Fitting parametric univariate distributions to non censored or censored data using the R `fitdistrplus` package

Marie Laure Delignette-Muller and Christophe Dutang

July 19, 2012

Contents

1 Introduction **1**
- 1.1 Overview . . . 1
- 1.2 Running examples . . . 2

2 Choice of candidate distributions **2**
- 2.1 Graphical display of the observed distribution . . . 3
- 2.2 Empirical basis for selecting candidate distributions . . . 3

3 Fit of a distribution by maximum likelihood estimation **6**
- 3.1 Parameter estimation . . . 6
- 3.2 Goodness-of-fit plots . . . 7
- 3.3 Measures of goodness-of-fit . . . 8
- 3.4 Goodness-of-fit tests . . . 9

4 The special case of censored data **10**
- 4.1 Graphical display of the observed distribution . . . 10
- 4.2 Maximum likelihood estimation . . . 10
- 4.3 Goodness-of-fit plot . . . 11

5 Alternative methods for parameter estimation **12**
- 5.1 Maximum goodness-of-fit estimation . . . 12
- 5.2 Moment matching estimation . . . 13
- 5.3 Quantile matching estimation . . . 13
- 5.4 Customization of the optimization algorithm . . . 14

6 Uncertainty in parameter estimates **15**
- 6.1 Bootstrap procedures . . . 15
- 6.2 Use of bootstrap samples . . . 16

1 Introduction

1.1 Overview

Fitting distributions to data is a very common task in statistics. It consists in choosing a probability distribution that gives a good representation of a statistical variable. It requires judgment and expertise and generally needs an iterative process of distribution choice, parameter estimation, and quality of fit evaluation. Function `fitdistr` in the R package `MASS` [15] is a well known general-purpose maximum-likelihood fitting routine for the parameter estimation step in R. Other steps of the process may be developed using R [12]. Our first objective by developping package `fitdistrplus` [6] was to provide R users a set of functions dedicated to help the overall process of fitting a univariate parametric distribution to data.

Function `fitdistr` estimates distribution parameters by maximizing the log-likelihood using function `optim`. In some cases, other estimation methods could be prefered, such as maximum goodness-of-fit estimation also commonly called minimum distance estimation, and proposed in package `actuar` with three different goodness-of-fit distances. While developping package `fitdistrplus`, our second objective was to extend function `fitdistr` by providing various estimation methods to fit distributions in addition to maximum likelihood. Functions were developped to enable matching moment estimation, matching quantile estimation, and maximum goodness-of-fit estimation (or minimum distance estimation) using eight different distances. Moreover, package `fitdistrplus` offers the possibility to specify a user-supplied function for optimization, useful in cases where optimization techniques not included in function `optim` may be more adequate.

In applied statistics, it is not uncommon to have to fit distributions to censored data. Function `fitdistr` does not enable maximum likelihood estimation from this type data. Some packages deal with censored data, especially survival data [13], but those packages generally focused on specific models, enabling the fit of only one distribution or a restricted family of distributions. Our third objective was thus to provide R users a function to estimate univariate distribution parameters from censored data, whatever the type of censoring.

This manuscript reviews the various features of version 0.3-4 of `fitdistrplus`. The package is available from the Comprehensive R Archive Network at `http://cran.r-project.org/package=fitdistrplus`. The development version of the package is located at R-forge as one the packages of the project "Risk Assessment with R" (`http://r-forge.r-project.org/projects/riskassessment/`) The following command will load the package.

```
> library(fitdistrplus)
```

1.2 Running examples

For illustrating the use of various functions of package `fitdistrplus`, we will use four examples published in various biological areas, corresponding to data sets included in the package.

The two first data sets correspond to the observation of a continuous variable on a random sample of a population of interest.

The "ground beef" data set contains values of serving sizes in grams, collected in a French survey, for ground beef patties consumed by children under 5 years old. This data set was used in a quantitative risk assessment published in a food microbiology journal ([5]).

```
> data(groundbeef)
> str(groundbeef)
```

```
'data.frame':        254 obs. of  1 variable:
 $ serving: num  30 10 20 24 20 24 40 20 50 30 ...
```

The "endosulfan" data set contains acute toxicity values for the organochlorine pesticide endosulfan (geometric mean of LC50 ou EC50 values in $\mu g.L^{-1}$), tested on Australian and non-Australian laboratory-species (arthropods, fish or nonarthropod invertebrates) ([8]).

```
> data(endosulfan)
> str(endosulfan)
```

```
'data.frame':        104 obs. of  3 variables:
 $ ATV       : num  0.6 2.8 182.2 0.8 478 ...
 $ Australian: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 1 ...
 $ group     : Factor w/ 3 levels "Arthropods","Fish",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The "Toxocara" data set corresponds to the observation of a discrete variable, the number of *Toxocara cati* parasites present in digestive tract, on a random sample of feral cats living on Kerguelen island ([7]).

```
> data(toxocara)
> str(toxocara)
```

```
'data.frame':        53 obs. of  1 variable:
 $ number: int  0 0 0 0 0 0 0 0 0 0 ...
```

The "smoked fish" data set corresponds to the observation of a continuous censored variable, the *Listeria monocytogenes* microbial concentration, on a random sample of smoked fish distributed on the Belgian market in the period 2005 to 2007 ([2]). Censored data are coded within 2 columns named left and right, describing each observed value of *Listeria monocytogenes* concentration (in $CFU.g^{-1}$) as an interval. The left column contains either `NA` for left censored observations, the left bound of the interval for interval censored observations, or the observed value for non-censored observations. The right column contains either `NA` for right censored observations, the right bound of the interval for interval censored observations, or the observed value for noncensored observations.

```
> data(smokedfish)
> str(smokedfish)
```

```
'data.frame':        103 obs. of  2 variables:
 $ left : num  NA NA NA NA NA NA NA NA NA NA ...
 $ right: num  0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 ...
```

2 Choice of candidate distributions

Before fitting one or more distributions to a data set, it is generally necessary to choose good candidates among a predefined family of distributions. To help the user in this preliminary task, we developed functions to plot and characterise empirical distributions.

2.1 Graphical display of the observed distribution

First of all, an empirical distribution may be plotted using classical R function or using Function `plotdist` which provides plots in density and in cdf as done in (Figure 1) for a continuous variable:

```
> plotdist(groundbeef$serving)
```

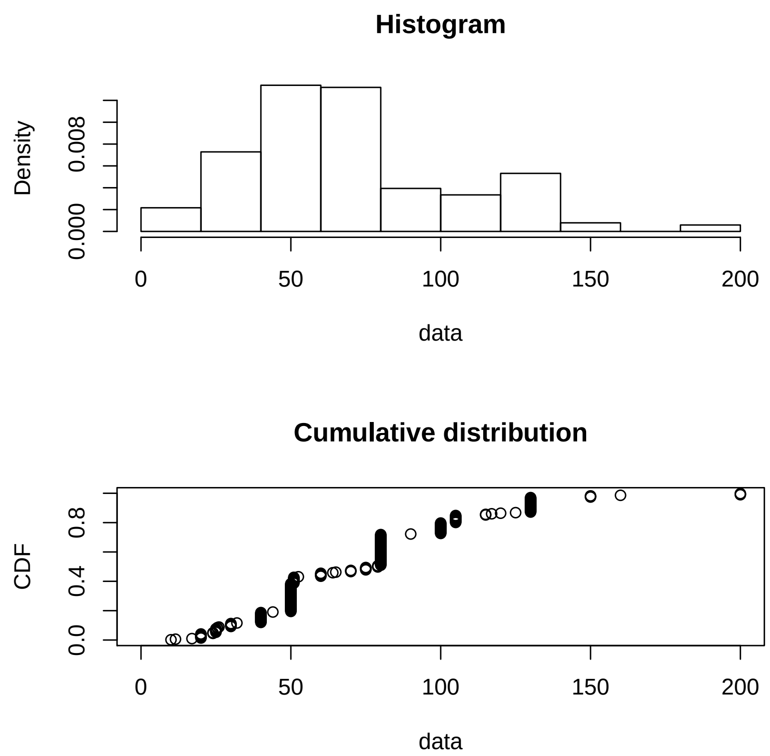

Figure 1: Density and cdf plots of an empirical distribution for a continuous variable (serving size from the "ground beef" data set)

In some cases a discrete variable may be plotted as a continuous one, for example for a large data set from a binomial distribution converging to a normal one, but Function `plotdist` also proposes specific plots in density and in cdf for discrete variables (Figure 2):

```
> plotdist(toxocara$number,discrete = TRUE)
```

2.2 Empirical basis for selecting candidate distributions

Descriptives statistics may help the choice of good candidates to describe an empirical distribution among a family of parametric distributions. Especially the skewness and kurtosis are useful for this purpose. The concept of skewness relates to deviations from symmetry of the distribution. The normal distribution has a skewness of zero. A positive (resp. negative) skewness indicates that the right (resp. left) tail of the distribution is more extended than the left (resp. right) one. The concept of kurtosis relates to the tail weight. The normal distribution has a kurtosis of 3. Distributions with a higher kurtosis are said to be leptokurtic, with heavier tails, such as the logistic distribution, while distributions with a smaller kurtosis are said platykurtic, with lighter tails, such as the uniform distribution.

Function `descdist` provides calculations of classical descriptive statistics (minimum, maximum, median, mean, sample sd) and by default unbiased estimations of skewness and Pearsons's kurtosis values. Nevertheless estimations of skewness and kurtosis are unbiased only for normal distributions and estimated values are thus only indicative. A skewness-kurtosis plot such as the one proposed by [3] is also provided for the empirical distribution (Figure 3). On this plot, values for common distributions are displayed as tools to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential for example), there is only one possible value for the skewness and the kurtosis and the distribution is thus represented by a point on the plot. For other distributions, areas of possible values are represented, consisting in lines (as for gamma and lognormal distributions), or larger areas (as for beta distribution).

Skewness and kurtosis are known not to be robust. In order to take into account the uncertainty of the estimated values of kurtosis and skewness from data, the data set may be boostraped by fixing the argument `boot` to an integer above 10. Values of skewness and kurtosis corresponding to bootstrap samples are then computed and reported on

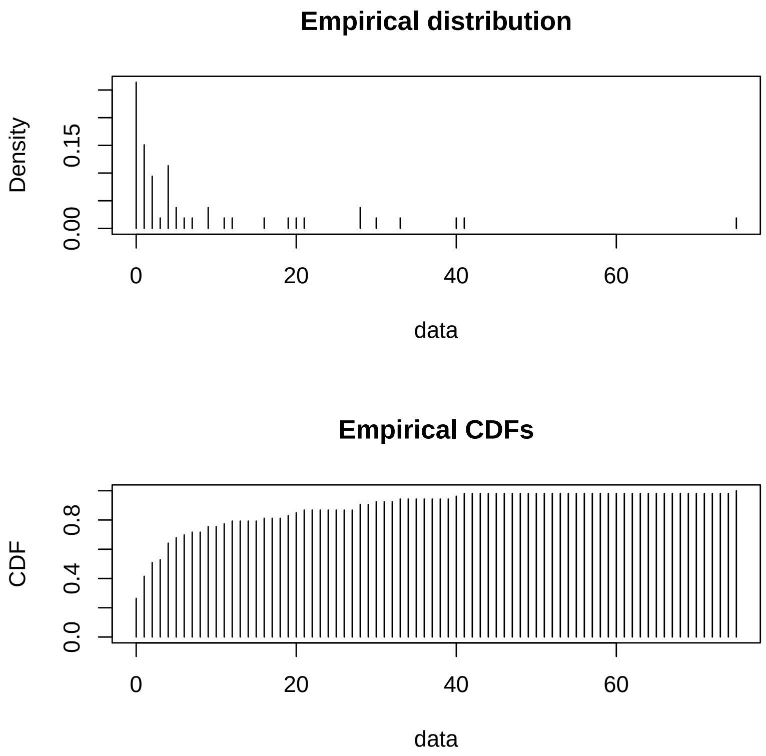

Figure 2: Density and cdf plots of an empirical distribution for a discrete variable (number of *Toxocara cati* parasites from the "Toxocara" data set)

the skewness-kurtosis plot. Below is a call to function `descdist` to describe the distribution of the serving size from the "ground beef" data set and to draw the corresponding skewness-kurtosis plot (Figure 3). Looking at the results on this example with a positive skewness and a kurtosis not far from 3, the fit of three common right-skewed distributions could be considered, Weibull, gamma and lognormal distributions.

```
> descdist(groundbeef$serving,boot=1000)

summary statistics
------
min:  10   max:  200
median:  79
mean:  73.6
estimated sd:  35.9
estimated skewness:  0.735
estimated kurtosis:  3.55
```

For discrete variables, such as the number of *Toxocara cati* parasites from the "Toxacara" data set, skewness and kurtosis values or set of values of Poisson and negative binomial distributions are represented in the skewness-kurtosis plot (Figure 4), together with values for the normal distribution, to which discrete distributions may converge. Looking at the skewness-kurtosis plot (Figure 4) obtained for the number of *Toxocara cati* parasites from the "Toxacara" data set, one could try the fit of Poisson and negative-binomial distributions.

