Additive survival least squares support vector machines

3rd October 2008

V. Van Belle, K. Pelckmans, J.A.K. Suykens and S. Van Huffel

K.U.Leuven, ESAT-SCD,
Kasteelpark Arenberg 10 bus 2446, B-3001 Leuven (Belgium),
vanya.vanbelle@esat.kuleuven.be

Abstract

This work studies a new survival modelling technique based on Least Squares Support Vector Machines (LS-SVM). We propose the use of a least squares support vector machine combining ranking and regression. The advantage of this kernel-based model is threefold: (i) the problem formulation is convex and can be solved conveniently by a linear system, (ii) non-linearity is introduced by using kernels, componentwise kernels in particular are useful to obtain interpretable results, (iii) introduction of ranking constraints makes is possible to handle censored data. In an experimental set-up, the model is used as a preprocessing step for the standard Cox proportional hazard regression by estimating the functional forms of the covariates, or for any linear survival model. The proposed model was compared to different survival models from the literature on the German Breast Cancer Study Group data.

1 Introduction

Assigning risk profiles to individual patients is of major relevance to decide upon the treatment of cancer patients. The acquisition of such a risk score requires a survival model, built on measurements depending on patient and tumor characteristics. Besides the risk score itself, the interpretation on how this score can be obtained is important to provide insight in the relevance of the different features. In breast cancer for instance, three variables (tumor size, number of lymph nodes and grade) are used to construct the Nottingham Prognostic Index (NPI) [1]. A high NPI indicates a higher risk to failure and is obtained for large, high grade tumors in patients with a higher number of positive lymph nodes.

Analysis of survival data is of interest within medical applications in general, and specifically in cancer studies [2, 3, 4, 5]. The most commonly used survival model is Cox' semi-parametric proportional hazard model [6]. Cox proposed to model the hazard (i.e. the instantaneous risk to experience the event at time t knowing that the event did not occur before time t) as a product of two terms. The first term is an arbitrary and unspecified baseline hazard function depending on time only. The dependence of the hazard function on the covariates is modelled through the second term. A recognized drawback of this model are the assumptions: (i) the proportional hazards assumption and (ii) linearity in the covariates.

Alternative models have been proposed in the literature. A first alternative concentrates on the linearity assumption. Instead of using the standard Cox model linear in the covariates, a general Cox model, allowing transformations of the covariates, is used. The difference in the proposed methods is found in the way the transformations are estimated. Hastie and Tibshirani [7] introduced generalized additive models in which the linear form is replaced by a sum of smooth functions. Splines were used in [3, 8, 9, 10, 11, 12]. Kernel Cox regression was proposed by [13], while functional ANOVA modelling was used in [14]. Neural networks were used to extend the Cox model towards non-linear covariate effects by [15, 16].

A second alternative uses flexible models without the assumption of proportional hazards. Biganzoli et al. [17] proposed a Partial Logistic Artificial Neural Network (PLANN) for the analysis of survival data. The time axis is divided in predefined intervals and the time is inserted as an extra covariate. Each sample is therefore replicated for each interval in which the sample is at risk. An artificial neural network is used to model a piecewise linear hazard rate. This approach is useful to model non-linear covariate effects, and is not restricted to proportional hazards.

The application of Support Vector Machines (SVMs) within survival analysis was previously presented in [18, 19], which is based on ranking SVMs [20]. We advance this work and according to the methodology of LS-SVMs [21, 22], we investigate formulations using regression and ranking constraints which can be solved by a linear system of moderate size. LS-SVMs are reformulations to the standard support vector machines [23] for classification and regression problems, where the solution is found by solving linear Karush-Kuhn-Tucker systems. Nonlinearity is introduced by using a feature map in the primal representation. Due to Mercer's theorem [23] it is not necessary to specify the feature map explicitly, instead one makes use of a positive definite kernel at the dual level. Incorporation of prior knowledge is realized by adding extra constraints.

The contribution of this work is the reformulation of the methodology of LS-SVMs towards censored data and survival analysis. Within the model specification two purposes are recognized: (i) ranking new instances according to their survival time and (ii) predicting their survival time. Within the first objective, censored and uncensored data are used, whereas the second only takes uncensored samples into account. Interpretability is obtained by using componentwise kernels [24, 25].

This paper is organized as follows. Section 2 discusses survival analysis in general and illustrates Cox' proportional hazard model [6], an artificial neural network approach (PLANN) as well as a previously reported support vector machine (SURSVM) based survival techniques [18, 19]. Section 3 indicates how LS-SVMs can be used to analyze survival data. In Section 4 we discuss results on artificial and real life data. Conclusions are given in section 5.

