Fitting species abundance models with maximum likelihood
Quick reference for `sads` package

Paulo Inácio Prado and Murilo Dantas Miranda
Theoretical Ecology Lab
LAGE at the Dep of Ecology, USP, Brazil
`http://ecologia.ib.usp.br/let/`
`prado@ib.usp.br`

April, 24, 2014

1 Introduction

Species abundance distributions (SADs) are one of the basic patterns of ecological communities (McGill et al., 2007). The empirical distributions are traditionally modelled through probability distributions. Hence, the maximum likelihood method can be used to fit and compare competing models for SADs. The package `sads` provides functions to fit the most used models to empirical SADs and also to evaluate fits and to compare competing models. The package also allows to simulate SADs expected from samples from communities, with and without aggregation of individuals of the same species.

2 Installation

The package is planned to be published at CRAN soon. Meanwhile you can install the working version from GitHub with the package `devtools`

```
> library(devtools)
> install_github('sads', 'piklprado')
```

And then load the package:

```
> library(sads)
```

3 Exploratory analyses

We'll use two data sets in the sads package:

```
> data(moths)# William's moth data
> data(ARN82.eB.apr77)# Arntz et al. benthos data
```

3.1 Octaves

Function `octav` tabulates the number of species in classes of logarithm of abundances at base 2 (Preston's octaves) and returns a data frame [1]:

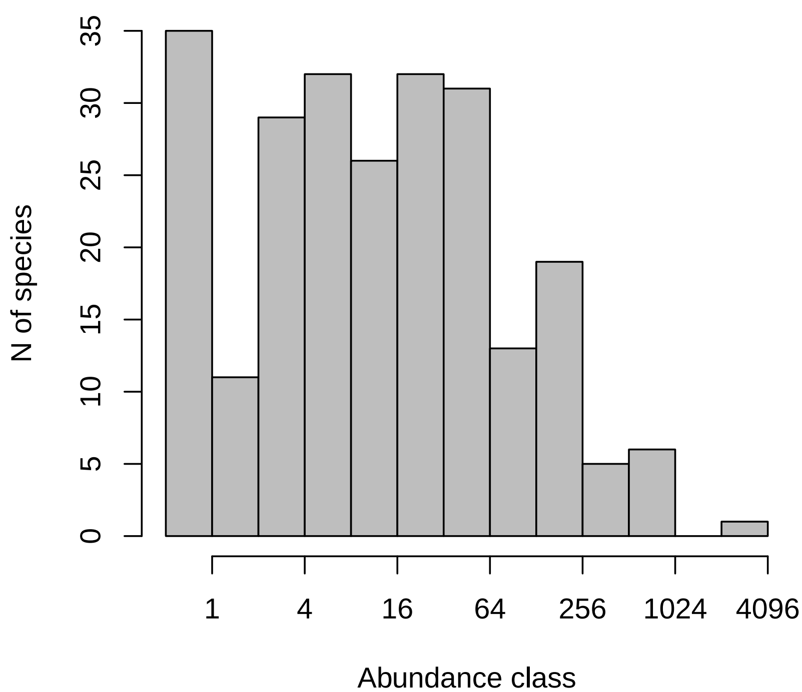
```
> (moths.oc <- octav(moths))

Object of class "octav"
   octave upper Freq
1       1     1   35
2       2     2   11
3       3     4   29
4       4     8   32
5       5    16   26
6       6    32   32
7       7    64   31
8       8   128   13
9       9   256   19
10     10   512    5
11     11  1024    6
12     12  2048    0
13     13  4096    1

> (arn.oc <- octav(ARN82.eB.apr77))

Object of class "octav"
   octave      upper Freq
1      -5   0.015625    3
2      -4   0.031250    5
3      -3   0.062500    4
4      -2   0.125000    6
5      -1   0.250000    3
6       0   0.500000    5
```

