

Two new defective distributions based on the Marshall–Olkin extension

Ricardo Rocha¹ · Saralees Nadarajah² ·
Vera Tomazella¹ · Francisco Louzada³

Received: 14 June 2014 / Accepted: 18 April 2015 / Published online: 8 May 2015
© Springer Science+Business Media New York 2015

Abstract The presence of immune elements (generating a fraction of cure) in survival data is common. These cases are usually modeled by the standard mixture model. Here, we use an alternative approach based on defective distributions. Defective distributions are characterized by having density functions that integrate to values less than 1, when the domain of their parameters is different from the usual one. We use the Marshall–Olkin class of distributions to generalize two existing defective distributions, therefore generating two new defective distributions. We illustrate the distributions using three real data sets.

Keywords Cure fraction · Defective models · Gompertz distribution · Inverse Gaussian distribution · Marshall–Olkin family · Survival analysis

1 Introduction

In survival data, the goal is to model the time until the occurrence of an event of interest. In the general theory, it is assumed that an element is always susceptible to a certain event of interest that will eventually occur. However, such an assumption is often invalid because many databases have what we call immune elements. For immune elements, the event will never occur because they are not susceptible to the event of interest.

✉ Ricardo Rocha
ricardorocha23@hotmail.com

¹ Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brasil

² School of Mathematics, University of Manchester, Manchester, UK

³ Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil

Thus, the study of the proportion of this population, or rather, the cure fraction, became one of the major subfields of survival analysis. To solve such problems (Berkson and Gage 1952), based on the work of (Boag 1949), a standard mixture model was proposed. The survival function is adjusted to

$$S(t) = p + (1 - p)S_0(t)$$

such that $S_0(t)$ is a proper survival function. In this way, $S(t)$ converges to p as the time goes by. Berkson and Gage (1952) analyzed a stomach cancer data set using the standard mixture model. Since then many other data sets have been analyzed using the standard mixture model. The most common choices for $S_0(t)$ have been the Weibull, log-logistic and lognormal distributions. Recently, different distributions have been proposed for this purpose, see Yakovlev and Tsodikov (1996), Chen et al. (1999) and Ibrahim et al. (2005).

However, here we use an alternative methodology based on defective models, a concept introduced by Balka et al. (2009). In these models it is possible to estimate a cure rate with the use of a naturally improper distribution. Instead of estimating the proportion p directly as a mixture model, we use a distribution by changing the domain of its parameters. It becomes a model with long-term duration. The integral of the density function does not result in 1, but in a value $p \in (0, 1)$. This is because the parameters of the distribution are no longer in their original domains. The cumulative distribution function no longer approaches to 1, but to p and, therefore, the survival function approaches to $1 - p$. Models that have this characteristic are called defective.

Obviously, the defective distribution is not proper. When used as a model for cure fraction, the proportion of the population that is immune is obtained by calculating the limit of the survival function with the estimated parameters. In the literature, there are two known distributions that can be used for this purpose: the inverse Gaussian and Gompertz distributions. The Gompertz distribution has two parameters, both positive. For negative values of the shape parameter, the distribution becomes defective. Haybittle (1959) fitted the Gompertz distribution to a breast cancer data set. Cantor and Shuster (1992) fitted a modified version of the same distribution to a pediatric cancer data set. Gieser et al. (1998) extended the distribution to include covariates.

The inverse Gaussian distribution is another distribution which can take a defective form. The inverse Gaussian distribution was derived by Schrödinger (1915) by calculating the probability of the first time passage of a one-dimensional Brownian process (also called Wiener process). A deeper study was made by Tweedie (1945). It is also where the name inverse Gaussian arose (Balka et al. 2009).

Application of survival models in terms of stochastic processes can be found in Whitmore et al. (1998), Lee and Whitmore (2004) and Lee and Whitmore (2006). Aalen and Gjessing (2001) suggested that more attention be given to these type of models. Whitmore (1979) used the inverse Gaussian distribution in its defective form. More concise studies from the point of view of survival analysis were presented in Balka et al. (2009) and Balka et al. (2011). Several scenarios involving the inverse Gaussian distribution as a cure fraction model were presented. Balka et al. (2011) presented real data applications with classical and Bayesian inferences.

The great advantage of the defective models compared to the standard mixture model is that the former have one less parameter to be estimated in the likelihood function, which can be quite relevant in terms of estimation, especially when the chosen models already have various parameters.

The aim of this paper is to propose two new defective distributions based on the Marshall–Olkin family of distributions (Marshall and Olkin 1997). This family is obtained by adding an extra parameter to a known distribution. Suppose $S(t)$ is a known survival function. Then, the extended survival function by the Marshall–Olkin family, $S^*(t)$, is

$$S^*(t) = \frac{rS(t)}{1 - (1 - r)S(t)}$$

for $r > 0$ and $t \in \mathbb{R}$. Simple algebraic manipulations determine the density function of the extended distribution:

$$f^*(t) = \frac{rf(t)}{[1 - (1 - r)S(t)]^2}. \quad (1)$$

The Marshall–Olkin extension for particular choices of $S(t)$ has been investigated by several authors, see Alice and Jose (2003), Alice and Jose (2005), Jose and Krishna (2011) and Cordeiro and Lemonte (2013). Barreto-Souza et al. (2013) investigated mathematical properties of the Marshall–Olkin family.

The main purpose of this paper is to propose two new defective distributions, extending the Gompertz and inverse Gaussian distributions through the Marshall–Olkin family. The details of these extensions including maximum likelihood estimation and the fact that S^* is defective if S is defective are shown in Sect. 2. Section 3 is a simulation study to assess the performance of the maximum likelihood estimators. Section 4 illustrates the proposed distributions using three real data sets.

2 Methodology

The Gompertz distribution is used for modeling survival data in various areas of knowledge (Gieser et al. 1998), especially where there is a suspicion of exponential hazard. The Gompertz density function is

$$f(t) = be^{at} e^{-\frac{b}{a}(e^{at}-1)} \quad (2)$$

for $a > 0$, $b > 0$ and $t > 0$. In this parameterization, a is the shape parameter and b is the location parameter. The survival function is

$$S(t) = e^{-\frac{b}{a}(e^{at}-1)}. \quad (3)$$

The defective Gompertz distribution is the distribution that allows for negative values for the parameter a . The proportion of immunity in the population is calculated as the limit of the survival function when $a < 0$:

$$p = \lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}} \in (0, 1).$$

Once the parameter values are estimated, one can easily compute the fraction of cure p . Figure 1 illustrates various scenarios for the density, survival and hazard functions of the Gompertz distribution.

The inverse Gaussian distribution arises as the first passage time of a Wiener process (Balka et al. 2009). Lee and Whitmore (2006) noted its potential as models for cure rate. Its density function is

$$f(t) = \frac{1}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{1}{2bt} (1 - at)^2 \right\} \quad (4)$$

for $a > 0, b > 0$ and $t > 0$. The inverse Gaussian distribution has survival function given by

$$S(t) = 1 - \left[\Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right], \quad (5)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable.

The inverse Gaussian distribution can be defective when $a < 0$. The fraction of cure, or the survival function limit, is

$$\begin{aligned} p &= \lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} 1 - \left[\Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right] \\ &= 1 - e^{2a/b} \in (0, 1). \end{aligned}$$

We estimate the cure fraction using the estimated parameters a and b . Figure 2 illustrates various scenarios for the density, survival and hazard functions of the inverse Gaussian distribution.

The great advantage of these distributions is that the proportion of cured is always estimated using a model with one parameter less than the standard mixture model, which brings plenty of benefits in terms of estimation.

We have been able to find only these two distributions (Gompertz and inverse Gaussian) that can be adapted to being defective. This does not mean there are not others. In order to construct other distributions of this kind, we propose the use of the Marshall–Olkin class (Marshall and Olkin 1997) to generalize a given distribution by adding an extra parameter.

The main result of this paper is that if a given distribution is defective, then its extension under the Marshall–Olkin family will be defective as well.

