

# CCREPE: Compositionality Corrected by Permutation and Renormalization

Emma Schwager, George Weingart, Craig Bielski, Curtis Huttenhower

May 2, 2014

## Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>ccrepe</b>	<b>2</b>
2.1	General functionality . . . . .	2
2.2	Arguments . . . . .	2
2.3	Output . . . . .	3
2.4	Usage . . . . .	3
2.5	Example 1 . . . . .	3
2.6	Example 2 . . . . .	5
2.7	Example 3 . . . . .	7
2.8	Example 4 . . . . .	9
2.9	Example 5 . . . . .	10
<b>3</b>	<b>nc.score</b>	<b>11</b>
3.1	General Functionality . . . . .	11
3.2	Arguments . . . . .	12
3.3	Output . . . . .	12
3.4	Usage . . . . .	12
3.5	Example 1 . . . . .	13
3.6	Example 2 . . . . .	14
3.7	Example 3 . . . . .	14
<b>4</b>	<b>References</b>	<b>15</b>

## 1 Introduction

---

*ccrepe* 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 `ccrepe` and `nc.score` and is maintained by the Huttenhower lab ([ccrepe-users@googlegroups.com](mailto:ccrepe-users@googlegroups.com)).

## 2 ccrepe

---

`ccrepe` 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.

### 2.2 Arguments

`x` 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.

`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")
```

Note that this example corresponds to the default behavior.

`min.subj` Minimum number (count) of samples that must be nonzero 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` Subset of columns from `x` to work on. The default (NULL) uses all columns. Note that all features are used for normalization, but calculations are performed only with the requested subset.

`subset.cols.y` Subset of columns from `y` to work on. The default (NULL) uses all columns. Note that all features are used for normalization, but calculations are performed only with the requested subset.

`errthresh1` Maximum allowable probability of getting all zeros in a given bootstrapped column for the first dataset `x`. If a feature has a number of zeros that makes the probability of obtaining all zeros when sampling with replacement greater than this value, that feature will be excluded from the subsequent analysis. This is to ensure that the standard deviation of the bootstrap sample is non-zero. (default: 0.0001).

`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.

`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` or `subset.cols.y`. If TRUE but `subset.cols.y=NA`, returns all comparisons involving any features in `subset.cols.x`.

`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, errthresh1 = 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, 1)
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
```

