Artificial Intelligence in Medicine 53 (2011) 107–118

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Support vector methods for survival analysis: a comparison between ranking and regression approaches

Vanya Van Belle [a,*], Kristiaan Pelckmans [b], Sabine Van Huffel [a], Johan A.K. Suykens [a]

[a] Department of Electrical Engineering (ESAT), Division SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b] Department of Information Technology, University of Uppsala, SE-751 05 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 23 December 2009
Received in revised form 11 May 2011
Accepted 18 June 2011

Keywords:
Support vector machines
Concordance index
Survival analysis
Cancer prognosis

ABSTRACT

Objective: To compare and evaluate ranking, regression and combined machine learning approaches for the analysis of survival data.
Methods: The literature describes two approaches based on support vector machines to deal with censored observations. In the first approach the key idea is to rephrase the task as a ranking problem via the concordance index, a problem which can be solved efficiently in a context of structural risk minimization and convex optimization techniques. In a second approach, one uses a regression approach, dealing with censoring by means of inequality constraints. The goal of this paper is then twofold: (i) introducing a new model combining the ranking and regression strategy, which retains the link with existing survival models such as the proportional hazards model via transformation models; and (ii) comparison of the three techniques on 6 clinical and 3 high-dimensional datasets and discussing the relevance of these techniques over classical approaches fur survival data.
Results: We compare SVM-based survival models based on ranking constraints, based on regression constraints and models based on both ranking and regression constraints. The performance of the models is compared by means of three different measures: (i) the concordance index, measuring the model's discriminating ability; (ii) the logrank test statistic, indicating whether patients with a prognostic index lower than the median prognostic index have a significant different survival than patients with a prognostic index higher than the median; and (iii) the hazard ratio after normalization to restrict the prognostic index between 0 and 1. Our results indicate a significantly better performance for models including regression constraints above models only based on ranking constraints.
Conclusions: This work gives empirical evidence that SVM-based models using regression constraints perform significantly better than SVM-based models based on ranking constraints. Our experiments show a comparable performance for methods including only regression or both regression and ranking constraints on clinical data. On high dimensional data, the former model performs better. However, this approach does not have a theoretical link with standard statistical models for survival data. This link can be made by means of transformation models when ranking constraints are included.

1. Introduction

Survival studies arise in different areas. Although they are most well known in medical and in particular in cancer studies, they also occur in economics (e.g. prediction of bankruptcy of factories), in mechanics (e.g. failure of airplanes, breakdown of engines, etc.), electronics (e.g. lifetime of electrical components), social sciences (e.g. estimating the time from marriage to divorce) and many other topics. Depending on the question at study one is interested in risk groups (which group of patients/components is more likely to experience the event?) or time predictions (before which time should the engine be replaced to decrease the risk of failure?).

The survival literature describes different models to answer these questions. Many common methods including the proportional hazard model (COX model) and log-odds model are transformation models (TM) [1–6]. This type of models assemble a prognostic index based on the covariates and link this index to the observed event times by means of a monotonic transformation function in a second step. TMS for survival analysis mainly focus on the first step. The standard COX model [7] for example avoids the second step by assuming that the hazard (the instantaneous risk to observe the event now, knowing that the event did not occur before) is proportional to an unspecified baseline hazard. Other models assume a fixed transformation function h. The accelerated failure time model is one example which assumes that the

* Corresponding author. Tel.: +32 16 32 10 65; fax: +32 16 32 19 86.
E-mail address: vanya.vanbelle@esat.kuleuven.be (V. Van Belle).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.06.006

Table 1
Overview of different kernel-based survival methods and their properties, in case a linear kernel is used.

Model	Reference	Ranking constraint	Regression constraint	# Tuning parameters
RANKSVMC	[9–11]	√		1
SVCR	[12]		√	1
SVRC	[13]		√	4
MODEL 1	[8]	√		1
MODEL 2		√	√	2

transformation function $h(y)$, with y the outcome under study, equals the logarithmic function, and the proportional odds model takes $h(y) = \text{logit}(y)$ [6].

Survival models based on support vector machines (SVM) [8] are able to incorporate non-linearities in an automatic way and using non-additive kernels, interactions are automatically incorporated. These methods use an approach which is different from the standard statistical approach. SVM-based models do not assume a true underlying function for which the parameters need to be estimated. Instead the empirical risk of misranking two instances with regard to their failure time, is minimized [9]. The survival problem was therefore reformulated as a ranking problem. To reduce the computational load, a simplified version comparing each observation only with its closest neighbor instead of with all other observations, was proposed in [10]. A more theoretical framework was provided in [8]. We will refer to the survival model proposed in the latter work as MODEL 1. In this work, we ask ourselves whether the inclusion of regression constraints can improve the performance. Therefore, the performance of MODEL 1 is compared with that of MODEL 2, including ranking and regression constraints. The proposed model is compared with survival methods only including ranking constraints (see [9–11]) and only including regression constraints (see [12,13]). Table 1 gives an overview of the different models handled in this work, their constraints, the number of tuning parameters in case of a linear kernel and how the ranking constraints are defined.

This paper is organized as follows. Section 2 gives an overview of transformation models in survival analysis. Section 3 starts with a summary of existing SVM-based survival methods, followed by the introduction of a new model, proposed by the authors. Section 4 compares the different SVM-based survival models on 8 different datasets. In addition to the methods mentioned before, the experiments include the performance of the cox model for comparison.

The following notations are used throughout the text. $\mathcal{D}$ denotes the set of observations $\{x_i, y_i, \delta_i\}_{i=1}^n$, where x_i is a d-dimensional covariate vector, y_i is the corresponding survival time and δ_i denotes whether an event was observed ($\delta_i = 1$) or the observation was right censored (δ_i=0). For notational convenience, it is assumed that the observations in $\mathcal{D}$ are sorted such that for two observations $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ with $j < i$, it applies that $y_j < y_i$.

2. Transformation models

A TM models a possibly unknown transformation of the outcome instead of the outcome itself as a function of the covariates. Initially, TMs were introduced in regression problems where the normality assumption on the distribution of the errors and the constant variance were not satisfied. A standard regression model for example tries to model the outcome y as a linear combination of the covariates:

$$y = w^T x + \epsilon, \tag{1}$$

where w is a coefficient vector and ϵ is the error variable. In cases where y is not normally distributed, the regression can be improved by transforming the dependent variable.

The general idea of transformation models can be summarized as follows [1]. Consider a model

$$h(y) = w^T x + \epsilon, \tag{2}$$

where w represents the regression parameters and ϵ is the error variable with density $f_\epsilon(\cdot)$. A TM specifies that some monotone function h of the response variable is linearly related to the regression variables (see Eq. (2)). In short, TMs are two-step models, estimating the ranking of observations in a first step by means of the utility $w^T x$ and linking this utility to the outcome by means of the transformation function h. Generalization towards non-linear functions $u(x) = w^T \varphi(x)$, with φ the feature map, of the covariates leads to

$$h(y) = u(x) + \epsilon. \tag{3}$$

TMs are also used when analyzing survival data. A large family of TMs in survival analysis are based on survival time distribution models, where one replaces the parameters of the model by a linear function of the covariates. The Weibull regression model and the accelerated failure time model are two examples of this type [14]. However, the most commonly used TM is the proportional hazard model (cox model) [7]. In this model the hazard at time t, defined as

$$\lambda(t|x) = \lim_{\Delta t \to 0} \frac{P(t \le y < t + \Delta t | x, y \ge t)}{\Delta t}, \tag{4}$$

is modelled as the product of an undefined baseline hazard $\lambda_0(t)$ and the exponential of a linear combination of the covariates:

$$\lambda(x, t) = \lambda_0(t) \exp(w^T x). \tag{5}$$

This leads to estimating the survival function $S(x, t) = P(y > t \mid x)$ as

$$S(x, t) = S_0(t)^{\exp(w^T x)}. \tag{6}$$

Taking $\ln(-\ln(\cdot))$ of both sides leads to

$$\ln(-\ln(S(x, t))) = \ln(-\ln(S_0(t))) - w^T x. \tag{7}$$

Written in the form of a TM this becomes

$$\epsilon = h(y) - u(x). \tag{8}$$

The difference between this model and the parametric transformation models discussed previously is that in the previous models the transformation function $h(\cdot)$ was fixed, which is not the case in the cox model.

Although standard TMs provide a *linear* relation between the explanatory variables and a function of the expected value of the dependent variable, these methodologies can be generalized to *non-linear* relations. An example is the extension of the cox model with smoothing splines [15,16].

3. Kernel-based survival models

This section starts with a brief discussion of existing survival models based on SVMs. In a second subsection, a new method is proposed. Since the outcome of this type of survival models can, in general, not be interpreted as a failure time, we will denote the outcome of the model as the prognostic index $u(x)$ instead of the prediction of the model. For the cox model this corresponds to $u(x) = w^T x$.