[1] actually an object of class *octav* which inherits from class *dataframe*

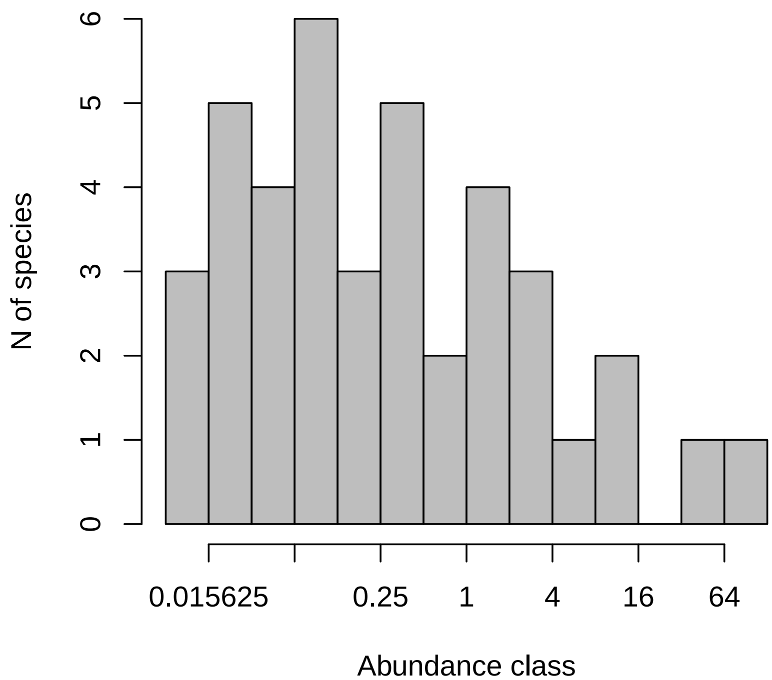
```
7        1   1.000000    2
8        2   2.000000    4
9        3   4.000000    3
10       4   8.000000    1
11       5  16.000000    2
12       6  32.000000    0
13       7  64.000000    1
14       8 128.000000    1
```

A logical argument `preston` allows to smooth the numbers as proposed by Preston (1948). The octave number is the lower limit of the class in log2 scale. Hence, for abundance values smaller than one (*e.g.* biomass data) the octave numbers are negative numbers. A Preston plot is a histogram of this table, that you get applying the function `plot` to the data frame:

```
> plot(moths.oc)
```



```
> plot(arn.oc)
```



3.2 Rank-abundance plots

Function `rad` returns a data frame of sorted abundances and their ranks [2]:

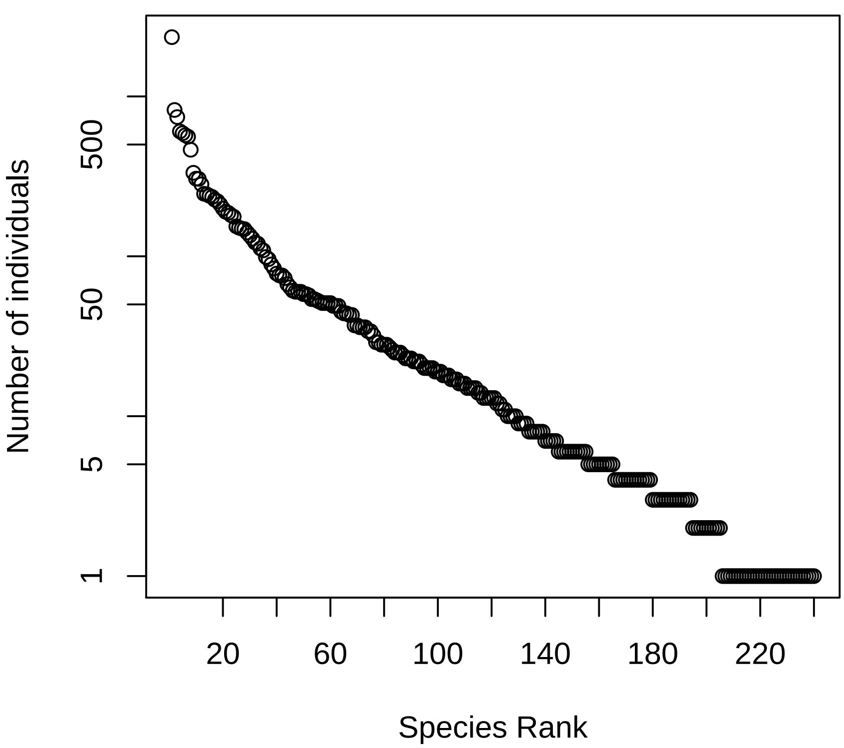
```
> head(moths.rad <- rad(moths))

  rank abund
1    1  2349
2    2   823
3    3   743
4    4   604
5    5   589
6    6   572
```

