AISM (2006) 58: 341–355
DOI 10.1007/s10463-005-0004-x

Srabashi Basu · Ayanendranath Basu · M. C. Jones

Robust and efficient parametric estimation for censored survival data

Received: 27 October 2003 / Revised: 21 December 2004 / Published online: 3 June 2006

Abstract We fit parametric models to survival data in the case of censoring and (outlier) contamination. To do so, we adapt the robust density power divergence methodology of Basu, Harris, Hjort, and Jones (*Biometrika*, *85*, 549–559, 1998) to the case of censored survival data. Asymptotic properties, simulation performance and application to data are provided.

Keywords Density power divergence · Kaplan–Meier · L_2-estimator · M-estimator

1 Introduction

Survival data are commonly encountered in biomedical or industrial settings where n individuals are followed until occurrence of a particular event of interest or n items are put on test until failure. Analysis of survival data is typically complicated by various censoring mechanisms. Since the failure times in many cases are not

S. Basu
Skytech Solutions Pvt Ltd,
BIPPL Bldg A, Salt Lake Electronics Complex,
Kolkata 700 091, India
E-mail: srabashi@hotmail.com

A. Basu
Applied Statistics Unit, Indian Statistical Institute,
203 B. T. Road, Kolkata 700 108, India
E-mail: ayanbasu@isical.ac.in

M. C. Jones (✉)
Department of Statistics, The Open University,
Walton Hall, Milton Keynes,
MK7 6AA, UK
E-mail: m.c.jones@open.ac.uk

observable and the censoring mechanism may or may not be known, a host of semi-parametric procedures have been developed for survival analysis. However, it is well known that the semi-parametric procedures are not as efficient as the maximum likelihood approach (or other efficient parametric methods) if the specified parametric form is valid.

On the other hand, when the underlying model is misspecified or contaminated the maximum likelihood or other classical parametric methods may be severely affected and lead to very poor results. In the presence of censoring the nature and amount of contamination can be very difficult to detect. Therefore, robust methods, which automatically discount the effects of contamination and model misspecification, can serve to provide a compromise between efficient classical parametric methods and the semi-parametric approach provided they are reasonably efficient at the model. In this paper, we consider parametric estimation for right censored data with and without contamination, and try to balance the dual aims of robustness and efficiency using a density-based minimum divergence procedure.

Basu et al. (1998) introduced a family of density-based divergence measures indexed by a tuning parameter α. The population parameters of interest are estimated by minimising a data-based estimate of the proposed divergence between the density underlying the data and the assumed model density. The trade-off between robustness and asymptotic efficiency of the parameter estimators is controlled by α. When $\alpha = 0$, the density power divergence is the Kullback–Leibler divergence (Kullback and Leibler, 1951) and the method is maximum likelihood estimation; when $\alpha = 1$, the divergence is the mean squared error, and a robust but relatively inefficient minimum mean squared error estimator ensues (Scott, 2001). Basu et al. (1998) have shown that the estimators with small $\alpha > 0$ have strong robustness properties with little loss in asymptotic efficiency relative to maximum likelihood under model conditions.

Here, we extend the estimator developed by Basu et al. to estimation of the population parameters under a parametric approach in the context of right censored data. The method has the great advantage that it does not require any nonparametric smoothing for producing a data-based estimate of the true density function, the empirical distribution function alone being used to approximate the appropriate divergence in the case of independently and identically distributed (i.i.d.) data. For the right censoring situation, we take advantage of the well known Kaplan–Meier estimator (Kaplan and Meier 1958), appropriately modified to make it complete, and the substitution of this in place of the empirical distribution function leads to our objective function which is minimised to generate robust parameter estimates.

The rest of the paper is organised as follows. In Sect. 2 we provide a brief review of the density power divergence and related inference in the case of i.i.d. data, and propose a modified estimator to handle right censored data. In Sect. 3 we consider the asymptotic properties of the proposed estimator. Proof of part of our theorem is given in the Appendix. Some simulation results involving exponential and Weibull distributions are presented in Sect. 4, along with robust fitting of Weibull models to data from Efron (1988). Throughout the rest of the paper we will denote distributions by upper case letters and their densities by corresponding lower case ones. The distribution generating the data will be denoted by G, having density g, and will be called the target distribution.

2 The density power divergence and right censored data

2.1 The density power divergence for i.i.d. data

Consider a parametric family of models $\{F_t\}$, indexed by an unknown parameter vector $t \in \Theta \subset \mathbb{R}^s$, possessing densities $\{f_t\}$ with respect to the dominating measure, and let $\mathcal{G}$ be the class of all distributions having densities with respect to that measure. Define the divergence $d_\alpha(g, f)$ between g and another density function f to be

$$d_\alpha(g, f) = \int \left\{ f^{1+\alpha}(z) - \left(1 + \frac{1}{\alpha}\right) g(z) f^\alpha(z) + \frac{1}{\alpha} g^{1+\alpha}(z) \right\} \mathrm{d}z \quad \text{for } \alpha > 0. \quad (1)$$

When $\alpha = 0$, the divergence $d_0(g, f)$ is defined as

$$d_0(g, f) = \lim_{\alpha \to 0} d_\alpha(g, f) = \int \left[g(z) \log(g(z)/f(z)) + (f(z) - g(z)) \right] \mathrm{d}z,$$

which is a version of the Kullback–Leibler divergence. The choice $\alpha = 1$ generates the mean squared error or L_2 distance $\int \{g(z) - f(z)\}^2 \mathrm{d}z$. We eschew values of $\alpha > 1$ in practice in view of their diminished efficiency. Basu et al. (1998) show that for all $\alpha \geq 0$, $d_\alpha(g, f)$ is a divergence in that it is nonnegative for all $g, f \in \mathcal{G}$ and is equal to zero if and only if $f \equiv g$ almost everywhere. A simple consequence of the latter fact is that for any given α the minimum density power divergence functional $T_\alpha(G)$ at G, defined by the requirement $d_\alpha(g, f_{T_\alpha(G)}) = \min_{t \in \Theta} d_\alpha(g, f_t)$, is Fisher consistent.

