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

Marie Laure Delignette-Muller and Christophe Dutang

February 7, 2013

TODO abstract

Contents

1 Introduction

Fitting distributions to data is a very common task in statistics and consists in choosing a probability distribution that gives a good representation of a statistical variable as well as finding parameter estimates of that distribution. 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** [42] is a well known general-purpose maximum-likelihood fitting routine for the parameter estimation step in R [34]. Other steps of the process may be developed using R [35]. In this paper, we present our package **fitdistrplus** [13] for the statistical software R. Our first objective by developping this package **fitdistrplus** 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 see [14]. 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 [40], 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.

Few packages on CRAN provide estimation procedures for a general distribution and a general type of data. The **distrMod** package of [25] provides an object-oriented (S4) implementation of probability models and includes distribution fitting procedures for a given minimization criterion. In **fitdistrplus**, we use the standard S3 class system, we believe, simpler than the full object-oriented S4 model for most R users. Furthermore, the **distrMod** package does not allow to fit censored data. The `mle` function of **stats4** package provides a procedure for maximum likelihood estimation whose output has class `"mle"`. Many generic methods are implemented for this type of object, e.g. `confint`, `logLik`,... When designing the **fitdistrplus** package, we also take this into account. Finally, various packages provide functions to estimate the mode, the moments or the L-moments of a distribution, see the reference manuals of packages **modeest**, **lmomco** and **Lmoments**.

This manuscript reviews the various features of version 1.0-0 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)
```

2 Fitting distributions to continuous non-censored data

2.1 Choice of candidate distributions

For illustrating the use of various functions of package **fitdistrplus** to help the fit of a distribution to continuous data, we will first use a data set named "ground beef" 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. This data set is used in a quantitative risk assessment published in the international journal of food microbiology journal [12].

```
> 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 family of distributions. This choice may be guided by the knowledge of stochastic processes governing the modelled variable, but also by the observation of the empirical distribution. To help the user in this observation, we developed functions to plot and characterise the empirical distribution.

First of all, the empirical distribution and density functions may be plotted using the classical R functions `ecdf` and `hist` or using Function `plotdist`. This function provides two plots (Figure 1): the left-hand plot is the histogram (on a density level) and the right-hand plot is the empirical cumulative distribution function (cdf).

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

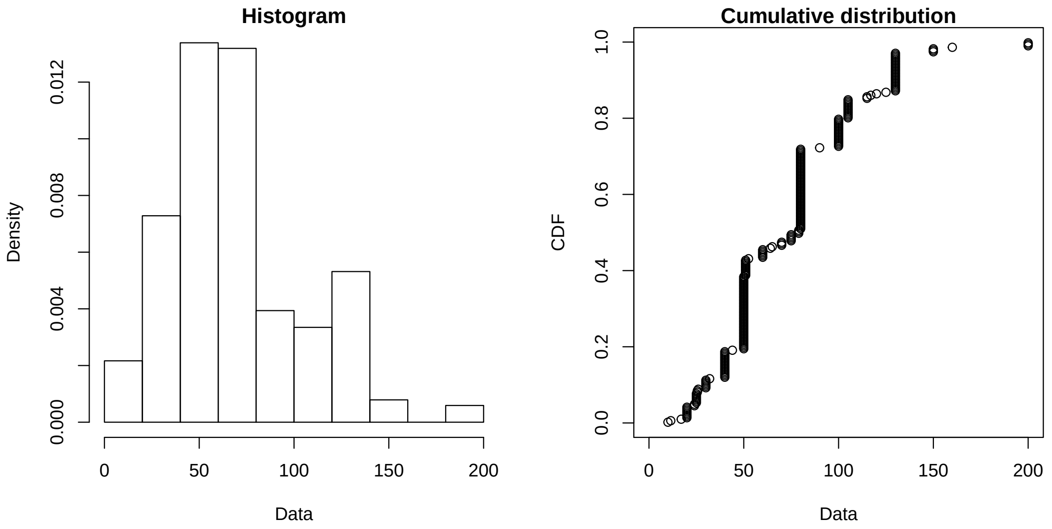

Figure 1: Density and cdf plots of an empirical distribution for a continuous variable (serving size from the "ground beef" data set)

In addition to empirical plots, descriptive statistics may help to choose good candidates to describe a distribution among a family 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 is equal to 3.

Function `descdist` provides calculations of classical descriptive statistics (minimum, maximum, median, mean, standard deviation) and skewness and Pearsons's kurtosis. By default unbiased estimations of the three last statistics are provided but the argument `method` may be used to obtain them without correction for bias. Skewness and kurtosis with their corresponding unbiased estimator of a sample $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$ 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 $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$, with x_i the n observations of variable x and $\overline{x}$ their mean value.

A skewness-kurtosis plot such as the one proposed by [10] is provided by the function `descdist` for the empirical distribution (see Figure 2 for the `groundbeef` data set). 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, a bootstrap procedure can be performed by fixing the argument `boot` to an integer above 10. `boot` bootstrap samples of the same size of the original data set are then constructed by random sampling with replacement from that original data set. Values of skewness and kurtosis are computed on that bootstrap samples and reported on 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 2). 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)
```

Together with the use of functions plotdist and `descdist` to characterize the empirical distribution, the properties of the modeled variable should be considered, especially its range, in order to choose good candidates for the next step : fit of distributions and comparison of goodness-of-fits.

2.2 Fit of distributions by maximum likelihood estimation

Once selected, one or more parametric distributions $f(.|\theta)$ may be fitted to the dataset, one at a time, using Function `fitdist`. Under the i.i.d. sample assumption, distribution parameters θ are by default estimated by maximizing the

```
summary statistics
------
min:  10   max:  200
median:  79
mean:  73.65
estimated sd:  35.88
estimated skewness:  0.7353
estimated kurtosis:  3.551
```

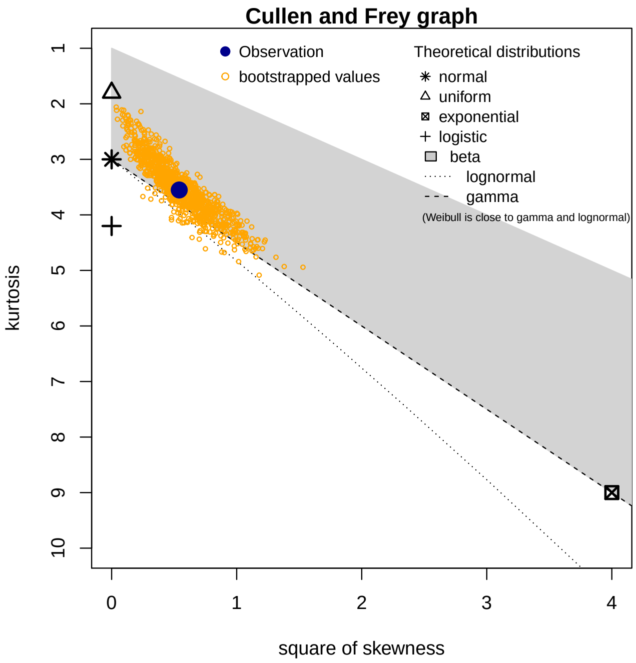

Figure 2: Skewness-kurtosis plot for a continuous variable (serving size from the `groundbeef` data set)

likelihood 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 3.1.

Function `fitdist` returns the results of the fit of any parametric distribution to a dataset as an S3 class object that may be easily printed, summarized or plotted. The parametric distribution must be a classically defined R distributions, with at least d, p and q functions respectively for the density, cumulative distribution 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, loglikelihood, and Akaike and 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” dataset.

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

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
      estimate Std. Error
shape    2.186     0.1046
scale   83.348     2.5269
Loglikelihood:  -1255   AIC:  2514   BIC:  2522
Correlation matrix:
       shape  scale
shape 1.0000 0.3218
scale 0.3218 1.0000
```

