Rao Standardization

Jari Oksanen

October 26, 2024

1 Introduction

Rao (1982) introduced perhaps the most often used measure of functional or phylogenetic diversity, Rao quadratic entropy. The quadratic entropy is a generalization of Simpson or Gini-Simpson index of diversity that takes into account the non-independence of species. If species are all functionally (in traits) or phylogenetically related, the community is less diverse than a community with dissimilar species. The dissimilarity measure weights taxa by their similarity, and two communities sharing no species can be similar to each other if the species are similar in traits or related in traits.

There are several `R` functions that implement Rao's quadratic diversity and some that also implement the related dissimilarity. Perhaps the most well-known are those in the **ade4** package (functions `divc` and `disc`, both by Sandrine Pavoine). I have also implemented them as `qrao` and `distrao` in **natto**. These are single-purpose functions to perform exactly this task. This documents inspects the possibility of implementing Rao's measures as a general standardization of the community data. This would allow implementing Rao's method with minimal intrusion in **vegan** diversity and dissimilarity functions, and also in general data analysis, such as in redundancy analysis.

2 Rao's Diversity and Distance

Rao (1982) defined quadratic entropy H for a community as

$$H = \sum_{j=1}^{S}\sum_{k=1}^{S} p_j p_k d_{jk}\,, \tag{1}$$

where p is the proportion of species j and k of the community, and S is the number of species, so that $\sum_{i=1}^{S} p_i = 1$, and d is the dissimilarity among species. In this assay we only study the case where dissimilarities are bounded in $(0, 1)$ where 1 means completely different and independent species, and 0 means identical species. The dissimilarity matrix has zero diagonal: species is always identical to itself and does not contribute to the quadratic entropy. If all species are completely and equally different, matrix $\mathbf{D} = \{d_{jk}\}$ has zero diagonal and the off-diagonal elements are ones. In that case Simpson-Gini diversity index is $1 - H$.

Rao (1982) defined a dissimilarity index that was based on eq. 1, but the species indexed there as j and k come from different communities, and we have their cross product. With this model, the quadratic entropies (diversities) for communities i and j are denoted as H_i and H_j and their cross product as H_{ij}. The distance between two communities is defined as (Rao, 1982, eq. 2.1.3)

$$\delta_{ij} = H_{ij} - \tfrac{1}{2}(H_i + H_j)\,. \tag{2}$$

Rao (1982) calls this Jensen distance, and we see later that it is actually one half of squared Euclidean distance.

Rao (1982) does not use matrix notation, but matrix $\mathbf{H} = \{H_{ij}\}$ can be found as

$$\mathbf{H} = \mathbf{PDP}'\,, \tag{3}$$

where $\mathbf{P}$ is a matrix of proportions of species in sites and $\mathbf{D}$ is the among species dissimilarity matrix. The diagonal of $\mathbf{H}$ gives the quadratic entropies H_i and H_j, and the off-diagonal elements the cross products H_{ij}.

3 Incorporation of Rao's Method in vegan

In this section we study how to implement Rao's method in **vegan** in a non-intrusive way. We study

specifically **vegan** function `designdist` that is the most flexible and configurable dissimilarity function in **vegan**. Dissimilarity functions need both the Rao entropy and cross product terms, or diagonal and off-diagonal elements of H (eq. 2). If we can implement dissimilarity measures, we can also implement Rao entropy.

