BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 87–94
doi:10.1093/bioinformatics/btq617

*Gene expression*

Advance Access publication November 8, 2010

# Improved performance on high-dimensional survival data by application of Survival-SVM

V. Van Belle[1,*], K. Pelckmans[2], S. Van Huffel[1] and J. A. K. Suykens[1]

[1]Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium and [2]Department of Information Technology, Division Syscon, Uppsala University, ITC Building 2, SE-751 05 Uppsala, Sweden

Associate Editor: John Quackenbush

## ABSTRACT

**Motivation:** New application areas of survival analysis as for example based on micro-array expression data call for novel tools able to handle high-dimensional data. While classical (semi-) parametric techniques as based on likelihood or partial likelihood functions are omnipresent in clinical studies, they are often inadequate for modelling in case when there are less observations than features in the data. Support vector machines (SVMs) and extensions are in general found particularly useful for such cases, both conceptually (non-parametric approach), computationally (boiling down to a convex program which can be solved efficiently), theoretically (for its intrinsic relation with learning theory) as well as empirically. This article discusses such an extension of SVMs which is tuned towards survival data. A particularly useful feature is that this method can incorporate such additional structure as additive models, positivity constraints of the parameters or regression constraints.

**Results:** Besides discussion of the proposed methods, an empirical case study is conducted on both clinical as well as micro-array gene expression data in the context of cancer studies. Results are expressed based on the logrank statistic, concordance index and the hazard ratio. The reported performances indicate that the present method yields better models for high-dimensional data, while it gives results which are comparable to what classical techniques based on a proportional hazard model give for clinical data.

**Contact:** vanya.vanbelle@esat.kuleuven.be

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on July 8, 2010; revised on October 13, 2010; accepted on October 29, 2010

## 1 INTRODUCTION

Survival analysis concerns the study of the time to occurrence of a certain event. It is best known in cancer studies where one is interested in characterizing which patients will relapse after surgery or pass away. Other applications can be found in electronic or mechanic components where the characterization of the lifetime is useful for optimizing maintenance strategies. Typically, survival data contain censored observations. Censoring indicates a lack of information on the outcome. For example, in a clinical study examining relapse of breast cancer patients, where patients are included between 1990 and 2000 and followed until 2008, not all patients will have experienced relapse. For these patients, the failure time is right censored.

The statistical literature describes different models for the analysis of failure time data, an overview of which can be found in Kalbfleisch and Prentice (2002). The largest breakthrough in modelling survival data came in 1972 when Cox proposed his proportional hazard model (PH) (Cox, 1972). The PH model is a semi-parametric model assuming that the hazard of an observation (the instantaneous risk of occurrence of the event given that the event did not occur before) is proportional to a 'baseline' hazard common to all observations. Proportionality is modelled as the exponential of a linear function of the covariates. The semi-parametric character of this model comes from the fact that the baseline hazard is left unspecified. Success of this model is to a certain extend due to the description and analysis of a corresponding partial likelihood function whose properties are proven to be quite similar to ordinary likelihood functions.

Although the PH model is perhaps the most common survival model, some drawbacks remain. At first, the model is based on the assumption that hazards for different subjects are proportional to each other, an assumption which is not always realistic. A second drawback is the restrictive parametric form in which the variables affect the outcome. During the last decade, different methods dealing with one or both of those drawbacks have been proposed. The linear parametric form was generalized by means of ANOVA models, splines and artificial neural networks [see Bakker *et al.* (2004); Huang *et al.* (2000); Leblanc and Crowley (1999) and references therein]. Models dealing with non-linear covariate effects and not imposing the proportional hazards assumption were proposed in the field of artificial neural networks (Biganzoli *et al.*, 1998; Faraggi and Simon, 1995; Lisboa *et al.*, 2003). Due to the good performances obtained with machine learning methods in regression and classification (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Suykens *et al.*, 2002; Vapnik, 1998), ideas underlying methods of machine learning started to pervade also traditional modelling areas. Support vector machine (SVM) regression for censored data was proposed by Shivaswamy *et al.* (2007) and a rank regression approach was given in Evers and Messow (2008); Van Belle *et al.* (2007, 2009a). In Van Belle *et al.* (2010a), a least-squares SVMs approach making use of ranking and regression constraints was used.

This work compares an SVM-based method incorporating ranking and regression constraints (survival support vector machines: SSVM)

*To whom correspondence should be addressed.

as proposed in Van Belle *et al.* (2009b) with the linear and non-linear PH model (Cox, 1972) and with the partial logistic artificial neural network model with automatic relevance detection (PLANNARD) (Lisboa *et al.*, 2003). Due to the growing interest in micro-array data within survival studies, the need for survival methods which perform well on high-dimensional data is increasing. Therefore, SSVM is compared with other survival methods, specifically adapted for high-dimensional data. Merely a few papers till date approach this problem in the context of SVMS, see for example Evers and Messow (2008). Many earlier published studies propose to reduce the problem of estimating a good prognostic index to a classification problem. Essentially, they discriminate between observations with (i) a poor prognosis (non-survivors) and (ii) a good prognosis (survivors). Although this approach is plausible in case all observations are followed for an equally long follow-up period, it remains open how this approach can be applied when censoring can occur at arbitrary moments. For more details, we refer to Callas *et al.* (1998); Green and Symons (1983).

This article is organized as follows. Section 2 describes the setup of survival analysis and gives some more details about the PH, PLANN and SSVM methods. After a short introduction of the SSVM method, a feature selection technique is proposed. This method results in a sparse coefficient vector and will turn out to be useful for high-dimensional datasets. In Section 3, a comparison is made between SSVM, the PH and the PLANNARD approaches. In a first experiment, the methods are compared on clinical cancer datasets. A second experiment involves high-dimensional micro-array data. The article ends with a discussion and some concluding remarks.

## 2 METHODS

This section describes three different approaches for estimating prognostic indices for survival problems. In the remainder of this text, the $p$-th covariate of observation $i$ will be denoted by $x_i^p$; the vector containing all covariates of the $i$-th observation is represented as $x_i$. The $p$-th element of a vector $w$ will be denoted by $w_p$.