Theorem 1 *If $S(t)$ is defective then $S^*(t)$ is also defective.*

Proof Suppose the limit of $S(t)$ is equal to $p_0 \in (0, 1)$. Then

$$\lim_{t \rightarrow \infty} S^*(t) = \lim_{t \rightarrow \infty} \frac{rS(t)}{1 - (1 - r)S(t)} = \frac{rp_0}{1 - (1 - r)p_0} = \frac{rp_0}{rp_0 + 1 - p_0}. \quad (6)$$

Since $1 - p_0$ is positive, it is easy to see that the last expression in (6) takes a value in $(0, 1)$. The proof is complete. \square

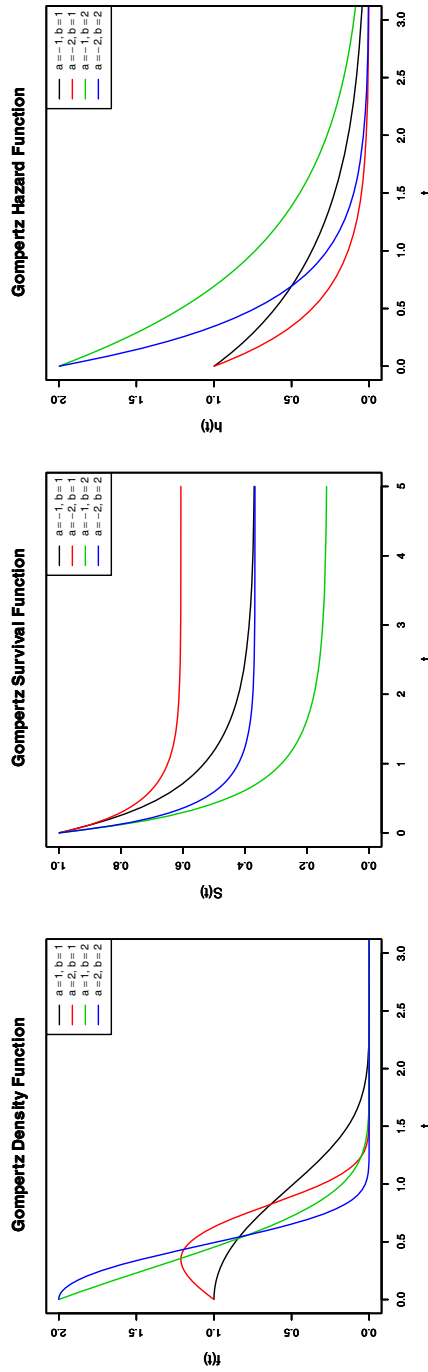


Fig. 1 Density, survival and hazard functions of the defective Gompertz distribution

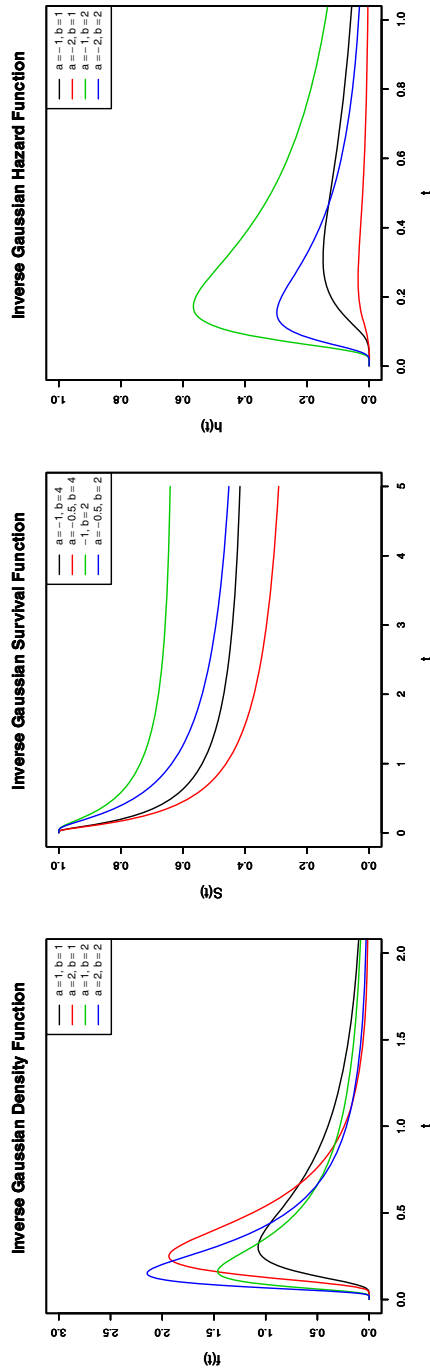


Fig. 2 Density, survival and hazard functions of the defective inverse Gaussian distribution

We propose now two new defective distributions: the Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions.

2.1 The Marshall–Olkin Gompertz distribution

Using (1) with density function in (2) and survival function in (3), we obtain the Marshall–Olkin Gompertz density function

$$f(t) = \frac{b \cdot r \cdot \exp\left(\frac{b - b \exp(at)}{a} + at\right)}{\left[r - (r - 1) \exp\left(\frac{b - b \exp(at)}{a}\right)\right]^2} \tag{7}$$

for $a > 0, b > 0, r > 0$ and $t > 0$. The corresponding survival function is

$$S(t) = \frac{r \exp\left[-\frac{b}{a} (\exp(at) - 1)\right]}{1 - (1 - r) \exp\left[-\frac{b}{a} (\exp(at) - 1)\right]}. \tag{8}$$

Figure 3 illustrates various scenarios for the density, survival and hazard functions of the Marshall–Olkin Gompertz distribution. As in the Gompertz distribution, if $a < 0$ then the Marshall–Olkin Gompertz distribution is defective. Its cure fraction is

$$\lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} 1 - \frac{1}{r e^{\frac{b(e^{at}-1)}{a}} - r + 1} = \frac{rp_0}{1 - (1 - r)p_0} = \frac{rp_0}{rp_0 + 1 - p_0} = p,$$

where p_0 is the cure fraction of the defective Gompertz distribution.

2.2 The Marshall–Olkin inverse Gaussian distribution

Using (1) with density and survival functions of the inverse Gaussian distribution given by (4) and (5), respectively, we obtain the density function of the Marshall–Olkin inverse Gaussian distribution as

$$f(t) = \frac{r \exp\left(-\frac{(at - 1)^2}{2bt}\right)}{\sqrt{2\pi} \sqrt{bt}^3 \left[(r - 1) \Phi\left(\frac{at - 1}{\sqrt{bt}}\right) + (r - 1) e^{\frac{2a}{b}} \Phi\left(-\frac{at + 1}{\sqrt{bt}}\right) - r \right]^2} \tag{9}$$

for $a > 0, b > 0$ and $r > 0$. The corresponding survival function is

$$S(t) = \frac{r \left[1 - \Phi\left(\frac{-1 + at}{\sqrt{bt}}\right) - e^{2a/b} \Phi\left(\frac{-1 - at}{\sqrt{bt}}\right) \right]}{1 - (1 - r) \left[1 - \Phi\left(\frac{-1 + at}{\sqrt{bt}}\right) - e^{2a/b} \Phi\left(\frac{-1 - at}{\sqrt{bt}}\right) \right]}. \tag{10}$$

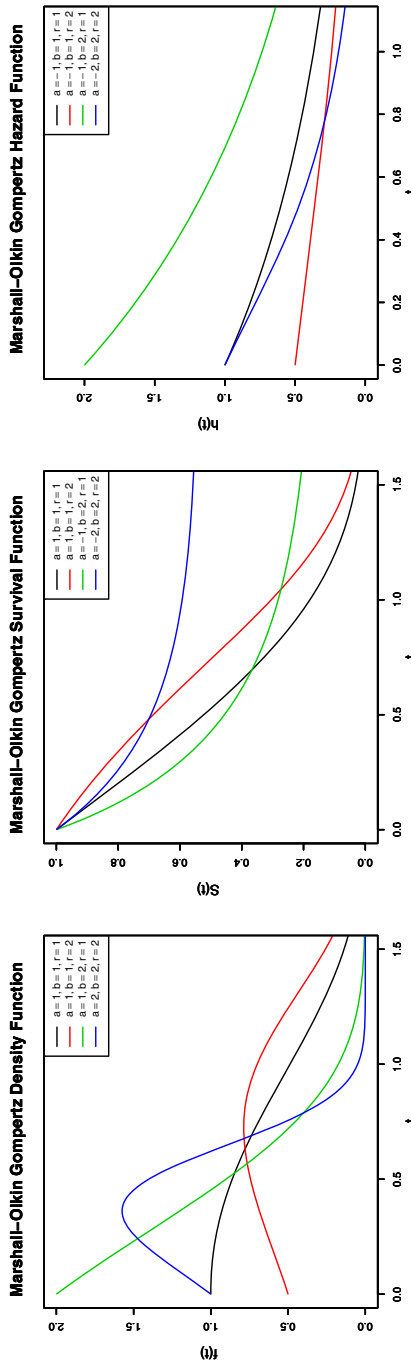


Fig. 3 Density, survival and hazard functions of the defective Marshall-Olkin Gompertz distribution

