CompLognormal: An R Package for Composite Lognormal Distributions

by S. Nadarajah and S. A. A. Bakar

Abstract In recent years, composite models based on the lognormal distribution have become popular in actuarial sciences and related areas. In this short note, we present a new R package for computing the probability density function, cumulative density function, and quantile function, and for generating random numbers of any composite model based on the lognormal distribution. The use of the package is illustrated using a real data set.

Introduction

Two-piece composite distributions arise in many areas of the sciences. The first two-piece composite distribution with each piece described by a normal distribution appears to have been used by Gibbons and Mylroie (1973). Recently, two-piece composite distributions with the first piece described a lognormal distribution (which we refer to as composite lognormal distributions) have proved popular in insurance and related areas. Cooray and Ananda (2005) introduced such distributions recently, but one of the referees has pointed out that composite lognormal distributions have been known before at least as early as Barford and Crovella (1998) in areas like modeling workloads of web servers.

Cooray and Ananda (2005) showed that composite lognormal distributions can give better fits than standard univariate distributions. They illustrate this fact for the Danish fire insurance data by introducing the *composite lognormal-Pareto distribution*, a distribution obtained by piecing together lognormal and Pareto probability density functions (pdfs). Scollnik (2007) improved the composite lognormal-Pareto distribution by using mixing weights as coefficients for each pdf replacing the constant weights applied earlier by Cooray and Ananda (2005). Scollnik (2007) also employed generalized Pareto distribution in place of Pareto distribution used by Cooray and Ananda (2005). Pigeon and Denuit (2011) provided further extensions of composite lognormal distributions by choosing the cutoff point (the point at which the two pdfs are pieced together) as a random variable. The most recent extensions of composite lognormal distributions are provided in Nadarajah and Bakar (2012).

Composite lognormal distributions have attracted considerable attention in spite of being introduced only in 2005. Some applications in the last three years include: estimation of insurance claim cost distributions (Bolancé et al., 2010), inflation and excess insurance (Fackler, 2010), large insurance claims in case reserves (Lindblad, 2011), statistical mechanics (Eliazar and Cohen, 2012), and modeling of mixed traffic conditions (Dubey et al., 2013).

The aim of this short note is to present a new contributed package **CompLognormal** for R that computes basic properties for *any* composite lognormal distribution. The properties considered include the pdf, cumulative distribution function (cdf), quantile function, and random numbers. Explicit expressions for these properties are given in Section "Properties". Two illustrations of the practical use of the new R package are given in Section "Illustrations".

Properties

Let $f_i(\cdot)$, $i = 1, 2$ be valid pdfs. Let $F_i(\cdot)$, $i = 1, 2$ denote the corresponding cdfs. In general, the pdf of a two-piece composite distribution is given by

$$f(x) = \begin{cases} a_1 f_1^*(x), & \text{if} \quad 0 < x \leq \theta, \\ a_2 f_2^*(x), & \text{if} \quad \theta < x < \infty, \end{cases} \tag{1}$$

where $f_1^*(x) = f_1(x)/F_1(\theta)$, $f_2^*(x) = f_2(x)/\{1 - F_2(\theta)\}$, θ is the cutoff point, and a_1, a_2 are non-negative weights summing to one. As suggested in Nadarajah and Bakar (2012), we take $a_1 = \frac{1}{1+\phi}$ and $a_2 = \frac{\phi}{1+\phi}$ for $\phi > 0$.

The cdf corresponding to (1) is

$$F(x) = \begin{cases} \dfrac{1}{1+\phi} \dfrac{F_1(x)}{F_1(\theta)}, & \text{if} \quad 0 < x \leq \theta, \\ \dfrac{1}{1+\phi} \left[1 + \phi \dfrac{F_2(x) - F_2(\theta)}{1 - F_2(\theta)}\right], & \text{if} \quad \theta < x < \infty. \end{cases} \tag{2}$$

The quantile function corresponding to (1) is

$$Q(u) = \begin{cases} F_1^{-1}\left(u(1+\phi)F_1(\theta)\right), & \text{if} \quad 0 < u \leq \frac{1}{1+\phi}, \\ F_2^{-1}\left(F_2(\theta) + (1-F_2(\theta))\left(\frac{u(1+\phi)-1}{\phi}\right)\right), & \text{if} \quad \frac{1}{1+\phi} < u < \infty. \end{cases} \tag{3}$$

This quantile function can be used to generate random numbers from the two-piece composite distribution.

