Fitting distributions by MLE to non-censored data

Marie Laure Delignette Muller, Christophe Dutang

2016-04-15

Contents

1 Choice of candidate distributions

For illustrating the use of various functions of the **fitdistrplus** package with continuous non-censored data, we will first use a data set named `groundbeef` which is included in our package. This data set contains pointwise values of serving sizes in grams, collected in a French survey, for ground beef patties consumed by children under 5 years old. It was used in a quantitative risk assessment published by (M. L. Delignette-Muller, Cornu, and AFSSA-STEC-Study-Group {2008}).

```
set.seed(1234)
library("fitdistrplus")
data("groundbeef")
str(groundbeef)
```

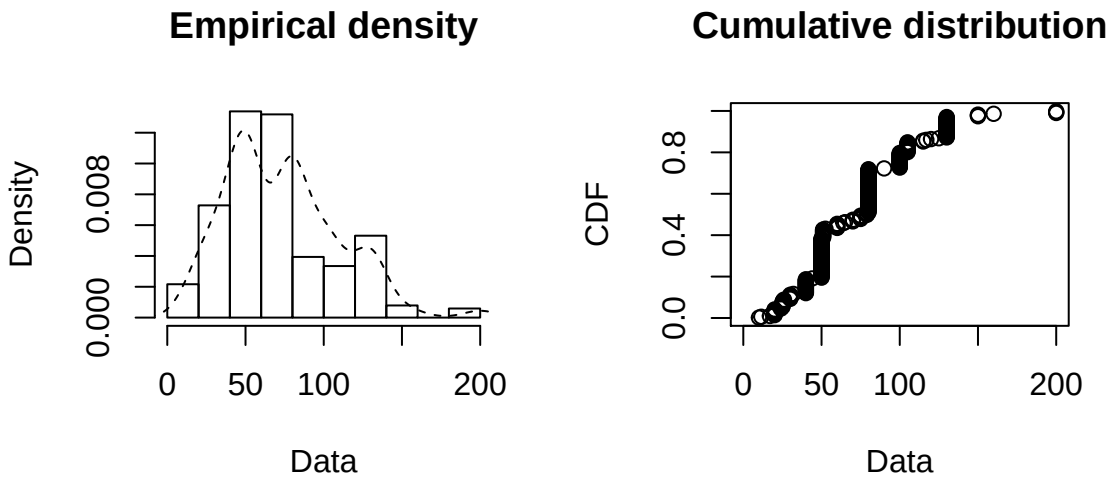
```
## 'data.frame':    254 obs. of  1 variable:
##  $ serving: num  30 10 20 24 20 24 40 20 50 30 ...
```

Before fitting one or more distributions to a data set, it is generally necessary to choose good candidates among a predefined set of distributions. This choice may be guided by the knowledge of stochastic processes governing the modelled variable, or, in the absence of knowledge regarding the underlying process, by the observation of its empirical distribution. To help the user in this choice, we developed functions to plot and characterize the empirical distribution.

First of all, it is common to start with plots of the empirical distribution function and the histogram (or density plot), which can be obtained with the `plotdist` function of the **fitdistrplus** package. This function

provides two plots (see Figure~**??**): the left-hand plot is by default the histogram on a density scale (or density plot of both, according to values of arguments `histo` and `demp`) and the right-hand plot the empirical cumulative distribution function (CDF).

```
plotdist(groundbeef$serving, histo = TRUE, demp = TRUE)
```



In addition to empirical plots, descriptive statistics may help to choose candidates to describe a distribution among a set of parametric distributions. Especially the skewness and kurtosis, linked to the third and fourth moments, are useful for this purpose. A non-zero skewness reveals a lack of symmetry of the empirical distribution, while the kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3. The skewness and kurtosis and their corresponding unbiased estimator (Casella and Berger 2002) from a sample $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$ with observations $(x_i)_i$ are given by

$$sk(X) = \frac{E[(X - E(X))^3]}{Var(X)^{\frac{3}{2}}} \ , \ \widehat{sk} = \frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{m_2^{\frac{3}{2}}}, \tag{1}$$

$$kr(X) = \frac{E[(X - E(X))^4]}{Var(X)^2} \ , \ \widehat{kr} = \frac{n-1}{(n-2)(n-3)}((n+1) \times \frac{m_4}{m_2^2} - 3(n-1)) + 3, \tag{2}$$