Figure 4 illustrates various scenarios for the density, survival and hazard functions of the Marshall–Olkin inverse Gaussian distribution. As in the inverse Gaussian distribution, if $a < 0$ then the Marshall–Olkin inverse Gaussian distribution is also defective. Its cure fraction is

$$\lim_{t \rightarrow \infty} S(t) = \frac{rp_0}{rp_0 + 1 - p_0} = p,$$

where p_0 is the cure fraction of the defective inverse Gaussian distribution.

2.3 Inference

Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)^T$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ are the censored failure times. The δ_i is equal to 1 if a failure is observed and 0 otherwise.

Suppose that the data are independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T$ denotes a vector of parameters. The likelihood function of $\boldsymbol{\theta}$ can be written as (see Klein and Moeschberger 2003)

$$L(\boldsymbol{\theta}, \mathbf{D}) \propto \prod_{i=1}^n \left[f(t_i, \boldsymbol{\theta})^{\delta_i} S(t_i, \boldsymbol{\theta})^{1-\delta_i} \right].$$

The corresponding log-likelihood function is

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^n (1 - \delta_i) \log S(t_i, \boldsymbol{\theta}).$$

For the Marshall–Olkin Gompertz distribution given by (7) and (8),

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} &+ \sum_{i=1}^n \delta_i \left[\log \left(br \exp \left(\frac{b - b \exp(at)}{a} + at \right) \right) \right] \\ &- \sum_{i=1}^n \delta_i \left[\log \left(\left[r - (r - 1) \exp \left(\frac{b - b \exp(at)}{a} \right) \right]^2 \right) \right] \\ &+ \sum_{i=1}^n (1 - \delta_i) \left[\log \left(r \exp \left[-\frac{b}{a} (\exp(at) - 1) \right] \right) \right] \\ &- \sum_{i=1}^n (1 - \delta_i) \left[\log \left(1 - (1 - r) \exp \left[-\frac{b}{a} (\exp(at) - 1) \right] \right) \right]. \end{aligned} \tag{11}$$

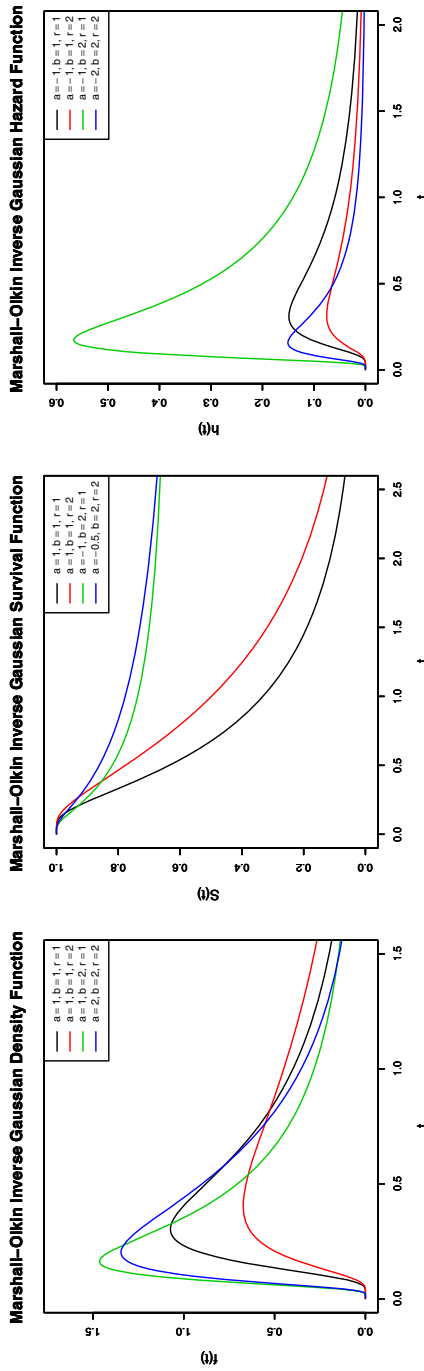


Fig. 4 Density, survival and hazard functions of the defective Marshall–Olkin inverse Gaussian distribution

For the Marshall–Olkin inverse Gaussian distribution given by (9) and (10),

$$\begin{aligned}
 \log L(\boldsymbol{\theta}, \mathbf{D}) = & \text{const} + \sum_{i=1}^n \delta_i \log \left(r \exp \left(-\frac{(at-1)^2}{2bt} \right) \right) \\
 & - \sum_{i=1}^n \delta_i \log \left(\sqrt{bt}^3 \left[(r-1) \Phi \left(\frac{at-1}{\sqrt{bt}} \right) \right. \right. \\
 & \left. \left. + (r-1) e^{\frac{2a}{b}} \Phi \left(-\frac{at+1}{\sqrt{bt}} \right) - r \right]^2 \right) \\
 & + \sum_{i=1}^n (1-\delta_i) \log \left(r \left[\Phi \left(\frac{at-1}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(-\frac{at+1}{\sqrt{bt}} \right) - 1 \right] \right) \\
 & - \sum_{i=1}^n (1-\delta_i) \log \left(-(r-1) \Phi \left(\frac{at-1}{\sqrt{bt}} \right) \right) \\
 & + (r-1) e^{\frac{2a}{b}} \left(\Phi \left(\frac{at+1}{\sqrt{bt}} \right) + 1 \right) - 1. \tag{12}
 \end{aligned}$$

The log likelihoods, (11) and (12), can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization. We used the routine `optim` in the R software (R Core Team 2014). Numerical calculations not reported here showed that the surfaces of (11) and (12) were smooth. The routine `optim` was able to locate the maximum for a wide range of starting values. The solution for the maximum was unique for all starting values. In the simulations and real data applications presented in Sects. 3 and 4, the routine `optim` converged all the time, giving unique maximum likelihood estimates. In all cases considered, `optim` did not take more than five seconds for convergence.

Confidence intervals for the parameters were based on asymptotic normality. If $\widehat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\theta}$ then it is well known that the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ can be approximated by a q -variate normal distribution (where q denotes the length of the vector $\boldsymbol{\theta}$ as defined above) with zero means and covariance matrix $\mathbf{I}(\widehat{\boldsymbol{\theta}})$, where $\mathbf{I}(\boldsymbol{\theta})$ denotes the observed information matrix defined by

$$\mathbf{I}(\boldsymbol{\theta}) = - \begin{pmatrix} \frac{\partial^2 \log L}{\partial \theta_1^2} & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_q} \\ \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_2^2} & \dots & \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_q} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L}{\partial \theta_q \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_q \partial \theta_2} & \dots & \frac{\partial^2 \log L}{\partial \theta_q^2} \end{pmatrix}.$$

So, an approximate $100(1 - \alpha)$ percent confidence interval for θ_i is $(\widehat{\theta}_i - z_{\alpha/2} \sqrt{I^{ii}}, \widehat{\theta}_i + z_{\alpha/2} \sqrt{I^{ii}})$, where I^{ii} denotes the i th diagonal element of the

inverse of \mathbf{I} and z_a denotes the $100(1 - a)$ percentile of a standard normal random variable.

We have supposed the usual asymptotes of the maximum likelihood estimates hold. However, defective distributions like the mixture model are not proper distributions. The checking of regularity conditions for the asymptotes by analytical means is not easy. Such conditions have not been checked even for the standard mixture model. We suggest analytical checking of the regularity conditions as a possible future work.

