Risk group detection and survival function estimation for interval coded survival methods

Vanya Van Belle[1,2], Patrick Neven[3,4], Vernon Harvey[5], Sabine Van Huffel[1], Johan A. K. Suykens[1], and Stephen Boyd[6]

[1] Department of Electrical Engineering (ESAT-SCD), KU Leuven/iMinds Future Health Department, Leuven, Belgium
[2] Department of Mathematics and Statistics, Liverpool John Moores University, Liverpool, UK
[3] Department of Gynaecological Oncology, University Hospitals Leuven, Leuven, Belgium
[4] Multidisciplinary Breast Centre (MBC), University Hospitals Leuven, Leuven, Belgium
[5] Regional Cancer Centre, Auckland City Hospital, Auckland, New Zealand
[6] Department of Electrical Engineering, Stanford University, Stanford, California, United States of America

Abstract. The highly flexible model structure of methods in data mining and machine learning results in models that are often difficult to interpret. Their use in domains where interpretability is an issue is therefore hampered. In order to bridge the gap between advanced modeling techniques and their use in domains that demand interpretable results, the interpretability aspect should be included in the design of the technique. The Interval Coded Score index (ICS) is a recently proposed model that satisfies this condition and automatically detects thresholds on variables to generate score systems. The method was extended for censored data (ICSc) but two problems remain: (i) Given a prognostic index, how can observations be grouped in different risk groups; (ii) Given the risk groups, how can survival curves be estimated for survival models based on support vector machines or ICS models.
This work offers solutions to both of these problems. The ICSc model is used on the prognostic index to detect thresholds on this index. A grouped index, that can be interpreted as a risk group indicator, is the result. The method is then modified to ensure that observations with a lower prognostic index are allocated to higher risk groups. The second problem is tackled by simultaneously estimating multiple Kaplan-Meier curves, taking into account that the estimated survival curve for higher risk groups should always be lower than the curve for lower risk groups. The proposed approach is illustrated on the prognosis of breast cancer patients and compared with the proportional hazard model. Both models are comparable w.r.t. discrimination, but calibration is better for the ICSc risk groups.

1 Introduction

Methods within artificial intelligence and machine learning [1–4] have proved their use in many domains, including clustering, classification, regression [5] and prognosis [6–9]. Their ability to model complex data and to deal with the curse of dimensionality have made them very popular data modeling tools. However, in domains where interpretability is an issue, the black-box nature of these methods hampers their use in practice. In order to introduce the use of more complex mathematical methods [1, 2, 4, 8, 10–12] in these domains, different methods to obtain a score system have been proposed: optimal cut-points methods [13, 14], post-processing of previously built regression models [15], classification and regression trees [16], adaptive index models [17] and rule extraction methods [18]. All these methods result in categorizations of variables, with differences on whether the thresholds are defined before, during or after modeling, feature selection is included, thresholds are defined sequentially or simultaneously and whether predictions are given. The resulting models are easy to apply, but performance, interpretability of the results and use for different data types differ and

disadvantages remain: dependency on the choice and number of the thresholds, dependency of later thresholds on the choice of former thresholds, multi-testing problems, no optimal trade-off between sparsity and performance. The Interval Coded Score (ICS) method was recently proposed to solve these issues for classification problems [19] and survival data [20] (ICSc). The approach is based on transformation models [21–23] where a prognostic index is trained to be as concordant with the observed outcome as possible. It is assumed that there exists a monotonic relationship between this index and the outcome of interest. ICS models additionally assume additive models and restrict the functional forms of the inputs to step functions. As such, score models are generated with an automatic detection of the number and position of thresholds.

For use in real-life applications, the resulting score should be accompanied by an estimated survival function. Two approaches are possible. A first one assumes a baseline survival function, that can be changed according to the observed variables. The advantage is that a survival function can be estimated for each observation. However, an additional assumption, such as the proportional hazards assumption in a Cox model [10], is needed in this case. In this work, an alternative approach is taken. Observations are divided into different risk groups depending on their prognostic index. For each risk group, a survival function is estimated. In order to use this approach, two problems need to be tackled: (i) how to select the number of risk groups and how to define the thresholds in the prognostic index to allocate each observation to a risk group; (ii) Given the risk groups, how to estimate a survival function for each group, taking into account the assumptions of the model used to derive the prognostic index. To solve the first issue, standard clustering methods can not be used since they are unable to deal with censored data. Additionally, the number of clusters should be known in advance (see [24] for an exception). This work proposes to use the ICSc method with the previously developed prognostic index as a single input variable. A new index, which will be a grouped version of the prognostic index, is the result and can be interpreted as a risk group indicator. The method is modified with the inclusion of a monotonicity constraint (mICSc) to ensure that observations with a lower prognostic index are allocated to a higher risk group.
Once the risk groups are defined, a survival function for each of these could be estimated by means of a Kaplan-Meier (KM) curve [25]. However, this does not take into account that the model assumed non-crossing survival curves when training the prognostic index. A modified KM estimator is therefore proposed. The method is based on the inverse-probability-of-censoring weighted average estimator [26]. The resulting step functions are then smoothened by means of a monotonic regressor. An overview of this work is presented in Figure 1. The approach of this paper is summarized in algorithm 1. All methods were implemented in matlab[7] using `CVX`[8] [27].

The remainder of the paper is organized as follows. Section 2 starts with the description of support vector machines for survival analysis. Section 3 discusses how this method can be adapted to automatically obtain score systems for censored data (ICSc). Section 4 proposes a modification of ICSc to allow to cluster survival data after a survival model has generated a prognostic index, such that risk groups can be obtained. Section 5 proposes a new method to estimate survival curves, that takes the monotonicity assumption of the ICSc method into account. Section 6 illustrates the latter method on artificial data before applying the whole