[2] actually an object of class *rad* which inherits from class *dataframe*

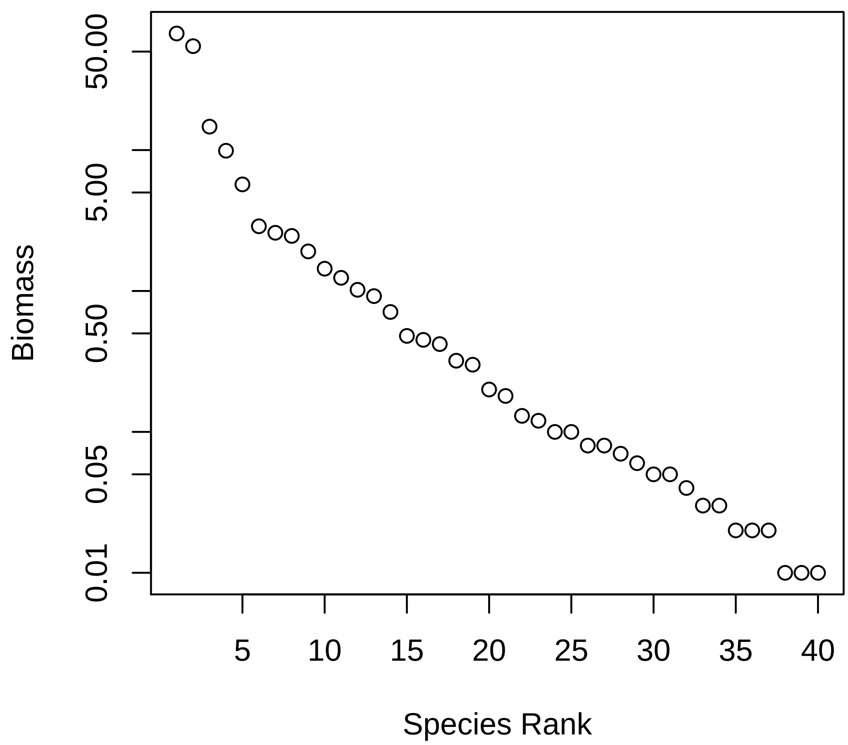
```
> head(arn.rad <- rad(ARN82.eB.apr77))

     rank abund
sp17    1 67.21
sp11    2 54.67
sp33    3 14.67
sp9     4  9.90
sp30    5  5.71
sp10    6  2.88
```

To get the rank-abundance or Whitaker's plot apply the function `plot` on the data frame:

```
> plot(moths.rad, ylab="Number of individuals")
```



```
> plot(arn.rad, ylab="Biomass")
```



4 Model fitting

The *sads* package provides maximum-likelihood fits of many probability distributions to empirical sads. The working horse is the functions `fitsad` for fitting species abundance distributions and `fitrad` for fitting rank-abundance distributions. The first argument of these functions is the vector of observed abundances and the second argument is the name of the model to be fitted. Please refer to the help page of the functions for details on the models. For more information on the fitting procedure see also the vignette of the *bbmle* package, on top of which the package *sads* is built.
To fit a logseries distribution use the argument `sad='ls'`:

```
> (moths.ls <- fitsad(moths,'ls'))
```

```
Call:
mle2(minuslogl = LL, start = list(alpha = alfa), method = "Brent",
    data = list(x = x), lower = 0, upper = upper)

Coefficients:
   alpha
40.24728

Log-likelihood: -1087.71
```

The resulting model object inherits from *mle2* (Bolker & R Development Core Team, 2012), and all usual methods for model objects, such as summaries, log-likelihood, and AIC values:

```
> summary(moths.ls)

Maximum likelihood estimation

Call:
mle2(minuslogl = LL, start = list(alpha = alfa), method = "Brent",
    data = list(x = x), lower = 0, upper = upper)

Coefficients:
      Estimate Std. Error z value     Pr(z)
alpha   40.247      6.961  5.7818 7.391e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 2175.425

> coef(moths.ls)

   alpha
40.24728

> logLik(moths.ls)

'log Lik.' -1087.713 (df=1)

> AIC(moths.ls)

[1] 2177.425
```