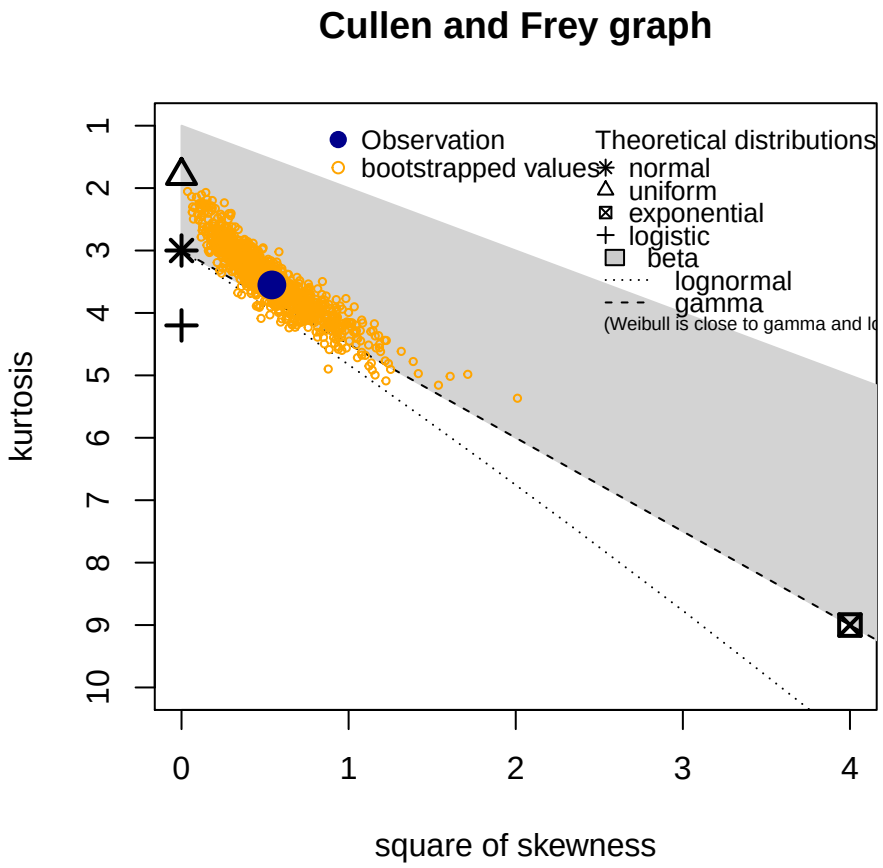
where m_2, m_3, m_4 denote empirical moments defined by $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$, with x_i the n observations of variable x and $\overline{x}$ their mean value.

The `descdist` function provides classical descriptive statistics (minimum, maximum, median, mean, standard deviation), skewness and kurtosis. By default, unbiased estimations of the three last statistics are provided. Nevertheless, the argument `method` can be changed from `"unbiased"` (default) to `"sample"` to obtain them without correction for bias. A skewness-kurtosis plot such as the one proposed by (Cullen and Frey 1999) is provided by the `descdist` function for the empirical distribution (see Figure~**??** for the `groundbeef` data set). On this plot, values for common distributions are displayed in order to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential), there is only one possible value for the skewness and the kurtosis. Thus, the distribution is represented by a single 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, a nonparametric bootstrap procedure (Efron and Tibshirani 1994) can be performed by using the argument `boot`. %to an integer above 10. Values of skewness and kurtosis are computed on bootstrap samples (constructed by random sampling with replacement from the original data set) and reported on the skewness-kurtosis plot. Nevertheless, the user needs to know that