[7] `http://www.mathworks.nl/products/matlab/`
[8] `http://cvxr.com/cvx/`

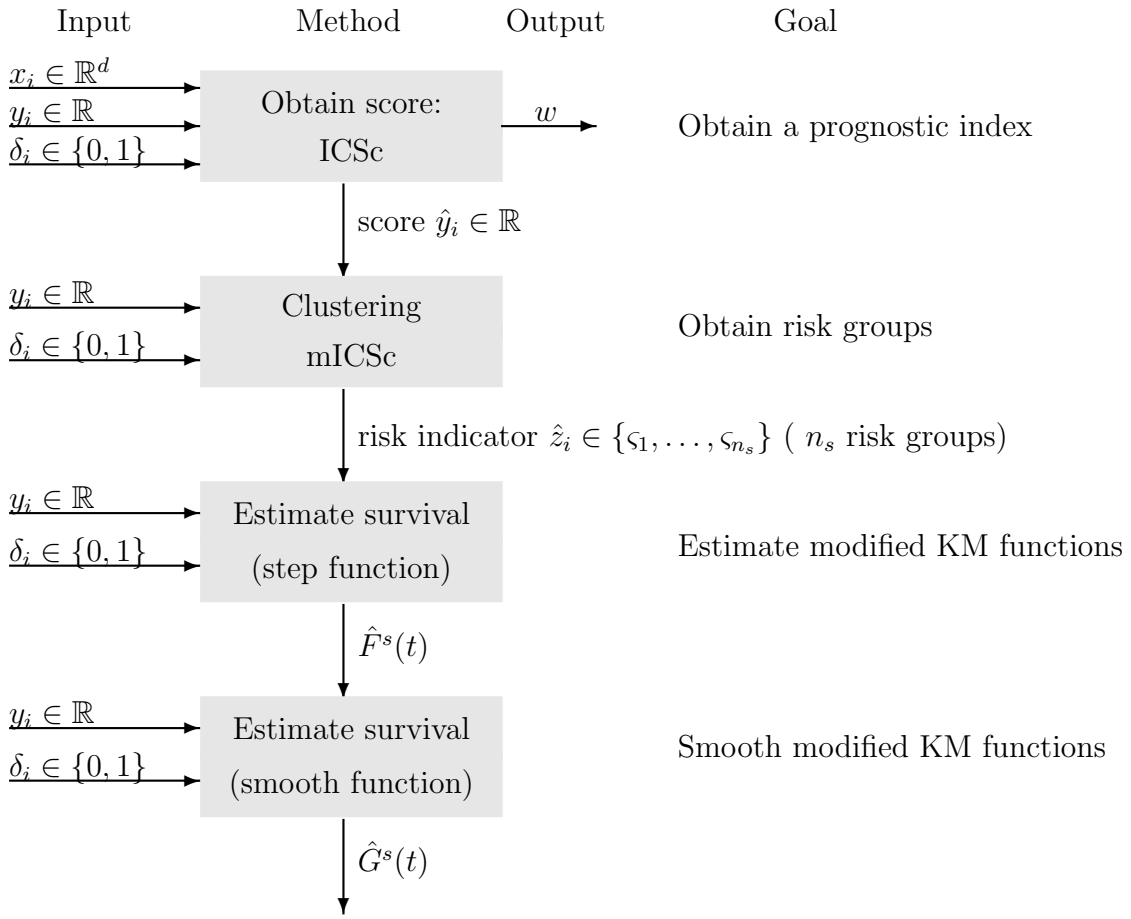


Fig. 1. Overview of this work. The data set $\mathcal{S} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$, representing the input variables ($x_i \in \mathbb{R}^d$), the outcome ($y_i \in \mathbb{R}$) and the censoring indicator ($\delta_i \in \{0,1\}$) are transformed into a score or prognostic index $\hat{y}_i \in \mathbb{R}$ by means of the ICSc approach given in Section 3. The method also results in a vector w indicating the selected set of variables, the selected intervals and their contribution to the score. In order to obtain a small number of risk groups, the scores are used as input variables for the mICSc approach (see Section 4). This results in a risk indicator $\hat{z}_i \in \mathbb{R}$, a clustered version of the scores $\hat{y}_i$, that can be interpreted as risk groups. For each risk group s, a step function $\hat{F}^s(t)$ is calculated using the approach of Section 5.1 to obtain an estimate of the survival function as $1 - \hat{F}^s(t)$. These step functions are smoothened according to the method of Section 5.2 to obtain smooth survival estimates $1 - \hat{G}^s(t)$.

procedure (see algorithm 1) on a large breast cancer dataset [28]. The results are compared with a proportional hazard model (PH model) [10]. Section 7 finalizes the paper.

Throughout the paper, the following notation will be adopted. Let $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$ be a dataset with $x_i \in \mathbb{R}^d$ a vector containing all input variables for observation i, $y_i = \min(t_i, c_i)$ the observed failure times, with t_i and c_i the true failure and censoring time, respectively. δ_i is an event indicator equal to $\delta_i = \mathcal{I}[y_i \leq c_i]$, with $\mathcal{I}[z]$ the indicator function equal to 1 when z is true and zero otherwise. The p^{th} input variable is denoted as x^p.

2 Support vector machines for censored data

Support vector machines (SVM) for classification or regression can not be used for the analysis of survival data due to the occurrence of censored data. The outcome of survival analysis is

Algorithm 1 Necessary steps to automatically obtain a score system with survival estimates for different risk groups.

1: Given the training data $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$, train the ICSc model to obtain a score $\hat{y}_i$ for each observation i.
2: Use the mICSc model with the scores $\hat{y}_i$ as a single input variable to obtain risk groups indicators $\hat{z}_i$ that can only take values in $\{\varsigma_1, \ldots, \varsigma_{n_s}\}$, where n_s is obtained from the method.
3: Determine step functions $\hat{F}^s(t), s = 1, \ldots, n_s$, that are monotonic w.r.t. the risk groups s as an estimate of the cumulative distribution for all risk groups simultaneously.
4: Smoothen the step functions $\hat{F}^s(t)$ to obtain smooth estimates $\hat{G}^s(t)$ of the cumulative distribution functions that are monotonic w.r.t. the risk groups.

the time until a predefined event occurs. However, observations can drop out of the study and the outcome will not be observed exactly: the outcome is censored. The most frequent censoring type is right censoring, and occurs when a lower bound on the outcome is known. In the remainder of this work, only right censoring will be considered.

In order to deal with censored data, support vector machines for survival analysis take a two-step approach [8]. In a first step, a prognostic index (also called utility, latent variable or score) that is as concordant as possible with the observed survival times, is trained under the assumption that a monotonic relation exists between the prognostic index and the outcome of interest. The prognostic index is optimized such that as many comparable pairs as possible are concordant. A pair of observations (x_i, y_i, δ_i) and (x_j, y_j, δ_j) is comparable when both observations have an observed event time, or when only one of them is censored and the censoring occurs later than the event. More formally, a pair $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ is comparable if:

$$(\delta_i = 1 \ \& \ \delta_j = 1)$$
$$\text{or}$$
$$(\delta_i = 1 \ \& \ \delta_j = 0 \ \& \ y_i \leq y_j)\,.$$

A comparable pair is considered concordant when the ranking in observed survival time y_i and y_j is the same as the ranking in the prognostic index.

This work is based on the SVM survival model with ranking and regression constraints as proposed in [9]:

$$\min_{w,b,\epsilon,\xi,\xi^*} \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$
$$\text{subject to}$$
$$\begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall\ i = 1, \ldots, n, \\ w^T \varphi(x_i) + b \geq y_i - \xi_i, & \forall\ i = 1, \ldots, n, \\ -\delta_i (w^T \varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\ i = 1, \ldots, n, \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n, \\ \xi_i \geq 0, & \forall\ i = 1, \ldots, n, \\ \xi_i^* \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \tag{1}$$

with

$$\bar{j}(i) = \underset{j}{\arg\max} \quad j$$
$$\text{subject to} \ \begin{cases} (x_i, y_i, \delta_i) \text{ and } (x_j, y_j, \delta_j) \text{ comparable} \\ y_j < y_i\,. \end{cases} \tag{2}$$

The first constraint in (1) is a ranking constraint. The second and third constraint are the regression constraints. $\bar{j}(i)$ indicates the observation within the training set, that is comparable with observation i, with a survival time the closest to that of observation i. The estimated outcome $\hat{y}_\star$ for a new point $x_\star$ is then found as $\hat{y}_\star = w^T\varphi(x_\star) + b$.

3 Interval coded score system for censored data

In order to obtain a score system, model (1) is adapted in three ways [19, 20]. Firstly, the model is constrained to be additive [29]. Secondly, the estimated functional forms are restricted to be step functions, closely related to constant B-spline functions [30]. The range of each variable x^p is divided into k_p consecutive intervals. The functional form of this variable is then defined as $\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}[\theta_{p,l-1} \leq x_i^p < \theta_{p,l}]$, namely a linear combination of binary indicators denoting whether the value of the variable is within each of the k_p intervals. Lastly, in order to obtain a sparse model representation, the total variation of the coefficients vector w is minimized [31]. The problem to be optimized then becomes:

$$\min_{w,\hat{y},b,\epsilon,\xi,\xi^*} \sum_{p=1}^{d}\sum_{l=1}^{k_p+1} \chi_{p,l}|w_{p,l} - w_{p,l-1}| + \gamma\sum_{i=1}^{n}\epsilon_i + \mu\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to

$$\begin{cases} \hat{y}_i = \sum_{p=1}^{d}\left(\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}[\theta_{p,l-1} \leq x_i^p < \theta_{p,l}]\right) + b, \forall\, i = 1,\ldots,n\,, \\ \hat{y}_i - \hat{y}_{\bar{j}(i)} \geq y_i - y_{\bar{j}(i)} - \epsilon_i, \qquad \forall\, i = 1,\ldots,n, \\ \hat{y}_i \geq y_i - \xi_i, \qquad \forall\, i = 1,\ldots,n, \\ -\delta_i\hat{y}_i \geq -\delta_i y_i - \xi_i^*, \qquad \forall\, i = 1,\ldots,n, \\ \epsilon_i \geq 0, \qquad \forall\, i = 1,\ldots,n, \\ \xi_i \geq 0, \qquad \forall\, i = 1,\ldots,n, \\ \xi_i^* \geq 0, \qquad \forall\, i = 1,\ldots,n\,. \end{cases} \tag{3}$$

Note that $\hat{y}$ is eliminated before solving the model in $w, b, \epsilon, \xi, \xi^*$. In first instance $\chi_{p,l} = 1,\ \forall\ p = 1,\ldots,d,\ \forall\ l = 1,\ldots,k_p+1$. In order to further improve the sparsity of the model, the method is iteratively reweighted [32] with $\chi_{p,l} = \frac{1}{\varepsilon + a|w_{p,l} - w_{p,l-1}|}$. Here ε is a small positive value (e.g. 0.0005) and the value of a is optimized for the problem at hand.