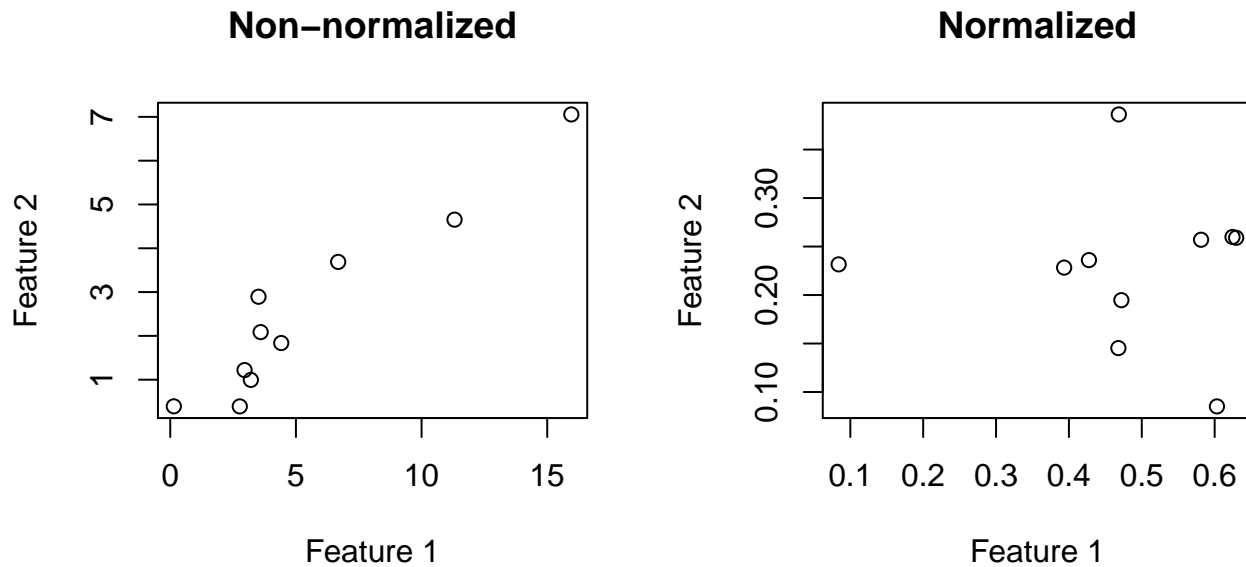


Figure 1: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive sim.score value in the [1,2] element of test.output\$sim.score and the small q-value in the [1,2] element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

```
## [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")

test.output

## $p.values
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA  0.04094  0.8792  0.6567
## Feature 2  0.04094          NA  0.5617  0.2790
## Feature 3  0.87915  0.56170          NA  0.4992
## Feature 4  0.65673  0.27899  0.4992          NA
##
## $z.stat
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA   2.0442  0.1520  0.4444
```

```
## Feature 2    2.0442      NA   -0.5803   -1.0826
## Feature 3    0.1520   -0.5803      NA   -0.6758
## Feature 4    0.4444   -1.0826   -0.6758      NA
##
## $q.values
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1      NA    0.5819    2.083    1.867
## Feature 2    0.5819      NA    1.996    1.983
## Feature 3    2.0827    1.9960      NA    2.365
## Feature 4    1.8669    1.9828    2.365      NA
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1      NA    0.2727   -0.3333   -0.3212
## Feature 2    0.2727      NA   -0.4424   -0.4788
## Feature 3   -0.3333   -0.4424      NA   -0.3939
## Feature 4   -0.3212   -0.4788   -0.3939      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, 1)
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, 1)
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

## [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, 7)))

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

```

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")

```

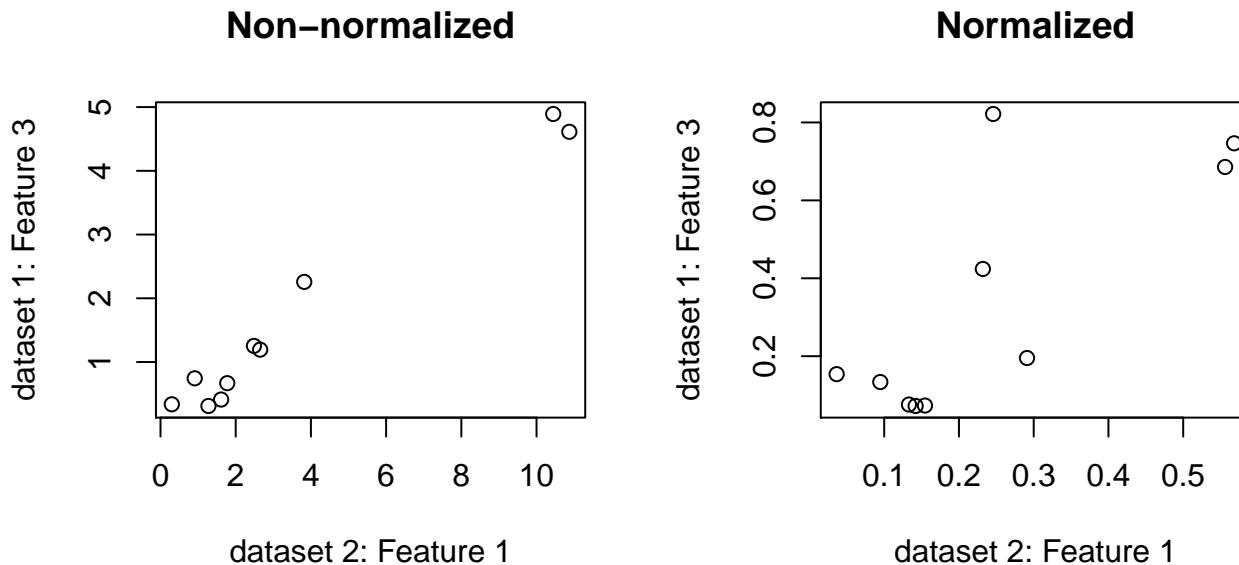


Figure 2: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive `sim.score` value in the [1,2] element of `test.output$sim.score` and the small `q-value` in the [1,2]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

