

# CCREPE: Compositionality Corrected by PERmutation and RENormalization

*Emma Schwager, George Weingart, Craig Bielski, Curtis Huttenhower*

October 27, 2020

## Contents

|     |                                 |    |
|-----|---------------------------------|----|
| 1   | Introduction . . . . .          | 2  |
| 2   | ccrepe . . . . .                | 2  |
| 2.1 | General functionality . . . . . | 2  |
| 2.2 | Arguments . . . . .             | 2  |
| 2.3 | Output . . . . .                | 4  |
| 2.4 | Usage . . . . .                 | 4  |
| 2.5 | Example 1 . . . . .             | 5  |
| 2.6 | Example 2 . . . . .             | 6  |
| 2.7 | Example 3 . . . . .             | 8  |
| 2.8 | Example 4 . . . . .             | 10 |
| 2.9 | Example 5 . . . . .             | 12 |
| 3   | nc.score . . . . .              | 12 |
| 3.1 | General Functionality . . . . . | 13 |
| 3.2 | Arguments . . . . .             | 13 |
| 3.3 | Output . . . . .                | 13 |
| 3.4 | Usage . . . . .                 | 14 |
| 3.5 | Example 1 . . . . .             | 14 |
| 3.6 | Example 2 . . . . .             | 15 |
| 3.7 | Example 3 . . . . .             | 16 |
| 4   | References . . . . .            | 17 |

## 1 Introduction

---

`ccepe` is a package for analysis of sparse compositional data. Specifically, it determines the significance of association between features in a composition, using any similarity measure (e.g. Pearson correlation, Spearman correlation, etc.) The CCREPE methodology stands for Compositionality Corrected by Renormalization and Permutation, as detailed below. The package also provides a novel similarity measure, the N-dimensional checkerboard score (NC-score), particularly appropriate to compositions derived from microbial community sequencing data. This results in p-values and false discovery rate q-values corrected for the effects of compositionality. The package contains two functions `ccepe` and `nc.score` and is maintained by the Huttenhower lab ([ccepe-users@googlegroups.com](mailto:ccepe-users@googlegroups.com)).

## 2 ccepe

---

`ccepe` is the main package function. It calculates compositionality-corrected p-values and q-values for a user-selected similarity measure, operating on either one or two input matrices. If given one matrix, all features (columns) in the matrix are compared to each other using the selected similarity measure. If given two matrices, each feature in the first are compared against all features in the second.

### 2.1 General functionality

Compositional data induces spurious correlations between features due to the nonindependence of values that must sum to a fixed total. CCREPE abrogates this when determining the significance of a similarity measure for each feature pair using two main steps, permutation/renormalization and bootstrapping. First, given two features to compare, CCREPE generates a null distribution of the similarity expected just due to compositionality by iteratively permuting one feature, renormalizing all samples in the composition to their previous sum, and computing the resulting similarity measures. Second, CCREPE bootstraps over sample subsets in order to assess confidence in the "true" similarity measure. Finally, the two resulting distributions are compared using a pooled-variance Z-test to give a compositionality-corrected p-value. False discovery rate q-values are additionally calculated using the Benjamin-Hochberg-Yekutieli procedure. For greater detail, see [Faust et al. \[2012\]](#) and [Schwager and Colleagues](#).

CCREPE employs several filtering steps before the data are processed. It removes any missing subjects using `na.omit`: in the two dataset case, any subjects missing in *either* dataset will be removed. Any subjects or features which are all zero are removed as well: an all-zero subject cannot be normalized (its sum is 0) and an all-zero feature has standard deviation 0 (in addition to being uninteresting biologically).

### 2.2 Arguments

- ✕ First *dataframe* or *matrix* containing relative abundances. Columns are features, rows are samples. Rows should therefore sum to a constant. Row names are used for identification if present.

## CCREPE: Compositionality Corrected by PERmutation and RENormalization

**y** Second *dataframe* or *matrix* (optional) containing relative abundances. Columns are features, rows are samples. Rows should therefore sum to a constant. If both **x** and **y** are specified, they will be merged by row names. If no row names are specified for either or both datasets, the default is to merge by row number.

**sim.score** Similarity measure, such as **cor** or **nc.score**. This can be either an existing R function or user-defined. If the latter, certain properties should be satisfied as detailed below (also see examples). The default similarity measure is Spearman correlation.