The plot of an object of class `"fitdist"` provides four classical goodness-of-fit plots [10]:

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

Functions `denscomp`, `cdfcomp`, `qqcomp` and `ppcomp`, enable to separately plot each of these four plots, in order to compare the empirical and various theoretical distributions fitted on a same dataset. These functions must be called with a first argument corresponding to a list of objects of class `fitdist`, and optionaly further arguments to customize the plot (see the reference manual [13] for lists of arguments that may be changed for each plot), as in the following example comparing the fit of Weibull, lognormal and gamma distributions to `groundbeef` dataset (Figure 3).

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

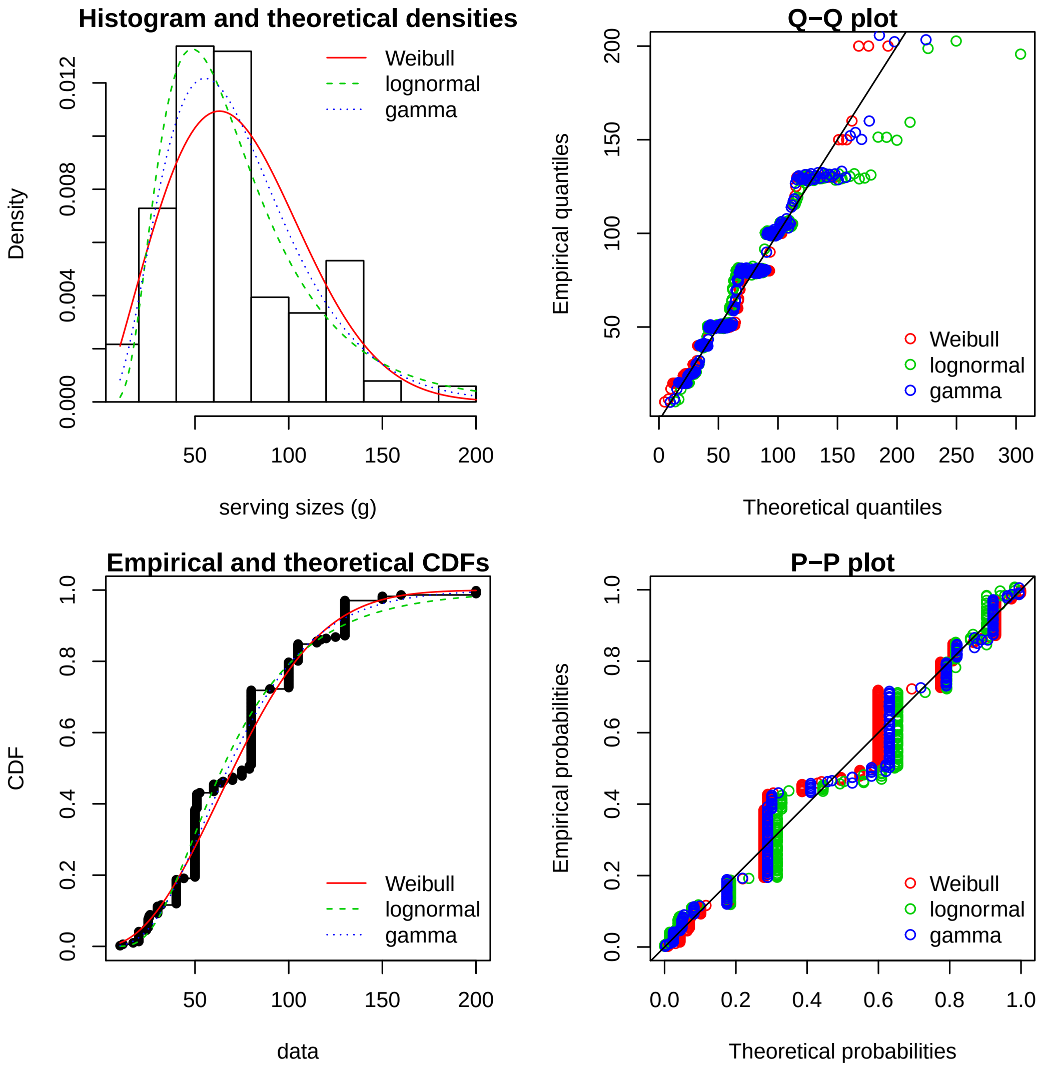

Figure 3: Four Goodness-of-fit plots for various distributions fitted on continuous data (Weibull, gamma and lognormal distributions fitted to serving sizes from the “groundbeef” dataset)

For cdf, Q-Q and P-P plots, the probability plotting position is defined as recommended by Blom [4], by default using the Hazen’s rule, with probability points of the empirical distribution defined as `(1:n - 0.5)/n`. This plotting position can be easily changed using arguments `use.ppoints` and `a.ppoints`. When `use.ppoints = TRUE`, the argument `a.ppoints` is passed to Function `ppoints` from the **stats** package to define the probability points of the empirical distribution as `(1:n - a.ppoints)/(n - 2a.ppoints + 1)`. When `use.ppoints = FALSE`, the probability points are simply defined as `1:n / n`.

If the density plot and the cdf plot are the most classical goodness-of-fits 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. As an example in Figure 3, none of the three fitted distributions describes the center of the distribution rather better than the two others, but the Weibull and gamma distributions should be prefered for their better description of the right tail of the empirical distribution, especially if the weight of this tail is important in the use of the fitted distribution.