We are interested in a two-piece composite distribution, where the first piece is specified by the lognormal distribution. So, we take

$$f_1(x) = \frac{1}{x\sigma}\psi\left(\frac{\ln x - \mu}{\sigma}\right) \qquad \text{and} \qquad F_1(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right),$$

where $\psi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and the standard normal cdf, respectively. For this choice, (1), (2), and (3) reduce to

$$f(x) = \begin{cases} \frac{\psi\left((\ln x - \mu)/\sigma\right)}{(1+\phi)\sigma x \Phi\left((\ln\theta - \mu)/\sigma\right)}, & \text{if} \quad 0 < x \leq \theta, \\ \frac{\phi f_2(x)}{(1+\phi)\left(1-F_2(\theta)\right)}, & \text{if} \quad \theta < x < \infty, \end{cases} \tag{4}$$

$$F(x) = \begin{cases} \frac{\Phi\left((\ln x - \mu)/\sigma\right)}{(1+\phi)\Phi\left((\ln\theta - \mu)/\sigma\right)}, & \text{if} \quad 0 < x \leq \theta, \\ \frac{1}{1+\phi}\left[1 + \phi\frac{F_2(x) - F_2(\theta)}{1-F_2(\theta)}\right], & \text{if} \quad \theta < x < \infty, \end{cases} \tag{5}$$

and

$$Q(u) = \begin{cases} \exp\left\{\mu + \sigma\Phi^{-1}\left[u(1+\phi)\Phi\left(\frac{\ln\theta - \mu}{\sigma}\right)\right]\right\}, & \text{if} \quad 0 < u \leq \frac{1}{1+\phi}, \\ F_2^{-1}\left(F_2(\theta) + (1-F_2(\theta))\left(\frac{u(1+\phi)-1}{\phi}\right)\right), & \text{if} \quad \frac{1}{1+\phi} < u < \infty, \end{cases} \tag{6}$$

respectively. We shall refer to the distribution given by (4), (5), and (6) as the *composite lognormal distribution*. Random numbers from the composite lognormal distribution can be generated as

$$x_i = Q\left(u_i\right) \tag{7}$$

for $i = 1, 2, \ldots, n$, where u_i, $i = 1, 2, \ldots, n$ are random numbers from a uniform $[0, 1]$ distribution. Further statistical properties of the composite lognormal distribution including moment properties can be found in Bakar (2012).

The pdf of two-piece composite distributions are in general not continuous or differentiable at the cutoff point θ. To have these properties satisfied, we impose the conditions $a_1 f_1^*(\theta) = a_2 f_2^*(\theta)$ and $a_1 d f_1^*(\theta)/d\theta = a_2 d f_2^*(\theta)/d\theta$. Nadarajah and Bakar (2012) have shown that these conditions are equivalent to the following for any composite lognormal distribution:

$$\mu = \ln\theta + \sigma^2 + \theta\sigma^2\frac{f_2'(\theta)}{f_2(\theta)}, \qquad \text{and} \qquad \phi = \frac{f_1(\theta)\left[1 - F_2(\theta)\right]}{f_2(\theta)F_1(\theta)}. \tag{8}$$

The smoothness conditions in (8) are imposed for technical reasons—the respective parametric models then become more tractable.