In the next section, we perform an extensive simulation study partly to check the asymptotes of the maximum likelihood estimates. Simulations have been used in many papers to assess the behavior of maximum likelihood estimates, especially when an analytical investigation is intractable.

3 Simulation study

Here, we perform three simulation experiments. The first one is to assess the performance of the maximum likelihood estimates with respect to sample size. The second one is a comparison of defective and mixture models in terms of AIC and cure rate estimates when the data were generated from a defective model. The third one is the same as the second one, but the data were generated from a mixture model. All computations were performed in R Core Team (2014).

Consider that the time of occurrence of an event of interest has cumulative distribution function $F(t)$. Suppose we want to simulate a random sample of size n containing real times, censored times and a cure fraction of p . An algorithm to generate data from the defective model is:

1. Determine the desired parameter values, as well as the value of the cure fraction p ;
2. Generate $M_i \sim \text{Bernoulli}(1 - p)$;
3. If $M_i = 0$ set $t'_i = \infty$. If $M_i = 1$ take t'_i as the root of $F(t) = u$, where $u \sim \text{uniform}(0, 1 - p)$;
4. Generate $u'_i \sim \text{uniform}(0, \max(t_i))$, considering only the finite t_i ;
5. Calculate $t_i = \min(t'_i, u'_i)$. If $t_i < u'_i$ set $\delta_i = 1$, otherwise set $\delta_i = 0$.

In this first experiment, we simulated one thousand random samples each of size $n = 20, 40, \dots, 1000$. The random samples were taken to come from (i) the Marshall–Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$; (ii) the Marshall–Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$. We computed the maximum likelihood estimates, $\hat{a}, \hat{b}, \hat{r}$ and \hat{p} , and their standard errors for each sample. These were used to compute the bias, the mean squared error, the coverage probability and the coverage length of $\hat{a}, \hat{b}, \hat{r}$ and \hat{p} for each n .

Figures 5 and 6 show the plots of the mean squared errors, the biases, the coverage probabilities and the coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions.

We can observe the following from the figures: (i) the mean squared errors for all parameters generally decrease to zero with increasing n ; (ii) the mean squared errors for all parameters appear reasonably close to zero for all $n \geq 600$; (iii) the mean squared errors appear smallest for the parameter, p ; (iv) the mean squared

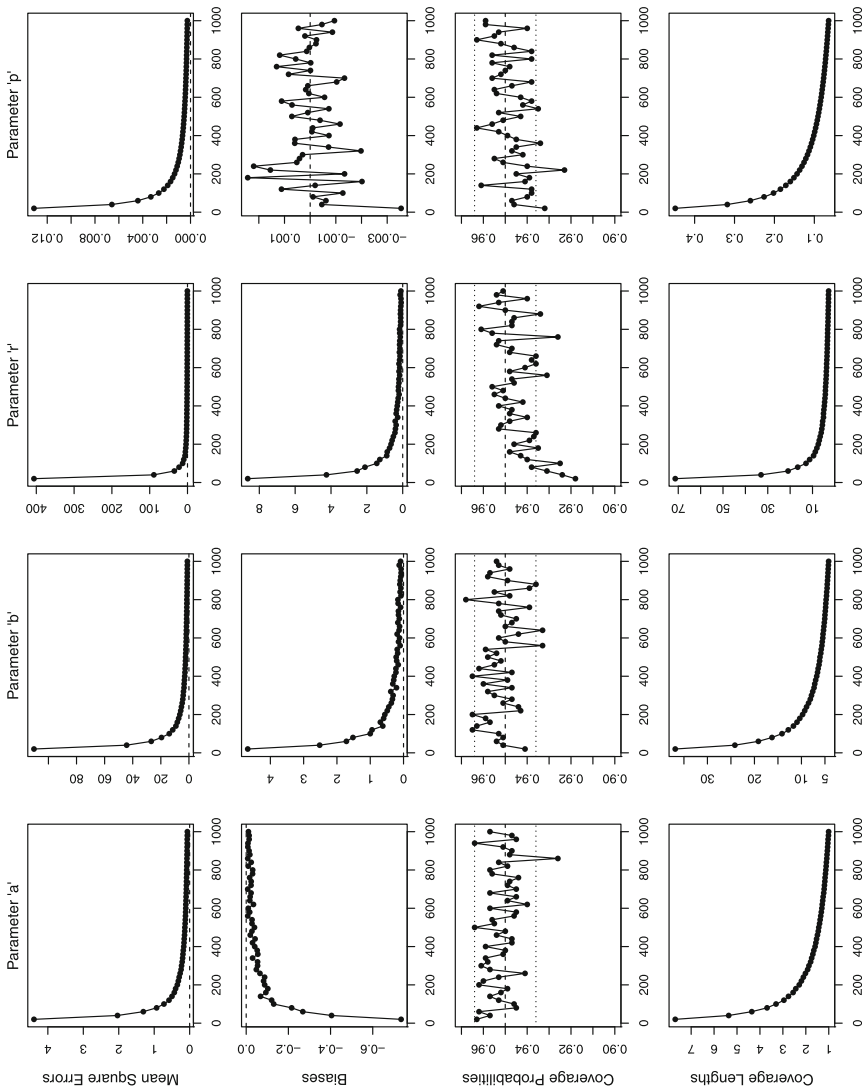


Fig. 5 Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall–Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$

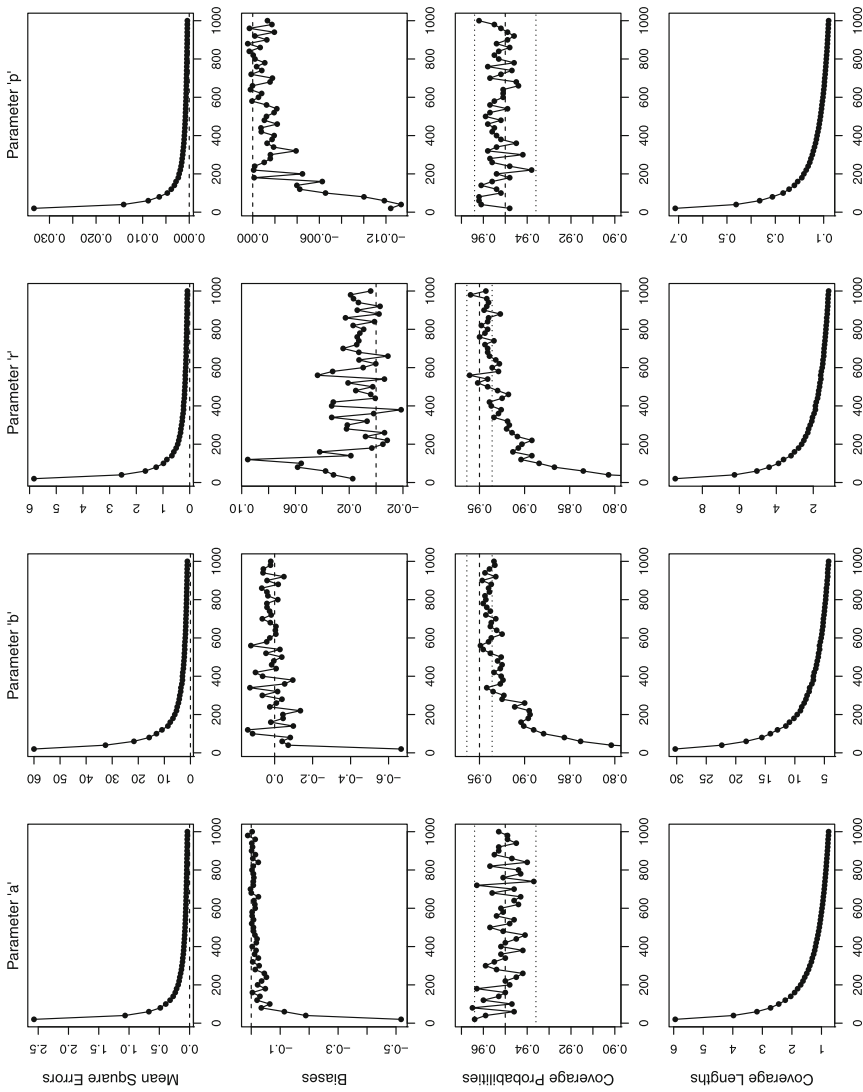


Fig. 6 Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{\gamma})$ versus n for simulated data from the Marshall–Olkin inverse Gaussian distribution with $(a, b, r, \gamma) = (-2, 10, 2, 0.4958)$

