CCREPE: Compositionality Corrected by PErmutation and REnormalization

Emma Schwager, George Weingart, Craig Bielski, Curtis Huttenhower

October 14, 2015

Contents

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`).

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.

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

`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")
```

`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.

`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 simliarity 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")
```

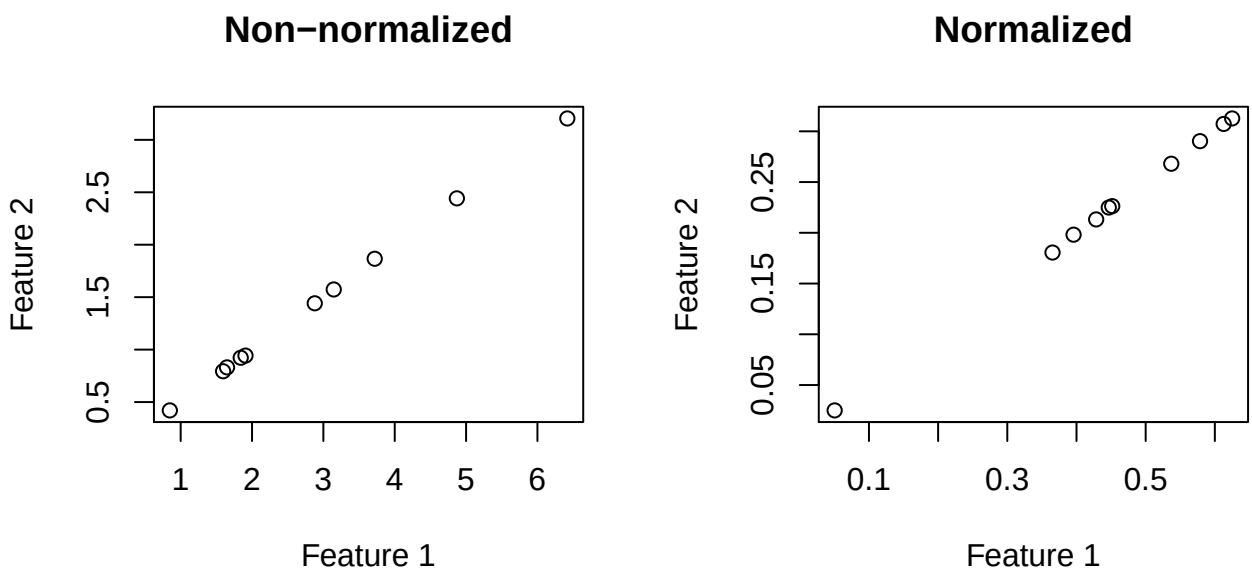

Figure 1: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive sim.score value in the [1,2] element of test.output$sim.score and the small q-value in the [1,2] element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

```
test.output
```

```
## $p.values
##              Feature 1    Feature 2 Feature 3 Feature 4
## Feature 1           NA 7.638084e-09 0.5909586 0.3831658
## Feature 2 7.638084e-09           NA 0.9947114 0.0252325
## Feature 3 5.909586e-01 9.947114e-01        NA 0.2235729
## Feature 4 3.831658e-01 2.523250e-02 0.2235729        NA
##
## $z.stat
##            Feature 1    Feature 2    Feature 3  Feature 4
## Feature 1         NA  5.776256641  0.537447384 -0.8720781
## Feature 2  5.7762566           NA  0.006628372 -2.2378245
## Feature 3  0.5374474  0.006628372           NA -1.2170822
## Feature 4 -0.8720781 -2.237824486 -1.217082181         NA
##
## $sim.score
##            Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1         NA  0.9999278 -0.1469420 -0.8891205
## Feature 2  0.9999278         NA -0.1509028 -0.8872599
## Feature 3 -0.1469420 -0.1509028         NA -0.3220418
## Feature 4 -0.8891205 -0.8872599 -0.3220418         NA
##
## $q.values
##              Feature 1    Feature 2 Feature 3 Feature 4
## Feature 1           NA 1.085666e-07  1.679960 1.3615653
## Feature 2 1.085666e-07           NA  2.356446 0.1793255
## Feature 3 1.679960e+00 2.356446e+00        NA 1.0592775
```