Maximum likelihood estimation is a common method for estimation. Suppose we have a random sample $x_1, x_2, \ldots, x_n$ from (1). Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_q)$ be the parameters of $f_2(\cdot)$. Suppose also that μ and ϕ can be expressed as $\mu = \mu(\sigma, \theta, \boldsymbol{\lambda})$ and $\phi = \phi(\sigma, \theta, \boldsymbol{\lambda})$, respectively. Then the log-likelihood function is

$$\begin{aligned} \ln L\left(\sigma, \theta, \boldsymbol{\lambda}\right) = &-n\ln(1+\phi) + \sum_{x_i \leq \theta} \ln\psi\left(\frac{\ln x_i - \mu}{\sigma}\right) - \sum_{x_i \leq \theta} \ln\left(\sigma x_i\right) \\ &- M\ln\Phi\left(\frac{\ln\theta - \mu}{\sigma}\right) + \sum_{x_i > \theta} \ln f_2\left(x_i\right) \\ &- m\ln\left[1 - F_2(\theta)\right] + m\ln\phi, \end{aligned} \tag{9}$$

where $M = \sum_{i=1}^{n} I\{x_i \leq \theta\}$ and $m = \sum_{i=1}^{n} I\{x_i > \theta\}$. It is clear that the maximum likelihood estimators of $(\sigma, \theta, \boldsymbol{\lambda})$ cannot be obtained in closed form. They have to be obtained numerically.

The new package **CompLognormal**, available from CRAN, computes (4), (5), and (6) for any given

Quantity	Calling sequence
$f(x)$ in (4)	`dcomplnorm(x,spec,sigma=1,theta=1,...)`
$F(x)$ in (5)	`pcomplnorm(x,spec,sigma=1,theta=1,...)`
$Q(u)$ in (6)	`qcomplnorm(p,spec,sigma=1,theta=1,...)`
x_i in (7)	`rcomplnorm(n,spec,sigma=1,theta=1,...)`

Table 1: Quantity and calling sequence for the composite lognormal distribution.

composite lognormal distribution. It also generates random numbers from the specified composite lognormal distribution. Table 1 summarizes the functions implemented in the package **CompLognormal** along with their arguments.

The input argument `spec` (a character string) specifies the distribution of the second piece of the composite lognormal distribution. The distribution should be one that is recognized by R. It could be one of the distributions implemented in the R base package or one of the distributions implemented in an R contributed package or one freshly written by a user. In any case, there should be functions `dspec`, `pspec`, `qspec` and `rspec`, computing the pdf, cdf, qf and random numbers of the distribution.

Some examples of `spec` are: `spec = "norm"`, meaning that $f_2(x) = (1/\sigma)\psi((x-\mu)/\sigma)$ and $F_2(x) = \Phi((x-\mu)/\sigma)$; `spec = "lnorm"`, meaning that $f_2(x) = \{1/(\sigma x)\}\,\psi((\ln x - \mu)/\sigma)$ and $F_2(x) = \Phi((\ln x - \mu)/\sigma)$; `spec = "exp"`, meaning that $f_2(x) = \lambda\exp(-\lambda x)$ and $F_2(x) = 1 - \exp(-\lambda x)$.

`dcomplnorm`, `pcomplnorm`, `qcomplnorm` and `rcomplnorm` can also take additional arguments in the form of `...`. These arguments could give inputs (for example, parameter values) for the distribution of the second piece of the composite lognormal distribution. For example, if `spec = "norm"` then `...` can include `mean = 1,sd = 1` to mean that $f_2(x) = \psi(x-1)$ and $F_2(x) = \Phi(x-1)$; if `spec = "lnorm"` then `...` can include `meanlog = 1,sdlog = 1` to mean that $f_2(x) = \psi(\ln x - 1)$ and $F_2(x) = \Phi(\ln x - 1)$; if `spec = "exp"` then `...` can include `rate = 1` to mean that $f_2(x) = \lambda\exp(-x)$ and $F_2(x) = 1 - \exp(-x)$. Further details about the calling sequences can be seen from the documentation for the **CompLognormal** package.