In order to make the score system easily applicable, the weights are normalized such that the smallest non-zero absolute value of the coefficients (ν) becomes 1. All other normalized coefficients are rounded to the nearest integer : $\tilde{w}_{p,l} = [w_{p,l}/\nu]$. The final score for a new observation $x_\star$ is then found as

$$\hat{y}_\star = \sum_{p=1}^{d}\left(\sum_{l=1}^{k_p+1} \tilde{w}_{p,l}\mathcal{I}[\theta_{p,l-1} \leq x_\star^p < \theta_{p,l}]\right) + b\,. \tag{4}$$

Application of this procedure results in the Interval Coded Score index for censored data (ICSc) .

4 Obtaining risk groups

Once a score or prognostic index is trained, risk groups need to be defined for application in practice. Most often, three risk groups are considered: a low, moderate and high risk group. However, the choice for three groups is artificial and no statistical ground exists to support this choice. A second problem is how the groups should be defined. A first possibility is to use clustering methods to define clusters using the inputs of the observations and/or survival time. However, these methods are based on a distance measure between all pairs of data points. Clustering survival data is therefore difficult since calculating a distance in survival time is not always possible due to censoring, corresponding to uncertainty about the survival time. Using clustering mechanisms on the score obtained from the ICSc method for example is not an option either. This score is only defined up to a monotonic relation and a distance of a between the scores of two observations thus has another meaning depending on the exact value of the score. Additionally, clustering mechanisms do not take into account that the prognostic index defines a ranking on the risk groups: the higher (lower) the index, the higher[9] (higher)[10] the risk. Another possibility is to define a grid of possible thresholds on the prognostic index and select the combination of thresholds that leads to the largest difference in survival curves (by means of the log-rank test for example) on bootstrap samples of the original data. However, the number of groups still needs to be defined in advance.

In order to solve these issues, another approach that uses a shrinkage mechanism for clustering as in [33], is proposed here. The goal is to find a categorization of the prognostic index that maintains the concordance with the outcome as much as possible, without the need to define the number of categories in advance. Since ICSc is a survival model that is able to define relevant intervals on each of the input variables, and controls for the loss of information as a result of the categorization, this method can be used as a clustering/categorization method for survival data when the prognostic index is used as a single input variable. However, an extra adaptation is necessary to ensure that the identified risk groups have a monotonic relationship with the prognostic index. The outcome w of the ICSc model represents the additional effect of each interval on the survival. To express that a higher prognostic index indicates a lower risk (for models modeling the survival such as ICSc) it is necessary that w increases with the prognostic index. The adapted method will therefore be referred to as the monotonic ICSc model (mICSc) and is obtained from:

[9] for methods that model the risk
[10] for methods that model the survival

$$\min_{w,\hat{z},b,\epsilon,\xi,\xi^*} \sum_{l=1}^{k+1} \chi_l |w_l - w_{l-1}| + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

subject to

$$\begin{cases} \hat{z}_i = \left(\sum_{l=1}^{k+1} w_l \mathcal{I}[\theta_{l-1} \leq \hat{y}_i < \theta_l] \right) + b, \forall\ i = 1, \ldots, n\,, \\ \hat{z}_i - \hat{z}_{\bar{j}(i)} \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall\ i = 1, \ldots, n, \\ \hat{z}_i \geq y_i - \xi_i, & \forall\ i = 1, \ldots, n, \\ -\delta_i \hat{z}_i \geq -\delta_i y_i - \xi_i^*, & \forall\ i = 1, \ldots, n, \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n, \\ \xi_i \geq 0, & \forall\ i = 1, \ldots, n, \\ \xi_i^* \geq 0, & \forall\ i = 1, \ldots, n, \\ w_l - w_{l-1} \geq 0, & \forall\ l = 2, \ldots, k\,, \end{cases} \tag{5}$$

where $\hat{y}_i$ denotes the value of the prognostic index for observation i and χ_l is defined as before. Again, $\hat{z}$ is eliminated before solving the model in $w, b, \epsilon, \xi, \xi^*$. Note that for models that are modeling the risk (e.g. the PH model), the last constraint in equation (5) becomes: $w_l - w_{l-1} \leq 0,\ \forall\ l = 2, \ldots, k$. The risk group indicator $\hat{z}_\star$ for a new point $x_\star$ with prognostic index $\hat{y}_\star$ is then defined as $\hat{z}_\star = \sum_{l=1}^{k+1} \tilde{w}_l \mathcal{I}[\theta_{l-1} \leq \hat{y}_\star < \theta_l] + b$, with $\tilde{w}$ defined as before. The risk groups are then defined by the unique $\hat{z}$ values, where the highest value corresponds to the first risk group (lowest risk/highest predicted survival). Figure 2 illustrates that this approach can be interpreted as clustering on the level of $\hat{z}_i$. Observations with a score $\hat{y}_i \leq -1$ all receive the same risk group indicator $\hat{z}_i = 0$ and form a cluster or risk group with the highest risk. Observations with a score $-1 < \hat{y}_i \leq -0.5$ all receive a value of $\hat{z}_i = 5$ and form another risk group.

5 Estimation of survival curves

As discussed before, support vector machines for survival analysis assume a monotonic relation between the score and the outcome of interest (here the survival function S, or cumulative distribution function $F = 1 - S$). To obtain an estimate of the survival function, a separate function needs to be estimated for each possible risk group. Since survival functions are non-increasing functions in time, the estimated functions should be monotonic in time and in risk groups (see Figure 3). Two different approaches will be provided in this Section. First, the inverse-probability-of-censoring weighted average estimator of the cumulative distribution ($\hat{F}_{RR}$) as proposed by Robins and Rotnitzky [26] will be adapted to include the monotonicity constraints. This method has several advantages but the estimated survival functions are step-functions. A second approach solves the problem in the dual space and smooth curves are obtained. Although this second approach is appealing, direct application of this method is time-consuming since it requires the estimation of $\mathcal{O}(nn_t)$ unknowns, with n_t the number of time-points at which the survival curve needs to be estimated. Using the first method first and smoothing the results by means of the second, only $\mathcal{O}(n_s n_t)$ unknowns need to be estimated, with n_s the number of different risk groups and $n_s \ll n$.

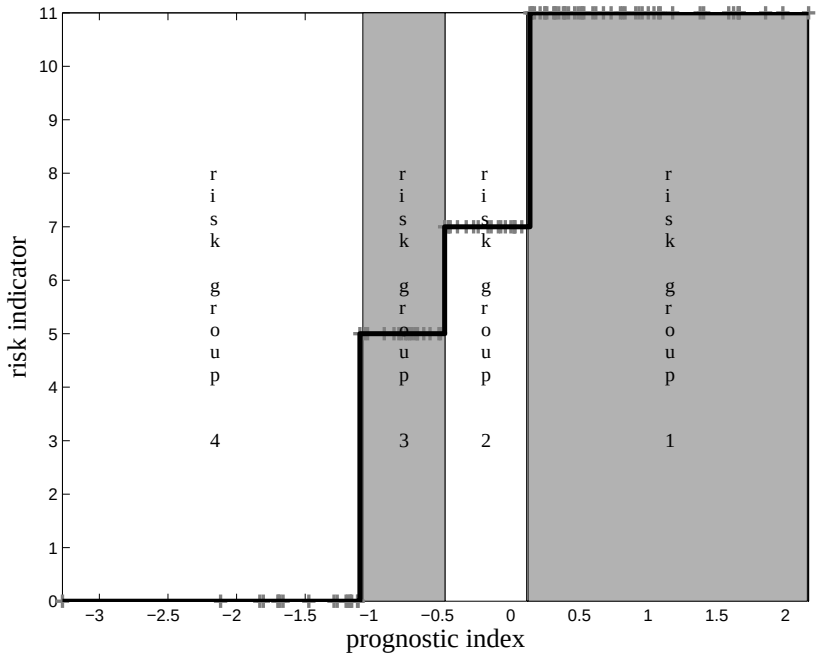


Fig. 2. Illustration of mICSc with a prognostic index (modeling survival) as a single input. The prognostic index or score $\hat{y}_i$ is mapped on a risk group indicator $\hat{z}_i = \sum_{l=1}^{k+1} \tilde{w}_l \mathcal{I}[\theta_{l-1} \leq \hat{y}_i < \theta_l] + b$, with a restricted set of possible values (here $\hat{z}_i \in \{\varsigma_1, \ldots, \varsigma_4\} = \{11, 7, 5, 0\}$). The pluses indicate the observed pairs $(\hat{y}_i, \hat{z}_i)$. The coefficients vector w is restricted to be positive such that $\hat{z}(\hat{y})$ is a monotonically increasing function of $\hat{y}$. The sparsity of the mICSc method makes it possible to interpret the results as risk groups.