Consider a dataset $\mathcal{D}=\{(x_i,y_i,\delta_i)\}_{i=1}^n$, where $x_i, y_i$ and $\delta_i$ represent the covariate vector, a positive survival time and an event indicator for observation $i$, respectively. The event indicator is equal to 1 if an event occurred, and zero if the subject is (right) censored at time $y_i$. The survival function $S(t)=P(y>t)$ is defined as the probability of not having experienced the event until time $t$. The hazard $\lambda(t)$ is defined as the risk of the event to occur at time $t$, given that the event did not occur before that time:

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{P(t\le y<t+\Delta t|y\ge t)}{\Delta t}. \tag{1}$$

### 2.1 The proportional hazards model

The proportional hazard model (Cox, 1972) is built on two assumptions: (i) proportional hazards and (ii) linear parametric form in the covariates. Let $\lambda(x,t)$ be the hazard, $x$ a specific covariate vector and $t$ the time at which one wants to estimate the hazard. The model becomes

$$\lambda(x,t)=\lambda_0(t)\exp(w^Tx). \tag{2}$$

The risk associated with an observation with covariates $x_i$ is related to $w^Tx_i$, also called a prognostic index. The parameters $w$ are inferred by maximizing the partial likelihood function defined as

$$L(w)=\prod_{j=1}^{k}\left(\frac{\exp(w^Tx_j)}{\sum_{l\in\mathcal{R}_j}\exp(w^Tx_l)}\right), \tag{3}$$

where $\mathcal{R}_j$ represents all patients at risk at the $j$-th failure time and $k$ is the number of distinct failure times. In practice, the log partial likelihood $\ell(w)=\log(L(w))$ is maximized.

*2.1.1 Penalized proportional hazard regression* In order to reduce the problem of overfitting, penalized forms of the likelihood function were introduced. A penalization term that is often used is called ridge regression or weight decay [see e.g. Hastie *et al.* (2001) and references therein]. The penalized log partial likelihood function then becomes

$$\ell_\lambda(w)=\ell(w)-\frac{\lambda}{2}w^Tw, \tag{4}$$

with $\lambda\ge 0$ a regularization constant. We will refer to this model as $\text{PH}_{l2}$ in the remainder of this work. A second penalized model ($\text{PH}_{l1}$) maximizes the log partial likelihood subject to $\sum_{p=1}^{d}|w_p|\le s$, where $s$ needs to be optimized [see Goeman (2010); Tibshirani (1997)].

*2.1.2 Including non-linearities: PH with penalized smoothing splines* To relax the PH model towards non-parametric effects of the covariates, the use of penalized splines in PH models was proposed by Eilers and Marx (1996), among others. The PH model is rewritten as

$$\lambda(x,t)=\lambda_0(t)\exp(f(x)), \tag{5}$$

where $f(x)$ represents a function of $x$. When using B-splines, this function is a linear combination of $m$ basis functions $B_a$:

$$f(x)=\sum_{a=1}^{m}w_aB_a(x). \tag{6}$$

The penalized partial likelihood function then becomes

$$\ell_\lambda(w)=\ell(w)-\frac{\lambda}{2}\sum_{a=3}^{m}(w_a-2w_{a-1}+w_{a-2})^2, \tag{7}$$

with $\lambda$ a positive regularization constant. This approach will be used for comparison and will be denoted by $\text{PH}_{pspline}$.

*2.1.3 Practical approaches for handling high-dimensional data* When dealing with high-dimensional data, the PH model needs to be adapted to avoid overfitting as before. Three different practical adaptations of the standard PH model are studied and implemented in Bøvelstad *et al.* (2007), and matlab code was provided by the authors. A first method (PCR) uses (unsupervised) principal component analysis (PCA) to select a number of principal components of the expression data accounting for the largest variation in gene expression profiles. The selected principal components are then used as covariates (Hastie *et al.*, 2001) in a standard PH model. The SPCR method selects a subset of genes which correlate best with the observed survival using a univariate PH model and applies PCR on the resulting genes (Bair *et al.*, 2006). The PLS method creates new features as a linear combination of the covariates and uses these as input variables for the PH model (Nygård *et al.*, 2008).

### 2.2 Multi-layer perceptron models

*2.2.1 The partial logistic artificial neural network model* Biganzoli *et al.* (1998) proposed a partial logistic artificial neural network (PLANN) for the analysis of survival data. This multi-layer perceptron contains three layers: (i) an input layer containing a neuron for each input $p=1,\ldots,d$, and one neuron for the time variable; (ii) a hidden layer containing $h=1,\ldots,H$ hidden neurons; and (iii) an output layer with one output neuron. The model is trained as follows. First the time in which observations were followed is divided into time intervals $[t_{k-1},t_k], k=1,\ldots,K$. The goal is then to estimate the chance that an observation will experience the event at study within each of these intervals, given that they did not relapse before. Therefore, each data point is replicated for each time interval in which the outcome is known. The input of the model consists of two parts: the covariate vector and a time variable indicating the time interval. The output is one if the patient experienced the event under study within the considered time interval, zero otherwise. For discrete time studies, the output of the model will represent the predicted hazard within each time interval.

*2.2.2 Feature selection: PLANN model with automatic relevance detection* Lisboa *et al.* (2003) proposed to incorporate Bayesian automatic relevance determination (ARD) in PLANN. A penalization term $\alpha$ is linked with each parameter of PLANN and Bayes' theorem is used as a regularization framework. As the PH model, PLANN estimates parameters by optimizing the likelihood function of the parameters $w$, given the data $\mathcal{D}$, the penalization terms $\alpha$ and the model hypothesis space $\mathcal{H}$. According to Bayes' theorem, this can be expressed as:

$$P(w|\mathcal{D},\alpha,\mathcal{H}) = \frac{P(\mathcal{D}|w,\alpha,\mathcal{H})P(w|\alpha,\mathcal{H})}{P(\mathcal{D}|\alpha,\mathcal{H})}. \tag{8}$$

The PLANNARD procedure has three levels. On the first level, the regularization parameters $\alpha$ are assumed to be fixed and the prior distribution $P(w|\alpha,\mathcal{H})$ of the weights $w$ is assumed to be normal, centered at zero, with variance $1/\alpha$. $P(w|\mathcal{D},\alpha,\mathcal{H})$ can then be optimized in function of $w$. On a second level, the regularization parameters are estimated. On a third level, the evidence in support of a particular model hypothesis $\mathcal{H}$ is estimated. The prior distribution $P(\alpha|\mathcal{H})$ is assumed to be log-normal. A flat prior is assumed for the model space. The optimal parameters are found by iterating over the three levels. The values of $\alpha$ are inversely proportional to the relevance of the variables. See Lisboa *et al.* (2003) for more details.

## 2.3 Support vector machines

*2.3.1 Survival support vector machines* An SVM-based method formulating the problem of estimating a prognostic index as a ranking problem was proposed in Evers and Messow (2008); Van Belle *et al.* (2007), and computationally simplified in Van Belle *et al.* (2008). Van Belle *et al.* (2009b) proposed the use of an SVM approach for survival analysis with ranking and regression constraints. Comparing these methods with those using only ranking constraints, the former performed significantly better. In this article, we refer to MODEL 2, as discussed in the latter article as to survival SVM (SSVM).