The code for `dcomplnorm`, `pcomplnorm` and `qcomplnorm` currently uses constructs of the form

```
do.call(paste("d", spec, sep = ""), list(z}, \code{...))
```

This technique was necessary as in the R base package distributions are no data type but rather are given by the respective four constituent functions `<prefix><name>`, where `<prefix>` is one of `r`, `d`, `p`, `q` and `<name>` is the name of the distribution. A way to circumvent such constructions would be a data type "distribution". This has been available on CRAN for quite a while within the `distr` family of packages, with corresponding S4 classes for distributions. Recently, another approach based on S5-classes has been pursued in the package **poweRlaw** (Gillespie, 2013). We leave these as future work.

Illustrations

We describe two simple illustrations of the new package **CompLognormal**. The following packages should be loaded (after installing them if necessary) in advance:

```
library(CompLognormal)
library(actuar)
library(SMPracticals)
library(evd)
library(fitdistrplus)
library(stats4)
```

Illustration 1

The illustration presented here plots the pdf and the cdf of the composite lognormal-loglogistic distribution for varying parameter values.

```
curve(dcomplnorm(x, "llogis", 0.4, 0.5, shape = 1, scale = 0.8), xlim = c(0, 5),
      ylim = c(0, 0.7), xlab = "x", ylab = "f(x)", n = 250, col = "black", lty = 1)
d1 <- dcomplnorm(0.5, "llogis", 0.4, 0.5, shape = 1, scale = 0.8)
segments(0.5, 0, 0.5, d1, col = "black", lty = 2)
```

```
curve(dcomplnorm(x, "llogis", 0.3, 0.2, shape = 0.2, scale = 0.5), add = TRUE,
      col = "red", lty = 1)
d2 <- dcomplnorm(0.2, "llogis", 0.3, 0.2, shape = 0.2, scale = 0.5)
segments(0.2, 0, 0.2, d2, col = "red", lty = 2)
curve(dcomplnorm(x, "llogis", 0.5, 0.4, shape = 0.5, scale = 0.5), add = TRUE,
      col = "blue", lty = 1)
d3 <- dcomplnorm(0.4, "llogis", 0.5, 0.4, shape = 0.5, scale = 0.5)
segments(0.4, 0, 0.4, d3, col = "blue", lty = 2)
legend(1.5, 0.5, legend =
    c(expression(paste(sigma==0.4, ",", theta==0.5, ",", shape==1, ",", scale==0.8)),
      expression(paste(sigma==0.3, ",", theta==0.2, ",", shape==0.2, ",", scale==0.5)),
      expression(paste(sigma==0.5, ",", theta==0.4, ",", shape==0.5, ",", scale==0.8))),
      col = c("black", "red", "blue"), lty = 1)

curve(pcomplnorm(x, "llogis", 0.4, 0.5, shape = 1, scale = 0.8), xlim = c(0, 5),
      ylim = c(0, 1), xlab =  "x", ylab = "F(x)", n = 250, col = "black", lty = 1)
d1 <- pcomplnorm(0.5, "llogis", 0.4, 0.5, shape = 1, scale = 0.8)
segments(0.5, 0, 0.5, d1, col = "black", lty = 2)
curve(pcomplnorm(x, "llogis", 0.3, 0.2, shape = 0.2, scale = 0.5), add = TRUE,
      col = "red", lty = 1)
d2 <- pcomplnorm(0.2, "llogis", 0.3, 0.2, shape = 0.2, scale = 0.5)
segments(0.2, 0, 0.2, d2, col = "red", lty = 2)
curve(pcomplnorm(x, "llogis", 0.5, 0.4, shape = 0.5, scale = 0.5), add = TRUE,
      col = "blue", lty = 1)
d3 <- pcomplnorm(0.4, "llogis", 0.5, 0.4, shape = 0.5, scale = 0.5)
segments(0.4, 0, 0.4, d3, col = "blue", lty = 2)
```

Figure 1 shows the pdfs and cdfs of the composite lognormal-loglogistic distribution. The parameters, σ and θ, the scale parameter of the lognormal distribution and the cutoff point, respectively, are as defined in Section "Properties". The parameters, `shape` and `scale`, are the shape and scale parameters of the loglogistic distribution, as defined in the R base package. The vertical lines in the figures correspond to the values for θ, the cutoff point.

