Cluster Analysis based on beta or gamma Diversity

Jari Oksanen

August 27, 2024

Contents

1 Introduction

The **natto** package includes function `diverclust` that introduces a potentially new method of beta and gamma diversity clustering. The idea of this new method is obvious and I have long assumed that such a method must have been invented previously. However, I have failed to find such a method, and therefore I added the function in **natto**. This document describes the method as implemented in **natto** for two purposes: to explain what was done in **natto** and to help finding its predecessors.

2 The Method

2.1 The Stages of Clustering

The main steps are:

1. Initially estimate diversity of all sampling units (§2.2).
2. First evaluate the *increase* in diversity when you pool together two sampling units (optionally with equalization, §2.3). The increase in diversity in pooling is conventionally called as beta diversity.
3. Select the smallest of pooled beta diversity values and make it a cluster.
4. Pool the sampling units of the new merged cluster by summing up abundance values (optionally with equalization §2.3), and re-evaluate its beta diversities with all other units.
5. Go back to step 3 until all sampling units are merged.

2.2 Measurement of Diversity

Diversity indices are based on proportional abundances p_j of for species $j = 1 \dots S$, where S is the number of species. The species proportions are normally found from community matrix $\{x_{ij}\}$ by dividing by the sum of row sums $p_j = x_{ij} / \sum_{i=1}^{N} x_{ij}$. The `diverclust` function uses **vegan** function `renyi` to evaluate any Rényi diversity or the corresponding Hill number. Rényi diversity of order a is (Hill, 1973):

$$H_a = \frac{1}{1-a} \log \sum_{j=1}^{S} p_j^a \,, \tag{1}$$

and the corresponding Hill number is $N_a = \exp(H_a)$. Many common diversity indices are special cases of Hill numbers: $N_0 = S$ is the number of species, $N_1 = \exp(H')$ is exponent of the Shannon diversity, $N_2 = 1/\sum_{j=1}^{S} p_j^2$ is the Simpson diversity, and $N_\infty = 1/(\max p_j)$ is the Berger-Parker index. The corresponding Rényi diversities

are $H_0 = \log(S)$, $H_1 = H'$, $H_2 = -\log(\sum p_j^2)$, and $H_\infty = -\log(\max p_j)$.

The beta diversity is defined as an increase of diversity when pooling sampling units. Pooling is performed by summing up abundance values of species, and the beta diversity β for a cluster of M sampling units is defined as a difference of the pooled gamma diversity and mean of alpha diversities $\beta = \gamma - \bar{\alpha}$:

$$\beta_a = \underbrace{H_a\left(\sum_{i=1}^{M} x_{ij}\right)}_{\gamma} - \underbrace{\frac{1}{M}\sum_{i=1}^{M} H_a(x_{ij})}_{\bar{\alpha}}, \qquad (2)$$

Alternatively we can use Hill numbers N_a in place of H_a. The equation implies additive partitioning of diversity. However, Rényi diversities are on log scale which translates a multiplicative model into an additive model. If you use Hill numbers in eq. 2, the model is truly additive (cf. eqs. 11 and 12).

2.3 Equalizing Observations

The total pooled diversity (γ) in eq. 2 includes both the within-unit diversity (α) and between-unit differences (β). This requires that pooled diversity really is additive: it maintains the within-unit diversity and adds between-unit differences. This seems to be the case with Rényi diversities, but not necessarily for the Hill numbers beyond $N_0 = S$ (species richness). However, if the sampling units are equalized, the additivity is greatly improved. Equalization scales units with different total abundances to a more similar magnitude for pooling, but does not influence their alpha diversities. The suggested equalization is dependent on the scale a of the Rényi diversity, and all values for a unit are divided by weight

$$w_i = \left(\sum_{j=1}^{S} x_{ij}^a\right)^{1/a}. \qquad (3)$$

For $a = 1$ (Shannon diversity) w scales to unit sum, for $a = 2$ (Simpson diversity) w scales to unit sum of squares (norm), and for $a = \infty$ (Berger-Parker diversity) w is the maximum for each sampling unit. No equalization is needed for $a = 0$ (species richness).

2.4 Clustering Based on gamma Diversity

If we omit the $\bar{\alpha}$ term for average alpha diversity in eq. 2, we can base the clustering on total or gamma diversity. It may be difficult to see what would be the utility of such a clustering, but perhaps it could be used to form low-diversity classes that differ from each other.