```
> descdist(toxocara$number,discrete = TRUE,boot=1000)

summary statistics
------
min:  0   max:  75
median:  2
mean:  8.68
estimated sd:  14.3
estimated skewness:  2.63
estimated kurtosis:  11.4
```

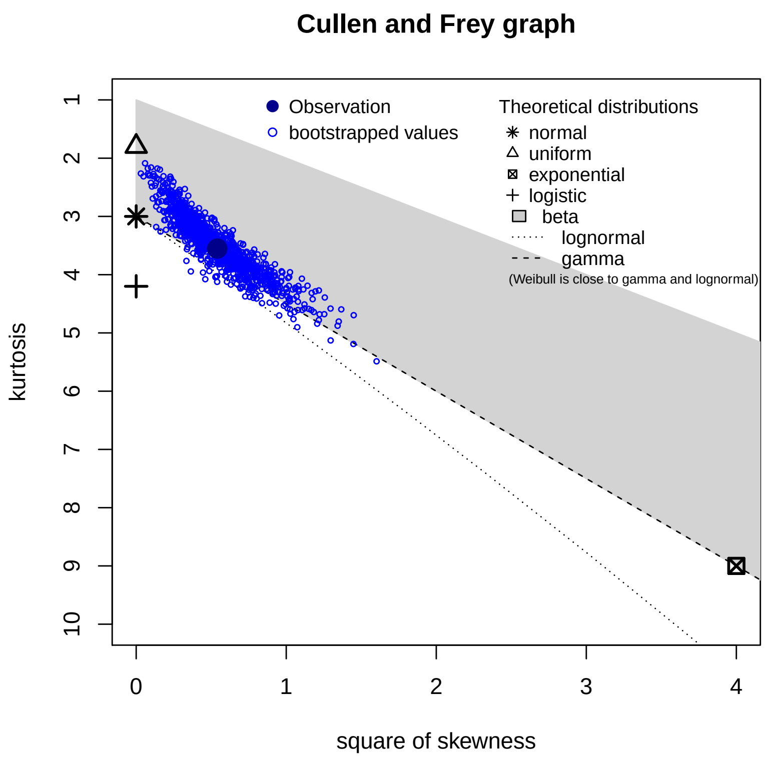

Figure 3: Skewness-kurtosis plot for a continuous variable (serving size from the “ground beef” data set)

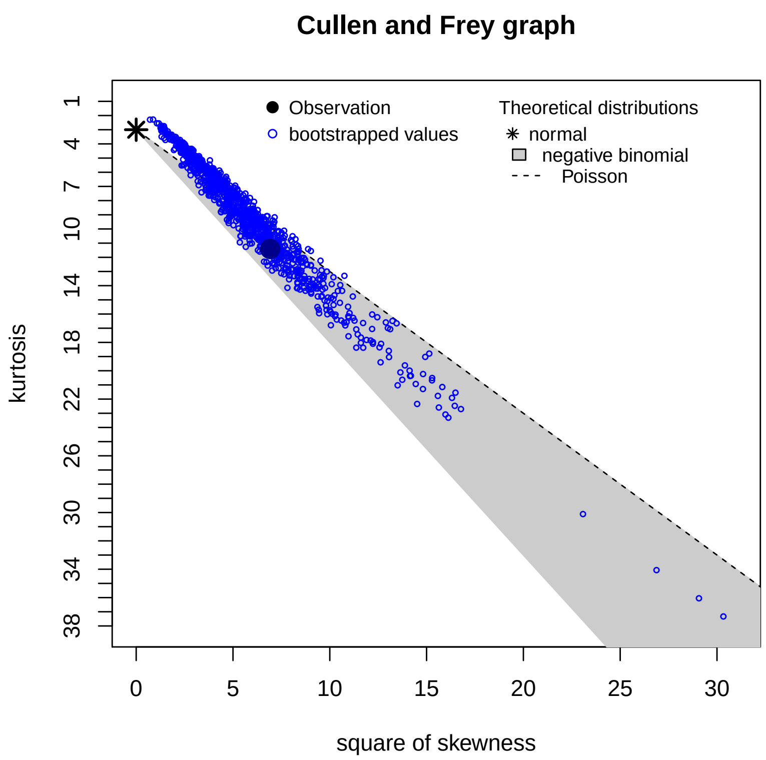

Figure 4: Skewness-kurtosis plot for a discrete variable (number of *Toxocara cati* parasites from the “Toxocara” data set)

3 Fit of a distribution by maximum likelihood estimation

3.1 Parameter estimation

Once selected, one or more parametric distributions may be fitted to the data set, one at a time, using Function `fitdist`. By default, distribution parameters θ are estimated by maximizing the likelihood defined as:

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) \tag{1}$$

with y_i the n observations of variable y and f the density function of the fitted parametric distribution.

The other proposed estimation methods are described in Section 5.

Function `fitdist` returns the results of the fit of any parametric distribution to a data set as an S3 class object that may be easily printed, summarized or plotted (see Figure 5 in Section 3.2). The parametric distribution must be a classically defined R distributions, with at least d, p and q functions respectively for the density cdf and quantile functions (for example `dnorm`, `pnorm` and `qnorm` for the normal distribution). The name of the fitted distribution is specified in the first argument by its classical abbreviation used as the second part of d, p and q functions (for example "norm" for the normal distribution. Numerical results returned by Function `fitdist` are parameter estimates with estimated standard errors computed from the estimate of the Hessian matrix at the maximum likelihood solution, correlation matrix between parameter estimates, the loglikelihood, the Akaike and the Schwarz information criteria (so called AIC and BIC). Below is a call to function `fitdist` to fit a Weibull distribution to the serving size in the "ground beef" data set.

```
> fw <- fitdist(groundbeef$serving,"weibull")
> print(fw)

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
      estimate Std. Error
shape     2.19      0.105
scale    83.35      2.527

> summary(fw)

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
      estimate Std. Error
shape     2.19      0.105
scale    83.35      2.527
Loglikelihood:  -1255   AIC:  2514   BIC:  2522
Correlation matrix:
      shape scale
shape 1.000 0.322
scale 0.322 1.000
```

The same procedure is required to fit a discrete distribution. As an example, using "toxocara" data set, Poisson and negative distributions may be easily fitted and AIC values compared, in this case giving the preference to the negative binomial distribution, with a much smaller AIC value.

```
> (fp <- fitdist(toxocara$number,"pois"))

Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
       estimate Std. Error
lambda     8.68      0.405

> (fnb <- fitdist(toxocara$number,"nbinom"))

Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters:
     estimate Std. Error
size    0.397     0.0829
mu      8.680     1.9350

> fp$aic

[1] 1017
```

```
> fnb$aic
```

```
[1] 323
```

For some distributions (see the help of `fitdist` for details), it is necessary to specify initial values for the distribution parameters in the argument `start` when using the maximum likelihood method. `start` must be a named list of parameters initial values. The names of the parameters in `start` must correspond exactly to their definition in R or in a user-supplied R code. Function `plotdist` (see Section 3.2), which can plot any parametric distribution with specified parameter values in argument `para` may help to find correct initial values for the distribution parameters in non trivial cases, by iterative calls if necessary (see [6] for examples).

3.2 Goodness-of-fit plots

The plot of an object of class `fitdist` provides two types of results depending of the nature of the distribution, continuous or discrete. For continuous distributions, four goodness-of-fit plots are provided : a draw of pdf curve and histogram together, an cdf plot of both empirical and theoretical distributions, a Q-Q plot (plot of the quantiles of the theoretical fitted distribution (x-axis) against the empirical quantiles of the data) and a P-P plot (i.e. for each value of the data set, plot of the cumulative density function of the fitted distribution (x-axis) against the empirical cumulative density function (y-axis)) are also given ([3]). The Q-Q plot emphasizes the lack-of-fit at the distribution tails while the P-P plot emphasizes the lack-of-fit at the distribution center. As an example, let us look at the plot of the previous fit of a Weibull distribution to the"groundbeef" data set (Figure 5). The fit is not perfect, especially in the center of the distribution, but seems correct while looking at the tails.