Illustration 2

The illustration presented here fits the composite lognormal-Fréchet distribution to the Danish fire insurance data by the method of maximum likelihood, see (9). The data were obtained from the R package **SMPracticals** (Davison, 2012).

```
x <- danish[1:2492]
nlm(function(p) {
        -sum(dcomplnorm(x, "frechet", sigma = exp(p[1]),
             theta = exp(p[2]), scale = exp(p[3]), shape = exp(p[4]), log = TRUE))
    }, p = c(0, 0, 0, 0))
```

The output will be

```
$minimum
[1] 3859.293

$estimate
[1] -1.7178497  0.1176224 -0.2872701  0.4130180

$gradient
[1] -0.0008301586  0.0031436684  0.0001900844 -0.0004574758

$code
[1] 1

$iterations
[1] 23
```

The number 2492 refers to the actual length of the Danish data set. This output shows that the maximized log-likelihood is -3859.293, the estimated σ is $\exp(-1.7178497) = 0.1794516$, the estimated θ is $\exp(0.1176224) = 1.124819$, the estimated scale parameter of the Fréchet distribution

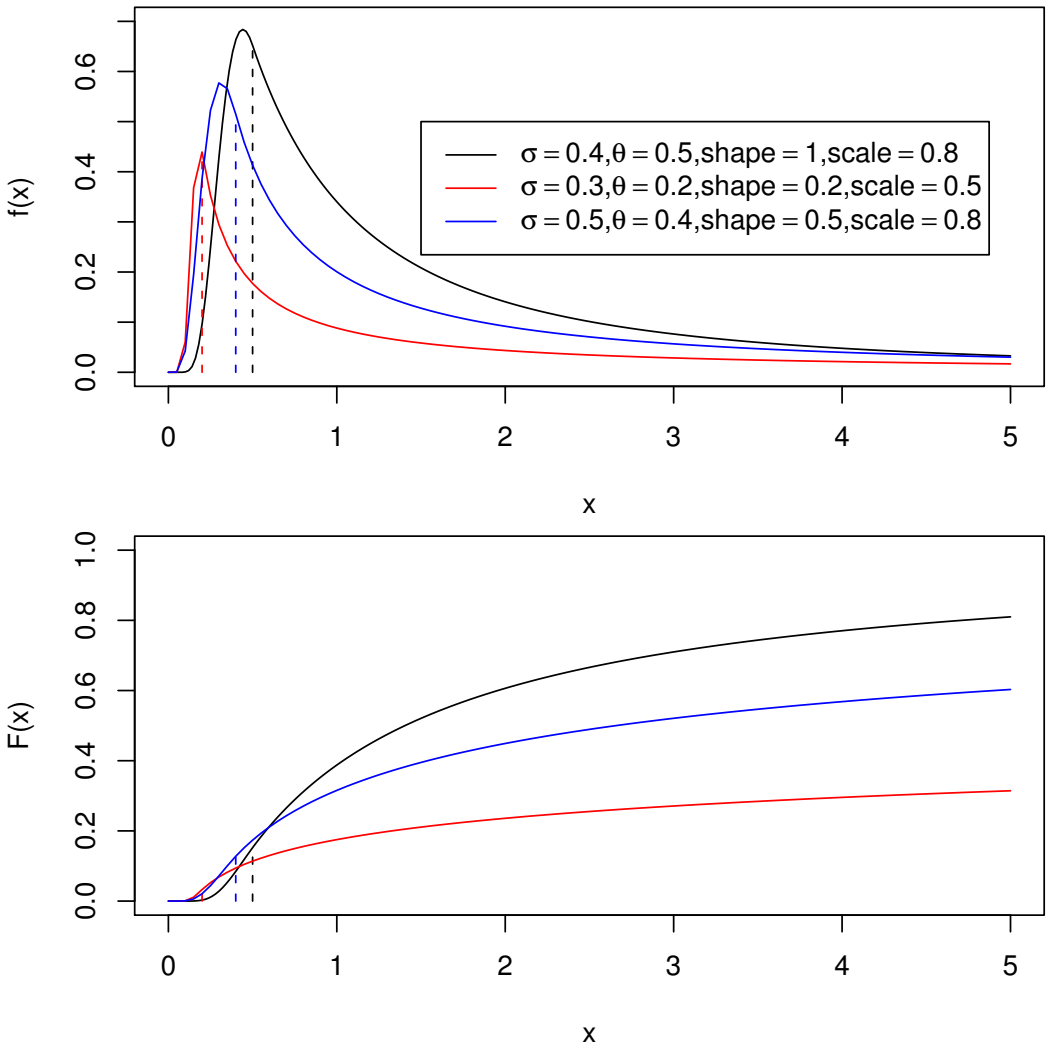

Figure 1: Pdfs (top) and cdfs (bottom) of the composite lognormal-loglogistic distribution. The top legend applies to both plots.

is $\exp(-0.2872701) = 0.750309$, and the estimated shape parameter of the Fréchet distribution is $\exp(0.4130180) = 1.511372$. The standard errors of these estimates can be computed by adding `hessian = TRUE` to the `nlm` command.

We have chosen the composite lognormal-Fréchet distribution for simplicity of illustration. The purpose of this illustration was not to find the "best fitting" model for the Danish data. Of course, many other composite lognormal distributions can be expected to give better fits than the composite lognormal-Fréchet distribution. Nadarajah and Bakar (2012) modeled the Danish data using a large class of composite lognormal distributions, including the composite lognormal-Burr, the