Note that the density power divergence $d_\alpha(g, f_t)$ between the target density g and the model density f_t can be represented as $\int f_t^{1+\alpha}(z)\mathrm{d}z - (1 + 1/\alpha) \int f_t^\alpha(z) \times \mathrm{d}G(z) + \beta$. The quantity β is independent of the parameter t, so does not affect the minimisation procedure, and the first term is known, given t. Using a random sample $X_1, \ldots, X_n$ from the target distribution G one can actually minimise

$$\int f_t^{1+\alpha}(z)\mathrm{d}z - (1 + 1/\alpha) \int f_t^\alpha(z)\mathrm{d}G_n(z)$$
$$= \int f_t^{1+\alpha}(z)\mathrm{d}z - (1 + 1/\alpha)\, n^{-1} \sum_{i=1}^{n} f_t^\alpha(X_i) \quad (2)$$

with respect to t, where G_n is the empirical distribution function, to obtain the minimum density power divergence estimator of the parameter vector. Notice that this method has the greatly appealing advantage that it does not require a smooth nonparametric estimate of g which is necessary, or unnecessarily imposed, in other robust density-based minimum divergence approaches (e.g. Beran, 1977; Cao et al. 1995; Heathcote, 1977); thus the bandwidth selection problem and rate of convergence results for the kernel density estimator are not relevant. As discussed in Basu et al. (1998, Sect. 3.4) in general the density power divergence estimator is equivariant only to linear data transformations and not to more general ones; however, this covers the important case in survival analysis of timescale changes. It is possible that $\int f_t^{1+\alpha}(z)\mathrm{d}z$ is infinite for some t, but this just means that Eq. (2) will not be minimised at such t.

Under differentiability of the model and appropriate regularity conditions, the minimum density power divergence estimators can be obtained by solving the estimating equation

$$\int u_t(z) f_t^{1+\alpha}(z)\mathrm{d}z - n^{-1}\sum_{i=1}^{n} u_t(X_i) f_t^{\alpha}(X_i) = 0,$$

where $u_t(z) = \partial \log f_t(z)/\partial t$ is the maximum likelihood score function. Note that the above estimating equation is unbiased when $g = f_t$. If, for example, $\{F_t\}$ is a location model, with location parameter t, the minimum density power divergence estimator is the maximiser of $\sum_i f_t^{\alpha}(X_i)$, with corresponding estimating equation $\sum_i u_t(X_i) f_t^{\alpha}(X_i) = 0$. This contrasts with the maximum likelihood estimator which maximises $\sum_i \log f_t(X_i)$, with the corresponding estimating equation being $\sum_i u_t(X_i) = 0$. For several parametric models such as the normal, $u_t(z) f_t^{\alpha}(z)$ is a *bounded* function of z for fixed t and for all $\alpha > 0$, although $u_t(z)$ itself is not. Thus the estimating equation of the minimum density power divergence estimator downweights the score function in a probabilistic manner. Basu et al. (1998) have shown that the estimators corresponding to small values of α combine strong robustness properties together with reasonably high efficiency.

2.2 The density power divergence for right censored data

Now let (X_i, C_i) , $i = 1, \ldots, n$, be n i.i.d. pairs of random variables. The variables X_i are randomly generated from the target distribution G which is modeled by the parametric family $\{F_t\}$. The sequence of variables $\{C_i\}$ are censoring variables, so that one actually observes $Y_i = \min(X_i,\ C_i)$ and the indicator function δ_i, where $\delta_i = 1$ if $X_i < C_i$, $\delta_i = 0$ otherwise. Although in most conceivable applications the distributions G and H of X_i and C_i, respectively, will both be absolutely continuous with respect to Lebesgue measure, our results will also hold if they are not but do not have any jump points in common. Throughout the rest of the paper we will assume that the variable of interest X and the censoring variable C are independent.

Kaplan and Meier (1958) developed a nonparametric estimator for the survival function $S(x) = 1 - G(x)$ as

$$\widehat{S}_n(x) = \begin{cases} \Pi_{i:Y_{(i)} \le x} \left(\frac{n-i}{n-i+1}\right)^{I(\delta_{(i)}=1)} & \text{if } x \le Y_{(n)}, \\ 0 & \text{if } \delta_{(n)} = 1 \text{ for } x > Y_{(n)}, \\ \text{undefined} & \text{if } \delta_{(n)} = 0 \text{ for } x > Y_{(n)}, \end{cases}$$

where $(Y_{(i)},\ \delta_{(i)})$, $i = 1, 2, \ldots, n$, are the n pairs of observations ordered over the $Y_{(i)}$. The Kaplan–Meier estimator is a step function with positive mass points at those observations X_i for which $\delta_i = 1$, i.e. only if X_i is a failure; if $\delta_i = 1$ for all i, the Kaplan–Meier estimator reduces to the ordinary empirical survivor function $1 - G_n$. In the case where the largest observation is censored, the Kaplan–Meier estimator is undefined after the largest failure point. It is convenient to artificially complete $\widehat{S}_n$ by distributing the leftover mass equally among all the censored observations greater than the largest failure. In this paper we have followed this convention.

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimator of the underlying survival function. It is a strongly consistent estimator of the target survival function $S = 1 - G$, so that $\widehat{S}_n(x) \to S(x)$ almost surely under appropriate conditions (see Peterson, 1977; Miller, 1981), the most important one of which is formally stated in assumption A6 in Sect. 3 (implicitly, we assume this to be true for the rest of the paper). When these conditions are satisfied, the method described in the previous paragraph of artificially completing the Kaplan–Meier estimator has no role in its asymptotic properties; the adjustment is simply a tool to make the method work in small samples.