The approach taken in SSVM is quite different from the other methods discussed in the previous subsections. Instead of inferring the hazard directly, a utility function of the covariates is searched such that the resulting utility values are as concordant as possible with the corresponding observed failures. For a thorough investigation of this reasoning and its relation with the technique of maximizing the margin in standard SVMs, see Van Belle *et al.* (2009a). A good concordance is found by implementing ranking constraints, penalizing misranking between pairs of observations. In addition, regression constraints are included. Those direct the estimate towards prediction of the events. In Van Belle *et al.* (2009b), it is empirically observed that such mechanism yields improved estimates over a pure ranking-based approach. Technically, consider a transformation with feature map $\varphi(x)$ of the covariates $x$ such that the utility estimate for observation $i$ equals $u(x_i)=w^T\varphi(x_i)$. The problem is then formulated as:

$$\min_{w,\epsilon_{i,i-1},\xi_i,\xi_i^*} \frac{1}{2}w^Tw+\gamma\sum_{i=2}^{n}\epsilon_{i,i-1}+\mu\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$
$$\text{s.t.}\begin{cases} w^T\varphi(x_i)-w^T\varphi(x_{i-1})+\epsilon_{i,i-1}\geq y_i-y_{i-1} & \\ & \forall\, i=2,\ldots,n \\ w^T\varphi(x_i)+\xi_i\geq y_i & \forall\, i=1,\ldots,n \\ -\delta_i w^T\varphi(x_i)+\xi_i^*\geq -\delta_i y_i & \forall\, i=1,\ldots,n \\ \epsilon_{i,i-1}\geq 0 & \forall\, i=2,\ldots,n \\ \xi_i\geq 0 & \forall\, i=1,\ldots,n \\ \xi_i^*\geq 0 & \forall\, i=1,\ldots,n, \end{cases} \tag{9}$$

with $\gamma$ and $\mu$ positive regularization constants and $\epsilon_{i,i-1}, \xi_i$ and $\xi_i^*$ slack variables. In the above formulation, the data points are sorted according to their failure (or censoring) time such that $y_i \geq y_{i-1}$. The first set of constraints leads to a solution for which $u(x_i)\geq u(x_{i-1})$ if $y_i \geq y_{i-1}$ is mostly satisfied: those are referred to as to the *ranking constraints*. The second and third set of constraints are referred to as the *regression constraints*, where the equality between $y_i$ and $u(x_i)$ is desired only for non-censored observations. Since $\xi^*=0$ for right censored cases ($\delta_i=0$), the outcome is targeted higher than the survival time for these cases. Non-linearities are modelled by the choice of $\varphi(x)$. However, thanks to the kernel trick or $K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$, this transformation function does not need to be specified explicitly. Common kernels include: (i) the linear kernel: $K(x,z)=x^Tz$; (ii) the polynomial kernel of degree $a$: $K(x,z)=(\tau+x^Tz)^a$ with $\tau\geq 0$; and (iii) the RBF kernel defined as $K(x,z)=\exp\left(-\frac{||x-z||_2^2}{\sigma^2}\right)$. More recently, a kernel for clinical data (Daemen and De Moor, 2009) was proposed as an additive kernel $K(x,z)=\sum_{p=1}^{d}K^p(x,z)$, where $K^p(\cdot,\cdot)$ depends on the type of the variable $x^p$. For continuous and ordinal variables, $K^p(\cdot,\cdot)$ is defined as

$$K^p(x^p,z^p)=\frac{c-|x^p-z^p|}{c}, \tag{10}$$

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

$$K^p(x^p,z^p)=\begin{cases}1 \text{ if } x^p=z^p \\ 0 \text{ if } x^p\neq z^p.\end{cases} \tag{11}$$

The estimated outcome $u(x_\star)$ for any new observation $x_\star$ is then given as

$$u(x_\star) = \sum_{i=2}^{n}\alpha_i[K(x_i,x_\star)-K(x_{i-1},x_\star)] + \sum_{i=1}^{n}(\beta_i-\delta_i\beta_i^*)K(x_i,x_\star), \tag{12}$$

with $\{\alpha_i\}, \{\beta_i\}$ and $\{\beta_i^*\}$ the sets of Lagrange multipliers corresponding to the first, second and third set of constraints in (9). For more information on this subject, we refer to Van Belle *et al.* (2009a, b). The resulting optimization problem is a convex Quadratic Programming (QP) problem, which can be solved efficiently using contemporary solvers.

*2.3.2 Feature selection in linear models: positivity constraints* In clinical applications, one is not only interested in trying to find a 'good' prognostic index. Searching which variables/genes are relevant (and should be measured in the future) and which are not needed, is equally important. Feature selection is included in SSVM by constraining the weights $w$ to positive weights and will be denoted as SSVMP. If the true parameters were positive, this constraint would not introduce extra bias unlike an $L_1$ approach. This estimate is obtained by solving the problem

$$\min_{w,\epsilon_{i,i-1},\xi_i,\xi_i^*} \frac{1}{2}w^Tw+\gamma\sum_{i=2}^{n}\epsilon_{i,i-1}+\mu\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$
$$\text{s.t.}\begin{cases} w^Tx_i-w^Tx_{i-1}+\epsilon_{i,i-1}\geq y_i-y_{i-1} & \\ & \forall\, i=2,\ldots,n \\ w^Tx_i+\xi_i\geq y_i & \forall\, i=1,\ldots,n \\ -\delta_i w^Tx_i+\xi_i^*\geq -\delta_i y_i & \forall\, i=1,\ldots,n \\ w_p\geq 0 & \forall\, p=1,\ldots,d \\ \epsilon_{i,i-1}\geq 0 & \forall\, i=2,\ldots,n \\ \xi_i\geq 0 & \forall\, i=1,\ldots,n \\ \xi_i^*\geq 0 & \forall\, i=1,\ldots,n. \end{cases} \tag{13}$$

This set of constraints is only relevant after preprocessing the data, in order to ensure that negative relations are not ignored. Therefore, the concordance (Harrell *et al.*, 1988) between each variable and the failure time is calculated. Each variable $x^p$ with a concordance less than 0.5 is switched as $x^p \leftarrow -x^p$. Constraining the weights to be positive will then lead to weights close to zero for irrelevant variables, and higher weights for relevant variables. This technique is closely related with the non-negative least squares estimator, and with the non-negative garotte estimator (Breiman, 1995).