The following notations are used throughout the paper. A lower case b will denote a vector or a constant (depending on the context), an upper case B will denote a matrix. The covariates $x_i \in \mathbb{R}^d$ of dimension $d \in \mathbb{N}$ of each subject $i = 1, \ldots, n$ in the dataset can be organized in a matrix $X \in \mathbb{R}^{n \times d}$, $X = [x_1 \ldots x_n]^T$. The failure times $t_i \geq 0$ are arranged in a column vector $t \in \mathbb{R}^n$.

2 Survival models

2.1 Survival data and censoring mechanisms

Failure time data is obtained by observing subjects from a certain initial event to either the occurrence of a predefined event or the end of the study for this subject. The latter case results in incomplete or censored data. Common types of censoring include (Figure 1) [26]:

Right censoring occurs when the event of interest did not occur until the last follow-up time. This type of censoring typically occurs at the end of the study period. Although the exact failure time is not known in this case, the failure time is known to be later than the date of the last follow-up. Random censorship deals with the cases where the censoring times are independent of the failure times and the censoring times can have an arbitrary distribution. Examples of random censoring are patients lost to follow-up, drop out in which case patients decide to stop joining the study. Another example of right censoring is type I censoring. In this case the event is only observed when it occurs before a certain time. An example is found in studies where every object is followed for a predefined time. Examples of type I censoring are commonly found in industrial testing where n identical items are tested simultaneously and observed on a fixed interval. Type II censoring is the type of censoring occurring when the study is stopped after the observation of a predefined number of events. All objects without event at that time are censored. This type of censorship is also most common in industrial life testing.

We denote a sample $i = 1, \ldots, n$ with covariates x_i as (x_i, t_i, δ_i). The observed failure t_i can be expressed as $t_i = \min(c_i, t_i')$, with t_i' the true failure time, c_i the censoring time and $\delta_i = I(t_i \leq c_i)$.

Left censoring deals with the case when the failure is known to have happened before a certain time. An example of left censoring arises in case a variable can only be measured when its value is above a certain level. A typical

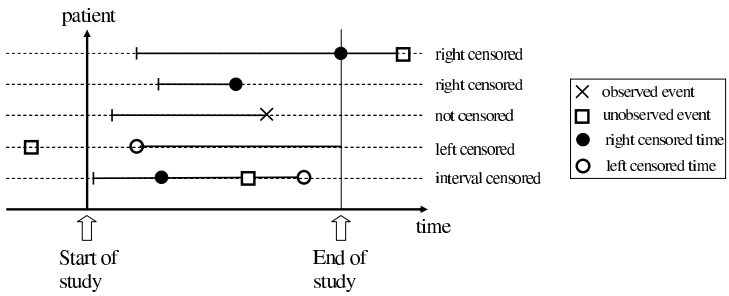


Figure 1: Types of censoring. In case the event occurs later than the last follow-up time, the observed failure time is called right censored: the event time is larger than the last follow-up time. In case the event is observed within the study period, the survival time is uncensored. A left censored survival time occurs in case the event already took place before the patient was followed: the survival time is less than the time at which the patient entered the study. Interval censoring is a combination of right and left censoring. In this case we know that the event takes place within a certain interval $[t_1, t_2]$. Because the survival time is later than t_1, the latter time is right censored. Because the survival time is prior to t_2, t_2 is left censored.

example of left censoring used in literature is the descent of the baboons as described in [27]. Baboons in the Amboseli Reserve, Kenya, sleep in the trees and descend for foraging at some time of the day, defined as the time when half of the members of the group has descended. Those days that observers arrive later than this time of descent they can only ascertain that descent took place before the time the observers arrived. At these days the descent times are left censored.

Interval censoring is a combination of the previous two censoring types. In this case the failure time is not known exactly, instead an interval including the failure time is indicated. This type of censoring is often found in medical studies where the patients are subject to regular check up times [28]. In cancer studies for example, patients are asked to come for check-ups, after they are tumor free (thanks to an operation), every three months. If a new tumor is found at the third check-up, the time of cancer recurrence lies between the second and the third check-up time. Another example is found in AIDS studies [29]. When an individual is infected by the HIV virus at the beginning of the study, the infection time is determined using a retrospective study of the person's history. Therefore, the infection time is known to lie between the last negative and the first positive HIV test.

For notational convenience we consider only the case of right censoring, but the described methodology can be applied straightforwardly to other cases as well.

