Mathematical Biosciences and Engineering

MBE, 17(2): 1616–1636.
DOI: 10.3934/mbe.2020085
Received: 30 August 2019
Accepted: 22 November 2019
Published: 10 December 2019

http://www.aimspress.com/journal/MBE

Research article

Cluster validity indices for mixture hazards regression models

Yi-Wen Chang [1], Kang-Ping Lu [2] and Shao-Tung Chang [1,*]

[1] Department of Mathematics, National Taiwan Normal University, Taipei, Taiwan
[2] Department of Applied Statistics, National Taichung University of Science and Technology, Taichung, Taiwan

*** Correspondence:** Email: schang@math.ntnu.edu.tw.

Abstract: In the analysis of survival data, the problems of competing risks arise frequently in medical applications where individuals fail from multiple causes. Semiparametric mixture regression models have become a prominent approach in competing risks analysis due to their flexibility and easy interpretation of resultant estimates. The literature presents several semiparametric methods on the estimations for mixture Cox proportional hazards models, but fewer works appear on the determination of the number of model components and the estimation of baseline hazard functions using kernel approaches. These two issues are important because both incorrect number of components and inappropriate baseline functions can lead to insufficient estimates of mixture Cox hazard models. This research thus proposes four validity indices to select the optimal number of model components based on the posterior probabilities and residuals resulting from the application of an EM-based algorithm on a mixture Cox regression model. We also introduce a kernel approach to produce a smooth estimate of the baseline hazard function in a mixture model. The effectiveness and the preference of the proposed cluster indices are demonstrated through a simulation study. An analysis on a prostate cancer dataset illustrates the practical use of the proposed method.

Keywords: mixture regression model; Cox proportional hazards model; EM-algorithm; kernel estimator; validity indices

1. Introduction

Survival analysis is a branch of statistics for analyzing time-to-event data. When looking into survival data, one frequently encounters the problem of competing risks in which samples are subject to multiple kinds of failure. The Cox proportional hazards model, introduced by Cox [1], is popular

in survival analysis for describing the relationship between the distributions of survival times and covariates and is commonly employed to analyze cause-specific survival data. The traditional approach is to separately fit a Cox proportional hazards model to the data for each failure type, after considering the data with other kinds of failure censored. However, this conventional method encounters problems like the estimates are hard to interpret and the confidence bands of estimated hazards are wide, because the method does not cover all failure types [2,3].

An alternative approach is to fit competing risks data by using a mixture model that incorporates the distinct types of failure to partition the population into groups, and it assumes an individual will fail from each risk with the probabilities being attributed to the proportions of each group, respectively. Moreover, the mixture approach is helpful for estimating the effects of covariates in each group through parametric proportional hazard regressions such as Cox's model. McLachlan and Peel [4] noted that a mixture model is allowed for both dependent and independent competing risks and it can improve a model's fit to the data than the traditional approach in which the causes of failure are assumed to be independent. Mixture models are popular in competing risks analysis, because their resultant estimates are easy to interpret [2], although complex.

Semi-parametric mixture models are a generalization of parametric mixture models and have become a prominent approach for modelling data with competing risks. Semiparametric approaches to mixture models are preferable for their ability to adjust for the associated variables and allow for assessing the effects of these variables on both the probabilities of eventual causes of failure through a logistic model and the relevant conditional hazard functions by applying the Cox proportional hazards model (cf. [2]). Below, we review the existing semiparametric methods of mixture models for competing risks data.

Ng and McLachlan [5] proposed an ECM-based semi-parametric mixture method without specifying the baseline hazard function to analyze competing risks data. They noted that when the component-baseline hazard is not monotonic increasing their semi-parametric approach can consistently produce less biased estimates than those done by fully parametric approaches. Moreover, when the component-baseline hazard is monotonic increasing, the two approaches demonstrate comparable efficiency in the estimation of parameters for mildly and moderately censoring. Chang et al. [6] studied non-parametric maximum-likelihood estimators through a semiparametric mixture model for competing risks data. Their model is feasible for right censored data and can provide estimates of quantities like a covariate-specific fatality rate or a covariate-specific expected time length. Moreover, Lu and Peng [7] set up a semiparametric mixture regression model to analyze competing risks data under the ordinary mechanism of conditional independent censoring. Choi and Huang [8] offered a maximum likelihood scheme for semiparametric mixture models to make efficient and reliable estimations for the cumulative hazard function. One advantage with their approach is that the joint estimations for model parameters connect all considered competing risks under the constraint that all the probabilities of failing from respective causes sum to 1 given any covariates. Other research studies for competing risks data are based on semiparametric mixture models, e.g. [5–8].

Although the mixture hazard model is preferable to direct approaches, two important but challenging issues frequently encountered in the applications are the determination of the number of risk types and the identification of the failure type of each individual.

It is understandable that the results of a mixture model analysis highly depend on the number of components. It is also conceivably hard to cover all types of competing risks in a mixture model.

Validity indices are a vital technique in model selection. The cluster validity index is a kind of criterion function to determine the optimal number of clusters. Some cluster validity indices presented by [9–11] are designed to find an optimal cluster number for fuzzy clustering algorithms; some are only related to the membership, while some take into account the distance between the data sets and cluster centers. Wu et al. [12] proposed median-type validity indices, which are robust to noises and outliers. Zhou et al. [13] introduced a weighted summation type of validity indices for fuzzy clustering, but they are unfeasible for mixture regression models. Conversely, Henson et al. [14] evaluated the ability of several statistical criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to produce a proper number of components for latent variable mixture modeling. However, AIC and BIC may present the problem of over- and under-estimation on the number of components [15], respectively.

As to the identification of failure types, many studies on the problems of competing risks like [5–8] assumed that the failure type of an individual is known if the subject's failure time is observed, but if an individual is censored and only the censored time is known, then which failure type the subject fails from is unknown. In fact, even if one observes the failure time, then the true cause of failure might not be clear and needs further investigations. Thus, deciding the number of competing risks and recognizing the failure type of each individual are critical in competing risks analysis, but scant work has been done on them.

Besides the above problems, another critical issue existing in mixture Cox hazard models particularly is the estimation of the baseline hazard function. The Cox proportional hazards model consists of two parts: the baseline hazard function and the proportional regression model. Bender et al. [16] assumed that the baseline hazard function follows a specific lifetime distribution, but this assumption is obviously restrictive. A single lifetime distribution may not adequately explain all data —for example, the failure rate is not monotonic increasing or decreasing. Alternatively, some scholars adopted nonparametric approaches to estimate the baseline hazard function that are more flexible. Ng and McLachlan [5] assumed the baseline hazard function to be piecewise constant by treating each observed survival time as a cut-off point, but the piecewise constant assumption has the disadvantage that the estimated curve is not smooth, while smoothing is required in several applications [17]. In fact, our simulation results also show that the derived estimates based on a piecewise constant hazard function are not sufficient in some cases (e.g. model 4 in Figure 4). Understandably, an inadequate estimation of the baseline function affects the selection of the number of model components; and hence leads to insufficient estimates of the model parameters.

In order to solve the above mentioned problems with the Cox mixture hazard modelling for competing risks data, we propose four indices and the kernel estimation for the base line function in this paper. Validity indices are a vital technique in model selection, but they have been less utilized for deciding the number of components of a mixture regression model. By using posterior probabilities and residual functions, we propose four validity indices that are applicable to regression models in this study. Under the EM-based mixture model, the posterior probabilities play an important role in classifying data, which take role of data memberships in fuzzy clustering. Unlike the traditional regression model, the survival model does not meet the assumption that survival time variation is constant for each covariate. Therefore, we incorporate the functions of posterior probabilities and the sum of standard residuals to constitute the new validity indices and verify the effectiveness of the proposed new indices through extensive simulations. Moreover, we extend the kernel method of Guilloux et al. [18] to estimate the baseline hazard function smoothly and hence

more accurately.

The remainder of this paper is organized as follows. Section 2 introduces the mixture Cox proportional hazards regression model, develops an EM-based algorithm to estimate the model parameters, and also discusses kernel estimations for the baseline hazard function. Section 3 constructs four validity indices for selecting the number of model components in a mixture Cox proportional hazards regression model. Section 4 carries out several simulations and assesses the effectiveness of our validity indices. Section 5 analyzes a practical data set of prostate cancer patients treated with different dosages of the drug diethylstilbestrol. Finally, Section 6 states conclusions and a discussion.