```
## Feature 4 1.361565e+00 1.793255e-01  1.059278        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
```

```
##  [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",se

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

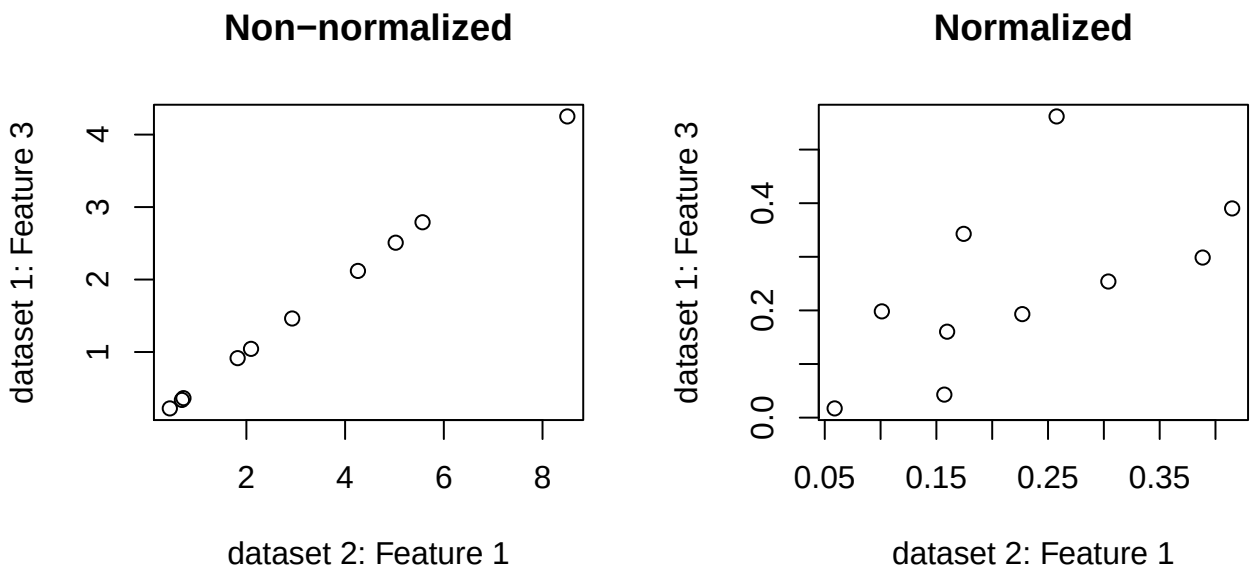

Figure 2: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive sim.score value in the [1,2] element of test.output$sim.score and the small q-value in the [1,2] element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

```
test.output.two.datasets
```

```
## $p.values
##            Feature 1 Feature 2  Feature 3 Feature 4 Feature 5  Feature 6
## Feature 1 0.32389253 0.6409007 0.16425440 0.2565426 0.6451894 0.85842939
## Feature 2 0.13376616 0.9355872 0.02920411 0.4870899 0.4047420 0.84744765
## Feature 3 0.04588306 0.8276569 0.07285171 0.3420442 0.2928063 0.24870421
## Feature 4 0.74813661 0.5790255 0.82824599 0.9658858 0.9488420 0.03732019
##            Feature 7
## Feature 1 0.09022877
## Feature 2 0.06454453
## Feature 3 0.20736690
## Feature 4 0.94602399
##
## $z.stat
##            Feature 1   Feature 2  Feature 3   Feature 4   Feature 5
## Feature 1 -0.9864904  0.46643989  1.3909045  1.13460086  0.46045540
## Feature 2 -1.4994144  0.08081735  2.1807203  0.69494480  0.83318209
## Feature 3  1.9964675  0.21770761 -1.7937586 -0.95013403 -1.05198535
## Feature 4 -0.3210974 -0.55480877 -0.2169517 -0.04276888  0.06416109
##            Feature 6   Feature 7
## Feature 1 -0.1783738 -1.69419218
## Feature 2 -0.1923761 -1.84839976
## Feature 3 -1.1535025  1.26083890
## Feature 4  2.0822440 -0.06770058
##
## $sim.score
##            Feature 1   Feature 2   Feature 3   Feature 4   Feature 5
```