4.1 Model diagnostics

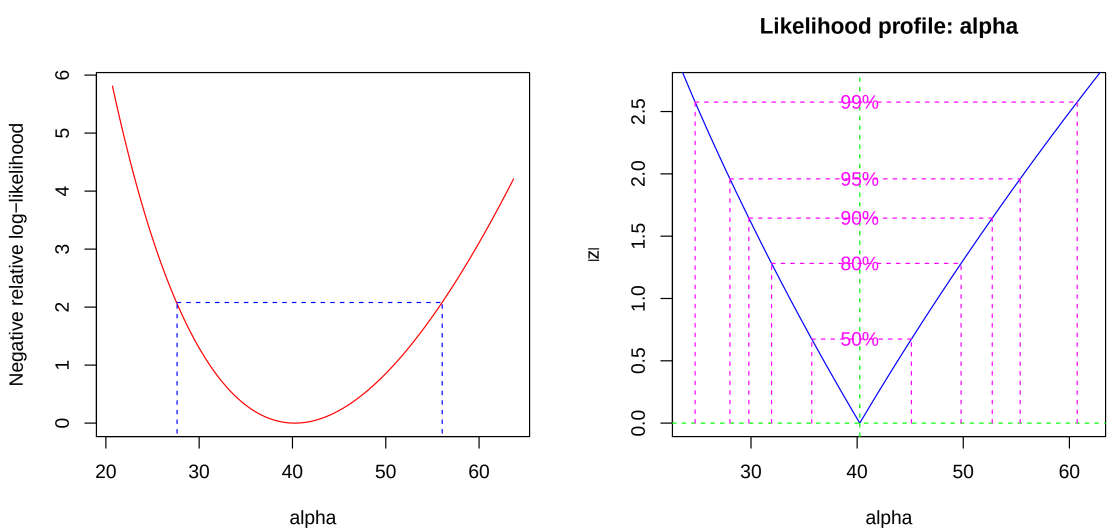
Many other diagnostic and functions are available for sad and rad models. To get likelihood profiles and confidence intervals use:

```
> moths.ls.prf <- profile(moths.ls)
> confint(moths.ls.prf) # conf intervals
```

```
   2.5 %   97.5 %
28.01537 55.36267
```

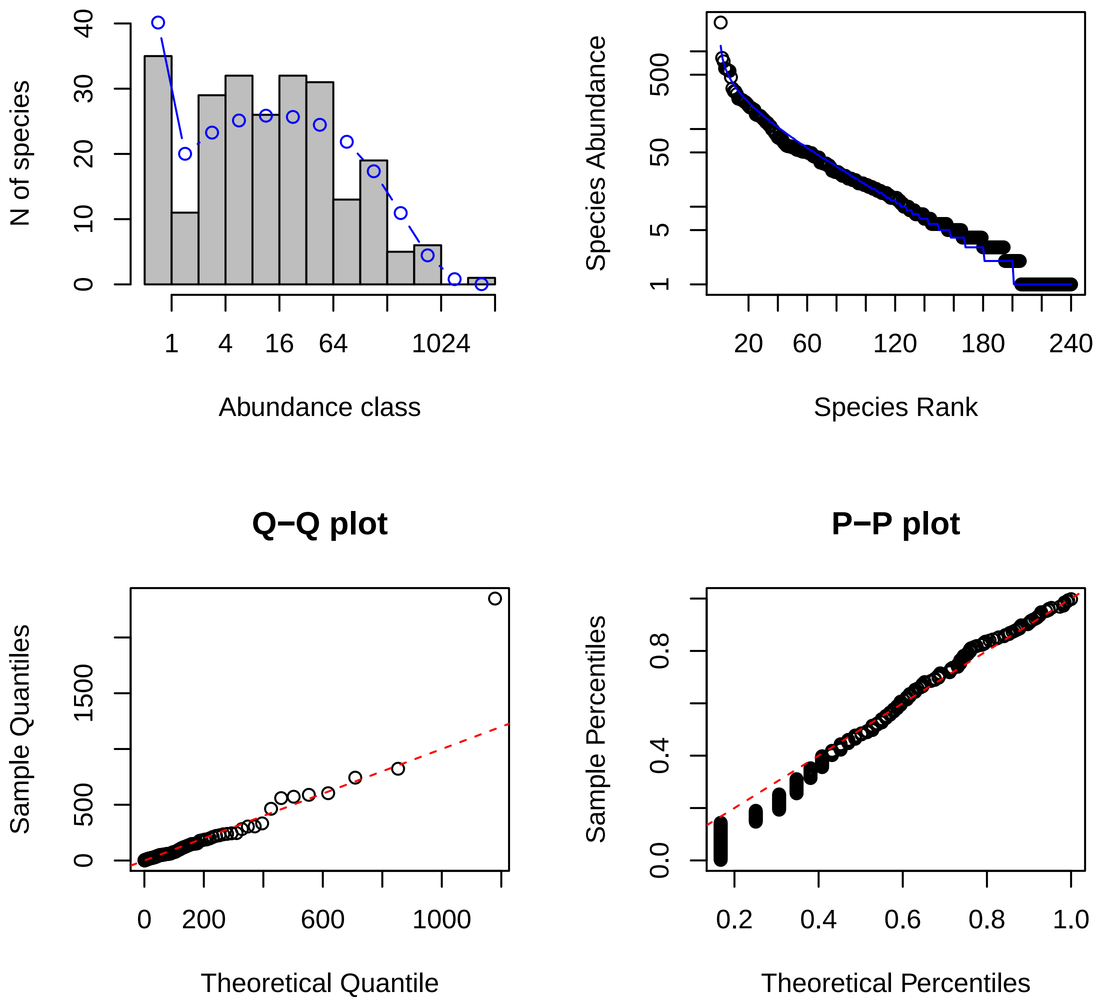
And then use `plotprofmle` to plot likelihood profiles at the original scale (relative negative log-likelihood) and function `plot` to get plots at chi-square scale (square-root of twice the relative log-likelihood):

```
> par(mfrow=c(1,2))
> plotprofmle(moths.ls.prf)# log-likelihood profile
> plot(moths.ls.prf)# z-transformed profile
> par(mfrow=c(1,1))
```