Let z be a random variable. One possible way to present survival data is to calculate the survival (distribution) function (SDF) $S(t)$ [26]:

$$S(t) = 1 - F(t) = P(z > t)\,, \tag{1}$$

with $F(t)$ the cumulative distribution function and assume this function possesses first order derivatives. The hazard function expresses the instantaneous failure rate at time t :

$$\lambda(t) = \frac{f(t)}{S(t)}\,, \tag{2}$$

with $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$ representing the probability density function.

2.2 Cox' proportional hazard model

Let $\lambda(t, x)$ represent the hazard function at time t for a subject with covariates x. The proportional hazard model [6] models the hazard function at time t as a time-depending arbitrary unspecified baseline hazard function λ_0 multiplied with an exponential function, depending linearly on the covariates x:

$$\lambda(t, x) = \lambda_0(t) \exp(\beta^T x)\,. \tag{3}$$

To estimate the coefficients vector β the partial likelihood, as introduced by Cox [6] is maximized.

2.3 Partial Logistic Artificial Neural Networks

Biganzoli *et al.* [17] proposed an artificial neural network (ANN) for the analysis of survival data. Within this partial logistic ANN (PLANN) the hazard function is estimated as a piecewise linear function. The time axis is divided into equally long time intervals. Within every time interval the hazard rate is defined as a conditional failure probability. This approximation to the hazard function gets better when the time intervals become smaller. Considering only right censored values, the likelihood can be expressed as

$$L(w, b) = \prod_{i=1}^{n} \prod_{l=1}^{l_i} \lambda_{w,b}(x_i, t_l)^{\delta_i(t_l)} (1 - \lambda_{w,b}(x_i, t_l))^{\delta_i(t_l)}\,, \tag{4}$$

with $i = 1, \ldots, n$ a sample number, $l = 1, \ldots, l_i$ the time intervals in which subject i is at risk and $\delta_i(t_l)$ an event indicator equal to 1 if patient i had an event in the l-th time interval and 0 otherwise. PLANN uses one input for each covariate and an additional input for the time interval. The multilayer perceptron structure is

$$\frac{\lambda_{w,b}(x, t_k)}{1 - \lambda_{w,b}(x, t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h g\left(\sum_{i=1}^{N_i} w_{ih} x_i + w t_k + b_h \right) + b \right), \tag{5}$$

where the indices i and h denote the input and hidden node layers respectively and $g(.)$ is the hidden layer activation function. We use the maximum likelihood $\max_{w,b} L(w,b)$ to estimate w and b. The output of the output neuron can be used to explore the shape of the hazard function depending on time and covariates.

Lisboa *et al.* [4] added a Bayesian framework for doing automatic relevance determination to PLANN (PLANN-ARD) for weighting each covariate for its relevance.

2.4 Survival svm

The methodology using support vector machines (SVM) for the analysis of survival data, starts from a different perspective [18, 19]. Instead of modelling the hazard function, the concordance between the observed failure times and the model's output is optimized. The model's output is a prognostic function (also called utility function and more specific within medical studies prognostic index or health function) $u : \mathbb{R}^d \to \mathbb{R}$ defined as

$$u(x) = w^T \varphi(x)\,. \tag{6}$$

The concordance index (c-index) as introduced by Harrell [30], measures the concordance between the ranking function and the observed failure times using censored as well as uncensored observations.

Definition 1 (Concordance Index (c-index)) *Let $u : \mathbb{R}^d \to \mathbb{R}$ be the predicted utility of an instance with covariates x. Assuming that there are no ties is the data, the c-index measures the concordance between the random variables $\{u(x_i)\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$*

$$c(u) = E(I((u(x_i) - u(x_j))(t_i - t_j) > 0)) \tag{7}$$

where $I(z)$ is the indicator function equal to 1 if $z = 1$ and 0 otherwise. Given the set of observations denoted as $\mathcal{S} = \{(x_i, t_i, \delta_i)\}_{i=1}^n$, where δ indicates whether the failure time is observed exactly ($\delta = 1$) or right censored ($\delta = 0$), the empirical c-index between the survival times t and the predictions $u(x)$ is defined as:

$$c_n(u) = \frac{\sum_{i,j=1}^n v_{ij} I\Big(\big(u(x_j) - u(x_i)\big)(t_j - t_i) > 0\Big)}{\sum_{i,j=1}^n v_{ij}} \tag{8}$$

where the v_{ij} values indicate whether the samples (x_i, t_i, δ_i) and (x_j, t_j, δ_j) are comparable. The v_{ij} values are defined as

$$v_{ij} = \begin{cases} 1 & t_i < t_j,\ \delta_i = 1 \\ 0 & t_i < t_j,\ \delta_i = 0 \end{cases} \tag{9}$$

A pair is comparable when the ranking in the time domain is known. Examples of comparable datapoints are (i) two uncensored samples and (ii) an uncensored

and a right censored sample, for which the latter is censored later than the event of the first is observed. Note that the v_{ij} values can be generalized to other censoring schemes.