A user-defined similarity measure should mimic the interface of **cor**:

1. Take either two *vector* inputs one *matrix* or *dataframe* input.
2. In the case of two inputs, return a single number.
3. In the case of one input, return a matrix in which the  $(i,j)$ th entry is the similarity score for column **i** and column **j** in the original matrix.
4. The resulting matrix (in the case of one input) must be symmetric.
5. The inputs must be named **x** and **y**.

**sim.score.args** An optional list of arguments for the measurement function. When given, they are passed to the **sim.score** function directly. For example, in the case of **cor**, the following would be acceptable:

```
sim.score.args = list(method="spearman", use="complete.obs")
```

**min.subj** Minimum number (count) of samples that must be non-missing in order to apply the similarity measure. This is to ensure that there are sufficient samples to perform a bootstrap (default: 20).

**iterations** The number of iterations for both bootstrap and permutation calculations (default: 1000).

**subset.cols.x** A vector of column indices from **x** to indicate which features to compare

**subset.cols.y** A vector of column indices from **y** to indicate which features to compare

**errthresh** If feature has number of zeros greater than  $errthresh^{1/n}$ , that feature is excluded

**verbose** If **TRUE**, print periodic progress of the algorithm through the dataset(s), as well as including more detailed debugging output. (default: **FALSE**).

**iterations.gap** If **verbose=TRUE**, the number of iterations between issuing status messages (default: 100).

**distributions** Optional output file for detailed log (if given) of all intermediate permutation and renormalization distributions.

## CCREPE: Compositionality Corrected by PERmutation and REnormalization

**compare.within.x** A boolean value indicating whether to do comparisons given by taking all subsets of size 2 from `subset.cols.x` or to do comparisons given by taking all possible combinations of `subset.cols.x` and `subset.cols.y`. If TRUE but `subset.cols.y=NA`, returns all comparisons involving any features in `subset.cols.x`. This argument is only used when `y=NA`.

**concurrent.output** Optional output file to which each comparison will be written as it is calculated.

**make.output.table** A boolean value indicating whether to include table-formatted output.

### 2.3 Output

`ccrepe` returns a *list* containing both the calculation results and the parameters used:

**sim.score** *matrix* of similarity scores for all requested comparisons. The  $(i,j)$ th element corresponds to the similarity score of column `i` (or the `i`th column of `subset.cols.1`) and column `j` (or the `j`th column of `subset.cols.1`) in one dataset, or to the similarity score of column `i` (or the `i`th column of `subset.cols.1`) in dataset `x` and column `j` (or the `j`th column of `subset.cols.2`) in dataset `y` in the case of two datasets.

**p.values** *matrix* of the corrected p-values for all requested comparisons. The  $(i,j)$ th element corresponds to the p-value of the  $(i,j)$ th element of `sim.score`.

**q.values** *matrix* of the Benjamini-Hochberg-Yekutieli corrected p-values. The  $(i,j)$ th element corresponds to the p-value of the  $(i,j)$ th element of `sim.score`.

**z.stat** *matrix* of the z-statistics used in generating the p-values for all requested comparisons. The  $(i,j)$ th element corresponds to the z-statistic generating the  $(i,j)$ th element of `p.values`.

### 2.4 Usage

```
ccrepe(  
  x = NA,  
  y = NA,  
  sim.score = cor,  
  sim.score.args = list(),  
  min.subj = 20,  
  iterations = 1000,  
  subset.cols.x = NULL,  
  subset.cols.y = NULL,  
  errthresh = 1e-04,  
  verbose = FALSE,  
  iterations.gap = 100,  
  distributions = NA,  
  compare.within.x = TRUE,  
  concurrent.output = NA,
```

```
make.output.table = FALSE)
```

## 2.5 Example 1

An example of how to use `ccrepe` with one dataset.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.rowsum <- apply(data,1,sum)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized

