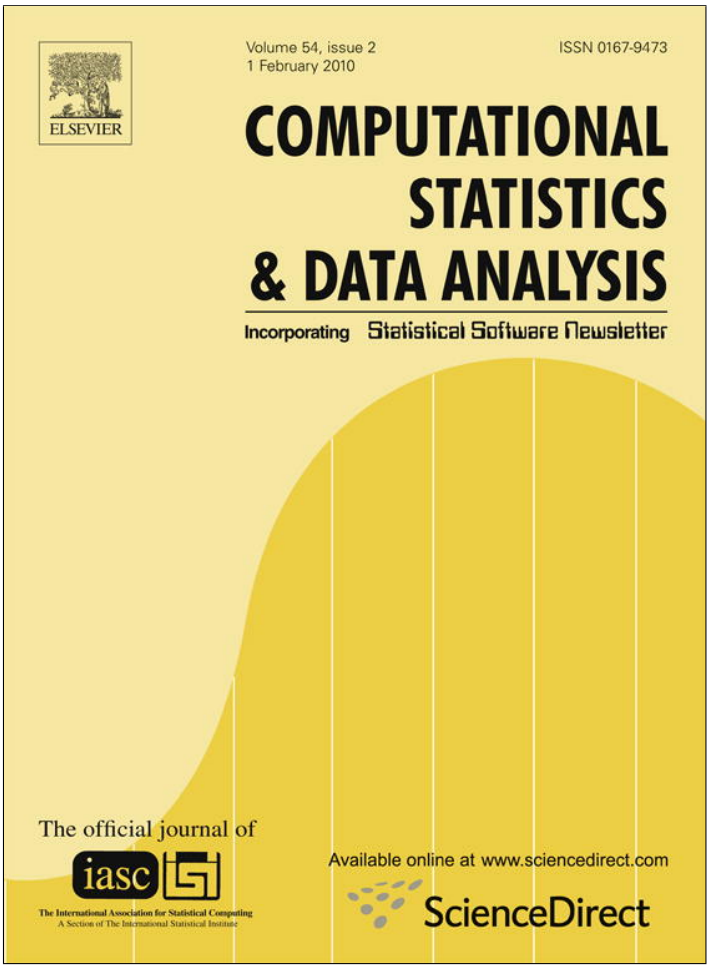

Computational Statistics and Data Analysis 54 (2010) 416–428

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A two-component Weibull mixture to model early and late mortality in a Bayesian framework

Alessio Farcomeni [a,*], Alessandra Nardi [b]

[a] University of Rome "La Sapienza", piazzale Aldo Moro, 5 00186 Roma, Italy
[b] University of Rome "Tor Vergata", via della Ricerca Scientifica, 00133 Roma, Italy

ARTICLE INFO

Article history:
Received 21 December 2008
Received in revised form 4 September 2009
Accepted 5 September 2009
Available online 12 September 2009

ABSTRACT

A two-component parametric mixture is proposed to model survival after an invasive treatment, when patients may experience different hazards regimes: a risk of early mortality directly related to the treatment and/or the treated condition, and a risk of late death influenced by several exogenous factors. The parametric mixture is based on Weibull distributions for both components. Different sets of covariates can affect the Weibull scale parameters and the probability of belonging to one of the two latent classes. A logarithmic function is used to link explanatory variables to scale parameters while a logistic link is assumed for the probability of the latent classes. Inference about unknown parameters is developed in a Bayesian framework: point and interval estimates are based on posterior distributions, whereas the Schwarz criterion is used for testing hypotheses. The advantages of the approach are illustrated by analyzing data from an aorta aneurysm study.

1. Introduction

In some survival studies, in particular in clinical trials characterized by invasive treatments, patients experience a very high risk of early, post-operative death followed by several years of reduced, but still not negligible, risk of late mortality. Examples include studies of survival after organ transplantation and situations where an emergency surgery is needed such as the aorta aneurysm study discussed in Section 6. In these cases it is common practice to base the statistical analysis for the comparison of treatments and for the identification of prognostic factors on the Cox model (Cox, 1972). However, the Cox model may not be the best choice in such a multiple-component risk setting. Different mechanisms often underpin early and late events, and different exogenous factors can affect each component differently. Even the same explanatory variable might have a different impact on short and long term survival. This implies a violation of the proportionality assumption underlying the Cox model, leading to concerns regarding the robustness of inference with respect to the shape of the baseline. Nor the implementation of separate analyses should be expected to provide more satisfactory results, since inference for late mortality is often complicated by the limited number of patients still at risk and by the possibility of serious selection bias. A suitable mixture model lends itself as a more appropriate solution, allowing simultaneous inference for the potentially different effects of different explanatory variables. For an introduction to mixture models in survival, see Farewell (1982) and Mc Lachlan and Mc Giffin (1994) (see also Greenhouse and Silliman (1996) and Yu and Peng (2008)). Pack and Morgan (1990) use a similar mixture model for the analysis of quantal assay data and Pocock et al. (1982) use it for the analysis of curability of breast cancer. A review of survival analysis using gene expression levels as covariates is given in van Wieringen et al. (2009).

* Corresponding author. Tel.: +39 06 49974679; fax: +39 06 49970103.
E-mail addresses: alessio.farcomeni@uniroma1.it (A. Farcomeni), alenardi@mat.uniroma2.it (A. Nardi).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.09.007

There are many possible choices of mixture models for the analysis of survival data (see Section 3). However the specialised literature focuses on semi-parametric models while the fully parametric approach has received less attention. Our goal here is to show that a simple parametric mixture model with a small number of parameters is flexible enough to satisfactorily accommodate the observed data, thus avoiding instability or even identifiability problems. We propose a two-component parametric mixture with early and late mortality modelled by two Weibull families of distributions. Extension of the model to the general case of $p \geq 2$ mixing components is straightforward, but in our experience seldom results in a significantly improved fit. The approach is demonstrated on data from an original study on aorta aneurysm. Results from the proposed mixture model are compared with those from a Cox model. Although this comparison is complicated by the intrinsically different modelling frameworks, we believe it highlights some important issues. Based in particular on the case-study illustrated in Section 6, it emphasises that a Cox semi-parametric modelling approach does not provide a satisfactory fit to the data at hand.

The outline of the paper is as follows. Section 2 presents a partial characterization of the hazard function of a mixture of two Weibull distributions. In Section 3 we give further details on the proposed model and in Section 4 we describe an ad hoc Gibbs sampling scheme. An approximation to the Bayes factor for hypothesis testing is discussed in Section 5. An application to an aorta aneurysm study is given in Section 6 whereas Section 7 reports some final remarks. Technical details are deferred to the Appendix.