Now, the reason why the minimum density power divergence method is able to avoid the use of a smooth nonparametric density estimate is that in expression (1) for the divergence the target distribution appears only in a linear functional (except in the part which is independent of the unknown parameter and which therefore has no role in the optimisation). Thus, in the right censoring context described above, we can replace G_n in Eq. (2) by $\widehat{G}_n(x) = 1 - \widehat{S}_n(x)$ which provides a consistent estimator of the true distribution function in this context, and which is the Kaplan–Meier estimator of the distribution function G. Therefore, for the right censoring situation we generate the sample version of the density power divergence between the model density f_t and the target density g, minus the term β independent of t, as

$$D_n(t) = \int f_t^{1+\alpha}(z)\mathrm{d}z - \left(1 + \frac{1}{\alpha}\right) \int f_t^{\alpha}(z)\mathrm{d}\widehat{G}_n(z). \tag{3}$$

The corresponding estimating equation for the unknown parameter is then given by

$$U_n(t) = \int u_t(z) f_t^{1+\alpha}(z)\mathrm{d}z - \int u_t(z)\mathrm{d}\widehat{G}_n(z) = 0. \tag{4}$$

3 Consistency and asymptotic normality

Here we establish the consistency and asymptotic normality of the minimum density power divergence estimator in the right censored situation when the data are generated from the target distribution G. In the following theorem, θ represents the best fitting parameter, whereas t denotes a generic element of Θ. The best fitting parameter is the minimiser of $D(t) = \int f_t^{1+\alpha}(z)\mathrm{d}z - (1 + 1/\alpha) \int f_t^{\alpha}(z)\mathrm{d}G(z)$ with respect to t which will be assumed to exist and be unique. Let $\widehat{\theta}$ be the minimiser of $D_n(t)$ given by Eq. (3) One can represent $D_n(t)$ as $\int V_t(z)\mathrm{d}\widehat{G}_n(z)$, where

$$V_t(x) = \int f_t^{1+\alpha}(z)\mathrm{d}z - \left(1 + \frac{1}{\alpha}\right) f_t^{\alpha}(x).$$

We assume that G and the censoring distribution H have no common points of discontinuity.

The minimum density power divergence estimator which is obtained as the solution of $\int \psi(z, t)\mathrm{d}\widehat{G}_n(z) = 0$, where $\psi(x, t) = (\psi_1(x, t), \ldots, \psi_s(x, t))^T = (\partial V_t(x)/\partial t_1, \ldots, \partial V_t(x)/\partial t_s)^T$, is also a particular form of M-estimator for censored data; for the latter, see e.g., Reid (1981) and Wang (1999). The difficulty in developing the asymptotic properties of M-estimators for censored data in general has been the absence of a law of large numbers and central limit theorem results

for general functionals $\int \phi \, \mathrm{d}\widehat{G}_n$ of the Kaplan–Meier estimator. However, during the last decade or so the works of W. Stute and J.L. Wang (Stute and Wang, 1993; Stute, 1995; Wang, 1995, 1999) have laid down just such a theoretical framework and obtained strong consistency and asymptotic normality results.

Below, we present a theorem on the asymptotic properties of our minimum divergence estimator; the theorem has two parts, (1) consistency, (2) asymptotic normality. It turns out that, since we do not assume the components of $\psi(x, t)$ to be bounded, it is easiest to adapt Lehmann's (1983) Theorem 6.4.1(i) on consistency to the censored data case; this is done in the Appendix. Had we assumed boundedness, as would often be the case, then the consistency result of part (1) of our theorem would follow from Wang (1999, Theorem 3(i)). Part (2) of our theorem follows directly from Theorem 5 of Wang (1999). For any given α, we first make the following assumptions:

A1: The distributions F_t have common support, so that the set $A = \{z | f_t(z) > 0\}$ is independent of t. The true distribution G is also supported on A, on which $g > 0$.
A2: There is an open subset ω of the parameter space Θ containing the best fitting parameter θ such that $\int f_t^{1+\alpha}(z)\mathrm{d}z < \infty$ and, for almost all $z \in A$ and all $t \in \omega$, $f_t(z)$ is three times differentiable with respect to t and the third partial derivatives are continuous with respect to t.
A3: The integrals $\int f_t^{1+\alpha}(z)\mathrm{d}z$ and $\int f_t^{\alpha}(z)\mathrm{d}G(z)$, when finite, can be differentiated three times, and the derivatives can be taken under the integral sign. $E_G\left\{\partial V_t(X)/\partial t_k|_{t=\theta}\right\} < \infty$ for all $k = 1, \ldots, s$.
A4: The $s \times s$ matrix $J(G, t)$ is defined by

$$J_{kl}(G, t) = E_G\left\{\frac{\partial^2 V_t(X)}{\partial t_k \partial t_l}\right\}, \quad k, l = 1, \ldots, s.$$

All elements of $J(G, \theta)$ are finite and the matrix is positive definite.
A5: For all $k, l, m = 1, \ldots, s$, there exist functions $M_{klm}(x)$ such that

$$\left|\frac{\partial^3 V_t(x)}{\partial t_k \partial t_l \partial t_m}\right| \leq M_{klm}(x)$$

for all $t \in \omega$, where $E_G[M_{klm}(X)] = m_{klm} < \infty$ for all $k, l, m = 1, \ldots, s$.
A6: For a distribution L, let $\tau_{\mathrm{L}} = \sup\{x\colon L(x) < 1\}$ denote the upper bound of the support of L. Then, $\tau_G \leq \tau_H$, where equality may hold except when H is continuous at τ_G, and $G(\tau_G) - G(\tau_G-) > 0$.
A7: The conditions of Lemma 1 and Theorem 5 of Wang (1999) hold. These are essentially conditions on the first two moments of $\psi_k(x, t)$, $k = 1, ..., s$, when $t = \theta$ and on the continuity of their derivatives with respect to elements of t at $t = \theta$.