composite lognormal-inverse Burr, the composite lognormal-F, the composite lognormal-Fréchet, the composite lognormal-generalized Pareto, the composite lognormal-inverse Pareto, and the composite lognormal-loglogistic distributions. The composite lognormal-Burr distribution was shown to give the best fit.

In Illustration 2, we used `nlm` for optimization of the likelihood function. R has many other model fitting routines like `fitdistr` from the package **MASS** (Venables and Ripley, 2002), `fitdist` from the package **fitdistrplus** (Delignette-Muller et al., 2013), `mle` from the package **stats4**, and `MLEstimator`, `MDEstimator` from the package **distrMod** (Kohl and Ruckdeschel, 2013). For instance, `fitdist` can be used to estimate the parameters of the composite lognormal-Fréchet distribution as follows:

```
dclnormf <- function(x, logsigma, logtheta, logscale, logshape) {
    dcomplnorm(x, spec = "frechet", sigma = exp(logsigma), theta = exp(logtheta),
               scale = exp(logscale), shape = exp(logshape))
}
pclnormf <- function(q, logsigma, logtheta, logscale, logshape) {
```

```
    pcomplnorm(q, spec = "frechet", sigma = exp(logsigma), theta = exp(logtheta),
               scale = exp(logscale), shape = exp(logshape))
}
qclnormf <- function(p, logsigma, logtheta, logscale, logshape) {
    qcomplnorm(p, spec = "frechet", sigma = exp(logsigma), theta = exp(logtheta),
               scale = exp(logscale), shape = exp(logshape))
}
fitdist(danish[1:2492], "clnormf", start = list(logsigma = -1.718,
        logtheta = 0.118, logscale = -0.287, logshape = 0.413))
```

The output will be

```
Fitting of the distribution ' clnormf ' by maximum likelihood
Parameters:
           estimate Std. Error
logsigma -1.7180280 0.05737326
logtheta  0.1176365 0.02413085
logscale -0.2877565 0.16564834
logshape  0.4130318 0.03704017
```