errors appear largest for the parameters, b and r ; (v) the biases for all parameters generally approach zero with increasing n ; (vi) the biases for all parameters appear reasonably close to zero for all $n \geq 600$; (vii) the biases appear generally negative for the parameter, a ; (viii) the biases appear generally positive for the parameter, r ; (ix) the biases appear smallest for the parameter, p ; (x) the coverage probabilities for all parameters generally approach the nominal level with increasing n ; (xi) the coverage probabilities for all parameters appear reasonably close to the nominal level for all $n \geq 800$; (xii) the coverage probabilities appear furthest from the nominal level for the parameter, r ; (xiii) the coverage lengths for all parameters generally decrease with increasing n ; (xiv) the coverage lengths appear smallest for the parameter, p ; (xv) the coverage lengths appear largest for the parameters, b and r .

These observations are for the Marshall–Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$ and for the Marshall–Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$. But many of the observations were the same when the simulations were repeated for a wide range other values of (a, b, r, p) for both the Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions.

We also noted that the decrease in coverage lengths with increasing n was slow. Indeed, some of the coverage lengths in Figs. 5 and 6 do appear large even for a sample of size 200. Some of the confidence intervals reported in Sect. 4 appear large too. This suggests a very large sample size may be needed in order to have reliable interval estimates. It is comforting however two of the three real data sets considered in Sect. 4 have sizes over one thousand.

The second experiment is to compare the performance of the defective models versus their respective mixture models when the data were generated from defective models. The Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian defective distributions were simulated using $(a, b, r, p) = (-3, 4, 2, 0.4172)$ and $(a, b, r, p) = (-2, 10, 2, 0.4958)$, respectively. They were compared to the corresponding mixture versions. Figures 7 and 8 (left) provide a comparison in terms of

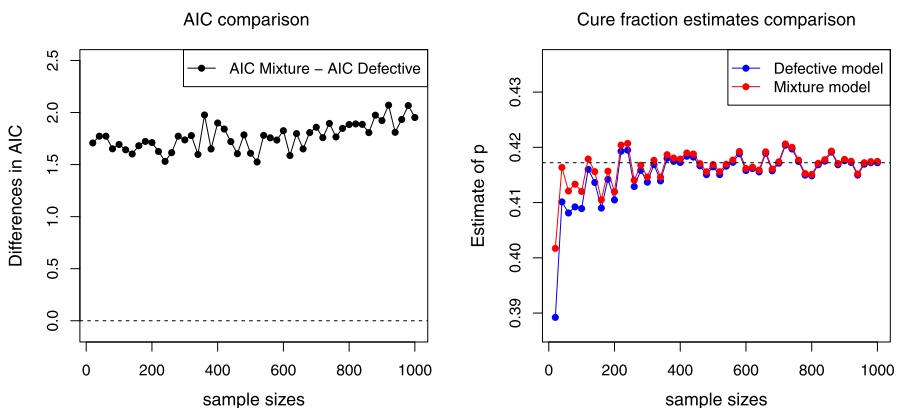


Fig. 7 In the left, the *plotted line* represents the difference between the AIC values obtained under the Marshall–Olkin Gompertz mixture and defective models, when the data were generated from a defective model. In the right, the corresponding estimates of p

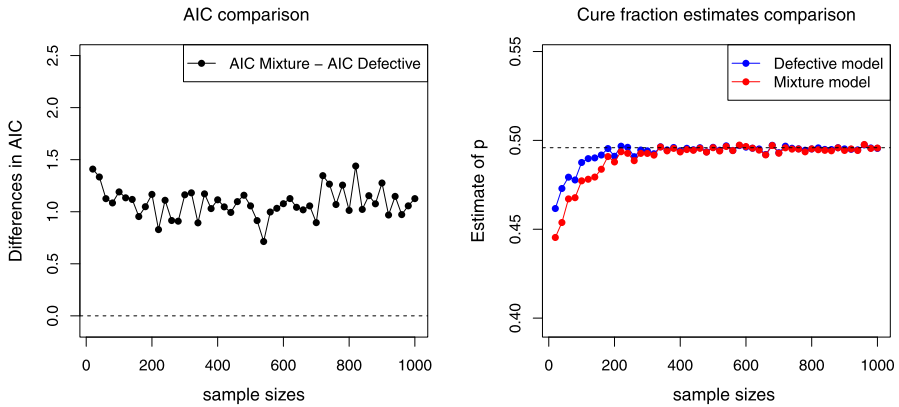


Fig. 8 In the left, the *plotted line* represents the difference between the AIC values obtained under the Marshall–Olkin inverse Gaussian mixture and defective models, when the data were generated from a defective model. In the right, the corresponding estimates of p

the AIC. The black line represents the difference between the AIC of the mixture model and that of the defective model. The difference is positive for all samples sizes, meaning that the AIC of the defective model is always smaller. On average, the AIC of the defective model is 1.7704 smaller than the AIC of the mixture model for the Marshall–Olkin Gompertz distribution. On average, the AIC of the defective model is 1.0865 smaller for the Marshall–Olkin inverse Gaussian distribution.

Figures 7 and 8 (right) compare the cure rate estimates for mixture and defective models. We have not compared other parameters since they do not directly relate to the proposed distributions. The estimates of p under both models appear good for the Marshall–Olkin Gompertz distribution, see Fig. 7. The quadratic error sum for the defective model is 0.00130 and that for the mixture model is 0.00049. This gives a slight advantage for the mixture model. The estimates of p under the defective and mixture models appear good also for the Marshall–Olkin inverse Gaussian distribution, see Fig. 8. The quadratic error sum for the defective model is 0.00256 and that for the mixture model is 0.00727. Again a small difference but now in favour of the defective model.

The third and the last experiment is to compare the performance of the defective models versus their respective mixture models when the data were generated from mixture models. Mixture versions of the Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions were simulated using $(a, b, r, p) = (0.2, 0.2, 0.2, 0.5)$ and $(a, b, r, p) = (2, 2, 0.5, 0.5)$, respectively. They were compared to the corresponding defective versions. Figures 9 and 10 (left) compare the models in terms of the AIC. The black line again represents the difference between the AIC of the mixture model and that of the defective model. The differences decrease as n increases for the Marshall–Olkin Gompertz distribution and become less than zero only when $n > 960$, see Fig. 9. The differences appear positive for all sample sizes for the Marshall–Olkin inverse Gaussian distribution, see Fig. 10. On average, the AIC of the defective model is 0.8196 smaller than the AIC of the mixture model for the Marshall–Olkin Gompertz distribution. On average, the AIC of the defective model is 0.7511 smaller for the Marshall–Olkin inverse Gaussian distribution.

