A SIMULATION STUDY OF A MINIMUM DISTANCE ESTIMATOR FOR FINITE MIXTURES UNDER CENSORING

RENÉ FERLAND AND SORANA FRODA
Université du Québec à Montréal

JEAN LAVIGNE
MDS Pharma Services

In this paper we introduce a minimum distance estimator of the weights in a finite mixture when the data are censored. Our estimator is a natural extension of the estimator of Choi and Bulgren (1968). In order to check the robustness of this minimum distance estimator we perform a simulation study.

1. Introduction

In this paper we consider data $Y_1, \dots, Y_n$ from the following finite mixtures model

$$(1.1) \qquad H(t) = \sum_{j=1}^{k} \pi_j F_j(t) = \sum_{j=1}^{k} \pi_j F(t, \theta_j), \quad \pi_j \geq 0, \sum_{j=1}^{k} \pi_j = 1,$$

where k is known, and the parameters θ_j, $j = 1, \dots, k$ are either known or unknown. The data Y_i, $i = 1, \dots, n$ are times (to relapse or recovery, e.g.) and therefore it makes sense to assume that $F_j(t)$ is some typical distribution in survival analysis, e.g. exponential, Weibull or Gompertz. In medical studies, many illnesses are actually mixtures of two or more conditions. The rate of survival (relapse, failure) can be different for each of these conditions, but the conditions or cause of death may be hard to identify. In a cohort, one may have both categorized and uncategorized data (patients e.g.): the uncategorized data form a mixture of unknown proportions or weights, weights to be estimated, while the categorized ones can offer initial estimates of the parameter values, θ_j, $j = 1, \dots, k$. McLachlan and Basford (1988, Chapter 4) describe various practical settings where one has some knowledge of the components F_j in (1.1) and is interested in estimating the proportions π_j. In particular, they quote Choi (1979) who proposes mixture models in the case of two mutually exclusive causes of failure (competing risks).

In the medical (lifetime) context it is typical for the data to be censored, i.e., often one observes (T_i, δ_i) where, for each $i = 1, \dots, n$, $T_i = \min(Y_i, C_i)$, with C_i a censoring time independent of Y_i, and δ_i an indicator variable such that

$$\delta_i = \begin{cases} 1 & \text{if } Y_i \leq C_i \\ 0 & \text{if } Y_i > C_i. \end{cases}$$

Keywords and phrases: finite mixture; minimum distance estimator; censored data; robustness.

In the context of finite mixture models, the important monograph by Titterington et al. (1985) and the very recent one by McLachlan and Peel (2000) give both the theoretical and the applied background in the field. In particular, McLachlan and Peel (2000, Chapter 10) present mixture models for failure-time data, with emphasis on mixture models for competing risks. A monograph on mixtures which discusses many practical issues is McLachlan and Basford (1988). Titterington et al. (1985) describe in detail maximum likelihood (Section 4.3) and minimum distance estimation based on the distribution function (Section 4.5). They consider either estimation of the weights π_j only, or of both weights and parameters θ_j, discuss various algorithms, and give the asymptotic properties of these estimators (which are quite similar, with the maximum likelihood estimator being the most efficient at the true model). In maximum likelihood estimation, the EM algorithm is a very flexible tool and is widely used; on the other hand, as pointed out in McLachlan and Basford (1988, Section 1.7), it is very slow and quite sensitive to a poor choice of initial values. When it comes to estimating the weights only, minimum distance estimation based on quadratic distances (like the one considered in this paper) is easy to compute and converges fast. Moreover, minimum distance estimation is considered to be more robust than maximum likelihood, and this seems to be also true when estimating weights in finite mixtures (see, e.g., Woodward et al., 1984, who treated the case of normal mixtures). Thus, it may be of interest and practical worth to use minimum distance estimation in order to obtain good starting values for the EM algorithm, an idea implicit in McLachlan and Basford (1988, Section 4.1). In other words, one could exploit the merits of both estimation methods, and combine them: first use minimum distance, as a simple, fast and robust estimator of the weights, when a first "guess" of the parameters θ_j is available, and further use these estimates of π_j and θ_j, $j = 1, \ldots, k$ as starting values in the EM algorithm. (A good "guess" can be an estimate from former studies, training sample, categorized data in the actual sample, etc.) One objective of this paper is to check empirically how good is a minimum distance estimate of the weights π_j, $j = 1, \ldots, k$ when an initial estimate of θ_j (of course different from the true value) is available. Another objective of the paper is to check the influence of censoring.