2. Mixtures of two Weibull distributions

The choice of any parametric family of distributions should be motivated by prior assumptions on the possible shape of the hazard functions. For example, in the study of aorta aneurysm, the occurrence of an emergency surgical treatment naturally leads to an increased chance of early death, likely by several years of lower and stable, yet appreciable hazard. This scenario is examined and validated through the estimates (and tests) reported in Section 6. The presence of a monotone increasing or decreasing risk of late mortality may also be considered, including a possible second peak in the hazard. On the basis of these considerations we propose to represent such a pattern in the hazard rate via a mixture of two Weibull distributions. To give the reader an idea about the flexibility, but also the limitations, of the proposed parametric model, we present a partial characterization of the corresponding hazard function. For the sake of simplicity, we assume no censoring and homogeneous observations in this section. In the next section we will extend our model to a more general case.

Let T be a non-negative random variable denoting the failure time of interest. The survival function for a mixture of two Weibull distributions is

$$S(t) = pS_1(t) + (1-p)S_0(t), \tag{1}$$

with $0 \leq p \leq 1$ and

$$S_i(t) = \exp[-(\eta_i t)^{\gamma_i}]. \tag{2}$$

Such a mixture is characterized by five parameters: the shape γ_i and scale η_i parameters modelling early and late failure times and the parameter p defining the mixture. The corresponding hazard function can thus be expressed as

$$h(t) = w(t)h_1(t) + (1-w(t))h_0(t), \tag{3}$$

where $h_i(t) = \gamma_i \eta_i^{\gamma_i} t^{\gamma_i - 1}$ is the hazard function associated with $S_i(t)$ and

$$w(t) = p\frac{S_1(t)}{S(t)}. \tag{4}$$

Jiang and Murthy (1998) proved that for either small or large t the shape of $h(t)$ is similar to that for $h_1(t)$. As a consequence, $h(t)$ can be either increasing or decreasing for small and large t, excluding the possibility of a U-shape. The authors also showed that the failure rate of an n-fold mixture of Weibull distributions is never U-shaped.

The case of an either increasing or decreasing hazard can still be obtained by assuming $\gamma_1 = \gamma_0$, coupled with suitable values for the ratio η_1/η_0. However a sharply decreasing hazard followed by a slowly decaying tail can be more flexibly obtained by combining a decreasing risk function with a constant one. An example is shown in Fig. 1a. Alternative shapes for the hazard function can be obtained from the combination of a constant or decreasing hazard ($0 < \gamma_1 \leq 1$) with an increasing one ($\gamma_0 > 1$). The situation of a high initial peak of risk followed by an almost constant hazard is shown in Fig. 1b; interestingly this follows from assuming a constant hazard for the first component of the mixture in combination with an increasing one for the second component.

The presence of a second peak in the hazard function can be modelled combining a decreasing hazard for the first component with an increasing risk for the second one (Fig. 1c). The choice of an increasing shape for both the mixture components allows the risk of late death to increase as time increases (Fig. 1d). Clearly a variety of shapes for the hazard function can be obtained by suitable choices of the five parameters; a more detailed characterization can be found in Jiang and Murthy (1998). The paper is aimed at showing that a mixture of two Weibull distributions offers a flexible modelling approach to describe non-trivial hazard functions as that featured in the illustrated case-study, while still being characterized by a manageable number of parameters.

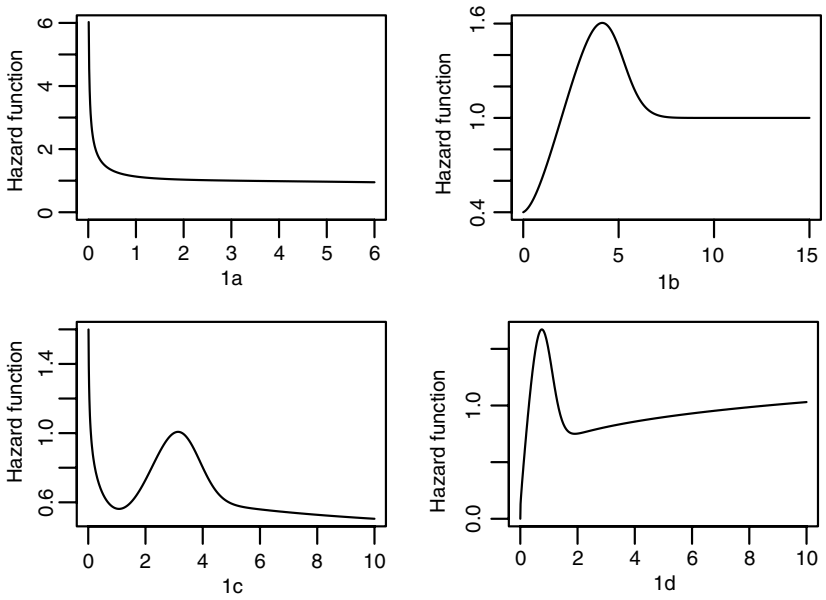

Fig. 1. Mixtures of two Weibull distributions assuming: 1(a) $\gamma_1 = 0.6$, $\gamma_0 = 1$, $\eta_1 = 5$, $\eta_0 = 1$, $p = 0.6$; 1(b) $\gamma_1 = 1$, $\gamma_0 = 2$, $\eta_1 = 1$, $\eta_0 = 0.5$, $p = 0.4$; 1(c) $\gamma_1 = 0.8$, $\gamma_0 = 3$, $\eta_1 = 1$, $\eta_0 = 0.4$, $p = 0.8$; 1(d) $\gamma_1 = 2$, $\gamma_0 = 1.2$, $\eta_1 = 1.5$, $\eta_0 = 0.6$, $p = 0.6$.

3. Extension to non-homogeneous and/or censored survival times

Typically, observations arising from medical research are not homogeneous and evaluating the effect of different factors on survival is one of the goals of the analysis.

Let $\mathbf{X}_j$ be a vector of d explanatory variables and prognostic factors observed on the jth patient in a sample of size n. Their effect on the risk of early and late death can be separately modelled by reparametrizing the scale parameters of two Weibull distributions as follows:

$$\log(\eta_i) = \beta_i^t \mathbf{X}_j \quad i = 0, 1. \tag{5}$$

This assumption corresponds to the classical structure of accelerated failure time models, where the effect of covariates is to accelerate or decelerate a baseline survival time T_0 by a factor $\exp(-\beta_i^t \mathbf{X}_j)$. Negative values of $\beta_i^t \mathbf{X}_j$ lead to a better survival prognosis while positive values result in an increased risk of failure.

A set of explanatory variables can also be assumed to affect the probability p of allocation to the first component of the mixture, i.e. in the case-study presented in Section 6 the probability to die of surgery-related causes. For the sake of simplicity we continue to use the notation $\mathbf{X}_j$. This effect can be modelled using a logistic link function of the form

$$p_j = \exp\{\alpha^t \mathbf{X}_j\}(1 + \exp\{\alpha^t \mathbf{X}_j\})^{-1}. \tag{6}$$

The classical assumption of independent and non-informative censoring completes the specification of the likelihood function.

It should be noted that the choice of Weibull hazard models for the components of the mixture is here motivated on the grounds of the simplicity and flexibility of such parametric family of distributions, as discussed in Section 2. While parametric (log-logistic, log-normal, etc.) or semi-parametric alternatives exist, these are not further considered din the paper. One may also use the generalized modified Weibull (Carrasco et al., 2008) or the lifetime distribution proposed by Tahmasbi and Rezaei (2008).