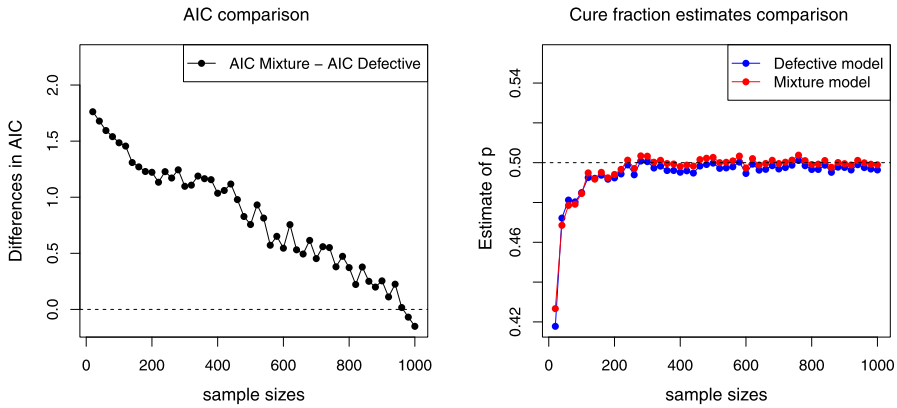


Fig. 9 In the left, the *plotted line* represents the difference between the AIC values obtained under the Marshall–Olkin Gompertz mixture and defective models, when the data were generated from a mixture model. In the right, the corresponding estimates of p

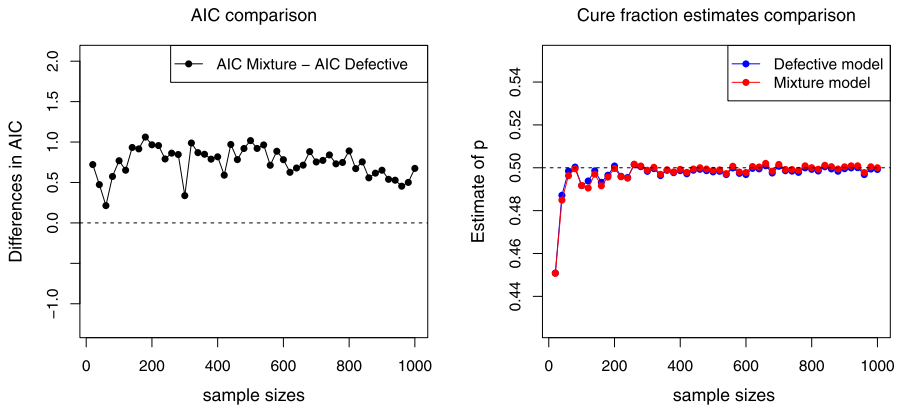


Fig. 10 In the left, the *plotted line* represents the difference between the AIC values obtained under the Marshall–Olkin inverse Gaussian mixture and defective models, when the data were generated from a mixture model. In the right, the corresponding estimates of p

Figures 9 and 10 (right) compare the cure rate estimates for mixture and defective models. The estimates of p under both models appear good for both Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions. The quadratic error sums for the defective and mixture models are 0.00715 and 0.00782, respectively, for the Marshall–Olkin Gompertz distribution. The quadratic error sums for the defective and mixture models are 0.00288 and 0.00302, respectively, for the Marshall–Olkin inverse Gaussian distribution.

The differences found in the second and third experiments are small, but they show clearly that the defective model is better. The results remained the same for a wide range of other parameter choices. That is, the AIC values and the quadratic error sums were smaller for the defective model most of the time for a wide range of parameter choices and for the two distributions. Hence, the defective model can be considered a viable alternative for the mixture model.

Section 4 presents three real data applications. The sample size for the first data set is forty four. The sample size for the second data set is over one thousand. The sample size for the third data set is over one thousand eight hundred. Hence, the given point as well as interval estimates for the second and third data sets can be considered accurate enough. But those for the first data set must be treated conservatively.

4 Applications

To illustrate the distributions presented we are going to use three data sets. The first one relates to a study of recurrence of leukemia in patients who were submitted to a certain kind of transplantation. Leukemia is a type of cancer that affects the white blood cells produced by the bone marrow and can take several forms. The data set has forty four observations with 20.45 percent censoring (nine in total). The maximum observation time was approximately five years. For details of this data set, see [Kersey et al. \(1987\)](#).

The second data set relates to the time of birth of a second child for a couple and is based on medical records of births in Norway in 1997. The observed time is the gap between the birth of the first child and the birth of the second child for the same couple. The data set consists of 53543 women who had their first child between 1983 and 1997. The censoring indicates whether the woman had a second child, the event of interest, or if she did not before the end of the study. The data set was previously analyzed by [Aalen et al. \(2008\)](#). For illustrative purposes, we took a random sample accounting for 2 percent of the data set, totalling 1071 observations with 69.74 percent censoring (747 in total).

The third data set arises from one of the first successful trials of adjuvant chemotherapy for colon cancer. The event of interest here is the recurrence or death for the individual under the proposed treatment. The data set has 1858 observations and 50.58 percent censoring (938 in total). The data set is available in R Core Team (2014) in the survival package. Details of this data set can be found in [Laurie et al. \(1989\)](#).

The three data sets represent three different real scenarios (see the Kaplan–Meier curves later). They were chosen carefully to test the flexibility of the proposed distributions under different conditions. The first and third data sets are about the recurrence of a type of cancer. For these data sets, it is fair to assume that there are individuals who will never have the cancer again, implying a cure rate. For the second data set, the presence of a cure rate is even more obvious: the immune elements are simply those couples who do not plan to have a second child.

The Gompertz, the Marshall–Olkin Gompertz, the inverse Gaussian and the Marshall–Olkin inverse Gaussian distributions were fitted to the data set via maximum likelihood. The variance of the cure fraction was estimated by using the delta method. Software packages in the R Core Team (2014) environment were used for optimization of functions of interest and other computations. The summary of the fitted Gompertz and Marshall–Olkin Gompertz distributions is shown in Tables 1, 2 and 3. The summary of the fitted inverse Gaussian and Marshall–Olkin inverse Gaussian distributions is shown in Tables 4, 5 and 6.

The fitted survival curves of the proposed distributions for the leukemia data set are shown in Fig. 11. Those for the birth data set are shown in Fig. 12. Those for the colon

Table 1 MLEs for the fits of the Gompertz and Marshall–Olkin Gompertz distributions for the leukemia data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Gompertz	<i>a</i>	−1.5103	0.3696	−2.4399	−0.9349
	<i>b</i>	2.3767	0.5171	1.5517	3.6405
	<i>p</i>	0.2073	0.0611	0.0875	0.3271
Marshall–Olkin Gompertz	<i>a</i>	−4.0973	0.7898	−5.9783	−2.8082
	<i>b</i>	25.6059	9.1558	12.7051	51.6061
	<i>r</i>	121.9638	150.2286	10.9085	1363.6302
	<i>p</i>	0.191	0.0593	0.0748	0.3071

Table 2 MLEs for the fits of the Gompertz and Marshall–Olkin Gompertz distributions for the birth data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Gompertz	<i>a</i>	2.4401	0.3178	1.8172	3.063
	<i>b</i>	1.0025	0.0865	0.8329	1.172
	<i>p</i>	–	–	–	–
Marshall–Olkin Gompertz	<i>a</i>	−8.6164	0.6121	−9.9036	−7.4965
	<i>b</i>	84.5282	11.5758	64.6298	110.5529
	<i>r</i>	9449.995	7210.3028	2118.2085	42159.4024
	<i>p</i>	0.3416	0.0145	0.3132	0.37

Table 3 MLEs for the fits of the Gompertz and Marshall–Olkin Gompertz distributions for the colon data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Gompertz	<i>a</i>	−2.3372	0.1772	−2.7117	−2.0145
	<i>b</i>	2.0014	0.1025	1.8103	2.2127
	<i>p</i>	0.4247	0.0115	0.4022	0.4472
Marshall–Olkin Gompertz	<i>a</i>	−4.6989	0.3527	−5.4436	−4.0560
	<i>b</i>	11.1570	1.7778	8.1642	15.2469
	<i>r</i>	8.7515	2.2139	5.3304	14.3685
	<i>p</i>	0.4732	0.0116	0.4505	0.4959

data set are shown in Fig. 13. Table 7 presents the AIC values for all four of the fitted distributions. Figure 14 plots the Kaplan–Meier estimates of the survival function versus the predicted values from the proposed distributions. There is a diagonal line in each plot. The closer the points to this line the better the fit.