5.1 Estimation of the survival curve by means of step functions

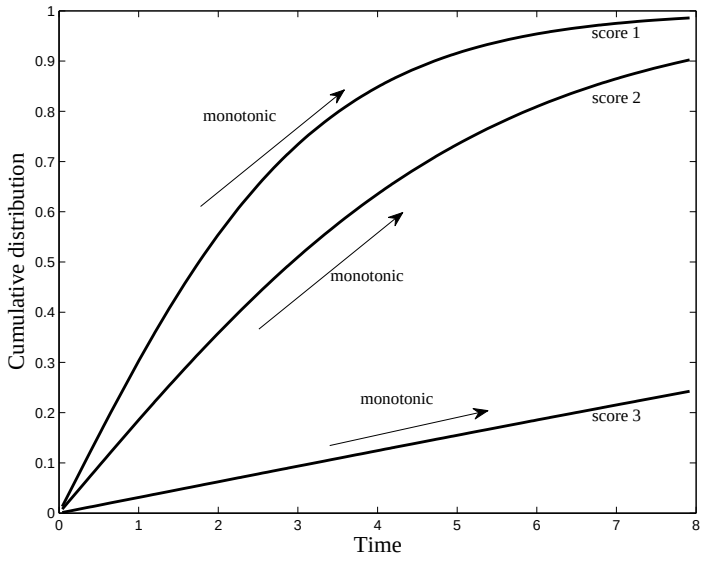
This Section proposes an approach to estimate survival curves for each of the risk groups obtained by using any type of score system that assumes a monotonic relation between scores and survival. The approach of Robins and Rotnitzky [26] to calculate the Kaplan-Meier estimate of the survival function is first described. An alternative implementation is proposed and adapted to simultaneously estimate n_s survival curves that are monotonic w.r.t. the risk groups. Section 5.2. proposes a method to smoothen the results from this Section.

Estimation of a single survival curve The inverse-probability-of-censoring weighted average estimator of the cumulative distribution ($\hat{F}_{RR}$) [26] is defined as

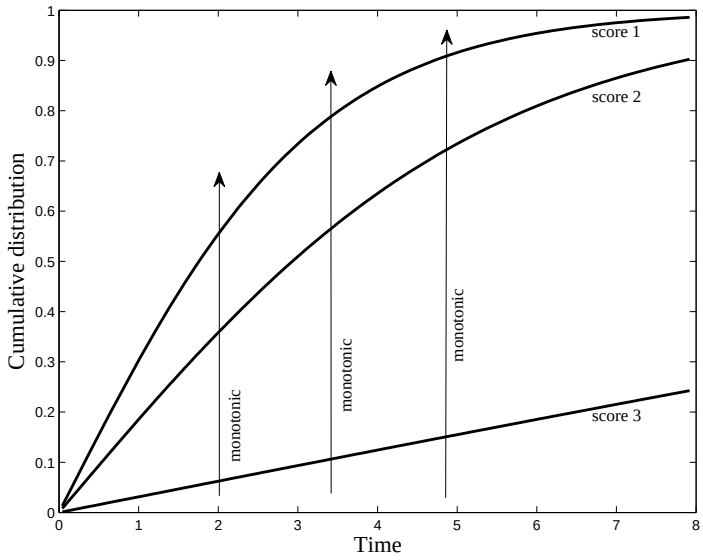
$$\hat{F}_{RR}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathcal{I}[y_i \leq t]\delta_i}{\hat{C}(y_i^-)} \,. \tag{6}$$

$\hat{F}_{RR}$ is a monotonically increasing step function, changing value at discrete time points τ_j, $j = 1, \ldots, n_t$. $\hat{C}$ is the Kaplan-Meier estimate of the censoring distribution, and $\hat{C}(y_i^-)$ is the function value at $y_i^- = \max_{\tau_j \leq y_i} \tau_j$. In [34] it is proven that $\hat{F}_{RR} = 1 - \hat{S}_{KM}$, with $\hat{S}_{KM}$ the Kaplan-Meier estimator when the unique event times are used as time points τ_j.

Proposition (Estimation of $\hat{F}_{RR}$ as an optimization problem)**.** *Given a dataset* $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$ *and* $\tau_j, j = 1, \ldots, n_t$ *the unique event times in* $\mathcal{D}$*, sorted in ascending order. Then,* $\hat{F}_{RR}(t)$



(a) monotonic w.r.t. time



(b) monotonic w.r.t. the score

Fig. 3. Illustration of the monotonicity constraints.

equals

$$\hat{F}(t) = \sum_{j=1}^{n_t} \hat{v}_j \mathcal{I}[\tau_j \leq t],\ j = 1, \ldots, n_t\,, \tag{7}$$

with $\hat{v} = [\hat{v}_1, \ldots, \hat{v}_{n_t}]^T$ *equal to*

$$\hat{v} = \arg\min_{v} \sum_{i=1}^{n} \sum_{j=1}^{n_t} \left(\frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}(y_i^-)} - \sum_{j'=1}^{j} v_{j'} \right)^2 . \tag{8}$$

Proof. $\hat{F}_{RR}(t)$ can only change value at $t = \tau_j,\ j = 1, \ldots, n_t$. Since $\hat{F}(t)$ is a step function with steps at $t = \tau_j, j = 1, \ldots, n_t$, it can only change value at $t = \tau_j, j = 1, \ldots, n_t$ (see Figure 4). The proposition is therefore proven if we can proof that $\hat{F}_{RR}(\tau_j)$ equals $\hat{F}(\tau_j)$, for all $j = 1, \ldots, n_t$.
The estimate of $v_j, \forall\ j = 1, \ldots, n_t$ is found by taking the derivative of the cost function

$$\mathcal{J} = \sum_{i=1}^{n} \sum_{j=1}^{n_t} \left(\frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}(y_i^-)} - \sum_{j'=1}^{j} v_{j'} \right)^2$$

w.r.t. v_j. The optimal value of v_j is then found as the value for which this derivative equals zero:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial v_j} &= -2 \sum_{i=1}^{n} \sum_{j=1}^{n_t} \left(\frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}(y_i^-)} - \sum_{j'=1}^{j} v_{j'} \right) = 0 \\ &\Rightarrow \sum_{j'=1}^{j} \hat{v}_{j'} = \frac{1}{n} \sum_{i=1} \frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}(y_i^-)} \,. \end{aligned}$$

The value of $\hat{F}(t)$ at $t = \tau_j$ then equals

$$\begin{aligned} \hat{F}(\tau_j) &= \sum_{j'=1}^{n_t} \hat{v}_{j'} \mathcal{I}[\tau_{j'} \leq \tau_j] \\ &= \sum_{j'=1}^{j} \hat{v}_{j'} \\ &= \frac{1}{n} \sum_{i=1} \frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}(y_i^-)} \,, \end{aligned}$$

which equals $\hat{F}_{RR}(\tau_j),\ \forall\ j = 1, \ldots, n_t$.

Note that the proposition above is not needed to estimate $\hat{F}_{RR}(t)$ since it can be estimated using equation (6). However, when two or more related cumulative distribution functions need to be estimated, equation (6) can no longer be used. Additionally remark that v_j will always be positive since $\hat{F}_{RR}(t)$ is a cumulative distribution function.