2. Mixture Cox proportional hazards model with kernel estimation

2.1. Mixture Cox proportional hazards model

For mixture model analysis, suppose each member of a population can be categorized into g mutually exclusive clusters according to its failure type. Let $D=\left\{ (t_j, \boldsymbol{X}_j^T, \delta_j): j=1,\text{L } ,n\right\}$, be a sample drawn from this population where T denotes the transpose of a vector, t_j is the failure or right censoring time, $\boldsymbol{X}_j=(x_{j1}, x_{j2},..., x_{jd})^T$ is a d-dimensional vector of covariates, and:

$$\delta_j=\begin{cases}1, & \text{if the } j\text{-th individaul is uncensored,}\\ 0, & \text{if the } j\text{-th individaul is censored.}\end{cases}$$

The mixture probability density function (pdf) of t is defined by:

$$f(t)=\sum_{i=1}^{g} p_i\cdot f_i(t)\text{, subject to } \sum_{i=1}^{g} p_i=1, \tag{1}$$

where p_i is the mixing probability of failure due to the i^{th} type of risk and g is the number of model components.

In the i^{th} component, the hazard function $h_i(t\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)$ given covariate $\boldsymbol{X}_j$ follows a Cox proportional hazards model defined by

$$h_i(t\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)=h_{0i}(t)\exp(\boldsymbol{X}_j{}^T\boldsymbol{\beta}_i), \tag{2}$$

where $\boldsymbol{\beta}_i=(\beta_{i1},\beta_{i2},...,\beta_{id})^T$ is the vector of regression coefficients, and $h_{0i}(t)$ is the baseline hazard function of the i^{th} component. We define the i^{th} component-survival function and pdf by:

$$S_i(t\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)=\exp\left[-H_{0i}(t)\exp(\boldsymbol{X}_j{}^T\boldsymbol{\beta}_i)\right]$$

and

$$f_i(t\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)=h_{0i}(t)\exp\left[\boldsymbol{X}_j{}^T\boldsymbol{\beta}_i-H_{0i}(t)\exp(\boldsymbol{X}_j{}^T\boldsymbol{\beta}_i)\right],$$

where $H_{0i}(t)=\int_0^t h_{0i}(s)\,\mathrm{d}s$ is the i^{th} component-cumulative baseline hazard function.

The unknown parameters are the mixing probabilities $\boldsymbol{p}=(p_1,p_2,...,p_{g-1})^T$ and regression coefficients $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T,...,\boldsymbol{\beta}_g{}^T)^T$, where $\boldsymbol{\beta}_i=(\beta_{i1},\beta_{i2},...,\beta_{id})^T$. Based on (1) and Zhou [19], the log-likelihood function under the mixture hazards model with right censored data is given by

$$l(\boldsymbol{p},\boldsymbol{\beta})=\sum_{j=1}^{n}\sum_{i=1}^{g}\left\{\delta_j\log\left[p_i\cdot f_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]+(1-\delta_j)\log\left[p_i\cdot S_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]\right\},$$

where $f(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}\right.)=\sum_{i=1}^{g}p_i\cdot f_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)$ and $S(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}\right.)=\sum_{i=1}^{g}p_i\cdot S_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)$.

Assume that the true causes of failure for an individual are unobserved, and hence the data are incomplete. We introduce the latent variable z_{ij} as:

$$z_{ij}=\begin{cases}1, & \text{if } j\text{-th individual fails due to } i\text{-th type of risk;}\\ 0, & \text{otherwise.}\end{cases}$$

The complete-data log-likelihood function is given by:

$$l_c(\boldsymbol{p},\boldsymbol{\beta})=\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\left\{\delta_j\log\left[p_i\cdot f_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]+(1-\delta_j)\log\left[p_i\cdot S_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]\right\}. \tag{3}$$

Subsequently, the parameters are estimated through the expectation and maximization (EM) algorithm.

E-step: On the $(k+1)^{\text{th}}$ iteration of E-step, we calculate the conditional expectation of the complete-data log-likelihood (3) given the current estimates of the parameters, i.e.:

$$\begin{aligned}E\left[l_c(\boldsymbol{p},\boldsymbol{\beta})\left|\boldsymbol{p}^{(k)},\boldsymbol{\beta}^{(k)}\right.\right]&=\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}{}^{(k)}\left\{\delta_j\log\left[p_i\cdot f_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]+(1-\delta_j)\log\left[p_i\cdot S_i(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_i\right.)\right]\right\}\\&=\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}{}^{(k)}\log p_i\\&\quad+\sum_{j=1}^{n}z_{1j}{}^{(k)}\left[\delta_j\log f_1(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_1\right.)+(1-\delta_j)\log S_1(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_1\right.)\right]\\&\quad+\ \mathrm{M}\\&\quad+\sum_{j=1}^{n}z_{gj}{}^{(k)}\left[\delta_j\log f_g(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_g\right.)+(1-\delta_j)\log S_g(t_j\left|\boldsymbol{X}_j,\boldsymbol{\beta}_g\right.)\right]\\&=Q_0+Q_1+\mathrm{L}\ +Q_g\,.\end{aligned} \tag{4}$$

Here, $\boldsymbol{p}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ are the estimates of $\boldsymbol{p}$ and $\boldsymbol{\beta}$, respectively, in the k^{th} iteration. By Baye's

Theorem, we have:

$$z_{ij}^{(k)} = E\left(z_{ij} \left| \boldsymbol{p}^{(k)}, \boldsymbol{\beta}^{(k)} \right.\right) = \frac{p_i^{(k)} f_i(t_j \left| \boldsymbol{X}_j, \boldsymbol{\beta}_i^{(k)} \right.)^{\delta_j} S_i(t_j \left| \boldsymbol{X}_j, \boldsymbol{\beta}_i^{(k)} \right.)^{1-\delta_j}}{\sum_{l=1}^{g} p_l^{(k)} f_l(t_j \left| \boldsymbol{X}_j, \boldsymbol{\beta}_l^{(k)} \right.)^{\delta_j} S_l(t_j \left| \boldsymbol{X}_j, \boldsymbol{\beta}_l^{(k)} \right.)^{1-\delta_j}} \tag{5}$$

which is the posterior probability that the j^{th} individual with survival time t_j fails due to the i^{th} type of risk.

M-step: The $(k+1)^{\text{th}}$ iteration of M-step provides the updated estimates $\boldsymbol{p}^{(k+1)}$ and $\boldsymbol{\beta}^{(k+1)}$ that maximizes (4) with respect to $\boldsymbol{p}$ and $\boldsymbol{\beta}$.

Under the constraints $\sum_{i=1}^{g} p_i = 1$, to maximize $Q_0 = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}^{(k)} \log p_i$ from (4), we obtain the estimation of mixing probability with

$$p_i^{(k+1)} = \frac{\sum_{j=1}^{n} z_{ij}^{(k)}}{n}. \tag{6}$$

The equation Q_i from (4) for $i = 1,...,g$ can be written as:

$$Q_i = \sum_{j=1}^{n} z_{ij}^{(k)} \left\{ \delta_j \left[\log h_{0i}(t_j) + \boldsymbol{X}_j{}^T \boldsymbol{\beta}_i \right] - \exp(\boldsymbol{X}_j{}^T \boldsymbol{\beta}_i) H_{0i}(t_j) \right\}. \tag{7}$$

Define the score vector $U(\boldsymbol{\beta}_i)$ for $i = 1,...,g$ as the first derivate of (7) with respect to the vector $\boldsymbol{\beta}_i$ given $H_{0i}(t)$ fixed at $H_{0i}^{(k+1)}(t)$, and the estimation $\boldsymbol{\beta}_i^{(k+1)}$ satisfies the equation:

$$U(\boldsymbol{\beta}_i) = \left. \frac{\partial Q_i}{\partial \boldsymbol{\beta}_i} \right|_{H_{0i}(t_j) = H_{0i}^{(k+1)}(t_j)} = \sum_{j=1}^{n} z_{ij}^{(k)} \left[\delta_j - \exp(\boldsymbol{X}_j{}^T \boldsymbol{\beta}_i) H_{0i}^{(k+1)}(t_j) \right] \boldsymbol{X}_j = 0. \tag{8}$$

2.2. Kernel estimation for the baseline hazard function

To estimate the baseline hazard function under the mixture hazards model, we propose the kernel estimator. Define $N_j(t) = I(t_j \le t \wedge \delta_j = 1)$ as an event counting process and $Y_j(t) = I(t_j \ge t)$ as risk process. The updated kernel estimator of i^{th} component-baseline hazard function $h_{0i}(t)$ on the $(k+1)^{\text{th}}$ iteration is defined by:

$$h_{0i}^{(k+1)}(t\,|\,\boldsymbol{X},\boldsymbol{Z}_i^{(k)},\boldsymbol{\beta}_i^{(k)},b^{(k)})=\frac{1}{b^{(k)}}\int_0^{\tau}K\left(\frac{t-u}{b^{(k)}}\right)dH_{0i}^{(k+1)}(u\,|\,\boldsymbol{X},\boldsymbol{Z}_i^{(k)},\boldsymbol{\beta}_i^{(k)}),\quad \tau\geq 0, \tag{9}$$

where $K:\ \rightarrow$ is a kernel function, and $b^{(k)}$ is a positive parameter called the bandwidth. There are several types of kernel functions commonly used, appearing in Table 1 and Figure 1. We try these kernel functions in the simulated examples and find no significant differences. In this paper, we choose biweight as the kernel function to estimate the baseline hazard function.