skewness and kurtosis, like all higher moments, have a very high variance. This is a problem which cannot be completely solved by the use of bootstrap. The skewness-kurtosis plot should then be regarded as indicative only. The properties of the random variable should be considered, notably its expected value and its range, as a complement to the use of the `plotdist` and `descdist` functions. Below is a call to the `descdist` function to describe the distribution of the serving size from the `groundbeef` data set and to draw the corresponding skewness-kurtosis plot (see Figure~??). 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:  73.64567
## estimated sd:  35.88487
## estimated skewness:  0.7352745
## estimated kurtosis:  3.551384
```

2 Fit of distributions by MLE

Once selected, one or more parametric distributions $f(.|\theta)$ (with parameter $\theta \in \mathbb{R}^d$) may be fitted to the data set, one at a time, using the `fitdist` function. Under the i.i.d. sample assumption, distribution parameters θ are by default estimated by maximizing the likelihood function defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3}$$

with x_i the n observations of variable X and $f(.|\theta)$ the density function of the parametric distribution. The other proposed estimation methods are described in Section~**??**.

The fit of a distribution using `fitdist` assumes that the corresponding `d`, `p`, `q` functions (standing respectively for the density, the distribution and the quantile functions) are defined. Classical distributions are already defined in that way in the **stats** package, e.g., `dnorm`, `pnorm` and `qnorm` for the normal distribution (see `?Distributions`). Others may be found in various packages (see the CRAN task view: Probability Distributions at http://cran.r-project.org/web/views/Distributions.html). Distributions not found in any package must be implemented by the user as `d`, `p`, `q` functions. In the call to `fitdist`, a distribution has to be specified via the argument `dist` either by the character string corresponding to its common root name used in the names of `d`, `p`, `q` functions (e.g., `"norm"` for the normal distribution) or by the density function itself, from which the root name is extracted (e.g., `dnorm` for the normal distribution). Numerical results returned by the `fitdist` function are (1) the parameter estimates, (2) the estimated standard errors (computed from the estimate of the Hessian matrix at the maximum likelihood solution), (3) the loglikelihood, (4) Akaike and Bayesian information criteria (the so-called AIC and BIC), and (5) the correlation matrix between parameter estimates. Below is a call to the `fitdist` function to fit a Weibull distribution to the serving size from the `groundbeef` data set.

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

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters:
##        estimate Std. Error
## shape  2.185885  0.1045755
## scale 83.347679  2.5268626
```

3 Generic functions for a `"fitdist"` object

The `fitdist` function returns an S3 object of class `"fitdist"` for which `print`, `summary` and `plot` functions are provided.

```
print(fw)
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters:
##        estimate Std. Error
## shape  2.185885  0.1045755
## scale 83.347679  2.5268626
```

```
summary(fw)
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape  2.185885  0.1045755
## scale 83.347679  2.5268626
```

```
## Loglikelihood:  -1255.225   AIC:  2514.449   BIC:  2521.524
## Correlation matrix:
##          shape    scale
## shape 1.000000 0.321821
## scale 0.321821 1.000000
```

The plot of an object of class `"fitdist"` provides four classical goodness-of-fit plots (Cullen and Frey 1999) presented on Figure~**??**:

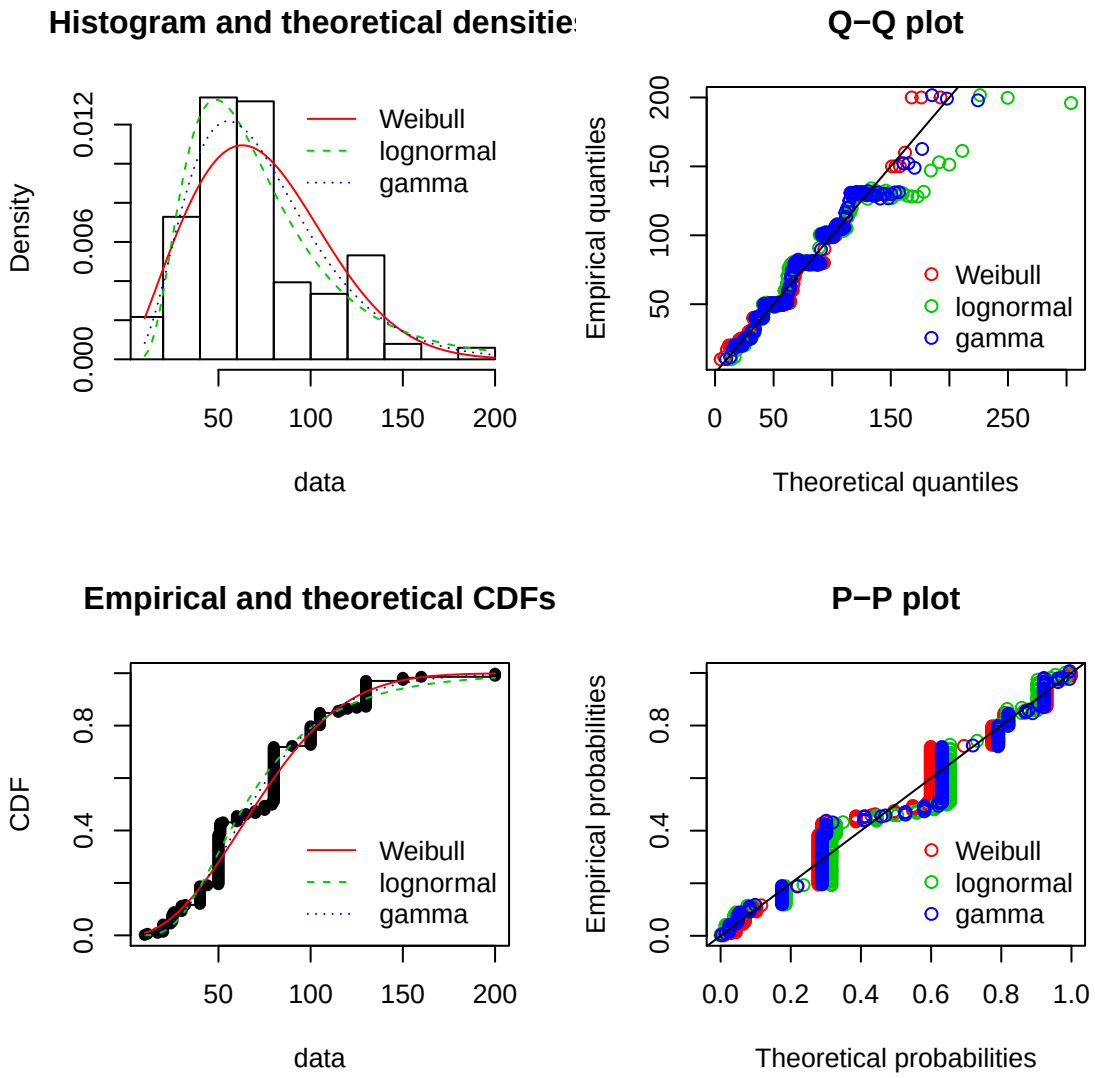
- a density plot representing the density function of the fitted distribution along with the histogram of the empirical distribution,
- a CDF plot of both the empirical distribution and the fitted distribution,
- a Q-Q plot representing the empirical quantiles (y-axis) against the theoretical quantiles (x-axis)
- a P-P plot representing the empirical distribution function evaluated at each data point (y-axis) against the fitted distribution function (x-axis).

For CDF, Q-Q and P-P plots, the probability plotting position is defined by default using Hazen's rule, with probability points of the empirical distribution calculated as `(1:n - 0.5)/n`, as recommended by (Blom 1959). This plotting position can be easily changed (see the reference manual for details (M. Delignette-Muller et al. 2014).

4 Additional graphic functions for a `"fitdist"` object

Unlike the generic `plot` function, the `denscomp`, `cdfcomp`, `qqcomp` and `ppcomp` functions enable to draw separately each of these four plots, in order to compare the empirical distribution and multiple parametric distributions fitted on a same data set. These functions must be called with a first argument corresponding to a list of objects of class `fitdist`, and optionally further arguments to customize the plot (see the reference manual for lists of arguments that may be specific to each plot (M. Delignette-Muller et al. 2014). In the following example, we compare the fit of a Weibull, a lognormal and a gamma distributions to the `groundbeef` data set (Figure~**??**).

```
fg <- fitdist(groundbeef$serving, "gamma")
fln <- fitdist(groundbeef$serving, "lnorm")
par(mfrow = c(2, 2))
plot.legend <- c("Weibull", "lognormal", "gamma")
denscomp(list(fw, fln, fg), legendtext = plot.legend)
qqcomp(list(fw, fln, fg), legendtext = plot.legend)
cdfcomp(list(fw, fln, fg), legendtext = plot.legend)
ppcomp(list(fw, fln, fg), legendtext = plot.legend)
```