Estimation of n_s different survival curves In case the observations in the dataset $\mathcal{D}$ are grouped into n_s groups, n_s different cumulative distribution functions $\hat{F}^s$, with $s = 1, \ldots, n_s$ need to be estimated. Let $\varsigma_s,\ s = 1, \ldots, n_t$ denote the unique group indicators. All cumulative distribution functions $\hat{F}^s, s = 1, \ldots, n_s$ can then be estimated independently using equations (7-8) as

$$\hat{F}^s(t) = \sum_{j=1}^{n_t} \hat{v}_j^s \mathcal{I}[\tau_j \leq t],\ j = 1, \ldots, n_t \,,$$

with $\hat{v}$ the solution of

$$\hat{v}^s = \arg\min_{v^s} \sum_{i=1}^{n} \sum_{j=1}^{n_t} \mathcal{I}[\hat{z}_i = \varsigma_s] \left(\frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}^s(y_i^-)} - \sum_{j'=1}^{j} v_{j'}^s \right)^2 .$$

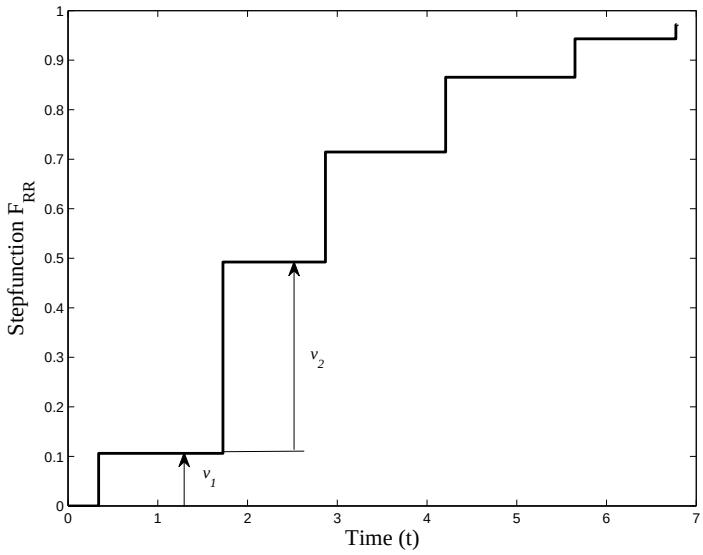


Fig. 4. Representation of a monotonic step function with n_t steps by means of positive constants $v_j,\ j = 1, \ldots, n_t$.

To ensure that the estimated cumulative distribution functions fulfill the assumption of monotonicity w.r.t. the group indicator, the optimization problems for all functions need to be solved simultaneously with addition of the following constraint

$$v_j^s - v_j^{s-1} \geq 0,\ \forall\ j = 1, \ldots, n_t; \forall\ s = 2, \ldots, n_s.$$

The steps $v_j^s,\ s = 1, \ldots, n_s,\ j = 1, \ldots, n_t$ are further restricted to be positive, to ensure that the solutions are valid cumulative distribution functions. All functions $F^s,\ s = 1, \ldots, n_s$ can then be estimated as

$$\hat{F}^s(t) = \sum_{j=1}^{n_t} \hat{v}_j^s \mathcal{I}[\tau_j \leq t],\ s = 1, \ldots, n_s\,, \tag{9}$$

where $\hat{v}$ is the solution of

$$\begin{aligned} \hat{v} = \underset{v}{\arg\min} & \sum_{i=1}^{n} \sum_{j=1}^{n_t} \sum_{s=1}^{n_s} \mathcal{I}[\hat{z}_i = \varsigma_s] \left(\frac{\mathcal{I}[y_i \leq \tau_j]\delta_i}{\hat{C}^s(y_i^-)} - \sum_{j'=1}^{j} v_{j'}^s \right)^2, \\ \text{subject to } & \begin{cases} v_j^s \geq 0, & \forall\ j = 1, \ldots, n_t; \forall\ s = 1, \ldots, n_s, \\ v_j^s - v_j^{s-1} \geq 0 & \forall\ j = 1, \ldots, n_t; \forall\ s = 2, \ldots, n_s\,. \end{cases} \end{aligned} \tag{10}$$

5.2 Smooth estimation of the survival function

The estimated cumulative distribution functions are step functions and a smoothing algorithm is needed to obtain smooth survival curves. Standard smoothers can not be used, since it can not be guaranteed that the smoothed versions of $\hat{F}^s(t)$ will remain monotonically increasing with the risk group. It is therefore necessary to define a smoothing algorithm that incorporates this monotonicity constraint.

The approach that we follow starts from a least-squares SVM (LS-SVM) regressor [3, **?**], with the values of $\hat{F}^s(\tau_j),\ \forall\ j = 1, \ldots, n_t;\ \forall\ s = 1, \ldots, n_s$, as outcomes and the times $\tau =$

$\tau_1, \cdots, \tau_{n_t}$ and unique risk group indicators $\varsigma_s,\ s = 1, \ldots, n_s$ as inputs. Let $\tilde{x}_l, l = 1, \ldots, n_s n_t$ be defined as $\tilde{x}_l = [\varsigma_s\ \tau_j]^T$, with $s = \lfloor (l-1)/n_t + 1 \rfloor$ and $j = l - n_t(s-1)$, where $\lfloor a \rfloor$ denotes the largest integer not larger than a. The standard LS-SVM formulation can then be used as follows

$$\begin{aligned} \min_{\tilde{v},b,\varepsilon} \quad & \frac{1}{2}\tilde{v}^T\tilde{v} + \frac{1}{2}\gamma \sum_{l=1}^{n_s n_t} \varepsilon_l^2 \\ \text{subject to } & \tilde{v}^T\varphi(\tilde{x}_l) + b = \Psi_l - \varepsilon_l,\ \forall\ l = 1, \ldots, n_s n_t\,, \end{aligned} \tag{11}$$

with $\varphi(\cdot)$ a feature map and $\Psi_l = \hat{F}^s(\tau_j)$, with s and j defined as before. The dual problem then becomes a set of linear equations:

$$\begin{bmatrix} \Omega + \frac{I}{\gamma} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} \Psi \\ 0 \end{bmatrix}, \tag{12}$$

where Ω is a matrix with elements $\Omega_{l,r} = k(\tilde{x}_l, \tilde{x}_r) = \varphi(\tilde{x}_l)^T\varphi(\tilde{x}_r)$, with $k(\cdot,\cdot)$ a kernel function. The monotonicity constraints are added to the dual problem formulation such that the desired result is obtained using the following parametric model:

$$\begin{aligned} \min_{\beta,\theta} \quad & \left|\left| \begin{bmatrix} \Omega + \frac{I}{\gamma} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta \\ b^* \end{bmatrix} - \begin{bmatrix} \Psi \\ 0 \end{bmatrix} \right|\right|_2 \\ \text{subject to } & \begin{cases} M(\Omega\beta + b^*) \geq 0 \\ \tilde{M}(\Omega\beta + b^*) \geq 0\,, \end{cases} \end{aligned} \tag{13}$$

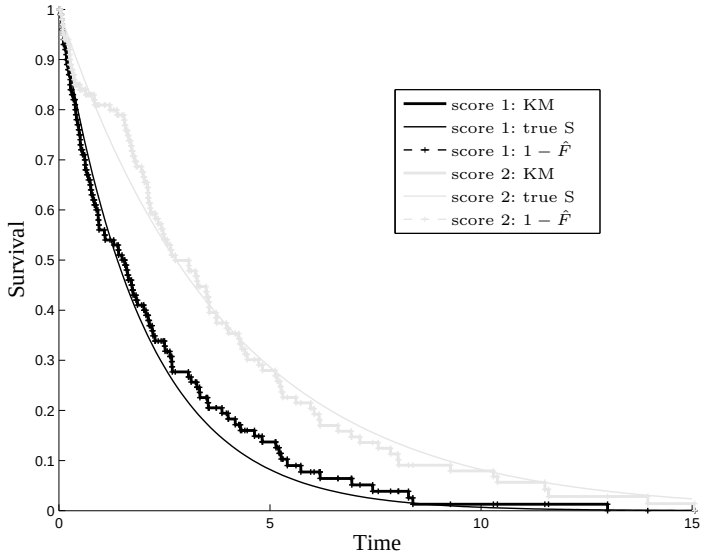
with $M \in \mathbb{R}^{(n_s-1)n_t \times n_s n_t}$ is a matrix with diagonal elements equal to 1 and elements on the n_t^{th} diagonal equal to -1 and all other elements equal zero; and $\tilde{M} \in \mathbb{R}^{n_s(n_t-1)\times n_s n_t}$ a matrix with diagonal elements equal to -1 and the elements on the first diagonal equal to 1. The first constraint enforces the monotonicity w.r.t. the risk groups and the second w.r.t. time. An estimate of the cumulative distribution function for a score ς_* at time τ_* can then be calculated as $\hat{G}([\varsigma_*\ \tau_*]^T) = \sum_{l=1}^{n_s n_t} \beta_l k(\tilde{x}_l, [\varsigma_*\ \tau_*]^T) + b^*$. An estimate of the survival curve for risk group s is then given by $\hat{S}^s(t) = 1 - \hat{G}^s(t)$.