```
> plot(fw)
```

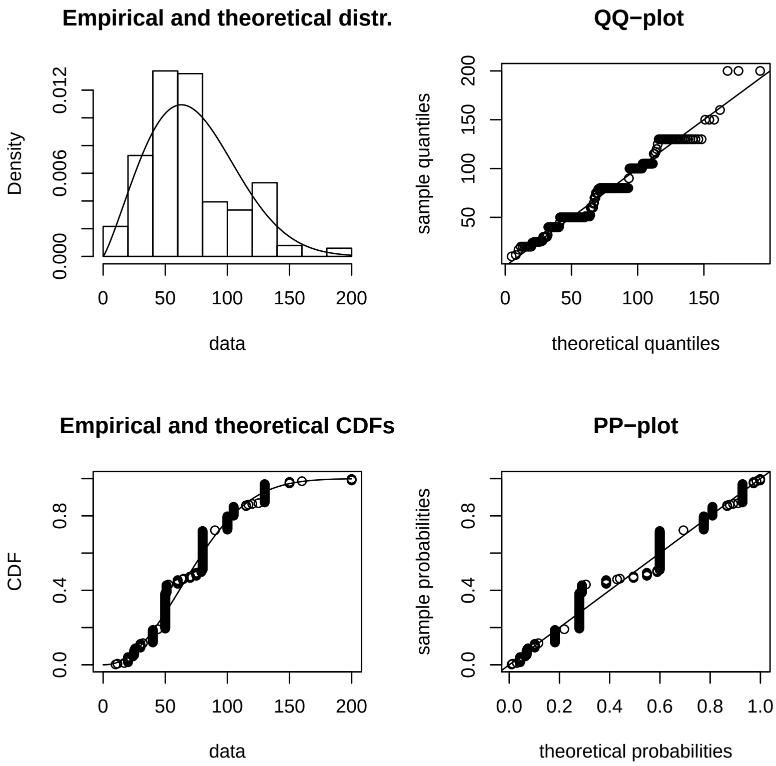

Figure 5: Plot of the fit of a continuous distribution (a Weibull distribution fitted to serving sizes from the "ground beef" data set)

For continuous distributions, Function `cdfcomp` enables the visual comparison of empirical and theoretical cumulative distributions for various distributions fitted on a same data set. Function `cdfcomp` must be called with a first argument corresponding to a list of objects of class `fitdist`, and optionaly further arguments to customize the plot, as in the following example comparing the fit of Weibull, lognormal and gamma distributions to "groundbeef" data set (Figure 6).

```
>      fg <- fitdist(groundbeef$serving,"gamma")
>      fln <- fitdist(groundbeef$serving,"lnorm")
>      cdfcomp(list(fw,fln,fg),legendtext=c("Weibull","lognormal","gamma"),
+      xlab="serving sizes (g)",lwd=2)
```

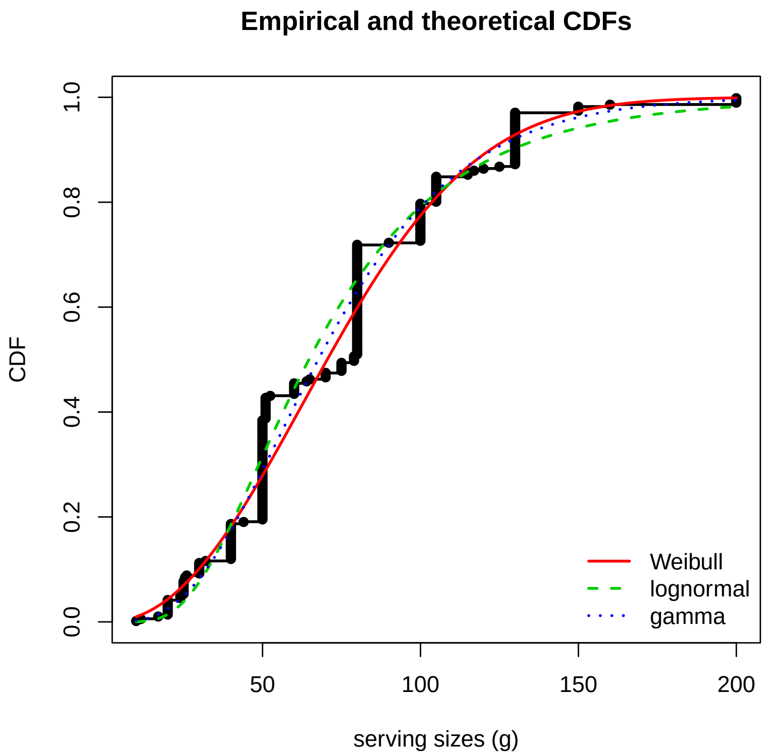

Figure 6: Comparison of CDF plots of various distributions fitted on continuous data (Weibull, gamma and lognormal distributions fitted to serving sizes from the “ground beef” data set)

In such a plot, data may be represented in a log scale when required, by just fixing the argument `xlogscale` to `TRUE` in the call to `cdfcomp`.

For discrete distributions, the plot of an object of class `fitdist` provides two goodness-of-fit plots comparing empirical and theoretical distributions in pdf and in cdf. As an exemple, let us look at the plot of the previous fit of a negative binomial distribution to “toxocara” data set.

```
> plot(fnb)
```

3.3 Measures of goodness-of-fit

When fitting continuous distributions, Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling statistics may be computed using the function `gofstat` as defined by Stephens ([4]).

```
> gofstat(fw)
```

```
Kolmogorov-Smirnov statistic:  0.14
Cramer-von Mises statistic:  0.684
Anderson-Darling statistic:  3.57
```

As giving more weight to distribution tails, Anderson-Darling statistics is of special interest where it is important to place equal emphasis on fitting a distribution at the tails as well as the main body, as it is often the case in risk assessment [3, 16]. Nevertheless, this statistics should be used cautiously when comparing fits of various distributions, keeping in mind that the weighting of each cdf quadratic difference is dependent of the theoretical distribution.

When fitting discrete distributions, the Chi-squared statistic is computed by Function `gofstat` using cells defined by the argument `chisqbreaks` or cells automatically defined from the data in order to reach roughly the same number of observations per cell, roughly equal to the argument `meancount`, or sligthly more if there are some ties. The choice to define cells from the empirical distribution (data) and not from the theoretical distribution was done to enable the comparison of Chi-squared values obtained with different distributions fitted on a same dataset. If arguments `chisqbreaks` and `meancount` are both omitted, `meancount` is fixed in order to obtain roughly $(4n)^{2/5}$ cells, with n the length of the dataset [16]. Using this default option with the fit of a negative binomial distribution to “toxocara” data set gives following results :