To illustrate other features of package **fitdistrplus**, we will now use another data set named "endosulfan", which is included in our 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 (arthropods, fish or nonarthropod invertebrates) ([21]).

```
> 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 2 1 ...
 $ group     : Factor w/ 3 levels "Arthropods","Fish",..: 1 1 1 1 1 1 1 1 1 1 1 ...
```

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, named the hazardous concentration 5% (HC5) This value is interpreted as the value of the pollutant concentration protecting 95% of the species. But the fit of a lognormal or a loglogistic distribution to the whole "endosulfan" data set is rather bad, especially due to a minority of very high values. We can try to fit this data set by the Pareto distribution or the three-parameter Burr distribution which is an extension of both the loglogistic and the Pareto distribution. Pareto and Burr distributions are provided in package **actuar**. For distributions defined in R packages and for some few other 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. Function `plotdist` 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 [13] for examples). Having defined reasonable starting values, we can fit various distributions and graphically compare their fits. On this example, the use of Functions `cdfcomp` and `qqcomp` is especially interesting to evaluate the goodness-of-fit on the tail of interest while defining an $HC5$ value.

```
> 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))
> par(mfrow=c(1, 2))
> cdfcomp(list(fendo.ln,fendo.ll,fendo.P,fendo.B),xlogscale=TRUE,
+           legendtext = c("lognormal","loglogistic","Pareto","Burr"))
> qqcomp(list(fendo.ln,fendo.ll,fendo.P,fendo.B),xlogscale=TRUE,ylogscale=TRUE,
+        legendtext = c("lognormal","loglogistic","Pareto","Burr"))
```

We can see in Figure 4 that none of the fitted distribution correctly describes the right tail observed in the data set, but that 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 generic Function quantile defined for an object of class `"fitdist"`.

```
> (HC5 <- quantile(fendo.B,probs = 0.05))

Estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2939
```

To go further in the comparison of various distributions, we propose the calculation of different goodness-of-fit statistics in our package. The purpose of goodness-of-fit statistics aims to measure the distance between the fitted parametric distribution and the empirical distribution. Such a distance can be measured beetween the cumulative distribution function F defined from the fitted parametric distribution with the empirical distribution function F_n

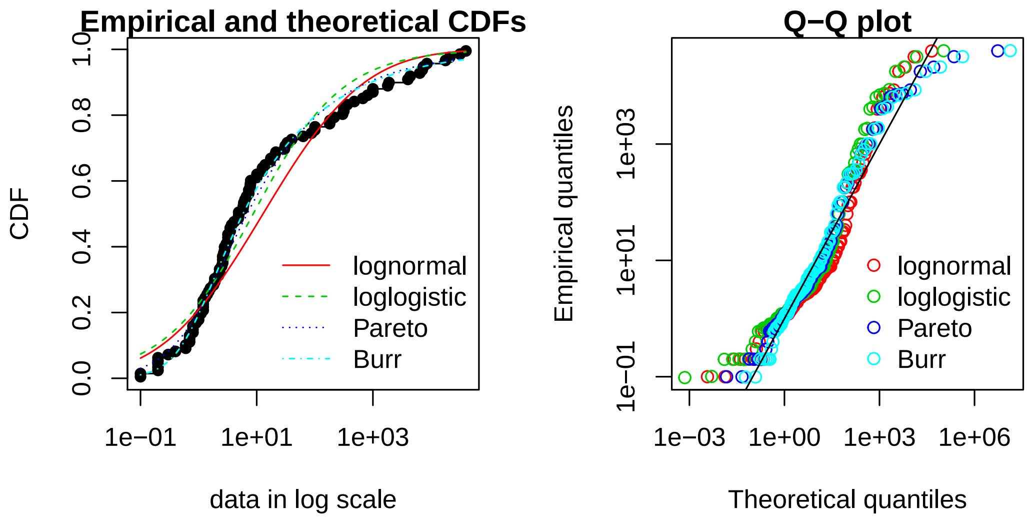

Figure 4: CDF and Q-Q plots to compare the fit of four distributions to acute toxicity values of various organisms for the organochlorine pesticide endosulfan ("endosulfan" dataset)

based on the data. When fitting continuous distributions, three goodness-of-fit statistics are classicaly considered: Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling statistics. They can be computed using the function `gofstat` as defined by Stephens [11]. 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.

Table 1: Goodness-of-fit statistics as defined by Stephens [11].

Statistic	General formula	Computational formula
Kolmogorov-Smirnov (KS)	$\sup \lvert F_n(x) - F(x) \rvert$	$\max(D^+, D^-)$ with $D^+ = \max\limits_{i=1,\dots,n} \left(\frac{i}{n} - F(x_i) \right)$; $D^- = \max\limits_{i=1,\dots,n} \left(F(x_i) - \frac{i-1}{n} \right)$
Cramer-von Mises (CvM)	n $\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum\limits_{i=1}^{n} \left(F(x_i) - \frac{2i-1}{2n} \right)^2$
Anderson-Darling (AD)	n $\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} \sum\limits_{i=1}^{n} ((2i-1)(\log(F(x_i)) + \log(1 - F(x_{n+1-i}))))$

```
> gofstatres <- cbind(
+   ln = c(aic=summary(fendo.ln)$aic,bic=summary(fendo.ln)$bic, unlist(gofstat(fendo.ln)[c("cvm","ks","ad"
+   ll = c(aic=summary(fendo.ll)$aic, bic=summary(fendo.ll)$bic, unlist(gofstat(fendo.ll)[c("cvm","ks","ad
+   P = c(aic=summary(fendo.P)$aic, bic=summary(fendo.P)$bic, unlist(gofstat(fendo.P)[c("cvm","ks","ad")])
+   B = c(aic=summary(fendo.B)$aic, bic=summary(fendo.B)$bic, unlist(gofstat(fendo.B)[c("cvm","ks","ad")])
+ )
```

```
> gofstatres
```

```
           ln        ll         P         B
aic 1068.8104 1069.2458 1.048e+03 1.046e+03
bic 1074.0992 1074.5346 1.053e+03 1.054e+03
cvm    0.6374    0.3827 1.393e-01 6.803e-02
ks     0.1672    0.1196 8.488e-02 6.155e-02
ad     3.4721    2.8316 8.921e-01 5.239e-01
```

REPLACE BY A CALL TO PRINT OF NEW FUNCTION gofstat TODO

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 [10, 43]. 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 is dependent of the theoretical distribution in its definition. Anderson-Darling statistics computed for several distributions fitted on the same data set are thus theoretically difficult to compare. Moreover, such a statistics, as Cramer-von Mises and Kolmogorov-Smirnov ones, does not take into account the complexity of the model. 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 seems thus also interesting, especially to discourage overfitting.

In the previous example, all the goodness-of-fit statistics based on the cdf distance encourage the choice of the Burr function, the only one characterized by three parameters, while Akaike and Schwarz information criteria (so called AIC and BIC) respectively gives 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 3.3 and the reference manual [13] for examples).

2.3 Uncertainty in parameter estimates

The uncertainty in the parameters of the fitted distribution may be simulated by parametric or nonparametric bootstrap using the `boodist` function for non censored data. These functions return the bootstrapped 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 (see Figure 5).

Below is an example of the use of the `bootdist` function with the previous of the Weibull distribution to `endosulfan` dataset.