6 Results

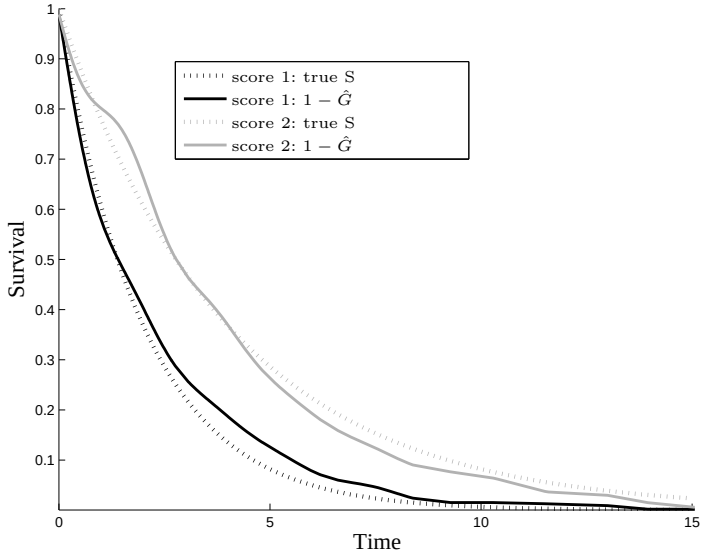
This Section starts with an illustration of the method to estimate survival curves for different risk groups on artificial data. It is shown that our first approach results in the Kaplan-Meier estimator for the different groups when the monotonicity constraints are valid. In case this assumption is violated, the method will result in coinciding survival curves. A real life application of the interval coded score system for survival analysis on the prognosis of breast cancer patients follows. The quality of the model is assessed in terms of discrimination [35] and calibration [36, **?**]. The proposed approach is compared with the proportional hazard model [10]. All parameters are tuned by means of 10-fold cross validation. The model selection criterion to obtain a score system is the c-index [35]. The model selection criterion to estimate survival curves is the Hosmer-Lemeshow χ^2 [36] at 2 and 5 years using 10 groups. The RBF kernel was used to obtain smooth estimates of the survival curves.

6.1 Artificial data

Consider a dataset with two groups of 100 observations. The true survival times of both groups are Weibull distributed ($f(t) = b_1^{-b_2} b_2 t^{b_2-1} \exp(-(t/b_1)^{b_2})$) with parameters $(2, 1)$ and $(4, 1)$ for both groups respectively. The censoring times have an exponential distribution ($f(t) = b_1 \exp(-b_1 t)$) with parameter $b_1 = 50$. Figure 5 illustrates the results. The results of model (9-10) coincide with the Kaplan-Meier estimates since these are already monotonic as a function of the scores.



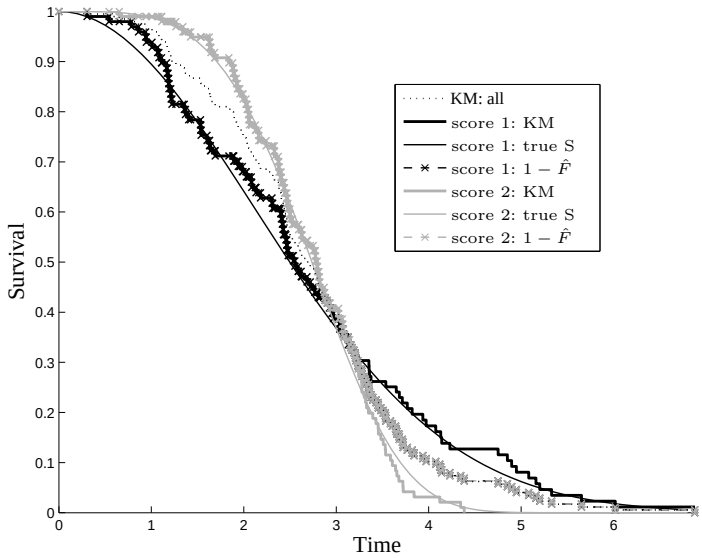
(a)



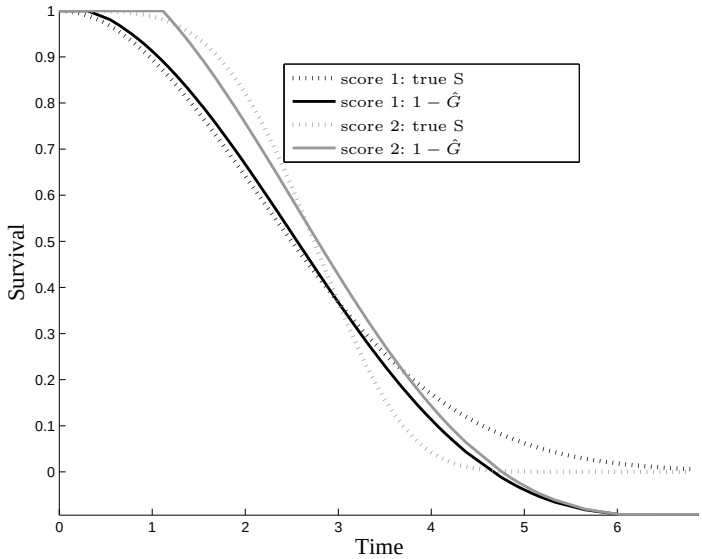
(b)

Fig. 5. Artificial example 1. (a) The true survival functions are monotonic as a function of the groups. The estimated survival curves $(1-\hat{F})$ coincide with the Kaplan-Meier estimators. (b) The smoothed versions $(1-\hat{G})$ align closely with the true survival curve.

In a second example, consider two groups (each containing 100 patients) with Weibull distributed survival times, with parameters $(3, 2)$ and $(3, 4)$, respectively. The true survival curves are thus non-monotonic in function of the scores. The censoring times have an exponential distribution with parameter $b_1 = 50$. Figure 6 illustrates the results. The estimates after model (9-10) coincide with the Kaplan-Meier estimates when the monotonicity constraints are valid. Violation of the constraint leads to equal estimated survival curves for both groups.



(a)



(b)

Fig. 6. Artificial example 2. (a) The true survival functions are non-monotonic as a function of the scores. The estimated survival curves $(1 - \hat{F})$ coincide with the Kaplan-Meier estimators as long as the monotonicity constraint holds. Afterwards, the estimated survival curves for both groups coincide and equal the Kaplan-Meier estimate of the whole dataset. (b) Smooth survival function estimates $1 - \hat{G}$.

6.2 Prognosis of breast cancer patients

The complete methodology (see algorithm 1) is illustrated on the prognosis of breast cancer patients. The training set consists of 1923 patients with complete information (age, tumor size, number of positive lymph nodes, expression of the progesterone (PR) and human epidermal growth factor receptor 2 (HER2) and tumor grade) which were diagnosed with primary operable breast cancer at the University Hospitals Leuven between January 2000 and June 2005. An external test set on 1192 patients containing complete information treated in New Zealand (Auckland Breast Cancer Registry) between January 2000 and December 2005 is available to test the resulting model. The obtained score model is summarized in Table 1 (see [20] for a figure-based representation).

Table 1. ICSc score system to obtain a prognostic index (step 1 in Figure 1) for the prognosis of primary operable breast cancer patients. If the answer on the question is yes, the points at the right of the question, need to be added to the score.

variable question	# points
Number of positive lymph nodes	
number of positive lymph nodes = 1	-1
2 $\leq$number of positive lymph nodes $\leq$ 3	-2
number of positive lymph nodes = 4	-3
5 $\leq$number of positive lymph nodes $\leq$ 6	-4
number of positive lymph nodes = 7	-10
number of positive lymph nodes >= 8	-17
Progesterone receptor	
positive PR	2
Human epidermal growth factor receptor	
positive HER2	-2
Tumor grade	
tumor grade = 2	-4
tumor grade = 3	-11

In order to divide the observations in a smaller number of risk groups, the prognostic index obtained from Table 1 is used as a single input in the mICSc model. The mICSc methodology will then automatically find the number of groups and the thresholds on the score. The c-index is again used as model selection criterion. Six different risk groups are identified in this way (see Figure 7). However, the group with the highest risk (risk group 6) contains only four patients and, since a survival curve can not be estimated accurately based an a small number of patients, is combined with risk group 5. Figure 8 (a) shows the estimated survival curves together with the Kaplan-Meier estimates for all five groups. The estimated survival curves and the Kaplan-Meier curves are very similar since the Kaplan-Meier estimates are already monotonically increasing with the scores. Table 2 summarizes the results.

The ICSc model is compared with the proportional hazard model [10], using the variables selected by the ICSc model. The mICSc method is applied with the prognostic index of the PH model as input to define risk groups. Five different risk groups are obtained, but the highest risk group contains only seven observations and is combined with risk group 4. The mean estimated survival curve for each group is given in Figure 8(b)). Table 3 compares both models. No significant differences are found between the discrimination abilities of the methods.