When applied on a sad model object the function `plot` returns four diagnostic plots:

```
> par(mfrow=c(2,2))
> plot(moths.ls)
> par(mfrow=c(1,1))
```



The first two plots (top right and left) are the octave and rank-abundance plots with the predicted values of number of species in each octave and abundances of each species. The two last plots (bottom) are quantile-quantile and percentile-percentile graphs of the observed vs. predicted abundances. The straight line indicates the expected relation in case of perfect fit.

4.2 Rank-abundance models

Species-abundance models assigns a probability to each abundance value. Hence these models are probability density functions (PDFs) of abundances of a species taken at random from the community. Rank-abundance models assigns a probability for each **abundance rank**. They are PDFs for rankings. The models are interchangeable (May, 1975), but currently only four rad models are available in package sads trough the argument `rad` of function `fitrad`:

- "gs": geometric series (which is NOT geometric PDF, available in `fitsads` as "geom";
- "rbs": broken-stick model (MacArthur, 1957; May, 1975)
- "zipf": zipf power-law distribution
- "mand": zipf-mandelbrot power-law distribution

Comparison to `radfit` from *vegan* package:
fits by `fitsad`, `fitrad` and `radfit` of *vegan* package provide similar estimates of model coefficients but not comparable likelihood values. This is because each function fit models that assigns probability to data in different ways. Function `fitsad` fit PDFs to observed abundances and `fitrad` fit PDFs to the ranks of the abundances. Finally, `radfit` fits a Poisson generalized linear model to the *expected abundances* deduced from rank-abundance relationships from the correspodending sads and rads models (Wilson, 1991). See also the help page of `radfit`. Therefore **likelihoods obtained from these three functions are not comparable**.

5 Model selection

You can fit other models to the same data set, such as the Poisson-lognormal and a truncated lognormal:

```
> (moths.pl <- fitsad(x=moths,sad="poilog"))#default is zero-truncated

Call:
mle2(minuslogl = LL, start = as.list(pl.par), data = list(x = x))

Coefficients:
      mu      sig
1.996469 2.187126

Log-likelihood: -1086.07
```

```
> (moths.ln <- fitsad(x=moths,sad="lnorm", trunc=0.5)) # lognormal truncated at 0.5

Call:
mle2(minuslogl = LL, start = list(meanlog = meanlog, sdlog = sdlog),
    data = list(x = x))

Coefficients:
 meanlog    sdlog
2.274346 2.039740

Log-likelihood: -1085.47
```

and then you can use the function `AICtab` and friends from the *bbmle* package to get a model selection table:

```
> AICtab(moths.ls, moths.pl, moths.ln, base=TRUE)

         AIC    dAIC   df
moths.ln 2174.9    0.0 2
moths.pl 2176.1    1.2 2
moths.ls 2177.4    2.5 1
```