```
> gofstat(fnb)
```

```
Chi-squared statistic:  7.49
```

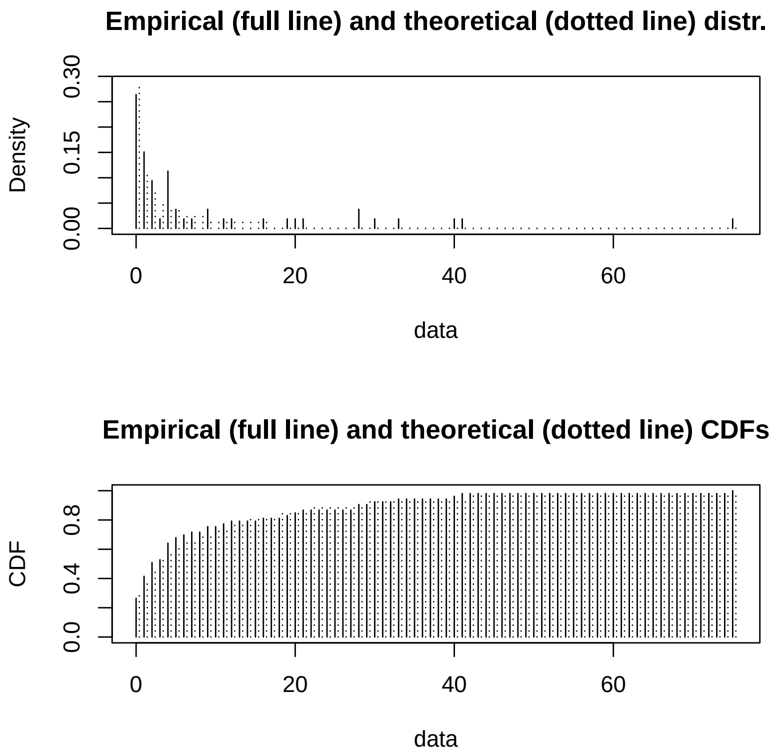

Figure 7: Plot of the fit of a discrete distribution (a negative binomial distribution fitted to numbers of *Toxocara cati* parasites from the "Toxocara" data set)

Among its returned values, Function `gofstat` provides a table with observed and theoretical counts used for the Chi-squared calculations:

```
> gofstat(fnb)$chisqtable

Chi-squared statistic:  7.49
      obscounts theocounts
<= 0      14.00      15.30
<= 1       8.00       5.81
<= 3       6.00       6.85
<= 4       6.00       2.41
<= 9       6.00       7.84
<= 21      6.00       8.27
> 21       7.00       6.54
```

Even if specifically recommended for discrete distributions, the Chi-squared statistic may also be used for continuous distributions (see [6] for examples).

3.4 Goodness-of-fit tests

For continuous distributions, an approximate Kolmogorov-Smirnov test is performed by assuming the distribution parameters known. The critical value defined by Stephens [4] for a completely specified distribution is used to reject or not the distribution at the significance level 0.05. Because of this approximation, the result of the test (decision of rejection of the distribution or not) is returned only for datasets with more than 30 observations. Note that this approximate test may be too conservative.

For datasets with more than 5 observations and for continuous distributions for which the test is described by Stephens [4] (normal, lognormal, exponential, Cauchy, gamma, logistic and Weibull), the Cramer-von Mises and Anderson-darling tests are performed as described by Stephens [4]. Those tests take into account the fact that the parameters are not known but estimated from the data. The result is the decision to reject or not the distribution at the significance level 0.05. Both tests are available only for maximum likelihood estimations.

When the Chi-squared statistic is computed (for discrete or optionnaly continuous distributions), and if the degree of freedom (nb of cells - nb of parameters - 1) of the corresponding distribution is strictly positive, the p-value of the Chi-squared test is returned.

Goodness-of-fit tests may be used carefully. As for any null-hypothesis significance test, the non reject of the null hypothesis dose not imply its acceptation. However, this misinterpretation of p-values is very common and comes from the wrong assumption that absence of evidence is evidence of absence [1]. On the contrary, in some cases, especially on very big datasets, even if the null hypothesis is rejected, a fitted distribution may be chosen as the best one among simple distributions to describe an empirical distribution, if the goodness-of-fit plots do not show strong differences between empirical and theoretical distributions.

Now let us look at the Chi-squared test results for the fit of a negative binomial distribution to "toxocara" data set :

```
> gofstat(fnb,print.test = TRUE)

Chi-squared statistic:  7.49
Degree of freedom of the Chi-squared distribution:  4
Chi-squared p-value:  0.112
   the p-value may be wrong with some theoretical counts < 5
```

A warning message appears as one of the theoretical counts is under 5 using the default breaks (see Section 3.3). In order to solve this problem, one may specify breaks more adapted for the realization of the test.

```
> gofstat(fnb,chisqbreaks=c(0,1,4,8,20),print.test=TRUE)$chisqtable

Chi-squared statistic:  3.42
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.332
      obscounts theocounts
<= 0      14.00      15.30
<= 1       8.00       5.81
<= 4      12.00       9.25
<= 8       4.00       6.63
<= 20      7.00       9.05
> 20       8.00       6.96
```

From goodness-of-fit graphs, Chi-squared statistics, AIC and BIC values, it seems better to choose the fit of a negative binomial distribution for this dataset even it has one more parameter than the Poisson one. This was not obvious while looking at the skewness-kurtosis graph. This graph must be used cautiously especially for continuous distributions far from the normal distribution or for discrete distributions. It is only indicative.

4 The special case of censored data

Censored data may contain left censored, right censored and interval censored values, with several lower and upper bounds. Data must be coded into a dataframe with two columns, respectively named `left` and `right`, describing each observed value as an interval. The `left` column contains either `NA` for left censored observations, the left bound of the interval for interval censored observations, or the observed value for non-censored observations. The `right` column contains either `NA` for right censored observations, the right bound of the interval for interval censored observations, or the observed value for non-censored observations.

4.1 Graphical display of the observed distribution

Using censored data such as those coded in the "smokedfish" data set, the empirical distribution may be plotted using the function `plotdistcens`. By default this function uses the EM approach of Turnbull [14] to compute the overall empirical cdf curve with confidence intervals, by calls to functions `survfit` and `plot.survfit` from the `survival` package. Let us see such a plot for "smokedfish" data set after classical transformation of microbial counts in decimal logarithm (Figure 8).

```
> log10C <- data.frame(left=log10(smokedfish$left),right=log10(smokedfish$right))
> plotdistcens(log10C)
```

4.2 Maximum likelihood estimation

As for non censored data, one or more parametric distributions may then be fitted to the censored data set, one at a time, but using in this case Function `fitdistcens`. This function estimates the distribution parameters by maximizing the likelihood given by following equation for censored data. As `fitdist`, it returns the results of the fit of any parametric distribution to a data set as an S3 class object that may be easily printed, summarized or plotted.

For "smokedfish" data set, a normal distribution may be fitted to log transformed data as commonly done for microbial count data.