In `designdist` we make a difference between *terms* (Table 1) and *formulae* (Table 2) using these terms. Terms can be binary, quadratic or first degree terms using minima between variables. The quadratic terms are estimated through matrix multiplication $\mathbf{XX}'$. With binary data, the multiplication gives the binary parameters of number of species in each community in the diagonal, and the numbers of shared species between communities in the off-diagonal elements. We can incorporate species dissimilarities in matrix multiplication similarly as in eq. 3. However, for standard dissimilarity measures we must have similarities $\mathbf{R} = 1 - \mathbf{D}$ instead of dissimilarities. For $\mathbf{D}$ bounded in $(0, 1)$, the diagonal elements of $\mathbf{R}$ are 1, and off-diagonal elements are complements of dissimilarity $\{r_{jk}\} = \{1 - d_{jk}\}$, and Rao style quadratic terms will be given by

$$\mathbf{Q} = \mathbf{PRP}', \tag{4}$$

and $\mathbf{Q} = 1 - \mathbf{H}$. The Jensen distance of eq. 2 will be expressed in the form given in Table 2, e.g., with reversal of signs. The complement $1-q_i$ of the cross product matrix is Rao's quadratic entropy, and-Simpson's diversity evaluated with eq. 4 is Rao's quadratic entropy.

The form of eq. 4 cannot be applied for the minimum terms (Table 1), and this would limit applying Rao methods to quadratic and binary terms. However, if we standardize data by

$$\mathbf{Z} = \mathbf{PR}^{1/2}, \tag{5}$$

then $\mathbf{Q} = \mathbf{ZZ}'$. Matrix $\mathbf{Z}$ has same rows and columns as data matrix $\mathbf{P}$, and it can be used to estimate the minimum terms of Table 1.

Matrix square root is not square root of its elements $\mathbf{D} \neq \{\sqrt{d_{jk}}\}$, but it is defined by matrix multiplication $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$. The matrix square root is easiest to find via eigen decomposition: If

$$\mathbf{R} = \mathbf{U\Lambda U}' \tag{6}$$

then

$$\mathbf{R}^{1/2} = \mathbf{U\Lambda}^{1/2}\mathbf{U}' \tag{7}$$

where $\mathbf{U}$ are orthonormal eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, and $\mathbf{\Lambda}^{1/2} = \mathrm{diag}(\sqrt{\lambda_i})$. For real valued matrix squareroot, all eigenvalues must be non-negative and the matrix $\mathbf{R}$ must be positive semidefinite. This is true of all correlation and covariance matrices, and it seems to be true of $\mathbf{R} = 1 - \mathbf{D}$ when $\mathbf{D}$ is Euclidean, or eigenvalues of $-\frac{1}{2}\bar{\mathbf{D}}^2$ are non-negative (Gower, 1966), where $\bar{\mathbf{D}}$ is the double-centred dissimilarity matrix. This is the same condition as for valid dissimilarities in Rao's quadratic entropy (Pavoine *et al.*, 2005).

I used notation $\mathbf{R}$ for similarities, because they are correlation-like and matrix takes the role of correlation structure in linear modelling (Pinheiro and Bates, 2000).

4 Implementation and Proof of the Concept

In this section I give the implementation of Rao standardization (eq. 5) and demonstrate that standardized data can be used to find Rao's quadratic entropy as Simpson diversity of standardized data, and Jensen distance from the Euclidean distance of standardized data.

Rao's quadratic entropy is estimated with function `qrao` and Rao's distance with `distrao`, both in the **natto** package. For other analyses I use function in base `R` and **vegan** package. I analyse Terschelling dune meadow vegetation, and I use coalescence ages from inferred phylogeny for among species dissimilarities (Fig. 1). The phylogeny is an ultrametric tree which guarantees that $\mathbf{D}$ is Euclidean and $\mathbf{R}$ positive semidefinite (Pavoine *et al.*, 2005).

Function for Rao standardization function is

```
raostand <-
function (x, d, propx = TRUE, dmax)
{
    TOL <- sqrt(.Machine$double.eps)
    x <- as.matrix(x)
    if (propx)
        x <- decostand(x, "tot")
    dn <- attr(x, "dimnames")
    d <- as.dist(d)
    if (anyNA(d))
```

Table 1: Terms used in defining formulae for dissimilarity functions.