The density plot and the CDF plot may be considered as the basic classical goodness-of-fit plots. The two other plots are complementary and can be very informative in some cases. 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. In the present example (in Figure~**??**), none of the three fitted distributions correctly describes the center of the distribution, but the Weibull and gamma distributions could be prefered for their better description of the right tail of the empirical distribution, especially if this tail is important in the use of the fitted distribution, as it is in the context of food risk assessment.

5 Fitting non-R-base distribution

The data set named `endosulfan` will now be used to illustrate other features of the **fitdistrplus** package. This 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 (Hose and Van den Brink {2004}).

In ecotoxicology, a lognormal or a loglogistic distribution is often fitted to such a data set in order to characterize the species sensitivity distribution (SSD) for a pollutant. A low percentile of the fitted distribution, generally the 5% percentile, is then calculated and named the hazardous concentration 5% (HC5). It is interpreted as the value of the pollutant concentration protecting 95% of the species (Posthuma, Suter, and Traas 2010). But

the fit of a lognormal or a loglogistic distribution to the whole `endosulfan` data set is rather bad (Figure~**??**), especially due to a minority of very high values.

The two-parameter Pareto distribution and the three-parameter Burr distribution (which is an extension of both the loglogistic and the Pareto distributions) have been fitted. Pareto and Burr distributions are provided in the package **actuar**. Until here, we did not have to define starting values (in the optimization process) as reasonable starting values are implicity defined within the `fitdist` function for most of the distributions defined in R (see `?fitdist` for details). For other distributions like the Pareto and the Burr distribution, initial values for the distribution parameters have to be supplied in the argument `start`, as a named list with initial values for each parameter (as they appear in the `d`, `p`, `q` functions). Having defined reasonable starting values[1], various distributions can be fitted and graphically compared. On this example, the function `cdfcomp` can be used to report CDF values in a logscale so as to emphasize discrepancies on the tail of interest while defining an HC5 value (Figure~**??**).

```
data("endosulfan")
ATV <-endosulfan$ATV
fendo.ln <- fitdist(ATV, "lnorm")
library("actuar")
fendo.ll <- fitdist(ATV, "llogis", start = list(shape = 1, scale = 500))
fendo.P <- fitdist(ATV, "pareto", start = list(shape = 1, scale = 500))
fendo.B <- fitdist(ATV, "burr", start = list(shape1 = 0.3, shape2 = 1,
  rate = 1))
```

None of the fitted distribution correctly describes the right tail observed in the data set, but as shown in Figure~**??**, the left-tail seems to be better described by the Burr distribution. Its use could then be considered to estimate the HC5 value as the 5% quantile of the distribution. This can be easily done using the `quantile` generic function defined for an object of class `"fitdist"`. Below is this calculation together with the calculation of the empirical quantile for comparison.

<>=

```
quantile(fendo.B, probs = 0.05)
```

```
## Estimated quantiles for each specified probability (non-censored data)
##             p=0.05
## estimate 0.2939259
```

```
quantile(ATV, probs = 0.05)
```

```
##  5%
## 0.2
```

In addition to the ecotoxicology context, the `quantile` generic function is also attractive in the actuarial–financial context. In fact, the value-at-risk VAR_α is defined as the $1-\alpha$-quantile of the loss distribution and can be computed with `quantile` on a `"fitdist"` object.