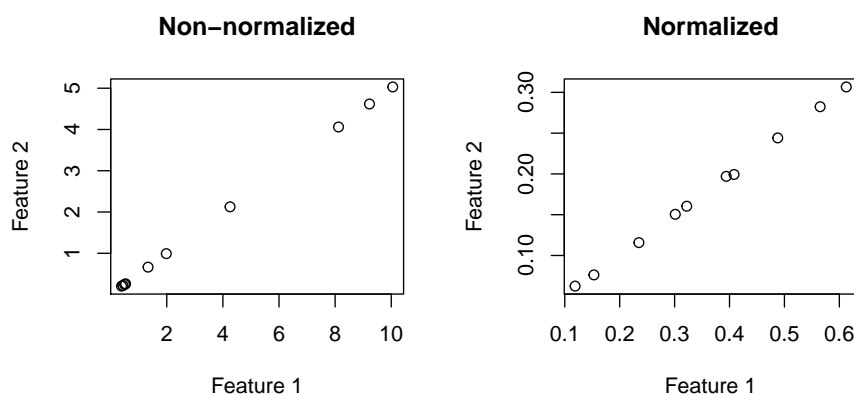
## [1] 1 1 1 1 1 1 1 1 1 1

test.input <- data.norm

dimnames(test.input) <- list(c(
  "Sample 1", "Sample 2", "Sample 3", "Sample 4", "Sample 5",
  "Sample 6", "Sample 7", "Sample 8", "Sample 9", "Sample 10"),
  c("Feature 1", "Feature 2", "Feature 3", "Feature 4"))

test.output <- ccrepe(x=test.input, iterations=20, min.subj=10)

par(mfrow=c(1,2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```



**Figure 1: Non-normalized and normalized associations between feature 1 and feature 2**

In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive `sim.score` value in the `[1,2]` element of `test.output$sim.score` and the small `q-value` in the `[1,2]` element of `test.output$q.values`.

```
test.output
```

## CCREPE: Compositionality Corrected by PERmutation and REnormalization

```
## $p.values
##           Feature 1    Feature 2  Feature 3 Feature 4
## Feature 1           NA 1.099714e-10 0.26824326 0.7617798
## Feature 2 1.099714e-10           NA 0.03863191 0.3201303
## Feature 3 2.682433e-01 3.863191e-02           NA 0.4532106
## Feature 4 7.617798e-01 3.201303e-01 0.45321057           NA
##
## $z.stat
##           Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1           NA 6.4525668 -1.1071172 -0.3031444
## Feature 2 6.4525668           NA -2.0680863 -0.9941902
## Feature 3 -1.1071172 -2.0680863           NA 0.7500733
## Feature 4 -0.3031444 -0.9941902 0.7500733           NA
##
## $sim.score
##           Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1           NA 0.9997268 -0.6116927 -0.6090706
## Feature 2 0.9997268           NA -0.6188943 -0.6017044
## Feature 3 -0.6116927 -0.6188943           NA -0.2548284
## Feature 4 -0.6090706 -0.6017044 -0.2548284           NA
##
## $q.values
##           Feature 1    Feature 2 Feature 3 Feature 4
## Feature 1           NA 1.563117e-09 1.2709232 1.804637
## Feature 2 1.563117e-09           NA 0.2745541 1.137571
## Feature 3 1.270923e+00 2.745541e-01           NA 1.288373
## Feature 4 1.804637e+00 1.137571e+00 1.2883735           NA
```

## 2.6 Example 2

An example of how to use `ccrepe` with two datasets.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.rowsum <- apply(data,1,sum)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized

## [1] 1 1 1 1 1 1 1 1 1 1

test.input <- data.norm

data2 <- matrix(rlnorm(105,meanlog=0,sdlog=1),nrow=15,ncol=7)
aligned.rows <- c(seq(1,4),seq(6,9),11,12) # The datasets dont need
                                           # to have subjects line up exactly
data2[aligned.rows,1] <- 2*data[,3] + rnorm(10,0,0.01)
data2.rowsum <- apply(data2,1,sum)
data2.norm <- data2/data2.rowsum
apply(data2.norm,1,sum) # The rows sum to 1, so the data are normalized
```

## CCREPE: Compositionality Corrected by PERmutation and RENormalization

```
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

test.input.2 <- data2.norm

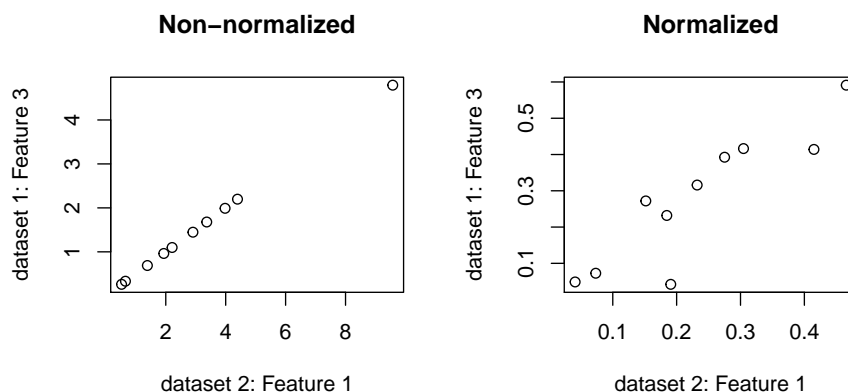
dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))
dimnames(test.input.2) <- list(paste("Sample",c(seq(1,4),11,seq(5,8),12,9,10,13,14,15)),paste("Feature",seq(1,4)))