Table 1. Different types of kernel function.

Kernel function	$K(u)$
Gaussian	$K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}\quad,-\infty<u<\infty$
Epanechnikov	$K(u)=\frac{3}{4}(1-u^2)\quad,\lvert u\rvert\leq 1$
Biweight	$K(u)=\frac{15}{16}(1-u^2)^2\quad,\lvert u\rvert\leq 1$
Triweight	$K(u)=\frac{35}{32}(1-u^2)^3\quad,\lvert u\rvert\leq 1$

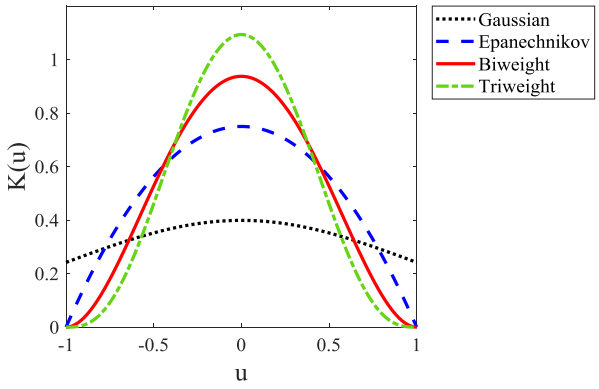

Figure 1. Plot of different types of kernel function.

Derived by smoothing the increments of the Breslow estimator, the kernel estimator (9) can be written as:

$$h_{0i}^{(k+1)}(t\,|\,\boldsymbol{X},\boldsymbol{Z}_i^{(k)},\boldsymbol{\beta}_i^{(k)},b^{(k)})=\frac{1}{b^{(k)}}\sum_{j=1}^{n}\int_0^{\tau}K\left(\frac{t-u}{b^{(k)}}\right)\frac{z_{ij}^{(k)}I(\bar{Y}(u)>0)}{S_{ni}(u\,|\,\boldsymbol{X},\boldsymbol{Z}_i^{(k)},\boldsymbol{\beta}_i^{(k)})}dN_j(u), \tag{10}$$

where $\bar{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_j$ and $S_{ni}(u\,|\,\boldsymbol{X},\boldsymbol{Z}_i,\boldsymbol{\beta}_i)=\sum_{j=1}^{n}z_{ij}\exp(\boldsymbol{X}_j{}^T\boldsymbol{\beta}_i)Y_j(u)$.

Horova et al. [20] and Patil [21] introduced the cross-validation method to select the bandwidth of

the kernel estimator. We define the cross-validation function for bandwidth b written as $CV(b)$ under our model as:

$$CV(b)=\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}^{(k)}\cdot\left[h_{0i}^{(k+1)}(t_j\,|\,\boldsymbol{X},\boldsymbol{Z}_i^{(k)},\boldsymbol{\beta}_i^{(k)},b^{(k)})\right]^2-2\sum_{i=1}^{g}\sum_{l\neq j}\sum \frac{1}{b^{(k)}}K\left(\frac{t_l-t_j}{b^{(k)}}\right)\frac{z_{il}^{(k)}\delta_l}{Y(t_l)}\frac{z_{ij}^{(k)}\delta_j}{Y(t_j)}\ .$$

The selection of bandwidth on the $(k+1)^{\text{th}}$ iteration is given by:

$$b^{(k+1)}=\arg\min_{b\in B_n} CV(b)\,, \tag{11}$$

where B_n cover the set of acceptable bandwidths.

The algorithm is shown as follows, where we fix n and g and set up initial values for mixing probabilities $\boldsymbol{p}^{(0)}$, which are usually $1/g$, regression coefficients $\boldsymbol{\beta}^{(0)}$, baseline hazard rates $\boldsymbol{h}_0^{(0)}$, bandwidth $b^{(0)}$ is 0.5, and a termination value, $\varepsilon>0$.

Set the initial counter $l=1$.

Step 1: Compute $\boldsymbol{Z}^{(l-1)}$ with $\boldsymbol{p}^{(l-1)}$, $\boldsymbol{h}_0^{(l-1)}$ and $\boldsymbol{\beta}^{(l-1)}$ by (5);

Step 2: Update $\boldsymbol{p}^{(l)}$ with $\boldsymbol{Z}^{(l-1)}$ by (6);

Step 3: Update $\boldsymbol{h}_0^{(l)}$ with $\boldsymbol{Z}^{(l-1)}$, $\boldsymbol{\beta}^{(l-1)}$ and $b^{(l-1)}$ by (10);

Step 4: Update bandwidth $b^{(l)}$ with $\boldsymbol{Z}^{(l-1)}$, $\boldsymbol{h}_0^{(l)}$ and $\boldsymbol{\beta}^{(l-1)}$ by (11);

Step 5: Update $\boldsymbol{\beta}^{(l)}$ with $\boldsymbol{Z}^{(l-1)}$, $\boldsymbol{h}_0^{(l)}$ and $\boldsymbol{\beta}^{(l-1)}$ by (8);

Step 6: IF $\left\|\boldsymbol{p}^{(l)}-\boldsymbol{p}^{(l-1)}\right\|_2+\left\|\boldsymbol{h}_0^{(l)}-\boldsymbol{h}_0^{(l-1)}\right\|_2+\left\|\boldsymbol{\beta}^{(l)}-\boldsymbol{\beta}^{(l-1)}\right\|_2<\varepsilon$, THEN stop;

ELSE let $l=l+1$ and GOTO Step 1.

Note that the superscript (.) represents the number of iterations, $\boldsymbol{h}_0^{(0)}=(\boldsymbol{h}_{01}^{(0)},\boldsymbol{h}_{02}^{(0)},...,\boldsymbol{h}_{0g}^{(0)})^T$ is a $g\times n$ matrix, where $\boldsymbol{h}_{0i}^{(0)}=(h_{0i}^{(0)}(t_1),h_{0i}^{(0)}(t_2),...,h_{0i}^{(0)}(t_n))^T$, and each row is initialized by a constant vector.

3. Validity indices

In traditional regression analysis, we select the best model by picking the one that minimizes the sum of squared residuals, but unlike the traditional regression model, the survival model does not meet the assumption that the standard deviation of the survival time is a constant at each covariate. From Figure 2(a), we see that the survival time with higher expectation has higher standard deviation. Therefore, to select the best model we need to adjust the standard deviation to avoid being greatly affected by data that have large standard deviations. Moreover, if the model fits the data well, then

each observed survival time will be close to the expectation of the component model, which has the largest posterior probability corresponding to one's risk type.