	Binary terms	Quadratic terms	Minimum terms
J	No. of shared species	$\sum_{i=1}^{S} x_{ij}x_{ik}$	$\sum_{i=1}^{S} \min(x_{ij}, x_{ik})$
A	No. of species in j	$\sum_{i=1}^{S} x_{ij}^2$	$\sum_{i=1}^{S} x_{ij}$
B	No. of species in k	$\sum_{i=1}^{S} x_{ik}^2$	$\sum_{i=1}^{S} x_{ik}$

Table 2: Formulae and common names for some popular dissimilarity measures using terms defined in Table 1.

	Binary terms	Quadratic terms	Minimum terms
$A+B-2J$	No. of different species	Squared Euclidean	Manhattan
$\frac{1}{2}(A+B)-J$	No name	Jensen	No name
$\frac{A+B-2J}{A+B}$	Sørensen	No name	Bray-Curtis
$\frac{A+B-2J}{A+B-J}$	Jaccard	Similarity Ratio	Quantitative Jaccard
$1-\frac{J}{\sqrt{AB}}$	Ochiai	Cosine complement	No name

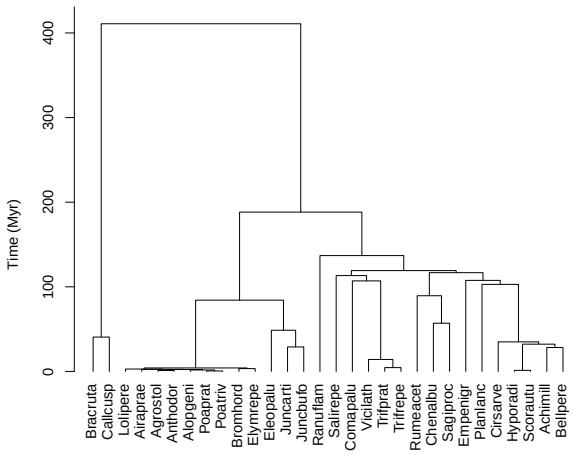

Figure 1: Dated phylogenetic tree of species in Dutch Dune meadows

```
        stop("missing values are not accepted")
    if (!missing(dmax)) {
        d <- d/dmax
        if (max(d) > 1)
            d[d > 1] <- 1
    }
    else if (max(d) > 1)
        d <- d/max(d)
    d <- as.matrix(d)
    d <- 1 - d
    if (any(abs(diag(d) - 1) > TOL))
        stop("'d' is not a valid dissimilarity object")
    e <- eigen(d)
    if (any(e$values < -TOL))
      stop("dissimilarities 'd' do not define Euclidean transformation")
    k <- e$values > TOL
    vec <- e$vectors[, k, drop = FALSE]
    ev <- e$values[k]
    d <- vec %*% (sqrt(ev) * t(vec))
    x <- x %*% d
    if (any(abs(x) < TOL))
        x[abs(x) < TOL] <- 0
    attr(x, "dimnames") <- dn
    x
}
```

Most of the functions take care that species distances are bounded in $(0, 1)$, and in diversity calculations, row sums are 1. However, we are explicit here and take care of this manually:

```
D <- dune.phylodis/max(dune.phylodis)
P <- as.matrix(decostand(dune, "tot"))
Z <- raostand(P, D)
```

The standardized matrix $\mathbf{Z} = \mathbf{P}\mathbf{R}^{1/2}$ has the following properties:

Standardization was applied to matrix `P` where each row sums up to unity. The Simpson index can be found directly from

```
tol <- sqrt(.Machine$double.eps)
all(abs(1 - rowSums(P^2) -
            diversity(dune, "simpson")) < tol)

## [1] TRUE
```

Rao's quadratic entropy can be found from $\mathbf{Z}$ similarly as the Simpson index:

$$1 - \sum_{j=1}^{S} z_i^2 = \sum_{j=1}^{S}\sum_{k=1}^{S} p_j p_k d_{jk} \tag{8}$$

```
all(abs(1 - rowSums(Z^2) - qrao(dune, D)) < tol)

## [1] TRUE
```