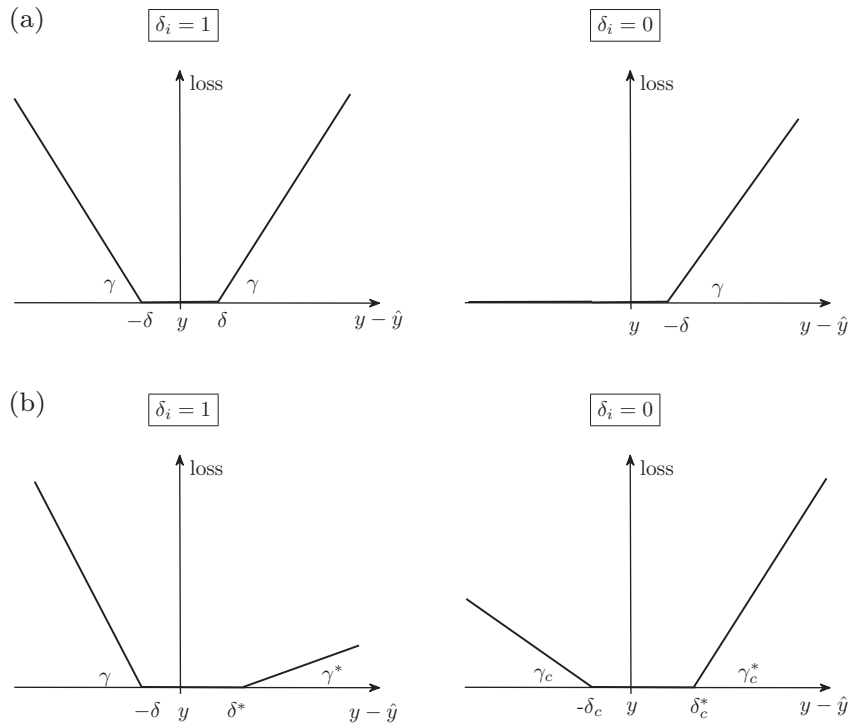


Fig. 1. Loss functions as defined by SVCR (a) [12] and SVRC (b) [13]. Both methods differ in the way the loss function is calculated. A major disadvantage of the SVCR method is the necessity to define 4 hyperparameters. In this paper we adopt the choice of setting the insensitive zone to width zero since this zone was of no practical relevance to the examples, and the formulations become much more convenient as such.

3.1. *Existing* SVM*-based survival models*

The excellent performance of support vector machines for classification and regression led to the question whether this type of models can be extended towards other statistical problems. When analyzing survival data, one is interested in the time between a certain starting point and the occurrence of a predefined event. A first approach one could think of is to use regression models to model the time of recurrence. However, survival data typically contain datapoints with incomplete information, called censored data. Although different types of censoring exist [6], we will restrict our attention to right censoring in this work. Right censored observations are observations for which a lower bond of the failure time is known instead of the failure time itself. Only incorporating observations for which the exact failure time is known would lead to underestimated survival times. Different proposals on how to model the survival problem by means of support vector machines are elaborated below. As the expert reader will notice, the width of the insensitive zones in the different SVR formulations was chosen to be zero (i.e. $\delta = \delta^*$ in Fig. 1). This choice reduces the complexity of the formulations substantially, while no consequent loss of performance was observed in the experiments.

3.1.1. *Support vector regression for censored data*

In standard support vector regression [17] the prediction is defined as a linear combination of a transformation of the variables $\varphi(x)$, with $\varphi(\cdot)$ the feature map, plus a constant $b : \hat{y} = w^T\varphi(x) + b$. To obtain this prediction, the coefficients w and the constant b need to be calculated. In order to obtain good generalizability properties, the coefficients are kept small. SVMs are formulated as optimizations problems, where a loss function should be minimized subject to certain constraints. The constraints include that the prediction $\hat{y}$ should be larger than $y - \epsilon$, with $\epsilon > 0$. Similarly $-\hat{y}$ should be larger than $-y - \epsilon^*$, with $\epsilon^* > 0$. The loss function penalizes large values of ϵ and ϵ^* such that the resulting predictions $\hat{y}$ will be close to the observed values y. Formally, the problem is formulated as:

$$\begin{aligned} &\min_{w,b,\epsilon,\epsilon^*} \quad \frac{1}{2}w^Tw + \gamma\sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*), \\ &\text{subject to} \quad \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall i = 1,\ldots,n \\ -w^T\varphi(x_i) - b \geq -y_i - \epsilon_i^*, & \forall i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall i = 1,\ldots,n. \end{cases} \end{aligned} \tag{9}$$

The parameter γ is a strict positive regularization constant and ϵ and ϵ^* are slack variables allowing for errors in the predictions of the training data. The predicted outcome for a new point $x^\star$ is then found as:

$$\hat{y}(x^\star) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varphi(x_i)^T\varphi(x^\star) + b, \tag{10}$$

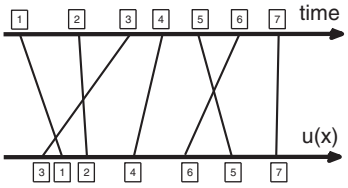
with α_i and α_i^* the Lagrange multipliers. An advantage of SVM-based models, is that the feature map $\varphi(\cdot)$ does not need to be defined explicitly. According to Mercer's theorem [18], $\varphi(x_i)^T\varphi(x_j)$ can be written as

$$k(x_i, x_j) = \varphi(x_i)^T\varphi(x_j), \tag{11}$$

provided that $k(\cdot,\cdot)$ is a positive definite kernel. Often used kernels are

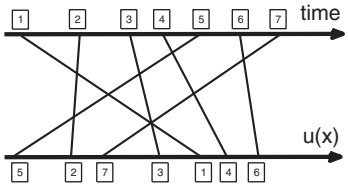
- the linear kernel: $k(x, z) = x^Tz$,
- the polynomial kernel of degree a: $k(x, z) = (\tau + x^Tz)^a$, with $\tau \geq 0$,
- the RBF kernel: $k(x, z) = \exp\left(-\frac{||x-z||_2^2}{\sigma^2}\right)$.

More recently, a kernel for clinical data [19] was proposed as an additive kernel $k(x, z) = \sum_{p=1}^{d} k_p(x^p, z^p)$, where $k_p(\cdot,\cdot)$ depends on



concordance	rank 1 2 3 4 5 6	ALL
2	1	1
3	0 0	0
4	1 1 1	3
5	1 1 1 1	4
6	1 1 1 1 0	4
7	1 1 1 1 1 1	6
		c-index
ALL	5 4 4 3 1 1	**18/21**

(a) first model



concordance	rank 1 2 3 4 5 6	ALL
2	0	0
3	0 1	1
4	1 1 1	3
5	0 0 0 0	0
6	1 1 1 1 1	1
7	0 1 0 0 1 0	2
		c-index
ALL	2 4 2 1 2 0	**11/21**

(b) second model

Fig. 2. Illustration of the search for a prognostic index which minimizes the empirical risk of misranking two observations. The upper part shows the ranking of 7 observations according to the survival time and according to two prognostic models $u(x)$. The lower part shows the calculation of the concordance index. Assuming that all survival times are known exactly, all pairs can be compared. A pair is concordant for a certain model whenever the ranking in time equals the ranking in model (no crosses in the upper part of the figure). (a) The first model results in a prognostic index which misranks 3 out of 21 possible pairs of datapoint. (b) The second model on the contrary misranks 10 pairs, resulting in a much lower concordance index.

the type of the p th variable. For continuous and ordinal variables, $k_p(\cdot,\cdot)$ is defined as

$$k_p(x^p, z^p) = \frac{c - |x^p - z^p|}{c}, \tag{12}$$

with x^p the pth covariate of observation x, $c = \max_p - \min_p$, with $\min_p$ and $\max_p$ the minimal and maximal value of the pth covariate in the given training dataset $\mathcal{D}$. For categorical and binary data $k_p(\cdot,\cdot)$ is defined as

$$k_p(x^p, z^p) = \begin{cases} 1 & \text{if } x^p = z^p \\ 0 & \text{if } x^p \neq z^p. \end{cases} \tag{13}$$

The problem of using support vector regression for censored data lies in the uncertainty about the outcomes y. The earliest regression approaches to censored data either omitted the censored observations, resulting in underestimated failure times, or treated the censored observations as non-events, resulting in biased models. In order to use all the available information Shivaswamy et al. [12] proposed a support vector regression approach to censored data (SVCR). For uncensored observations, the same constraints account as in the standard SVM regression model. For right censored observations, it is known that the failure did not occur until their censoring time. The first constraint in (9) is therefore still valid. However, the second constraint is too restrictive for right censored observations. The support vector regression model for censored data as proposed by Shivaswamy can therefore be formulated as:

SVCR:

$$\min_{w,b,\epsilon,\epsilon^*} \quad \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*),$$

$$\text{subject to} \quad \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \epsilon_i^*, & \forall i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall i = 1,\ldots,n. \end{cases} \tag{14}$$

The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_i (\alpha_i - \delta_i\alpha_i^*)\varphi(x_i)^T\varphi(x^\star) + b, \tag{15}$$

with α_i and α_i^* the Lagrange multipliers.

