Fitting distributions by MLE to censored data

Marie Laure Delignette Muller, Christophe Dutang

2016-04-15

Contents

1 Fitting distributions to other types of data

Analytical methods often lead to semi-quantitative results which are referred to as censored data. Observations only known to be under a limit of detection are left-censored data. Observations only known to be above a limit of quantification are right-censored data. Results known to lie between two bounds are interval-censored data. These two bounds may correspond to a limit of detection and a limit of quantification, or more generally to uncertainty bounds around the observation. Right-censored data are also commonly encountered with survival data (Klein and Moeschberger 2003). A data set may thus contain right-, left-, or interval-censored data, or may be a mixture of these categories, possibly with different upper and lower bounds. Censored data are sometimes excluded from the data analysis or replaced by a fixed value, which in both cases may lead to biased results. A more recommended approach to correctly model such data is based upon maximum likelihood [@].

Censored data may thus contain left-censored, right-censored and interval-censored values, with several lower and upper bounds. Before their use in package **fitdistrplus**, such 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. To illustrate the use of package **fitdistrplus** to fit distributions to censored continous data, we will use another data set from ecotoxicology, included in our package and named `salinity`. This data set contains acute salinity tolerance (LC50 values in electrical conductivity, $mS.cm^{-1}$) of riverine macro-invertebrates taxa from the southern Murray-Darling Basin in Central Victoria, Australia (Kefford et al. 2007).

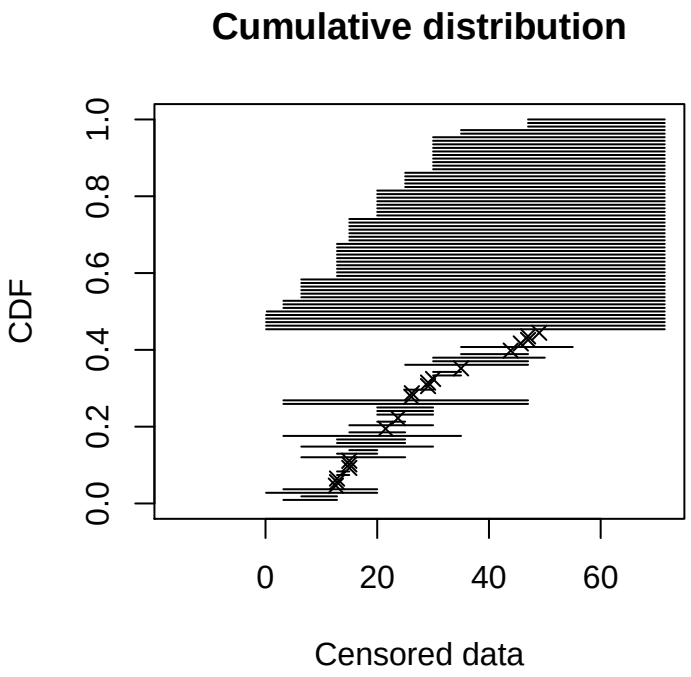
```
library(fitdistrplus)
data("salinity")
str(salinity)
```

```
## 'data.frame':    108 obs. of  2 variables:
##  $ left : num  20 20 20 20 20 21.5 15 20 23.7 25 ...
##  $ right: num  NA NA NA NA NA 21.5 30 25 23.7 NA ...
```

2 Graphical display of the observed distribution

Using censored data such as those coded in the `salinity` data set, the empirical distribution can be plotted using the `plotdistcens` function. By default, this function uses the Expectation-Maximization approach of (Turnbull {1974}) to compute the overall empirical cdf curve with optional confidence intervals, by calls to `survfit` and `plot.survfit` functions from the **survival** package (Figure~**??** shows the Turnbull plot of data together with two fitted distributions). A less rigorous but sometimes more illustrative plot can be obtained by fixing the argument `Turnbull` to `FALSE` in the call to `plotdistcens` (see Figure~**??** for an example and the help page of Function `plotdistcens` for details). This plot enables to see the real nature of censored data, as points and intervals.

```
plotdistcens(salinity,Turnbull = FALSE)
```



3 Maximum likelihood estimation

As for non censored data, one or more parametric distributions can be fitted to the censored data set, one at a time, but using in this case the 'fitdistcens} function. This function estimates the vector of distribution parameters θ by maximizing the likelihood for censored data defined as:

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

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`, `fitdistcens` returns the results of the fit of any parametric distribution to a data set as an S3 class object that can be easily printed, summarized or plotted. For the `salinity` data set, a lognormal distribution or a loglogistic can be fitted as commonly done in ecotoxicology for such data. As with `fitdist`,

for some distributions (see (Delignette-Muller et al. 2014) 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 a manual iterative use if necessary.

```
fsal.ln <- fitdistcens(salinity, "lnorm")
fsal.ll <- fitdistcens(salinity, "llogis",
  start = list(shape = 5, scale = 40))
summary(fsal.ln)
```

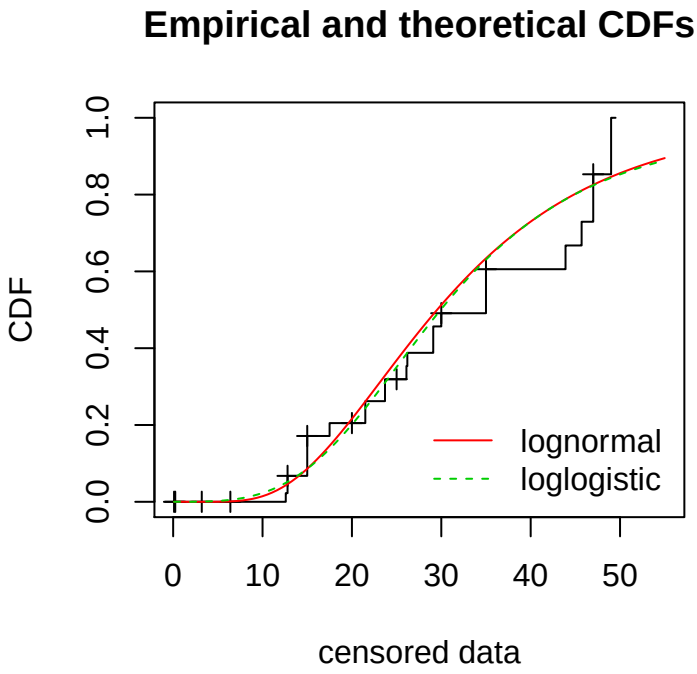
```
## Fitting of the distribution ' lnorm ' By maximum likelihood on censored data
## Parameters
##          estimate Std. Error
## meanlog 3.3854230 0.06486627
## sdlog   0.4961333 0.05455091
## Fixed parameters:
## data frame with 0 columns and 0 rows
## Loglikelihood:  -139.055   AIC:  282.1099   BIC:  287.4742
## Correlation matrix:
##           meanlog     sdlog
## meanlog 1.0000000 0.2938412
## sdlog   0.2938412 1.0000000
```

```
summary(fsal.ll)
```

```
## Fitting of the distribution ' llogis ' By maximum likelihood on censored data
## Parameters
##        estimate Std. Error
## shape  3.420545  0.4157984
## scale 29.930115  1.9447140
## Fixed parameters:
## data frame with 0 columns and 0 rows
## Loglikelihood:  -140.0717   AIC:  284.1433   BIC:  289.5076
## Correlation matrix:
##            shape      scale
## shape  1.0000000 -0.2021633
## scale -0.2021633  1.0000000
```

Computations of goodness-of-fit statistics have not yet been developed for fits using censored data but the quality of fit can be judged using Akaike and Schwarz's Bayesian information criteria (AIC and BIC) and the goodness-of-fit CDF plot, respectively provided when summarizing or plotting an object of class `"fitdistcens"`. Function `cdfcompcens` can also be used to compare the fit of various distributions to the same censored data set. Its call is similar to the one of `cdfcomp`. Below is an example of comparison of the two fitted distributions to the `salinity` data set (see Figure~**??**).

```
cdfcompcens(list(fsal.ln, fsal.ll), legendtext = c("lognormal", "loglogistic "))
```



Function `bootdistcens` is the equivalent of `bootdist` for censored data, except that it only proposes nonparametric bootstrap. Indeed, it is not obvious to simulate censoring within a parametric bootstrap resampling procedure. The generic function `quantile` can also be applied to an object of class `"fitdistcens"` or `"bootdistcens"`, as for continuous non-censored data.

References

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.

Kefford, B.J., E.J. Fields, C. Clay, and D. Nugegoda. 2007. "Salinity Tolerance of Riverine Macroinvertebrates from the Southern Murray-Darling Basin." *Marine and Freshwater Research* 58: 1019–31.

Klein, J.P., and M.L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data.* 2nd ed. Springer-Verlag.

Turnbull, B.W. {1974}. "Nonparametric Estimation of a Survivorship Function with Doubly Censored Data." *Journal of the American Statistical Association* 69 (345): 169–73.