```
## Feature 1 -0.3916373  0.10397798  0.62935628  0.23876635  0.21473190
## Feature 2 -0.3980905  0.10326128  0.62768339  0.25296629  0.20986829
## Feature 3  0.6105359 -0.03643048 -0.66778314 -0.35876969 -0.25328453
## Feature 4 -0.1883647 -0.08841107 -0.03996805  0.09565693  0.01637209
##             Feature 6  Feature 7
## Feature 1  0.07998547 -0.5337251
## Feature 2  0.08467410 -0.5318293
## Feature 3 -0.44863335  0.7233650
## Feature 4  0.39649728 -0.1405579
##
## $q.values
##           Feature 1 Feature 2 Feature 3 Feature 4 Feature 5 Feature 6
## Feature 1  2.727269  3.897522  2.247488  2.552920  3.717098  3.915288
## Feature 2  2.091793  4.096516  3.196792  3.332418  2.953639  4.033253
## Feature 3  1.674177  4.314212  1.594925  2.674389  2.670973  2.722410
## Feature 4  4.094692  3.728371  4.121042  3.776053  3.846808  2.042604
##           Feature 7
## Feature 1  1.646130
## Feature 2  1.766322
## Feature 3  2.522129
## Feature 4  3.982898
```

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
```

```
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")
```