A second proposal to SVM regression for censored data was made by Khan and Zubek [13] (SVRC). The difference between both models lies in the applied penalty or loss for incorrect prognostic indices (see Fig. 1), which can be interpreted as predicted failure times for regression approaches. The SVCR method penalizes incorrect predictions the same whether the prediction was higher or lower than the observed failure time and penalizes incorrect predictions for right censored data only if the prediction is lower than the observed censoring time. In addition, the penalty for wrong predictions is the same, whether it consists of a prediction for censored or observed failure times. On the contrary, SVMC applies different penalties for the four possible cases: (i) penalty γ for events with predicted survival less than the observed survival; (ii) penalty γ^* for events with predicted survival larger than the observed survival, with $\gamma^* > \gamma$; (iii) penalty γ_c for right censored data with predicted survival less than the observed censoring time; and (iv) penalty γ_c^* for right censored data with predicted survival larger than the observed censoring time, with $\gamma_c^* < \gamma_c$. Additionally, this model provides different ϵ-losses for events and right censored data and for predictions higher and lower than the observed time. Therefore, the major drawback of the latter method is the large number of hyperparameters.

3.1.2. Support vector machines based on ranking constraints

When not using regression models, survival problems are often translated into classification problems answering the question whether the patient survives a certain predefined time (e.g. surviving 5 years after surgery). However, to be able to include as much events as possible, this predefined time should be taken very late. However, an early time allows to retain more patients since all patients censored before that time will be lost for the analysis. A

second problem in this approach is that the validity of the method decreases when more and more patients are censored earlier. To overcome the problems described above, it was proposed to formulate the survival problem as a ranking problem in [9,11] and a computational simplification was proposed in [10], much in the sense of the transformation models described in Section 2. The idea behind formulating the survival problem as a ranking problem is that in clinical applications one is often interested in defining risks groups. One is not primarily interested in a prediction of the survival time, but in whether the patient has a high or low risk for the event to occur, such that appropriate treatment can be given. To obtain this goal, an SVM ranking method is used, similar to the ranksvm model for ranking or preference learning [20]. The method proposed by [9,11] involves a regularization as usual and a penalization for each comparable pair of datapoints for which the order in prognostic index differs from the observed order. A data pair $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ is said to be comparable whenever the order of their event times is known (e.g. two events, or an event and a right censored data point for which the censoring time of the latter is later than the failure time of the former). More formally, a comparability indicator for a pair of observations $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ is defined as

$$\mathrm{comp}(i,j) = \begin{cases} 1 & \text{if } \delta_i = 1 \text{ and } \delta_j = 1 \\ & \delta_i = 1 \text{ and } \delta_j = 0 \text{ and } y_i \le y_j \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

The model is then formulated as

RANKSVMC:

$$\min_{w,\epsilon} \quad \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \sum_{\substack{j\,:\,y_i > y_j \\ \mathrm{comp}(i,j)=1}} \epsilon_{ij},$$
$$\text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_j)) \ge 1 - \epsilon_{ij}, & \forall i = 1,\ldots,n;\ \forall j : y_i > y_j \text{ and } \mathrm{comp}(i,j) = 1 \\ \epsilon_{ij} \ge 0, & \forall i = 1,\ldots,n;\ \forall j : y_i > y_j \text{ and } \mathrm{comp}(i,j) = 1. \end{cases} \tag{17}$$

Note that addition of a constant term b is not useful here since only differences in prognostic indices are considered. The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \sum_{\substack{j\,:\,y_i > y_j \\ \mathrm{comp}(i,j)=1}} \alpha_{ij}(\varphi(x_i) - \varphi(x_j))^T \varphi(x^\star), \tag{18}$$

with α_{ij} the Lagrange multipliers.

A drawback of the method is that a quadratical programming (QP) problem of $\mathcal{O}(n^2)$ constraints needs to be solved. To overcome this problem, a computationally simplified approach solving a QP of $\mathcal{O}(n)$ constraints was proposed in [10]. The reduction was found by comparing each data point i with the comparable neighbor with the largest survival time smaller than y_i, which we will indicate with $\bar{j}(i)$ assuming such exists (i.e. the first observation is assumed to be an event), instead of comparing with all comparable data points. We will refer to this simplified model as RANKSVMC.

In [8], a survival model was presented similar to the RANKSVMC model, the difference lying in the right hand side of the first set of constraints in Eq. (17), where 1 is replaced by $y_{(i)} - y_{\bar{j}(i)}$. Remark that in the case without censoring one has $\mathrm{comp}(i,j) = 1, \forall i = 1,\ldots, n; j \neq i$, such that $\bar{j}(i) := i - 1$. This model is formulated as

MODEL 1:

$$\min_{w,\epsilon} \quad \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i,$$
$$\text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \ge y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall i = 1,\ldots,n \\ \epsilon_i \ge 0, & \forall i = 1,\ldots,n. \end{cases} \tag{19}$$

The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \alpha_i (\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T \varphi(x^\star), \tag{20}$$

with α_i the Lagrange multipliers. We refer the interested reader to Appendix A for a deviation of the model.

3.2. A new SVM-based model

In this section we propose a new TM making use of SVM survival models. The core idea of survival SVM in [8] is that a prognostic index can be found by minimizing the empirical risk of misranking two observations (see Fig. 2).

More formally, the model becomes

$$\min_{w,\epsilon} \quad \sum_{i=1}^{n} \epsilon_i$$
$$\text{subject to} \quad \begin{cases} u(x_i) - u(x_{\bar{j}(i)}) + \epsilon_i \ge \rho_i, & \forall i = 1,\ldots,n \\ \epsilon_i \ge 0, & \forall i = 1,\ldots,n, \end{cases} \tag{21}$$

where ρ_i are given positive constants and $u = w^T\varphi(x)$.

Different approaches result from different choices for ρ_i and the regularization method. It is seen that the RANKSVMC model is a special case of the above formulation taking $\rho_i = 1$ and regularizing using the maximal margin principle [17]. The method that we propose here, goes further along the work of [8], using the Lipschitz constant instead of the maximal margin for regularization purposes. Doing so solves the question as how to define ρ, since this methodology takes $\rho_i = y_i - y_{\bar{j}(i)}$.

Since MODEL 1 is built by minimizing the empirical risk on misrankings, the value of the prognostic model only has a relative importance. When one is interested in the prognostic model for which the prognostic index can be interpreted as times, one might include regression constraints as in [12,13]. MODEL 1 is therefore modified as described below:

MODEL 2:

$$\min_{w,\epsilon,\xi,\xi^*,b} \quad \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*),$$
$$\text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \ge y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall i = 1,\ldots,n \\ w^T\varphi(x_i) + b \ge y_i - \xi_i, & \forall i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \ge -\delta_i y_i - \xi_i^*, & \forall i = 1,\ldots,n \\ \epsilon_i \ge 0, & \forall i = 1,\ldots,n \\ \xi_i \ge 0, & \forall i = 1,\ldots,n \\ \xi_i^* \ge 0, & \forall i = 1,\ldots,n. \end{cases} \tag{22}$$

The first set of constraints is the same as in MODEL 1 and is the ranking constraint optimizing the concordance index. The second two sets of constraints are the same as for the regression based model (see Eq. (14)).

Table 2
Comparison of different survival SVM models and the cox model on 6 cancer datasets with clinical variables, using a linear kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models were tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly. No significant differences are noted with the cox model. However, the SVM-based models have the advantage that one does not have to check the linearity and proportional hazard assumption.