*2.3.3 Feature selection in non-linear models: a two-step approach* Since the above formulation cannot be extended directly to the case non-linear kernels are used, we propose a two-step approach in order to include feature selection [see also Van Belle *et al.* (2010a)]. In a first step, the method as described in (9) is solved. Using a non-linear but additive kernel $K(x_i,x_j)=\sum_{p=1}^{d}K^p(x_i^p,x_j^p)$, the estimated non-linear transformation of each covariate $p$ can be estimated as

$$\tilde{x}_\star^p = u^p(x_\star^p) = \sum_{i=2}^{n}\alpha_i\left[K^p(x_i^p,x_\star^p)-K^p(x_{i-1}^p,x_\star^p)\right] + \sum_{i=1}^{n}\left(\beta_i-\delta_i\beta_i^*\right)K^p(x_i^p,x_\star^p). \quad (14)$$

The second step performs feature selection by considering $\tilde{x}_\star^p$ as the covariates and applying method (13) with a linear kernel. We will refer to this approach using a clinical kernel in the first step as SSVMP$_{clinical}$.

## 3 RESULTS

In subsection 3.1, the performance of SSVM, PH and PLANNARD approaches are compared based on clinical datasets. In subsection 3.2, SSVM is compared with other methods, able to deal with high-dimensional data. The design parameters $\gamma$, $\mu$, $\lambda$ and $s$ are tuned using a 10-fold cross-validation criterion.

The different methods will be compared using three performance measures. Clinicians are typically interested in groups of patients with higher or lower risk profiles. Therefore, a first performance measure denotes the concordance between the predicted and observed order of relapse: the concordance index (c-index) (Harrell *et al.*, 1988). In the same perspective, patients are divided into two risk groups according to the prognostic index obtained with each model. The median prognostic index is used as threshold for defining the two groups. The logrank test $\chi^2$ statistic, measuring the difference in survival between both groups will be reported. As a third measure the hazard ratio as calculated by a univariate PH model using the estimated prognostic indices, normalized to the unit interval, is reported. For all three measures, a higher value corresponds to a better performance. In all experiments, datasets were 50 times randomly divided into training (2/3 of the data) and test (1/3 of the data) sets. Models were estimated based on the training data. Results are calculated based on the test data. Table 1 gives an overview of the different methods and their properties.

Complementary to the reported figures relating performance and clinical usefulness of the presented methods, a comparison regarding the effective CPU time required to train each method is reported in the Supplementary Material. Those results confirm the claim of fast and effective implementations compared with state-of-the-art approaches if a suitable contemporary optimization solver is used.

### 3.1 Clinical cancer data

This subsection compares performances obtained by SSVM, SSVMP, PH and PLANNARD methods applied on six different clinical datasets[1]:

- The leukemia dataset (Emerson and Banks, 1994) contains observations of 129 patients with leukemia. The available variables are as follows: treatment (daunorubicin or idarubicin), sex, age, Karnofsky score (indicating how well a patient having cancer is functioning, expressed on a scale from 0% to 100% ), baseline white blood cells, baseline platelets, baseline haemoglobin, kidney transplant (binary), complete remission and time until complete remission. In a first experiment, the endpoint is time until death (LD). In a second experiment, the endpoint is time until complete remission (LCR) and time until death is not taken into account.
- The Veteran's Administration Lung Cancer Trial (VLC; Kalbfleisch and Prentice, 2002; Prentice, 1974) incorporated 137 men with advanced inoperable lung cancer. Patients were randomized to a standard or test chemotherapy. Only nine patients were still alive at the end of the study. The available variables are as follows: the Karnofsky performance score, age, prior therapy (binary), histological type of the tumour, treatment and months between diagnosis and randomization.
- The data on prostatic cancer (PC; Byar and Green, 1980) contains 506 patients with four types of treatment (placebo, 0.2, 1.0 or 5.0 mg diethylstilbestrol daily). Although this dataset contains information on competing risks, we use it to estimate survival where death due to prostatic cancer is the event under study. The variables are as follows: treatment, age, weight index, performance index, history of cardiovascular disease, size of the tumour, a combined index of stage and histologic grade and serum haemoglobin. Of total, 483 patients had complete information on the variables mentioned above. In all, 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; Therneau and Grambsch, 2000) is a subset of data concerning advanced lung cancer patients, conducted at the North Central Cancer Treatment Group, Rochester Minnesota (Loprinzi *et al.*, 1994). The subset used here contains 167 patients with full information on the following variables: age, sex, the physician's estimate of the ECOG performance and Karnofsky score, patient's estimate of the Karnofsky score, calories consumed at meals and weight loss in the last 6 months. In all, 120 (72%) patients died during the study period.
- The German Breast Cancer Study (BC; Sauerbrei and Royston, 1999; Schumacher *et al.*, 1994) is a dataset containing observations of 720 breast cancer patients, who were recruited in 41 centres between July 1984 and December 1989. Available variables are as follows: hormonal therapy (binary), menopausal status (binary), patient's age at diagnosis, tumour grade, tumour size, the number of positive lymph nodes, expression of the progesterone and oestrogen receptors (in fmol). The study is performed on the 686 cases with complete data. In all, 299 (44%) of these patients had a breast cancer-related event (remission) during the study period.

The SSVM and SSVMP methods are tested using two different kernels: the linear kernel (SSVM$_{linear}$ and SSVMP$_{linear}$) and the clinical kernel (SSVM$_{clinical}$ and SSVMP$_{clinical}$). Earlier publications of the authors compare with results obtained with a RBF kernel, see Van Belle *et al.* (2010b). The PH model is tested using a linear parametric form of the covariates (PH$_{linear}$), using a ridge regression penalized partial likelihood function (PH$_{l2}$), using a LASSO penalization (PH$_{l1}$) and with penalized smoothing splines (PH$_{pspline}$). Table 2 illustrates the

[1]Data available on http://lib.stat.cmu.edu/datasets and http://cran.r-project.org/web/packages/survival/index.html.

**Table 1.** Summary of the different methods used in the experiments

| Model | (Equation) [References] | Non-linear | Covariate selection | High dim. | Hazard | Kernel | Software | Model selection |
|---|---|---|---|---|---|---|---|---|
| SSVM$_{linear}$ | (9) | | | ✓ | | Linear | mosek, matlab | CV |
| | (9) | ✓ | | ✓ | | Clinical | mosek, matlab | CV |
| SSVMP$_{linear}$ | (13) | | ✓ | ✓ | | Linear | mosek, matlab | CV |
| SSVMP$_{clinical}$ | (14) | ✓ | ✓ | ✓ | | Clinical | mosek, matlab | CV |
| PH$_{linear}$ | Cox (1972) | | | | ✓ | | matlab | |
| PH$_{l2}$ | Bøvelstad *et al.* (2007) | | | ✓ | ✓ | | matlab | CV |
| PH$_{l1}$ | Goeman (2010) | ✓ | ✓ | ✓ | ✓ | | R | CV |
| PH$_{pspline}$ | Eilers and Marx (1996) | ✓ | | | ✓ | | R | CV |
| PLANNARD | Lisboa *et al.* (2003) | ✓ | ✓ | | ✓ | | matlab | Bayesian |
| PCR | Hastie *et al.* (2001) | | | ✓ | ✓ | | matlab | CV |
| SPCR | Bair *et al.* (2006) | | | ✓ | ✓ | | matlab | CV |
| PLS | Nygård *et al.* (2008) | | | ✓ | ✓ | | matlab | CV |

It is indicated whether the methods are able (a) to model non-linear effects of the covariates, (b) to perform covariate selection, (c) to handle high-dimensional data well, (d) whether or not the hazard is estimated, (e) which kernel is used for kernel-based methods, (f) which software program was used and (g) how model selection was done in this article. CV: cross-validation.

results. None of the methods performs overall better or worse than the other methods.