6 Goodness-of-fit statistics

The computation of different goodness-of-fit statistics is proposed in the **fitdistrplus** package in order to further compare fitted distributions. The purpose of goodness-of-fit statistics aims to measure the distance

[1] The 'plotdist' function can plot any parametric distribution with specified parameter values in argument 'para'. It can thus help to find correct initial values for the distribution parameters in non trivial cases, by iterative calls if necessary (see the reference manual for examples [@fitdistrplus].

between the fitted parametric distribution and the empirical distribution: e.g., the distance between the fitted cumulative distribution function F and the empirical distribution function F_n. When fitting continuous distributions, three goodness-of-fit statistics are classicaly considered: Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling statistics (D'Agostino and Stephens 1986). Naming x_i the n observations of a continuous variable X arranged in an ascending order, Table 1 gives the definition and the empirical estimate of the three considered goodness-of-fit statistics. They can be computed using the function `gofstat` as defined by Stephens (D'Agostino and Stephens 1986).

```
gofstat(list(fendo.ln, fendo.ll, fendo.P, fendo.B),
  fitnames = c("lnorm", "llogis", "Pareto", "Burr"))
```

```
## Goodness-of-fit statistics
##                                  lnorm    llogis     Pareto       Burr
## Kolmogorov-Smirnov statistic 0.1672498 0.1195888 0.08488002 0.06154925
## Cramer-von Mises statistic   0.6373593 0.3827449 0.13926498 0.06803071
## Anderson-Darling statistic   3.4721179 2.8315975 0.89206283 0.52393018
##
## Goodness-of-fit criteria
##                                    lnorm   llogis   Pareto     Burr
## Aikake's Information Criterion 1068.810 1069.246 1048.112 1045.731
## Bayesian Information Criterion 1074.099 1074.535 1053.400 1053.664
```

Statistic	General formula	Computational formula
Kolmogorov-Smirnov (KS)	$\sup \lvert F_n(x) - F(x) \rvert$	$\max(D^+, D^-)$ with $D^+ = \max_{i=1,\ldots,n} \left(\frac{i}{n} - F_i\right)$ $D^- = \max_{i=1,\ldots,n} \left(F_i - \frac{i-1}{n}\right)$
Cramer-von Mises (CvM)	$\mathrm{n} \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^{n} \left(F_i - \frac{2i-1}{2n}\right)^2$
Anderson-Darling (AD)	$\mathrm{n} \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \log(F_i(1 - F_{n+1-i}))$
where $F_i \overset{\triangle}{=} F(x_i)$		

Table 1: Goodness-of-fit statistics as defined by Stephens [@Stephens86].

As giving more weight to distribution tails, the Anderson-Darling statistic is of special interest when it matters to equally emphasize the tails as well as the main body of a distribution. This is often the case in risk assessment [@]. For this reason, this statistics is often used to select the best distribution among those fitted. Nevertheless, this statistics should be used cautiously when comparing fits of various distributions. Keeping in mind that the weighting of each CDF quadratic difference depends on the parametric distribution in its definition (see Table 1), Anderson-Darling statistics computed for several distributions fitted on a same data set are theoretically difficult to compare. Moreover, such a statistic, as Cramer-von Mises and Kolmogorov-Smirnov ones, does not take into account the complexity of the model (i.e., parameter number). It is not a problem when compared distributions are characterized by the same number of parameters, but it could systematically promote the selection of the more complex distributions in the other case. Looking at classical penalized criteria based on the loglikehood (AIC, BIC) seems thus also interesting, especially to discourage overfitting.

In the previous example, all the goodness-of-fit statistics based on the CDF distance are in favor of the Burr distribution, the only one characterized by three parameters, while AIC and BIC values respectively give the preference to the Burr distribution or the Pareto distribution. The choice between these two distributions

seems thus less obvious and could be discussed. Even if specifically recommended for discrete distributions, the Chi-squared statistic may also be used for continuous distributions (see Section~**??** and the reference manual for examples (M. Delignette-Muller et al. 2014).