Dataset	Method	c-Index	Logrank χ^2	Hazard rate	# sv
VLC	MODEL 1	0.61 ± 0.07 ***	1.38 ± 2.19 ***	3.97 ± 4.29 ***	84.5 ± 7.75 ***
	MODEL 2	$\mathbf{0.69 \pm 0.03}$	5.27 ± 5.35	10.50 ± 13.19	91.0 ± 0.70
	RANKSVMC	0.62 ± 0.08 ***	3.02 ± 3.19 ***	5.03 ± 6.53 ***	$\mathbf{78.0 \pm 6.44}$ ***
	SVCR	$\mathbf{0.69 \pm 0.04}$	$\mathbf{5.75 \pm 4.37}$	9.50 ± 9.93	89.0 ± 1.21 ***
	PH_{linear}	0.68 ± 0.03	5.45 ± 3.66	$\mathbf{10.61 \pm 6.47}$	
LCR	MODEL 1	0.55 ± 0.07 ***	1.08 ± 2.23 *	2.40 ± 5.36 ***	73.5 ± 8.50 ***
	MODEL 2	$\mathbf{0.60 \pm 0.05}$	1.87 ± 2.51	$\mathbf{5.06 \pm 9.65}$	86.0 ± 0.00
	RANKSVMC	0.51 ± 0.07 ***	0.51 ± 2.81 ***	1.43 ± 5.60 ***	$\mathbf{72.5 \pm 6.55}$ ***
	SVCR	0.59 ± 0.04	$\mathbf{2.02 \pm 2.73}$	4.46 ± 2.90	80.0 ± 3.25 ***
	PH_{linear}	$\mathbf{0.60 \pm 0.04}$	1.60 ± 2.06	4.05 ± 4.80	
LD	MODEL 1	0.62 ± 0.07 ***	2.19 ± 2.92 ***	4.45 ± 6.53 ***	84.5 ± 7.97 **
	MODEL 2	0.68 ± 0.04	7.53 ± 5.18	13.8 ± 15.21	86.0 ± 2.21
	RANKSVMC	0.62 ± 0.08 ***	2.55 ± 3.57 ***	4.68 ± 15.96 ***	$\mathbf{80.0 \pm 4.24}$ ***
	SVCR	$\mathbf{0.69 \pm 0.05}$	6.70 ± 3.95	14.21 ± 12.70	81.0 ± 5.10 ***
	PH_{linear}	$\mathbf{0.69 \pm 0.05}$ *	$\mathbf{9.73 \pm 8.31}$ **	$\mathbf{22.21 \pm 32.09}$ *	
MLC	MODEL 1	0.60 ± 0.05 **	2.27 ± 3.77	3.22 ± 3.43	99.5 ± 7.50 ***
	MODEL 2	$\mathbf{0.62 \pm 0.05}$	$\mathbf{3.32 \pm 3.41}$	4.12 ± 5.44	108.0 ± 3.00
	RANKSVMC	0.59 ± 0.05 *	3.06 ± 2.91	3.20 ± 6.79	99.0 ± 4.45 ***
	SVCR	0.61 ± 0.04	2.27 ± 2.86	3.48 ± 4.77	$\mathbf{95.0 \pm 6.51}$ ***
	PH_{linear}	0.61 ± 0.04 *	2.89 ± 2.68	$\mathbf{6.04 \pm 5.55}$	
PC	MODEL 1	0.75 ± 0.04	10.05 ± 5.94	148.86 ± 340.89	$\mathbf{133.0 \pm 17.50}$ ***
	MODEL 2	0.76 ± 0.03	11.86 ± 5.69	231.79 ± 296.90	205.5 ± 23.50
	RANKSVMC	0.75 ± 0.03	$\mathbf{12.56 \pm 5.37}$	166.10 ± 340.56	255.5 ± 13.54 ***
	SVCR	0.76 ± 0.03	10.15 ± 5.64	198.45 ± 479.37	137.5 ± 28.72 **
	PH_{linear}	$\mathbf{0.77 \pm 0.03}$	11.39 ± 5.68	$\mathbf{266.26 \pm 478.54}$	
BC	MODEL 1	0.64 ± 0.04 ***	12.03 ± 9.57 *	36.02 ± 1657.30 **	411.5 ± 26.70 ***
	MODEL 2	$\mathbf{0.67 \pm 0.03}$	16.45 ± 8.06	70.14 ± 2876.42	446.0 ± 28.40
	RANKSVMC	0.62 ± 0.10 ***	9.99 ± 9.24 ***	21.36 ± 306.01 ***	399.5 ± 59.32 ***
	SVCR	$\mathbf{0.67 \pm 0.10}$	$\mathbf{20.17 \pm 8.74}$	67.29 ± 861.97	$\mathbf{366.0 \pm 72.09}$ ***
	PH_{linear}	$\mathbf{0.67 \pm 0.03}$	18.11 ± 8.76	$\mathbf{102.78 \pm 3182.89}$	

Boldface numbers indicate the highest performances which were obtained.
* $p < 0.05$ (Wilcoxon rank sum test).
** $p < 0.01$ (Wilcoxon rank sum test).
*** $p < 0.001$ (Wilcoxon rank sum test).

The prognostic index for a new observation $x^\star$ is then found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} (\alpha_i (\varphi(x_i) - \varphi(x_{\tilde{j}(i)}))^T + (\beta_i - \delta_i \beta_i^*) \varphi(x_i)^T) \varphi(x^\star) + b, \quad (23)$$

with α_i, β_i and β_i^* the Lagrange multipliers. See Appendix B for more information. Remark that MODEL 2 reduces to MODEL 1 for $\mu = 0$ and to SVCR for $\gamma = 0$.

An important advantage of RANKSVMC, MODEL 1, and MODEL 2 is their link with the proportional hazards model and the log-odds model. All these models are transformation models, estimating a utility in a first step under the assumption that this utility is monotonically related (by means if the transformation function) with the prediction of interest. In a second step this monotonic relation can be estimated (sometimes a specific parametric form is assumed). The ranking constraints induce the monotonicity of the transformation function without using a restrictive parametric representation. Additionally, these constraints coincide with optimizing the c-index.

4. Experiments

This section compares the performances of the discussed methods on 5 clinical data sets and 3 high dimensional data sets. A description of the data and the different performance measures is given first. Next the results on real data and on artificial data are discussed.

4.1. Data sets

We will use 5 publicly available data sets for 6 clinical experiments. The first data set concerns the prediction of complete remission and death of 129 patients with leukemia [21]. Information on the treatment, sex, age, performance score (Karnofsky score), white blood cells, platelets, hemoglobin and whether the patient received a kidney transplant was given. In a first experiment (LD) the event at study is death, while in a second experiment (LCR) the event is complete remission. A second data set concerns the veteran's administration lung cancer trial (VLC) [6,22]. Only 9 out of the 137 men with advanced inoperable lung cancer were alive at the end of the study. Patients were randomized into a standard or test chemotherapy. Information on the performance score, age, therapy, histology of the tumor, treatment and time between the diagnosis and the randomization was available. The third data set concerns 506 patients with prostatic cancer (PC) [23]. The variables used for this experiment are: treatment, age, weight index, performance index, history of cardiovascular disease, size of the tumor, a combined index of stage of histologic grade and serum hemoglobin. The analysis was performed on the 483 patients with complete information. 125 (26%) patients died due to prostatic cancer during the study. All other patients were right censored at their date of last follow-up. The Mayo clinic lung cancer data (MLC) [24] contains information on 167 patients with advanced lung cancer. The data set contains information on age, sex, two performance scores estimated by the physician and one estimated by the patient, the number of calories consumed and the weight loss in the last six months. During the study period, 72% of the patients died. A last data set contains information on 720 breast cancer patients

Table 3
Comparison of different survival SVM models on 6 cancer datasets with clinical variables, using a clinical kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models were tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly. No significant differences are noted with the COX model. However, the SVM-based models have the advantage that one does not have to check the linearity and proportional hazard assumption.