Figure 2(b) illustrate that observation A is closer to the mean line (red line) of the component model corresponding to risk type 1, say model 1, than to the mean line (blue line) of model 2. From (5), we see that the posterior probabilities of the observation A corresponding to the first type of risk (red line) will be much larger than that of the second type of risk (blue line). Hence, to build up validity indices for mixture models, we consider the posterior probabilities as the role of weights and define the mixture sum of standard absolute residuals (MsSAE) and mixture sum of standard squared residuals (MsSSE) as follows:

$$MsSAE = \sum_{i=1}^{g}\sum_{j=1}^{n} \frac{\hat{z}_{ij}\left|t_j - \hat{E}_i(t_j)\right|}{\sqrt{\hat{V}ar_i(t_j)}} \;;\; MsSSE = \sum_{i=1}^{g}\sum_{j=1}^{n} \frac{\hat{z}_{ij}\left[t_j - \hat{E}_i(t_j)\right]^2}{\hat{V}ar_i(t_j)},$$

where $\hat{E}_i(t_j) = \int_0^{\infty} \exp\left[-\hat{H}_{0i}(t)\right]^{\exp\left(x_j^T \hat{\beta}_i\right)} dt$ and $\hat{V}ar_i(t_j) = 2\int_0^{\infty} t \cdot \exp\left[-\hat{H}_{0i}(t)\right]^{\exp\left(x_j^T \hat{\beta}_i\right)} dt - \hat{E}_i(t_j)^2$. The squared distance is considered, because it is easier to catch an abnormal model.

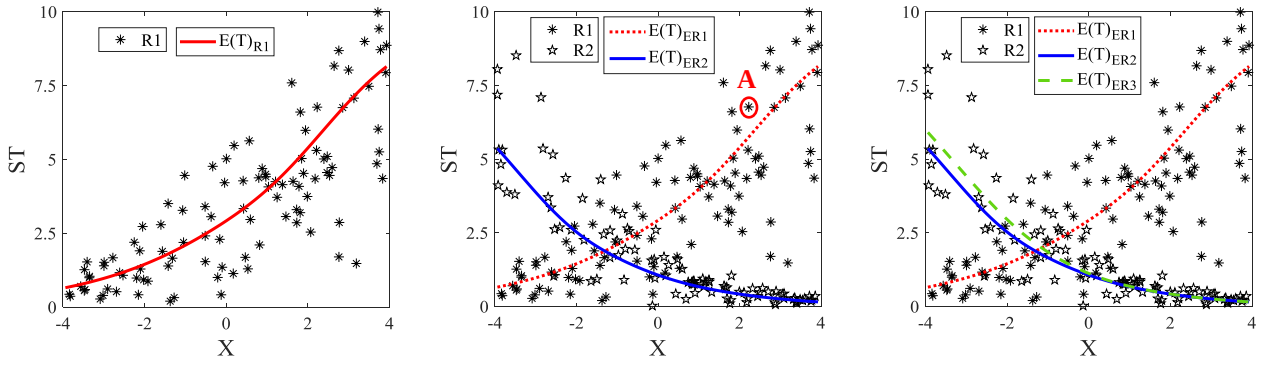

Figure 2. The scatter plot of the observed data and the expectation of the survival time E(T) from non-mixture model (a), mixture models with two types of risk (b), and three types of risk (c). Note that ER represents the type of risk for estimation.

From Figure 2(c) we can see that the expectation (green line) of the survival time according to the third type of risk (ER3) is close to that (blue line) corresponding to the second type of risk (ER2). In order to penalize the overfitting model, which is the model with too many model components, we consider the distance between the expectations of each survival time according to any two types of risk as the penalty. Define the average absolute separation $\overline{ASep}$, the average squared separation $\overline{SSep}$, the minimum absolute separation $\min ASep$ and the minimum squared

separation $\min SSep$ as:

$$\overline{ASep}=\frac{2}{g(g-1)}\sum_{i=1}^{g}\sum_{l>i}^{g}\sum_{j=1}^{n}\left|\hat{E}_i(t_j)-\hat{E}_l(t_j)\right|;\ \ \overline{SSep}=\frac{2}{g(g-1)}\sum_{i=1}^{g}\sum_{l>i}^{g}\sum_{j=1}^{n}\left[\hat{E}_i(t_j)-\hat{E}_l(t_j)\right]^2;$$

$$\min ASep=\min_{i\neq l}\sum_{j=1}^{n}\left|\hat{E}_i(t_j)-\hat{E}_l(t_j)\right|;\ \ \min SSep=\min_{i\neq l}\sum_{j=1}^{n}\left[\hat{E}_i(t_j)-\hat{E}_l(t_j)\right]^2.$$

A good model will possess small mixture standard residuals and large separation of expectations. Hence, based on the above-mentioned functions of residuals and separation of means, we propose four validity indices $V_1 \sim V_4$ for selecting the number of model components under the mixture hazards regression model.

(V_1). Absolute standard residuals and average separation V_{aRaS}

$$V_{aRaS}=\frac{MsSAE}{\overline{ASep}}=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}\left|t_j-\hat{E}_i(t_j)\right|\Big/\sqrt{\hat{V}ar_i(t_j)}}{\frac{2}{g(g-1)}\sum_{i=1}^{g}\sum_{l>i}^{g}\sum_{j=1}^{n}\left|\hat{E}_i(t_j)-\hat{E}_l(t_j)\right|}$$

We find an optimal number g of types of risk by solving $\min_{2\le g\le n-1}V_{aRaS}$.

(V_2). Squared standard residuals and average separation V_{sRaS}

$$V_{sRaS}=\frac{MsSSE}{\overline{SSep}}=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}\left[t_j-\hat{E}_i(t_j)\right]^2\Big/\hat{V}ar_i(t_j)}{\frac{2}{g(g-1)}\sum_{i=1}^{g}\sum_{l>i}^{g}\sum_{j=1}^{n}\left[\hat{E}_i(t_j)-\hat{E}_l(t_j)\right]^2}$$

We find an optimal number g of types of risk by solving $\min_{2\le g\le n-1}V_{sRaS}$.

(V_3). Absolute standard residuals and minimum separation V_{aRmS}

$$V_{aRmS}=\frac{MsSAE}{\min ASep}=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}\left|t_j-\hat{E}_i(t_j)\right|\Big/\sqrt{\hat{V}ar_i(t_j)}}{\min\limits_{i\neq l}\sum_{j=1}^{n}\left|\hat{E}_i(t_j)-\hat{E}_l(t_j)\right|}$$

We find an optimal number g of types of risk by solving $\min_{2\le g\le n-1}V_{aRmS}$.

(V_4). Squared standard residuals and minimum separation V_{sRmS}

$$V_{sRmS}=\frac{MsSSE}{\min SSep}=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}\left[t_j-\hat{E}_i(t_j)\right]^2\Big/\hat{V}ar_i(t_j)}{\min\limits_{i\neq l}\sum_{j=1}^{n}\left[\hat{E}_i(t_j)-\hat{E}_l(t_j)\right]^2}$$

We find an optimal number g of types of risk by solving $\min_{2\le g\le n-1} V_{sRmS}$.

4. Simulation

For the simulated data we consider four different models M_1~M_4. Under the mixture Cox proportional hazards model (2), the i^{th} component hazard function is:

$$h_i(t\,|\,\boldsymbol{X}_j,\boldsymbol{\beta}_i)=h_{0i}(t)\exp(x_j\beta_{i,1,1}+...+x_j{}^k\beta_{i,1,k}+...+x_j\beta_{i,d,1}+...+x_j{}^k\beta_{i,d,k}),$$

where d is the number of covariates, k is the degree of models, $\boldsymbol{\beta}_i=(\beta_{i,1,1},\beta_{i,1,k},...,\beta_{i,d,k},\beta_{i,d,k})^T$ is the vector of regression coefficients and $h_{0i}(t)$ is the i^{th} component-baseline hazard function.

Consider two common distributions for the baseline hazard functions, Weibull and Gompertz; the i^{th} component Weibull baseline and Gompertz baseline are defined by $h_{0i}(t)=\lambda_i\,\rho_i\,t^{\rho_i-1}$ and $h_{0i}(t)=\lambda_i\exp(\rho_i t_j)$, respectively, where λ_i and ρ_i are the scale and shape parameters. Let $\boldsymbol{\lambda}=(\lambda_1,...,\lambda_g)^T$, $\boldsymbol{\rho}=(\rho_1,...,\rho_g)^T$, and $\boldsymbol{\beta}=(\boldsymbol{\beta}_1{}^T,\boldsymbol{\beta}_2{}^T,...,\boldsymbol{\beta}_g{}^T)^T$. The covariates $\boldsymbol{X}=(x_1,x_2,...,x_n)^T$ in all cases are generated independently from a uniform distribution $U(-4,4)$. The information for four models is shown in Table 2, and the scatter plots of a sample dataset are presented in Figure 3.

Table 2. The information for models M_1~M_4 respectively.