test.output.two.datasets <- ccrepe(x=test.input, y=test.input.2, iterations=20, min.subj=10)

## Warning in preprocess_data(CA): Removing subjects Sample 11, Sample 12, Sample
13, Sample 14, Sample 15 from dataset y because they are not in dataset x.
```

Please note that we receive a warning because the subjects don't match - only paired observations.

```
par(mfrow=c(1,2))
plot(data2[aligned.rows,1],data[,3],xlab="dataset 2: Feature 1",ylab="dataset 1: Feature 3",main="Non-normalized")
plot(data2.norm[aligned.rows,1],data.norm[,3],xlab="dataset 2: Feature 1",ylab="dataset 1: Feature 3",
      main="Normalized")
```



**Figure 2: Non-normalized and normalized associations between feature 1 and feature 2**

In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive sim.score value in the [1,2] element of test.output\$sim.score and the small q-value in the [1,2] element of test.output\$q.values.

```
test.output.two.datasets

## $p.values
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1 3.538887e-01 0.5218125 0.7978558 0.07655114 0.1123044 0.7456825
## Feature 2 9.236567e-01 0.7412677 0.9204031 0.19073837 0.4045824 0.9909446
## Feature 3 3.406880e-10 0.7201470 0.3119924 0.52795827 0.1767145 0.5103230
## Feature 4 9.963608e-04 0.7046326 0.4171961 0.27279862 0.8335264 0.5085601
##           Feature 7
## Feature 1 0.951722053
## Feature 2 0.602766485
## Feature 3 0.053502783
```

## CCREPE: Compositionality Corrected by PERmutation and RENormalization

```
## Feature 4 0.004731077
##
## $z.stat
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5
## Feature 1 -0.92707289 0.6405540 -0.25612303 -1.7710575 1.5879202
## Feature 2 -0.09582866 0.3301750 -0.09992593 -1.3083980 0.8334653
## Feature 3  6.27903177 -0.3582624 -1.01105029  0.6311258 -1.3509405
## Feature 4 -3.29155232 -0.3790744  0.81129540  1.0966404  0.2101810
##           Feature 6 Feature 7
## Feature 1  0.32433748 -0.0605444
## Feature 2 -0.01134951 -0.5204263
## Feature 3  0.65833480 -1.9308405
## Feature 4 -0.66108137  2.8247957
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1 -0.1907519 0.2296691 0.13747569 -0.3635626 0.3456475 -0.09050654
## Feature 2 -0.1899959 0.2337687 0.13358198 -0.3626916 0.3458904 -0.09515077
## Feature 3  0.8967466 -0.1185548 -0.16034562  0.1485712 -0.5199519  0.24523784
## Feature 4 -0.7083457 -0.1110182  0.02512497  0.2124515  0.1767906 -0.15399144
##           Feature 7
## Feature 1 -0.1334669
## Feature 2 -0.1292397
## Feature 3 -0.4556609
## Feature 4  0.5874774
##
## $q.values
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1 3.521635e+00 3.569972 3.797231 1.675915 2.048877 3.710237
## Feature 2 3.888728e+00 3.863903 4.030032 2.609867 3.690592 3.874019
## Feature 3 3.729299e-08 3.941500 3.415186 3.399547 2.763405 3.724125
## Feature 4 5.453270e-02 4.059565 3.512912 3.317952 3.801706 3.976350
##           Feature 7
## Feature 1 3.8584844
## Feature 2 3.6656160
## Feature 3 1.4641540
## Feature 4 0.1726272
```

## 2.7 Example 3

An example of how to use `ccrepe` with `nc.score` as the similarity score.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.rowsum <- apply(data,1,sum)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized
## [1] 1 1 1 1 1 1 1 1 1 1
```