2.5 Implementation

Function `diverclust` implements beta and gamma diversity clustering as described above. The following options can be used to modify its behaviour:

- The scale a of Rényi diversity (eq. 1) can be given.
- The equalization of eq. 3 can be turned off or on.
- beta or gamma diversity clustering can be selected.
- Either Rényi diversities or Hill numbers can be selected.

The following example performs diversity clustering with defaults: it uses beta diversity based on Rényi index H_1 (Shannon diversity) and equalizes sample plots, and displays the dendrogram (Fig. 1).

```
data(BCI)
## row names by most abundant species
colnames(BCI) <- make.cepnames(colnames(BCI))
dom <- colnames(BCI)[apply(BCI, 1, which.max)]
rownames(BCI) <- paste0(dom, 1:50)
## diversity clustering
cl <- diverclust(BCI, trace=FALSE)
plot(cl,hang = -1, cex=0.8)
```

3 Other Methods

One reason for writing this vignette was to find out if the `diverclust` function re-invents an existing method. Information Analysis is based on very similar reasoning as the diversity clustering (Williams *et al.*, 1966; Lance and Williams, 1966). To compare it against `diverclust`, it was implemented as function `infodist` in the **natto** package.

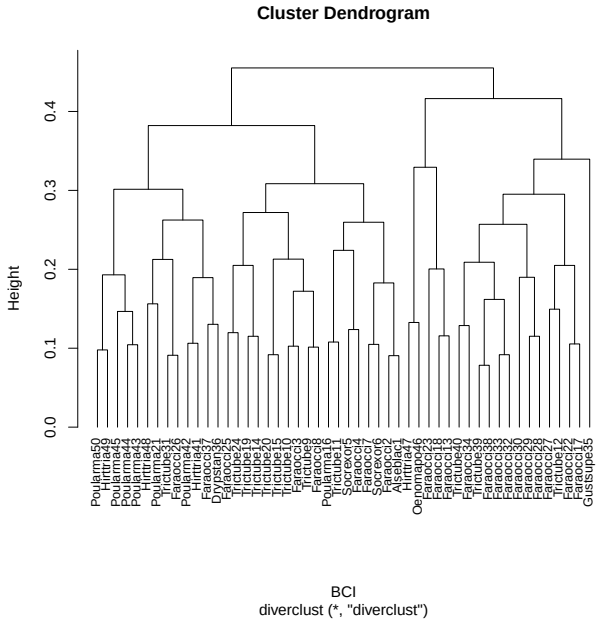

Figure 1: Diversity clustering of Barro Colorado Island forests based on H_1. The sampling units are named by their most abundant tree species.

3.1 Clustering Based on Information Analysis

Information analysis works on binary community data matrix. The information content I of a cluster of M sampling units is defined as (Williams *et al.*, 1966; Lance and Williams, 1966):

$$I = SM \log M - \sum_{j=1}^{S} Mq_j \log Mq_j + M(1-q_j) \log M(1-q_j), \quad (4)$$

where q_j is the relative frequency of species j in the M sampling units of the cluster. The contribution to I is 0 for species absent from the cluster ($q = 0$) and for species present on every unit in the cluster ($q = 1$), and it is maximal for species with frequency $q = 0.5$. The clustering minimizes the information criterion I and therefore it tries to produce clusters where species are either absent or nearly absent, or constant or nearly so. This often produces clusters that are easy to interpret in floristic terms.

The clusters are formed similarly as in `diverclust`: two sampling units or clusters are pooled, the species frequencies are recalculated, and the information content with the pooled group and all other groups are re-evaluated. However, the group selected for merging is not not the one with lowest I, but the group that gives the lowest increase ΔI. The information content of single sampling units is $I = 0$, but formed clusters have positive values of I, and the information values of the members of the cluster are subtracted from the value of the cluster, and the difference ΔI is used as the criterion of clustering. The merge still happens at the level of pooled new unit I, and the clusters are not formed in the order of their merge heights. This conflicts with R conventions, and the `infoclust` function updates the merge table to correspond to the merge heights.