Dataset	Method	c-Index	Logrank χ^2	Hazard rate	# sv
VLC	MODEL 1	$0.58 \pm 0.06^{***}$	$2.01 \pm 3.72^{***}$	$3.26 \pm 4.70^{***}$	91.0 ± 3.51
	MODEL 2	$\mathbf{0.69 \pm 0.03}$	6.27 ± 5.11	9.78 ± 7.46	91.0 ± 1.10
	RANKSVMC	$0.57 \pm 0.07^{***}$	$1.60 \pm 3.83^{***}$	$2.70 \pm 5.36^{***}$	$\mathbf{77.0 \pm 6.53}^{***}$
	SVCR	$\mathbf{0.69 \pm 0.04}$	$\mathbf{9.88 \pm 6.05}^{*}$	$\mathbf{11.12 \pm 9.75}$	$90.0 \pm 0.96^{***}$
	$PH_{pspline}$	$0.67 \pm 0.09^{**}$	4.52 ± 6.62	8.40 ± 19.14	
LCR	MODEL 1	$0.54 \pm 0.06^{***}$	1.04 ± 1.63	$1.98 \pm 2.85^{**}$	$81.0 \pm 1.00^{***}$
	MODEL 2	0.59 ± 0.04	1.08 ± 1.90	3.56 ± 3.50	86.0 ± 0.00
	RANKSVMC	$0.52 \pm 0.06^{***}$	0.37 ± 2.17	$1.38 \pm 2.88^{***}$	$\mathbf{78.0 \pm 7.22}^{***}$
	SVCR	$\mathbf{0.60 \pm 0.04}$	$\mathbf{1.94 \pm 1.72}^{*}$	$\mathbf{4.04 \pm 3.25}$	$80.5 \pm 3.67^{***}$
	$PH_{pspline}$	0.58 ± 0.05	1.39 ± 1.98	3.70 ± 5.83	
LD	MODEL 1	$0.66 \pm 0.07^{**}$	$3.60 \pm 4.35^{***}$	$5.97 \pm 12.37^{***}$	86.0 ± 2.47
	MODEL 2	$\mathbf{0.69 \pm 0.05}$	8.50 ± 7.48	17.23 ± 23.33	86.0 ± 1.79
	RANKSVMC	$0.65 \pm 0.08^{***}$	$2.68 \pm 6.60^{***}$	$5.12 \pm 8.95^{***}$	$\mathbf{76.0 \pm 5.04}^{***}$
	SVCR	$\mathbf{0.69 \pm 0.04}$	$\mathbf{10.7 \pm 6.51}$	$13.1 \pm 33.54^{***}$	$82.0 \pm 4.13^{**}$
	$PH_{pspline}$	0.67 ± 0.05	$4.89 \pm 5.45^{*}$	$\mathbf{18.35 \pm 62.26}$	
MLC	MODEL 1	0.61 ± 0.05	3.13 ± 3.74	3.48 ± 7.48	111.0 ± 0.00
	MODEL 2	0.61 ± 0.05	2.53 ± 2.76	3.28 ± 3.83	111.0 ± 0.00
	RANKSVMC	0.61 ± 0.05	2.55 ± 3.70	3.30 ± 7.42	$\mathbf{95.0 \pm 4.90}^{***}$
	SVCR	$\mathbf{0.62 \pm 0.04}$	$\mathbf{3.70 \pm 3.22}^{*}$	$\mathbf{4.58 \pm 5.30}^{*}$	$96.0 \pm 7.84^{***}$
	$PH_{pspline}$	$0.57 \pm 0.04^{***}$	2.04 ± 2.27	3.08 ± 3.46	
PC	MODEL 1	$0.76 \pm 0.04^{**}$	$11.22 \pm 5.71^{*}$	$108.39 \pm 107.79^{*}$	222.0 ± 21.00
	MODEL 2	$\mathbf{0.78 \pm 0.03}$	$\mathbf{14.56 \pm 6.47}$	148.08 ± 137.34	226.5 ± 41.50
	RANKSVMC	0.76 ± 0.03	11.89 ± 6.00	111.26 ± 116.90	$264.0 \pm 11.38^{**}$
	SVCR	$\mathbf{0.78 \pm 0.03}$	13.15 ± 6.60	150.66 ± 165.24	$\mathbf{160.0 \pm 62.93}^{***}$
	$PH_{pspline}$	0.77 ± 0.03	12.76 ± 5.89	$\mathbf{286.6 \pm 615.05}^{***}$	
BC	MODEL 1	$0.63 \pm 0.04^{***}$	$9.34 \pm 8.83^{***}$	$10.10 \pm 20.91^{***}$	$443.0 \pm 33.10^{***}$
	MODEL 2	$\mathbf{0.68 \pm 0.03}$	$\mathbf{22.62 \pm 9.09}$	24.23 ± 13.19	457.0 ± 16.50
	RANKSVMC	$0.62 \pm 0.10^{***}$	$8.93 \pm 8.83^{***}$	$11.20 \pm 15.23^{***}$	$\mathbf{404.0 \pm 58.26}^{***}$
	SVCR	$\mathbf{0.68 \pm 0.10}$	$19.51 \pm 7.89^{*}$	18.99 ± 12.02	$453.0 \pm 75.70^{**}$
	$PH_{pspline}$	0.67 ± 0.03	$17.36 \pm 8.24^{*}$	$\mathbf{98.93 \pm 4052.05}^{***}$	

Boldface numbers indicate the highest performances which were obtained.
* $p < 0.05$ (Wilcoxon rank sum test).
** $p < 0.01$ (Wilcoxon rank sum test).
*** $p < 0.001$ (Wilcoxon rank sum test).

(the German breast cancer study) (GBSG) [25,26]. Information was available on hormonal therapy, menopausal status, age, grade, tumor size, number of positive lymph nodes, progesterone and estrogene receptors. The experiment is based on the 686 cases with complete information. During the study period, 299 patients experienced a breast cancer related event. All these data sets can be found on the world-wide-web (http://lib.stat.cmu.edu/datasets (accessed 29 August 2008) and the R-package http://cran.r-project.org/web/packages/survival/index.html (accessed 20 November 2007)).

The methods will additionally be tested on three high dimensional data sets: the Dutch Breast Cancer Data set (DBCD), Norway/Stanford breast cancer data (NSBCD) and the diffuse large-B-cell lymphoma data (DLBCL). More information on these data sets can be found in [27].

4.2. Performance measures

In order to compare the different models, three performance measures are used in the experiments:

Table 4
Comparison of different survival SVM models on 3 high dimensional data sets, using a linear kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models was tested using the Wilcoxon rank sum test. The regression approach performs significantly better than both the other approaches.

Dataset	Method	c-Index	Logrank χ^2	Hazard rate	# sv
NSBCD	MODEL 1	0.66 ± 0.04	1.63 ± 1.12	13.93 ± 7.43	$\mathbf{25.0 \pm 4.00}^{***}$
	MODEL 2	0.65 ± 0.04	1.22 ± 0.92	11.19 ± 6.21	77.0 ± 0.00
	RANKSVMC	0.63 ± 0.06	1.34 ± 0.99	9.58 ± 6.24	$48.0 \pm 4.00^{***}$
	SVCR	$\mathbf{0.68 \pm 0.04}^{**}$	$\mathbf{2.44 \pm 1.64}^{*}$	$\mathbf{14.98 \pm 8.6}^{*}$	$45.0 \pm 2.00^{***}$
DBCD	MODEL 1	0.64 ± 0.04	3.26 ± 2.29	9.54 ± 5.71	$\mathbf{61.5 \pm 6.50}^{***}$
	MODEL 2	0.64 ± 0.04	3.14 ± 2.19	10.12 ± 6.12	197.0 ± 0.00
	RANKSVMC	0.61 ± 0.05	2.6 ± 2.06	6.1 ± 3.82	$110.0 \pm 7.50^{***}$
	SVCR	$\mathbf{0.71 \pm 0.03}^{***}$	$\mathbf{9.85 \pm 3.22}^{***}$	$\mathbf{18.72 \pm 7.8}^{***}$	$100.0 \pm 3.00^{***}$
DLBCL	MODEL 1	0.58 ± 0.04	2.43 ± 1.68	3.37 ± 1.57	$\mathbf{78.0 \pm 5.00}^{***}$
	MODEL 2	0.58 ± 0.04	2.41 ± 1.85	3.4 ± 1.61	160.0 ± 0.00
	RANKSVMC	0.59 ± 0.03	2.26 ± 1.72	4.35 ± 2.09	$117.0 \pm 4.00^{***}$
	SVCR	$\mathbf{0.62 \pm 0.03}^{***}$	$\mathbf{5.12 \pm 2.73}^{***}$	$\mathbf{8.88 \pm 4.64}^{***}$	$133.0 \pm 2.00^{***}$

Boldface numbers indicate the highest performances which were obtained.
* $p < 0.05$ (Wilcoxon rank sum test).
** $p < 0.01$ (Wilcoxon rank sum test).
*** $p < 0.001$ (Wilcoxon rank sum test).

Table 5
Comparison of different survival SVM models on 3 high dimensional data sets, using a clinical kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models were tested using the Wilcoxon rank sum test. The regression approach performs significantly better than both the other approaches.

