TCVS Package

Yicong Mao, Zhiwen Jiang, Tianying Wang, Yi-Juan Hu, and Xiang Zhan

August 27, 2024

1. Overview

The TCVS package implements the Tree-guided Compositional Variable Selection method to identify outcome-associated components within high-dimensional microbial compositional data. It accommodates compositional covariates (e.g., relative abundances of taxa). This method enhances variable selection by incorporating auxiliary knockoff copies of microbiome features, which are treated as noise, and utilizing hierarchical tree structures inherent to microbial taxa. By integrating these elements, TCVS refines the selection process to achieve more precise outcomes. This approach is particularly effective in enhancing selection accuracy, proving advantageous when the signal strength surpasses minimal thresholds. This dual strategy of leveraging tree structure and augmenting with knockoff features enables TCVS to provide robust insights into the complex interactions within microbiome data.

The core function of the TCVS package is **TCVS**, which is designed to identify OTUs (operational taxonomic units) associated with disease outcomes by performing statistical variable selection within a augmented regression model. The function **Group.lasso** computes the penalty term in the augmented problem incorporating auxiliary knockoff copies and tree structure with L2 norm (without square) for a given set of coefficients, groups, and a regularization parameter; The function **get.all.knockoff** generates model-X knockoff copies [Candes et al., 2018] for the covariates; The function **BIC_TCVS** calculates the Bayesian Information Criterion (BIC) for a given set of parameters using the CVXR package; The function **cvxr.result** performs an optimization to solve the augmented regression problem using the CVXR package.

Download the latest version of the package from GitHub at (https://github.com/Yicong1225/TCVS) to a local hard drive and install and load the package:

```
install.packages("TCVS_1.0.tar.gz", repos=NULL)
library(TCVS)
```

2. Examples

2.1 Generate the simulated data

We begin by loading the taxonomic tree structure (matrix form) for calculating the penalty term in the augmented problem. The first 2p rows of the taxonomic tree matrix **P** correspond to the l_1 penalty applied to the CLR (Centered Log-Ratio) covariates followed by their knockoff copies. The remaining rows correspond to the tree structure associated with the CLR covariates and their knockoff copies, following the same order of covariates first and knockoff copies second.

```
library(CVXR)
library(knockoff)
library(MASS)
library(vegan)
```

```
library(GUniFrac)
library(cluster)
library(dirmult)
data(P_60, package = "TCVS")
dim(P)
```

```
## [1] 248 120
```

```
data(throat.tree, package = "GUniFrac")
data(throat.otu.tab, package = "GUniFrac")
data(DirMultOutput)
```

We then generated a count data matrix $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times p}$ from a Dirichlet-Multinomial distribution, whose parameters were estimated from counts of 856 taxa in an upper-respiratory-tract microbiome dataset [Charlson et al., 2010, Chen et al., 2012]. To adjust for zero counts before CLR transformation, we added a pseudo-count of 0.5 to all counts before normalizing the count data matrix $\mathbf{W}$ into the compositional covariate matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ via $x_{ij} = w_{ij}/\sum_{l=1}^{p} w_{il}$.

```
n = 200
p = 60
setting = 1
fdr.normial = 0.05
method = "BIC"
loop_start_index = 1
seed = 2023
type = "Dirmult" # type to generate compositional data
normalizeMethod = "Rowsum"
beta.slack.factor = NULL
pseudocount = 0.5
maxlam = 0.1
minlam = 1e-7
nlam = 20
X <- get.all.OTU(n, p, trans = 0, type, normalizeMethod, pseudocount, seed = seed)
# clr transformation
Z <- get.all.OTU(n, p, trans = 2, type, normalizeMethod, pseudocount, seed = seed)
y <- get.all.y(Z, setting = setting, beta.slack.factor, seed = seed)
dim(Z)
```

```
## [1] 200  60
```

2.2 Perform the variable selection procedure

The core function that implemented the TCVS method is **TCVS**. We can use the TCVS function to select outcome-associated compositional components. The $\mathbf{X}$ (e.g., the OTU compositional matrix) should have rows corresponding to samples and columns corresponding to OTUs.

```
result.TCVS <- TCVS(
  X = X,
  Z = Z,
  y = y,
  P = P,
```

```
  method = method,
  maxlam,
  minlam,
  nlam,
  fdr = fdr.normial,
  seed = seed
)
selected.OTU = which(result.TCVS$S.TCVS!=0)
selected.OTU # selected OTUs using TCVS
```

```
##  [1]  1  2  4  5  8 10 11 12 18 19 20
```

References

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Emily S Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D Bushman, and Ronald G Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS one*, 5(12):e15216, 2010.

Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.