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

Parametric bootstrap medians and 95% percentile CI
       Median    2.5%  97.5%
shape1 0.1986 0.09096 0.3677
shape2 1.6081 1.04393 3.0045
rate   1.5184 0.68257 2.8134

> plot(bendo.B)
```

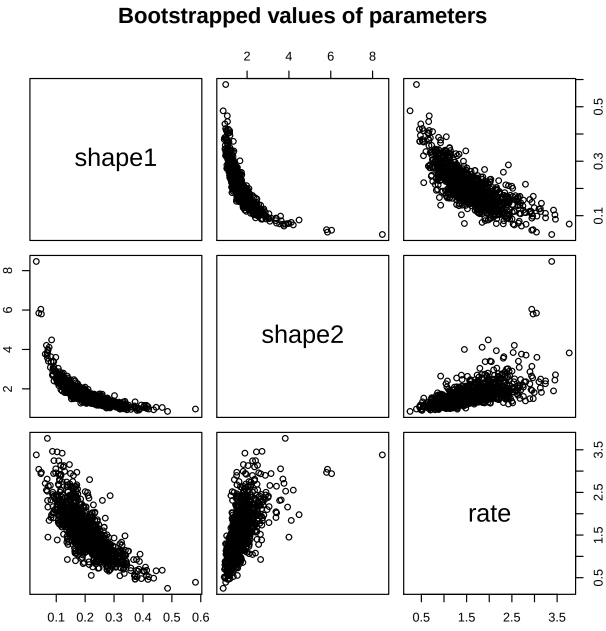

Figure 5: Bootstrappped values of parameters for a fit of a distribution characterized by three parameters (example on the fit of a Burr distribution to acute toxicity values from the `endosulfan` dataset)

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.

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** ([32]). One could refer to Pouillot *et al.* ([31]) for an introduction to the use of **mc2d** and **fitdistrplus** packages in the context of quantitative risk assessment.

Bootstrap 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% bootstrap confidence intervals of quantiles are provided. Going back to he previous axample from ecotoxicolgy, this function can be used to estimate the uncertainty associated to the HC5 estimation, for example from the previously fitted Burr distribution.

```
> quantile(bendo.B, probs = 0.05)
```

```
(original) estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2939
Median of bootstrap estimates
         p=0.05
estimate 0.3033

two-sided 95 % CI of each quantile
       p=0.05
2.5 %  0.1747
97.5 % 0.5025
```

3 Advanced topics

3.1 Alternative methods for parameter estimation

Despite maximum likelihood estimation is the default estimation proposed by `fitdist`, other classical estimation methods can be handled to estimate parameters for non-censored data. Thus, this subsection focuses on alternative estimation methods.

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 Table 1, or one of the variants of the Anderson-Darling distance proposed by [28] and defined in Table 2. 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.

Table 2: Modified Anderson-Darling statistics as defined by Luceno [28].

Statistic	General formula	Computational formula
Right-tail AD (ADR)	$\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{1-F(x)} dx$	$\frac{n}{2} - 2\sum_i F(x_i) - \frac{1}{n}\sum_i((2i-1)ln(1-F(x_{n+1-i})))$
Left-tail AD (ADL)	$\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(F(x))} dx$	$-\frac{3n}{2} + 2\sum_i F(x_i) - \frac{1}{n}\sum_i((2i-1)ln(F(x_i)))$
Right-tail AD 2nd order (AD2R)	$ad2r = \int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(1-F(x))^2} dx$	$ad2r = 2\sum_i ln(1-F(x_i)) + \frac{1}{n}\sum_i \frac{2i-1}{1-F(x_{n+1-i})}$
Left-tail AD 2nd order (AD2L)	$ad2l = \int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(F(x))^2} dx$	$ad2l = 2\sum_i ln(F(x_i)) + \frac{1}{n}\sum_i \frac{2i-1}{F(x_i)}$
AD 2nd order (AD2)	$ad2r + ad2l$	$ad2r + ad2l$

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.

Maximum goodness-of-fit estimation may be useful to give more weight to data at one tail of the distribution. Let us go back to the previous example from ecotoxicology. Instead of trying to find a less classical distribution to correctly fit the empirical distribution especially on its left tail, one could consider the fit of the classical lognormal distribution, but maximizing a goodness-of-fit distance giving more weight to the left tail of the empirical distribution, so as to correctly estimate the 5% percentile. In the following example of `endosulfan` dataset, we use left tail Anderson-Darling distances of first or second order (see Figure 6).

```
> fendo.ln.ADL <- fitdist(ATV,"lnorm",method="mge",gof="ADL")
> fendo.ln.AD2L <- fitdist(ATV,"lnorm",method="mge",gof="AD2L")
> cdfcomp(list(fendo.ln, fendo.ln.ADL, fendo.ln.AD2L),
+ xlogscale = TRUE, main = "",
+ legendtext = c("maximum likelihood",
+ "Left-tail Anderson-Darling", "Left tailed Anderson-Darling of second order"),cex=0.7,
+ xlegend = 500, ylegend = 0.15)
```

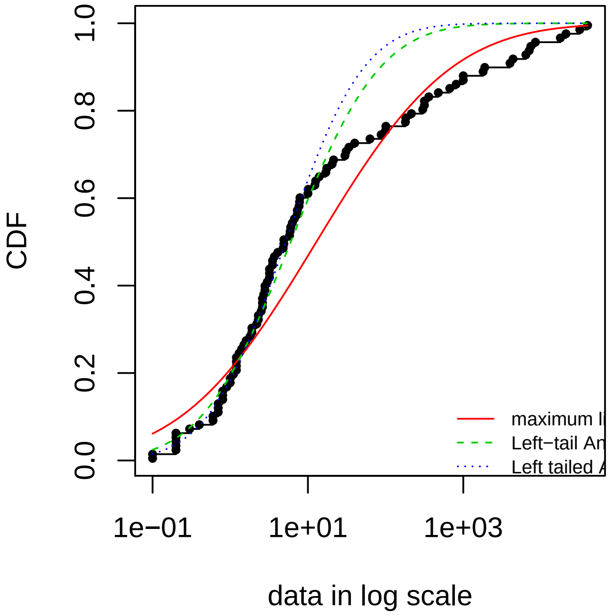

Figure 6: Comparison of one distribution fitted by maximum likelihood estimation and by maximum goodness-of-fit using two different goodness-of-fit distances (example of a lognormal distribution fitted to acute toxicity values from the `endosulfan` dataset)

Comparing the 5% percentiles (HC5) calculated using these three fits to the one calculated from the MLE fit of the Burr distribution, we can observe, on this example, that fitting the lognormal distribution by maximizing left tail Anderson-Darling distances of first or second order gives rather similar results than fitting the Burr distribution by maximum likelihood.

```
> (HC5.B.MLE <- quantile(fendo.B,probs = 0.05))

Estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2939

> (HC5.ln.MLE <- quantile(fendo.ln,probs = 0.05))

Estimated quantiles for each specified probability (non-censored data)
          p=0.05
estimate 0.07259

> (HC5.ln.ADL <- quantile(fendo.ln.ADL,probs = 0.05))

Estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.1959

> (HC5.ln.AD2L <- quantile(fendo.ln.AD2L,probs = 0.05))

Estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2588
```

Another method commonly used to fit parametric distribution consists in estimating the parameters θ at the values that makes the first theoretical raw moments of the parametric distribution equal to the empirical moments (Equation 4).

$$E(X^k|\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{4}$$

for $k = 1, \dots, p$, with p the number of parameters to estimate and x_i the n observations of variable x. For moments of order greater or equal than 2, it is also relevant to match centered moments as given by Equation (5).

$$E\left((X - E(X))^k|\theta\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k \tag{5}$$

This method called moment matching estimation can be performed fixing the argument `method` to `"mme"` in the call to fitdist. The estimate is computed by a closed formula for following distributions: normal, lognormal, exponential, Poisson, gamma, logistic, negative binomial, geometric, beta and uniform distributions (i.e. base R distributions). In this case, 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 (see e.g. [43]). Otherwise, for not so-common distributions, the equation of moments is solved numerically using the `optim` function by minimizing the sum of squared differences between observed and theoretical moments (see the **fitdistrplus** reference manual [13] for technical details).

To illustrate this method, we will use a classical dataset from the Danish insurance industry published in [30]. In **fitdistrplus**, the dataset is stored in `danishuni` and `danishmulti` for univariate and multivariate versions, respectively.