2. The method

Choi and Bulgren (1968) propose and study a minimum distance estimator of the weights π_j, $j = 1, \ldots, k$ in the finite mixture (1.1) where k is known, and the cumulative distribution functions (c.d.f.) $F_j(t)$, $j = 1, \ldots, k$ are known. If the ordered data are $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, let $\widehat{H}$ be the empirical distribution function, i.e., $\widehat{H}(t) = i/n$, $t_{(i-1)} \leq t < t_{(i)}$ and 0 otherwise. For fixed t, $\widehat{H}(t)$ is an unbiased, consistent and asymptotically normal estimator of the mixture distribution, $H(t)$. The Choi and Bulgren method consists of minimizing, for $G = \{\pi_j, j = 1, \dots, k\}$,

$$Q_n(G) = \int \{P_G(t) - \widehat{H}(t)\}^2 \, d\widehat{H}(t) = \sum_{i=1}^{n} d_i \Big\{ \sum_{j=1}^{k} \pi_j F_j(t_{(i)}) - \widehat{H}(t_{(i)}) \Big\}^2 = \sum_{i=1}^{n} d_i \Big\{ \sum_{j=1}^{k} \pi_j F_j(t_{(i)}) - \frac{i}{n} \Big\}^2, \tag{2.1}$$

where $d_i = d\widehat{H}(t_{(i)}) = 1/n$, $i = 1, \dots, n$. For the mixture weights π_j, $j = 1, \dots, k$ we have to impose the constraints

$$\pi_j \geq 0, \tag{2.2a}$$

and

$$\sum_{j=1}^{k} \pi_j = 1. \tag{2.2b}$$

Quadratic programming algorithms are available for this minimization under constraints problem (see, e.g., Wolfe, 1962). As noted in Titterington et al. (1985, Chapter 4) if one is prepared to risk the possibility of negative weights, the problem becomes a nonrestricted one, since condition (2.2b) can be easily replaced by an unrestricted estimation of the $(k-1)$-dimensional vector $(\pi_1, \dots \pi_{k-1})$, followed by $\widehat{\pi}_k = 1 - \sum_{j=1}^{k-1} \widehat{\pi}_j$. Lavigne (1995) used Wolfe's minimization under constraints algorithm. In this paper we report results based on unrestricted estimation (see Section 3 for details). A variant of the distance (2.1) is proposed in Macdonald (1971), who suggests replacing i/n with $(i - 1/2)/n$. (Woodward et al. (1984) use the same distance.) For uncensored data and finite samples Macdonald (1971) brings empirical evidence that replacing i/n with $(i-1/2)/n$ reduces the bias, but, obviously, the asymptotic properties of the estimators based on one distance or the other are the same. Further discussion of this point is detailed in Section 3.

Firstly, we propose to adapt the estimator proposed by Choi and Bulgren (1968) to the case where the data are censored. Under censoring, some values $t_{(i)}$ are censored and, therefore, one has to replace both $\widehat{H}(t)$ and the jumps $1/n$. The natural approach is to consider the Kaplan–Meier estimator of the survival function $S(t) = 1 - H(t)$, i.e., to compute the step function

$$\widetilde{S}(t) = \prod_{t_{(r)} \leq t} \left[\frac{(n-r)}{(n-r+1)} \right]^{\delta_{(r)}} \tag{2.3}$$

where $\delta_{(r)}$ is the indicator variable corresponding to $t_{(r)}$. Then in (2.1) one can replace $\widehat{H}(t)$ with $\widetilde{H}(t) = 1 - \widetilde{S}(t)$. Note that the weights d_i, $i = 1, \ldots, n$ are variable in this case, since they are the irregular jumps of the step function $\widetilde{H}(t)$. Indeed, both the location and the size of the jumps of $\widetilde{H}(t)$ depends on the particular pattern of censored values and the jumps are different from 0 only at values $t_{(r)}$ which are not censored.

Secondly, we propose to investigate the case where the cumulative distribution functions $F_j(t)$ are known up to parameters θ_j, $j = 1, \ldots, k$ which may need also to be estimated, i.e., we suppose that $F_j = F(t, \theta_j)$ with θ_j, $j = 1, \ldots, k$ unknown. On the other hand, rather than proposing simultaneous estimation of π_j and θ_j, $j = 1, \ldots, k$, as considered in Choi (1969), we assume that one has obtained a preliminary estimate of the unknown parameters θ_j, $j = 1, \ldots, k$, from training, and not mixed, samples (as is often the case in practice). With this preliminary estimate in hand, one can proceed to an initial estimate of the weights only, by treating θ_j as known, and letting $\theta_j \approx \widehat{\theta}_j$.