The survival SVM (SURSVM) is designed to optimize the c-index directly as described in [18]. A computational simplification was presented in [19]. Knowing that t_i is less than t_j, we want $u(x_i)$ to be less than $u(x_j)$. The SURSVM model goes further by implying that this constraint should be met with a certain margin, resulting in the constraints

$$u(x_j) - u(x_i) \geq 1, \ \forall \ i < j \, . \tag{10}$$

Slack variables $\xi_{ij} \geq 0$ can be introduced to allow for misranking. In order to work with right censored failure times, a parameter v_{ij} is added as in (9). The model is formulated as follows:

$$\begin{aligned} \min_{w,\xi} \quad & \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \textstyle\sum_{i<j} v_{ij} \xi_{ij} \\ \text{s.t.} \quad & \begin{cases} w^T \varphi(x_j) - w^T \varphi(x_i) \geq 1 - \xi_{ij}, \ \forall \ i < j \\ \xi_{ij} \geq 0, \ \forall \ i < j \, , \end{cases} \end{aligned} \tag{11}$$

with $\gamma \geq 0$ a regularization constant.

The solution to (11) is given by

$$\begin{aligned} \min_{\alpha} \quad & \tfrac{1}{2} \alpha^T D K D^T \alpha - 1_n^T \alpha \\ \text{s.t.} \quad & 0 \leq \alpha \leq \gamma v \, , \end{aligned} \tag{12}$$

where $\alpha = [\alpha_1 \ ... \ \alpha_{n_c}]^T$, $v = [v_1 \ \dots \ v_{n_c}]^T$ and $K_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ denote the elements of the kernel matrix and n_c denotes the number of comparable pairs in the training set. $K(x, z)$ denotes a positive definite kernel function evaluated at points x and z. $D \in \{-1, 0, 1\}^{(n_c \times n)}$ is defined such that

$$DX = \begin{bmatrix} x_1 - x_2 \\ \vdots \\ x_1 - x_n \\ \vdots \\ x_{n-1} - x_n \end{bmatrix} , \tag{13}$$

with $X = [x_1 \ \dots \ x_n]^T \in \mathbb{R}^{n \times d}$. 1_n represents a column vector of size n with 1 at every position. The value of the health function for a new point x^* is given by

$$u(x^*) = \hat{\alpha}^T D K_n(x^*) \, , \tag{14}$$

with $K_n(x^*) \in \mathbb{R}^n$ and $K_n(x^*) = [K(x^*, x_1) \dots \ K(x^*, x_n)]^T$. Censoring is handled by the choice of the v value. Censoring reduces the computational load as elaborated in [18].

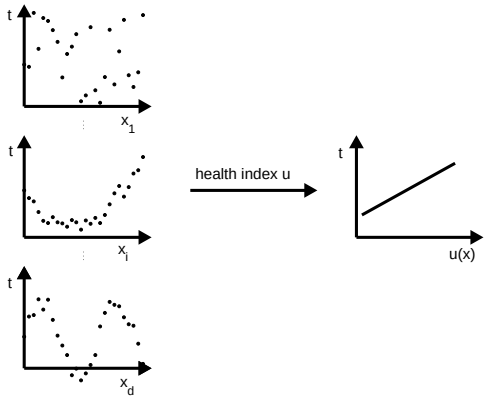


Figure 2: Goal: A model $u(x)$ is trained starting from measurements of covariates x, which influence the failure time t in an unknown manner. A good model is able to predict the ranking of failure times. Therefore, the interval in which the function $u(x)$ increases monotonically as a function of the failure times t should be as wide as possible.

3 A Survival Least Squares Support Vector Machine

This section studies the use of least squares support vector machines [21, 22] for kernel-based modelling of failure time data. A model with health index $u(x)$ is trained starting from measurements of covariates x, which influence the failure time t in an unknown manner. The goal of the model is to order new samples correctly with respect to their survival time. Therefore, the interval in which the function $u(x)$ increases monotonically as a function of the failure times t should be as wide as possible (Figure 2). As in [18, 19] we base our method on maximizing the concordance index (see Definition 1).