Model	n[1]	g[2]	d[3]	k[4]	$\boldsymbol{p}=\begin{bmatrix}p_1\\ \mathrm{M}\\ p_g\end{bmatrix}$	BH[5]	$\boldsymbol{\lambda}=\begin{bmatrix}\lambda_1\\ \mathrm{M}\\ \lambda_g\end{bmatrix}$	$\boldsymbol{\rho}=\begin{bmatrix}\rho_1\\ \mathrm{M}\\ \rho_g\end{bmatrix}$	$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1{}^T\\ \mathrm{M}\\ \boldsymbol{\beta}_g{}^T\end{bmatrix}$	U_i[6]
M_1	200	2	1	1	$\begin{bmatrix}0.7\\0.3\end{bmatrix}$	Weibull	$\begin{bmatrix}0.005\\1.5\end{bmatrix}$	$\begin{bmatrix}3.0\\2.0\end{bmatrix}$	$\begin{bmatrix}0.3\\0.5\end{bmatrix}$	$U_1(5,9)$ $U_2(2,6)$
M_2	200	2	1	2	$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	Gompertz	$\begin{bmatrix}0.2\\0.7\end{bmatrix}$	$\begin{bmatrix}1.5\\2.0\end{bmatrix}$	$\begin{bmatrix}0.8 & 0.1\\ -0.6 & 0.1\end{bmatrix}$	$U_1(4,9)$ $U_2(4,9)$
M_3	400	2	2	1	$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	Weibull	$\begin{bmatrix}0.003\\0.002\end{bmatrix}$	$\begin{bmatrix}0.5\\0.7\end{bmatrix}$	$\begin{bmatrix}0.8 & -0.5\\ -0.6 & 0.5\end{bmatrix}$	$U_1(12,15)$ $U_2(10,13)$
M_4	400	3	1	1	$\begin{bmatrix}0.35\\0.30\\0.35\end{bmatrix}$	Gompertz	$\begin{bmatrix}0.0002\\0.002\\0.0003\end{bmatrix}$	$\begin{bmatrix}0.7\\2.0\\0.8\end{bmatrix}$	$\begin{bmatrix}-0.8\\0.2\\1.0\end{bmatrix}$	$U_1(10,15)$ $U_2(4,6)$ $U_3(15,20)$

1: sample size; 2: number of risk types; 3: number of covariates; 4: degree of models; 5: baseline hazard function; 6: censored times are generated from a uniform distribution $U_i(a,b)$ for i=1,…,g.

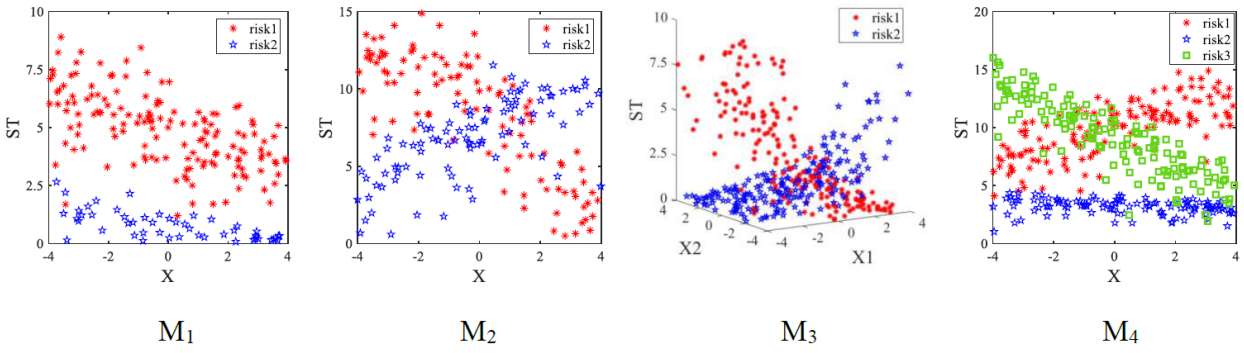

Figure 3. The scatter plot of a sample data set for models M_1~M_4 respectively.

4.1. Compare two methods of estimating the baseline hazard function

We consider an EM-based semi-parametric mixture hazards model to analyze simulated data, and compare the two methods of estimating the baseline hazard function. For the first method proposed by Ng and McLachlan [5], they assume the baseline hazard function is piecewise constant and calculate this function using maximum likelihood estimation (MLE). For the second method introduced in this paper, we use a kernel estimator to estimate the baseline hazard rates and choose biweight as the kernel function.

In order to graphically demonstrate the results, we first show the results for a single run of simulation in Table 3 and Figure 4. The correct rate (*CR*) in Table 3 is the percentage of individuals matched into the true attributable type of risk. According to the results of the estimation, we match the individuals into one type of risk with largest posterior probability. Thus, this correct rate is defined as:

$$CR = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{g} I\left\{ j \in risk(i) \cap \hat{z}_{ij} = \max_{i}(\hat{\boldsymbol{Z}}_j) \right\} \quad \text{where} \quad \hat{\boldsymbol{Z}}_j = (\hat{z}_{1j}, \hat{z}_{2j}, ..., \hat{z}_{gj})^T .$$

When using a piecewise constant estimator under M_1, from the estimated mixing probabilities, *CR* in Table 3 and the expectation line in Figure 4(M_1-1), it can be seen that we will misclassify some data into the 2nd type of risk where their true risk type is the 1st one. As a result, the estimates of regression coefficients in Table 3 and the cumulative baseline hazard rate in Figure 4(M_1-2) are not close to the true model. Furthermore, under M_4, from the expectation line according to the 1st and 2nd types of risk in Figure 4(M_4-1), it can be seen that we will misclassify some data between the 1st and 2nd types of risk when using piecewise constant estimator. The estimates of regression coefficients in Table 3 and the cumulative baseline hazard rate in Figure 4(M_4-2) are mismatched with the real model. It is obvious that using the kernel procedure for the baseline hazard estimation will increase CR compared to using the piecewise constant procedure.

We next show the results for 1000 simulations in Table 4. The absolute relative bias (ARB) for parameter θ is defined by:

$$ARB(\theta) = \left| \frac{E(\hat{\theta}) - \theta}{E(\hat{\theta})} \right| .$$

In Table 4 the mean absolute relative bias ($\overline{ARB}$) of the model with k parameters is defined by $\overline{ARB} = \sum_{i=1}^{k} ARB(\theta_i)/k$. Moreover, $\overline{CR}$ and $\overline{MsSSE}$ are the mean of CR and $MsSSE/n$ for each simulation. Table 4 presents that the $\overline{ARB}$ and $\overline{MsSSE}$ of the kernel estimate are smaller than those of the piecewise constant estimate. Moreover, the $\overline{CR}$ of the kernel estimate is larger than that of the piecewise constant estimate in all cases considered. Thus, the model with the baseline hazard functions estimated by the kernel method fits the data better than that with piecewise constant baseline.

Table 3. The estimation of a simulated series by models M_1~M_4 respectively.

		p	**β**	*CR*	*MsSSE*/*n*
M_1	True[1]	$\begin{bmatrix}0.7\\0.3\end{bmatrix}$	$\begin{bmatrix}0.3\\0.5\end{bmatrix}$		
	Piecewise[2]	$\begin{bmatrix}0.561\\0.439\end{bmatrix}$	$\begin{bmatrix}0.528\\0.851\end{bmatrix}$	0.860	0.810
	Kernel[3] ,bw[4]=1.0	$\begin{bmatrix}0.672\\0.328\end{bmatrix}$	$\begin{bmatrix}0.336\\0.586\end{bmatrix}$	0.945	0.659
M_2	True	$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	$\begin{bmatrix}0.8 & 0.1\\-0.6 & 0.1\end{bmatrix}$		
	Piecewise	$\begin{bmatrix}0.641\\0.958\end{bmatrix}$	$\begin{bmatrix}0.674 & 0.136\\-1.136 & 0.298\end{bmatrix}$	0.705	0.963
	Kernel, bw=0.5	$\begin{bmatrix}0.523\\0.476\end{bmatrix}$	$\begin{bmatrix}0.738 & 0.078\\-0.762 & 0.146\end{bmatrix}$	0.855	0.910
M_3	True	$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	$\begin{bmatrix}0.8 & -0.5\\-0.6 & 0.5\end{bmatrix}$		
	Piecewise	$\begin{bmatrix}0.508\\0.491\end{bmatrix}$	$\begin{bmatrix}0.993 & -0.562\\-0.562 & 0.608\end{bmatrix}$	0.838	1.240
	Kernel, bw = 0.4	$\begin{bmatrix}0.478\\0.522\end{bmatrix}$	$\begin{bmatrix}0.885 & -0.534\\-0.628 & 0.521\end{bmatrix}$	0.843	1.142
M_4	True	$\begin{bmatrix}0.35\\0.30\\0.35\end{bmatrix}$	$\begin{bmatrix}-0.8\\0.2\\1.0\end{bmatrix}$		
	Piecewise	$\begin{bmatrix}0.399\\0.265\\0.335\end{bmatrix}$	$\begin{bmatrix}-0.938\\0.920\\1.137\end{bmatrix}$	0.693	1.211
	Kernel, bw=0.9	$\begin{bmatrix}0.368\\0.306\\0.325\end{bmatrix}$	$\begin{bmatrix}-0.806\\0.192\\0.927\end{bmatrix}$	0.873	0.828

1: true parameters; 2: piecewise constant estimates; 3: kernel estimates; 4: bandwidth.