To summarize: we propose to study how the estimates $\widehat{\pi}_j$, $j = 1, \ldots, k$ are affected by the error in the initial estimation of θ_j, and also how sensitive $\widehat{\pi}_j$ are to the amount of censoring.

3. Simulation study

For our computer simulations, we retain two types of survival distributions which are popular in practice and quite flexible: the Weibull (1939, 1951) and the Gompertz (1825), whose densities depend on two parameters. The Weibull distribution was initially used for modelling failure times in engineering (strength of materials), while the Gompertz distribution is extremely common in actuarial studies (for modelling mortality). Over time, both distributions have found wide applications in the medical context as well. The Weibull distribution is found in any standard textbook on survival data. In mixture models, the Gompertz distribution was used by Gordon (1990). Recent research involving data analyses combining the Gompertz and the Weibull distributions is cited in McLachlan and Peel (2000, Section 10.3.3).

The Weibull distribution has density and hazard rate given by (respectively)

$$f(t) = (\lambda^\gamma \gamma) t^{\gamma-1} \exp(-(\lambda t)^\gamma), \; r(t) = (\lambda^\gamma \gamma) t^{\gamma-1} = (\lambda^\gamma \gamma) \exp[(\gamma - 1)\log(t)], \quad \lambda > 0, \gamma > 0,$$

while the Gompertz distribution has density and hazard rate given by (respectively)

$$f(t) = \lambda^* \exp(\gamma^* t) \exp\left[\frac{\lambda^*}{\gamma^*}(1 - \exp(\gamma^* t))\right], \; r(t) = \lambda^* \exp(\gamma^* t), \qquad \lambda^* > 0, \gamma^* \in \mathbb{R}.$$

Our parametrization of the Gompertz distribution is slightly different from the standard one; usually the parameters are $(\gamma^*, \widetilde{\lambda})$ with $\widetilde{\lambda} = \log(\lambda^*)$. Note that the hazard rates are increasing for $\gamma > 1$ (respectively $\gamma^* > 0$). Moreover, when $\gamma > 1$, the Weibull has a unique mode (local maximum) at the positive value $t_0 = [(\gamma - 1)/\gamma]^{1/\gamma}/\lambda$. Further, we choose the parameters so that they satisfy the following set of relations:

$$\lambda^* = (\lambda^\gamma \gamma), \quad \gamma^* = (\gamma - 1), \quad \gamma - 1 > 0, \quad \frac{\gamma^*}{\lambda^*} > 1. \tag{3.1}$$

If $\gamma - 1 > 0$ and $\gamma^* = \gamma - 1$, both the Gompertz and the Weibull have an increasing hazard rate. The last inequality in (3.1) ensures that the unique mode of the Gompertz density is $t_0 = \log(\gamma^*/\lambda^*)/\gamma^* > 0$. The first two equalities in (3.1) ensure that the hazard rates have the same parameters. However, the Weibull progresses at a slower rate (as a polynomial in t) while the Gompertz increases exponentially in t. In what follows, we consider that the data come from various mixtures of two components of the same type, i.e., both Weibull or both Gompertz. The components have same parameter γ (respectively γ^*) and differ only in λ (respectively λ^*). We use $\pi = \pi_1$, $1 - \pi = \pi_2$, and $\pi F_1 + (1 - \pi) F_2$. Such mixtures are identifiable (see Titterington et al., 1985 and references therein, Gordon, 1990).

For our simulations we took:

(a) three choices of pairs of weights $(\pi, 1 - \pi)$, given by $\pi = 0.3$, 0.5, 0.7;

(b) for each weight π, a mixture of

  (i) two Weibull distributions, with $(\gamma = 3, \lambda_1 = 0.2)$ and $(\gamma = 3, \lambda_2 = 0.5)$, and

  (ii) two Gompertz distributions, with $(\gamma^* = 2, \lambda_1^* = 0.024)$ and $(\gamma^* = 2, \lambda_2^* = 0.375)$.