The Marshall–Olkin Gompertz distribution is a clear improvement over the Gompertz distribution for all three data sets. The fitted survival curve for the former captures the Kaplan–Meier curve much better, see Figs. 11, 12, 13 and 14. For all data sets, the

Table 4 MLEs for the fit of the Marshall–Olkin inverse Gaussian distribution for the leukemia data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Inverse Gaussian	a	-0.0003	0.0141	-0.0279	0.0273
	b	3.3612	0.7169	2.2128	5.1057
	p	0.0002	0.0021	0.0000	0.0044
Marshall–Olkin inverse Gaussian	a	-1.3387	0.4147	-2.4567	-0.7294
	b	1.0507	0.2182	0.6993	1.5786
	r	0.0226	0.0247	0.0027	0.1918
	p	0.2107	0.0615	0.0902	0.3312

Table 5 MLEs for the fit of the Marshall–Olkin inverse Gaussian distribution for the birth data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Inverse Gaussian	a	2.1169	0.1277	1.8666	2.3673
	b	1.5312	0.101	1.3332	1.7293
	p	–	–	–	–
Marshall–Olkin inverse Gaussian	a	-1.5842	0.6094	-3.3668	-0.7454
	b	1.063	0.09	0.9004	1.2549
	r	0.0161	0.017	0.002	0.1274
	p	0.2318	0.0129	0.2065	0.2571

Table 6 MLEs for the fit of the Marshall–Olkin inverse Gaussian distribution for the colon data set

Distribution	Parameter	Point estimate	SD	Low 95 % CI	Upper 95 % CI
Inverse Gaussian	a	-1.6688	0.1568	-2.0063	-1.3881
	b	7.3406	0.2901	6.7936	7.9317
	p	0.3653	0.0112	0.3435	0.3872
Marshall–Olkin inverse Gaussian	a	-0.0012	0.0160	-0.0326	0.0302
	b	12.3160	1.0183	10.4734	14.4827
	r	2.8375	0.2190	2.4392	3.3009
	p	0.0005	0.0005	-0.0005	0.0016

Marshall–Olkin Gompertz distribution estimates a by a negative value with a negative confidence interval. The Gompertz distribution gives a negative interval for a for the leukemia and colon data sets but estimates a by a positive value for the birth data set. So, the birth data set is an example, where the baseline distribution does not yield a defective model, while the Marshall–Olkin extension gives a much better fit as a defective model. All but the Marshall–Olkin inverse Gaussian distribution appear to estimate the cure fraction in the expected range in relation to the Kaplan–Meier curve. The Marshall–Olkin inverse Gaussian distribution appears to underestimate the cure fraction for the birth and colon data sets.

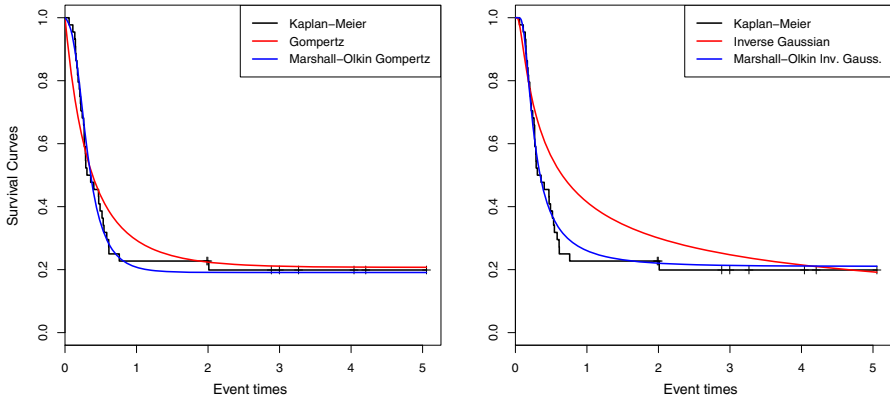


Fig. 11 Survival curves for the fitted Gompertz, Marshall–Olkin Gompertz, inverse Gaussian and Marshall–Olkin inverse Gaussian distributions for the leukemia data set

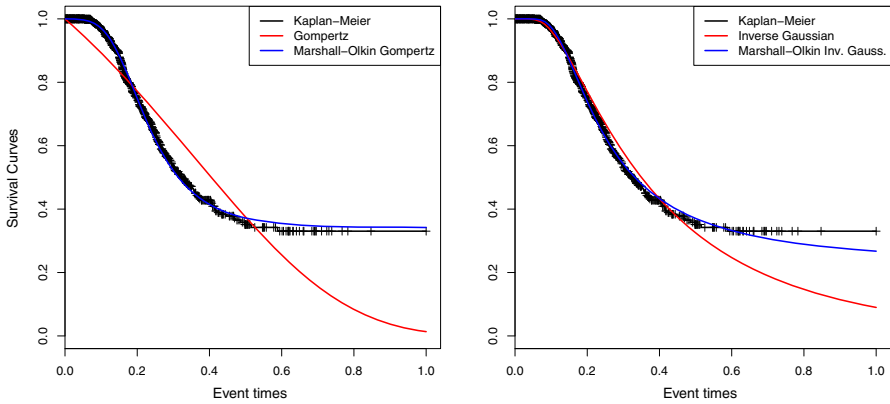


Fig. 12 Survival curves for the fitted Gompertz, Marshall–Olkin Gompertz, inverse Gaussian and Marshall–Olkin inverse Gaussian distributions for the birth data set

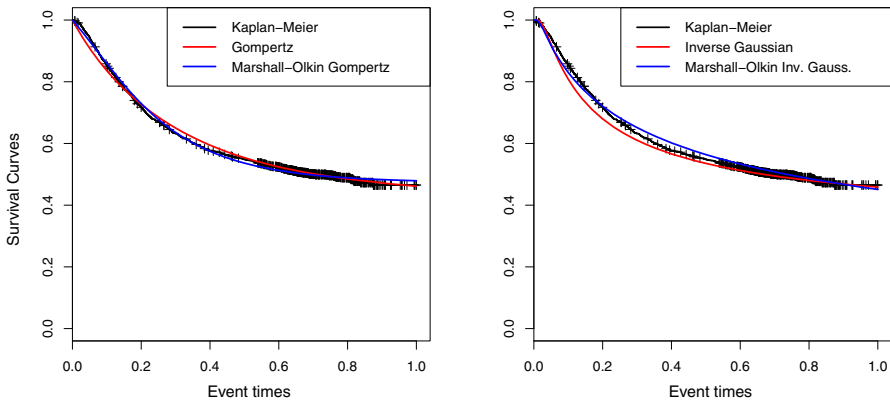


Fig. 13 Survival curves for the fitted Gompertz, Marshall–Olkin Gompertz, inverse Gaussian and Marshall–Olkin inverse Gaussian distributions for the colon data set

Table 7 AIC values for the fitted defective distributions compared with their respective mixture models

Distribution	Leukemia		Birth		Colon	
	Defective	Mixture	Defective	Mixture	Defective	Mixture
Gompertz	52.58	50.74	321.74	197.17	1518.02	1520.02
MO Gompertz	37.16	37.88	80.56	136.75	1488.64	1484.79
Inv. Gaussian	51.38	36.43	99.94	114.36	1597.47	1668.36
MO Inv. Gaussian	35.35	38.34	72.54	109.73	1529.34	1601.22

The bold numbers represent the smallest value between the defective and mixture models in a given data set

The Marshall–Olkin inverse Gaussian distribution is a clear improvement over the inverse Gaussian distribution for all data sets, especially for the leukemia data set. The fitted survival curve for the former captures the Kaplan–Meier curve much better. For the birth data set, both distributions appear to perform equally well at first, but as time increases the tail of the inverse Gaussian distribution gets distanced from the Kaplan–Meier curve while that of the Marshall–Olkin inverse Gaussian distribution keeps close. For the leukemia data set, the inverse Gaussian distribution estimates a by a very small negative value, giving a very small estimate of the cure fraction not significantly different from zero. For the leukemia and birth data sets, the Marshall–Olkin inverse Gaussian distribution estimates a by a negative value with a negative confidence interval. The estimate of a for the colon data set is close to zero, leading to a very small cure fraction.

The estimate of r for Marshall–Olkin distributions is significantly different from 1, meaning that those distributions provide better fits. This can also be checked in Fig. 14. The Marshall–Olkin distributions have points closer to the diagonal line than the baseline distributions.

Table 7 shows there is a big reduction in AIC values when the Marshall–Olkin Gompertz and Gompertz distributions are compared and when the Marshall–Olkin inverse Gaussian and inverse Gaussian distributions are compared.