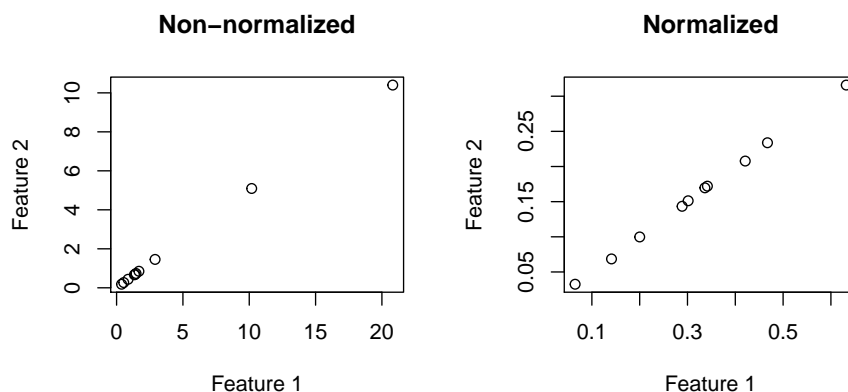
## CCREPE: Compositionality Corrected by PERmutation and RENormalization

```
test.input <- data.norm

dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

test.output.nc.score <- ccrepe(x=test.input, sim.score=nc.score, iterations=20, min.subj=10)

par(mfrow=c(1,2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```



**Figure 3: Non-normalized and normalized associations between feature 1 and feature 2**

In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive `sim.score` value in the [1,2] element of `test.output$sim.score` and the small `q`-value in the [1,2] element of `test.output$q.values`. In this case, however, the `sim.score` represents the NC-Score between two features rather than the Spearman correlation.

```
test.output.nc.score

## $p.values
##           Feature 1   Feature 2 Feature 3 Feature 4
## Feature 1          NA 1.949897e-05 0.2446170 0.6964397
## Feature 2 1.949897e-05          NA 0.1608701 0.9227462
## Feature 3 2.446170e-01 1.608701e-01          NA 0.7181142
## Feature 4 6.964397e-01 9.227462e-01 0.7181142          NA
##
## $z.stat
##           Feature 1   Feature 2 Feature 3 Feature 4
## Feature 1          NA 4.27055206 -1.1635240 0.39013096
## Feature 2 4.270552          NA -1.4021512 0.09697507
## Feature 3 -1.163524 -1.40215117          NA 0.36098023
## Feature 4 0.390131 0.09697507 0.3609802          NA
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA 1.00000 -0.59375 0.34375
```

## CCREPE: Compositionality Corrected by PERmutation and REnormalization

```
## Feature 2  1.00000      NA -0.59375  0.34375
## Feature 3 -0.59375 -0.59375      NA -0.34375
## Feature 4  0.34375  0.34375 -0.34375      NA
##
## $q.values
##           Feature 1  Feature 2 Feature 3 Feature 4
## Feature 1      NA 0.0002771554  1.158983  2.474773
## Feature 2 0.0002771554      NA  1.143292  2.185963
## Feature 3 1.1589830571 1.1432918578      NA  2.041434
## Feature 4 2.4747725224 2.1859627896  2.041434      NA
```

## 2.8 Example 4

An example of how to use `ccrepe` with a user-defined `sim.score` function.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.rowsum <- apply(data,1,sum)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized
## [1] 1 1 1 1 1 1 1 1 1 1

test.input <- data.norm

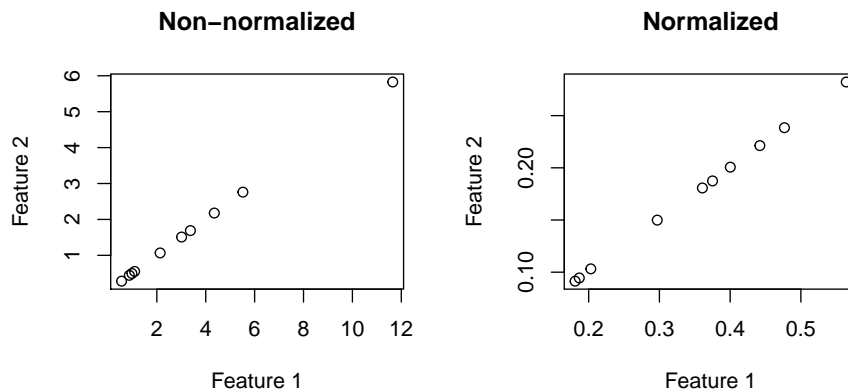
dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

my.test.sim.score <- function(x,y=NA,constant=0.5){
  if(is.vector(x) && is.vector(y)) return(constant)
  if(is.matrix(x) && is.na(y)) return(matrix(rep(constant,ncol(x)^2),ncol=ncol(x)))
  if(is.data.frame(x) && is.na(y)) return(matrix(rep(constant,ncol(x)^2),ncol=ncol(x)))
  else stop('ERROR')
}

test.output.sim.score <- ccrepe(x=test.input, sim.score=my.test.sim.score, iterations=20, min.subj=10, sim)

par(mfrow=c(1,2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```

## CCREPE: Compositionality Corrected by PERmutation and REnormalization



**Figure 4: Non-normalized and normalized associations between feature 1 and feature 2**

In this case we would expect feature 1 and feature 2 to be associated. Note that the values of sim.score are all 0.6 and none of the p-values are very small because of the arbitrary definition of the similarity score.