```
> data(danishuni)
> str(danishuni)

'data.frame':        2167 obs. of  2 variables:
 $ Date: Date, format: "1980-01-03" "1980-01-04" ...
 $ Loss: num  1.68 2.09 1.73 1.78 4.61 ...

> data(danishmulti)
> str(danishmulti)

'data.frame':        2167 obs. of  5 variables:
 $ Date    : Date, format: "1980-01-03" "1980-01-04" ...
 $ Building: num  1.1 1.76 1.73 0 1.24 ...
 $ Contents: num  0.586 0.337 0 1.305 3.367 ...
 $ Profits : num  0 0 0 0.474 0 ...
 $ Total   : num  1.68 2.09 1.73 1.78 4.61 ...
```

Our first example of fitting a lognormal distribution on `danish` dataset uses a closed formula. Comparing the two fitted distributions functions, we observe on Figure 7 that the moment matching estimation is far more conservative than the maximum likelihood estimation, which is also more conservative than goodness-of-fit estimation.

```
> flndanishMLE <- fitdist(danishuni$Loss, "lnorm")
> flndanishMME <- fitdist(danishuni$Loss, "lnorm", method="mme", order=1:2)
> cdfcomp(list(flndanishMME, flndanishMLE),
+         legend=c("MME", "MLE"), main="Fitting lognormal distribution",
+         xlogscale=TRUE, datapch="*")
```

Our second example is the fitting of a Pareto type II distribution. We use the implementation of **actuar** package providing moments and limited expected value for that distribution (in addition to `d`, `p`, `q` and `r` functions, see [19]). Fitting a heavy-tailed distribution for which the first and the second moments do not exist for certain values of the shape parameter requires some cautiousness. This is carried out by providing a lower and an upper bounds for the optimization by `optim`. Our call below immadiately calls the L-BFGS-B optimization method, since this quasi-Newton allows box constraints[1]. We also observe that the fitting is relatively good when comparing empirical and fitted moments. Note that we have to pass a function for computing the empirical raw moment to `fitdist`.

```
> library(actuar)
> memp <- function(x, order) ifelse(order == 1, mean(x), sum(x^order)/length(x))
> fparedanishMME <- fitdist(danishuni$Loss, "pareto", method="mme", order=1:2,
+       memp="memp", start=c(shape=10, scale=10), lower=2+1e-6, upper=Inf)
> c(theo = mpareto(1, fparedanishMME$estimate[1], fparedanishMME$estimate[2]),
+ emp = memp(danishuni$Loss, 1))

 theo   emp
3.448 3.385

> c(theo = mpareto(2, fparedanishMME$estimate[1], fparedanishMME$estimate[2]),
+ emp = memp(danishuni$Loss, 2))
```

[1] That's what the B stands for.

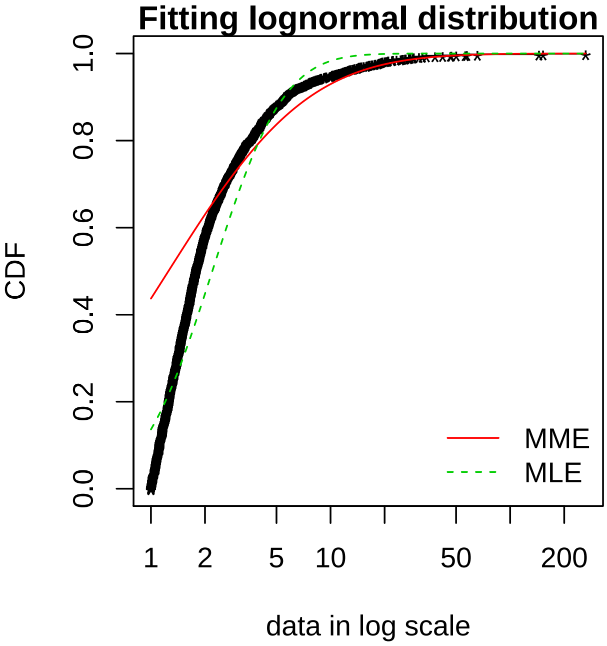

Figure 7: Comparison between MME and MLE when fitting lognormal distribution on `danishuni`

```
theo  emp
83.8 83.8
```

Fitting of a parametric distribution may also be done by matching theoretical quantiles of the parametric distributions (for specified probabilities) to the empirical quantiles. Equation (6) below is thus similar to Equations (4) and (5)

$$F^{-1}(p^k|\theta) = Q_{n,p_k} \tag{6}$$

for $k = 1, \ldots, p$, with p the number of parameters to estimate (dimension of θ if there is no fixed parameters) and Q_{n,p_k} the empirical quantiles calculated from data for specified probabilities p_k.

Quantile matching can be performed by 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. Empirical quantiles are computed using the `quantile` function of the **stats** package using the `type` argument equal to 7 by default, but the type of quantile can be easily changed by using the `qty` argument in the call to the `qme` function. The quantile matching is carried out numerically, by minimizing the sum of squared differences between observed and theoretical quantiles.

```
> flndanishQME1 <- fitdist(danishuni$Loss, "lnorm", method="qme", probs=c(1/3, 2/3))
> flndanishQME2 <- fitdist(danishuni$Loss, "lnorm", method="qme", probs=c(3/4, 4/5))
> cdfcomp(list(flndanishQME1, flndanishQME2, flndanishMLE),
+         legend=c("QME(1/3, 2/3)", "QME(3/4, 4/5)", "MLE"), main="Fitting lognormal distribution",
+         xlogscale=TRUE, datapch="*")
```

Above is an example of fitting of a lognormal distribution to `danishuni` dataset by matching probabilities ($p_1 = 1/3, p_2 = 2/3$) and ($p_1 = 3/4, p_2 = 4/5$). As expected, the second QME fit is more conservative when looking at the tail of the distributions. Compared to the maximum likelihood estimation, the second QME fit is also more conservative, whereas the first QME fit is less conservative. The quantile matching estimation is of particular interest when we need a good precision around particular quantiles, e.g. $p = 99.5\%$ for Solvency II insurance context.

3.2 Customization of the optimization algorithm

Each time a numerical minimization (or maximization) is carried out using `fitdist`, the `optim` function of the **stats** package 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 `optim` function or to use another optimization function than `optim` to maximize the likelihood or to minimize a squared difference.

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 derivative-free Nelder and Mead method), `"BFGS"` (the BFGS quasi-Newton method), `"CG"` (the conjugate gradient hessian-free method), `"SANN"` (a variant of (stochastic) simulated annealing) or `"L-BFGS-B"` (a modification of the BFGS quasi-Newton method which enables box constraints optimization and limited-memory usage). For the use of the last method the arguments `lower`

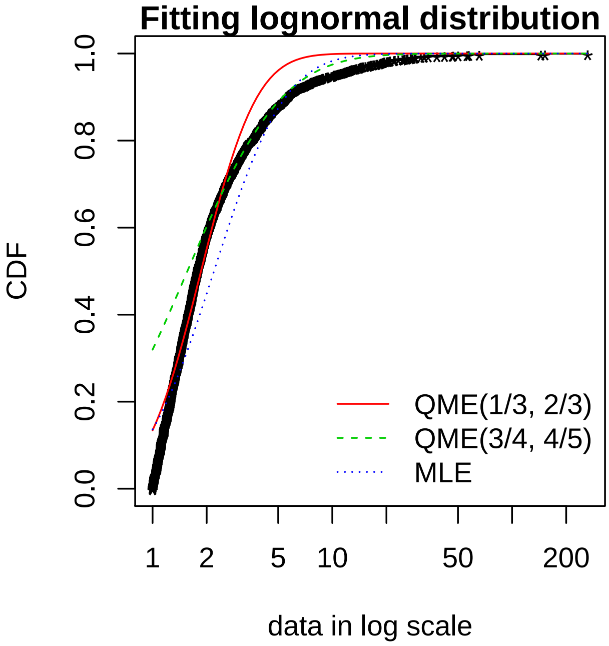

Figure 8: Comparison between QME and MLE when fitting lognormal distribution on `danishuni`

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 `groundbeef` dataset with various options of `optim`. Note that the conjugate gradient algorithm needs far more iterations to converge (around 2500 iterations) compared to other algorithms (converging in less than 100 iterations).