Furthermore, the number of components need not be confined to two. While extending the model to the general case is straightforward, in our experience the two Weibull mixture is already flexible enough. Use of different parametric choices for the two components makes interpretation of the parameter estimates harder, while semi-parametric choices may result in too much flexibility, and thus lack of model structure.

In order to draw inferences in a Bayesian framework, we must specify priors $\pi(\cdot)$ for the parameters in the model. If available, prior information should be summarized and properly taken into account when choosing the priors. The general approach we propose is to use independent zero-centered Normal priors for the α and β regression parameters, and Log-Normals or Inverse Gammas for the shape parameters γ. A prior sensitivity analysis can then be conducted on the posterior estimates (see Section 6).

4. Model formulation and ad hoc Gibbs strategy

In the following we use the short-hand notation θ for the vector of parameters.

Let Z_j be a latent indicator, with $Z_j = 0$ if the jth patient belongs to one population and $Z_j = 1$ if the jth patient belongs to the other. Let $P(Z_j = 1) = p$. Let δ_j be the event indicator, where $\delta_j = 0$ denotes a censored survival time. The complete data likelihood can then be written as

$$L(\theta) = \prod_{j=1}^{n} [pL_{1j}(\theta)]^{I(z_j=1)} [(1-p)L_{0j}(\theta)]^{I(z_j=0)}; \tag{7}$$

where

$$L_{ij}(\theta) = \exp[-(e^{\beta_i^t X_j} t_j)^{\gamma_i}](e^{\gamma_i \beta_i^t X_j} \gamma_i t_j^{\gamma_i - 1})^{\delta_j}, \tag{8}$$

and $I(\cdot)$ is the indicator function. If the mixture weight is assumed to depend on some explanatory variable, then $P(Z_i = 1|X_j) = p_j$ as in (6). In this case, p is simply substituted by its reparametrization in the complete data likelihood.

The model can be fit adapting an ad hoc Gibbs sampling scheme for mixture models; see Diebolt and Robert (1994) and the book by Robert and Casella (2000). In this scheme, the latent indicators are sampled from their posterior distributions, and the other parameters are sampled from the complete likelihood conditionally on the latent indicators. Let $\pi(\cdot)$ denote the prior for a specific parameter or vector of parameters. The general iteration of the Gibbs sampling scheme we use for the proposed mixture model is as follows:

1. Sample $Z_j, j = 1, \ldots, n$ from

$$\tilde{z}_j = P(Z_j = 1|\theta) = \frac{p_j L_{1j}(\theta)}{p_j L_{1j}(\theta) + (1-p_j)L_{0j}(\theta)}, \tag{9}$$

   where θ is as in the current iteration of the sampler;
2. Sample α from

$$\pi(\alpha|Z) \propto \pi(\alpha) \prod_{j=1}^{n} \frac{e^{I(Z_j=1)\alpha^t X_j}}{1 + e^{\alpha^t X_j}}.$$

   Compute $p = \{p_j, j = 1, \ldots, n\}$ as in (6).
3. Sample β_i and γ_i from

$$\pi((\beta_0, \beta_1, \gamma_0, \gamma_1)|T, Z) \propto L(\theta)\pi(\beta_0, \beta_1, \gamma_0, \gamma_1).$$

There are many difficulties associated with sampling the four Weibull parameters at Step 3. A Metropolis Hastings (MH) formulation is not directly available. Key to success for MH is linked to an efficient candidate transition kernel, which is not readily available here. Furthermore, the full conditional distribution is also potentially multi-modal, and even if a good candidate transition kernel were available tuning of the MH routine would be complicated by the presence of the second component in the mixture and volatility in Z. In order to avoid such difficulties, we sample the parameters in θ simultaneously by Adaptive Rejection Metropolis Sampling (ARMS), see Gilks et al. (1995). In order to keep the model identifiable and prevent any label switching problem we impose $\gamma_0 > \gamma_1$ at each iteration (see Jiang and Murthy (1998)). This is straightforward to do with ARMS.

Moreover we note that it is straightforward to check the conditional independence conditions featured in Step 2. Sampling from $\pi(\alpha|Z)$ can also be performed via ARMS, although there are many different possible alternative approaches for this standard problem. We finally note that in certain cases one may not wish to let the weight p depend on covariates. Then, a Beta(ν_1, ν_2) prior can be assumed for p. The full conditional distribution at Step 2 of the MCMC algorithm would then simply be a Beta with parameters $\nu_1 + \sum_{j=1}^{n} I(Z_j = 1)$ and $\nu_2 + \sum_{j=1}^{n} I(Z_j = 2)$.

5. The Schwarz criterion for hypothesis testing

A problem that naturally arises when using the proposed model is how to test hypotheses about the parameters. Hypotheses of interest may include

- whether there is actually a two-component structure;
- whether a covariate is significantly related to survival time;
- whether a covariate significantly affects the probability of being assigned to one or the other hazard regime.

The first problem can be formulated as a test of the hypothesis $H_0 : p = 0$, which would imply a single Weibull model if accepted. The second and third problems instead reduce to testing $H_0 : \beta_{ik} = 0$ for the kth covariate of interest, $i = 1, 2$, or $H_0 : \alpha_k = 0$ respectively.

The classical frequentist approach to testing defines a rejection region and reports associated (pre-inferential) error probabilities of incorrect rejection of the null and of the alternative. Of course, such error probabilities cannot be used as post-experimental evidence, since they do not reflect information given by data. The most common alternative to this approach

Table 1
Cox model fit to the aorta aneurysm data.

	β	exp(β)	S.E.	χ^2	p
Age (years)	0.06568	1.068	0.01259	27.2057	<0.0001
Gender (0 male, 1 female)	0.49550	1.641	0.26768	3.4265	0.0642
Shock (0 without shock, 1 with shock)	0.72443	2.064	0.20016	13.0985	0.0003

is, in practice, the use of p-values as data dependent measure of evidence. However, p-values are not a reliable measure of evidence and they have been shown to be, in certain cases, highly misleading (Berger and Selke, 1987; Royall, 1997).

In the Bayesian framework Bayes factors (BF) are used as measures of empirical evidence (Kass and Raftery, 1995). In this context evaluation of the full Bayes factor is difficult, due to the presence of censored data and to the inherent model complexity. In line with Greenhouse and Silliman (1996) we use the Schwartz criterion (Schwarz, 1978), which approximates the logarithm of the Bayes factor with

$$SC = \log(L(\hat{\theta}_{H_1})) - \log(L(\hat{\theta}_{H_0})) - 0.5d\log(n), \tag{10}$$

where d denotes the difference in number of parameters between the two models, and $\log(L(\hat{\theta}_{H_i}))$ is the maximum of the log-likelihood under H_i, $i = 0, 1$. It can be shown that the Schwarz criterion approximates the Bayes factor as the number of observations grows, and that the rate of convergence is $O(n^{-1/2})$ whenever the models are nested with a reasonable specification of the priors.