Dataset	Method	c-Index	Logrank χ^2	Hazard rate	# sv
NSBCD	MODEL 1	0.65 ± 0.08 *	2.7 ± 3.71 **	11.85 ± 36.22 **	**18.0 ± 2.54*****
	MODEL 2	0.61 ± 0.10	1.15 ± 2.64	5.34 ± 236.25	77.0 ± 0.27
	RANKSVMC	0.63 ± 0.08	2.3 ± 3.38	8.27 ± 41.79 *	27.0 ± 3.82 ***
	SVCR	**0.69 ± 0.07*****	**2.78 ± 2.94***	**12.94 ± 41.26*****	45.0 ± 3.12 ***
DBCD	MODEL 1	0.61 ± 0.09 *	1.84 ± 4.11 *	5 ± 39.92 *	**43.5 ± 4.04*****
	MODEL 2	0.64 ± 0.04	3.49 ± 2.79	7.55 ± 7.4	197.0 ± 0.00
	RANKSVMC	0.59 ± 0.09 ***	2.32 ± 3.75	4.08 ± 20.05 *	65.5 ± 6.11 ***
	SVCR	**0.71 ± 0.04*****	**9.85 ± 4.92*****	**18.72 ± 30.54*****	100.0 ± 5.95 ***
DLBCL	MODEL 1	0.54 ± 0.04 ***	0.75 ± 1.4 ***	2.07 ± 1.3 ***	**87.0 ± 4.46*****
	MODEL 2	0.61 ± 0.04	4.32 ± 3.75	6.15 ± 8.66	160.0 ± 0.00
	RANKSVMC	0.55 ± 0.04	0.9 ± 1.9	2.22 ± 1.75	97.5 ± 6.42 ***
	SVCR	**0.62 ± 0.04*****	**5.12 ± 4.49*****	**9.02 ± 15.34*****	98.5 ± 5.77 ***

Boldface numbers indicate the highest performances which were obtained.
* $p < 0.05$ (Wilcoxon rank sum test).
** $p < 0.01$ (Wilcoxon rank sum test).
*** $p < 0.001$ (Wilcoxon rank sum test).

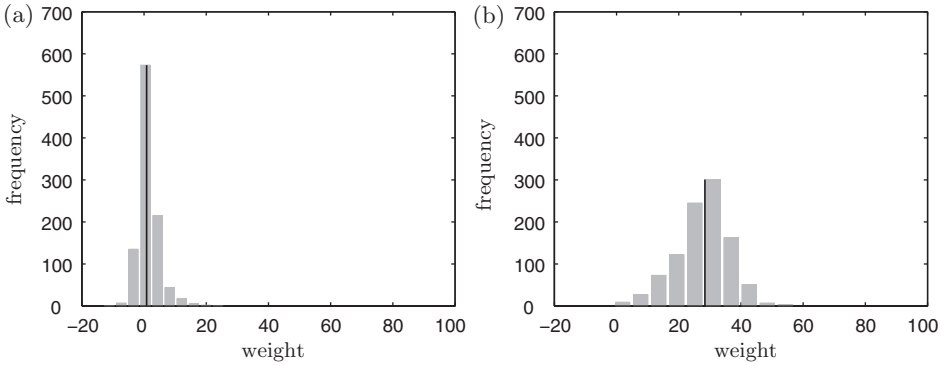


Fig. 3. Distribution of the estimated weights (using a linear kernel) of the Karnofsky score on 1000 bootstraps of one particular training set of the VLC dataset: (a) MODEL 1 and (b) MODEL 2. MODEL 1 has a wide distribution without a clear estimate (estimates with different signs). Addition of the regression constraints in MODEL 2 results in an undoubted positive relation between the Karnofsky score and the survival time (weight > 0).

Concordance index (c-index): The concordance index is the ratio of the number of concordant pairs and the number of comparable pairs:

$$\text{c-index}(u) = \frac{\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{I}[(u(x_i)-u(x_j))(y_i-y_j)\geq 0]}{\sum_{i=1}^{n}\sum_{j\neq i}\text{comp}(i,j)}, \qquad (24)$$

where $\mathcal{I}$ is the indicator function and $\text{comp}(i,j)$ is defined as before.
Logrank χ^2statistic: Clinicians are often interested in the allocation of risk profiles to patients. In this context it is relevant to test how well the developed prognostic index is able to differentiate high and low risk patients. The patients are therefore divided into a low and high risk group according to whether their prognostic index is below or above the median value. The logrank test is used to test significant differences in survival between both groups. Instead of working with the resulting p-value, we will report on the χ^2-value of the test. The higher this value, the better the separating ability of the model.
Hazard ratio: Estimated by a univariate cox model. All produced prognostic indices, after normalization to restrict its value between 0 and 1, will be fed to the cox model and the resulting hazard ratio will be calculated. The higher the hazard ratio, the better the index.

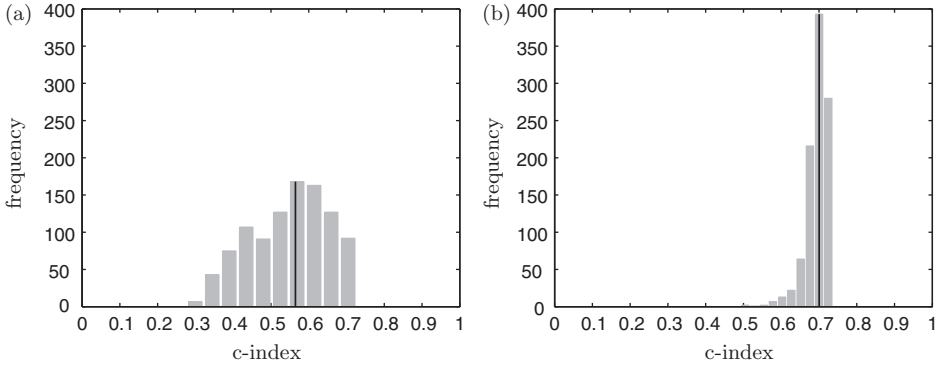


Fig. 4. Distribution of the concordance index on test on 1000 bootstraps of one particular training set of the VLC dataset: (a) MODEL 1 and (b) MODEL 2. MODEL 1 has a wide distribution from which one cannot decide whether the estimated prognostic index is related or inversely related to the failure time. Addition of the regression constraints in MODEL 2 leads to a narrowing of this distribution. This results in an clear positive relation between the prognostic index and the survival time (c-index > 0.5).

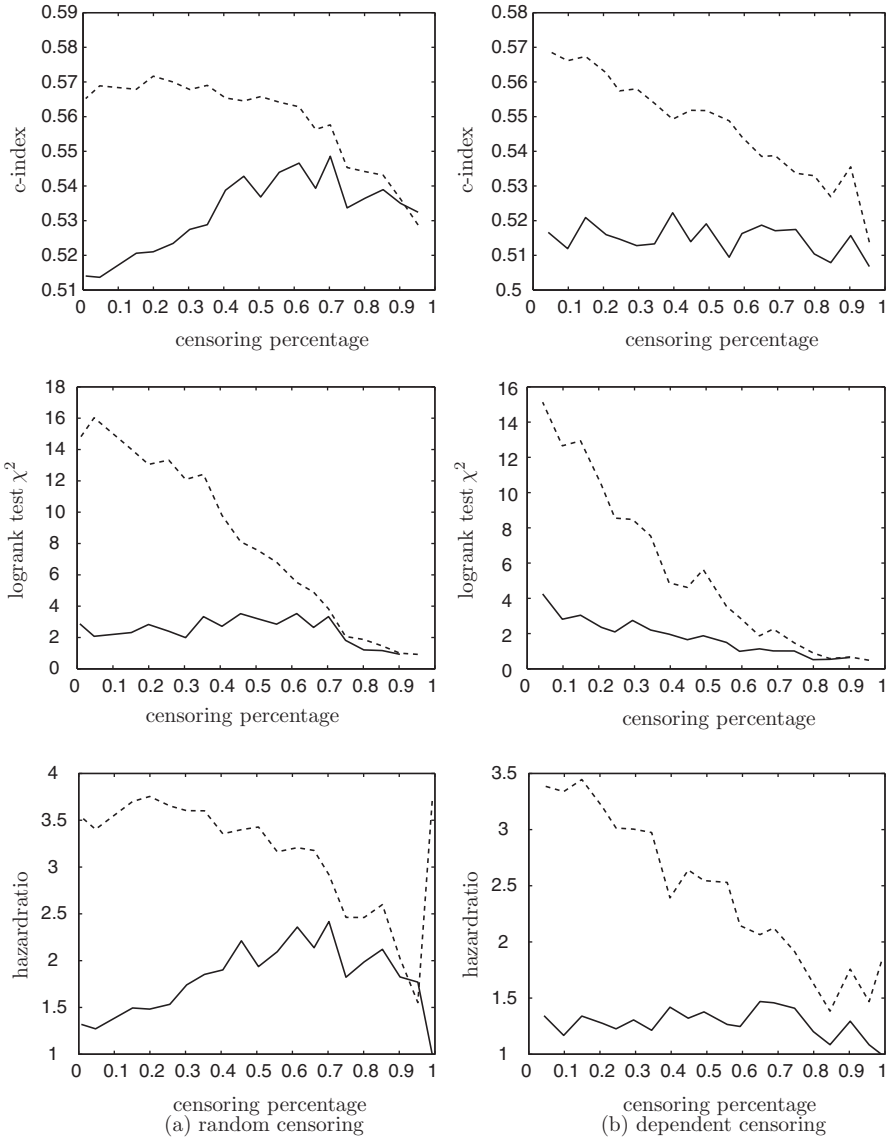


Fig. 5. Comparison between models MODEL 1 (solid line) and MODEL 2 (dashed line) for random (a) and informative (b) censoring in a 10-dimensional artificial example. All variables were normally distributed with zero mean and variance equal to 1. All variables were uncorrelated except for 4 pairs of variables with correlations of 0.7, 0.3, −0.7 and −0.3. The fixed coefficient vector w was taken. The failure time was sampled from an exponential distribution with a parameter equal to $w^T x$, x being the variable vector. The percentage of censored observations was varied from 0 to 100%. The largest performance differences are noted for low censoring percentages. However, MODEL 2 seems to perform better for all censoring percentages, both for random and informative censoring.