```
> data(groundbeef)
> fNM <- fitdist(groundbeef$serving, "gamma", optim.method="Nelder-Mead")
> fBFGS <- fitdist(groundbeef$serving, "gamma", optim.method="BFGS")
> fSANN <- fitdist(groundbeef$serving, "gamma", optim.method="SANN")
> fCG <- try(fitdist(groundbeef$serving, "gamma", optim.method="CG", control=list(maxit=10000)))
> if(class(fCG) == "try-error")
+   fCG <- list(estimate=NA)
```

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. We assume that `custom.optim` should carry out a MINIMIZATION and must 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 wrap `genoud` function from **rgenoud** package in order to respect our optimization “template”. The **rgenoud** package implements the genetic (stochastic) algorithm.

```
> 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 `groundbeef` dataset. Note that in this example various arguments are also passed from `fitdist` to `genoud` : `nvars`, `Domains`, `boundary.enforcement`, `print.level` and `hessian`. The code below compare all the parameter estimates by the different algorithms: shape and rate parameters are relatively the same.

```
> fgenoud <- mledist(groundbeef$serving, "gamma", custom.optim= mygenoud, nvars=2,
+     max.generations=10, Domains=cbind(c(0,0), c(10,10)), boundary.enforcement=1,
+     hessian=TRUE, print.level=0, P9=10)
> cbind(NM=fNM$estimate,
+ BFGS=fBFGS$estimate,
+ SANN=fSANN$estimate,
+ CG=fCG$estimate,
+ fgenoud=fgenoud$estimate)

           NM    BFGS    SANN      CG fgenoud
shape 4.00825 4.22848 3.97890 4.12891 4.00834
rate  0.05442 0.05742 0.05418 0.05607 0.05443
```

3.3 Fitting distributions to other types of data

WRITE AN INTRODUCTION, change the example for censored data, and revise all this part .... TODO

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.

The `smokedfish` dataset, included in the package, 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 ([6]). 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 ...
```

3.3.1 Graphical display of the observed distribution

Using censored data such as those coded in the `smokedfish` dataset, the empirical distribution may be plotted using the `plotdistcens` function. By default, this function uses the EM approach of Turnbull [41] to compute the overall empirical cdf curve with confidence intervals, by calls to `survfit` and `plot.survfit` functions from the **survival** package. Let us see such a plot for `smokedfish` dataset after classical transformation of microbial counts in decimal logarithm (Figure 9).

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

3.3.2 Maximum likelihood estimation

As for non censored data, one or more parametric distributions may then be fitted to the censored dataset, one at a time, but using in this case the `fitdistcens` function. This function estimates distribution parameters θ by maximizing the likelihood for censored data defined as:

$$L(\theta) = \prod_{i=1}^{N_{nonC}} f(x_i|\theta) \times \prod_{j=1}^{N_{leftC}} F(x_j^{upper}|\theta) \times \prod_{k=1}^{N_{rightC}} (1 - F(x_k^{lower}|\theta)) \times \prod_{m=1}^{N_{intC}} (F(x_m^{upper}|\theta) - F(x_j^{lower}|\theta)) \qquad (7)$$

with x_i the N_{nonC} non-censored observations, x_j^{upper} upper values defining the N_{leftC} left-censored observations, x_k^{lower} lower values defining the N_{rightC} right-censored observations, $[x_m^{lower} ; x_m^{upper}]$ the intervals defining the N_{intC} interval-censored observations, and F the cumulative distribution function of the parametric distribution.

As `fitdist`, it returns the results of the fit of any parametric distribution to a dataset as an S3 class object that may be easily printed, summarized or plotted. For “smokedfish” dataset, a normal distribution may be fitted to log transformed data as commonly done for microbial count data.

```
> flog10Cn <- fitdistcens(log10C, "norm")
> summary(flog10Cn)
```

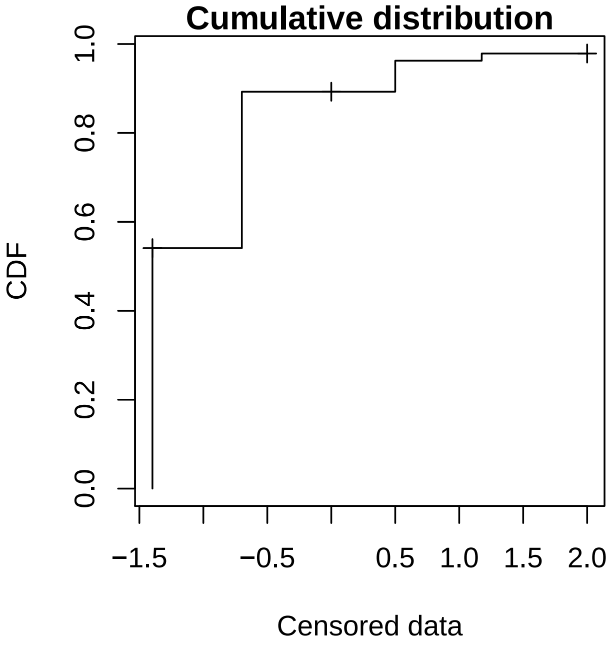

Figure 9: CDF plot of censored data (microbial counts from the `smokedfish` dataset)

```
FITTING OF THE DISTRIBUTION ' norm ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
     estimate Std. Error
mean   -1.575     0.2014
sd      1.539     0.2118
Loglikelihood:  -87.11   AIC:  178.2   BIC:  183.5
Correlation matrix:
        mean      sd
mean  1.0000 -0.4325
sd   -0.4325  1.0000
```

As with `fitdist`, for some distributions (see [13] for details), it is necessary to specify initial values for the distribution parameters in the argument `start`. The `plotdistcens` function can help to find correct initial values for the distribution parameters in non trivial cases, by an manual iterative use if necessary.

Only one goodness-of-fit plot is provided for censored data, corresponding to the theoretical cumulative distribution function added to the plot of censored data presented in Section 3.3.1. The `cdfcompcens` function can be used to compare the fit of various distributions to the same censored dataset. Its call is similar to the one `cdfcomp`. Below is an example of comparison of two fitted distribution to `smokedfish` dataset (see Figure 10).