### 3.2 High-dimensional micro-array data

This subsection reports performances of the described survival methods when used to model high-dimensional data. Three different micro-array datasets are used:

- The Dutch Breast Cancer Data (DBCD) from van Houwelingen *et al.* (2006) is a subset of the data from van de Vijver *et al.* (2002) and contains information on 4919 gene expression levels of a series of 295 women with breast cancer. Measurements were taken from the fresh-frozen tissue bank of The Netherlands Cancer Institute. All 295 tumours were primary invasive breast carcinoma less than 5 cm in diameter. The women were 52 years or younger. The diagnosis was made between 1984 and 1995 and there was no previous history of cancer, except non-melanoma skin cancer. In 79 (26.78%) patients, distant metastases were noted within the study period. The median follow-up was 6.7 years (range 0.05–18.3).
- The diffuse large B-cell lymphoma data (DLBCL) from Rosenwald *et al.* (2002) contains data of 240 patients with diffuse large B-cell lymphoma. For each patient, one observed 7399 different gene expression measurements. The median follow-up time was 2.8 years and 58% of the patients passed away during the study period.
- The Norway/Stanford breast cancer data (NSBCD) as presented in Sørlie *et al.* (2003) contains gene expression measurements of 115 women with breast cancer. Of total, 549 intrinsic genes introduced in Sørlie *et al.* (2003) were measured. Missing values were previously imputed using 10-nearest neighbour imputation. In all, 38 (33%) patients experienced an event within the study period.

The SSVM and SSVMP methods will be compared with PH models adapted for the high-dimensionality in these datasets. Since variable selection is a major issue in high-dimensional datasets, the number of coefficients $w_p$ with an absolute value larger than $10^{-8}$ is reported as # weights. Table 3 indicates that the clinical kernel performs better than the linear one. However, when including a feature selection algorithm, the results of the linear kernel become significantly better. The SSVMP method significantly outperforms all other tested methods. In addition to a better performance, the SSVMP method with a linear kernel results in a sparser model than the other methods. Due to the fact that most PH models approach dimensionality reduction by composing new features as a function of the measured covariates, nearly all gene expressions need to be obtained for any test case [see e.g. Nguyen and Rocke (2002); Park *et al.* (2002)]. The SSVMP method on the other hand is able to obtain a high performance based on less gene expressions.

## 4 DISCUSSION AND CONCLUSIONS

The PH model is most often used in clinical applications, thanks to its easy applicability and interpretability. The disadvantages are that it assumes in its standard form proportional hazards, as well as a linear parametric form of the covariates. Both assumptions can be relaxed, the first one e.g. by including time-dependent variables, the second one e.g. by using regression splines. The PLANN model is a multi-layer perceptron model incorporating non-linearities in the covariates as well as interactions. The main disadvantage of such models is the difficulty to interpret them. Clinicians are generally interested in the contribution of each covariate to the estimated risk. Due to the complex architecture of multi-layer perceptrons, this contribution is less straightforward to recover. Including automatic relevance determination in the PLANN framework is a step in the right direction, although two problems remain. The first disadvantage of PLANN and PLANNARD is that they both reduce the survival problem to time-dependent classification problems. Therefore, all patients need to be replicated at each time at which they are at risk. This replication leads to an exponential increase of the complexity of the estimation problem. The second problem occurs when dealing with high-dimensional data. For these datasets, the parameter space for multi-layer perceptrons becomes very large,