```

test.output.two.datasets

## $p.values
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1  0.38073   0.7068   0.2887   0.26502  0.10221  0.5740
## Feature 2  0.27567   0.3123   0.8869   0.78695  0.48098  0.5879
## Feature 3  0.08368   0.5591   0.4181   0.03748  0.01892  0.2084
## Feature 4  0.15102   0.1522   0.2775   0.70519  0.51157  0.4039
##           Feature 7
## Feature 1  0.9250
## Feature 2  0.7443
## Feature 3  0.2050
## Feature 4  0.4060
##
## $z.stat
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1  -0.8765  -0.3761   1.0610   1.1146   1.6342   0.5622
## Feature 2  -1.0901   1.0104  -0.1423   0.2703   0.7047   0.5418
## Feature 3   1.7297  -0.5841  -0.8098  -2.0805  -2.3472  -1.2579
## Feature 4  -1.4359   1.4318  -1.0859   0.3783  -0.6564   0.8347

```

```
##           Feature 7
## Feature 1 -0.09415
## Feature 2  0.32613
## Feature 3  1.26745
## Feature 4 -0.83102
##
## $q.values
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1  2.977    3.224    2.633    3.223    2.797    2.992
## Feature 2  3.018    2.630    3.596    3.313    2.925    2.925
## Feature 3  3.053    3.060    2.692    2.051    2.071    2.852
## Feature 4  3.306    2.777    2.762    3.356    2.947    2.948
##           Feature 7
## Feature 1  3.616
## Feature 2  3.259
## Feature 3  3.206
## Feature 4  2.777
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1 -0.4061  -0.1515  0.53939  0.5515  0.6485  0.2000
## Feature 2 -0.2848  0.3091  0.06667  0.1879  0.2121  0.1152
## Feature 3  0.6727  -0.1394 -0.39394  -0.6364  -0.5879  -0.5879
## Feature 4 -0.3212  0.6485  -0.36970  -0.1030  -0.3939  0.3818
##           Feature 7
## Feature 1 -0.22424
## Feature 2  0.10303
## Feature 3  0.56364
## Feature 4 -0.05455
```

## 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, 1)
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.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")
```