These parameters are such that (3.1) is satisfied. Moreover, the modes of the components are well separated, and $F_1(t) < F_2(t)$ for all $t > 0$. The modes are $(4.3679, 1.7471)$ for the Weibull and $(2.211424, 0.8369)$ for the Gompertz. The graphs of the six mixtures are given in Figure 1. We assumed that censoring might be present, and we simulated data from the 3×2 models considered above with 0%, 20%, 40%, and 60% censoring and exponential censoring times, C_i. Lavigne (1995) explains how such censoring can be simulated when C_i follows an exponential distribution, and how the appropriate parameter of the censoring time C_i should be computed (the parameter depends both on the mixture and the amount of censoring).

Once the data were generated, we estimated the weight π by minimizing the (unconstrained) distance

$$Q_n(\pi) = \sum_{i=1}^{n} \{[\pi F_1(t_{(i)}) + (1-\pi)F_2(t_{(i)})] - \widetilde{H}(t_{(i)})\}^2 \, d_i, \tag{3.2}$$

where d_i, $i = 1, \ldots, n$ is the jump of the Kaplan–Meier step function $\widetilde{H}(t)$ at $t_{(i)}$. We kept track of how many cases were out of bounds (i.e., that gave $\widehat{\pi} = 0$ or 1). F_j were either both Weibull with parameters $(3, \lambda_j)$ or both Gompertz with parameters $(2, \lambda^*{}_j)$, $j = 1, 2$.

We supposed λ_j and λ_j^* to be estimated as well, and we took, for each mixture:

(i) Weibull: $\widehat{\lambda}_1 = 0.2$, 0.22, 0.18, $\widehat{\lambda}_2 = 0.5$, 0.55, 0.45;

(ii) Gompertz: $\widehat{\lambda}_1^* = 0.024$, 0.0216, 0.0264, $\widehat{\lambda}_2^* = 0.375$, 0.3375, 0.4125.

In other words, first we considered the case where the parameters λ_j and λ_j^*, $j = 1, 2$ were known. Furthermore, we checked the robustness of $\widehat{\pi}$ when λ_j and λ_j^* were estimated, and we took a relative error of 10%, i.e., we assumed $|\widehat{\lambda}_j - \lambda_j|/\lambda_j = 10\%$ (same relative error for λ_j^*). In what follows we refer to these estimates of λ_j and λ_j^* as perturbed parameters. In all cases γ and γ^* are known. Note that in our examples $F_1 < F_2$, where F_1 is the component of weight π.

For each mixture, and each amount of censoring (12 cases for each probability distribution, 24 cases in all), we simulated $N = 1000$ samples of size $n = 100$ each and computed the mean $\overline{\widehat{\pi}}$ of the estimated $\widehat{\pi}_k$, $k = 1, 2, \ldots, N$, as well as the mean squared error $\text{MSE} = \sum_{k=1}^{N} (\widehat{\pi}_k - \pi)^2 / N$.

The results are given in Table 1 and Table 2, each composed of 3 subtables (one for each value of π). In the simulations we used the combined random number generators of L'Écuyer (1988), and the same data in each subtable. Note that the results for the Gompertz are better than those for the Weibull, which is not surprising: just look at their graphs in Figure 1. We do not report here the proportion of time with $\widehat{\pi}$ out of bounds, since there were very few (in most subtables none). In computing the mean $\overline{\widehat{\pi}}$ we set $\widehat{\pi} = 0$ whenever $\widehat{\pi} \leq 0$ and $\widehat{\pi} = 1$ whenever $\widehat{\pi} \geq 1$.

The way to read the tables is the following: the first row gives the effect of censoring only, while the first column gives the effect of perturbing the parameters λ_j when no censoring is present. The other entries give the