Theorem 3.1 *Under the above conditions, with probability tending to 1 as $n \to \infty$, there exist solutions $\widehat{\theta}_n$ of the density power divergence estimating Eq. (4) such that:*

(1) $\widehat{\theta}_n$ is consistent for estimating θ;
(2) $n^{1/2}(\widehat{\theta}_n - \theta)$ is multivariate normal with (vector) mean zero and covariance matrix $J(G, \theta)^{-1}\boldsymbol{C}(\psi, \theta, G, H)J(G, \theta)^{-1}$, where $\boldsymbol{C}(\cdot, \cdot, \cdot, \cdot)$ is as defined in equation (2.15) of Wang (1999) with our G and H replacing F and G in the notation of that paper.

To keep the description simple, we have limited the statement of the theorem to the existence of one sequence of consistent roots to the estimating equation. However, if there exists a compact set $K \subset I\!R^s$ such that $\inf_{t\in K} \left| \int \psi_j(z,t) \mathrm{d}G(z) \right| > 0$, for $1 \leq j \leq s$, then any sequence of solutions to the estimating equation converges to the best fitting parameter θ [Wang, 1999; Theorem 3(ii)]. Expressions (3.2) of Wang give the basic form of the influence functions for our estimators.

4 Numerical studies

4.1 Simulation results: exponential distributions

Consider the lifetime distribution to be the one parameter exponential with density $f_\lambda(x) = \lambda \mathrm{e}^{-\lambda x}$, $x \geq 0$, referred to as exp(λ). Hence, the first term in Eq. (2) is

$$\int_0^\infty f_\lambda^{1+\alpha}(z)\, \mathrm{d}z \;=\; \frac{\lambda^\alpha}{1+\alpha}.$$

The second term in Eq. (2) may be written as

$$\left(1+\frac{1}{\alpha}\right) \int f_\lambda^\alpha(z)\, \mathrm{d}\widehat{G}_n(z) = \left(1+\frac{1}{\alpha}\right) \lambda^\alpha \sum_j g_n(y_j)\, \mathrm{e}^{-\lambda\alpha y_j}.$$

Here, the y_js are the set of failure times plus all the censored values greater than the largest failure time and $g_n(y_j)$ is the mass attributed to y_j by the completed Kaplan–Meier estimator (recall that if $Y_{(n)}$ is not a failure, then the residual mass is assigned equally to all censored observations larger than $Y_{(n)}$). To obtain the minimum divergence estimator of λ, we therefore minimise

$$d_\alpha^*(g,f) = \frac{\lambda^\alpha}{1+\alpha} - \frac{1+\alpha}{\alpha}\lambda^\alpha \sum_j g_n(y_j)\mathrm{e}^{-\lambda\alpha y_j}$$

with respect to λ for fixed α. This leads to the estimating equation for λ,

$$\alpha \;-\; (1+\alpha)^2 \sum_j (1-\lambda y_j)\, g_n(y_j)\, \mathrm{e}^{-\lambda\alpha y_j} = 0,$$

which can be solved numerically, using, for example, the Newton–Raphson procedure.

A modest numerical study is performed to compare the performance of the maximum likelihood estimator (MLE) and the minimum divergence estimator (MDE) developed in this paper in the exponential model with or without contamination for various values of α. A sample is generated from exp(5) and 0, 5, 10, 15 or 20% of the observations are contaminated by exp(1.5) successively. We have used an exponential censoring scheme with the censoring rate determined so as to keep the expected proportions of censoring under the true distribution at 10 or 20%: when the true distribution is exp(5), to keep the expected proportion

of censoring at 10%, the censoring distribution is taken to be exp(5/9); when the expected proportion of censoring is 20%, the censoring distribution is exp(5/4). The values of α are chosen to be 0.001, 0.01, 0.1, 0.2, 0.25, 0.5, 0.75 and 1.0. For given levels of contamination and censoring, the MLE and MDE for each value of α are calculated for a randomly generated exponential sample of size 50 and the whole procedure is repeated 500 times. The mean squared errors between the MLE and the true parameter, MSE(MLE), and between the MDE and the true parameter, MSE(MDE), are computed. Empirical efficiency is defined to be the ratio of MSE(MLE):MSE(MDE) so that efficiency greater than 1 implies the density-based estimator is performing better than the MLE. The results of the simulation study are given in Table 1.

The general observations from the empirical efficiencies in Table 1 are as follows. Under pure data (no contamination) the MDEs are generally less efficient than the MLE, as one would expect. In this case the efficiencies are generally decreasing functions of α. As the contamination proportion increases, however, the MLE gets progressively worse. Even for moderate contaminations at 5–10% levels, the estimators for relatively small values of α (say between 0.1 and 0.25) are superior to the MLE. For the largest contamination proportion considered here (20%) all estimators corresponding to $\alpha \geq 0.1$ outperform the MLE, some of them substantially. However, the gains from using the robust estimates are reduced somewhat when the censoring proportion increases.

An interesting suggestion made by a referee was to repeat the above exercise with 'short-tailed contamination' in the form of an exp(15) distribution. This we did although detailed results have not been added to the paper. Basically, this exercise emphasises that the MDE, in line with most other robust estimators, is driven by downweighting regions of low density. As the contamination here is in a high density area, the MDE does not, in general, gain over the MLE. It retains good performance for small α but can deteriorate badly for large values of α, particularly for high amounts of contamination.