```
> flog10C <- fitdistcens(log10C,"norm")
> print(flog10C)
```

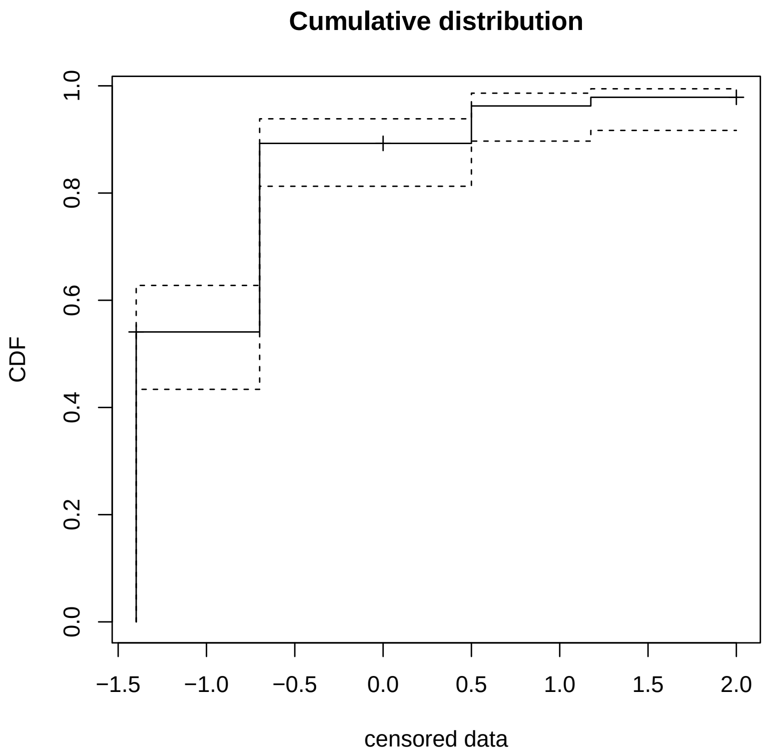

Figure 8: CDF plot of censored data (microbial counts from the "smoked fish" data set)

```
Fitting of the distribution ' norm ' on censored data by maximum likelihood
Parameters:
     estimate
mean    -1.58
sd       1.54
```

```
> summary(flog10C)
```

```
FITTING OF THE DISTRIBUTION ' norm ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
     estimate Std. Error
mean    -1.58      0.201
sd       1.54      0.212
Loglikelihood:  -87.1   AIC:  178   BIC:  183
Correlation matrix:
       mean     sd
mean  1.000 -0.433
sd   -0.433  1.000
```

As with `fitdist`, for some distributions (see [6] for details), it is necessary to specify initial values for the distribution parameters in the argument `start`. The function `plotdistcens` may help to find correct initial values for the distribution parameters in non trivial cases, by an manual iterative use if necessary.

4.3 Goodness-of-fit plot

Only one goodness-of-fit plot is provided for censored data (Figure 9), corresponding to the theoretical cumulative distribution function added to the plot of censored data presented in Section 4.1.

```
> plot(flog10C)
```

Computations of goodness of fit statistics have not yet been developed for fits using censored data, so the quality of fit may only be estimated from the loglikelihood and the goodness-of-fit CDF plot.

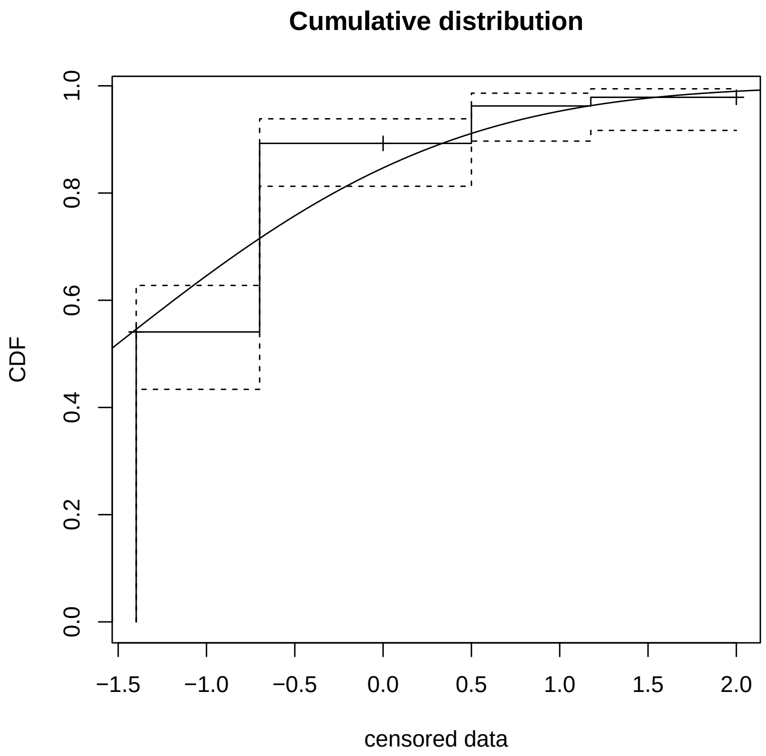

Figure 9: Goodness-of-fit CDF plot for a fit of a continuous distribution on censored data (a lognormal distribution fitted to microbial counts from the "smoked fish" data set)

5 Alternative methods for parameter estimation

5.1 Maximum goodness-of-fit estimation

Maximum likelihood is only the default estimation method proposed by Function `fitdist`, but other methods may be used to estimate parameters for non-censored data. One of the alternative for continuous distributions is the maximum goodness-of-fit estimation method also called minimum distance estimation method. In this package this method is proposed with eight different distances the three classical distances defined in [4], (Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling which gives more weight to the tails of the distribution), or one of the variants of this last distance proposed by [9]. The right-tail AD gives more weight only to the right tail, the left-tail AD gives more weight only to the left tail. Either of the tails, or both of them, can receive even larger weights by using second order Anderson-Darling Statistics.

To fit a distribution by maximum goodness-of-fit estimation, one needs to fix the argument `method` to "mge" in the call to `fitdist` and to specify the argument `gof` coding for the chosen goodness-of-fit distance. This function is intended to be used only with continuous variables and distributions. It may be useful to fit distributions for which maximum likelihood does not provide good estimations, such as the uniform distribution ([9]).

```
> u <- runif(50)
> fitdist(u,"unif",method="mge",gof="KS")
```

```
Fitting of the distribution ' unif ' by maximum goodness-of-fit
Parameters:
    estimate
min    0.109
max    1.097
```

Maximum goddness-of-fit estimation may also be useful to give more weight to data at one tail of the distribution. In ecotoxicology, species sensitivity distributions such as those presented in [8] are often fitted by a lognormal distribution (or another parametric distribution) so as to estimate a low percentile, often 5% percentile, named the hazardous concentration 5% (HC5). This value is then interpreted as a value of the contaminant concentration protecting 95% of the species. In this context, one may consider to fit the parametric distribution by giving more weight to the left tail of the empirical distribution such as in the following example using left tail Anderson-Darling distances of first or second order (Figure 10).