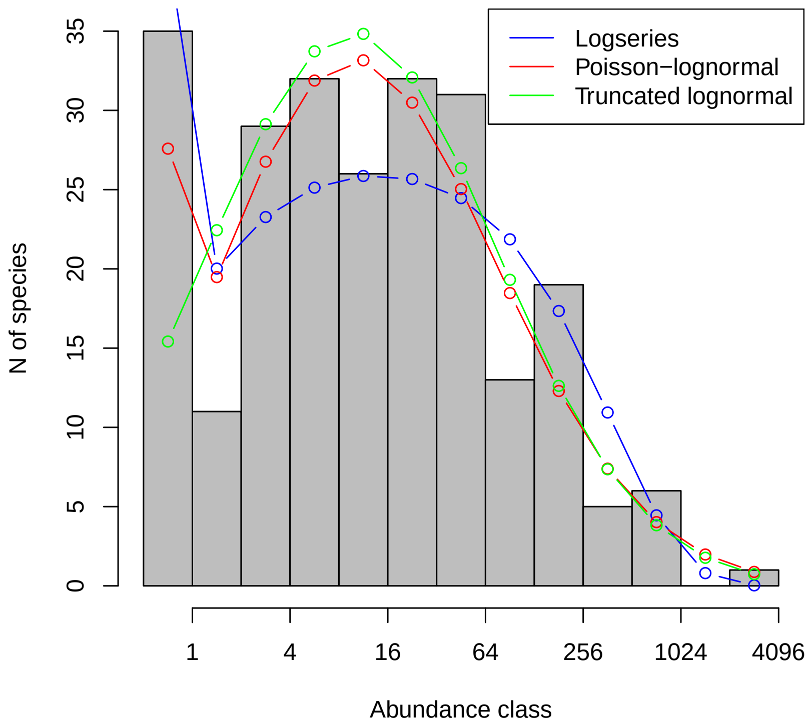
To compare visually fits first get octave tables:

```
> head(moths.ls.oc <- octavpred(moths.ls))

  octave upper     Freq
1      1     1 40.14377
2      2     2 20.02026
3      3     4 23.27123
4      4     8 25.12674
5      5    16 25.86285
6      6    32 25.67116

> head(moths.pl.oc <- octavpred(moths.pl))

  octave upper     Freq
1      1     1 27.58735
2      2     2 19.48216
3      3     4 26.76472
4      4     8 31.88374
5      5    16 33.16140
6      6    32 30.49061
```

```
> head(moths.ln.oc <- octavpred(moths.ln))

  octave upper     Freq
1      1     1 15.41886
2      2     2 22.44066
3      3     4 29.13034
4      4     8 33.72746
5      5    16 34.82976
6      6    32 32.08088
```

and then use `lines` to superimpose the predicted values in the octave plot:

```
> plot(moths.oc)
> lines(moths.ls.oc, col="blue")
> lines(moths.pl.oc, col="red")
> lines(moths.ln.oc, col="green")
> legend("topright",
        c("Logseries", "Poisson-lognormal", "Truncated lognormal"),
        lty=1, col=c("blue","red", "green"))
```



To do the same with rank-abundance plots get the rank-abundance objects:

```
> head(moths.ls.rad <- radpred(moths.ls))

  rank abund
1    1  1180
2    2   854
3    3   710
4    4   619
5    5   554
6    6   503

> head(moths.pl.rad <- radpred(moths.pl))

  rank abund
1    1  4348
```