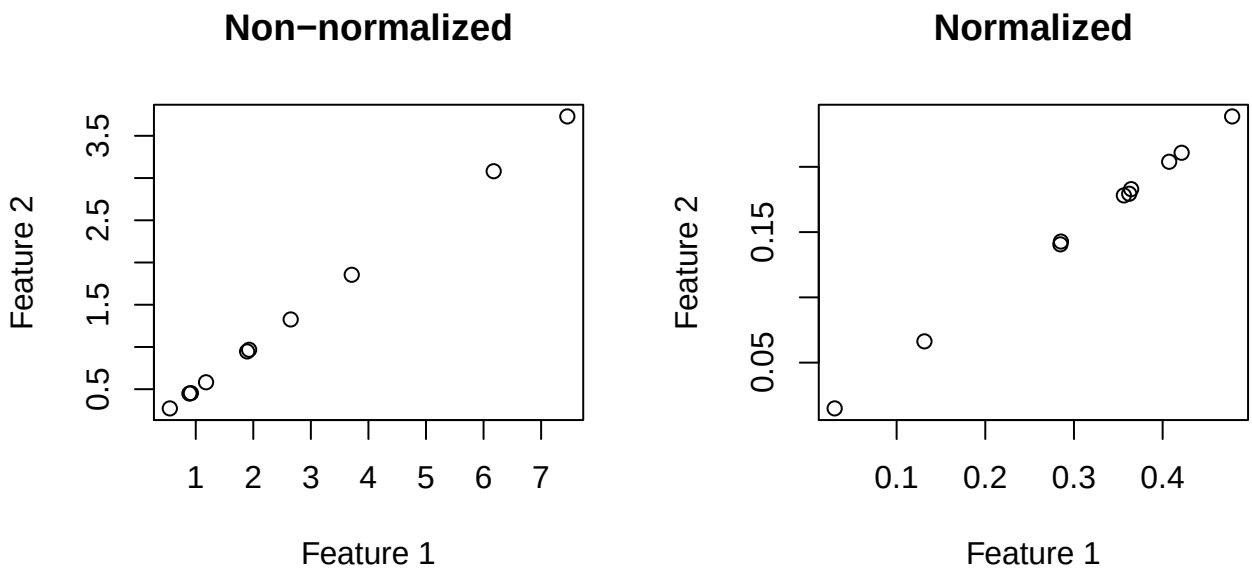

Figure 3: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. In the output we see this by the positive sim.score value in the [1,2] element of test.output$sim.score and the small q-value in the [1,2] element of test.output$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 3.653350e-07 0.6473090 0.9734842
## Feature 2 3.653350e-07           NA 0.6348547 0.4842268
## Feature 3 6.473090e-01 6.348547e-01        NA 0.7265988
## Feature 4 9.734842e-01 4.842268e-01 0.7265988        NA
##
## $z.stat
##            Feature 1  Feature 2  Feature 3   Feature 4
## Feature 1         NA  5.0861864  0.4575037  0.03323872
## Feature 2 5.08618641         NA  0.4749050 -0.69952044
## Feature 3 0.45750375  0.4749050         NA -0.34965346
## Feature 4 0.03323872 -0.6995204 -0.3496535          NA
##
## $sim.score
##           Feature 1 Feature 2 Feature 3 Feature 4
## Feature 1        NA   1.00000   0.21875  -0.65625
## Feature 2   1.00000        NA   0.21875  -0.65625
## Feature 3   0.21875   0.21875        NA  -0.50000
## Feature 4  -0.65625  -0.65625  -0.50000        NA
##
## $q.values
##              Feature 1    Feature 2 Feature 3 Feature 4
## Feature 1           NA 5.192817e-06  2.300188  2.306160
## Feature 2 5.192817e-06           NA  3.007910  3.441364
```

```
## Feature 3 2.300188e+00 3.007910e+00        NA  2.065553
## Feature 4 2.306160e+00 3.441364e+00  2.065553        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,
```

```
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")
```

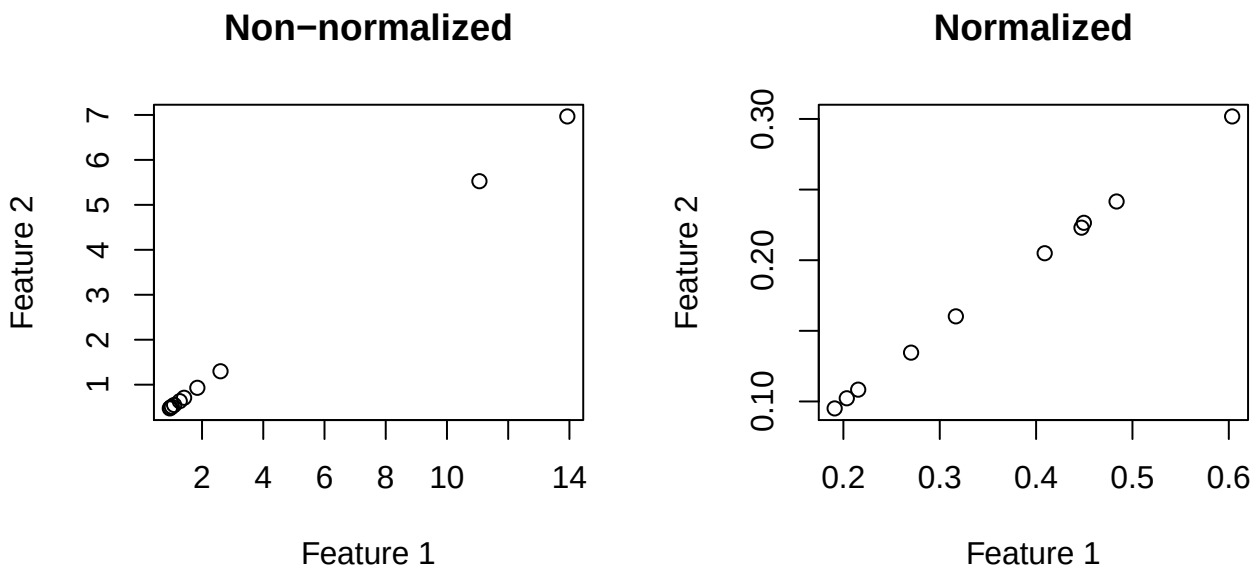

Figure 4: Non-normalized and normalized associations between feature 1 and feature 2. In this case we would expect feature 1 and feature 2 to be associated. Note that the values of sim.score are all 0.6 and none of the p-values are very small because of the arbitrary definition of the similarity score.

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

```
##            Feature 1  Feature 3
## Feature 1         NA -0.2716409
## Feature 3 -0.2716409         NA
```

```
test.output.1$sim.score
```

```
##            Feature 1  Feature 2  Feature 3  Feature 4
## Feature 1         NA -0.2136514 -0.2716409 -0.4076701
## Feature 2 -0.2136514         NA         NA         NA
## Feature 3 -0.2716409         NA         NA         NA
## Feature 4 -0.4076701         NA         NA         NA
```