```
> flog10Cl <- fitdistcens(log10C, "logis")
> cdfcompcens(list(flog10Cn, flog10Cl),
+     legendtext=c("normal distribution", "logistic distribution"),
+     xlab="bacterial concentration (log10[CFU/g])", ylab="F")
```

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.

The `toxocara` dataset 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 ([17]). We will use it in order to illustrate the case of discrete data.

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

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

In some cases a discrete variable may be plotted as a continuous one, for example for a large dataset 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 11):

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

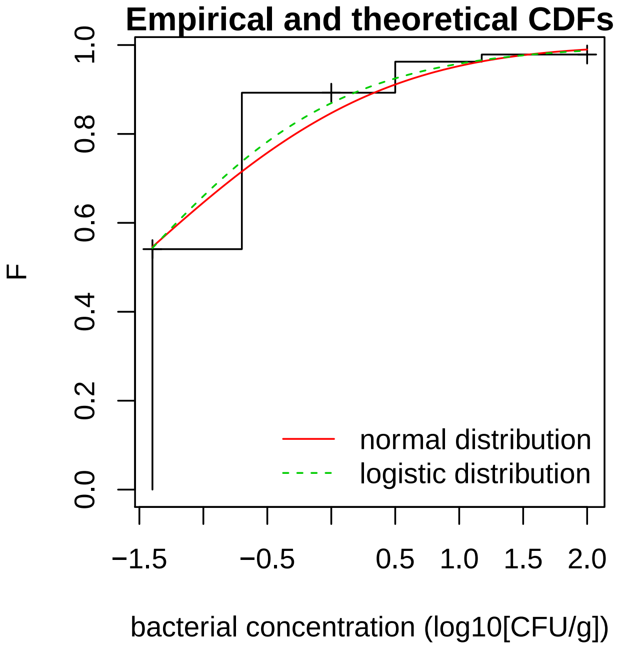

Figure 10: Goodness-of-fit CDF plots for fits of continuous distributions on censored data (Comparison of lognormal and loglogistic distributions fitted to microbial counts from the `smokedfish` dataset)

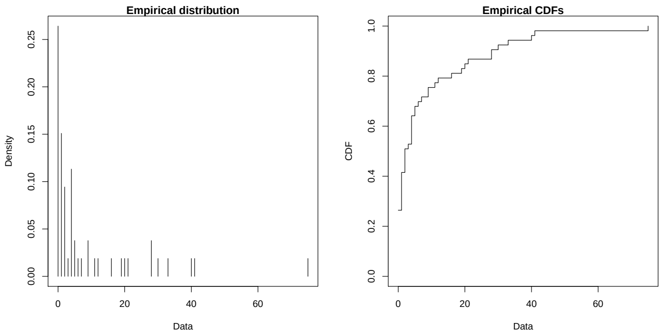

Figure 11: Density and cdf plots of an empirical distribution for a discrete variable (number of *Toxocara cati* parasites from the `toxocara` dataset)

As for continuous non-censored data (see Section **??**) Function `descdist` can be used, but with the argument `discrete` fixed to `TRUE`. This function will especially compute skewness and kurtosis values, and plot them in a skewness-kurtosis plot with skewness and kurtosis values or set of values of Poisson and negative binomial together with values for the normal distribution, to which discrete distributions may converge.

The fit of a discrete distribution to discrete data by maximum likelihood estimation requires the same procedure as for continuous non-censored data. As an example, using the `toxocara` dataset, 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")
> summary(fp)

Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
       estimate Std. Error
lambda    8.679     0.4047
Loglikelihood:  -507.5   AIC:  1017   BIC:  1019

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

Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
     estimate Std. Error
size   0.3971    0.08289
```

```
mu     8.6803    1.93501
Loglikelihood:  -159.3   AIC:  322.7   BIC:  326.6
Correlation matrix:
           size         mu
size  1.0000000 -0.0001039
mu   -0.0001039  1.0000000
```

For discrete distributions, the plot of an object of class `"fitdist"` simply 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 the `toxocara` dataset.

```
> plot(fnb)
```

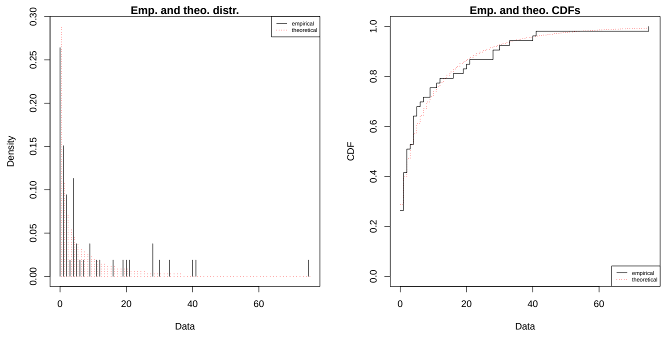

Figure 12: Plot of the fit of a discrete distribution (a negative binomial distribution fitted to numbers of *Toxocara cati* parasites from the `toxocara` dataset)

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 [43]. Using this default option with the fit of a negative binomial distribution to `toxocara` dataset gives following results :

```
> gofstat(fnb)
```

```
[1] "1-mle-nbinom"
Chi-squared statistic:  7.486
Degree of freedom of the Chi-squared distribution:  4
Chi-squared p-value:  0.1123
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theocounts
<= 0     14.000     15.295
<= 1      8.000      5.809
<= 3      6.000      6.845
<= 4      6.000      2.408
<= 9      6.000      7.835
<= 21     6.000      8.271
> 21      7.000      6.537
```

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

```
> gofstat(fnb)$chisqtable
```

```
[1] "1-mle-nbinom"
      obscounts theocounts
<= 0     14.000     15.295
<= 1      8.000      5.809
<= 3      6.000      6.845
<= 4      6.000      2.408
<= 9      6.000      7.835
<= 21     6.000      8.271
> 21      7.000      6.537
```

4 Conclusion

Papers citing **fitdistrplus** are [26, 7, 37, 27, 22, 44, 5, 1, 39, 38, 33, 29, 24, 20, 23, 18, 16, 9, 8, 3, 2, 15, 36]

References

[1] Özlem Aktaş and Maria Sjöstrand. Cornish-fisher expansion and value-at-risk method in application to risk management of large portfolios. Master's thesis, School of Information Science, Computer and Electrical Engineering, Halmstad University, 2011. only mentionned page 83. 18

[2] Praveen Anand, Kalidas Yeturu, and Nagasuma Chandra. Pocketannotate: towards site-based function annotation. *Nucleic Acids Research*, 40(W1 W400-W408):1–9, 2012. C-F graph and plotdist graph - page 5. 18