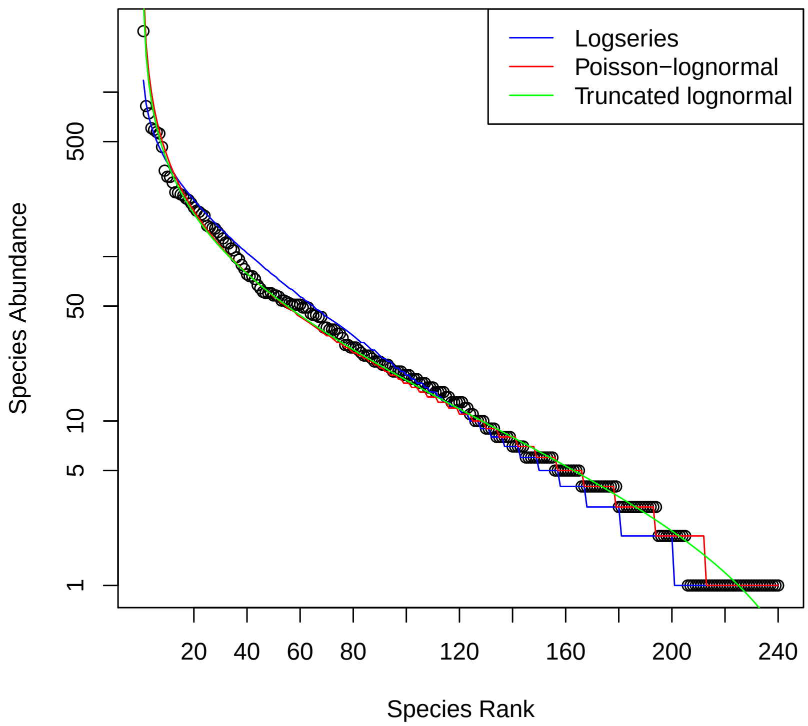
```
2     2  1973
3     3  1322
4     4  1001
5     5   807
6     6   676
```

```
> head(moths.ln.rad <- radpred(moths.ln))
```

```
  rank     abund
1    1 3524.2394
2    2 1674.8603
3    3 1148.3539
4    4  883.6309
5    5  720.7864
6    6  609.2707
```

and then plot observed and predicted values:

```
> plot(moths.rad)
> lines(moths.ls.rad, col="blue")
> lines(moths.pl.rad, col="red")
> lines(moths.ln.rad, col="green")
> legend("topright",
        c("Logseries", "Poisson-lognormal", "Truncated lognormal"),
        lty=1, col=c("blue","red", "green"))
```



6 Simulations

The function `rsad` returns random samples of a community which has S species. The mean abundance of the species in the communities are independent identically distributed (iid) variables that follow a given probability distribution. The sample simulates a given number of draws of a fraction a of the total number of individuals of the community. For instance, to simulate two Poisson samples of 10% of a community with 10 species that follows a lognormal distribution with parameters $\mu = 3$ and $\sigma = 1.5$ use:

```
> set.seed(42)# fix random seed to make example reproducible
> (samp1 <- rsad(S = 10, frac = 0.1, sad = "lnorm", zeroes=TRUE,
                 ssize = 2, meanlog = 3, sdlog = 1.5))
```

```
   sample species abundance
1       1       1        20
```

```
2          1         2           4
3          1         3           7
4          1         4           2
5          1         5           4
6          1         6           1
7          1         7          25
8          1         8           3
9          1         9          45
10         1        10           1
11         2         1          17
12         2         2           2
13         2         3           0
14         2         4           3
15         2         5           6
16         2         6           2
17         2         7          18
18         2         8           0
19         2         9          53
20         2        10           4
```

The function returns a data frame with a sample numeric label, species numeric label and species abundance in each sample. By default, `rsad` returns a vector of abundances of single Poisson sample with zeroes ommited:

```
> (samp2 <- rsad(S = 100, frac=0.1, sad="lnorm",
               meanlog=5, sdlog=2))

 [1]  155  697    4    7   48    5   40   56  105    8   48
[12]    1    3    1   14   21    6   66    2    3   32  259
[23]    8   51   21    1  312   42   23   20   48   12   28
[34]   14   20   40  267    5  209   36  107   93   58    1
[45]    7   39    2    7   56   70   31    3    4  305   25
[56]   15   12    3   48    8   12  101   69  255    5   51
[67]  253    4    1    2   17   49  187  121  599    3   23
[78]   12    9   16   21   10   17    3    5    2    9    5
[89] 3214    1   19    1   31
```

Once this is a Poisson sample of a lognormal community, the abundances in the sample should follow a Poisson-lognormal distribution with parameters $\mu + \log a$ and σ (Grøtan & Engen, 2008). We can check this by fitting a Poisson-lognormal model to the sample:

```
> (samp2.pl <- fitsad(samp2, 'poilog'))

Call:
mle2(minuslogl = LL, start = as.list(pl.par), data = list(x = x))

Coefficients:
      mu      sig
2.709138 1.884220

Log-likelihood: -453.22

> ## checking correspondence of parameter mu
> coef(samp2.pl)[1] - log(0.1)

      mu
5.011723
```