Table 1 Empirical efficiencies of MDE in exponential case

Censoring	α	Contamination				
		None	5%	10%	15%	20%
10%	0.001	0.9738	0.9378	0.9083	0.8921	0.8944
	0.01	0.9777	0.9544	0.9278	0.9091	0.9100
	0.1	0.9708	1.0845	1.1215	1.1095	1.0967
	0.2	0.8976	1.0808	1.2356	1.3266	1.3240
	0.25	0.8500	1.0439	1.2446	1.4062	1.4227
	0.5	0.6314	0.8189	1.0799	1.4659	1.6300
	0.75	0.5118	0.6844	0.9261	1.3508	1.5822
	1.0	0.4477	0.6096	0.8401	1.2635	1.5120
20%	0.001	0.9342	0.9406	0.9194	0.9085	0.9115
	0.01	0.9413	0.9502	0.9303	0.9183	0.9199
	0.1	0.9722	1.0230	1.0348	1.0307	1.0206
	0.2	0.9335	1.0214	1.0993	1.1539	1.1493
	0.25	0.8945	0.9895	1.0990	1.2003	1.2103
	0.5	0.6748	0.7649	0.9291	1.2128	1.3453
	0.75	0.5459	0.6356	0.7951	1.1198	1.3121
	1.0	0.4787	0.5692	0.7239	1.0539	1.2640

4.2 Simulation results: Weibull distributions

Now consider the lifetime distribution to be the two parameter Weibull with density given by

$$f_{a,b}(x) = \left(\frac{b}{a}\right)\left(\frac{x}{a}\right)^{b-1} \mathrm{e}^{-(x/a)^b}, \quad a, b > 0, \quad x \geq 0,$$

denoted Weibull (a, b); a and b are scale and shape parameters, respectively. For the Weibull density, the first term in Eq. (2) may be written as

$$\left(\frac{b}{a}\right)^{\alpha} \int_0^{\infty} z^{\alpha(1-\frac{1}{b})} \mathrm{e}^{-(1+\alpha)z} \mathrm{d}z$$

which equals

$$\left(\frac{b}{a}\right)^{\alpha} \left(\frac{1}{1+\alpha}\right)^{\alpha(1-\frac{1}{b})+1} \Gamma\left(\alpha\left(1-\frac{1}{b}\right)+1\right)$$

provided that $b > \alpha/(1+\alpha)$ (which is assured for $b > 1/2$ if $0 \leq \alpha \leq 1$). For $b \leq \alpha/(1+\alpha)$ this term is infinite, which means that for any given α, the MDE of b will always be greater than $\alpha/(1+\alpha)$. The second term in Eq. (2) reduces to

$$\left(\frac{1+\alpha}{\alpha}\right)\left(\frac{b}{a}\right)^{\alpha} \sum_j g_n(y_j) \left(\frac{y_j}{a}\right)^{\alpha(b-1)} \mathrm{e}^{-\alpha(y_j/a)^b}.$$

Random samples of size $n = 50$ are generated from a Weibull(2, 5) distribution and the censoring scheme is taken to be exp(0.0575) for an expected censoring proportion of 10% and exp(0.1222) for an expected censoring proportion of 20%. Contamination is introduced through exp(1.5) and the contaminating proportion is varied as in the case of exponential lifetime distribution; note that the contamination is (largely) to the left of the true distribution in this case. The MLE of b is found by solving the appropriate likelihood estimating equation utilizing the bisection method. Once b is estimated, the MLE of a is immediately available. To obtain the MDEs of a and b, we do a bivariate grid search. Mean squared errors and their ratio, the empirical efficiency, are calculated separately for the scale and shape parameters and the results are presented in Tables 2 and 3, respectively. The number of replications is again 500.

For the Weibull distribution, in general, the performance of the MDE compared to MLE is superior than in the case of the exponential distribution, both in magnitude and scope. At higher levels of contamination and larger values of α the MDE is between 2.5 to 8 times better than the MLE in terms of empirical efficiency. Even at moderate levels of contamination, the superiority of the MDEs, including those for larger values of α, are clearly apparent.

Table 2 Empirical efficiencies of MDE for the scale parameter a in the Weibull case.

Censoring	α	Contamination				
		None	5%	10%	15%	20%
10%	0.001	0.9839	0.9984	0.9890	0.9831	0.9763
	0.01	0.9802	1.0412	1.0435	1.0291	1.0153
	0.1	0.9697	1.5182	1.7027	1.6979	1.5073
	0.2	0.9543	1.8063	2.4346	2.8238	2.4815
	0.25	0.9306	1.8509	2.6907	3.3910	3.1885
	0.5	0.8348	1.7944	3.0735	4.8071	7.0624
	0.75	0.7329	1.6132	2.8926	4.7488	8.4722
	1.0	0.6408	1.4474	2.6568	3.2915	6.3208
20%	0.001	0.9834	0.9976	0.9849	0.9588	0.9491
	0.01	0.9845	1.0448	1.0302	0.9974	0.9822
	0.1	0.9660	1.4707	1.6550	1.5893	1.4610
	0.2	0.9395	1.6887	2.3168	2.6244	2.3812
	0.25	0.9139	1.7302	2.5336	3.2183	3.0225
	0.5	0.8023	1.6512	2.8039	5.0254	6.2407
	0.75	0.6963	1.4828	2.6251	5.1667	7.3060
	1.0	0.6028	1.3020	2.3426	3.7405	5.6013

Table 3 Empirical efficiencies of MDE for the shape parameter b in the Weibull case

