User Guide: Dirichlet_LRT_test.R

October 2018

This guide outlines use of the *Dirichlet_LRT_test.R* function from DirtyGenes to perform likelihood ratio tests on Dirichlet distributed compositional data using chi-squared and randomization procedures. The function also performs goodness-of-fit testing to the Dirichlet distribution and estimates the power of future tests, assuming that the maximum likelihood estimate of Dirichlet parameters under the alternative hypothesis is true. The guide will discuss usage, dependencies, inputs and outputs.

Usage

```
dirichlet_LRT_test = function(df,randomise = NULL, ever.present = TRUE,
min.proportion = 0, col = 1, gof.sims = NULL, power.sims = NULL,
power.threshold = 0.05, power.plot.obs = NULL)
```

Dependencies

This function requires the *gtools* and *plyr* packages to be installed in R before use.

Input Arguments

- `df` The only non-optional argument. A *data.frame* in which each row provides information from a single sample. One column (preferably the first) should be of the type *factor* and give the environment the sample is taken from. The remaining columns should represent the classes of the data and contain either counts or relative abundances for each class within the given sample. Data given in the *DirtyGenes_workspace* available in this repository are all examples of acceptable inputs.

- `randomise`[1]Number of trials used to run the randomization test. If unused, the randomization test will not be performed.

- `ever.present` Logical argument to determine whether only classes that are ever-present in the population should be used. If *TRUE*, classes that are not ever-present will be aggregated into the *LRT.Other* class, created by the function. If set to *FALSE*, results may be unreliable.

- `min.proportion` Value between 0 and 1. Any class which does not meet this proportion threshold for at least one sample in the data will be aggregated into the *LRT.Other* class.

- `col` Index of the *factor* column within `df`. Set to be 1 by default.
- `gof.sims`[1] Number of trials used to run goodness-of-fit testing to the Dirichlet distribution. If unused, goodness-of-fit testing will not be performed.
- `power.sims`[1] Number of simulations used to estimate the power of the chi-squared test, assuming that the maximum likelihood estimate or Dirichlet parameters under the alternative hypothesis are true. If unused, power testing will not be performed and the `power.threshold` and `power.plot.obs` arguments become redundant.
- `power.threshold` Value between 0 and 1. Maximum p-value for which the null hypothesis is rejected for power testing simulations. Set to 0.05 by default.
- `power.plot.obs` Vector containing integer values. For each value, n, within the vector, `power.sims` simulations will be used to estimate the power of the chi-squared test assuming n samples from *every* environment. Use of this argument will also produce a bar chart showing the estimated power for each value in the vector. If this argument is unused but `power.sims` is, then a power estimate will only be given for the same sample structure as `df`.

Output

- `LRT.stat` The raw test statistic from performing the likelihood ratio that is used in both the chi-squared and randomization tests. The value is labelled D in the DirtyGenes paper.
- `chisq.p` p-value from the chi-squared test.
- `rand.p` p-value from the randomization test. *NULL* if `randomise` is unused.
- `null.parameters` Maximum likelihood estimate of Dirichlet distribution parameters assuming no difference between environments (H_0).
- `alt.parameters` Maximum likelihood estimate of Dirichlet distribution parameters assuming differences between environments (H_1).
- `null.gof.p` p-value from goodness-of-fit testing to Dirichlet distribution using `null.parameters`. *NULL* if `gof.sims` is unused.
- `alt.gof.p` p-value from goodness-of-fit testing to Dirichlet distribution using `alt.parameters`. *NULL* if `gof.sims` is unused.
- `power.estimate` Estimate of the power of the chi-squared test assuming `alt.parameters` are correct. *NULL* if `power.sims` is unused.

[1]These arguments may significantly increase the running time of the code. In each case the function code in the relevant section contains a commented progress tracker line. If these arguments are used, it may be worth uncommenting these lines of code to get an indication of how quickly the function is running.

- `power.est.vector` Estimates of the power of the chi-squared test assuming `alt.parameters` are correct under each value given in `power.plot.obs` for the number of samples per environment. *NULL* if `power.sims` or `power.plot.obs` is unused.