```
test.output.sim.score

## $p.values
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1         NA        NaN        NaN        NaN
## Feature 2         NaN         NA        NaN        NaN
## Feature 3         NaN         NaN         NA        NaN
## Feature 4         NaN         NaN         NaN         NA
##
## $z.stat
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1         NA        NaN        NaN        NaN
## Feature 2         NaN         NA        NaN        NaN
## Feature 3         NaN         NaN         NA        NaN
## Feature 4         NaN         NaN         NaN         NA
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1         NA         0.6         0.6         0.6
## Feature 2         0.6         NA         0.6         0.6
## Feature 3         0.6         0.6         NA         0.6
## Feature 4         0.6         0.6         0.6         NA
##
## $q.values
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1         NA        NaN        NaN        NaN
## Feature 2         NaN         NA        NaN        NaN
## Feature 3         NaN         NaN         NA        NaN
## Feature 4         NaN         NaN         NaN         NA
```

## 2.9 Example 5

An example of how to use `ccrepe` when specifying column subsets.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data.rowsum <- apply(data,1,sum)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized
## [1] 1 1 1 1 1 1 1 1 1 1
test.input <- data.norm

dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

test.output.1.3 <- ccrepe(x=test.input, iterations=20, min.subj=10, subset.cols.x=c(1,3))
test.output.1 <- ccrepe(x=test.input, iterations=20, min.subj=10, subset.cols.x=c(1), compare.within.x=
test.output.12.3 <- ccrepe(x=test.input, iterations=20, min.subj=10, subset.cols.x=c(1,2),subset.cols.y=c
test.output.1.3$sim.score

##           Feature 1  Feature 3
## Feature 1          NA -0.5214169
## Feature 3 -0.5214169          NA

test.output.1$sim.score

##           Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1          NA -0.3722563 -0.5214169 -0.4928654
## Feature 2 -0.3722563          NA          NA          NA
## Feature 3 -0.5214169          NA          NA          NA
## Feature 4 -0.4928654          NA          NA          NA

test.output.12.3$sim.score

##           Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1          NA          NA -0.5214169          NA
## Feature 2          NA          NA -0.4164992          NA
## Feature 3 -0.5214169 -0.4164992          NA          NA
## Feature 4          NA          NA          NA          NA
```

## 3 nc.score

The `nc.score` similarity measure is an N-dimensional extension of the checkerboard score particularly suited to similarity score calculations between compositions derived from ecological relative abundance measurements. In such cases, features typically represent species abundances, and the NC-score discretizes these continuous values into one of N bins before computing a normalized similarity of co-occurrence or co-exclusion. This can be used as a standalone function or with `ccrepe` as above to obtain compositionality-corrected p-values.

### 3.1 General Functionality

The NC-score is an extension to Diamond's checkerboard score (see [Cody and Diamond \[1975\]](#)) to ordinal data, and simplifies to a calculation of Kendall's  $\tau$  on binned data instead of ranked data. Let two features in a dataset with  $n$  subjects be denoted by

$$\begin{bmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \end{bmatrix}.$$

The binning function maps from the original data to  $b$  numbered bins in  $\{1, \dots, b\}$ . Let the binning function be denoted by  $B(\cdot)$ . The co-variation and co-exclusion patterns are the same as concordant and discordant pairs in Kendall's  $\tau$ . Considering a  $2 \times 2$  submatrix of the form

$$\begin{bmatrix} B(x_i) & B(x_j) \\ B(y_i) & B(y_j) \end{bmatrix},$$

a co-variation pattern is counted when  $(B(x_i) - B(x_j))(B(y_i) - B(y_j)) > 0$  and a co-exclusion pattern, conversely, when  $(B(x_i) - B(x_j))(B(y_i) - B(y_j)) < 0$ . The NC-score statistic for features  $x$  and  $y$  is then defined as

$$(\text{number of co-variation patterns}) - (\text{number of co-exclusion patterns}),$$

normalized by the Kendall's  $\tau$  normalization factor accounting for ties described in [Kendall \[1970\]](#).

