ecostructure - Grade of Membership Model and Visualization for ecological species abundance data

May 31, 2017

1 Introduction

The **ecostructure** package is an R package that replicates the statistical analysis in this paper, but its toolbox of functions is generic enough in handling and analyzing other species abundance data. The package provides functions for fitting the Grade of Membership (GoM) model, along with the visualization of model fit using Block Structure plot **????**. The package comes with the raw taxonomic data saved as an ExpressionSet object and provides a pipeline for reading and processing counts data for different dimensions of diversity, e.g. - phylogenetic, regional and functional, which serve as readymade input for the GoM model. This package is an upgraded version of the CountClust package due to Dey et al **??** for fitting GoM models on RNA-seq data

2 Installation

The package is available on Github and can be installed as follows

```
library(devtools)
install_github("kkdey/ecostructure")
```

Load the package as

```
library(ecostructure)
library(Biobase)
```

to use the GoM model, the user needs to install the **maptpx** package

```
library(devtools)
install_github("kkdey/maptpx")
```

3 Data Preparation

One can load the taxonomic data, together with the grid metadata ans species metadata as an ExpressionSet object as follows

```
data <- get(load(system.file("extdata", "HimalayanBirdsData.rda",
                             package = "ecostructure")))
taxonomic_counts <- exprs(data)
taxonomic_counts[1:5,1:5]
```

```
##                           A2 A3 A4 A6 A7
## Macropygia_unchall         0  0  0  0  0
## Streptopelia_chinensis     0  0  0  0  0
## Streptopelia_senegalensis  0  0  0  0  0
## Columba_pulchricollis      0  0  0  0  2
## Streptopelia_orientalis    0  0  0  0  0
```

The corresponding grid metadata can be read as

```
grid_metadata <- pData(phenoData(data))
head(grid_metadata)
```

```
##    Elevation North East WorE
## A2       198  27.0 92.9    E
## A3       734  27.0 92.4    E
## A4      1243  27.0 92.4    E
## A6      2629  27.1 92.5    E
## A7      2340  27.1 92.4    E
## A8       300  27.0 93.0    E
```

The species metadata can be read as follows

```
species_metadata <- pData(featureData(data))
head(species_metadata)
```

```
##                           bill_length bill_width bill_depth wing tarsus  mass
## Macropygia_unchall              11.08       4.26       4.97  198   26.3 168.0
## Streptopelia_chinensis          10.77       3.50       3.87  140   22.6 159.0
## Streptopelia_senegalensis        9.23       2.88       3.27  130   20.4  83.9
## Columba_pulchricollis           12.98       5.59       5.68  203   25.4 330.0
## Streptopelia_orientalis         10.88       4.09       3.90  192   25.5 233.0
## Chalcophaps_indica              11.77       3.61       4.42  151   26.3 121.0
```

Along with the taxonomic data and metadata, the package provides the phylogenetic tree data for the bird species as a **.tre** file that can be loaded as follows, using the package **ape**.

```
phylo_tree <- ape::read.tree(system.file("extdata", "AllHim_Mar_27_2015.tre",
                             package = "ecostructure"))
```

```
phylo_tree
```

```
##
## Phylogenetic tree with 589 tips and 588 internal nodes.
##
## Tip labels:
##  Macropygia_unchall, Streptopelia_chinensis, Streptopelia_senegalensis, Columba_hodgson
##
## Rooted; includes branch lengths.
```

The shape files for the regional motif analysis can be loaded as follows

```
shp_file <- ape::read.tree(system.file("extdata", "AllHim_Mar_27_2015.tre",
                                package = "ecostructure"))
shp_file
```

```
##
## Phylogenetic tree with 589 tips and 588 internal nodes.
##
## Tip labels:
##  Macropygia_unchall, Streptopelia_chinensis, Streptopelia_senegalensis, Columba_hodgson
##
## Rooted; includes branch lengths.
```

4 Grade of Membership Model and Visualization

Here we illustrate how one can fit the Grade of Membership model and perform the visualization of the model fit using the Block Structure plot. Here we present a case study with number of clusters chosen to be between 2 and 4.

```
elevation_metadata=grid_metadata$Elevation;
east_west_dir = grid_metadata$WorE;
gom_fit <- CountClust::FitGoM(t(taxonomic_counts), K=2:4, tol=0.1)
```

```
##
## Estimating on a 38 document collection.
## Fit and Bayes Factor Estimation for K = 2
## log posterior increase: 2703.9, 19.3, 42.2, 0.4, done.
## log BF( 2 ) = 1950.27
##
## Estimating on a 38 document collection.
## Fit and Bayes Factor Estimation for K = 3
## log posterior increase: 3213.1, 153.7, 2.6, done.
## log BF( 3 ) = 2017.37
##
```

```
## Estimating on a 38 document collection.
## Fit and Bayes Factor Estimation for K = 4
## log posterior increase: 4543.2, 8.7, 12.3, 1.6, 11.7, 0.6, 0.1, 0.5, 1.1, 4, 25.8, 23.8
## log BF( 4 ) = 2150.14
```

gom_fit is a list of size 3, with each component representing the model fit for the cluster k, varying from 2 to 4. The two main components of the model fit are the membership proportion matrix ω, given by gom_fit[[k]]$omega and the motif matrix gom_fit[[k]]$theta. Examples for $K = 2$ are

```
omega <- gom_fit[[2]]$omega
head(omega)
```

```
##          topic
## document        1        2        3
##       A2 0.993020 0.006869 1.11e-04
##       A3 0.998389 0.000790 8.21e-04
##       A4 0.863243 0.106883 2.99e-02
##       A6 0.000173 0.000299 1.00e+00
##       A7 0.000087 0.000387 1.00e+00
##       A8 0.982291 0.017615 9.38e-05
```

```
rowSums(omega)
```

```
## A2 A3 A4 A6 A7 A8 B1 B2 B3 B4 B5 D1 D3 G1 J1 J2 J4 J5 J6 K1 K2 K4 K5 K6 L1 M1
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## M2 M3 M4 N1 N2 N3 S1 U3 U4 MA U1 U2
##  1  1  1  1  1  1  1  1  1  1  1  1
```

```
theta <- gom_fit[[2]]$theta
head(theta)
```

