Nouveaux graphes d'ajustement pour les données censurées dans le package **fitdistrplus**

M.L. Delignette-Muller [a], C. Dutang [b] and A. Siberchicot [c]

a, c Université de Lyon, Université Lyon 1, CNRS, VetAgro Sup, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France

a `marielaure.delignettemuller@vetagro-sup.fr` , c `aurelie.siberchicot@univ-lyon1.fr`

b Université Paris-Dauphine, PSL Research University, CNRS, CEREMADE, F-75016 Paris

`christophe.dutang@dauphine.fr`

Mots clefs : distribution, données censurées, graphes d'ajustement.

Le package **fitdistrplus** est un package dédié à l'ajustement de distributions univariées sur divers types de données : données continues éventuellement censurées et données discrètes [1]. Il propose en particulier diverses méthodes d'ajustement de distributions paramétriques aux données et des outils pour juger de la qualité d'un ajustement et comparer l'ajustement de plusieurs distributions candidates à un même jeu de données : graphes et statistiques d'ajustement. Dans le cadre de l'ajustement de distributions sur des données continues non censurées, quatre graphes d'ajustement sont proposés :

- un graphe en densité, comparant les courbes de densité de probabilité des lois ajustées à l'histogramme et/ou à l'estimateur à noyau de la densité empirique (fonction `denscomp`),
- un graphe en fonction de répartition, comparant les fonctions de répartition des lois ajustées à la fonction de répartition empirique (fonction `cdfcomp`),
- un graphe Quantile-Quantile (Q-Q plot) représentant les quantiles observés en fonction des quantiles des lois ajustées (fonction `qqcomp`),
- un graphe Probabilité-Probabilité (P-P plot) représentant la fonction de répartion empirique en chaque point observé en fonction des fonctions de répartition ajustées (fonction `ppcomp`)

Le package **fitdistrplus** permet de gérer tout type de censures (à gauche, à droite et par intervalle) avec des niveaux de censure variables d'un point observé à l'autre au sein d'un même jeu de données. Dans ce cadre aucune des représentations graphiques citées précédemment n'est triviale à réaliser et nous ne proposions jusque là qu'une représentation de la fonction de répartition empirique utilisant le package **survival** et son implémentation de l'algorithme de Turnbull [2] (Figure 1, représentation en haut à gauche).

Turnbull [2] fut le premier à proposer un algorithme d'estimation non paramétrique par maximum de vraisemblance (NPMLE) de la fonction de répartition empirique pour des données censurées de tout type. Néanmoins il a été montré depuis que l'algorithme proposé par Turnbull et implémentée dans le package **survival** pouvait dans certains cas converger vers une fonction de répartition ne correspondant pas au maximum de vraisemblance [3]. Depuis de nombreux algorithmes NPMLE ont été proposés et mis à disposition notamment dans les packages **Icens**, **interval** et **npsurv**. Leurs auteurs ont soit corrigé l'algorithme de Turnbull pour assurer sa bonne convergence, soit développé des algorithmes complètement différents. Ces algorithmes donnent des estimations généralement très proches mais pas toujours identiques et ont des performances très variables en terme de temps de calcul. Tous ont adopté une nouvelle présentation de la fonction de répartition cumulée faisant explicitement apparaître sous forme de rectangles coloriés les zones d'indétermination de la fonction de répartition empirique.

Nous avons choisi d'utiliser la fonction `npsurv` du package **npsurv** qui utilise des algorithmes performants récemment publiés [4-6]. Le graphe d'ajustement en fonction de répartition empirique proposé par défaut dans **fitdistrplus** utilise donc maintenant cette fonction, et nous avons pu à partir de celle-ci proposer assez naturellement des graphes de type QQ-plot et PP-plot (Figure 1).