Not bad. By repeating the sampling and the fit many times you can evaluate the bias and variance of the maximum likelihood estimates:

```
> results <- matrix(nrow=500,ncol=2)
> for(i in 1:500){
     x <- rsad(S = 100, frac=0.1, sad="lnorm",
               meanlog=5, sdlog=2)
     y <- fitsad(x, "poilog")
     results[i,] <- coef(y)
 }
> results[,1] <- results[,1]-log(0.1)
```

Bias is estimated as the difference between the mean of estimates and the value of parameters

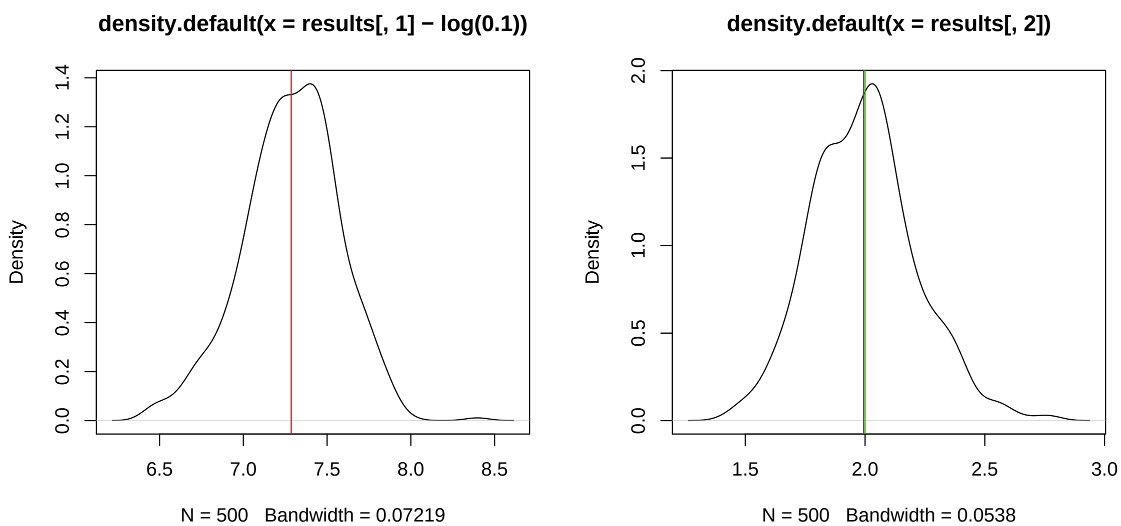
```
> ##Mean of estimates
> apply(results,2,mean)

[1] 4.983124 1.993854

> ## relative bias
> (c(5,2)-apply(results,2,mean))/c(5,2)

[1] 0.003375221 0.003072756
```

And the precision of the estimates are their standard deviations

```
> ##Mean of estimates
> apply(results,2,sd)

[1] 0.2869798 0.2180105

> ## relative precision
> apply(results,2,sd)/apply(results,2,mean)

[1] 0.05759033 0.10934122
```

Finally, a density plot with lines indicating the mean of estimates and the values of parameters:

```
> par(mfrow=c(1,2))
> plot(density(results[,1]-log(0.1)))
> abline(v=c(mean(results[,1]-log(0.1)),5), col=2:3)
> plot(density(results[,2]))
> abline(v=c(mean(results[,2]), 2), col=2:3)
> par(mfrow=c(1,1))
```



References

Bolker, B. & R Development Core Team, 2014. bbmle: Tools for general maximum likelihood estimation. R package version 1.0.16.

Grøtan, V. & S. Engen, 2008. poilog: Poisson lognormal and bivariate Poisson lognormal distribution. R package version 0.4.

MacArthur, R., 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America* **43**:293.

May, R. M., 1975. Patterns of species abundance and diversity. In M. L. Cody & J. M. Diamond, editors, *Ecology and Evolution of Communities*, chapter 4, pages 81–120. Harvard University Press, Cambridge, MA.

McGill, B., R. Etienne, J. Gray, D. Alonso, M. Anderson, H. Benecha, M. Dornelas, B. Enquist, J. Green, F. He, A. Hurlbert, A. E. Magurran, P. Marquet, B. Maurer, A. Ostling, C. Soykan, K. Ugland, & E. White, 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters* **10**:995–1015.

Preston, F. W., 1948. The commonness and rarity of species. *Ecology* **29**:254–283.

Wilson, J., 1991. Methods for fitting dominance/diversity curves. *Journal of Vegetation Science* **2**:35–46.