### 3.2 Arguments

- x** First numerical *vector*, or single *dataframe* or *matrix*, containing relative abundances. If the latter, columns are features, rows are samples. Rows should therefore sum to a constant.
- y** If provided, second numerical *vector* containing relative abundances. If given, **x** must be a *vector* as well.
- nbins** A non-negative integer of the number of bins to generate (cutoffs will be generated by the `discretize` function from the `infotheo` package).
- bin.cutoffs** A list of values demarcating the bin cutoffs. The binning is performed using the `findInterval` function.
- use** An optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs".

### 3.3 Output

**nc.score** returns either a single number (if called with two vectors) or a *matrix* of all pairwise scores (if called with a *matrix*) of normalized scores. This behaviour is precisely analogous to the `cor` function in R

### 3.4 Usage

```
nc.score(
  x,
  y = NULL,
  use = "everything",
  nbins = NULL,
  bin.cutoffs=NULL)
```

### 3.5 Example 1

An example of using `nc.score` to get a single similarity score or a matrix.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data.rowsum <- apply(data,1,sum)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized

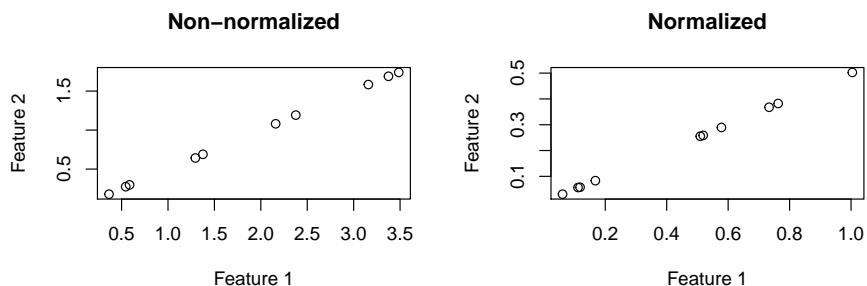
## [1] 1.3754546 1.3031599 1.7117218 1.4873440 1.5863092 1.4465728 0.7687089
## [8] 0.5560650 1.0734369 0.5277281

test.input <- data.norm

dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

test.output.matrix <- nc.score(x=test.input)
test.output.num <- nc.score(x=test.input[,1],y=test.input[,2])

par(mfrow=c(1, 2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```



**Figure 5: Non-normalized and normalized associations between feature 1 and feature 2 of the second example**

Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion patterns. This is shown by the positive and relatively high value of the [1,2] element of `test.output.matrix` (which is identical to `test.output.num`)

```
test.output.matrix
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1    1.00000    1.00000    0.00000   -0.34375
## Feature 2    1.00000    1.00000    0.00000   -0.34375
## Feature 3    0.00000    0.00000    1.00000   -0.59375
## Feature 4   -0.34375   -0.34375   -0.59375    1.00000

test.output.num
## [1] 1
```

## 3.6 Example 2

An example of using `nc.score` with an arbitrary bin number.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data.rowsum <- apply(data,1,sum)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized

## [1] 1.5831343 1.5160173 0.6318757 0.7662618 0.9313612 1.2519830 1.5139171
## [8] 1.1566765 1.4101321 1.1074489

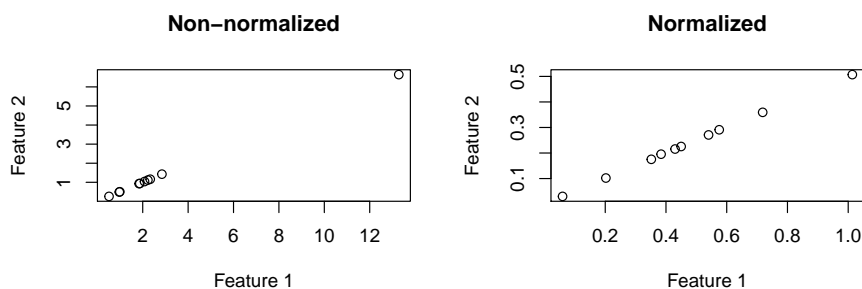
test.input <- data.norm

dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

test.output <- nc.score(x=test.input,nbins=4)