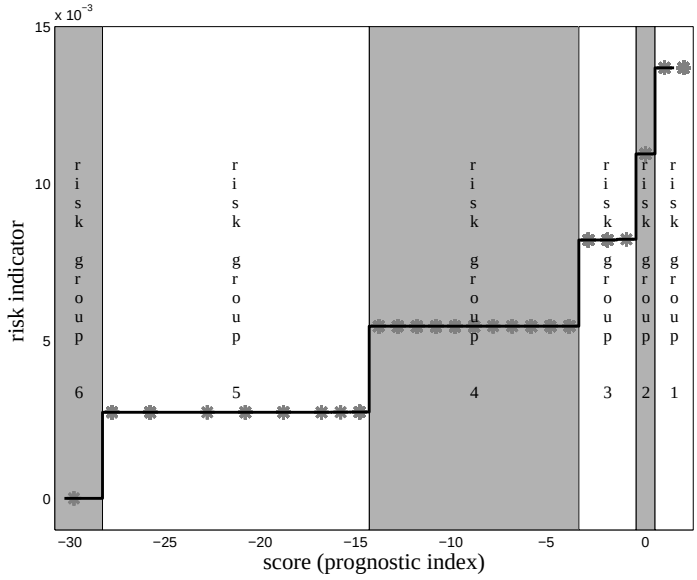


Fig. 7. The prognostic index $\hat{y}_i$ of the breast cancer patients is mapped onto a set of 6 unique values $\hat{z}_i$ by means of mICSc (step 2 in Figure 1). The observations are represented by means of the stars. Patients with the same values for their risk indicator (stars on the same step) are considered to belong to the same risk group.

Table 2. Risk groups obtained by means of the mICSc model (step 2 in Figure 1). Six risk groups were obtained, but due to a low number of observations in the highest risk group (score $\hat{y}_i$ lower than -29), this group was merged with the risk group containing observations with scores ranging from -29 to -15. The predicted survival $(1 - \hat{G})$ at 2 and 5 years of follow-up are reported for each risk group.

risk group	score	$\hat{S}$(2 year)	$\hat{S}$(5 year)
5	$\leq$ -15	0.75	0.62
4	-14 to -4	0.95	0.86
3	-3 to -1	0.98	0.94
2	0	0.99	0.95
1	≥ 1	≥ 0.99	0.98

Figure 9 illustrates the calibration results on the test set. The ICSc model is well calibrated. The PH model overestimates the survival in the risk group with the highest risk. This was also noted on the training set. This is due to the proportional hazard assumption that restricts the differences between the predicted survival curves. Since ICSc only assumes non-crossing survival curves, this method has more flexibility in the estimation of the different survival curves.

7 Conclusions

This work started with the introduction of a survival model that automatically leads to an easily applicable score system. In contrast to existing score models, the number and position of the thresholds are determined automatically by means of an incorporated control mechanism, making the trade-off between performance and categorization. Secondly, this method was adapted to define a clustering method for survival data, such that the number of clusters/risk groups are automatically determined. Thirdly, the inverse-probability-of-censoring weighted average estimator of the cumulative distribution was adapted to allow for the simul-

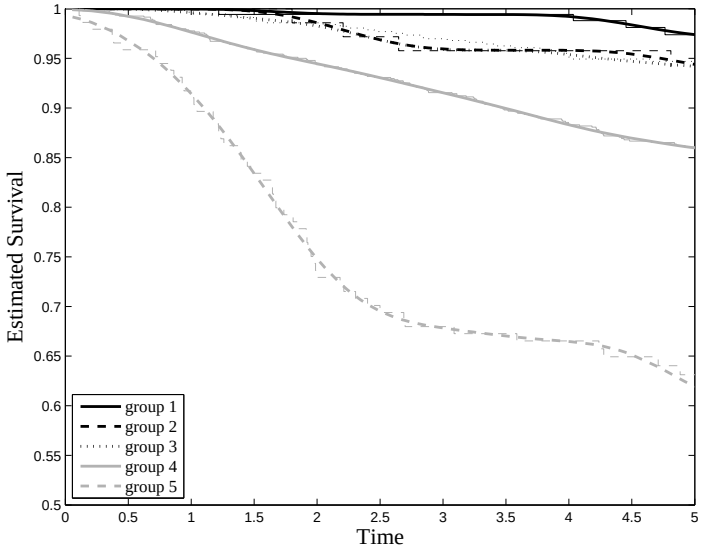
Table 3. Comparison of ICSc with the PH model

model	c-index	95% CI	c-index	95% CI
	training set		test set	
ICSc	0.711	0.676-0.741	0.724	0.687-0.758
PH	0.710	0.678-0.740	0.716	0.680-0.752
ICSc risk groups	0.687	0.659-0.715	0.702	0.669-0.731
PH risk groups	0.669	0.641-0.695	0.688	0.658-0.716

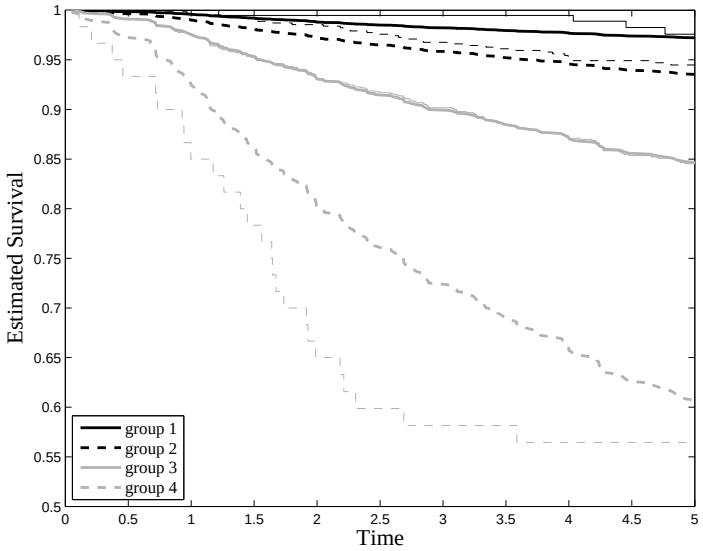
taneous estimation of different survival curves that are monotonic w.r.t. the risk groups. The method was illustrated on artificial and real-life data. The results of the proposed method are comparable with the PH model w.r.t. discrimination (c-index), but calibration is better for the ICSc approach. Additional advantages of the ICSc methodology are the incorporated feature selection and the automatic generation of the thresholds such that the performance of the resulting score is not dependent on the model developer.

Acknowledgments

Research supported by Research Council KUL: GOA MaNet, PFV/10/002 (OPTEC), several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, G.0108.11 (Compressed Sensing), G.0869.12N (Tumor imaging) , IWT: TBM070706-IOTA3, PhD Grants; iMinds2012; Belgian Federal Science Policy Office: IUAP P7/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); EU: RECAP 209G within INTERREG IVB NWE programme, EU HIP Trial FP7-HEALTH/ 2007-2013 (n 260777), ERC AdG A-DATADRIVE-B. VVB is a postdoctoral fellow of the Research Foundation - Flanders (FWO). SVH is a full professor and JS is a professor at the KU Leuven, Belgium.