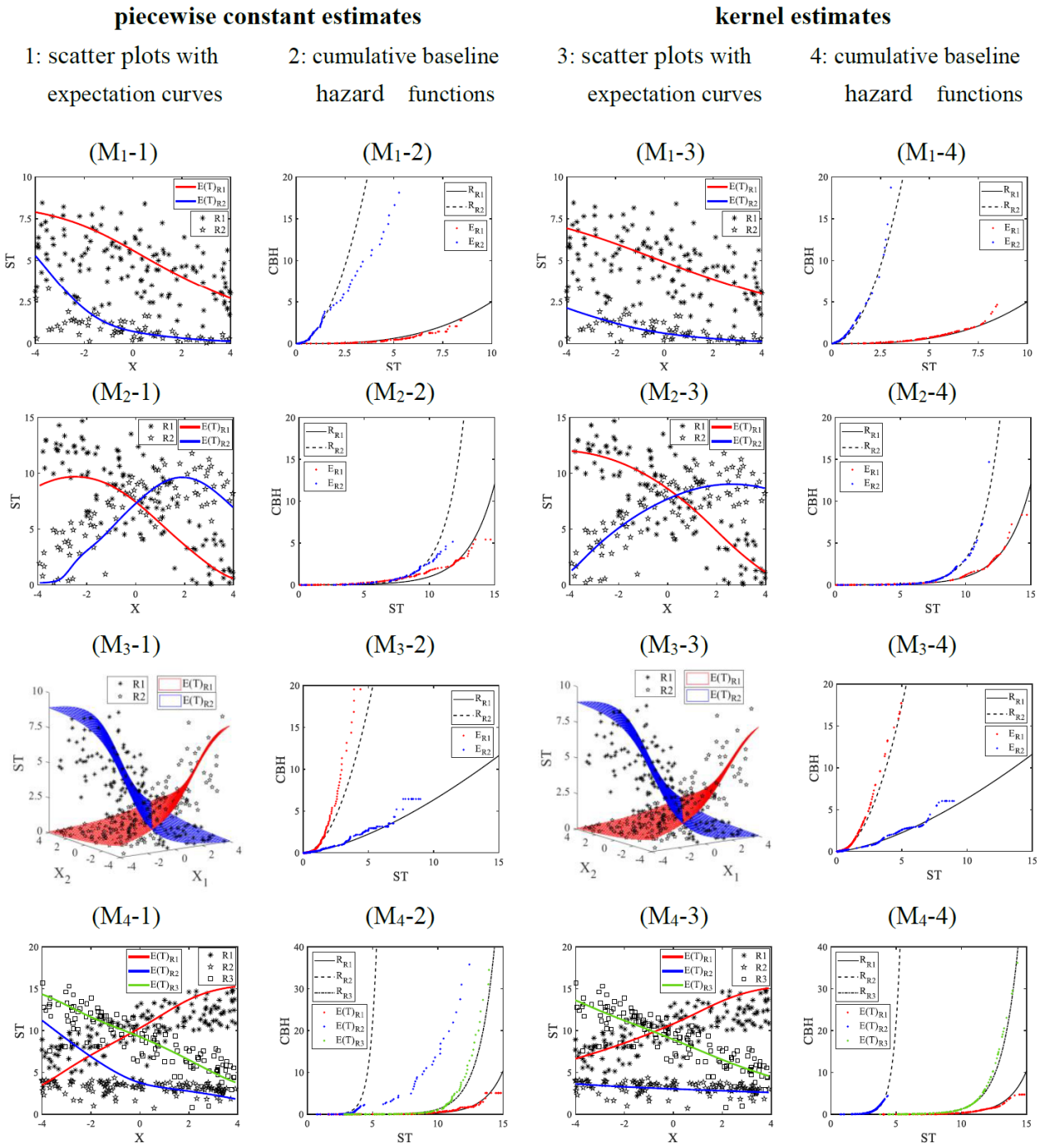

Figure 4. Plots of a single run simulated series by models M_1~M_4 respectively; $E(T)_{Ri}$: the i^{th} type of risk for estimation; CBH: cumulative baseline hazard rate; R_{Ri}: real cumulative baseline hazard function for the i^{th} type of risk; (M_i-j): model M_i is fitted and j = 1, 3: the scatter plots of the observed data and the estimated expectation curves; j = 2, 4: plots of the estimated cumulative baseline hazard functions; j = 1, 2: estimated by piecewise constant assumption; j = 3,4: estimated by kernel method.

Table 4. The estimation of 1000 simulated series by models M_1~M_4 respectively.

		bias_$\mathbf{p}$ [3]	MSE_$\mathbf{p}$ [4]	bias_$\boldsymbol{\beta}$ [5]	MSE_$\boldsymbol{\beta}$ [6]	$\overline{ARB}$	$\overline{CR}$	$\overline{MsSSE}$
M_1	Piecewise[1]	$\begin{bmatrix}-0.160\\0.160\end{bmatrix}$	$\begin{bmatrix}0.026\\0.026\end{bmatrix}$	$\begin{bmatrix}0.088\\0.275\end{bmatrix}$	$\begin{bmatrix}0.020\\0.076\end{bmatrix}$	0.401	0.699	0.796
	Kernel[2]	$\begin{bmatrix}-0.035\\0.035\end{bmatrix}$	$\begin{bmatrix}0.002\\0.002\end{bmatrix}$	$\begin{bmatrix}-0.073\\-0.007\end{bmatrix}$	$\begin{bmatrix}0.007\\0.000\end{bmatrix}$	0.107	0.856	0.653
M_2	Piecewise	$\begin{bmatrix}0.132\\-0.132\end{bmatrix}$	$\begin{bmatrix}0.017\\0.017\end{bmatrix}$	$\begin{bmatrix}-0.097 & 0.041\\-0.652 & 0.172\end{bmatrix}$	$\begin{bmatrix}0.010 & 0.001\\0.429 & 0.029\end{bmatrix}$	0.646	0.680	1.329
	Kernel	$\begin{bmatrix}0.089\\-0.089\end{bmatrix}$	$\begin{bmatrix}0.008\\0.008\end{bmatrix}$	$\begin{bmatrix}-0.123 & 0.054\\-0.311 & 0.017\end{bmatrix}$	$\begin{bmatrix}0.018 & 0.006\\0.124 & 0.000\end{bmatrix}$	0.292	0.774	1.009
M_3	Piecewise	$\begin{bmatrix}0.028\\-0.028\end{bmatrix}$	$\begin{bmatrix}0.000\\0.000\end{bmatrix}$	$\begin{bmatrix}0.167 & -0.091\\-0.079 & 0.046\end{bmatrix}$	$\begin{bmatrix}0.028 & 0.008\\0.006 & 0.002\end{bmatrix}$	0.122	0.847	1.271
	Kernel	$\begin{bmatrix}-0.006\\0.006\end{bmatrix}$	$\begin{bmatrix}0.000\\0.000\end{bmatrix}$	$\begin{bmatrix}0.033 & -0.020\\0.069 & -0.051\end{bmatrix}$	$\begin{bmatrix}0.001 & 0.000\\0.004 & 0.002\end{bmatrix}$	0.054	0.849	1.097
M_4	Piecewise	$\begin{bmatrix}0.043\\-0.055\\0.012\end{bmatrix}$	$\begin{bmatrix}0.001\\0.003\\0.000\end{bmatrix}$	$\begin{bmatrix}-0.003\\0.791\\0.251\end{bmatrix}$	$\begin{bmatrix}0.002\\0.627\\0.063\end{bmatrix}$	0.766	0.646	0.737
	Kernel	$\begin{bmatrix}0.018\\-0.042\\0.023\end{bmatrix}$	$\begin{bmatrix}0.000\\0.001\\0.000\end{bmatrix}$	$\begin{bmatrix}0.032\\0.071\\-0.014\end{bmatrix}$	$\begin{bmatrix}0.002\\0.009\\0.000\end{bmatrix}$	0.112	0.799	0.565

1: piecewise constant estimates; 2: kernel estimates; 3: bias of $\mathbf{p}$; 4: mean square error (MSE) of $\mathbf{p}$; 5: bias of $\boldsymbol{\beta}$; 6: mean square error (MSE) of $\boldsymbol{\beta}$.