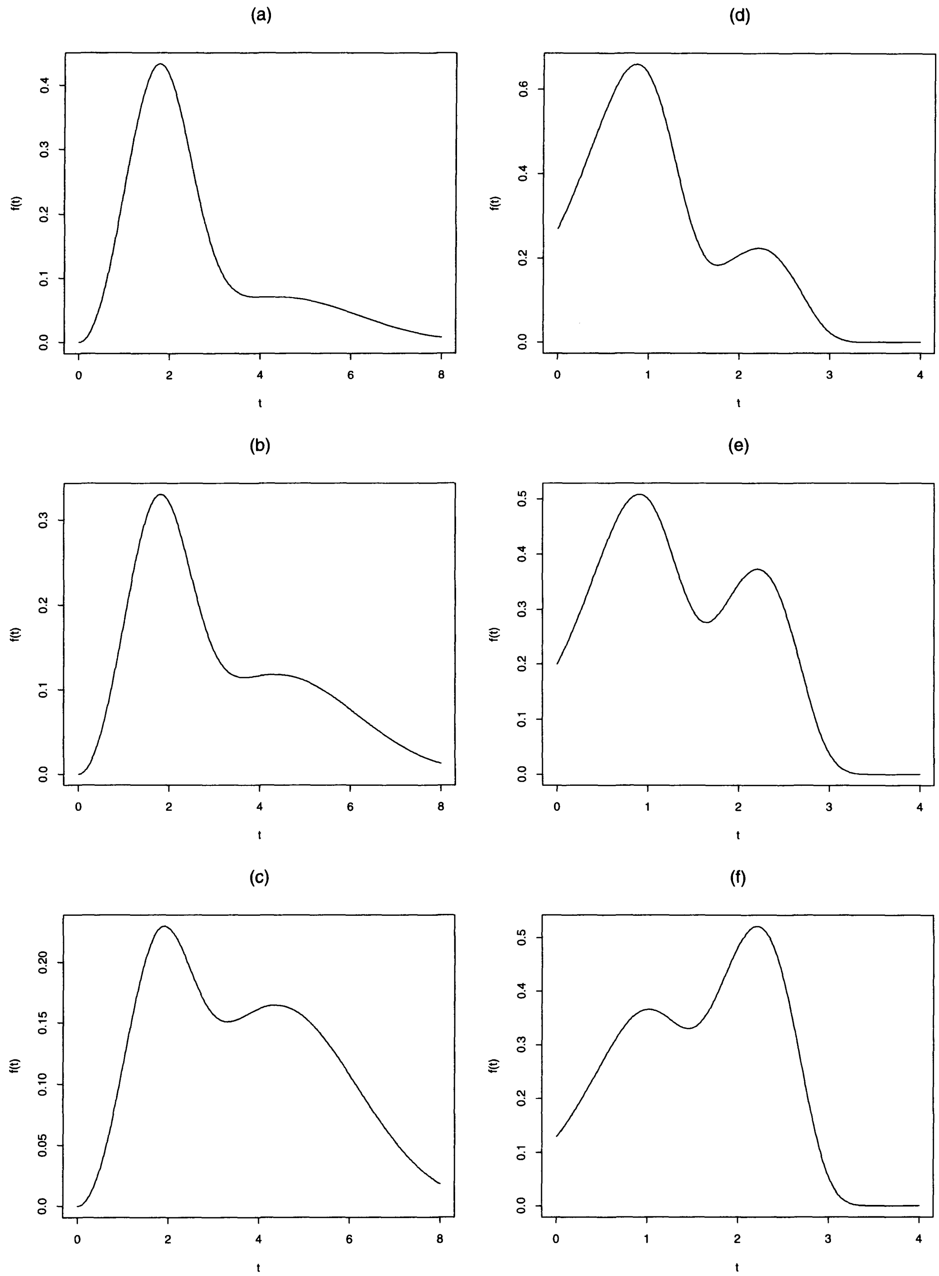

Figure 1. Graphs of the mixture densities. Weibull: (a) $\pi = 0.3$, (b) $\pi = 0.5$, (c) $\pi = 0.7$. Gompertz: (d) $\pi = 0.3$, (e) $\pi = 0.5$, (f) $\pi = 0.7$.

combined effect of censoring and perturbation. In what follows we discuss how censoring, perturbation, and the interaction of both affect the quality of the estimation.

Table 1. Results for the Weibull mixture $\pi f_W(x \mid 3; 0.2) + (1-\pi) f_W(x \mid 3; 0.5)$

(a) $\pi = 0.3$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE
0.2, 0.5	0.2890	0.004065	0.2860	0.004826	0.2842	0.005486	0.2636	0.007825
0.18, 0.45	0.1782	0.019452	0.1750	0.021058	0.1725	0.022409	0.1497	0.029907
0.18, 0.55	0.3380	0.003895	0.3352	0.004102	0.3333	0.004319	0.3164	0.004188
0.22, 0.45	0.2032	0.016286	0.1999	0.018149	0.1985	0.019504	0.1743	0.026637
0.22, 0.55	0.3845	0.010599	0.3816	0.010720	0.3805	0.011038	0.3628	0.009481