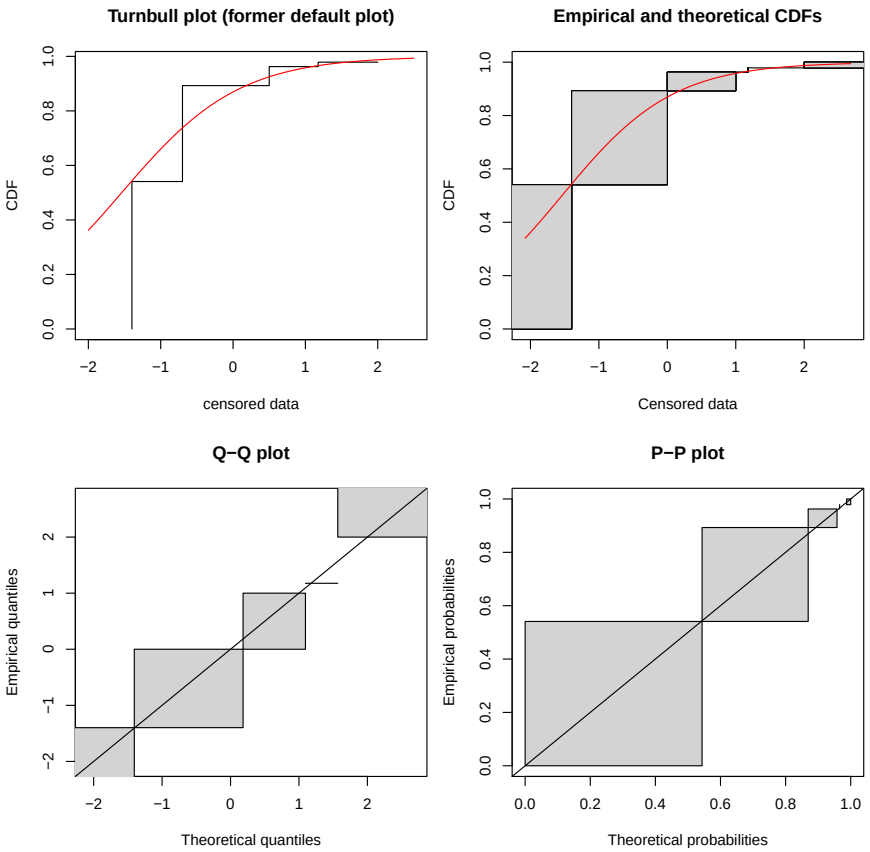

Figure 1 : l'ancien graphe d'ajustement (en haut à gauche) et les trois nouveaux graphes d'ajustement - exemple de l'ajustement d'un loi logistique (en rouge) sur des données de contamination de saumon fumé par *Listeria monocytogenes* (jeu de données `smokedfish` après transformation logarithmique).

Parallèlement le package **fitdistrplus** s'est aussi enrichi d'une nouvelle représentation du graphe d'ajustement d'une distribution sur données discrètes en densité de probabilité et d'une option d'utilisation du package **ggplot2** pour chaque graphe d'ajustement.

Références

[1] Delignette-Muller, M.L., Dutang, C. (2015). fitdistrplus: an R package for fitting distributions. Journal of Statistical Software, 64(4), 1-34.

[2] Turnbull BW (1974). Nonparametric estimation of a survivorship function with doubly censored data. Journal of American Statistical Association, 69, 169-173.

[3] Gentleman, R., & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. Biometrika, 81(3), 618-623.

[4] Wang, Y. (2008). Dimension-reduced nonparametric maximum likelihood computation for interval-censored data. Computational Statistics & Data Analysis, 52(5), 2388-2402.

[5] Wang, Y., & Taylor, S. M. (2013). Efficient computation of nonparametric survival functions via a hierarchical mixture formulation. Statistics and Computing, 23(6), 713-725.

[6] Wang, Y., & Fani, S. (2018). Nonparametric maximum likelihood computation of a U-shaped hazard function. Statistics and Computing, 28(1), 187-200.