Also `mle` can be used to estimate the parameters of the composite lognormal-Fréchet distribution as follows:

```
nllh <- function(p1, p2, p3, p4) {
    -sum(dcomplnorm(danish[1:2492], spec = "frechet",
         sigma = exp(p1), theta = exp(p2), scale = exp(p3), shape = exp(p4), log = TRUE))
}
mle(nllh, start = list(p1 = -1.718, p2 = 0.118, p3 = -0.287, p4 = 0.413))
```

The output will be

```
Call:
mle(minuslogl = nllh, start = list(p1 = -1.718, p2 = 0.118, p3 = -0.287,
    p4 = 0.413))

Coefficients:
        p1         p2         p3         p4
-1.7178735  0.1176081 -0.2870383  0.4130586
```

Conclusions

We have developed a new R package for computing quantities of interest for *any* composite lognormal distribution. The computed quantities include the pdf, cdf, qf, and random numbers. Although the package is specifically designed for composite lognormal distributions, it can be easily altered for any other composite distribution.

Acknowledgments

The authors would like to thank the Editor and the two referees for careful reading and for their comments which greatly improved the paper.

Bibliography

S. Bakar. *Some Contributions to Actuarial Parametric Modeling*. PhD thesis, University of Manchester, UK, 2012. [p98]

P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. *ACM SIGMETRICS Performance Evaluation Review*, 26:151–160, 1998. [p97]

C. Bolancé, M. Guillén, and J. Nielsen. Transformation kernel estimation of insurance claim cost distributions. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 43–51, 2010. [p97]

K. Cooray and M. Ananda. Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal*, pages 321–334, 2005. [p97]

A. Davison. *SMPracticals: Practicals for Use with Davison (2003) Statistical Models*, 2012. URL http://CRAN.R-project.org/package=SMPracticals. R package version 1.4-1. [p100]

M. Delignette-Muller, R. Pouillot, J.-B. Denis, and C. Dutang. *fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*, 2013. URL http://cran.r-project.org/web/packages/fitdistrplus/fitdistrplus.pdf. [p101]

S. Dubey, B. Ponnu, and S. Arkatkar. Time gap modeling using mixture distributions under mixed traffic conditions. *Journal of Transportation Systems Engineering and Information Technology*, 13:91–98, 2013. [p97]

I. Eliazar and M. Cohen. A Langevin approach to the log-Gauss-Pareto composite statistical structure. *Physica A: Statistical Mechanics and Its Applications*, 391:5598–5610, 2012. [p97]

M. Fackler. Inflation and excess insurance, 2010. ASTIN / AFIR Colloquium, Madrid. [p97]

J. Gibbons and S. Mylroie. Estimation of impurity profiles in ionimplanted amorphous targets using joined half-Gaussian distributions. *Applied Physics Letters*, 22:568–569, 1973. [p97]

C. Gillespie. *Fitting Heavy Tailed Distributions: The poweRlaw Package*, 2013. URL http://cran.r-project.org/web/packages/poweRlaw/index.html. [p99]

M. Kohl and P. Ruckdeschel. *distrMod: Object Oriented Implementation of Probability Models*, 2013. URL http://cran.r-project.org/web/packages/distrMod/index.html. [p101]

K.-S. Lindblad. How big is large? A study of the limit for large insurance claims in case reserves. Master's thesis, KTH/Matematisk statistik, 2011. [p97]

S. Nadarajah and S. Bakar. New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal*, 2012. DOI: 10.1080/03461238.2012.695748. [p97, 98, 101]

M. Pigeon and M. Denuit. Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal*, pages 177–192, 2011. [p97]

D. Scollnik. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal*, pages 20–33, 2007. [p97]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0. [p101]

S. Nadarajah
School of Mathematics
University of Manchester
Manchester M13 9PL, UK mbbsssn2@manchester.ac.uk

S. A. A. Bakar
Institute of Mathematical Sciences
University of Malaya
50603 Kuala Lumpur, Malaysia saab@um.edu.my