Censoring	α	Contamination				
		None	5%	10%	15%	20%
10%	0.001	1.0646	1.0969	1.0864	1.0709	1.0517
	0.01	1.0601	1.1573	1.1282	1.0959	1.0633
	0.1	1.0438	2.0538	1.8054	1.4986	1.2292
	0.2	0.9570	2.8471	3.0339	2.4424	1.6174
	0.25	0.9098	3.0811	3.6374	3.0703	1.9448
	0.5	0.6915	2.8991	4.8740	5.6297	4.2573
	0.75	0.5397	2.4816	4.7189	6.0842	5.6305
	1.0	0.4657	2.1752	4.3537	5.9878	6.1775
20%	0.001	1.0534	1.2029	1.1737	1.1270	1.1041
	0.01	1.0552	1.2730	1.2174	1.1481	1.1163
	0.1	1.0022	2.2253	1.9601	1.4758	1.2933
	0.2	0.9235	2.9974	3.2399	2.2520	1.7134
	0.25	0.8795	3.1039	3.8048	2.8244	2.0623
	0.5	0.6900	2.8162	4.8515	5.4431	4.4253
	0.75	0.5573	2.4046	4.4592	6.1037	5.7111
	1.0	0.4663	2.1928	4.2664	6.3243	6.1143

4.3 Data example

Next we apply this procedure to a real example taken from Efrom (1988). Data are available from a study comparing radiation therapy alone (arm A) and radiation therapy and chemotherapy (arm B) for the treatment of head and neck cancer. There were 51 patients assigned to arm A of the study of which 9 were lost to follow-up and, therefore, censored; alternatively, 45 patients were assigned to arm B of the study of which 14 were lost to follow-up. Note that censoring levels are fairly high in these data sets, approximately 20 and 30%, respectively. Efron (1988) makes various analyses of these data, which show radiation and chemotherapy B to be more effective in terms of survival times. Our focus here is on the appropriateness

or otherwise of certain parametric models for these data, basing our analysis on a standard model for such data, the Weibull distribution.

The MLEs and the MDEs of the two Weibull parameters are given for various values of the tuning parameter α in Tables 4 and 5. There are very considerable changes in both parameter estimates with α including, importantly, a change from $\hat{b} < 1$ (MLE and small α MDE) to $\hat{b} > 1$ (larger α MDE).

Figures 1 and 2 illustrate the results for Arms A and B, respectively. On each figure is shown: a kernel density estimate formed by kernel smoothing the Kaplan–Meier estimator (e.g. Wand and Jones, Sect. 6.2.3), using a bandwidth subjectively chosen not to oversmooth the data; the Weibull model fitted by MLE; the Weibull model fitted by MDE with $\alpha = 1$; and a further curve to be discussed below. What is clearly shown by the kernel density estimate in each case is a main body of data to the left, together with some much more long-lived individuals to the right. (The precise nature of the long tail may not be very well reflected by the kernel density estimate, especially when there are several large censored observations as in Arm B.)

The MLE Weibull fits are monotone decreasing because $\hat{b} < 1$, striking an uneasy compromise between accommodating the main body and the long tail of the data, and consequently failing to capture either. The robust Weibull fits, with $\hat{b} > 1$, provide a wholly better fit to the main body of the data at the expense of essentially ignoring the long tail. As such, this is entirely successful from the robust fitting viewpoint taken by this paper (and the whole of the robustness literature).

Table 4 Analysis of Efron data assuming Weibull model: Arm A

	α	Scale, $\hat{a}$	Shape, $\hat{b}$
MLE	0	399.24	0.91
MDE	0.001	418.18	0.98
	0.01	417.72	0.98
	0.1	412.72	0.99
	0.2	402.51	1.00
	0.25	395.31	1.02
	0.5	321.90	1.16
	0.75	252.85	1.44
	1.0	249.47	1.47

Table 5 Analysis of Efron data assuming Weibull model: Arm B

	α	Scale, $\hat{a}$	Shape, $\hat{b}$
MLE	0	925.45	0.76
MDE	0.001	789.23	0.91
	0.01	790.07	0.91
	0.1	791.81	0.90
	0.2	789.26	0.90
	0.25	785.13	0.90
	0.5	726.72	0.93
	0.75	551.53	1.03
	1.0	343.07	1.31

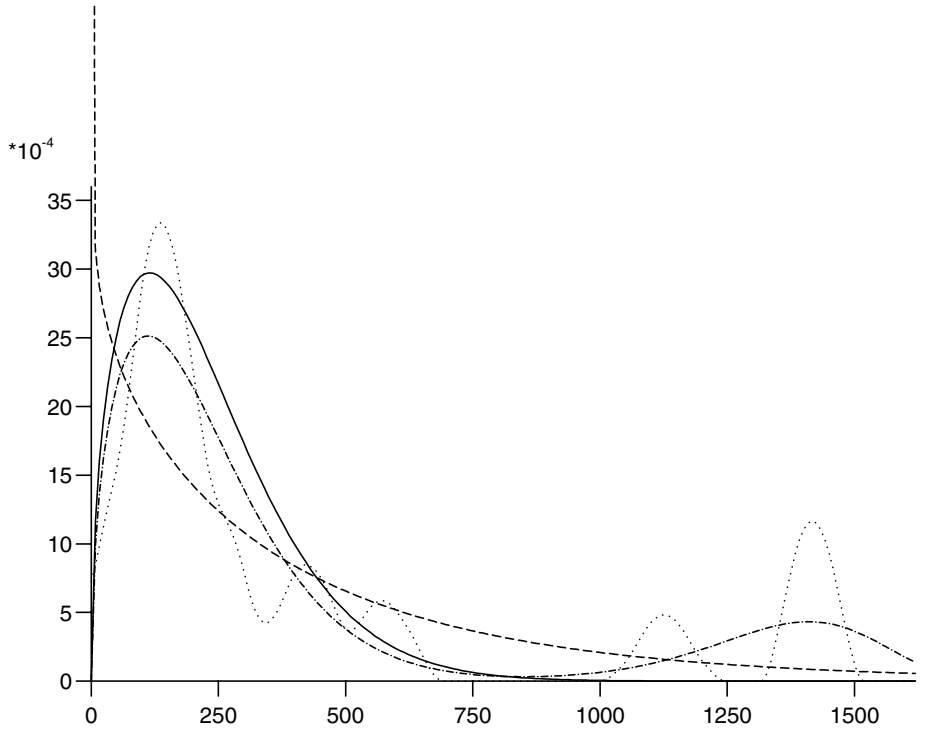

Fig. 1 Kernel density estimate (dotted line), MLE Weibull fit (dashed line), MDE $\alpha = 1$ Weibull fit (solid line) and MLE two component mixed Weibull fit (dot-dashed line) for Arm A of the Efron (1988) data