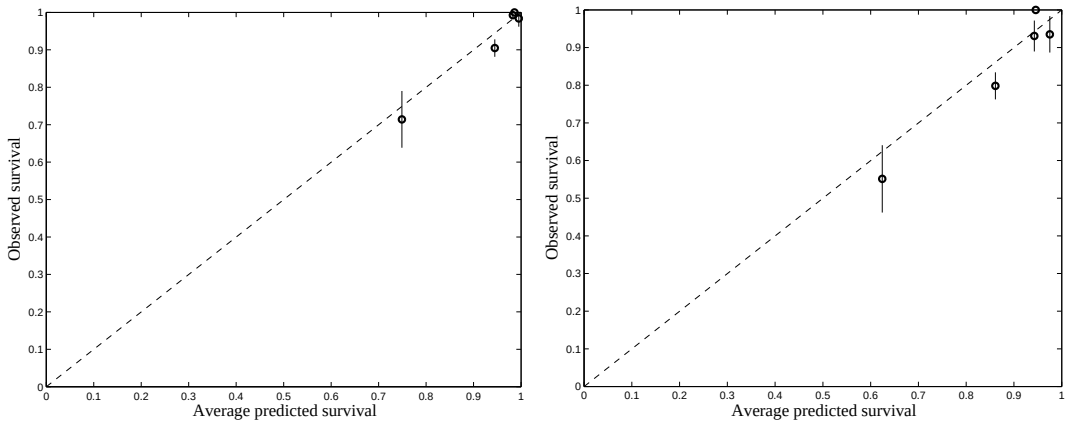


(a)



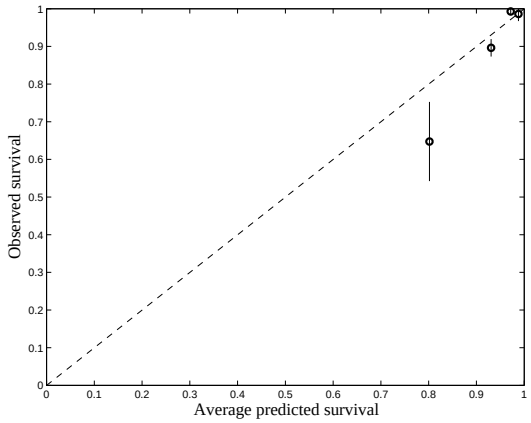
(b)

Fig. 8. Estimation of the survival curves for the different risk groups. (a) Estimated survival curves ($1 - \hat{G}$, steps 3-4 in Figure 1) for the ICSc risk groups; (b) mean estimated survival within the PH risk group. The stair functions indicate the Kaplan-Meier estimators.

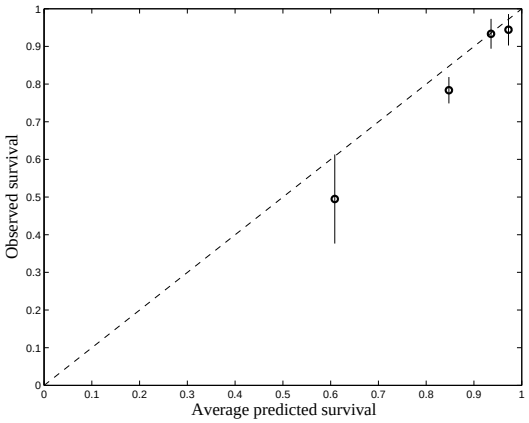


(a) ICSc, 2 year

(b) ICSc, 5 year



(c) PH model, 2 year



(d) PH model, 5 year

Fig. 9. Calibration plots for the risk groups on the test set after ICSc (a-b) and proportional hazard regression (c-d). The ICSc model is well calibrated. The PH model overestimates the survival in the risk group with the highest risk. This is the result of the proportional hazard assumption. Since ICSc only assumes non-crossing survival curves, this method has more flexibility in the estimation of the different survival curves.

References

1. V. Vapnik, Statistical Learning Theory., Wiley and Sons, New York, 1998.
2. C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
3. J. A. K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (3) (1999) 293–300.
4. J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines., World Scientific, Singapore, 2002.
5. J. Luts, F. Ojeda, R. Van De Plas, B. De Moor, S. Van Huffel, J. A. K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Analytica Chimica Acta 665 (2) (2010) 129–145.
6. E. Biganzoli, P. Boracchi, L. Mariani, E. Marubini, Feedforward neural networks for the analysis of censored survival data: a partial logistic regression approach, Statistics in Medicine 17 (10) (1998) 1169–1186.
7. P. Lisboa, H. Wong, P. Harris, R. Swindell, A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer, Artificial Intelligence in Medicine 28 (1) (2003) 1–25.
8. V. Van Belle, K. Pelckmans, J. A. K. Suykens, S. Van Huffel, Learning Transformation Models for Ranking and Survival Analysis, Journal of Machine Learning Research 12 (2011) 819–862.
9. V. Van Belle, K. Pelckmans, S. Van Huffel, J. A. K. Suykens, Support vector methods for survival analysis: a comparison between ranking and regression approaches, Artificial Intelligence in Medicine 53 (2) (2011) 107–118.
10. D. R. Cox, Regression models and life-tables (with discussion), Journal of the Royal Statistical Society, Series B 34 (2) (1972) 187–220.
11. D. W. Hosmer, S. Lemeshow, Applied Survival Analysis: Regression Modeling of Time to Event Data, Wiley-Interscience, New York, NY, USA, 1999.
12. D. W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd Edition, Wiley-Interscience, New York, NY, USA, 2000.
13. N. Holländer, M. Schumacher, On the problem of using 'optimal' cutpoints in the assessment of quantitative prognostic factors, Onkologie 24 (2) (2001) 194–199.
14. N. J. Perkins, E. F. Schisterman, The inconsistency of optimal cut-points using two roc based criteria., American Journal of Epidemiology 163 (7) (2006) 670–675.
15. L. M. Sullivan, J. M. Massaro, R. B. D'Agostino, Presentation of multivariate data for clinical use: The framingham study risk score functions, Statistics in Medicine 23 (10) (2004) 1631–1660.
16. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, New York: Chapman and Hall, 1984.
17. L. Tian, R. Tibshirani, Adaptive index models for marker-based risk stratification, Biostatistics 12 (1) (2011) 68–86.
18. N. H. Barakat, A. P. Bradley, Rule-extraction from support vector machines: A review, Neurocomputing 74 (2010) 178–190.
19. V. Van Belle, B. Van Calster, D. Timmerman, T. Bourne, C. Bottomley, L. Valentin, P. Neven, S. Van Huffel, J. A. K. Suykens, S. Boyd, A mathematical model for interpretable clinical decision support with applications in gynecology, PLoS One 7 (3) (2012) e34312.
20. V. Van Belle, S. Van Huffel, J. A. K. Suykens, S. Boyd, Interval coded scoring systems for survival analysis, in: M. Verleysen (Ed.), Proceedings of the European Symposium on Artificial Neural Networks, 2012, pp. 173–178.
21. S. C. Cheng, L. J. Wei, Z. Ying, Predicting Survival Probabilities with Semiparametric Transformation Models., Journal of the American Statistical Association 92 (437) (1997) 227–235.
22. D. M. Dabrowska, K. A. Doksum, Partial likelihood in transformation models with censored data, Scandinavian Journal of Statistics 15 (1) (1988) 1–23.
23. R. Koenker, O. Geling, Reappraising medfly longevity: A quantile regression survival analysis, Journal of the American Statistical Association 96 (2001) 458–468.
24. C. Alzate, J. A. K. Suykens, Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 335 –347.
25. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations., Journal of the American Statistical Association 53 (1958) 457–481.
26. J. M. Robins, A. Rotnitzky, Recovery of information and adjustment for dependent censoring using surrogate markers, AIDS Epidemiology - Methodological Issues (1992) 297–331.
27. M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 1.21 (Apr. 2011).

28. V. Van Belle, B. Van Calster, O. Brouckaert, I. Vanden Bempt, S. Pintens, V. Harvey, P. Murray, B. Naume, G. Wiedswang, R. Paridaens, P. Moerman, F. Amant, K. Leunen, A. Smeets, R. Drijkoningen, H. Wildiers, M. R. Christiaens, I. Vergote, S. Van Huffel, P. Neven, Qualitative assessment of the progesterone receptor and HER-2 improve the Nottingham Prognostic Index for short term breast cancer prognosis, Journal of Clinical Oncology 28 (27) (2010) 4129–4134.
29. T. Hastie, R. Tibshirani, Generalized additive models, Chapman and Hall, 1990.
30. C. de Boor, A Practical Guide to Splines, Springer, Berlin, 1978.
31. L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D: Nonlinear Phenomena 60 (1-4) (1992) 259–268.
32. E. J. Candès, M. B. Wakin, S. Boyd, Enhancing sparsity by reweighted l_1 minimization, Journal of Fourier Analysis and Applications 14 (5-6) (2008) 877–905.
33. K. Pelckmans, J. De Brabanter, J. A. K. Suykens, B. De Moor, Convex clustering shrinkage, in: Workshop on Statistics and Optimization of Clustering (PASCAL), 2005.
34. G. A. Satten, S. Datta, The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average, The American Statistician 55 (3) (2001) 207–210.
35. F. E. Harrell, Jr, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the yield of medical tests, JAMA: The Journal of the American Medical Association 247 (18) (1982) 2543–2546.
36. S. Lemeshow, D. W. Hosmer, A review of goodness of fit statistics for use in the development of logistic regression models, American Journal of Epidemiology 116 (1) (1982) 92–106.