The maximum of the likelihood can be found with an EM algorithm (Dempster et al., 1977), which is outlined in Appendix, or more easily with a numerical Newton–Raphson or Genetic Algorithm method, initialised at the observed maximum in the sampled chain. When using flat prior distributions, the starting point is likely to be close to the true global maximum. The EM procedure in Appendix can also be used to compute maximum likelihood estimates for the parameters and make inference in a classical framework. Note that in the classical framework it is not straightforward to test hypotheses on the boundary of the parameter space, like $H_0 : p = 0$, whereas in the Bayesian framework it is conceptually straightforward.

The Schwarz criterion is particularly suitable for choosing among hypotheses since it naturally penalizes for model complexity (as measured by the number of parameters). It is also acceptable from a frequentist point of view, and it is equivalent to minus twice the difference between the Bayesian Information Criterion (BIC) computed for each model. Hence, the BIC criterion and the Schwarz criterion are equivalent and always lead to the same model choice.

Testing is easily performed since, when the exponential of (10) is greater than 1, the data provide more evidence in favor of H_1 relative to H_0, with vice versa the opposite conclusion holding otherwise. Larger values are interpreted as stronger evidence in favor of H_1. Jeffreys (1961) proposed an empirical scale for classifying evidence provided by a Bayes factor, which can also be used for the Schwartz criterion. For instance, values larger than 3 (smaller than 1/3) are judged as moderate but clear evidence in favor of the alternative (null) hypothesis.

6. Application to an aorta aneurysm study

In the period from April 1998 through January 2000, 196 patients affected by rupture of an aorta aneurysm underwent surgery at the University of Vienna. Post-operative survival was the main clinical endpoint. The median follow-up is 1.69 months, resulting in 40.63% of censoring. Age ranged from 47 to 94 years with a mean value of about 71 years (s.d. 8.55), 168 patients where male (87.5%), all patients experienced a vessel rupture but only in 90 patients (46.88%) this was complicated by the occurrence of shock.

The analysis of this data set was performed in two separate steps. A preliminary analysis was carried out using the Cox model, with the aim of illustrating that even such a flexible model fails to provide a satisfactory fit in this multiple-component risk setting. Special attention was devoted to residual analysis. The proposed mixture model was subsequently fitted and results were compared with those obtained from the Cox model.

Results from a preliminary analysis using the Cox model are shown in Table 1.

Graphical inspection of martingale residuals (Therneau et al., 1990) does not show any departure from a linear effect for age. Conversely inspection of Schoenfeld residuals (Schoenfeld, 1982) detects a possible time dependent effect of Shock and Age.

Figs. 2 and 3 show a smoothed plot of scaled Schoenfeld residuals versus respectively observed failure times and their ranks. Both plots suggest that the occurrence of a shock initially leads to an increased risk of mortality that eventually wears off with time. The extreme concentration of events immediately after surgery is a serious issue when one tries to model the time dependency. Spurious effects tend to manifest unless a consistent degree of smoothing is applied, as indicated by the dashed and dotted lines in Fig. 2. However the same high degree of smoothing might introduce a significant bias in local estimates. Agreeing, albeit less clear evidence is further provided by the scaled Schoenfeld residuals for Age, illustrated in Figs. 4 and 5. Apparently, the effect of age sharply increases after surgery, reaching a peak and thereafter decreases; as seen before, the precise shape over long survival times is hard to infer due to the very limited number of failures. However, when Schoenfeld residuals are plotted against ranks, the interpretation becomes even less clear.

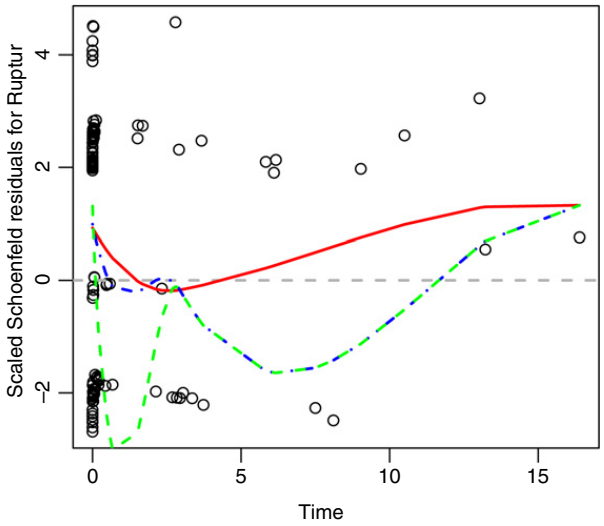

Fig. 2. Scaled Schoenfeld residuals for shock versus failure times.

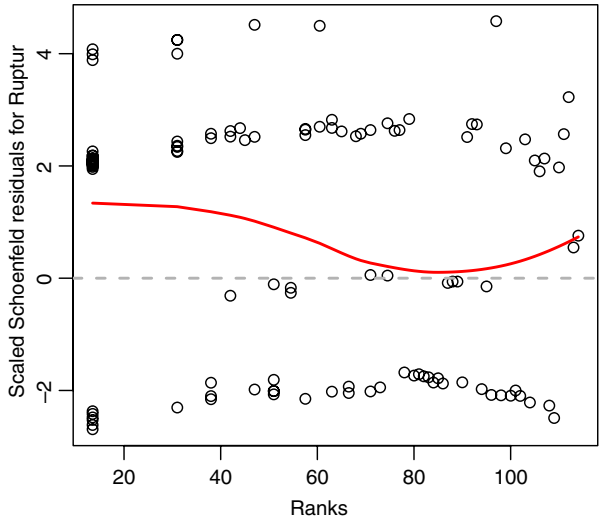

Fig. 3. Scaled Schoenfeld residuals for rupture versus ranks of failure times.

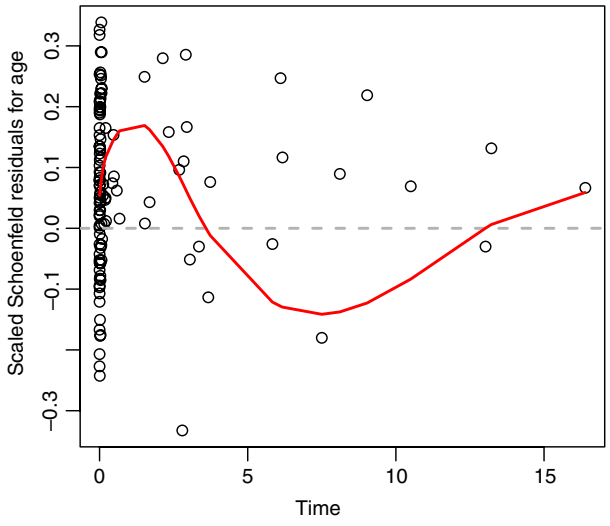

Fig. 4. Scaled Schoenfeld residuals for age versus failure times.