[3] Anurag Bagaria, Victor Jaravine, Y.J. Huang, G.T. Montelione, and Peter Güntert. Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Science*, 21(2):229–238, 2012. 18

[4] G. Blom. *Statistical Estimates and Transformed Beta Variables.* Wiley, New York, 1959. 5

[5] J.P. Brooks, D.J. Edwards, T.P. Sorrell, S. Srinivasan, and R.L. Diehl. Simulating calls for service for an urban police department. In *the 2011 Winter Simulation Conference*, pages 1770–1777, 2011. cited page 7. 18

[6] 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. 14

[7] 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*, 138:260–269, 2010. cited page 261. 18

[8] Natalie Commeau, Eric Parent, Marie-Laure Delignette-Muller, and Marie Cornu. Fitting a lognormal distribution to enumeration and absence/presence data. *International Journal of Food Microbiology*, 155:146–152, 2012. censored data. 18

[9] Nicholas J. Croucher, Simon R. Harris, Lars Barquist, Julian Parkhill, and Stephen D. Bentley. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog*, 8(6):e1002745, 2012. fit and confint. 18

[10] A.C. Cullen and H.C. Frey. *Probabilistic Techniques in Exposure Assessment.* Plenum Publishing Co., New York, first edition, 1999. 3, 4, 7

[11] R.B. D'Agostino and M.A. Stephens. *Goodness-of-Fit Techniques.* Dekker, New York, first edition, 1986. 7

[12] 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. 2

[13] 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. 2, 5, 6, 8, 11, 15

[14] C. Dutang, V. Goulet, and M. Pigeon. actuar: an R package for Actuarial Science. *Journal of Statistical Software*, 25(7), 2008. 2

[15] Marika Eik and Heiko Herrmann. Raytraced images for testing the reconstruction of fibre orientation distributions. In *the Estonian Academy of Sciences*, volume 61, pages 128–136, 2012. cited page 9. 18

[16] Martin Eling. Fitting insurance claims to skewed distributions: Are the skew-normal and the skew-student good models? *Insurance: Mathematics and Economics*, 51(2012):239–248, 2012. cited page 241. 18

[17] 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. 15

[18] C.H.Y. Fu, H. Steiner, and S.G. Costafreda. Predictive neural biomarkers of clinical response in depression: A meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. *Neurobiology of Disease*, 2012. cited in page 3 for grouped data. 18

[19] V. Goulet. *actuar: An R Package for Actuarial Science, version 1.1-5.* École d'actuariat, Université Laval, 2012. 11

[20] K. Hoelzer, R. Pouillot, D. Gallagher, M.B. Silverman, J. Kause, and S. Dennis. Estimation of Listeria monocytogenes transfer coefficients and efficacy of bacterial removal through cleaning and sanitation. *International Journal of Food Microbiology*, 157(2):267–277, 2012. Cited in page 9 for parametric bootstraping. 18

[21] 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. 6

[22] S. Jaloustre, M. Cornu, E. Morelli, V. Noel, and M.L. Delignette-Muller. Bayesian modeling of Clostridium perfringens growth in beef-in-sauce products. *Food microbiology*, 28(2):311–320, 2011. cited in page 4 for MLE fit. 18

[23] I. Jongenburger, M.W. Reij, E.P.J. Boer, M.H. Zwietering, and L.G.M. Gorris. Modelling homogeneous and heterogeneous microbial contaminations in a powdered food product. *International Journal of Food Microbiology*, 157(1):35–44, 2012. cited in page 4 for MLE censord fit. 18

[24] F.H. Koch, D. Yemshanov, R.D. Magarey, and W.D. Smith. Dispersal of invasive forest insects via recreational firewood: A quantitative analysis. *Journal of Economic Entomology*, 105(2):438–450, 2012. cited in page 4 for MLE fit. 18

[25] M. Kohl and P. Ruckdeschel. R package distrMod: S4 Classes and Methods for Probability Models. *Journal of Statistical Software*, 35(10), 2010. 2

[26] M. Kohl and P. Ruckdeschel. R package distrMod: S4 classes and methods for probability models. *Journal of Statistical Software*, 35(10):1–27, 2010. cited in page 17 for MLE fit. 18

[27] A. Leha, T. Beissbarth, and K. Jung. Sequential interim analyses of survival data in DNA microarray experiments. *BMC Bioinformatics*, 12(127):1–14, 2011. cited in page 10 for MLE censored fit. 18

[28] A. Luceno. Fitting the Generalized Pareto Distribution to Data using Maximum Goodness-of-fit Estimators. *Computational Statistics and Data Analysis*, 51(2):904–917, NOV 15 2006. 9

[29] N. Marquetoux, M. Paul, S. Wongnarkpet, C. Poolkhet, W. Thanapongtham, F. Roger, C. Ducrot, and K. Chalvet-Monfray. Estimating spatial and temporal variations of the reproduction number for highly pathogenic avian influenza H5N1 epidemic in Thailand. *Preventive Veterinary Medicine*, 106(2):143–151, 2012. cited inp age 3 for MLE fit. 18

[30] A.J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bull.*, 1997. 11

[31] 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. 9

[32] 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. 9

[33] R. Pouillot, K. Hoelzer, Y. Chen, and S. Dennis. Estimating probability distributions of bacterial concentrations in food based on data generated using the most probable number (MPN) method for use in risk assessment. *Food Control*, 29(2):350–357, 2012. 18

[34] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0. 2

[35] Ricci, V. Fitting distributions with r. Contributed Documentation available on CRAN, 2005. 2

[36] A. S. Rosa. Funções de predição espacial de propriedades do solo. Master's thesis, Universidade Federal de Santa Maria, 2012. cited in page 72 for descdist and plotdist. 18

[37] H. Sak and C. Haksoz. A copula-based simulation model for supply portfolio risk. *Journal of Operational Risk*, 2011. cited in page 9 for MLE fit. 18

[38] C.F. Scholl, C.C. Nice, J.A. Fordyce, Z. Gompert, and M.L. Forister. Larval performance in the context of ecological diversification and speciation in lycaeides butterflies. *International Journal of Ecology*, 2012(2012):1–13, 2012. cited in page 4 for MLE fit. 18

[39] J.P. Suuronen, Aki Kallonen, Marika Eik, Jari Puttonen, Ritva Serimaa, and Heiko Herrmann. Analysis of short fibres orientation in steel fibre-reinforced concrete (SFRC) by X-ray tomography. *Journal of Materials Science*, 2012. cited in page 5 for MLE fit. 18

[40] T. Therneau. *survival: Survival Analysis, Including Penalized Likelihood*, 2011. R package version 2.36-9. 2

[41] BW Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, 69(345):169–173, 1974. 14

[42] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, 4 edition, 2010. 2

[43] D. Vose. *Quantitative Risk Analysis. A Guide to Monte Carlo Simulation Modelling.* Wiley, New York, first edition, 2010. 7, 11, 17

[44] T. Wilson. What were they thinking: modeling think times for performance testing. *CMG Journal Information*, 2011. cited page 9 for MLE fit. 18