For the leukemia and birth data sets, the best fitting defective model is the Marshall–Olkin inverse Gaussian distribution, the second best fitting model is the Marshall–Olkin Gompertz distribution, the third best fitting model is the inverse Gaussian distribution and the worst fitting model is the Gompertz distribution. For the colon data set, the best fitting defective model is the Marshall–Olkin Gompertz distribution, the second best fitting model is the Gompertz distribution, the third best fitting model is the Marshall–Olkin inverse Gaussian distribution and the worst fitting model is the inverse Gaussian distribution.

Table 7 also compares the AIC values between the defective and mixture models based on the Gompertz, Marshall–Olkin Gompertz, inverse Gaussian and Marshall–Olkin inverse Gaussian distributions. The bold value represents the smaller value in the comparison. The defective model performs better for the leukemia data set when based on the Marshall–Olkin Gompertz and Marshall–Olkin inverse Gaussian distributions. The defective model is better for the birth data set when based on all but the Gompertz

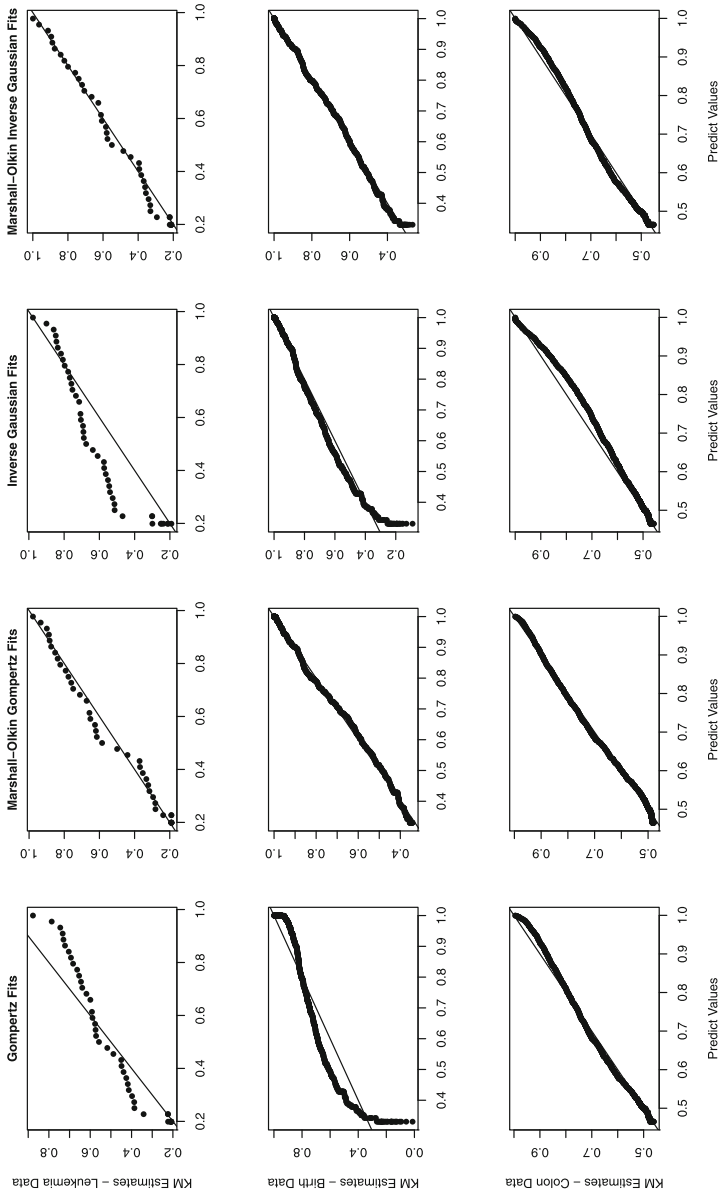


Fig. 14 Plots of the Kaplan–Meier estimates of the survival function versus the predicted values from the proposed distributions. The *top four plots* are for the birth data set. The *middle four plots* are for the leukemia data set. The *bottom four plots* are for the colon data set

distribution. The defective model is better for the colon data set when based on all but the Marshall–Olkin Gompertz distribution.

5 Conclusions

We have proposed two new distributions by using an idea due to Marshall and Olkin. These distributions can assume a defective form. In this way, the cure rate can be estimated by models having one less parameter than the usual standard mixture models.

Three real data applications have shown that Marshall–Olkin distributions perform much better than known deflection distributions in terms of likelihood values, proximity to the Kaplan–Meier curve and AIC values. Further investigations are needed to verify the potential of such distributions as cure fraction models.

Acknowledgments The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil) for financial support during the course of this project. The authors also thank the Associate Editor and the two referees for carefully reading and for comments which greatly improved the paper.

References

- Aalen O, Borgan O, Gjessing H (2008) *Survival and event history analysis: a process point of view*. Springer, New York
- Aalen OO, Gjessing HK (2001) Understanding the shape of the hazard rate: a process point of view (with comments and a rejoinder by the authors). *Stat Sci* 16(1):1–22
- Alice T, Jose KK (2003) Marshall–Olkin pareto processes. *Far East J Theor Stat* 9(2):117–132
- Alice T, Jose KK (2005) Marshall–Olkin logistic processes. *STARS Int J* 6:1–11
- Balka J, Desmond AF, McNicholas PD (2009) Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime Data Anal* 15(2):147–176
- Balka J, Desmond AF, McNicholas PD (2011) Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *J Appl Stat* 38(1):127–144
- Barreto-Souza W, Lemonte AJ, Cordeiro GM (2013) General results for the marshall and olkin's family of distributions. *Anais da Academia Brasileira de Ciências* 85(1):3–21
- Berkson J, Gage RP (1952) Survival curve for cancer patients following treatment. *J Am Stat Assoc* 47(259):501–515
- Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Ser B (Methodol)* 11(1):15–53
- Cantor AB, Shuster JJ (1992) Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat Med* 11(7):931–937
- Chen MH, Ibrahim JG, Sinha D (1999) A new bayesian model for survival data with a surviving fraction. *J Am Stat Assoc* 94(447):909–919
- Cordeiro GM, Lemonte AJ (2013) On the Marshall–Olkin extended weibull distribution. *Stat Papers* 54(2):333–353
- Gieser PW, Chang MN, Rao PV, Shuster JJ, Pullen J (1998) Modelling cure rates using the gompertz model with covariate information. *Stat Med* 17(8):831–839
- Haybittle JL (1959) The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br J Radiol* 32(383):725–733
- Ibrahim JG, Chen MH, Sinha D (2005) *Bayesian survival analysis*. Wiley Online Library, New York
- Jose KK, Krishna E (2011) Marshall–Olkin extended uniform distribution. *Probab Stat Optim* 4:78–88
- Kersey JH, Weisdorf D, Nesbit ME, LeBien TW, Woods WG, McGlave PB, Kim T, Vallera DA, Goldman AI, Bostrom B (1987) Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *N Engl J Med* 317(8):461–467

- Klein JP, Moeschberger ML (2003) *Survival analysis: statistical methods for censored and truncated data*. Springer, New York
- Laurie JA, Moertel CG, Fleming TR, Wieand HS, Leigh JE, Rubin J, McCormack GW, Gerstner JB, Krook JE, Malliard J (1989) Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *J Clin Oncol* 7(10):1447–1456
- Lee MLT, Whitmore GA (2004) First hitting time models for lifetime data. *Handb Stat* 23:537–543
- Lee MLT, Whitmore GA (2006) Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Stat Sci* 21(4):501–513
- Marshall AW, Olkin I (1997) A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika* 84(3):641–652
- R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Schrödinger E (1915) Zur theorie der fall-und steigversuche an teilchen mit brownischer bewegung. *Phys Z* 16:289–295
- Tweedie MCK (1945) Inverse statistical variates. *Nature* 155(3937):453–453
- Whitmore GA (1979) An inverse gaussian model for labour turnover. *J R Stat Soc Ser A (Gener)* 142(4):468–478
- Whitmore GA, Crowder MJ, Lawless JF (1998) Failure inference from a marker process based on a bivariate wiener model. *Lifetime Data Anal* 4(3):229–251
- Yakovlev AY, Tsodikov AD (1996) *Stochastic models of tumor latency and their biostatistical applications*, vol 1. World Scientific, Singapore