4.2. Select appropriate number of model components

In section 4.2, we consider an EM-based semi-parametric mixture hazards model to analyze simulated data under models M_1~M_4 by considering several possible number of risk types, that is model components, and use the kernel estimator to estimate the baseline hazard rates with biweight as the kernel function. Next, we use validity indices to select the optimal number of model components. The following six validity indices are used to compare with the validity indices we have come up with (V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS}).

1. Partition coefficient V_{PC} proposed by Bezdek [22].

2. Normalized partition coefficient V_{NPC} proposed by Dave [23].

3. Partition entropy V_{PE} proposed by Bezdek [24].

4. Normalized partition entropy V_{NPE} proposed by Dunn [25].

5. Akaike information criterion AIC.
6. Bayesian information criterion BIC.

It is well known that memberships play an important role in fuzzy clustering. Similarly, under the EM-based mixture model, the posterior probabilities are closely related to the role of memberships. Therefore, we replace the role of memberships with posterior probabilities in the validity indices V_{PC}, V_{NPC}, V_{PE}, and V_{NPE}. Moreover, the formulas for *AIC* and *BIC* are computed by

$$AIC = -2 \cdot l_c(\hat{\boldsymbol{p}}, \hat{\boldsymbol{\beta}}) + 2k; \quad BIC = -2 \cdot l_c(\hat{\boldsymbol{p}}, \hat{\boldsymbol{\beta}}) + k \log(n),$$

where $l_c(\hat{\boldsymbol{p}}, \hat{\boldsymbol{\beta}})$ is the complete-data log-likelihood (3) given the estimated parameters, and *k* is the number of parameters for estimation.

All in all we consider ten indices, including V_{PC}, V_{NPC}, V_{PE}, V_{NPE}, *AIC*, *BIC*, V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS}, to select the optimal number of model components. Table 5 shows the proportion of choosing the correct number of model components over 1000 simulation runs based on the considered indices respectively. In each simulation run, each model of M_1~M_4 is fitted for components 2, 3, and 4 separately. Note that we assume the number of model components is greater than one for satisfying the requirement of the proposed validity indices. We define the proportion of choosing the correct number of risk types by each index in Table 5 as:

$$\frac{\#(\text{choose correct g by index})}{\#(\text{simulaiton})}.$$

Table 5. The proportion of choosing the correct g by each index for 1000 simulation runs under models M_1~M_4 respectively.

	V_{PC}	V_{NPC}	V_{PE}	V_{NPE}	*AIC*	*BIC*	V_{aRaS}	V_{sRaS}	V_{aRmS}	V_{sRmS}
M_1	0.962	0.880	0.976	0.894	0.964	0.950	0.896	0.896	0.984	0.992
M_2	0.954	**0.564**	0.963	**0.485**	**0.524**	**0.631**	0.863	0.851	0.981	0.990
M_3	1.000	0.798	1.000	0.868	0.998	0.998	0.994	0.998	1.000	1.000
M_4	**0.486**	0.780	**0.413**	0.810	**0.646**	**0.660**	0.923	0.916	0.813	0.703

Table 5 shows that the proportion of choosing the correct g by traditional indices V_{PC}, V_{NPC}, V_{PE}, V_{NPE}, *AIC*, and *BIC* are not consistent under models M_1~M_4, where at least one model is not performing well (denoted by fluorescent yellow color in the table). On the other hand, the proposed indices (V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS}) are consistent and possess high proportions for every model, except that the proportion of V_{sRmS} under M_4 is 0.703, which is slightly low, but it is still higher than

that of most traditional indices. Hence, the proposed validity indices are superior than others in selecting the correct number of components.

5. Analysis of prostate cancer data

As a practical illustration of the proposed EM-based semi-parametric mixture hazard model, we consider the survival times of 506 patients with prostate cancer who entered a clinical trial during 1967–1969. These data were randomly allocated to different levels of treatment with the drug diethylstilbestrol (DES) and were considered by Byar and Green [26] and published by Andrews and Herzberg [27]. Kay [28] analyzed a subset of the data by considering eight types of risk, defined by eight categorical variables: drug treatment (RX: 0, 0.0 or 0.2 mg estrogen; 1, 1.0 or 5.0 mg estrogen); age group (AG: 0, < 75 years; 1, 75 to 79 years; 2, > 79 years); weight index (WT: 0, > 99 kg; 1, 80 to 99 kg; 2, < 80 kg); performance rating (PF: 0, normal; 1, limitation of activity); history of cardiovascular disease (HX: 0, no; 1, yes); serum haemoglobin (HG: 0, > 12 g/100 ml; 1, 9 to 12 g/100 ml; 2, < 9 g/100 ml); size of primary lesion (SZ: 0, < 30 cm^2; 1, ≥ 30 cm2), and Gleason stage/grade category (SG: 0, ≤ 10; 1, > 10). Cheng et al. [26] classified this dataset with three types of risk as: (1) death due to prostate cancer; (2) death due to cardiovascular (CVD) disease; and (3) other causes.

We analyze the same dataset with eight categorical variables (RX, AG, WT, PF, HX, SZ, SG). There are 483 patients with complete information on these covariates, and the proportion of censored observations is 28.8%. We ignore the information about the risk factors and use indices, including V_{PC}, V_{NPC}, V_{PE}, V_{NPE}, *AIC*, *BIC*, V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS} to select the optimal number of risk types. From Table 6, the number of risk types selected by V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS} is three, and that selected by other indices is two. The number of model components selected by the indices we have proposed is the same as that in the previous studies introduced by Cheng et al. [3].

Table 6. The value of each index with different number of risk types under prostate cancer data.

	V_{PC}	V_{NPC}	V_{PE}	V_{NPE}	*AIC*	*BIC*	V_{aRaS}	V_{sRaS}	V_{aRmS}	V_{sRmS}
$g = 2$	**0.7813**	**0.5626**	**0.3369**	**0.5720**	**4.1518**	**4.2989**	0.5894	0.4437	0.5894	0.4437
$g = 3$	0.6684	0.5027	0.5260	0.6135	4.5012	4.7262	**0.3783**	**0.1974**	**0.5016**	**0.2943**
$g = 4$	0.5581	0.4109	0.7564	0.7075	4.7967	5.0996	0.4746	57.572	0.6123	98.534

Note: (1) g represents the number of risk types when estimating. (2) The optimal values of g according to each index are highlighted in bold.

From existing medical experience and a previous study, we presume that these model components may agree with some particular types of risk and thus can decide whether there are significant relationships between the covariates and the survival times by using the Wald statistical test. Based on the cause-specific hazard approach, Cheng et al. [3] found that treatment with a higher DES dosage (RX = 1) significantly reduces the risk of death due to prostate cancer. Table 7 shows that the DES dosage has a significant effect on time to death due to the 1st type of risk, and that the estimated regression coefficients of RX is negative. Byar and Green [26] found that patients with a

big size of primary lesion (SZ = 1) and high-grade tumors (SG = 1) are at greater risk of prostate cancer death. Table 7 lists that SZ and SG have a significant effect on time to death due to the 1st type of risk, and that the estimated regression coefficients are all positive. Accordingly, we presume the 1st type of risk relates to prostate cancer. Furthermore, based on the cause-specific hazard approach, Cheng et al. [3] found that treatment with a higher DES dosage (RX = 1) significantly increases the risk of death due to CVD. From Table 7, we see that DES dosage has a significant effect on time to death due to the 2nd and 3rd types of risk, and that the estimated regression coefficient of RX is positive.