4.3. Experiments

This subsection compares models described in the literature (COX model, RANKSVMC, SVCR, and MODEL 1) with the model proposed in this paper (MODEL 2) on 5 clinical, low dimensional, 3 high dimensional and 1 artificial data set. 10-fold cross-validation with coupled simulated annealing [28] was used to tune the hyperparameters. The concordance index was used as model selection criterion. All models were implemented in matlab,[1] using the mosek[2] optimization toolbox for matlab.

4.3.1. Real data

MODEL 2 is compared with MODEL 1, RANKSVMC and SVCR on 50 randomizations between training (two thirds of the data) and test set (remaining third of the data) using a linear and a clinical [19] kernel. Tables 2 and 3 summarize the results. Statistically significant differences between MODEL 2 and the other models were calculated using the Wilcoxon rank sum test. The COX model is tested linearly in the covariates (PH_{linear}) and with penalized smoothing splines ($\text{PH}_{pspline}$) [15], for comparison. It is seen that both MODEL 1 and RANKSVMC are performing significantly less than the other models, independently of the used kernel. Addition of the regression constraints improves the performance significantly. Tables 4 and 5 report the results on the high dimensional data. Here, SVCR outperforms the other methods.

To investigate the difference in performance between models with and without regression constraints, 1000 bootstrap samples from one training set of the VLC dataset were taken. Using a linear kernel, the estimated weights for each covariate and the resulting concordance index on the test set were calculated. Figs. 3 and 4 show the distribution of the estimated weights of the Karnofsky performance score and the concordance index, respectively. The

[1] http://www.mathworks.com/products/matlab/
[2] http://www.mosek.com/

model only taking ranking constraints into account results in high uncertainties about the weight estimates and therefore high differences in concordance index occur for slightly different training sets. Inclusion of regression constraints results in more stable estimates of weights and concordance index.

4.3.2. Artificial data

In this last example, we investigate whether the improvement of MODEL 2 over MODEL 1 is dependent on the censoring percentage of the problem. In this experiment, the clinical setting is mimicked. Therefore, 100 datasets are constructed with 10 variables, 100 training and 500 test observations. All variables were normally distributed with zero mean and variance equal to 1. The correlations between the variables were zero except for the first and second covariate, the third and fourth, the fifth and sixth, and the seventh and eight, having a correlation coefficient of 0.7, 0.3, −0.7 and −0.3, respectively. The vector of regression coefficients w was taken to be $w = [-0.1, 1, 0.3, 0.4, 0.2, 0.03, 0.02, -0.01, -0.02, 0.01]^T$. The failure time was exponentially distributed with parameter equal to $w^T x$, with x the covariate vector. In a first example, the failure time was randomly censored (mimicking administrative censoring); in a second example the censoring time was sampled from the same distribution as the survival time $F_Y(x)$, the spread of the distribution was changed to create different censoring percentages ($G_Y(x) = aF_Y(x)$). Fig. 5 illustrates the results. For both the random and informative censoring, the largest performance differences between both models are noted for low censoring percentages. However, MODEL 2 performs better than the MODEL 1 for all censoring percentages.

5. Conclusions

This work compared different methods for survival analysis based on support vector machines. Three different approaches were discussed: (i) the ranking approach, (ii) the regression approach and (iii) the combined approach. On a theoretical basis, the first and third methods are preferred since they can be linked with well known statistical models for survival analysis. However, the experiments revealed that the ranking approach performs significantly less than both other approaches. Additionally, experiments on high dimensional data showed an increased performance of the regression approach over the combined approach. In conclusion, the authors state that when performance is the first interest of the study, the regression approach should be preferred. However, when a more theoretical basis is preferred, the combined approach might be considered. Thanks to this theoretical basis the link between the ranking and combined approaches and the two most commonly used models for survival analysis within statistics, the proportional hazards model and the log-odds model, is made.

Acknowledgments

This research is supported by Research Council KUL: GOA AMBioRICS, GOA MANET, CoE EF/05/006, IDO 05/010, IOF KP06/11, IOF SCORES4CHEM, several PhD, postdoc and fellow grants; Flemish Government: FWO: PhD and postdoc grants, IBBT, G.0407.02, G.0360.05, G.0519.06, G.0321.06, G.0341.07 and projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0302.07; IWT: PhD Grants, McKnow-E, Eureka-Flite; Belgian Federal Science Policy Office: IUAP P6/04; EU: FP6-2002 LIFESCIHEALTH 503094, IST 2004-27214, FP6-MC-RTN 035801; Prodex-8 C90242; EU: ERNSI. V. Van Belle is supported by a grant from IWT and a post-doctoral grant (BOF) from K.U. Leuven, Belgium.

Appendix A. MODEL 1

The method proposed in this paper is based on the survival model proposed in [8]. We give the deviation of the model in this first appendix. The survival model was written as a ranking problem as

MODEL 1:

$$\begin{aligned} &\min_{w,\epsilon} \quad \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i, \\ &\text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\tilde{j}(i)})) \geq y_i - y_{\tilde{j}(i)} - \epsilon_i, & \forall i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall i = 1, \ldots, n. \end{cases} \end{aligned} \tag{25}$$

Formulating the Lagrangian of (25)

$$\begin{aligned} \mathcal{L}(w, \epsilon; \alpha, \beta) = \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} \alpha_i w^T(\varphi(x_i) - \varphi(x_{\tilde{j}(i)})) \\ - \sum_{i=1}^{n} \alpha_i(-y_i + y_{\tilde{j}(i)} + \epsilon_i) - \sum_{i=1}^{n} \beta_i \epsilon_i, \end{aligned} \tag{26}$$

and solving the Karush–Kuhn–Tucker (KKT) [29] equations

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \displaystyle\sum_{i=1}^{n} \alpha_i(\varphi(x_i) - \varphi(x_{\tilde{j}(i)})) \\ \dfrac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \gamma = \alpha_i + \beta_i, & \forall i = 1, \ldots, n \\ \alpha_i \geq 0, & \forall i = 1, \ldots, n \\ \beta_i \geq 0, & \forall i = 1, \ldots, n \\ \displaystyle\sum_{i=1}^{n} \alpha_i(w^T(\varphi(x_i) - \varphi(x_{\tilde{j}(i)})) - y_i + y_{\tilde{j}(i)} + \epsilon_i) = 0, \\ \displaystyle\sum_{i=1}^{n} \beta_i \epsilon_i = 0, \end{cases} \tag{27}$$

leads, after elimination of the primal variables and β_i to the dual

$$\begin{aligned} &\min_{\alpha_i} \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k (\varphi(x_i) - \varphi(x_{\tilde{j}(i)}))^T (\varphi(x_k) - \varphi(x_{\tilde{l}(k)})) \\ &\qquad - \sum_{i=1}^{n} \alpha_i (y_i - y_{\tilde{j}(i)}), \\ &\text{subject to} \quad 0 \geq \alpha_i \geq \gamma, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{28}$$

The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \alpha_i (\varphi(x_i) - \varphi(x_{\tilde{j}(i)}))^T \varphi(x^\star). \tag{29}$$

Appendix B. MODEL 2

This appendix contains the derivation for the newly proposed model. We start from the model formulation as given in Section 3, describe the Lagrangian and derive the KKT conditions for optimality. MODEL 2 was defined as

MODEL 2:

$$\min_{w,\epsilon,\xi,\xi^*,b} \quad \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}\epsilon_i + \mu\sum_{i=1}^{n}(\xi_i+\xi_i^*),$$

$$\text{subject to} \quad \begin{cases} w^T(\varphi(x_i)-\varphi(x_{\tilde{j}(i)})) \geq y_i - y_{\tilde{j}(i)} - \epsilon_i, & \forall i=1,\ldots,n \\ w^T\varphi(x_i)+b \geq y_i - \xi_i, & \forall i=1,\ldots,n \\ -\delta_i(w^T\varphi(x_i)+b) \geq -\delta_i y_i - \xi_i^*, & \forall i=1,\ldots,n \\ \epsilon_i \geq 0, & \forall i=1,\ldots,n \\ \xi_i \geq 0, & \forall i=1,\ldots,n \\ \xi_i^* \geq 0, & \forall i=1,\ldots,n. \end{cases} \tag{30}$$