(b) $\pi = 0.5$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE
0.2, 0.5	0.4902	0.004381	0.4874	0.005213	0.4881	0.005682	0.4644	0.008604
0.18, 0.45	0.3938	0.015670	0.3906	0.017202	0.3908	0.017634	0.3643	0.026144
0.18, 0.55	0.4916	0.002710	0.4891	0.003232	0.4889	0.003547	0.4692	0.005466
0.22, 0.45	0.4763	0.008071	0.4728	0.009621	0.4745	0.010388	0.4449	0.016181
0.22, 0.55	0.5755	0.009858	0.5732	0.010225	0.5741	0.010830	0.5530	0.009827

(c) $\pi = 0.7$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE	$\bar{\hat{\pi}}$	MSE
0.2, 0.5	0.6911	0.004052	0.6898	0.004248	0.6839	0.005236	0.6677	0.008394
0.18, 0.45	0.5909	0.015493	0.5896	0.015956	0.5831	0.018206	0.5660	0.024810
0.18, 0.55	0.6450	0.005420	0.6440	0.005662	0.6384	0.006788	0.6249	0.010100
0.22, 0.45	0.7453	0.008965	0.7433	0.009114	0.7370	0.010130	0.7165	0.013031
0.22, 0.55	0.7799	0.010661	0.7784	0.010632	0.7734	0.010743	0.7579	0.011024

First consider the effect of censoring only, as reflected in the first row of each subtable. As expected, the estimates are worse when censoring increases. Still, the estimates are surprisingly robust to censoring, as they don't change much up to 40% censoring. In all cases there is a drop in quality at 60% censoring but the relative error is never bigger than 13%.

As far as censoring is concerned the really intriguing result is the tendency for the estimates to decrease with the amount of censoring. We propose a simple and general heuristics as follows. Under censoring, rather than having data from the original c.d.f., F_Y, the data come from $T = \min(C, Y)$, of survival function $1 - F_T$, where

$$1 - F_T = (1 - F_Y)(1 - F_C) < (1 - F_Y) \iff F_T > F_Y = \pi F_1 + (1 - \pi) F_2.$$

Table 2. Results for the Gompertz mixture $\pi\, f_G(x \mid 2; 0.024) + (1-\pi)\, f_G(x \mid 2; 0.375)$

(a) $\pi = 0.3$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE
0.024, 0.375	0.2890	0.004065	0.2860	0.004857	0.2821	0.005471	0.2592	0.008267
0.0216, 0.3375	0.2525	0.006416	0.2495	0.007483	0.2452	0.008475	0.2208	0.013352
0.0216, 0.4125	0.3073	0.003397	0.3044	0.003956	0.3007	0.004339	0.2794	0.005923
0.0264, 0.3375	0.2645	0.005986	0.2614	0.007094	0.2573	0.008033	0.2325	0.012597
0.0264, 0.4125	0.3215	0.004230	0.3185	0.004790	0.3150	0.005117	0.2933	0.006255

(b) $\pi = 0.5$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE
0.024, 0.375	0.4902	0.004381	0.4874	0.005146	0.4881	0.005714	0.4612	0.008884
0.0216, 0.3375	0.4582	0.006074	0.4552	0.007058	0.4558	0.007601	0.4279	0.012726
0.0216, 0.4125	0.4908	0.003709	0.4882	0.004354	0.4885	0.004843	0.4633	0.007571
0.0264, 0.3375	0.4862	0.005301	0.4832	0.006241	0.4843	0.006904	0.4554	0.010849
0.0264, 0.4125	0.5190	0.004608	0.5163	0.005196	0.5171	0.005795	0.4911	0.007333

(c) $\pi = 0.7$

Censoring	0%		20%		40%		60%	
$\hat{\lambda}_1,\ \hat{\lambda}_2$	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE	$\bar{\tilde{\pi}}$	MSE
0.024, 0.375	0.6911	0.004052	0.6898	0.004293	0.6841	0.005205	0.6674	0.008445
0.0216, 0.3375	0.6578	0.005630	0.6564	0.005962	0.6505	0.007253	0.6335	0.011635
0.0216, 0.4125	0.6742	0.004007	0.6729	0.004261	0.6674	0.005214	0.6517	0.008527
0.0264, 0.3375	0.7068	0.004792	0.7053	0.005022	0.6994	0.005937	0.6816	0.009150
0.0264, 0.4125	0.7211	0.004533	0.7197	0.004689	0.7143	0.005292	0.6979	0.007500