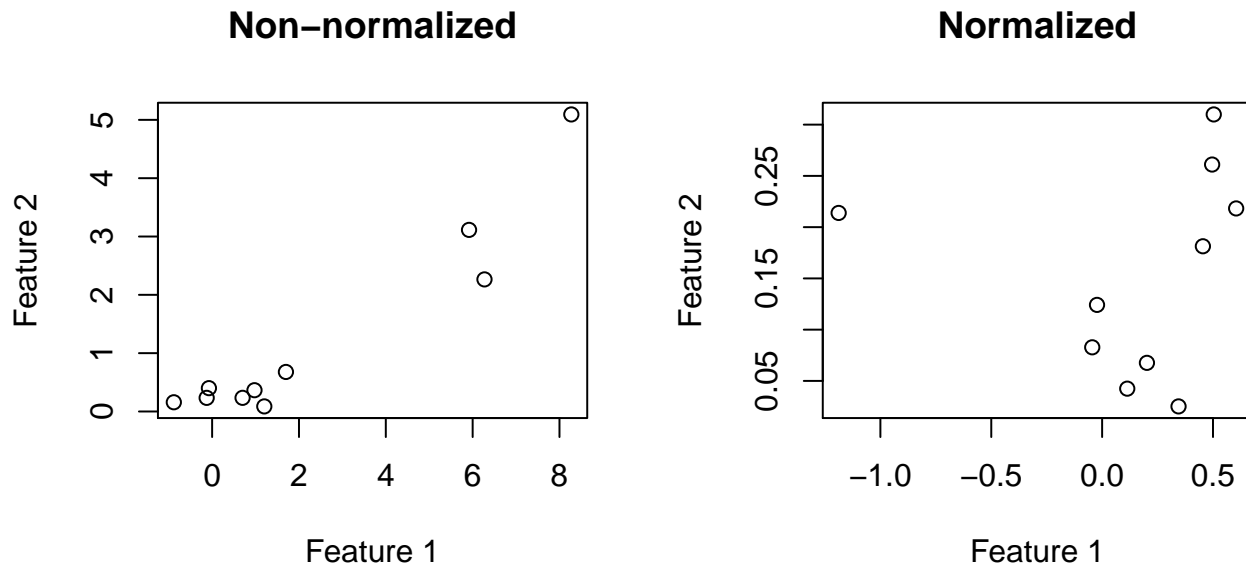


Figure 3: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive `sim.score` value in the [1,2] element of `test.output$sim.score` and the small `q`-value in the [1,2] element of `test.output$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  0.001086   0.8115  0.074876
## Feature 2  0.001086      NA   0.1951  0.006541
## Feature 3  0.811457  0.195093      NA  0.213255
## Feature 4  0.074876  0.006541   0.2133      NA
##
## $z.stat
##      Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1      NA    3.267  -0.2385  -1.781
## Feature 2    3.2673      NA  -1.2957  -2.719
## Feature 3  -0.2385  -1.296      NA    1.245
## Feature 4  -1.7812  -2.719   1.2447      NA
##
## $q.values
##      Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1      NA  0.01543   1.9223  0.35476
## Feature 2  0.01543      NA   0.6933  0.04649
## Feature 3  1.92232  0.69326      NA  0.60623
## Feature 4  0.35476  0.04649   0.6062      NA
##
## $sim.score
##      Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1      NA    0.3810  -0.7489  -0.8139
## Feature 2  0.3810      NA  -0.2857  -0.5887
```



```
## Feature 3  -0.7489  -0.2857      NA   0.3810
## Feature 4  -0.8139  -0.5887   0.3810      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, 1)
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.score.args = list(constant = 0.6))

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")
```

```
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
##
```

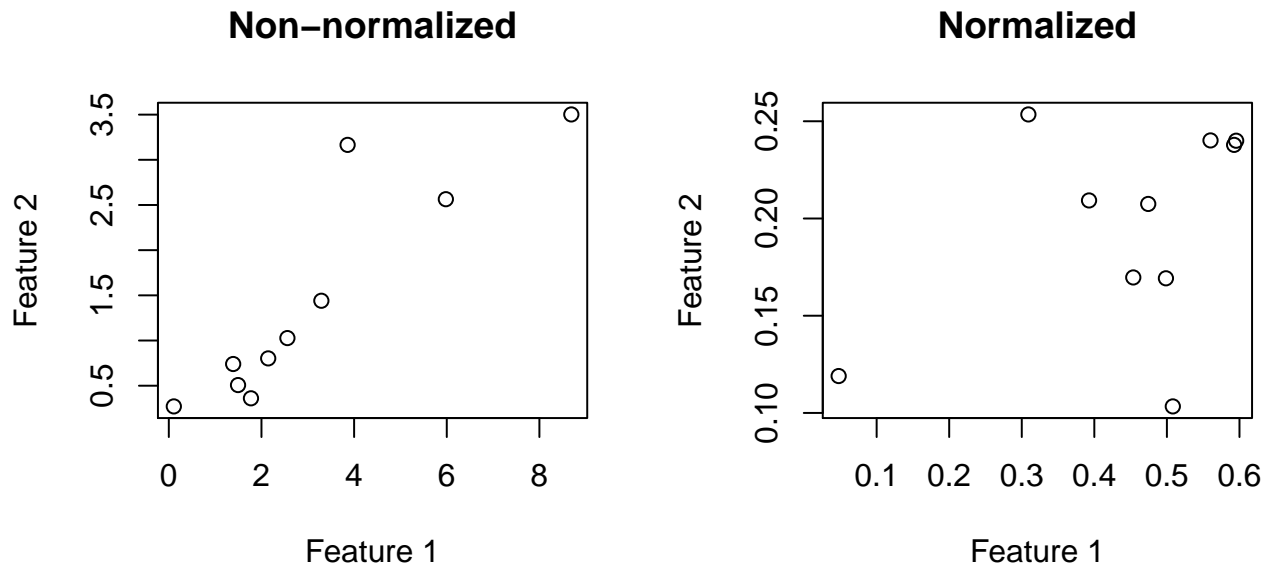


Figure 4: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. Note that the values of sim.score are all 0.6 and none of the p-values are very small because of the arbitrary definition of the similarity score.