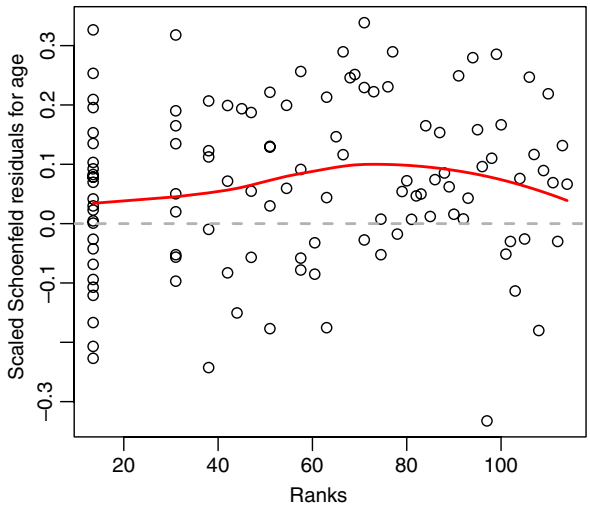

Fig. 5. Scaled Schoenfeld residuals for age versus ranks of failure times.

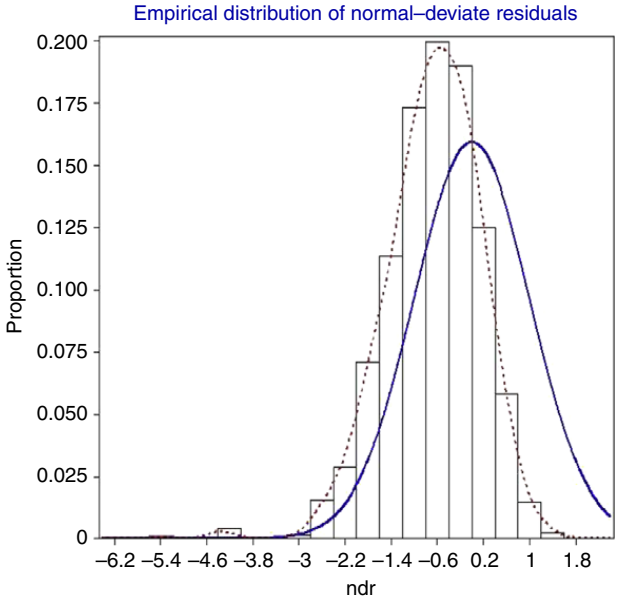

Fig. 6. Empirical distribution of the normal deviate residuals for the Cox model.

A further evidence of lack of fit is offered by Normal deviate residuals (Nardi and Schemper, 1999), plotted in Fig. 6 (showing the empirical distribution of the residuals) and Fig. 7 (instead showing the corresponding qq-plot). It is worth recalling that Normal deviate residuals are defined as the probit transformation of the estimated survival function evaluated at the observed failure times. Hence Normal deviate residuals are 0 if the observed failure time coincides with the median failure time estimated from the fitted model; increasing departures from the predicted median time are reflected by increasing absolute values. Thus these residuals can be regarded as a "distance" of an observed failure time from a predicted one, resembling the classical definition of residual in the general linear model. Note that large negative and positive residuals respectively identify too long and too short survival times, i.e. individuals who survived much longer or died much too early with respect to their predicted median failure time. Under the null hypothesis of a correctly specified model, Normal deviate residuals should approximately follow the standard Normal distribution. In order to avoid spurious evidence of lack of fit arising from censored data, corresponding residuals were imputed under the null hypothesis of a correctly specified model.

The analysis of the empirical distribution of Normal deviate residuals from the fitted Cox model overall suggests a lack of global fit, with an excess of large negative residuals corresponding to patients surviving much longer than their predicted median failure time. This tendency of the Cox model to overestimate the risk of death – especially with reference to long-lasting survivors – is confirmed by the plots in Fig. 10, which show the baseline survival estimate from the Cox model compared with the corresponding Kaplan–Meier estimate. Survival estimates from the fitted model are biased downwards after the drop due to high early risk.

In conclusion, while the Cox model at a first glance appears to adequately describe the examined data, it actuallymisses important features of the underlying hazard pattern and is affected by significant bias induced by the early exposure to considerable risk.

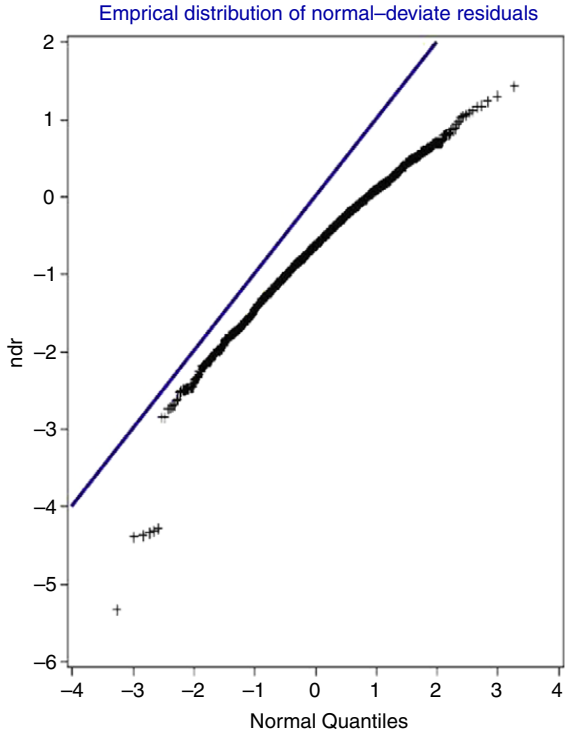

Fig. 7. QQ-plot for the normal deviate residuals for the Cox model.

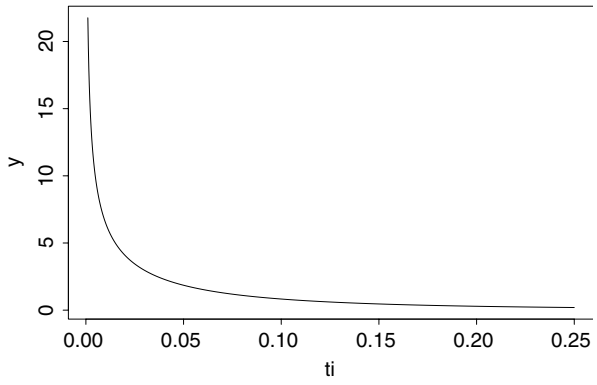

Fig. 8. Risk estimate for the two Weibull mixture model at baseline.

We now fit the proposed two-component Weibull mixture model, and approximate the Bayes factors for testing the hypothesis of lack of effect from each of the covariates. The model was estimated by running three parallel chains from different starting values, with two sets of starting values generated at random and the third consisting of the maximum likelihood estimates. After a burn-in of about 5000 iterations, the three parallel chains were found to be close along their marginal distributions. MCMC sampling was continued for another 5000 iterations which were then used for posterior summaries.

The estimated baseline hazard function shown in Fig. 8 confirms the prior assumption of a considerably high post-operative risk followed by a much lower and stable, yet not negligible, hazard. Table 2 shows the posterior means with corresponding 95% credibility intervals, together with the associated approximate Bayes factors.