**Table 2.** Performances for different survival methods for red6 clinical datasets

| Dataset | Method | c-index | Logrank $\chi^2$ | Hazard ratio |
|---|---|---|---|---|
| VLC | SSVM$_{linear}$ | $0.68 \pm 0.02^{***}$ | $4.19 \pm 3.17^{***}$ | $7.07 \pm 2.75^{***}$ |
| | SSVM$_{clinical}$ | $0.70 \pm 0.02$ | $7.84 \pm 2.85$ | $10.24 \pm 4.36^{*}$ |
| | SSVMP$_{linear}$ | $\mathbf{0.71 \pm 0.02}$ | $\mathbf{8.13 \pm 3.71}$ | $12.75 \pm 4.65$ |
| | SSVMP$_{clinical}$ | $0.69 \pm 0.02^{*}$ | $6.84 \pm 2.93$ | $7.80 \pm 3.20^{***}$ |
| | PH$_{linear}$ | $0.68 \pm 0.03^{***}$ | $5.45 \pm 1.97^{*}$ | $10.61 \pm 2.99^{**}$ |
| | PH$_{l2}$ | $0.70 \pm 0.02$ | $6.17 \pm 0.98$ | $\mathbf{31.26 \pm 7.47}^{*}$ |
| | PH$_{l1}$ | $0.70 \pm 0.02^{*}$ | $7.04 \pm 2.85^{**}$ | $11.41 \pm 4.47^{**}$ |
| | PH$_{pspline}$ | $0.70 \pm 0.03$ | $8.11 \pm 3.29^{***}$ | $14.29 \pm 6.76^{**}$ |
| | PLANNARD | $\mathbf{0.71 \pm 0.03}$ | $6.76 \pm 2.82$ | $11.18 \pm 6.25$ |
| LCR | SSVM$_{linear}$ | $0.59 \pm 0.02$ | $1.50 \pm 1.30$ | $3.57 \pm 1.54$ |
| | SSVM$_{clinical}$ | $0.58 \pm 0.04$ | $1.05 \pm 0.96$ | $2.79 \pm 1.36$ |
| | SSVMP$_{linear}$ | $0.56 \pm 0.11$ | $1.37 \pm 1.08$ | $\mathbf{4.30 \pm 3.19}$ |
| | SSVMP$_{clinical}$ | $0.59 \pm 0.03$ | $1.14 \pm 0.97$ | $3.10 \pm 1.43$ |
| | PH$_{linear}$ | $\mathbf{0.60 \pm 0.02}$ | $1.60 \pm 1.18$ | $4.05 \pm 1.34$ |
| | PH$_{l2}$ | $\mathbf{0.60 \pm 0.02}$ | $\mathbf{1.80 \pm 1.58}$ | $3.66 \pm 1.04$ |
| | PH$_{l1}$ | $0.57 \pm 0.03^{*}$ | $1.11 \pm 0.89$ | $3.22 \pm 1.37$ |
| | PH$_{pspline}$ | $0.56 \pm 0.03^{*}$ | $1.12 \pm 0.91$ | $3.42 \pm 1.69$ |
| | PLANNARD | $0.56 \pm 0.03$ | $0.72 \pm 0.68$ | $2.22 \pm 1.16$ |
| LD | SSVM$_{linear}$ | $0.67 \pm 0.03^{***}$ | $5.94 \pm 3.31^{***}$ | $16.04 \pm 9.27$ |
| | SSVM$_{clinical}$ | $0.69 \pm 0.03^{*}$ | $9.06 \pm 3.99^{*}$ | $20.98 \pm 9.90$ |
| | SSVMP$_{linear}$ | $\mathbf{0.72 \pm 0.03}$ | $\mathbf{12.54 \pm 3.90}$ | $18.45 \pm 8.43$ |
| | SSVMP$_{clinical}$ | $0.70 \pm 0.03$ | $7.52 \pm 4.06^{***}$ | $\mathbf{29.39 \pm 13.20}^{*}$ |
| | PH$_{linear}$ | $0.69 \pm 0.03^{**}$ | $9.73 \pm 4.23$ | $22.21 \pm 13.02$ |
| | PH$_{l2}$ | $0.68 \pm 0.02$ | $7.13 \pm 3.48$ | $13.9 \pm 7.00$ |
| | PH$_{l1}$ | $0.69 \pm 0.03^{*}$ | $6.97 \pm 3.93$ | $21.13 \pm 12.06^{*}$ |
| | PH$_{pspline}$ | $0.67 \pm 0.03$ | $5.1 \pm 3.56$ | $14.96 \pm 8.80$ |
| | PLANNARD | $0.69 \pm 0.03^{**}$ | $8.07 \pm 3.86^{*}$ | $11.95 \pm 7.20^{*}$ |
| MLC | SSVM$_{linear}$ | $0.62 \pm 0.03$ | $2.56 \pm 1.64$ | $4.28 \pm 2.21$ |
| | SSVM$_{clinical}$ | $0.61 \pm 0.03$ | $2.51 \pm 2.00$ | $3.53 \pm 1.27^{**}$ |
| | SSVMP$_{linear}$ | $\mathbf{0.63 \pm 0.03}$ | $3.78 \pm 2.37$ | $5.75 \pm 2.66$ |
| | SSVMP$_{clinical}$ | $0.60 \pm 0.03^{**}$ | $1.64 \pm 1.03^{*}$ | $3.06 \pm 1.20^{***}$ |
| | PH$_{linear}$ | $0.61 \pm 0.03^{*}$ | $2.89 \pm 1.95$ | $\mathbf{6.04 \pm 2.67}$ |
| | PH$_{l2}$ | $0.62 \pm 0.03$ | $4.10 \pm 2.61$ | $3.97 \pm 1.74$ |
| | PH$_{l1}$ | $0.60 \pm 0.03^{***}$ | $1.46 \pm 1.32^{**}$ | $3.55 \pm 2.16$ |
| | PH$_{pspline}$ | $0.56 \pm 0.03^{***}$ | $1.06 \pm 0.92^{**}$ | $2.41 \pm 1.29^{***}$ |
| | PLANNARD | $\mathbf{0.63 \pm 0.02}$ | $\mathbf{4.46 \pm 2.20}$ | $3.81 \pm 1.50$ |
| PC | SSVM$_{linear}$ | $0.76 \pm 0.02$ | $11.27 \pm 3.34$ | $239.91 \pm 123.05$ |
| | SSVM$_{clinical}$ | $\mathbf{0.78 \pm 0.02}$ | $\mathbf{13.80 \pm 4.49}^{*}$ | $149.23 \pm 79.55^{*}$ |
| | SSVMP$_{linear}$ | $0.77 \pm 0.02$ | $10.48 \pm 3.22$ | $250.09 \pm 131.39$ |
| | SSVMP$_{clinical}$ | $\mathbf{0.78 \pm 0.02}$ | $13.61 \pm 3.91^{**}$ | $144.33 \pm 73.35^{***}$ |
| | PH$_{linear}$ | $0.77 \pm 0.02$ | $11.39 \pm 3.46$ | $\mathbf{266.26 \pm 150.48}$ |
| | PH$_{l2}$ | $0.76 \pm 0.02$ | $11.19 \pm 3.44$ | $228.04 \pm 107.23$ |
| | PH$_{l1}$ | $0.76 \pm 0.02$ | $10.49 \pm 3.27$ | $247.56 \pm 128.35$ |
| | PH$_{pspline}$ | $\mathbf{0.78 \pm 0.02}$ | $12.61 \pm 4.26$ | $207.16 \pm 88.44$ |
| | PLANNARD | $0.76 \pm 0.02$ | $10.87 \pm 3.80$ | $49.44 \pm 31.14^{***}$ |
| BC | SSVM$_{linear}$ | $0.67 \pm 0.02$ | $16.97 \pm 4.84$ | $79.38 \pm 49.57$ |
| | SSVM$_{clinical}$ | $\mathbf{0.68 \pm 0.02}$ | $20.93 \pm 4.64^{**}$ | $22.21 \pm 8.04^{***}$ |
| | SSVMP$_{linear}$ | $0.67 \pm 0.01$ | $17.64 \pm 2.61$ | $73.57 \pm 37.03$ |
| | SSVMP$_{clinical}$ | $\mathbf{0.68 \pm 0.02}$ | $\mathbf{24.14 \pm 7.56}^{**}$ | $20.27 \pm 6.95^{**}$ |
| | PH$_{linear}$ | $0.67 \pm 0.02$ | $18.11 \pm 4.83$ | $102.78 \pm 61.38$ |
| | PH$_{l2}$ | $\mathbf{0.68 \pm 0.02}$ | $20.16 \pm 4.2$ | $48.19 \pm 18.49^{**}$ |
| | PH$_{l1}$ | $0.67 \pm 0.02$ | $15.20 \pm 3.99$ | $\mathbf{193.77 \pm 139.29}^{**}$ |
| | PH$_{pspline}$ | $\mathbf{0.68 \pm 0.02}$ | $21.99 \pm 6.69^{**}$ | $158.69 \pm 107.79$ |
| | PLANNARD | $0.67 \pm 0.02$ | $15.00 \pm 5.11$ | $19.63 \pm 9.28^{***}$ |

Median and median absolute deviation of 50 random splits into train-test sets are given. Statistically significant differences between SSVMP (reference model: indicated in grey) and the other methods are indicated as: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. The best results are typeset in bold.