```
## $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
##
## $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
```

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

##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA          NA  -0.5758          NA
## Feature 2          NA          NA          NA          NA
## Feature 3  -0.5758          NA          NA          NA
## Feature 4          NA          NA          NA          NA

test.output.1$sim.score

##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA  -0.2485  -0.5758  -0.6242
## Feature 2  -0.2485          NA          NA          NA
## Feature 3  -0.5758          NA          NA          NA
## Feature 4  -0.6242          NA          NA          NA

test.output.12.3$sim.score

##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1          NA          NA  -0.5758          NA
## Feature 2          NA          NA  -0.1515          NA
## Feature 3  -0.5758  -0.1515          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. The generalization of the checkerboard score is through defining general co-variation and co-exclusion patterns in ordinal data. With ordinal data, the checkerboard considers the  $2 \times 2$  submatrices which are of the form:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

In the cases that  $a < b$  and  $c < d$  (or  $a > b$  and  $c > d$ ), the submatrix is considered a co-variation pattern; conversely, if  $a > b$  and  $c < d$  (or  $a < b$  and  $c > d$ ) then the submatrix is considered a co-exclusion pattern. The number of co-exclusion patterns for  $n$  bins is adjusted based on the expected ratio of co-variation to co-exclusion patterns in any

given sample. The adjustment factor is given by:

$$1.5 * \frac{n(n-1)}{n^2 - n + 1},$$

with the complete derivation being left to Bielski [2013].

The function as implemented here first performs basic quality control filtering of input relative abundance data. It then transforms relative abundances to ordinal values based on user-provided or default bin thresholds. It then computes a raw NC-score:

$$C_{ij} - 1.5 * \frac{n(n-1)}{n^2 - n + 1} * D_{ij}$$

where  $C_{ij}$  and  $D_{ij}$  are the total number of co-variation and co-exclusion patterns, respectively, between species  $i$  and  $j$  and where  $n$  is the number of bins.

The raw NC-score is then normalized to between -1 and 1 for  $n$  bins and  $s$  samples by dividing by the maximum possible NC-score, which is based on the maximum number of co-variation patterns possible. The normalized NC-score is then analogous to Pearson's  $\rho$ , where positive values indicate more co-variation than co-exclusion patterns and the magnitude of the score indicates the strength of the association between the species. The derivation of this normalization factor can be found in Bielski [2013]. The normalization factor is given by:

$$\binom{s}{2} - \left[ s \bmod n * \binom{\lfloor \frac{s}{n} \rfloor + 1}{2} + (n - s \bmod n) * \binom{\lfloor \frac{s}{n} \rfloor}{2} \right].$$

## 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.

**bins** Either the number of bins to use or a *vector* specifying bin edges. If a single number is given, this is used as the number of bins with the `discretize` function of the package *infortheo*. If a *vector* is specified, the function `findInterval` is used to discretize the data. The default behavior is to use the defaults for the `discretize` function.

**verbose** Request verbose output.

**min.abundance** Minimum abundance threshold for quality control filtering. For a feature to be included, it must take a value of at least `min.abundance` in at least `min.samples` percent of samples.

**min.samples** Minimum sample threshold for quality control filtering. For a feature to be included, it must take a value of at least `min.abundance` in at least `min.samples` percent of samples.

## 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.

## 3.4 Usage