```
test.output.12.3$sim.score
```

```
##            Feature 1  Feature 2  Feature 3 Feature 4
## Feature 1         NA         NA -0.2716409        NA
## Feature 2         NA         NA -0.5838431        NA
## Feature 3 -0.2716409 -0.5838431         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 τ 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 $\{1, ..., b\}$. 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 τ. Considering a 2×2 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 τ 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 givinga method for computing covariances in the presence of missing values. This must be (an abbreviaion of) on 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] 1.4453950 1.6874322 1.0769828 0.7776566 1.7657644 0.9700891 1.2354783
##  [8] 1.0584721 0.5501648 0.9547496
```

```
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")
```

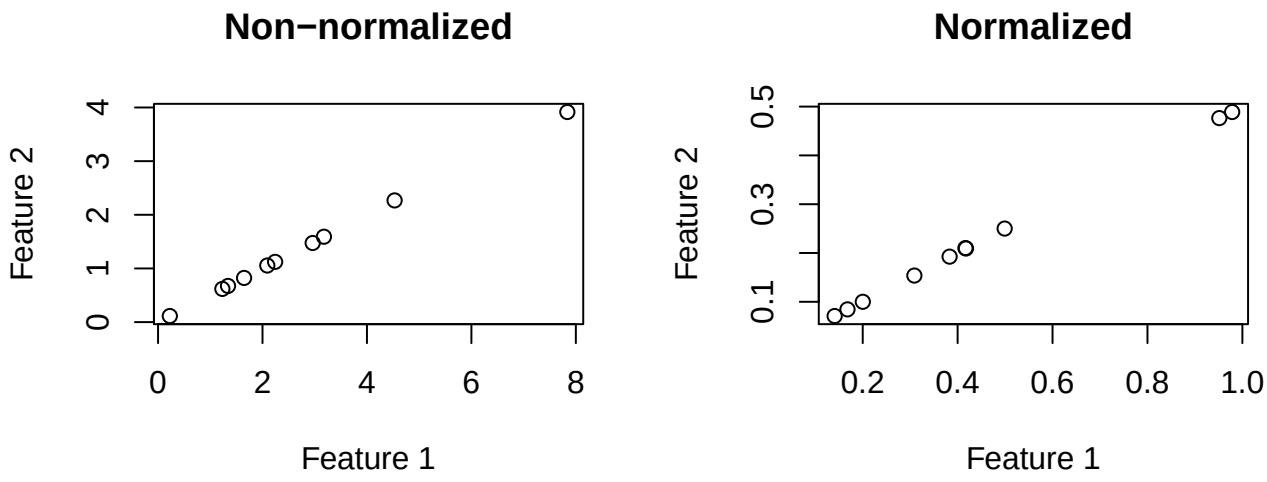

Figure 5: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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    1.0000   0.81250  -0.12500    -0.250
## Feature 2    0.8125   1.00000  -0.34375     0.000
## Feature 3   -0.1250  -0.34375   1.00000    -0.625
## Feature 4   -0.2500   0.00000  -0.62500     1.000
```

```
test.output.num
```

```
## [1] 0.8125
```

3.6 Example 2

An example of using `nc.score` with an aribitrary 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] 1.1748823 1.1501046 0.9682690 1.5749504 1.5322169 1.3085182 1.0806221
##  [8] 1.2468083 0.2669631 0.1386566
```