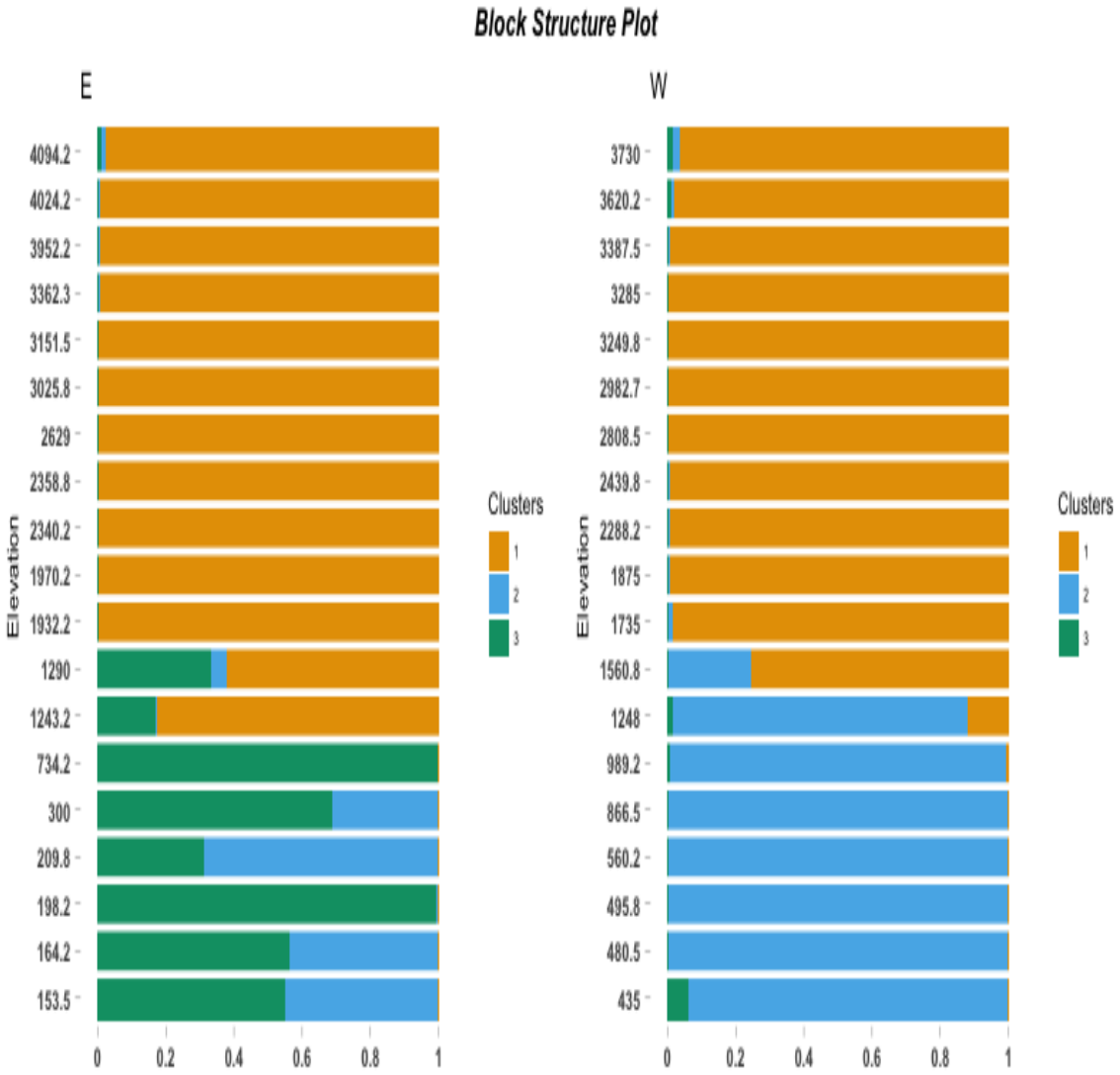
```
##                             topic
## phrase                             1        2        3
##   Macropygia_unchall        5.11e-07 9.32e-07 2.44e-03
##   Streptopelia_chinensis    5.11e-07 1.19e-02 6.70e-07
##   Streptopelia_senegalensis 5.11e-07 5.09e-03 6.70e-07
##   Columba_pulchricollis     5.10e-07 9.32e-07 1.22e-03
##   Streptopelia_orientalis   2.24e-03 1.63e-02 1.31e-06
##   Chalcophaps_indica        2.79e-03 1.04e-06 6.96e-07
```

```
colSums(theta)
```

```
## 1 2 3
## 1 1 1
```

Using the grid metadata, we provide a Block Structure Plot representation of the membership proportion matrix. In a block Structure plot representation, one metadata is used for forming blocks (here the East/West direction) and in each block, the the samples (along the rows of the Structure plot) are arranged by a second metadata (say Elevation).

```
BlockStructure(omega, blocker_metadata = east_west_dir,
               order_metadata = elevation_metadata,
               yaxis_label = "Elevation",
               levels_decreasing = FALSE)
```



5 Processing motif data