Alternatively, Rao's quadratic entropy can be found from the crossproduct of standardized data $\mathbf{Z}$:

$$1 - \mathrm{diag}(\mathbf{Z}\mathbf{Z}') = \sum_{j=1}^{S}\sum_{k=1}^{S} p_j p_k d_{jk} \tag{9}$$

```
all(abs(1 - diag(tcrossprod(Z)) -
            qrao(dune, D)) < tol)

## [1] TRUE
```

The Jensen distances δ (eq. 2) can be found from the elements $\{g_{ij}\} = \mathbf{G} = \mathbf{Z}\mathbf{Z}'$ with reversal of signs

$$\tfrac{1}{2}(g_{ii} + g_{jj}) - g_{ij} = H_{ij} - \tfrac{1}{2}(H_i + H_j) \tag{10}$$

and the Eulidean distances of $\mathbf{Z}$ are $(2\delta_{ij})^{1/2}$:

```
all(abs(distrao(dune, D) - dist(Z)^2/2) < tol)

## [1] TRUE
```

All postulated equalities were true which shows that the suggested standardization indeed works. We can add Rao's method to any **vegan** function with minimal changes in the code. In dissimilarity functions (`vegdist`, `designdist`) the input data must be tranformed with `raostand`, but the data need not have unit row totals. This allows using the Rao method also with minimum terms, such as with the popular Bray-Curtis and Jaccard indices. It is possible to add Rao's quadratic entropy and its species equivalent into `diversity` by applying `raostand` after transforming rows to unit totals. However, the standardization does work with Shannon index.

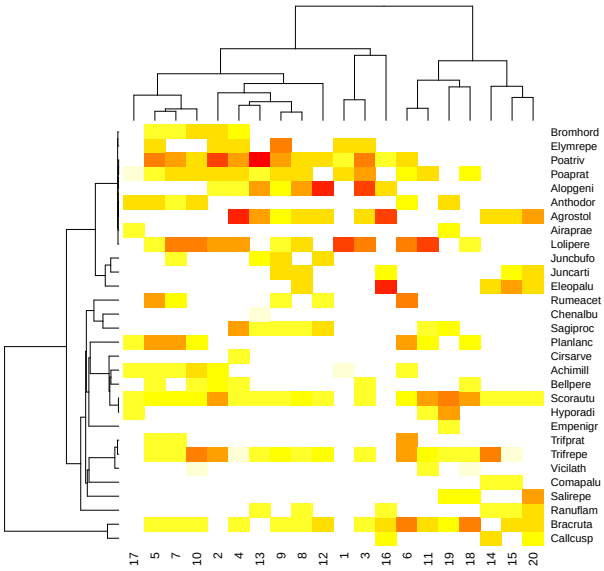

Figure 2: Data table ordered by species phylogeny and clustering based on phylogenetic Bray-Curtis dissimilarity

5 Extension and Example

In this section we apply Rao standardization for analyses based on phylogenetic dissimilarities. The data can be tabulated using clustering based on phylogenetic Bray-Curtis dissimilarity (Fig. 2):

```
tabasco(dune, hclust(vegdist(Z)), hclust(D))
```

Unconstrained ordination based on phylogenetic Bray-Curtis ordination can be performed with NMDS (Fig. 3):

```
ord <- metaMDS(Z, trace=FALSE)
ord

##
```