```
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")
```

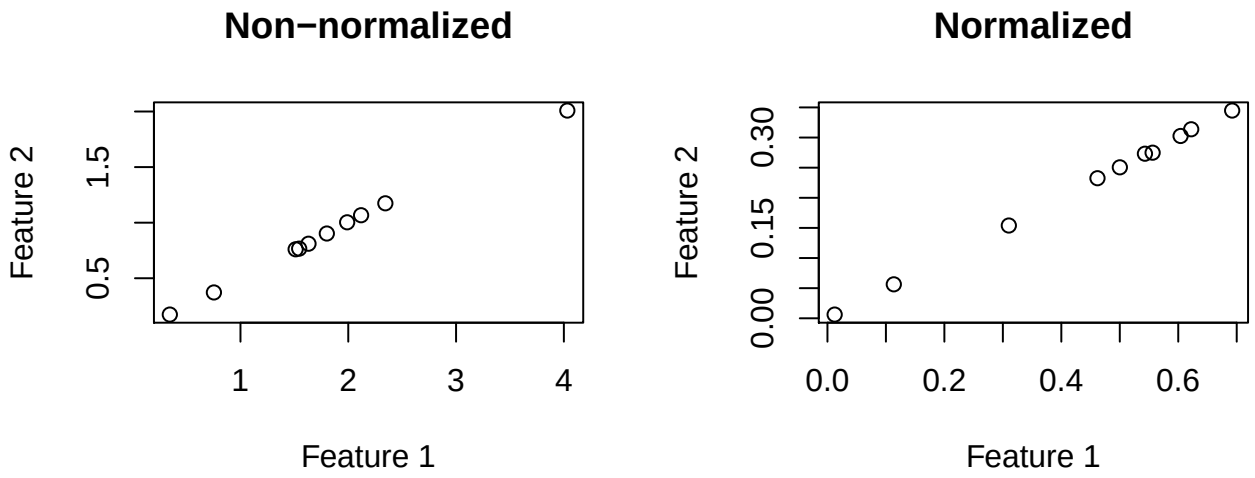

Figure 6: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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 1.0000000 1.0000000  0.2571429  0.6000000
## Feature 2 1.0000000 1.0000000  0.2571429  0.6000000
## Feature 3 0.2571429 0.2571429  1.0000000 -0.1142857
## Feature 4 0.6000000 0.6000000 -0.1142857  1.0000000
```

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] 1.8313386 0.5177427 2.0603321 2.0015454 1.9539657 0.5583940 1.3960397
##  [8] 1.9816574 1.4502365 1.6233236
```

```
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")
```

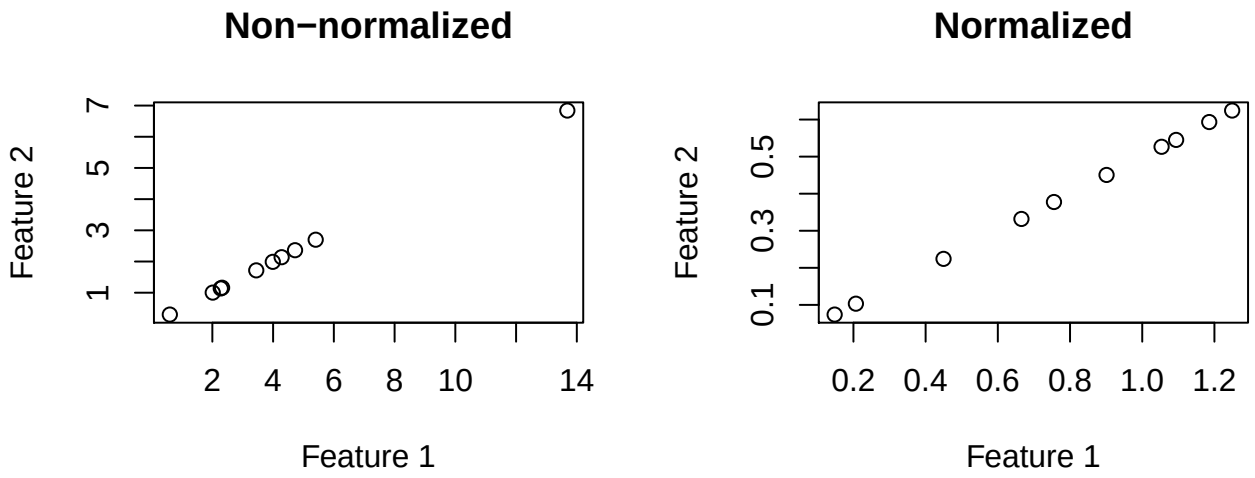

Figure 7: Non-normalized and normalized associations between feature 1 and feature 2 of the second example. Again, we expect to observe a positive association between feature 1 and feature 2. In terms of generalized checkerboard scores, we would expect to see more co-variation patterns than co-exclusion 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 1.0000000  0.8416254  0.12862394  0.00000000
## Feature 2 0.8416254  1.0000000  0.28867513 -0.20701967
## Feature 3 0.1286239  0.2886751  1.00000000 -0.08964215
## Feature 4 0.0000000 -0.2070197 -0.08964215  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 associtations between sparse and high dimensional compositional data. In Progress.