```
> data(endosulfan)
> ATV <-subset(endosulfan,group == "NonArthroInvert")$ATV
```

```
> flnMGEKS <- fitdist(ATV,"lnorm",method="mge",gof="KS")
> flnMGEAD <- fitdist(ATV,"lnorm",method="mge",gof="AD")
> flnMGEADL <- fitdist(ATV,"lnorm",method="mge",gof="ADL")
> flnMGEAD2L <- fitdist(ATV,"lnorm",method="mge",gof="AD2L")
> cdfcomp(list(flnMGEKS,flnMGEAD,flnMGEADL,flnMGEAD2L),
+ xlogscale = TRUE,main="",
+ legendtext = c("Kolmogorov-Smirnov (KS)","Anderson-Darling",
+ "Left-tail Anderson-Darling","Left tailed Anderson-Darling of second order"),cex=0.7,
+ xlegend = 500, ylegend = 0.15)
```

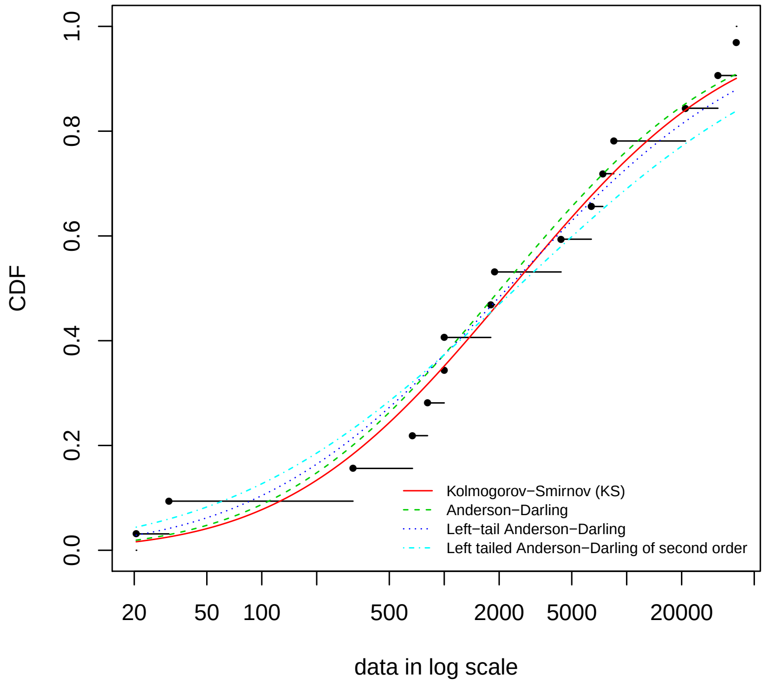

Figure 10: Comparison of one distribution fitted by maximum goodness-of-fit using various goodness-of-fit distances (a lognormal distribution fitted to acute toxicity values from the "endosulfan" data set)

5.2 Moment matching estimation

Moment matching estimation may also be performed fixing the argument `method` to "mme" in the call to fitdist.

```
> fitdist(u,"unif",method="mme")

Fitting of the distribution ' unif ' by matching moments
Parameters:
  estimate
1   0.0953
2   1.0780
```

The estimate is computed by a closed formula for following distributions: normal, lognormal, exponential, Poisson, gamma, logistic, negative binomial, geometric, beta and uniform distributions. For distributions characterized by one parameter (geometric, Poisson and exponential), this parameter is simply estimated by matching theoretical and observed means, and for distributions characterized by two parameters, these parameters are estimated by matching theoretical and observed means and variances ([16]).

For other distributions, Function `fitdist` carries out the matching numerically, by minimization of the sum of squared differences between observed and theoretical moments (see [6] for technical details).

5.3 Quantile matching estimation

Quantile matching may also be performed fixing the argument `method` to "qme" in the call to fitdist and adding an argument `probs` defining the probabilities for which the quantile matching is performed. The length of this vector must be equal to the number of parameters to estimate. The quantile matching is carried out numerically, by minimizing the sum of squared differences between observed and theoretical quantiles. Here is an example of fit of a uniform distribution by matching first and third quartiles.

```
> fitdist(u,"unif",method="qme",probs=c(0.25,0.75))

Fitting of the distribution ' unif ' by matching quantiles
Parameters:
    estimate
min    0.173
max    1.012
```

5.4 Customization of the optimization algorithm

Each time a numerical minimization (or maximization) is carried out using `fitdist`, Function `optim` of the package `stats` is used by default with the "Nelder-Mead" method for distributions characterized by more than one parameter and the "BFGS" method for distributions characterized by only one parameter

Sometimes the default algorithm fails to converge. It may then be interesting to change some options of the function `optim` or to use another optimization function than `optim` to maximize the likelihood.

The argument `optim.method` may be used in the call to `fitdist` or `fitdistcens`. It will internally be passed to `mledist` and to `optim`. This argument may be fixed to "Nelder-Mead" (the robust Nelder and Mead method), "BFGS" (the BFGS quasi-Newton method), "CG" (a conjugate gradients method), "SANN" (a variant of simulated annealing) or "L-BFGS-B" (a modification of the BFGS quasi-Newton method which enables box constraints optimization). For the use of the last method the arguments `lower` and/or `upper` also have to be passed. More details on these optimization functions may be found in the help page of `optim` from the package `stats`.

Here are examples of fits of a gamma distribution to "ground beef" data set with various options of `optim`.

```
> fitdist(groundbeef$serving,"gamma",optim.method="Nelder-Mead")

Fitting of the distribution ' gamma ' by maximum likelihood
Parameters:
      estimate Std. Error
shape   4.0083    0.34134
rate    0.0544    0.00494

> fitdist(groundbeef$serving,"gamma",optim.method="BFGS")

Fitting of the distribution ' gamma ' by maximum likelihood
Parameters:
      estimate Std. Error
shape   4.2285     0.3608
rate    0.0574     0.0052

> fitdist(groundbeef$serving,"gamma",optim.method="SANN")

Fitting of the distribution ' gamma ' by maximum likelihood
Parameters:
      estimate Std. Error
shape    3.973     0.3382
rate     0.054     0.0049
```

You may also want to use another function than `optim` to maximize the likelihood. This optimization function has to be specified by the argument `custom.optim` in the call to `fitdist` or `fitdistcens`. But before that, it is necessary to customize this optimization function : `custom.optim` function must have (at least) the following arguments, `fn` for the function to be optimized, `par` for the initialized parameters. It is assumed that `custom.optim` should carry out a MINIMIZATION. Finally, it should return at least the following components: `par` for the estimate, `convergence` for the convergence code, `value` for `fn(par)` and `hessian`.

Below is an example of code written to customize `genoud` function from `rgenoud` package.

```
mygenoud <- function(fn, par, ...)
{
   require(rgenoud)
   res <- genoud(fn, starting.values=par, ...)
   standardres <- c(res, convergence=0)
   return(standardres)
}
```

The customized optimization function may then be passed as the argument `custom.optim` in the call to `fitdist` or `fitdistcens`. The following code may for example be used to fit a gamma distribution to the "ground beef" data set. Note that in this example various arguments are also passed from `fitdist` to `genoud` : `nvars`, `Domains`, `boundary.enforcement`, `print.level` and `hessian`.