**Table 3.** Performances obtained on three micro-array datasets

| Dataset | Method | c-index | Logrank $\chi^2$ | Hazard ratio | # weights |
|---|---|---|---|---|---|
| nsbcd | SSVM$_{linear}$ | $0.60 \pm 0.08^{***}$ | $0.85 \pm 0.79^{***}$ | $4.35 \pm 3.70^{***}$ | |
| | SSVM$_{clinical}$ | $0.71 \pm 0.03^{***}$ | $4.46 \pm 2.31$ | $16.90 \pm 9.46^{***}$ | |
| | SSVMP$_{linear}$ | **$0.75 \pm 0.04$** | **$5.40 \pm 2.84$** | **$75.65 \pm 55.35$** | $171 \pm 16$ |
| | SSVMP$_{clinical}$ | $0.68 \pm 0.04^{***}$ | $3.72 \pm 2.10$ | $16.49 \pm 10.18^{***}$ | $288 \pm 183^{***}$ |
| | PCR | $0.72 \pm 0.03^{**}$ | $4.67 \pm 2.51$ | $24.23 \pm 14.84^{***}$ | $549 \pm 0^{***}$ |
| | SPCR | $0.71 \pm 0.03^{***}$ | $4.88 \pm 2.66$ | $13.55 \pm 7.46^{***}$ | $137 \pm 136^{**}$ |
| | PLS | $0.73 \pm 0.04^{*}$ | $4.00 \pm 2.09^{*}$ | $21.39 \pm 11.58^{***}$ | $549 \pm 0^{***}$ |
| | PH$_{l2}$ | $0.66 \pm 0.05^{***}$ | $1.91 \pm 1.45^{***}$ | $11.76 \pm 6.97^{***}$ | $549 \pm 0^{***}$ |
| | PH$_{l1}$ | $0.71 \pm 0.04^{***}$ | $4.73 \pm 1.89$ | $17.97 \pm 8.21^{***}$ | **$119.5 \pm 14^{***}$** |
| dlbcl | SSVM$_{linear}$ | $0.61 \pm 0.03^{***}$ | $4.14 \pm 2.41^{***}$ | $6.22 \pm 3.21^{***}$ | |
| | SSVM$_{clinical}$ | $0.63 \pm 0.03^{***}$ | $5.97 \pm 3.42^{***}$ | $7.45 \pm 3.15^{***}$ | |
| | SSVMP$_{linear}$ | **$0.76 \pm 0.02$** | **$22.47 \pm 5.72$** | **$224.78 \pm 143.10$** | **$2871 \pm 66$** |
| | SSVMP$_{clinical}$ | $0.62 \pm 0.02^{***}$ | $5.71 \pm 2.50^{***}$ | $7.53 \pm 3.02^{***}$ | $7027 \pm 266^{***}$ |
| | PCR | $0.60 \pm 0.02^{***}$ | $2.88 \pm 1.83^{***}$ | $4.66 \pm 1.90^{***}$ | $7399 \pm 0^{***}$ |
| | SPCR | $0.59 \pm 0.03^{***}$ | $2.08 \pm 1.84^{***}$ | $4.49 \pm 2.09^{***}$ | $7399 \pm 0^{***}$ |
| | PLS | $0.58 \pm 0.03^{***}$ | $1.65 \pm 1.50^{***}$ | $3.49 \pm 1.63^{***}$ | $7399 \pm 0^{***}$ |
| | PH$_{l2}$ | $0.64 \pm 0.03^{***}$ | $6.04 \pm 2.64^{***}$ | $8.71 \pm 4.27^{***}$ | $7399 \pm 0^{***}$ |
| | PH$_{l1}$ | $0.63 \pm 0.03^{***}$ | $5.41 \pm 2.85^{***}$ | $7.57 \pm 3.83^{***}$ | $3960.5 \pm 2477$ |
| dbcd | SSVM$_{linear}$ | $0.64 \pm 0.02^{***}$ | $3.49 \pm 1.48^{***}$ | $7.74 \pm 2.62^{***}$ | |
| | SSVM$_{clinical}$ | $0.75 \pm 0.03^{***}$ | $13.30 \pm 4.00^{***}$ | $33.61 \pm 15.88^{***}$ | |
| | SSVMP$_{linear}$ | **$0.82 \pm 0.02$** | **$18.66 \pm 4.97$** | **$360.47 \pm 233.84$** | $973 \pm 19$ |
| | SSVMP$_{clinical}$ | $0.75 \pm 0.03^{***}$ | $12.57 \pm 4.18^{***}$ | $33.78 \pm 17.00^{***}$ | $4232 \pm 271^{***}$ |
| | PCR | $0.73 \pm 0.02^{***}$ | $9.63 \pm 2.68^{***}$ | $28.10 \pm 16.96^{***}$ | $4919 \pm 0^{***}$ |
| | SPCR | $0.73 \pm 0.03^{***}$ | $10.31 \pm 3.02^{***}$ | $24.34 \pm 12.09^{***}$ | **$860 \pm 848$** |
| | PLS | $0.74 \pm 0.02^{***}$ | $11.23 \pm 3.97^{***}$ | $10.21 \pm 3.34^{***}$ | $4919 \pm 0^{***}$ |
| | PH$_{l2}$ | $0.71 \pm 0.03^{***}$ | $10.7 \pm 3.41^{***}$ | $23.06 \pm 12.38^{***}$ | $4919 \pm 0^{***}$ |
| | PH$_{l1}$ | $0.71 \pm 0.06^{***}$ | $11.14 \pm 6.51^{***}$ | $15.58 \pm 13.37^{***}$ | $1786 \pm 896$ |

The SSVM and SSVMP methods are compared with four PH models with different regularization mechanisms to deal with high-dimensional data. Median and median absolute deviations of the performances for 50 randomly splits in train-test sets are given. Statistically significant differences between SSVMP (reference model: indicated in grey) and the other methods are denoted as: $^{*}P < 0.05$; $^{**}p < 0.01$; $^{***}P < 0.001$. The best results are typeset in bold.

with an increasing risk of overfitting the training data. Thanks to regularization mechanisms, the number of effective parameters can be much lower than the number of weights. Nevertheless, for high-dimensional data, training remains difficult and/or time consuming, as illustrated in the Supplementary Material report. Finally, multi-layer perceptrons are known to be non-convex and the optimal solution will only be locally optimal.