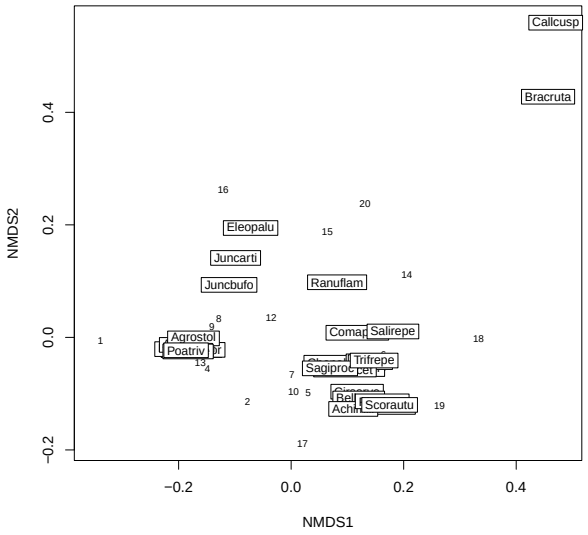

Figure 3: NMDS based on phylogenetic Bray-Curtis dissimilarity

```
## Call:
## metaMDS(comm = Z, trace = FALSE)
##
## global Multidimensional Scaling using monoMDS
##
## Data:     Z
## Distance: bray
##
## Dimensions: 2
## Stress:     0.05941756
## Stress type 1, weak ties
## Best solution was repeated 3 times in 20 tries
## The best solution was from try 17 (random start)
## Scaling: centring, PC rotation, halfchange scaling
## Species: expanded scores based on 'Z'

plot(ord, type="n")
text(ord, dis="si", cex=0.7)
ordilabel(ord, dis="sp", priority=colSums(dune),
          cex=0.8)
```

The species are strongly clustered by their phylogeny. In particularly grasses form a very compact group, and so do major clades in Dicots (Fig. 3).

Redundancy Analysis (RDA) is based on Euclidean distances, and when performed on Rao standardized data, the analysis will be based on phylogenetic distances among sample plots. In the following, we use automatic procedure to build a constrained model (Fig. 4):

```
m0 <- rda(Z ~ 1, dune.env)
m1 <- rda(Z ~ ., dune.env)

##
## Some constraints or conditions were aliased because they were
## redundant. This can happen if terms are linearly dependent
## (collinear): 'Manure^4'

(m <- ordistep(m0, formula(m1)))

##
## Start: Z ~ 1
##
##              Df     AIC      F Pr(>F)
## + Management  3 -66.672 4.2149  0.005 **
## + Manure      4 -66.567 3.6309  0.005 **
## + Use         2 -61.105 1.9237  0.070 .
## + A1          1 -61.231 2.0997  0.165
## + Moisture    3 -59.765 1.4263  0.210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step: Z ~ Management
##
##              Df     AIC      F Pr(>F)
## - Management  3 -61.025 4.2149  0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Some constraints or conditions were aliased because they were
## redundant. This can happen if terms are linearly dependent
## (collinear): 'Manure^4'

##
##            Df     AIC      F Pr(>F)
## + A1        1 -66.732 1.6274  0.160
## + Moisture  3 -66.028 1.3306  0.250
## + Manure    3 -65.000 1.0468  0.450
## + Use       2 -64.889 0.8206  0.575
## Call: rda(formula = Z ~ Management, data = dune.env)
##
## -- Model Summary --
##
##               Inertia Proportion Rank
## Total         0.04505    1.00000
## Constrained   0.01989    0.44143    3
## Unconstrained 0.02516    0.55857   16
##
## Inertia is variance
##
## -- Eigenvalues --
##
## Eigenvalues for constrained axes:
##     RDA1     RDA2     RDA3
## 0.017111 0.002186 0.000590
##
## Eigenvalues for unconstrained axes:
##      PC1      PC2      PC3      PC4      PC5      PC6
## 0.012086 0.008704 0.001416 0.001136 0.000673 0.000350
##      PC7      PC8      PC9     PC10     PC11     PC12
## 0.000287 0.000133 0.000126 0.000097 0.000076 0.000031
##     PC13     PC14     PC15     PC16
## 0.000028 0.000010 0.000007 0.000004

plot(m, scaling="sites")
```

The corresponding model with non-phylogenetic Euclidean distances has somewhat higher total inertia, but quite a large part of variation is also expressed by the phylogetic distances:

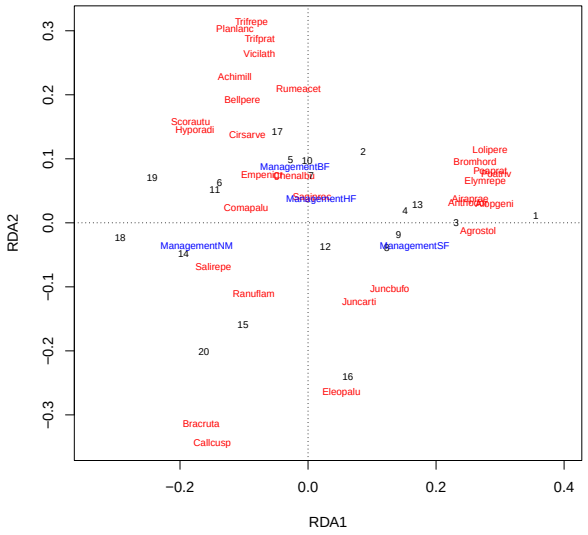

Figure 4: Constrained ordination based on phylogenetic Euclidean distances

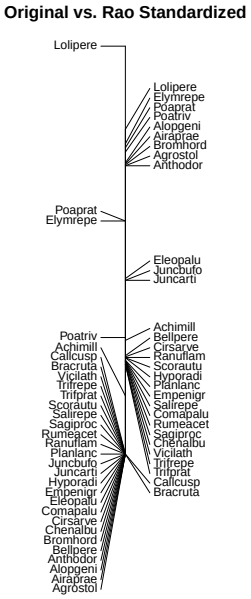

Figure 5: Raw data (on the left) and Rao standardized data for sample plot 1.

```
update(m, P ~ .)

## Call: rda(formula = P ~ Management, data = dune.env)
##
## -- Model Summary --
##
##               Inertia Proportion Rank
## Total         0.08043    1.00000
## Constrained   0.02291    0.28482    3
## Unconstrained 0.05752    0.71518   16
##
## Inertia is variance
##
## -- Eigenvalues --
##
## Eigenvalues for constrained axes:
##     RDA1     RDA2     RDA3
## 0.014092 0.006611 0.002205
##
## Eigenvalues for unconstrained axes:
##      PC1      PC2      PC3      PC4      PC5      PC6
## 0.019245 0.010710 0.006376 0.005499 0.004189 0.003009
##      PC7      PC8      PC9     PC10     PC11     PC12
## 0.001979 0.001561 0.001318 0.001171 0.000943 0.000652
##     PC13     PC14     PC15     PC16
## 0.000376 0.000272 0.000128 0.000093
```

Using explicit Rao standardization allows us to see how the data actually looks (Fig. 5). The displayed sample plot has only five species: four grasses (*Lolium perenne*, *Poa pratensis*, *Elymus repens* and *Poa trivialis*) and one species of Compositae (*Achillea millefolium*), and all other species have zero abundance. Rao standardization elevates all grasses and also other Monocots (*Eleocharis palustre*, *Juncus bufonius*, *J. articulatus*) to higher value than the only observed Dicot *A. millefolium*. The only missing species that remain at zero value are the two bryophytes (*Calliergonella cuspidata*, *Brachythecium rutabulum*) that are both maximally separated from the vascular plants (Fig. 1). The standardization may sound odd, but it must be understood that it only makes the effect transparent. Similar adjustment is done when phylogenetic or functional analysis is done without explicit standardization.

References

Gower JC (1966). "Some distance properties of latent root and vector methods used in multivariate analysis." *Biometrika*, **53**, 325–328.

Pavoine S, Ollier S, Pontier D (2005). "Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable?" *Theoretical Population Biology*, **67**, 231–239.

Pinheiro JC, Bates DM (2000). *Mixed-effect models in S and S-PLUS*. Springer.

Rao CR (1982). "Diversity and dissimilarity coefficients: A unified approach." *Theoretical Population Biology*, **21**, 24–43.