Williams *et al.* (1966) and Lance and Williams (1966) describe their method very briefly, and the current implementation is **natto** is based on the worked example of Legendre and Legendre (2012). The use is similar as for `diverclust`:

```
cli <- infoclust(BCI)
plot(cli, hang=-1, cex=0.8)
```

The clusters are often very compact (Fig. 2), and results differ from the diversity clustering. Fig. 1 was based on quantitative data, but clusterings are often very different with binary data as well. In this case, diversity clustering of binary data with Simpson index (N_2) and equalization gave most similar results to information clustering (Fig. 3).

```
library(dendextend) # tanglegram
cl2 <- diverclust(decostand(BCI, "pa"),
   renyi=2, hill=TRUE, equalize=TRUE, trace=FALSE)
tanglegram(untangle(as.dendrogram(cli),
   as.dendrogram(cl2), method="step2side"),
   main_left="Information", main_right="Rényi 2")
```

3.2 Dissimilarity Indices

The diversity clustering methods can also be expressed as dissimilarity measures for two pooled sampling units. Such dissimilarities do not produce the diversity clustering, because dissimilarities between clusters cannot be found from dissimilarities between sampling units, but the original data

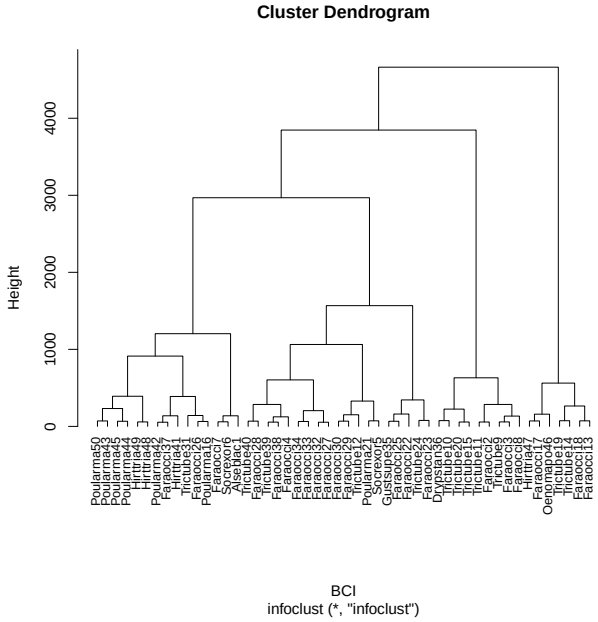

Figure 2: Information Analysis Clustering of the Barro Colorado Island forests.

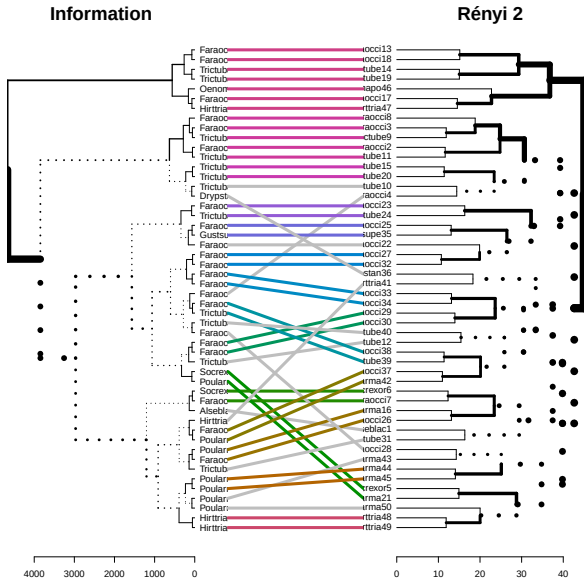

Figure 3: Information analysis (Fig. 2) against diversity clustering of Simpson Index (Hill number N_2) of binary data with equalization of sample plots. The dendrograms were rearranged to minimize entanglement.

must be pooled by clusters, and dissimilarites re-evaluated from the updated community data. However, it may be instructive to compare these indices with common dissimilarity measures. In this chapter we see how the diversity measures can be expressed as dissimilarities with binary data.

The **natto** package includes function `diverdist` that evaluates the pairwise diversity dissimilarities, and is used in the first step of diversity clustering to select first merged sampling units.