The Lagrangian becomes

$$\begin{aligned}\mathcal{L}(w,\epsilon,b;\alpha,\beta) = {} & \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}\epsilon_i + \mu\sum_{i=1}^{n}(\xi_i+\xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i(w^T(\varphi(x_i)-\varphi(x_{\tilde{j}(i)})) - y_i + y_{\tilde{j}(i)} + \epsilon_i) \\ & - \sum_{i=1}^{n}\beta_i(w^T\varphi(x_i)+b-y_i+\xi_i) \\ & - \sum_{i=1}^{n}\beta_i^*(-\delta_i(w^T\varphi(x_i)+b-y_i)+\xi_i^*) - \sum_{i=1}^{n}\eta_i\epsilon_i \\ & - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\nu_i^*\xi_i^*. \end{aligned} \tag{31}$$

The KKT conditions for optimality are

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w}=0 \rightarrow w = \sum_{i=1}^{n}\alpha_i(\varphi(x_i)-\varphi(x_{\tilde{j}(i)})) + \sum_{i=1}^{n}(\beta_i-\delta_i\beta_i^*)\varphi(x_i), \\ \frac{\partial\mathcal{L}}{\partial b}=0 \rightarrow \sum_{i=1}^{n}(-\beta_i+\delta_i\beta_i^*)=0 \\ \frac{\partial\mathcal{L}}{\partial\epsilon_i}=0 \rightarrow \gamma=\alpha_i+\eta_i, & \forall i=1,\ldots,n \\ \frac{\partial\mathcal{L}}{\partial\xi_i}=0 \rightarrow \mu=\beta_i+\nu_i, & \forall i=1,\ldots,n \\ \frac{\partial\mathcal{L}}{\partial\epsilon_i}=0 \rightarrow \gamma=\beta_i^*+\nu_i^*, & \forall i=1,\ldots,n \\ \alpha_i,\beta_i,\beta_i^*,\eta_i,\xi_i,\xi_i^* \geq 0, & \forall i=1,\ldots,n \\ \sum_{i=1}^{n}\alpha_i(w^T(\varphi(x_i)-\varphi(x_{\tilde{j}(i)}))-y_i+y_{\tilde{j}(i)}+\epsilon_i)=0, \\ \sum_{i=1}^{n}\beta_i\epsilon_i=0 \\ \sum_{i=1}^{n}\beta_i(w^T\varphi(x_i)+b-y_i+\xi_i)=0 \\ \sum_{i=1}^{n}\beta_i^*(-\delta_i(w^T\varphi(x_i)+b-y_i)+\xi_i^*)=0, \\ \sum_{i=1}^{n}\eta_i\epsilon_i=0 \\ \sum_{i=1}^{n}\nu_i\xi_i=0 \\ \sum_{i=1}^{n}\nu_i^*\xi_i^*=0. \end{cases} \tag{32}$$

After elimination of all unknowns except α_i, β_i and β_i^* the solution is found as

$$\begin{aligned}\min_{\alpha_i,\beta,\beta^*} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k(\varphi(x_i)-\varphi(x_{\tilde{j}(i)}))^T(\varphi(x_k)-\varphi(x_{\tilde{l}(k)})) \\ & + \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i\beta_k+\delta_i\delta_k\beta_i^*\beta_k^*)\varphi(x_i)^T\varphi(x_k) - \sum_{i=1}\alpha_i(y_i-y_{i-1}) \\ & + \sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i-\delta_k\beta_k^*)\alpha_k\varphi(x_i)^T(\varphi(x_k)-\varphi(x_{\tilde{l}(k)})) \\ & - \sum_{i=1}^{n}\sum_{k=1}^{n}\beta_i^*\delta_i\beta_i\varphi(x_i)^T\varphi(x_k) - \sum_{i=1}^{n}\beta_i y_i + \sum_{i=1}^{n}\delta_i\beta_i^* y_i, \end{aligned}$$

$$\text{subject to} \quad \begin{cases} 0 \geq \alpha_i \geq \gamma, & \forall i=1,\ldots,n \\ 0 \geq \beta_i \geq \mu, & \forall i=1,\ldots,n \\ 0 \geq \beta_i^* \geq \mu, & \forall i=1,\ldots,n. \end{cases} \tag{33}$$

The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n}(\alpha_i(\varphi(x_i)-\varphi(x_{\tilde{j}(i)})) + (\beta_i-\delta_i\beta_i^*)\varphi(x_i))^T\varphi_p(x^\star) + b. \tag{34}$$

References

[1] Kalbfleisch JD. Likelihood methods and nonparametric tests. Journal of the American Statistical Association 1978;73(361):167–70.
[2] Dabrowska DM, Doksum KA. Partial likelihood in transformation models with censored data. Scandinavian Journal of Statistics 1988;15(1):1–23.
[3] Doksum K, Gasko M. On a correspondence between models in binary regression and survival analysis. International Statistical Review 1990;58:243–52.
[4] Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika 1995;82(4):835–45.
[5] Cheng SC, Wei LJ, Ying Z. Predicting survival probabilities with semiparametric transformation models. Journal of the American Statistical Association 1997;92(437):227–35.
[6] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley Series in Probability and Statistics; 2002.
[7] Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 1972;34(2):187–220.
[8] Van Belle V, Pelckmans K, Suykens JAK, Van Huffel S. Learning transformation models for ranking and survival analysis. Journal of Machine Learning Research 2011;12:819–62.
[9] Van Belle V, Pelckmans K, Suykens JAK, Van Huffel S. Support vector machines for survival analysis. In: Ifeachor E, Anastasiou A, editors. Proceedings of the third international conference on Computational Intelligence in Medicine and Healthcare (CIMED). 2007. p. 1–8.
[10] Van Belle V, Pelckmans K, Suykens JAK, Van Huffel S. Survival SVM: a practical scalable algorithm. In: Verleysen M, editor. Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN), d-side, Evere. 2008. p. 89–94.
[11] Evers L, Messow CM. Sparse kernel methods for high-dimensional survival data. Bioinformatics 2008;24(14):1632–8.
[12] Shivaswamy PK, Chu W, Jansche M. A support vector approach to censored targets. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM). California: IEEE Computer Society; 2007. p. 655–60.
[13] Khan FM, Zubek VB. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In: Giannotti F, Gunopulos D, Turini F, Zaniolo C, Ramakrishnan N, Wu X, editors. Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM). California: IEEE computer society; 2008. p. 863–8.
[14] Lawless JF. Survival and event history analysis. In: Wiley reference series in biostatistics. Chapter: Parametric models in survival analysis. West Sussex, England: Wiley; 2006. p. 345–55.
[15] Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science 1996;11:89–121.
[16] Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. Journal of the Royal Statistical Society: Series B 1998;60:271–93.
[17] Vapnik V. Statistical learning theory. New York: Wiley and Sons; 1998.

[18] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society A 1909;209:415–46.
[19] Daemen A, De Moor B. Development of a kernel function for clinical data. In: Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). Piscataway: IEEE; 2009. p. 5913–7.
[20] Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers 2000: 115–32.
[21] Emerson SS, Banks PLC. Case studies in biometry. Chapter: Interpretation of a leukemia trial stopped early. New York: Wiley-Interscience; 1994. p. 275–99.
[22] Prentice RL. A log gamma model and its maximum likelihood estimation. Biometrika 1974;61(3):539–44.
[23] Byar D, Green S. Prognostic variables for survival in a randomized comparison of treatments for prostatic cancer. Bulletin du Cancer 1980;67: 477–90.
[24] Therneau TM, Grambsch PM. Modeling survival data: extending the cox model. 2nd ed. New York: Springer-Verlag; 2000.
[25] Schumacher M, Basert G, Bojar H, Huebner K, Olschewski M, Sauerbrei W, et al. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. Journal of Clinical Oncology 1994;12:2086–93.
[26] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society Series A 1999;162(1):71–94.
[27] Bøvelstad HMM, Nygård S, Størvold HLL, Aldrin M, Borgan O, Frigessi A, et al. Predicting survival from microarray data – a comparative study. Bioinformatics 2007;23(16):2080–7.
[28] Xavier de Souza S, Suykens JAK, Vandewalle J, Bolle D. Coupled simulated annealing. IEEE Transactions on Systems, Man, and Cybernetics – Part B 2010;40(2):320–35.
[29] Boyd S, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004.