Had we used the empirical distribution function, and had we tried to fit a mixture to data, we would have looked for the weight $\tilde{\pi}$ which satisfies the "equation" $F_T \approx \tilde{\pi} F_1 + (1-\tilde{\pi}) F_2$. In our case $F_1 < F_2$ and the estimated mixture would have put more weight on F_2 (and less weight on F_1) than F_Y does, because $F_T > F_Y$; therefore we would have obtained $\tilde{\pi} < \pi$. On the other hand, we do not use the empirical distribution function (which would have estimated F_T rather than F_Y) but the Kaplan-Meier estimator, which is supposed to correct for the effect of censoring but whose precision is affected by the amount of censoring. To summarize: if one has to fit a mixture of ordered c.d.f's to censored data, and therefore issued from F_T, one will necessarily end up in putting somewhat less weight on the smallest distribution function.

On the other hand, consider only the first entry in each subtable, i.e., 0% censoring (and true values λ_j or λ_j^*). At first sight one may wonder why π is *always* underestimated. This may be due to the fact that $\widehat{H}(t_{(i)}) = i/n$ and i/n is quite high for i close to n, in particular $\widehat{H}(t_{(n)}) = 1$, while the true value may be well under 1. Then the best fit puts more weight on the higher component, F_2, in order to lift the mixture up to $\widehat{H}(t_{(i)})$. In order to check if this negative bias can be reduced, we adapted Macdonald's (1971) idea mentioned in Section 2.1 and repeated some simulations with a slightly different Kaplan–Meier estimate, i.e., we used

$$\widetilde{S}_{\mathrm{mod}}(t) = \prod_{t_{(r)} \leq t} \left[\frac{(n-(r-1/2))}{(n-(r-1/2)+1)} \right]^{\delta_{(r)}}.$$

By replacing $\widetilde{S}$ with $\widetilde{S}_{\mathrm{mod}}$ we improved the estimate at the true values, but not necessarily at the perturbed ones. We do not report the results here.

Next, look at the first column in each subtable and consider the effect of perturbing the parameters (uncensored data). In this case not all results are that good, mainly for the Weibull mixture. Indeed, for the Weibull, the components are less well separated (see Figure 1). For both distributions we get the worst results for $\pi = 0.3$, i.e., when the component with the smallest distribution function receives less weight (and again the components are less well separated). Recall that we set a relative error of 10% for the estimates of λ_j and λ_j^*, $j = 1, 2$. For the Gompertz mixture, the relative errors of the estimates of π are below 10%, as all errors but one vary between 3% and 8.36%, which is good. The MSE is strictly below 0.007. In the case of the Weibull, the relative errors are much bigger as they generally vary between 11% and 28% with one exception where there is a sharp increase, namely a relative error of 40%. For the Weibull, the MSE is strictly below 0.02. In all, these values of the MSE are much lower than those reported by some other authors for the same sample sizes, even when the components are known (see, e.g., Choi and Bulgren, 1968). Still, in spite of the low values of the MSE, the great variations in the relative errors suggest that our study is not conclusive as far as robustness to perturbation is concerned.

To continue our analysis of perturbation, note that in both types of mixture the worst case occurs when both components are underestimated. In this case we obtain estimates of π which are smaller than the true values, a fact explained below (other cases can be treated in a similar way). Let $\widetilde{F}_j$, $j = 1, 2$ be the c.d.f. with perturbed parameters, and assume $\widetilde{F}_j < F_j$, $j = 1, 2$. In our examples $\widetilde{F}_1 < \widetilde{F}_2$. When we apply the minimization we fit a new weight, $\widetilde{\pi}$ say, and we look to "solve" in $\widetilde{\pi}$

$$\widetilde{\pi}\widetilde{F}_1 + (1-\widetilde{\pi})\widetilde{F}_2 \approx \pi F_1 + (1-\pi)F_2.$$

Since $\pi\widetilde{F}_1 + (1-\pi)\widetilde{F}_2 < \pi F_1 + (1-\pi)F_2$ and $\widetilde{F}_1 < \widetilde{F}_2$, we have to put more weight on $\widetilde{F}_2$ in order to obtain the desired "equality" of mixed cumulative distribution functions. In other words, the solution $\widetilde{\pi}$ must be such that $1-\widetilde{\pi} > 1-\pi$, or $\widetilde{\pi} < \pi$. The underestimation of π is more or less drastic depending on the probability distribution and on the parameter values.