```
fitdist(groundbeef$serving, "gamma", custom.optim=mygenoud, nvars=2,
   Domains=cbind(c(0,0), c(10, 10)), boundary.enforcement=1,
   print.level=1, hessian=TRUE)
```

6 Uncertainty in parameter estimates

6.1 Bootstrap procedures

The uncertainty in the parameters of the fitted distribution may be simulated by parametric or nonparametric boostrap using the function `boodist`. This function returns the boostrapped values of parameters in a S3 class object which may be plotted to visualize the bootstrap region. The medians and the 95 percent confidence intervals of parameters (2.5 and 97.5 percentiles) are printed in the summary. If inferior to the whole number of iterations, the number of iterations for which the function converges is also printed in the summary.

The plot of an object of class `bootdist` consists in a scatterplot or a matrix of scatterplots of the bootstrapped values of parameters providing a representation of the joint uncertainty distribution of the fitted parameters (Figure 11).

Below is an example of the use of this function with the previous of the Weibull distribution to "groundbeef" data set.

```
> bw<-bootdist(fw,niter=1001)
> plot(bw)
> summary(bw)

Parametric bootstrap medians and 95% percentile CI
      Median 2.5% 97.5%
shape    2.2  2.0  2.43
scale   83.4 78.7 88.45
```

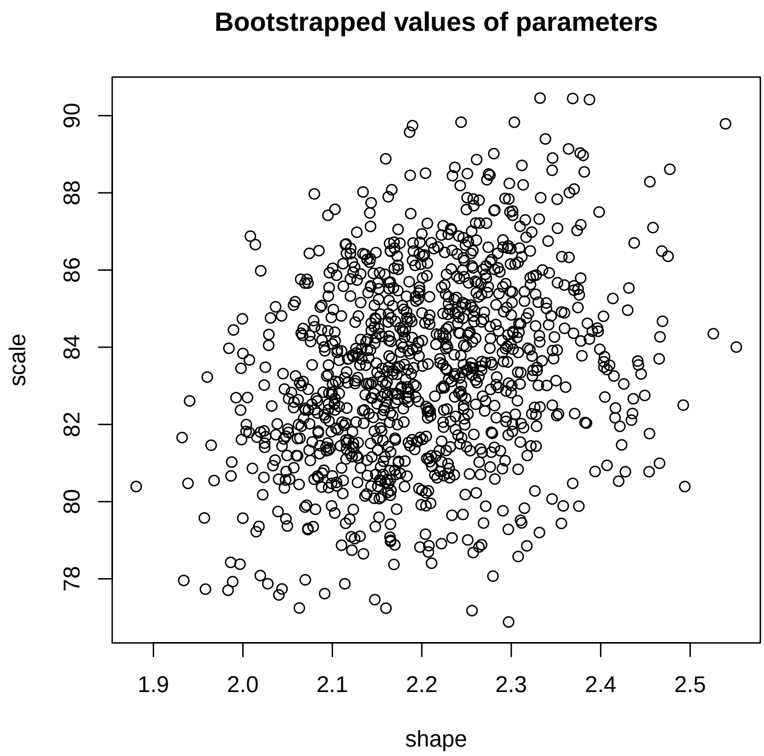

Figure 11: Bootstrappped values of parameters for a fit of a distribution characterized by two parameters (example on the fit of a Weibull distribution to serving sizes from the "ground beef" data set)

Function `boodistcens` provides the same type of results for fit on continuous distributions to censored data.

```
> blog10C<-bootdistcens(flog10C,niter=1001)
> summary(blog10C)

Nonparametric bootstrap medians and 95% percentile CI
     Median  2.5% 97.5%
mean  -1.56 -2.03 -1.23
sd     1.51  1.02  2.04
```

6.2 Use of bootstrap samples

Bootstrap samples of parameter estimates may be used to calculate confidence intervals on each parameter of the fitted distribution, but it is also interesting to look at the marginal distribution of the bootstrap values in a scatterplot (or a matrix of scatterplots if the number of parameters exceeds two), and especially to look at the potential structural correlation between parameters.

Moreover, the use of the whole bootstrap sample is of great interest in the risk assessment field. Its use enables the characterization of uncertainty in distribution parameters. It can be directly used within a second order Monte Carlo simulation framework, especially within the package `mc2d` ([11]). One could refer to Pouillot *et al.* ([10]) for an introduction to the use of `mc2d` and `fitdistrplus` packages in the context of quantitative risk assessment.

References

[1] DG Altman and JM Bland. Absence of Evidence is not Evidence of Absence. *AUSTRALIAN VETERINARY JOURNAL*, 74(4):311, OCT 1996.

[2] P. Busschaert, A. H. Geeraerd, M. Uyttendaele, and J. F. Van Impe. Estimating Distributions out of Qualitative and (Semi)Quantitative Microbiological Contamination Data for Use in Risk Assessment. *INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY*, 138(3):260–269, APR 15 2010.

[3] A.C. Cullen and H.C. Frey. *Probabilistic Techniques in Exposure Assessment.* Plenum Publishing Co., New York, first edition, 1999.

[4] R.B. D'Agostino and M.A. Stephens. *Goodness-of-Fit Techniques.* Dekker, New York, first edition, 1986.

[5] M. L. Delignette-Muller, M. Cornu, and AFSSA Stec Study Grp. Quantitative Risk Assessment for Escherichia coli O157:H7 in Frozen Ground Beef Patties Consumed by Young Children in French Households. *INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY*, 128(1, SI):158–164, NOV 30 2008. 5th International Conference on Predictive Modelling in Foods, Natl Tech Univ Athens, Athens, GREECE, SEP 16-19, 2007.

[6] M.L. Delignette-Muller, R. Pouillot, J.B. Denis, and C. Dutang. *fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*, 2011. R package version 0.3-4.

[7] E Fromont, L Morvilliers, M Artois, and D Pontier. Parasite Richness and Abundance in Insular and Mainland Feral Cats: Insularity or Density? *PARASITOLOGY*, 123(Part 2):143–151, AUG 2001.

[8] GC Hose and PJ Van den Brink. Confirming the Species-Sensitivity Distribution Concept for Endosulfan Using Laboratory, Mesocosm, and Field Data. *ARCHIVES OF ENVIRONMENTAL CONTAMINATION AND TOXICOLOGY*, 47(4):511–520, OCT 2004.

[9] A. Luceno. Fitting the Generalized Pareto Distribution to Data using Maximum Goodness-of-fit Estimators. *COMPUTATIONAL STATISTICS & DATA ANALYSIS*, 51(2):904–917, NOV 15 2006.

[10] R. Pouillot and M.L. Delignette-Muller. Evaluating Variability and Uncertainty Separately in Microbial Quantitative Risk Assessment using two R Packages. *INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY*, 142(3):330–340, SEP 1 2010.

[11] R. Pouillot, M.L. Delignette-Muller, and J.B. Denis. *mc2d: Tools for Two-Dimensional Monte-Carlo Simulations*, 2011. R package version 0.1-12.

[12] Ricci, V. Fitting distributions with r. Contributed Documentation available on CRAN, 2005.

[13] T. Therneau. *survival: Survival Analysis, Including Penalized Likelihood*, 2011. R package version 2.36-9.

[14] BW Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 69(345):169–173, 1974.

[15] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, 4 edition, 2010.

[16] D. Vose. *Quantitative Risk Analysis. A Guide to Monte Carlo Simulation Modelling.* Wiley, New York, first edition, 2010.