Table 6 Maximum likelihood parameter estimates for Efron data assuming two component Weibull mixture model

		Arm A	Arm B
	$\hat{p}$	0.82	0.35
First Component	scale, $\hat{a}$	241.53	156.00
	shape, $\hat{b}$	1.47	4.08
Second Component	scale, $\hat{a}$	1428.11	1800.00
	shape, $\hat{b}$	9.17	0.90

However, particularly in cases, as here, with substantial 'contamination', it can be argued that the contamination is in fact of interest too and should also be modelled. The results of fitting two component Weibull mixtures to the data by maximum likelihood are, therefore, also shown on Figs. 1 and 2. (The corresponding parameter estimates, in an obvious notation, are given in Table 6.) In Fig. 1, the mixed Weibull distribution confirms the robust Weibull fit as being appropriate for the main body of data and adds a small second component to cover the tail. In Fig. 2, the mixed Weibull distribution takes a rather different form, that of a narrow peak to the left and a long flat tail to the right. Further alternative parametric models

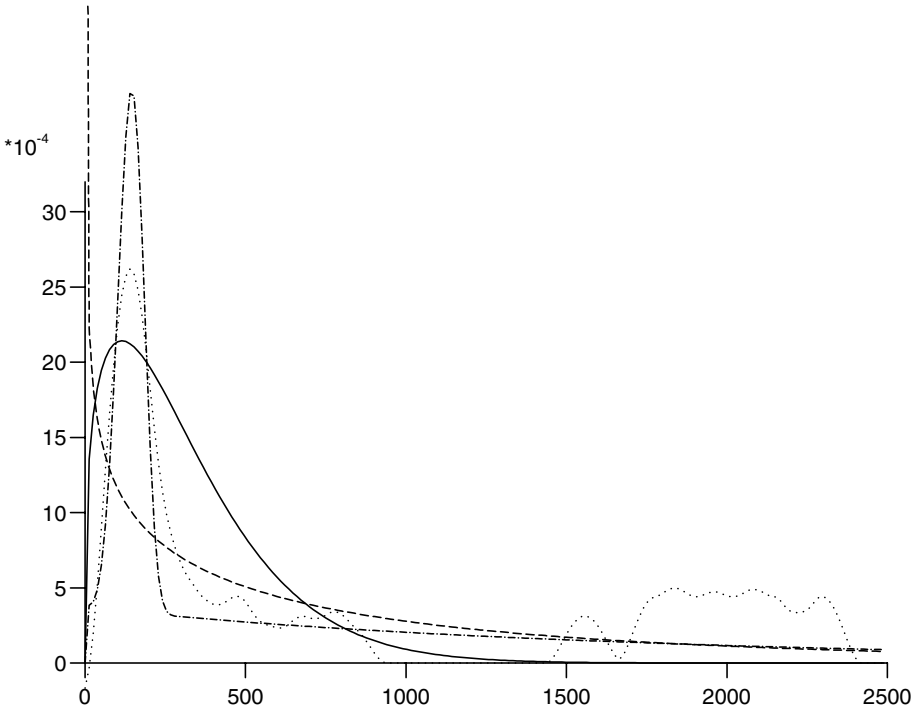

Fig. 2 Kernel density estimate (dotted line), MLE Weibull fit (dashed line), MDE $\alpha = 1$ Weibull fit (solid line) and MLE two component mixed Weibull fit (dot-dashed line) for Arm B of the Efron (1988) data

with heavier tails (e.g. the F?) might be an even better full-modelling way forward in this case.

This last part of our analysis reflects the usual tension between robust concentration on the majority of the data and full modelling of all aspects of the data that is ubiquitous in the literature.

5 Some remarks on choice of α

We have only little to contribute regarding appropriate choice of α. First, some insight can be gained by looking at parameter estimates (and corresponding fitted densities) for a range of values of α. Second, for low levels of censoring and contamination, small values of α, say between 0.05 and 0.25, tend to be appropriate. Third, the optimal value of α increases with increased levels of both censoring and, particularly, contamination. Values from 0.5 up to and including the 'L_2 estimation' choice of $\alpha = 1$ (Scott, 2001), as made in our data example, then seem more appropriate. Fourth, despite the negative views expressed in Sect. 5 of Basu et al. (1998), in the non-censored-survival situation, some progress has been made on automatic data-based selection of α; see Hong and Kim (2001) and Warwick

and Jones (2005). Unfortunately, it is not at all easy to adapt the latter approach to the survival data scenario, which is why we have not done so here.

Acknowledgements The authors are very grateful to two anonymous referees for suggestions which improved the quality and correctness of this paper.

6 Appendix

6.1 Proof of part (1) of Theorem 3.1

To prove the existence, with probability tending to 1, of a sequence of solutions to the estimating equation given in Eq. (4) which is consistent, we shall consider the behaviour of the density power divergence, given by Eq. (3), as a function of t, on a sphere Q_a with center at θ and radius a. We will show that for sufficiently small a the probability that $D_n(t) > D_n(\theta)$ tends to 1 for all points t on the surface of Q_a, and hence that $D_n(t)$ has a local minimum in the interior of Q_a. It will follow that for any $a > 0$, with probability tending to 1 as $n \to \infty$, the density power divergence estimating equations have a solution $\widehat{\theta}_n(a)$ within Q_a.