We note that no covariate seems to have effect on the time components, one of which is approximately exponential (the $\exp SC$ for $H_0 : \gamma_0 = 1$ versus $H_1 : \gamma_0 \neq 1$ is approximately 0.11), while the occurrence of shock and gender have a strong effect on the likelihood of being allocated to either of the two components. In fact, the Bayes factors in the weight component for these variables are very large. The occurrence of a shock increases the odds of an early death, with females in particular appearing more fragile than men. When it is recognised that two distinct population models are in place for these data, age is seen to be barely important.

It should also be noted that when a CI includes the value $\exp(\beta) = 1$, the Schwartz criterion (10) is small enough to convey small or no evidence in favor of the hypothesis $H_1 : \exp(\beta) \neq 1$.

As anticipated, Fig. 8 shows the estimated risk at baseline (male individual, no shock, average age).

To evaluate adequacy of model fit, the survival function estimated at the baseline was plotted for male individuals with no shock in Fig. 9 together with the corresponding Kaplan–Meier estimate. This can also be compared with the one obtained

Table 2
Two Weibull mixture model fit to the aorta aneurysm data.

	$Z_i = 0$				
γ_0 (95% CI)	1.009 (0.701, 1.317)				
	β	$\exp(\beta)$	$CI^{low}_{0.95}$	$CI^{up}_{0.95}$	e^{SC}
Intercept	−3.229	0.0396	0.0038	0.1529	
Age (years)	0.502	1.651	0.983	2.359	1.11
Gender (0 male, 1 female)	0.527	1.693	0.877	2.819	1.23
Shock (0 without shock, 1 with shock)	0.415	1.514	0.740	5.188	0.33
	$Z_i = 1$				
γ_1 (95% CI)	0.586 (0.482, 0.687)				
Intercept	3.319	27.633	15.259	43.183	
Age (years)	0.262	1.300	0.614	2.387	0.16
Gender (0 male, 1 female)	0.420	1.522	0.591	2.488	0.12
Shock (0 without shock, 1 with shock)	0.492	1.636	0.952	2.287	0.13
	Weight				
Intercept	−0.707	0.493	0.312	0.726	
Age (years)	0.260	1.297	0.844	2.146	0.12
Gender (0 male, 1 female)	1.042	2.835	1.616	5.236	446.28
Shock (0 without shock, 1 with shock)	0.969	2.635	1.582	4.367	308.26

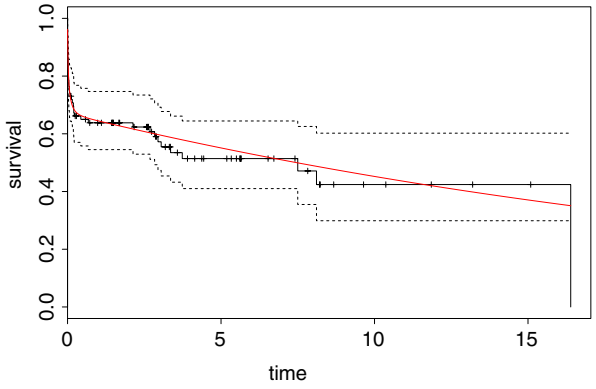

Fig. 9. Kaplan–Meier estimate compared with baseline survival estimate for the two Weibull mixture model.

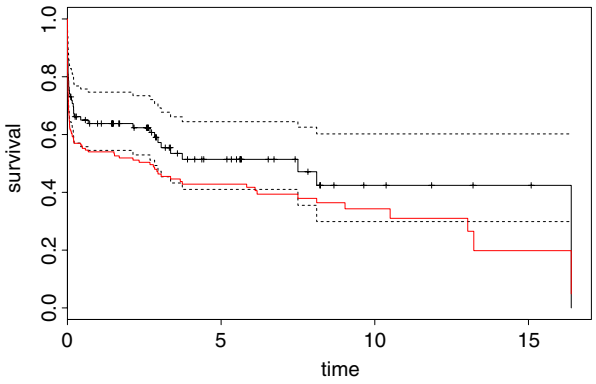

Fig. 10. Kaplan–Meier estimate compared with baseline survival estimate for Cox model.

from the Cox model, which is reported in Fig. 10. It can be seen that the proposed mixture model seems to provide a generally more satisfactory fit to the observed data than that achieved under a proportional hazards Cox assumption.

Fig. 11 plots the posterior probabilities of the component associated with early death, estimated for each subject as a function of the observed time of event or censoring. These probabilities are seen to be mostly extreme (close to zero or to one), suggesting a very small classification uncertainty, and are almost monotone over time suggesting very small overlap between the two components.

Despite the frequentist rules of residual analysis not yet being completely established under a Bayesian perspective, to facilitate a head-to-head comparison with the Cox model we plot both the empirical distribution (Fig. 12) and the

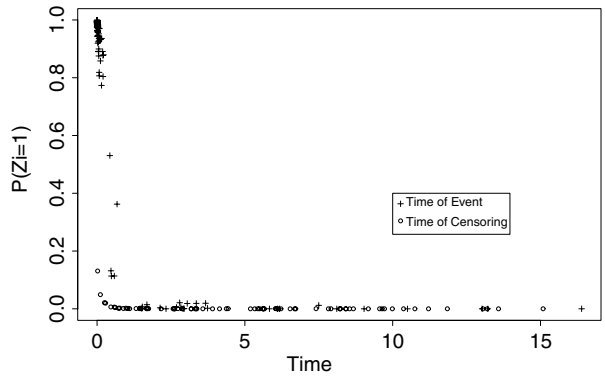

Fig. 11. Posterior probability $P(Z_i = 1|(T_i, \delta_i, X_i))$ as a function of time of event or censoring T_i.

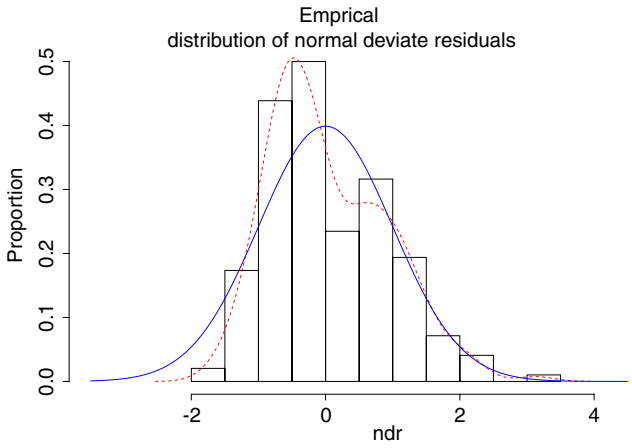

Fig. 12. Empirical distribution of the normal deviate residuals for the two Weibull mixture model.