The SSVM method does not assume a linear effect of the covariates, and has the additional computational advantage that data points do not need to be replicated. This is the result of reformulating the survival problem as a combined ranking-regression approach instead of as a time-dependent classification problem. The disadvantage of this approach is that estimation of the hazard is not directly incorporated in the model. The cumulative hazard can be estimated after categorizing patients into risk groups according to the estimated risk using the Nelson–Aalen estimator. The most important advantage of the SSVM method is the applicability and performance obtained on problems involving high-dimensional data. Additionally, since the estimation problem boils down to solving a convex QP standard efficient solvers can be used, and the estimate is a guaranteed global optimum.

The experiments give evidence that the SSVMP method outperforms standard techniques on high-dimensional micro-array data, while they perform similar to other methods on clinical data.

## ACKNOWLEDGEMENT

We kindly acknowledge the support and constructive remarks of E. Biganzoli and P. Boracchi.

*Funding*: Research Council KUL: GOA-AMBioRICS, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), several PhD/postdoc and fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0302.07 (SVM), research communities (ICCoS, ANMMM); IWT: TBM070706-IOTA3, PhD Grants; Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007–2011); EU: FAST (FP6-MC-RTN-035801), Neuromath (COST-BM0601). V.V.B. is supported by a grant from the IWT. K.P. is an associate professor ('Forskarassistent') at the University of Uppsala, Sweden. S.V.H. is a full professor and J.A.K.S. is a professor at the Katholieke Universiteit Leuven, Belgium.

*Conflict of Interest*: none declared.

## REFERENCES

Bair,E. *et al.* (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **101**, 119–137.

Bakker,B. *et al.* (2004) Improving Cox survival analysis with a neural-Bayesian approach. *Stat. Med.*, **23**, 2989–3012.

Biganzoli,E. *et al.* (1998) Feedforward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat. Med.*, **17**, 1169–1186.

Bøvelstad,H.M.M. *et al.* (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**, 2080–2087.

Breiman,L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.

Byar,D. and Green,S. (1980) Prognostic variables for survival in a randomized comparison of treatments for prostatic cancer. *Bull. Cancer*, **67**, 477–490.

Callas,P.W. *et al.* (1998) Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *Am. J. Indust. Med.*, **33**, 33–47.

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

Daemen,A. and De Moor,B. (2009) 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)*, IEEE, Piscataway, pp. 5913–5917.

Eilers,P.H. and Marx,B.D. (1996) Flexible smoothing with B-splines and penalties. *Stat. Sci.*, **11**, 89–121.

Emerson,S.S. and Banks,P.L.C. (1994) Interpretation of a leukemia trial stopped early. In Lange,N. *et al*. (eds) *Case Studies in Biometry*, Wiley-Interscience, New York, pp. 275–299.

Evers,L. and Messow,C.M. (2008) Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, **24**, 1632–1638.

Faraggi,D. and Simon,R. (1995) A neural network model for survival data. *Stat. Med.*, **14**, 73–82.

Goeman,J.J. (2010) L1 penalized estimation in the Cox proportional hazards model. *Biometr. J.*, **52**, 70–84.

Green,M.S. and Symons,M.J. (1983) A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J. Chronic Dis.*, **36**, 715–723.

Harrell,F. *et al.* (1988) Regression models in clinical studies: determining relationships between predictors and response. *J. Natl Cancer Inst.*, **80**, 1198–1202.

Hastie,T. *et al.* (2001). *The Elements of Statistical Learning*. Springer.

Huang,J.Z. *et al.* (2000) Functional ANOVA modeling for proportional hazards regression. *Ann. Stat.*, **28**, 961–999.

Kalbfleisch,J.D. and Prentice,R.L. (2002) *The Statistical Analysis of Failure Time Data*. Wiley series in probability and statistics, New York.

Leblanc,M. and Crowley,J. (1999) Adaptive regression splines in the Cox model. *Biometrics*, **55**, 204–213.

Lisboa,P. *et al.* (2003) A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Art. Intell. Med.*, **28**, 1–25.

Loprinzi,C.L. *et al.* (1994) Prospective evaluation of prognostic variables from patient-completed questionnaires: north central cancer treatment group. *J. Clin. Oncol.*, **12**, 601–607.

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Nygård,S. *et al.* (2008) Partial least squares Cox regression for genome-wide data. *Lifetime Data Anal.*, **14**, 179–195.

Park,P.J. *et al.* (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18** (Suppl. 1), S120–S127.

Prentice,R.L. (1974) A log gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539–544.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Sauerbrei,W. and Royston,P. (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J. R. Stat. Soc. Ser. A*, **162**, 71–94.

Schölkopf,B. and Smola,A. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge.

Schumacher,M. *et al.* (1994) Randomized $2 \times 2$ trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J. Clin. Oncol.*, **12**, 2086–2093.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

Shivaswamy,P.K. *et al.* (2007) A support vector approach to censored targets. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, CA, pp. 655–660.

Sørlie,T. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Suykens,J.A.K. (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore.

Therneau,T.M. and Grambsch,P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, 2nd edn. Springer, New York.

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 267–288.

Van Belle,V. *et al.* (2007) Support Vector Machines for Survival Analysis. In Ifeachor,E. and Anastasiou,A. (eds) *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED)*. University of Plymouth, Plymouth, UK.

Van Belle,V. *et al.* (2008) Survival SVM: a practical scalable algorithm. In Verleysen,M. (ed.) *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN)*. d-side, Evere, pp. 89–94.

Van Belle,V. *et al.* (2009a) Learning transformation models for ranking and survival analysis. *Technical report, 09-135, ESAT-SISTA, KULeuven (Leuven, Belgium) 2009*. Available at ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanbelle/reports/09-135.pdf (last accessed date November 19, 2010).

Van Belle,V. *et al.* (2009b) Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Technical report, 09-235, ESAT-SISTA, KULeuven (Leuven, Belgium) 2009*. Available at ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/vanbelle/reports/09-235.pdf (last accessed date November 19, 2010).

Van Belle,V. *et al.* (2010a) Additive survival least squares support vector machines. *Stat. Med.*, **29**, 296–308.

Van Belle,V. *et al.* (2010b) On the use of a clinical kernel in survival analysis. In Verleysen,M. (ed.) *Proceedings of the European Symposium on Artificial Neural Networks (ESANN2010)*, d-side, Evere, pp. 451–456.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van Houwelingen,H.C. *et al.* (2006) Cross-validated cox regression on microarray gene expression data. *Stat. Med.*, **25**, 3201–3216.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley and Sons, New York.