```
nc.score(x = NA, y = NA, bins = NA, verbose = FALSE, min.abundance = 1e-04, min.samples = 0.1)
```

### 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, 1)
data.norm <- data/data.rowsum
apply(data.norm, 1, sum) # The rows sum to 1, so the data are normalized

## [1] 1.2300 0.7188 1.5339 1.2907 1.0305 1.7123 2.2367 1.1729 0.9353 1.0363

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")
```

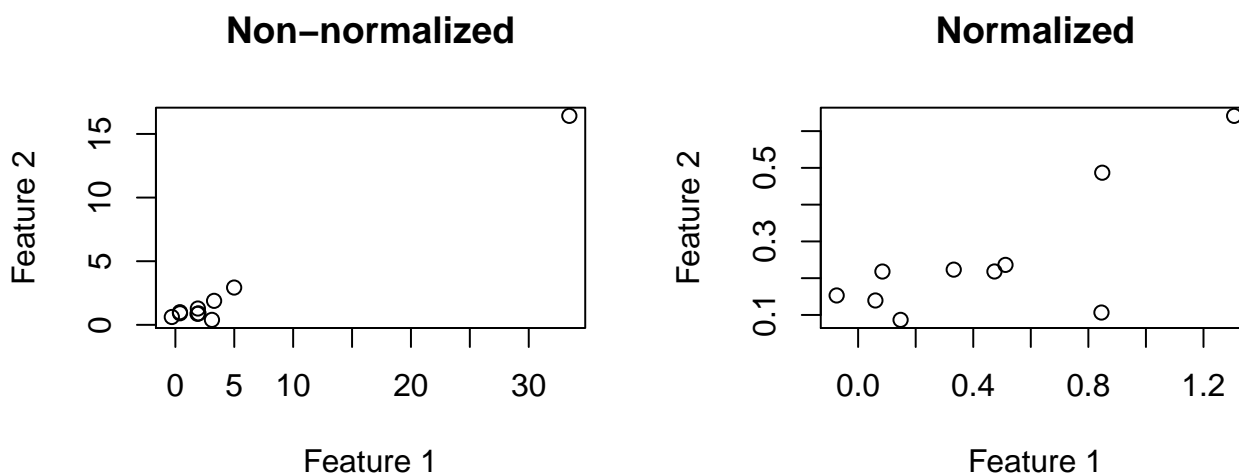


Figure 5: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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           NA    0.5887  -0.83550  -0.07792
```

```
## Feature 2  0.58874      NA -0.33333 -0.38095
## Feature 3 -0.83550  -0.3333      NA -0.07792
## Feature 4 -0.07792  -0.3810 -0.07792      NA

test.output.num

## [1] 0.5887
```

### 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, 1)
data.norm <- data/data.rowsum
apply(data.norm, 1, sum) # The rows sum to 1, so the data are normalized

## [1] 1.5571 0.4790 1.5085 1.3988 1.7848 1.2138 2.2280 1.6783 0.9496 1.3327

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, bins = 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")
```

```
test.output

##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1         NA         0.6         0.2         0.2
## Feature 2         0.6         NA         0.2        -0.2
## Feature 3         0.2         0.2         NA        -0.2
## Feature 4         0.2        -0.2        -0.2         NA
```

### 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, 1)
data.norm <- data/data.rowsum
apply(data.norm, 1, sum) # The rows sum to 1, so the data are normalized