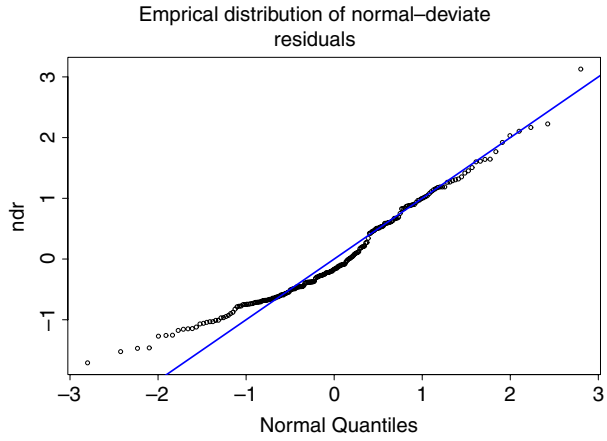

Fig. 13. QQ-plot for the normal deviate residuals for the two Weibull mixture model.

qq-plot (Fig. 13) of the Normal deviate residuals from the two-component Weibull mixture model. The empirical distribution of the residuals is much closer to the reference Normal density under the null hypothesis of perfect fit, in comparison with the corresponding empirical distribution obtained for the Cox model. The excess of residuals in the interval $(-1, 0)$ highlights the difficulty of achieving an adequate fit to the observed data in the presence of a high risk of very early death.

Furthermore we evaluate the robustness of posterior inferences with respect to the choice of prior distributions. Different priors are used for each set of parameters, with prior variances for the covariate-specific coefficients in each component taking values in the set $\{0.5, 1, 5, 10\}$. The same range is used for the variance of the Log-Normals used for the γ parameters. We finally end up with a set of 256 models, all of which are fitted starting the MCMC routine from the maximum likelihood

Table 3
Robustness of posterior means with respect to prior choice for the aorta aneurysm data.

Quantile	0%	25%	50%	75%	100%
	$Z_i = 0$				
γ_0	0.57	0.69	0.90	1.00	1.26
Intercept	−3.36	−3.22	−3.17	−3.13	−2.99
Age	0.35	0.57	0.64	0.72	0.90
Gender	0.47	0.60	0.68	0.82	1.07
Shock	0.17	0.30	0.36	0.47	0.62
	$Z_i = 1$				
γ_1	0.51	0.56	0.61	0.63	0.68
Intercept	3.24	3.56	3.68	3.84	4.16
Age	0.12	0.25	0.31	0.41	0.74
Gender	−0.48	−0.09	0.04	0.18	0.60
Shock	0.04	0.18	0.26	0.31	0.54
	Weight				
Intercept	−0.63	−0.76	−0.81	−0.89	−1.12
Age	0.43	0.27	0.24	0.22	0.16
Gender	1.27	1.20	1.17	1.08	0.99
Shock	1.20	1.07	1.04	0.97	0.89

estimates. After a burn-in of 1000 iterations each chain was left running for another 1000 iterations, and posterior means were recorded.

Table 3 shows the minimum, quartiles and maximum of the 256 posterior mean samples obtained. It can be seen that the model is robust with respect to prior inputs for most of the parameters, as the posterior means stretch over a short range taking values leading to substantially the same conclusions. The only exception is made by effect of gender on the component associated with late risk. This is more likely to be a consequence of the weak information provided by covariates with little predictive power, rather than of genuine sensitivity to prior inputs.

7. Discussion

The paper advocates utilizing a two-component mixture model as a more reliable alternative to the usual Cox model when dealing with data expressing complex hazard patterns, and argues that the fit and interpretation of the former is better compared to the latter at least for data sets where the event of interest may be caused by two distinct risk components.

Pending issues are nevertheless recognised, requiring further research: notably the issue of assessing the goodness of fit of a mixture model to the observed data in the context of survival analysis. The use of Normal deviate residuals was found to enable a direct comparison with the *frequentist* Cox model, although is not formalised with the Bayesian framework. We finally note that, especially when a parametric model structure is assumed, the availability of a measure of agreement between empirical information and model assumptions is crucial to investigate possible global or local weakness of the fitted model.

Acknowledgments

The authors are grateful to Prof. Giampaolo Scalia Tomba for advice on an earlier draft and to Dr. Stefano Conti for careful revision of the style of the presentation.

Appendix. EM algorithm for maximum likelihood estimation

The EM algorithm (Dempster et al., 1977) was used to maximize the likelihood (7). This is an iterative procedure which alternates the following steps until convergence:

- **E-step**: compute the conditional expected value of the complete data log-likelihood, given the current estimate of the parameter vector and the observed data;
- **M-step**: maximize the above expected value with respect to the parameter vector.

In the E-step, the indicator functions of the latent variables Z_j in the complete log-likelihood $l(\theta) = \log(L(\theta))$, where $L(\theta)$ is as in Eq. (7), are replaced by their expected values. It is straightforward to see that the conditional expected values $\tilde{z}_j$ for $I(Z_j = 1)$ are given by Eq. (9). We denote the expected complete log-likelihood by $\tilde{l}(\theta)$.

We can now set up an appropriate M-step.

A.1. M-step for a model with no covariates

If there are no covariates involved, the parameters can be easily maximized. The expected complete likelihood is maximized by equating its partial derivatives to zero. We must then solve the following system of equations:

$$\begin{cases} \dfrac{\partial \tilde{l}(\theta)}{\partial \eta_1} = \dfrac{\gamma_1}{\eta_1}\sum_j \tilde{z}_j\delta_j - \gamma_1\eta_1^{\gamma_1-1}\sum_j \tilde{z}_j t_j^{\gamma_1} = 0 \\ \dfrac{\partial \tilde{l}(\theta)}{\partial \eta_0} = \dfrac{\gamma_0}{\eta_0}\sum_j (1-\tilde{z}_j)\delta_j - \gamma_0\eta_0^{\gamma_0-1}\sum_j (1-\tilde{z}_j) t_j^{\gamma_0} = 0 \\ \dfrac{\partial \tilde{l}(\theta)}{\partial \gamma_1} = \dfrac{\sum_j \delta_j\tilde{z}_j}{\gamma_1} + \sum_j \delta_j\tilde{z}_j\log(t_j) - \eta_1^{\gamma_1}\sum_j \tilde{z}_j t_j^{\gamma_1}(\log(\eta_1)+\log(t_j)) = 0 \\ \dfrac{\partial \tilde{l}(\theta)}{\partial \gamma_0} = \dfrac{\sum_j \delta_j(1-\tilde{z}_j)}{\gamma_0} + \sum_j \delta_j(1-\tilde{z}_j)\log(t_j) - \eta_0^{\gamma_0}\sum_j (1-\tilde{z}_j) t_j^{\gamma_0}(\log(\eta_0)+\log(t_j)) = 0 \\ \dfrac{\partial \tilde{l}(\theta)}{\partial p} = \dfrac{\sum \tilde{z}_j}{p} - \dfrac{\sum(1-\tilde{z}_j)}{1-p} = 0. \end{cases}$$

We immediately get the explicit expression $\hat{p} = \frac{\sum \tilde{z}_j}{n}$. The second and third equations lead to the following expressions for η_0 and η_1 as functions of γ_0 and γ_1:

$$\eta_1 = \left(\frac{\sum_j \tilde{z}_j\delta_j}{\sum_j \tilde{z}_j t_j^{\gamma_1}}\right)^{\frac{1}{\gamma_1}}, \tag{11}$$

and

$$\eta_0 = \left(\frac{\sum_j (1-\tilde{z}_j)\delta_j}{\sum_j (1-\tilde{z}_j) t_j^{\gamma_0}}\right)^{\frac{1}{\gamma_0}}. \tag{12}$$

Now (11) and (12) can be plugged in the fourth and fifth equations, which are nonlinear in γ_0 and γ_1 but can be solved with a simple secant method. The estimated values $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are finally plugged in (11) and (12) to get $\hat{\eta}_0$ and $\hat{\eta}_1$.