As a last step in our analysis, consider the case where the parameters are perturbed and at the same time the data are censored. In these cases it seems hard to assess the quality of the estimation, simply because the effects of censoring and perturbation may cancel each other. At times we seem to get better estimates of π for the perturbed rather than the original parameters (in the same column). Still, most such anomalies occur with heavy censoring only (i.e., 60%) and the reduction in the MSE is of a few thousandths at most, which is quite small.

In view of all this, what can we make of our study? First, minimum distance estimates appear very robust with respect to censoring, but somewhat less with respect to the perturbation in the parameters λ_j or λ_j^*, $j=1,2$. When it comes to censoring, the study seems to point out some simple rules of thumb, e.g., that even 40% censoring can give very good results. The robustness with respect to censoring may be inherited from the Kaplan-Meier estimator, and this good feature suggests to use minimum distance in its own right for censored data (when components are known). On the other hand, it would be hard to state similar rules of thumb when it comes to replacing the true values λ_j or λ_j^*, $j=1,2$ with estimates. Indeed, in terms of relative error, the quality of the weight estimates appears to depend too much on the probability distribution and on the direction of the perturbation (over or underestimate). As far as the MSE is concerned, it was quite low for the perturbations considered in this paper, and this good behaviour of the MSE is an indicator of robustness. But, in all, further study may be needed before drawing definite conclusions on how robust is the estimation of the weights when one replaces the parameters of the components with estimates.

Acknowledgements. This research received financial support from NSERC. The authors thank the referees for very helpful suggestions which improved the presentation. The starting point of this paper is the M.Sc. thesis of J. Lavigne. The authors are happy to be able to contribute to the volume in honour of Constance van Eeden, a distinguished researcher and an exceptional graduate studies supervisor.

REFERENCES

Choi, K. (1969). Estimators for the parameters of a finite mixture of distributions. *Ann. Inst. Statist. Math.* 21, 117–125.

Choi, K. and Bulgren, W.B. (1968). An estimation procedure for mixtures of distributions. *J. Roy. Statist. Soc. Ser.* B 30, 444–460.

Choi, S.C. (1979). Two sample tests for compound distributions for homogeneity of mixing proportions. *Technometrics* 21, 361–365.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philos. Trans. Roy. Soc. London* 115, 513–583.

Gordon, N.H. (1990). Application of the theory of finite mixtures for the estimation of "cure" rates of treated cancer patients, *Statist. Med.* 9, 397-407.

Lavigne, J. (1995). *Estimation à distance minimum dans les lois de mélange.* M.Sc. Thesis, Université du Québec à Montréal, Montréal.

L'Écuyer, P. (1988). Efficient and portable combined random number generators. *Comm. ACM* 31, 742–749.

Macdonald, P.D.M. (1971). Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren, *J. Roy. Statist. Soc. Ser.* B 33, 326–329.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models. Inference and Applications to Clustering*, Statist. Textbooks Monogr., vol. 84, Marcel Dekker, New York.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, Wiley Ser. Probab. Statist. Appl. Probab. Statist., Wiley-Interscience, New York.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley Ser. Probab. Math. Statist. Appl. Probab. Statist., John Wiley, Chichester.

Weibull, W. (1939). A statistical theory of the strengths of materials, *Ing. Vetenskaps Akad. Handlingar*, No. 151, The phenomenon of rupture in solids, 293–297.

Weibull, W. (1951). A statistical distribution of wide applicability. *J. Appl. Mech.* 18, 293–297.

Woodward, W.A., Parr, W.C., Schucany, W.R., and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *J. Amer. Statist. Assoc.* 79, 590–598.

Wolfe, P. (1962). Recent developments in nonlinear programming, *Advances in Computers*, Vol. 3, Academic Press, New York, pp. 155–187.

René Ferland
Département de mathématiques
Université du Québec à Montréal
C.P. 8888, succ. Centre-ville
Montréal, QC H3C 3P8
Canada ferland@math.uqam.ca

Sorana Froda
Département de mathématiques
Université du Québec à Montréal
C.P. 8888, succ. Centre-ville
Montréal, QC H3C 3P8
Canada froda@math.uqam.ca

Jean Lavigne
MDS Pharma Services
2350, Cohen
Montréal, QC H4R 2N6
Canada
jean.lavigne@mdsps.com