## [1] 0.6506 1.7347 0.7661 1.4360 1.5307 0.7358 1.2605 1.5318 1.5810 1.3170
```

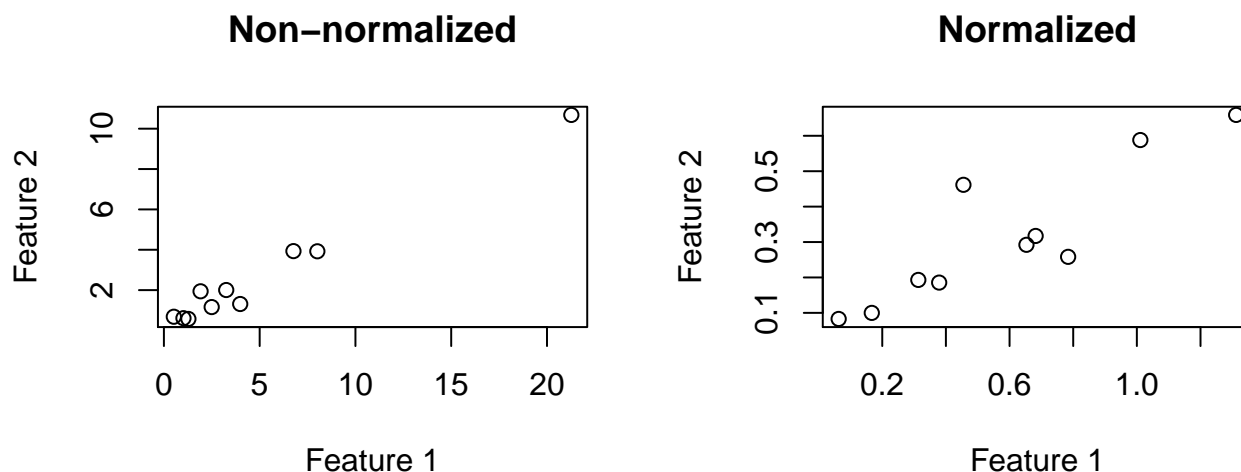


Figure 6: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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.input <- data.norm

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

test.output <- nc.score(x = test.input, bins = c(0.001, 0.1, 0.25, 0.6))

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")
```

```
test.output

##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1           NA  0.648649  0.147609 -0.002079
## Feature 2  0.648649           NA -0.008316 -0.147609
## Feature 3  0.147609 -0.008316           NA -0.399168
## Feature 4 -0.002079 -0.147609 -0.399168           NA
```

## 4 References

### References

Craig Bielski. Extending the Checkerboard Score to Ordinal Data: A Methodology for Detecting Species-Level Co-Variation and Co-Exclusion Patterns in the Human Microbiome. Master's thesis, Harvard School of Public Health, Boston, USA, 2013.

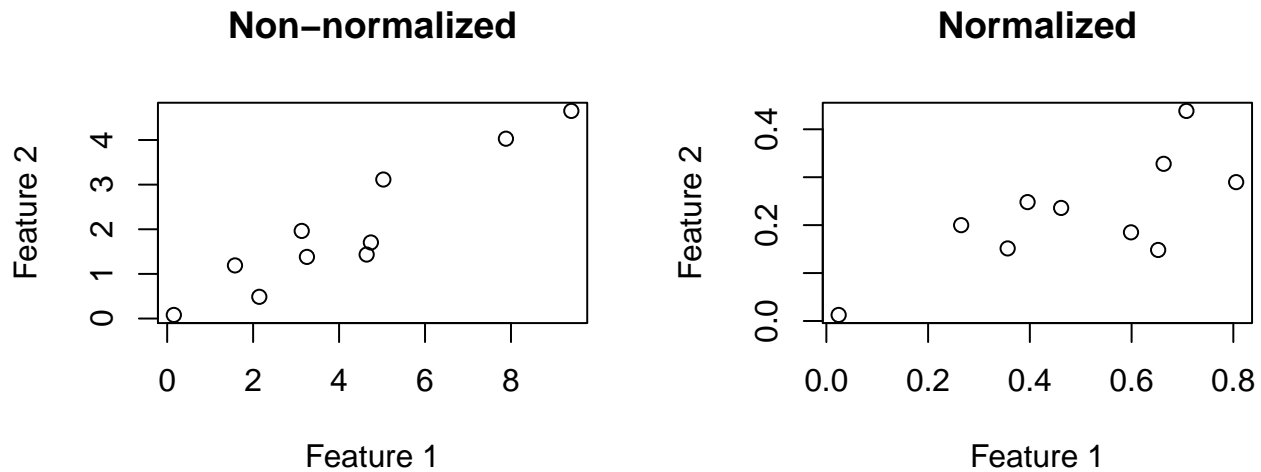


Figure 7: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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 ( $[0.6,1)$ ).

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.

Emma Schwager and Colleagues. Detecting statistically significant associations between sparse and high dimensional compositional data. In Progress.