In order to keep the model identifiable, as mentioned in Section 4 the larger estimate for the shape parameter is always labelled as $\hat{\gamma}_0$.

A.2. M-step for a model with covariates

When covariates are involved in the model, for simplicity we can split the complete log-likelihood into three summands – one part for each vector of parameters in each of the two Weibull components and one for the α parameters – to be maximized separately: $\tilde{l}(\theta) = \tilde{l}_0(\beta_0, \gamma_0) + \tilde{l}_1(\beta_1, \gamma_1) + \tilde{l}_2(\alpha)$, where

$$\tilde{l}_0(\beta_0, \gamma_0) = \sum (1-\tilde{z}_j)\log(L_{0j}(\theta))$$
$$\tilde{l}_1(\beta_1, \gamma_1) = \sum \tilde{z}_j \log(L_{1j}(\theta))$$
$$\tilde{l}_2(\alpha) = -\sum (1-\tilde{z}_j)\log(1+\exp\{\alpha^t\mathbf{X}_j\}) + \sum \tilde{z}_j(\alpha^t\mathbf{X}_j - \log(1+\exp\{\alpha^t\mathbf{X}_j\})).$$

To maximize $\tilde{l}_2(\alpha)$ a standard Newton–Raphson type iterative algorithm can be used.

In order to set up a Newton–Raphson algorithm to maximize $\tilde{l}_1(\beta_1, \gamma_1)$ we need the first and second derivatives with respect to each parameter:

$$\begin{cases} \dfrac{\partial l_1(\beta_1,\gamma_1)}{\partial \gamma_1} = -\sum \tilde{z}_j t_j^{\gamma_1}\exp(\gamma_1\beta_1^t\mathbf{X}_j)(\log(t_j)+\beta_1^t\mathbf{X}_j) + \sum \tilde{z}_j\delta_j(\log(t_j)+\beta_1^t\mathbf{X}_j) + \dfrac{\sum \tilde{z}_j\delta_j}{\gamma_1} \\ \dfrac{\partial l_1(\beta_1,\gamma_1)}{\beta_{1h}} = -\gamma_1\sum \tilde{z}_j t_j^{\gamma_1}\exp(\gamma_1\beta_1^t\mathbf{X}_j)x_{jh} + \gamma_1\sum \tilde{z}_j\delta_j x_{jh} \quad h = 1,\ldots,d; \end{cases}$$

and

$$\begin{cases} \dfrac{\partial l_1(\beta_1,\gamma_1)}{\partial \gamma_1^2} = -\sum \tilde{z}_j t_j^{\gamma_1}\exp(\gamma_1\beta_1^t\mathbf{X}_j)(\log(t_j)+\beta_1^t\mathbf{X}_j)^2 - \dfrac{\sum \tilde{z}_j\delta_j}{\gamma_1^2} \\ \dfrac{\partial l_1(\beta_1,\gamma_1)}{\partial \gamma_1\partial\beta_{1h}} = -\sum \tilde{z}_j t_j^{\gamma_1}\exp(\gamma_1\beta_1^t\mathbf{X}_j)(\gamma_1\log(t_j)+\gamma_1\beta_1^t\mathbf{X}_j+1)x_{jh} + \sum \tilde{z}_j\delta_j x_{jh} \quad h = 1,\ldots,p \\ \dfrac{\partial l_1(\beta_1,\gamma_1)}{\beta_{1h}\beta_{1r}} = -\gamma_1\sum \tilde{z}_j t_j^{\gamma_1}\exp(\gamma_1\beta_1^t\mathbf{X}_j)x_{jh}x_{jr} \quad h = 1,\ldots,d;\ r = 1,\ldots,d. \end{cases}$$

Similar expressions are obtained for the derivatives of $\tilde{l}_0(\beta_0, \gamma_0)$.

References

Berger, J.O., Selke, T., 1987. Testing a point null hypothesis: The irreconcilability of *p*-value and evidence. Journal of the American Statistical Association 82, 112–139.
Carrasco, J.M.F., Ortega, E.M.M., Cordeiro, G.M., 2008. A generalized modified Weibull distribution for lifetime modeling. Computational Statistics and Data Analysis 53, 450–462.
Cox, D.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, (Ser. B) 34, 187–220.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, (Ser. B) 39, 1–38.
Diebolt, J., Robert, C., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society (Ser. B) 56, 363–375.
Farewell, V.T., 1982. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38, 1041–1046.
Gilks, W.R., Best, N.G., Tan, K.K.C., 1995. Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46 p541-542 with Neal, R.M.). The American Statistician 44, 455–472.
Greenhouse, J.B., Silliman, N.P., 1996. Applications of a mixture survival model with covariates to the analysis of a depression prevention trial. Statistics in Medicine 15, 2077–2094.
Jeffreys, H., 1961. Theory of Probability. Oxford University Press, Oxford.
Jiang, R., Murthy, D.N.P., 1998. Mixture of Weibull distributions — Parametric characterization of failure rate function. Applied Stochastic Models and Data Analysis 14, 47–65.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Mc Lachlan, G.J., Mc Giffin, D.C., 1994. On the role of finite mixture models in survival analysis. Statistical Methods in Medical Research 3, 211–226.
Nardi, A., Schemper, M., 1999. New residuals for Cox regression and their application to outlier screening. Biometrics 55, 523–529.
Pack, S.E., Morgan, B.J.T., 1990. A mixture model for interval-censored time-to-response quantal assay data. Biometrics 46, 749–757.
Pocock, S.J., Gore, S.M., Kerr, G.R., 1982. Long term survival analysis: The curability of breast cancer. Statistics in Medicine 1, 93–106.
Robert, C.P., Casella, G., 2000. Monte Carlo Statistical Methods. Springer, New York.
Royall, M.R., 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall, London.
Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model. Biometrika 69, 239–241.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Tahmasbi, R., Rezaei, S., 2008. A two-parameter lifetime distribution with decreasing failure rate. Computational Statistics and Data Analysis 52, 3889–3901.
Therneau, T.M., Grambsch, P.M., Fleming, T., 1990. Martingale based residuals for survival models. Biometrika 77, 147–160.
van Wieringen, W.N., Kun, D., Hampel, R., Boulesteix, A-L., 2009. Survival prediction using gene expression data: A review and comparison. Computational Statistics and Data Analysis 53, 1590–1603.
Yu, B., Peng, Y., 2008. Mixture cure models for multivariate survival data. Computational Statistics and Data Analysis 52, 1524–1532.