7 Uncertainty in parameter estimates

The uncertainty in the parameters of the fitted distribution can be estimated by parametric or nonparametric bootstraps using the `boodist` function for non-censored data (Efron and Tibshirani 1994). This function returns the bootstrapped values of parameters in an S3 class object which can 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. When inferior to the whole number of iterations (due to lack of convergence of the optimization algorithm for some bootstrapped data sets), the number of iterations for which the estimation 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. Below is an example of the use of the `bootdist` function with the previous fit of the Burr distribution to the `endosulfan` data set (Figure~**??**).

```
bendo.B <- bootdist(fendo.ll, niter = 1001)
summary(bendo.B)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%      97.5%
## shape 0.5692948 0.4904687  0.6719985
## scale 8.8239357 4.8097437 15.2146461
```

```
plot(bendo.B, enhance=TRUE)
```

Bootstrap samples of parameter estimates are useful especially to calculate confidence intervals on each parameter of the fitted distribution from the marginal distribution of the bootstraped values. It is also interesting to look at the joint distribution of the bootstraped values in a scatterplot (or a matrix of scatterplots if the number of parameters exceeds two) in order to understand the potential structural correlation between parameters (see Figure~**??**).

The use of the whole bootstrap sample is also of 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** (Pouillot, Delignette-Muller, and Denis 2011). One could refer to (Pouillot and Delignette-Muller {2010}) for an introduction to the use of **mc2d** and **fitdistrplus** packages in the context of quantitative risk assessment.

The bootstrap method can also be used to calculate confidence intervals on quantiles of the fitted distribution. For this purpose, a generic `quantile` function is provided for class `bootdist`. By default, 95% percentiles bootstrap confidence intervals of quantiles are provided. Going back to the previous example from ecotoxicolgy, this function can be used to estimate the uncertainty associated to the HC5 estimation, for example from the previously fitted Burr distribution to the `endosulfan` data set.

```
quantile(bendo.B, probs = 5:10/100)
```

```
## (original) estimated quantiles for each specified probability (non-censored data)
##              p=0.05     p=0.06     p=0.07    p=0.08    p=0.09     p=0.1
## estimate 0.04833391 0.06792591 0.09084979 0.1171938 0.1470655 0.1805897
```

```
## Median of bootstrap estimates
##              p=0.05     p=0.06     p=0.07    p=0.08    p=0.09     p=0.1
## estimate 0.05066381 0.07115019 0.09488366 0.1230986 0.1542089 0.1894622
##
## two-sided 95 % CI of each quantile
##            p=0.05     p=0.06     p=0.07     p=0.08     p=0.09      p=0.1
## 2.5 %  0.01746072 0.02556766 0.03529812 0.04729221 0.06185176 0.07884498
## 97.5 % 0.12787847 0.17502172 0.22424292 0.28658646 0.35270136 0.41622832
```

References

Blom, G. 1959. *Statistical Estimates and Transformed Beta Variables*. 1st ed. John Wiley & Sons.

Casella, G., and R.L. Berger. 2002. *Statistical Inference*. 2nd ed. Duxbury Thomson Learning.

Cullen, A.C., and H.C. Frey. 1999. *Probabilistic Techniques in Exposure Assessment*. 1st ed. Plenum Publishing Co.

D'Agostino, R.B., and M.A. Stephens. 1986. *Goodness-of-Fit Techniques*. 1st ed. Dekker.

Delignette-Muller, M. L., M. Cornu, and AFSSA-STEC-Study-Group. {2008}. "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): 158–64.

Delignette-Muller, M.L., R. Pouillot, J.B. Denis, and C. Dutang. 2014. *Fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*. http://CRAN.R-project.org/package=fitdistrplus.

Efron, B., and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. 1st ed. Chapman & Hall.

Hose, G.C., and P.J. Van den Brink. {2004}. "Confirming the Species-Sensitivity Distribution Concept for Endosulfan Using Laboratory, Mesocosm, and Field Data." *Archives of Environmental Contamination and Toxicology* 47 (4): 511–20.

Posthuma, L., G.W. Suter, and T.P. Traas. 2010. *Species Sensitivity Distributions in Ecotoxicology*. Environmental and Ecological Risk Assessment Series. Taylor & Francis.

Pouillot, R., and M.L. Delignette-Muller. {2010}. "Evaluating Variability and Uncertainty Separately in Microbial Quantitative Risk Assessment using two R Packages." *International Journal of Food Microbiology* 142 (3): 330–40.

Pouillot, R., M.L. Delignette-Muller, and J.B. Denis. 2011. *Mc2d: Tools for Two-Dimensional Monte-Carlo Simulations*. http://CRAN.R-project.org/package=mc2d.