To study the behaviour of $D_n(t)$ on Q_a, we expand $D_n(t)$ around θ. Thus

$$\begin{aligned} D_n(\theta) - D_n(t) &= -\sum_{k=1}^{s} A_k(t_k - \theta_k) - \frac{1}{2}\sum_{k=1}^{s}\sum_{l=1}^{s} B_{kl}(t_k - \theta_k)(t_l - \theta_l) \\ &\quad + \frac{1}{6}\sum_{k=1}^{s}\sum_{l=1}^{s}\sum_{m=1}^{s}(t_k - \theta_k)(t_l - \theta_l)(t_m - \theta_m) \\ &\quad \times \int \gamma_{klm}(z) M_{klm}(z) \mathrm{d}\widehat{G}_n(z) \\ &= S_1 + S_2 + S_3, \end{aligned}$$

say, where

$$A_k = \frac{\partial}{\partial t_k} D_n(t)|_{t=\theta}, \quad B_{kl} = \frac{\partial^2}{\partial t_k \partial t_l} D_n(t)|_{t=\theta}, \quad 0 \le |\gamma_{klm}(x)| \le 1,$$

the last by assumption A5. First, note that

$$A_k = \int \frac{\partial}{\partial t_k} V_t(z) \mathrm{d}\widehat{G}_n(z),$$

so that it converges (using assumptions A3, A6 and Proposition 1 of Wang, 1999), with probability tending to 1, to $\{\partial D(t)/\partial t_k\}|_{t=\theta} = 0$ as $n \to \infty$. Similarly, $B_{kl} \to J_{kl}$ with probability tending to 1 (using assumptions A4, A6 and Proposition 1 of Wang, 1999). For any given a it follows that $|A_k| < a^2$ and hence $|S_1| < sa^3$ with probability tending to 1. Next,

$$2S_2 = -\sum_{k=1}^{s}\sum_{l=1}^{s} J_{kl}(t_k - \theta_k)(t_l - \theta_l) + \sum_{k=1}^{s}\sum_{l=1}^{s}(-B_{kl} + J_{kl})(t_k - \theta_k)(t_l - \theta_l).$$

It follows from an argument similar to that for S_1 that the absolute value of the second term of $2S_2$ is less that s^2a^4 with probability tending to 1. The first term of $2S_2$ is a negative (nonrandom) quadratic form in the variables $(t_k - \theta_k)$. By an orthogonal transformation this can be reduced to a diagonal form $\sum_i \lambda_i \xi_i^2$ with $\sum_i \xi_i^2 = a^2$. As the λs are all negative, by ordering them as $\lambda_s \leq \lambda_{s-1} \leq \ldots \leq \lambda_1 < 0$, one gets $\sum_i \lambda_i \xi_i^2 \leq \lambda_1 a^2$. Combining the first and the second terms, there exist $c > 0$, $a_0 > 0$ such that for $a < a_0$, $S_2 < -ca^2$, with probability tending to 1. Finally, with probability tending to 1, $\int \gamma_{klm}(z) M_{klm}(z) \mathrm{d}\widehat{G}_n(z) < 2m_{klm}$, and hence $|S_3| < ba^3$ on Q_a where $b = (\sum_k \sum_l \sum_m m_{klm})/3$. Combining these inequalities, we see that $\max(S_1 + S_2 + S_3) < -ca^2 + (b+s)a^3$, which is less than zero if $a < c/(b+s)$.

Thus, for sufficiently small a there exists a sequence of roots $\widehat{\theta}_n = \widehat{\theta}_n(a)$ such that $P(||\widehat{\theta}_n - \theta|| < a) \to 1$ where $||\cdot||$ represents the L_2 norm. It remains to show that we can determine such a sequence independently of a. Let θ_n^* be the root closest to θ. This exists because the limit of a sequence of roots is again a root by the continuity of $D_n(t)$ as a function of t. Then clearly $P(||\theta_n^* - \theta|| < a) \to 1$. This concludes the proof.

References

Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*, 549–559.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, *5*, 445–463.

Cao, R., Cuevas, A., Fraiman, R. (1995). Minimum distance density-based estimation. *Computational Statistics and Data Analysis*, *20*, 611–631.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association*, *83*, 414–425.

Heathcote, C.R. (1977). The integrated squared error estimation of parameters. *Biometrika*, *64*, 255–264.

Hong, C., Kim, Y. (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Association*, *30*, 453–465.

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Kullback, S., Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lehmann, E.L. (1983). *Theory of Point Estimation*. New York: Wiley.

Miller, R.G. (1981). *Survival Analysis*. New York: Wiley.

Peterson, A.V. (1977). Expressing the Kaplan–Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, *72*, 854–858.

Reid, N. (1981). Influence functions for censored data. *Annals of Statistics*, *9*, 78–92.

Scott, D.W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, *43*, 274–285.

Stute, W. (1995). The central limit theorem under random censorship. *Annals of Statistics*, *23*, 422–439.

Stute, W., Wang, J.L. (1993). The strong law under random censorship. *Annals of Statistics*, *21*, 1591–1607.

Wand, M.P., Jones, M.P. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wang, J.L. (1995). M-estimators for censored data: strong consistency. *Scandinavian Journal of Statistics*, *22*, 197–206.

Wang, J.L. (1999). Asymptotic properties of M-estimators based on estimating equations and censored data. *Scandinavian Journal of Statistics*, *26*, 297–318.

Warwick, J., Jones, M.C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, 75, 581–588.