencing data. As the metagenomics data, mOTUs profile contains numerous zero. Additionally, mOTU represents species level.

2) PANSS

The Positive and Negative Syndrome Scale (PANSS) is a medical scale used for measuring symptom severity of patients with schizophrenia.

For the original one, the patient is rated from 1 to 7 on 30 different symptoms based on the interview as well as reports of primary care hospital workers. It has 3 catalogs, positive scale (p1-p7 in our research), negative scale (m1-m7 in our research) and general psychopathology scale (g1-g16 in our research). The minimum score of PANSS total is 30, and the maximum score of PANSS is 210. It should be noted that 1 rather 0 is the lowest score for each item.

For the modified one, depending on the clinical significance, cooperating parties combined some items together as new items based on their features, which turned 30 items into 27 items.

Procedure

- Choosing penalty x and penalty z

  The data I used: mOTU profile and original PANSS(30 items).

  There are a function CCA.permute in PMA, which selects penalty x and penalty z by six steps :

  1. The samples in X are randomly permuted nperms times, to obtain matrices $X*_1,X*_2,...$.
  2. Sparse CCA is run on each permuted data set $(X*_i,Z)$ to obtain factors $(u*_i, v*_i)$.
  3. Sparse CCA is run on the original data (X,Z) to obtain factors u and v.

4. Compute $c*_i=cor(X*_i u*_i,Z v*_i)$ and $c=cor(Xu,Zv)$.
5. Use Fisher's transformation to convert these correlations into random variables that are approximately normally distributed. Let Fisher(c) denote the Fisher transformation of c.
6. Compute a z-statistic for Fisher(c), using $(Fisher(c)-mean(Fisher(c*)))/sd(Fisher(c*))$. The larger the z-statistic, the "better" the corresponding tuning parameter value.

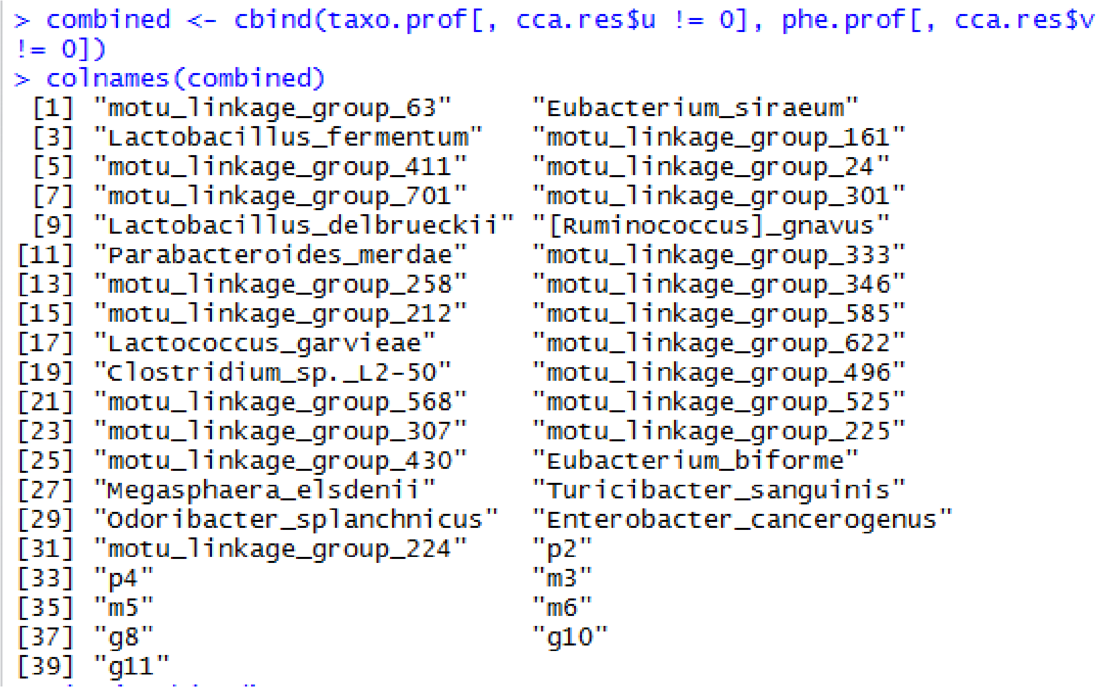
In this method, it just selects tuning parameter by the same order, so I add

for(i in 1:penaltyz) {
CCA.permute()}

to make more combinatory possibility. In fact, I split the vectors of tuning parameters and parallel process it to speed up which based on R package snow. In the end , I choose the penalty x and penalty z with largest z-statistic(Supplementary 1).

- sparse CCA

  According to the former results, the penalty x is 0.21, the penalty x is 0.4, the final u and v can be gotten by using CCA function. We got the following variables, which selected by sparse model.



```
> combined <- cbind(taxo.prof[, cca.res$u != 0], phe.prof[, cca.res$v
!= 0])
> colnames(combined)
 [1] "motu_linkage_group_63"     "Eubacterium_siraeum"
 [3] "Lactobacillus_fermentum"   "motu_linkage_group_161"
 [5] "motu_linkage_group_411"    "motu_linkage_group_24"
 [7] "motu_linkage_group_701"    "motu_linkage_group_301"
 [9] "Lactobacillus_delbrueckii" "[Ruminococcus]_gnavus"
[11] "Parabacteroides_merdae"    "motu_linkage_group_333"
[13] "motu_linkage_group_258"    "motu_linkage_group_346"
[15] "motu_linkage_group_212"    "motu_linkage_group_585"
[17] "Lactococcus_garvieae"      "motu_linkage_group_622"
[19] "Clostridium_sp._L2-50"     "motu_linkage_group_496"
[21] "motu_linkage_group_568"    "motu_linkage_group_525"
[23] "motu_linkage_group_307"    "motu_linkage_group_225"
[25] "motu_linkage_group_430"    "Eubacterium_biforme"
[27] "Megasphaera_elsdenii"      "Turicibacter_sanguinis"
[29] "Odoribacter_splanchnicus"  "Enterobacter_cancerogenus"
[31] "motu_linkage_group_224"    "p2"
[33] "p4"                        "m3"
[35] "m5"                        "m6"
[37] "g8"                        "g10"
[39] "g11"
```

Attention

All the R code and the supplementary 1 is available in https://github.com/juyanmei/Digging-into-metagenomic-data/tree/master/SZProjrct/metadata/sCCA
What's more, please download and open the supplementary 1 by using google chrome.

Reference

1. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515-34.