par(mfrow=c(1, 2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```

## CCREPE: Compositionality Corrected by PERmutation and REnormalization



**Figure 6: Non-normalized and normalized associations between feature 1 and feature 2 of the second example**

Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion patterns. This is shown by the positive and relatively high value in the [1,2] element of test.output. In this case, the smaller bin number yields a smaller NC-score because of the coarser partitioning of the data.

```
test.output

##           Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1  1.00000000  1.00000000 -0.4285714  0.05714286
## Feature 2  1.00000000  1.00000000 -0.4285714  0.05714286
## Feature 3 -0.42857143 -0.42857143  1.00000000  0.11428571
## Feature 4  0.05714286  0.05714286  0.1142857  1.00000000
```

### 3.7 Example 3

An example of using `nc.score` with user-defined bin edges.

```
data <- matrix(rlnorm(40,meanlog=0,sdlog=1),nrow=10,ncol=4)
data.rowsum <- apply(data,1,sum)
data[,1] = 2*data[,2] + rnorm(10,0,0.01)
data.norm <- data/data.rowsum
apply(data.norm,1,sum) # The rows sum to 1, so the data are normalized

## [1] 1.0546606 1.1114593 1.0421141 1.0598200 1.2797121 0.8195896 0.6051032
## [8] 1.6361260 2.0738000 1.1013499

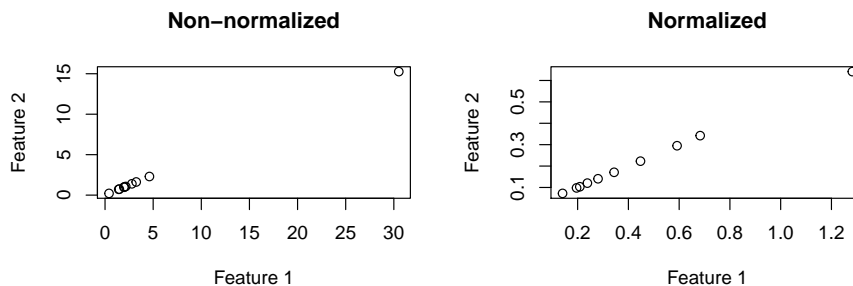
test.input <- data.norm

dimnames(test.input) <- list(paste("Sample",seq(1,10)),paste("Feature",seq(1,4)))

test.output <- nc.score(x=test.input,bin.cutoffs=c(0.1,0.2,0.3))

par(mfrow=c(1, 2))
plot(data[,1],data[,2],xlab="Feature 1",ylab="Feature 2",main="Non-normalized")
plot(data.norm[,1],data.norm[,2],xlab="Feature 1",ylab="Feature 2",
      main="Normalized")
```





**Figure 7: Non-normalized and normalized associations between feature 1 and feature 2 of the second example**

Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion patterns. This is shown by the positive and relatively high value in the [1,2] element of test.output. The bin edges specified here represent almost absent ([0,0.001)), low abundance ([0.001,0.1)), medium abundance ([0.1,0.25)), and high abundance ([0.6,1)).

test.output

| ##           | Feature 1   | Feature 2  | Feature 3   | Feature 4  |
|--------------|-------------|------------|-------------|------------|
| ## Feature 1 | 1.00000000  | 0.8381581  | -0.03035884 | -0.3439180 |
| ## Feature 2 | 0.83815808  | 1.00000000 | -0.16903085 | -0.4642071 |
| ## Feature 3 | -0.03035884 | -0.1690309 | 1.00000000  | 0.1765470  |
| ## Feature 4 | -0.34391797 | -0.4642071 | 0.17654697  | 1.00000000 |

## 4 References

### References

- Martin Leonard Cody and Jared Mason Diamond. *Ecology and evolution of communities*. Harvard University Press, 1975.
- Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*, 8(7):e1002606, 2012.
- M.G. Kendall. *Rank correlation methods*. Charles Griffin & Co., 1970.
- Emma Schwager and Colleagues. Detecting statistically significant associations between sparse and high dimensional compositional data. In Progress.