The diversity clustering and Rényi diversities are principally designed for quantitative data. For presence-absence dissimilarity indices we analyse binary data. Binary data defines a community with maximum equitability, so that all Rényi indices will be $N = \log S$ and Hill numbers $H = S$ irrespective of the Rényi scale. Average alpha diversity with Rényi ($\bar{\alpha}_H$) or Hill ($\bar{\alpha}_N$) indices for two sampling units eachs with S_1 and S_2 species is

$$\bar{\alpha}_H = \tfrac{1}{2}(\log S_1 + \log S_2) \tag{5}$$

$$\bar{\alpha}_N = \tfrac{1}{2}(S_1 + S_2)\,. \tag{6}$$

In two sampling units of S_1 and S_2 species and J shared species, there will be $S_1 + S_2 - 2J$ species that occur only one of the units and each at proportion $p = \frac{1}{S_1+S_2}$ and J species that occur in both sampling units at proportion $p = \frac{2}{S_1+S_2}$. The general equation (2) of distance based on Rényi beta diversity between two binary sampling units is:

$$d(H_a) = \frac{1}{1-a}[\log(S_1 + S_2 + (2^a - 2)J) - a\log(S_1 + S_2)] - \bar{\alpha}_H \tag{7}$$

and the corresponding formula for Hill numbers is:

$$d(N_a) = \left[\frac{S_1 + S_2 + (2^a - 2)J}{(S_1 + S_2)^a}\right]^{\frac{1}{1-a}} - \bar{\alpha}_N\,. \tag{8}$$

Some special cases simplify into more compact forms:

$$d(H_0) = \log(S_1 + S_2 - J) - \bar{\alpha}_H \quad (9)$$

$$= \log\left(\frac{A + B - J}{\sqrt{S_1 S_2}}\right) \quad (10)$$

$$d(N_0) = \tfrac{1}{2}(S_1 + S_2 - 2J) \quad (11)$$

$$= \bar{\alpha}_N - J \quad (12)$$

$$d(H_1) = \log(S_1 + S_2) - \frac{J \log 4}{S_1 + S_2} - \bar{\alpha}_H \quad (13)$$

$$d(H_2) = 2\log(S_1 + S_2) - \log(S_1 + S_2 + 2J) - \bar{\alpha}_H \quad (14)$$

$$d(N_2) = \frac{(S_1 + S_2)^2}{S_1 + S_2 + 2J} - \bar{\alpha}_N \quad (15)$$

$$d(H_\infty) = \log(S_1 + S_2) - \log 2 - \bar{\alpha}_H \quad (16)$$

$$d(N_\infty) = 0\,. \quad (17)$$

These may be regarded as "new" dissimilarity measures, although there hardly is a deficit of dissimilarity indices. Simplest case is $d(N_0)$ which is only half of the number of non-shared species, or half of the squared Euclidean distance between binary vectors. Most importantly, these are special cases for binary data and two pooled sampling units. Although the indices can be calculated for binary data, they really are intended for quantitative data. More importantly, these indices are only used for comparing a pair of unmerged sites, and they do not apply to comparisons involving clusters.

The implicit dissimilarity measure in information analysis is (Williams *et al.*, 1966; Lance and Williams, 1966):

$$d(I) = (S_1 + S_2 - 2J)\log(4)\,. \quad (18)$$

This is $d(N_0)$ with different multiplier. However, the similarity to $d(N_0)$ disappears when more than two sampling units are compared.

Dissimilarities based on diversity or information have been sometimes suggested. Koleff *et al.* (2003) suggest the following that they ascribe to Routledge (1984):

$$d(I) = \log(S_1 + S_2) - \frac{J \log 4}{S_1 + S_2} - \frac{S_1}{S_1 + S_2}\log S_1 - \frac{S_2}{S_1 + S_2}\log S_2\,. \quad (19)$$

The formulation was based on Koleff *et al.* (2003) who adapted it to binary data, and it was further rearranged to emphasize its resemblance to our $d(H_1)$ (eq. 13). The only difference is that $d(H_1)$ uses unweighted average alpha diversity $\bar{\alpha}_H$ (eq. 5), whereas eq. 19 weights sampling units by their species richness values. Using equalized pooling of binary data produces dissimilarities that are even more similar to this index.

References

Hill MO (1973). "Diversity and evenness: a unifying notation and its consequences." *Ecology*, **54**, 427–473.

Koleff P, Gaston KJ, Lennon JJ (2003). "Measuring beta diversity for presence-absence data." *Journal of Animal Ecology*, **72**, 367–382.

Lance GN, Williams WT (1966). "Computer programs for hierarchical polythetic classification ("similarity analyses")." *Computer Journal*, **9**, 60–64.

Legendre P, Legendre L (2012). *Numerical Ecology*. 3 edition. Elsevier.

Routledge RD (1984). "Estimating ecological components of biodiversity." *Oikos*, **42**, 23–29.

Williams WT, Lambert JM, Lance GN (1966). "Multivariate methods in plant ecology. V. Similarity analyses and information analysis." *Journal of Ecology*, **54**, 427–445.