We know that patients with a history of cardiovascular disease (HX = 1) have a higher probability of death due to CVD, compared to those patients without such a history. Table 7 shows that the estimated regression coefficient of HX is positive due to the 2nd type of risk. Hence, we presume the 2nd type of risk may relate to CVD. There is no explicit relationship between covariates and survival times adhering to the 3rd type of risk. Thus, we only presume the 3rd type of risk may relate to other death causes without specification. According to the significant relationship of covariates and survival times, we assess that the 1st, 2nd and 3rd types of risk for estimation from an EM-based semi-parametric mixture hazard model are classified to prostate cancer, CVD, and other unspecified causes, respectively.

Table 7. The model estimates (with standard errors) of prostate cancer data given the number of risk types equal to 3.

		1st type of risk	2nd type of risk	3rd type of risk
p		0.2132	0.3930	0.3936
β	RX	−0.0296*(0.1267)	0.3546*(0.1414)	0.7589*(0.1425)
	AG	0.3144*(0.1143)	1.7445*(0.1041)	1.8104*(0.1396)
	WT	−0.0817*(0.0916)	1.7915*(0.0967)	−0.5555*(0.1290)
	PF	1.4742*(0.2233)	0.1244*(0.2527)	1.6468*(0.3325)
	HX	3.0027*(0.1176)	1.2829*(0.1377)	−0.6092*(0.1486)
	HG	0.8489*(0.1536)	1.6074*(0.1669)	−5.2153*(0.7267)
	SZ	0.8567*(0.2119)	3.0334*(0.1998)	−3.2661*(0.4074)
	SG	4.3184*(0.1010)	−0.3907*(0.1419)	−0.9933*(0.1560)

*Note: * denotes P-value < 0.05.*

6. Conclusions and discussion

6.1. Conclusions

This study introduces four new validity indices, V_{aRaS}, V_{sRaS}, V_{aRmS}, and V_{sRmS}, for deciding the number of model components when applying an EM-based Cox proportional hazards mixture model to a dataset of competing risks. We incorporate the posterior probabilities and the sum of standard residuals to constitute the new validity indices. Moreover, our study sets up an extended kernel approach to estimate the baseline functions more smoothly and accurately. Extensive simulations show that the kernel procedure for the baseline hazard estimation is helpful for increasing the correct

rate of classifying individual into the true attributable type of risk. Furthermore, simulation results demonstrate that the proposed validity indices are consistent and have a higher percentage of selecting the optimal number of model components than the traditional competitors. Thus, the proposed indices are superior to several traditional indices such as the most commonly used in statistics, AIC and BIC. We also employ the propose method to a prostate cancer data-set to illustrate its practicability.

6.2. Discussion

It is obvious that if we apply the four new validity indices at the same time, then we have the best chance to select the optimal number of model components. One concern is picking the best one among the proposed validity indices. In fact, the average separation versions (V_{aRaS}, V_{sRaS}) easily neutralizes the effects of small and large distances among the expectations of component models. On the other hand, as long as there is a small distance among the expectations of component models, the minimum separation versions (V_{aRmS}, V_{sRmS}) will catch the information about the overfitting model. Under the analysis of prostate cancer data, we see that V_{aRmS} and V_{sRmS} behave more sensitively than V_{aRaS} and V_{sRaS} for detecting the overfitting models (i.e., the distances of indices between overfitting and optimal models are much larger than those between underfitting and optimal models). Furthermore, according to the simulation results, the index V_{sRmS} performs slightly poor on a certain model, we thus recommend employing V_{aRmS} if just one of the proposed validity indices is to be used.

In the future we may test the effectiveness of the proposed validity indices on statistical models other than the mixture Cox proportional hazards regression models. We could also advance the efficiency of the proposed indices in determining the number of components of mixture models. Another issue is to reduce the computation cost. For instance, the bandwidth of the kernel procedure for baseline hazard function estimates is recalculated on each iteration, which consumes computation time. All these factors need further investigation and will be covered in our future research.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and suggestions which have greatly improved this article. This work was partially supported by the Ministry of Science and Technology, Taiwan [Grant numbers MOST 108-2118-M-025-001 - and MOST 108-2118-M-003-003 -].

References

1. D. R. Cox, Regression models and life-tables with discussion, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **34** (1972), 187–220.

2. G. Escarela, and R. Bowater, Fitting a semi-parametric mixture model for competing risks in survival data, *Commun. Stat.-Theory Methods*, **37** (2008), 277–293.
3. S. C. Cheng, J. P. Fine, and L. J. Wei, Prediction of cumulative incidence function under the proportional hazards model, *Biometrics* **54** (1998), 219–228.
4. G. J. McLachlan and D. Peel, *Finite mixture models*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000.
5. S. K. Ng and G. J. McLachlan, An EM-based semi-parametric mixture model approach to the regression analysis of competing risks data, *Stat. Med.*, **22** (2003), 1097–1111.
6. I. S. Chang, C. A. Hsiung, C. C. Wen and W. C. Yang, Non-parametric maximum likelihood estimation in a semiparametric mixture model for competing risks data, *Scand. J. Stat.*, **34** (2007), 870–895.
7. W. Lu and L. Peng, Semiparametric analysis of mixture regression models with competing risks data, *Lifetime Data Anal.*, **14** (2008), 231–252.
8. S. Choi and X. Huang, Maximum likelihood estimation of semiparametric mixture component models for competing risks data, *Biometrics*, **70** (2014), 588–598.
9. A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
10. Y. G. Tang, F. C. Sun and Z. Q. Sun, Improved validation index for fuzzy clustering, in *American Control Conference, June 8-10, 2005. Portland, OR, USA*, (2005), 1120–1125.
11. W. Wang and Y. Zhang, On fuzzy cluster validity indices, *Fuzzy Sets Syst.*, **158** (2007), 2095–2117.
12. K. L. Wu, M. S. Yang and J. N. Hsieh, Robust cluster validity indexes, *Pattern Recognit.*, **42** (2009), 2541–2550.
13. K. L. Zhou, S. Ding, C. Fu and S. L. Yang, Comparison and weighted summation type of fuzzy cluster validity indices, *Int. J. Comput. Commun. Control*, **9** (2014), 370–378.
14. J. M. Henson, S. P. Reise and K. H. Kim, Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics, *Struct. Equ. Modeling*, **14** (2007), 202–226.
15. J. R. Busemeyer, J. Wang, J. T. Townsend and A. Eidels, *The Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press 2015.
16. R. Bender, T. Augustin and M. Blettner, Generating survival times to simulate Cox proportional hazards models, *Stat. Med.*, **24** (2005), 1713–1723.
17. P. Royston, Estimating a smooth baseline hazard function for the Cox model. Technical Report No. 314. University College London, London 2011. Available from: https://www.semanticscholar.org/paper/Estimating-a-smooth-baseline-hazard-function-for-Royston/2f329b48f674a74253eb428b71ff237365fd4051.
18. A. Guilloux, S. Lemler and M. L. Taupin, Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates, *J. Multivar. Anal.*, **148** (2016) 141–159.
19. M. Zhou, *Empirical likelihood method in survival analysis*, CRC Press 2016
20. I. Horova, J. Kolacek, and J. Zelinka, *Kernel smoothing in Matlab theory and practice of kernel smoothing*, World Scientific Publishing, 2012.
21. P. N. Patil, Bandwidth choice for nonparametric hazard rate estimation, *J. Stat. Plan. Infer.*, **35** (1993), 15–30.
22. J. C. Bezdek, Numerical taxonomy with fuzzy sets, *J. Math. Biol.*, **7** (1974), 57–71.

23. R. N. Dave, Validating fuzzy partition obtained through c-shells clustering, *Pattern Recognit. Lett.*, **17** (1996), 613–623.
24. J. C. Bezdek, Cluster validity with fuzzy sets, Journal of Cybernetics, 3 (1974), 58–73. DOI: 10.1080/01969727308546047.
25. J. C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, in *Fuzzy Automata and Decision Processes*, (ed. M.M. Gupta, Elsevier), NY, 1977.
26. D. P. Byar and S. B. Green, The choice of treatment for cancer patients based on covariate information: application to prostate cancer, *Bull. Cancer*, **67** (1980), 477–490.
27. D. F. Andrews and A. M. Herzberg, *Data: a collection of problems from many fields for the student and research worker*, Springer-Verlag, New York, (1985), 261–274.
28. R. Kay, Treatment effects in competing-risks analysis of prostate cancer data, *Biometrics*, **42** (1986), 203–211.

AIMS Press