3.1 Rank regression Least Squares Support Vector Machine

To obtain an LS-SVM formulation for the survival problem, equation (11) is modified in a similar way as in the case of regression SVMs to obtain regression LS-SVMs. Instead of the inequality constraints, equality constraints are taken where the value 1 at the right hand side is considered as a target value rather than a threshold. The error variable, similar to the slack variable in the SVM formulations, allows for misrankings. In the loss function the L_2-norm of the slack variables is included. Here we study a ranking function $u(x) : \mathbb{R}^d \to \mathbb{R}$

where $u(x) = w^T\varphi(x)$ which is trained as follows:

$$\begin{aligned} \min_{w,\xi} \quad & \tfrac{1}{2}w^T w + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 \\ \text{s.t.} \quad & w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij},\ \forall\ i,j = \ 1,\ldots,\ n\,, \end{aligned} \tag{15}$$

where $\gamma > 0$ is a fixed regularization constant trading model complexity and reconstruction error.

Within the training phase the samples can be sorted such that $t_i < t_j$ for $i < j$. Perfect ranking is then obtained if $w^T\varphi(x_j) - w^T\varphi(x_i) > 0$, $\forall\ i,j = 1,\ldots,n$. After constructing the Lagrangian and differentiating towards primal and dual variables, the solution is obtained by a linear set of equations.

Proposition 1 (Model 1: surlssvm) *The solution to (15) is given by*

$$\left[\gamma DKD^T + I_{n^2}\right]\alpha = \gamma 1_{n^2}\,, \tag{16}$$

with D as defined in (13), 1_{n^2} a vector of size n^2 containing ones, I_{n^2} the identity matrix with dimension equal to the number of order comparisons n^2 and $K \in \mathbb{R}^{n\times n}$ the kernel matrix with $K_{ij} = \varphi(x_i)^T\varphi(x_j)$ for all $i,j = 1,\ldots,n$. The prediction $u(x^*)$ for a new point x^* can be calculated as

$$\hat{u}(x^*) = \hat{\alpha}^T D K_n(x^*)\,, \tag{17}$$

with $K_n(x^*) \in \mathbb{R}^n$ and $K_n(x^*) = [K(x^*,x_1)\ \ldots\ K(x^*,x_n)]^T$ and $\hat{\alpha}$ the solution to (16).

Proof See appendix A.

The proposed learning scheme generates a model which predicts the ranking of new instances. In medical applications for instance ranking is important in order to predict which patients will suffer from a disease first (having a high risk) and which patients will develop the disease later or not (having a low risk). However, when considering the importance of different variables, one is not only interested in the ranking, but the questions rises how these variables influence the failure time. This question is important in understanding the underlying mechanisms of the disease. In order to answer this question we consider the following criterion based on the mse instead of the c-index.

Definition 2 (Mean Squared Error for censored data) *The standard definition of the mean squared error between the observations t_i and the corresponding prediction $\hat{t}_i$*

$$\text{mse}_{\text{standard}} = \frac{\sum_{i=1}^{n}(\hat{t}_i - t_i)^2}{n}\,, \tag{18}$$

where n is the number of datapoints in the set over which the measure is calculated, is useful for uncensored observations. In order to include information

contained in right censored observations, the mean squared error is adapted as:

$$\text{mse} = \frac{\sum_i q_i(\hat{t}_i - t_i)^2}{n}\,, \tag{19}$$

with q_i defined as

$$q_i = \delta_i + (1-\delta_i) I(\hat{t}_i - t_i < 0)\,, \tag{20}$$

*where I denotes the indicator function. The weight q_i is 1 for uncensored samples. For right censored samples, the weight is 0 in case the prediction is later than the observed failure time and 1 otherwise. Note that this measure is closely related to the censored Least Absolute Deviation (*CLAD*), as used in [31].*

Model (15) is then modified as follows:

$$\begin{array}{ll} \min\limits_{w,\xi,b,\chi} & \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} v_{ij}\xi_{ij}^2 + \dfrac{1}{2}\mu \sum\limits_{j=1}^{n}\chi_j^2 \\ \text{s.t.} & \begin{cases} w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij}, & \forall\ i,j = 1,\ldots,n \\ \delta_j t_j = \delta_j(w^T\varphi(x_j) + b) - \chi_j, & \forall\ j = 1,\ldots,n\,, \end{cases} \end{array} \tag{21}$$

where $\gamma \geq 0$ and $\mu \geq 0$ are fixed regularization parameters. The solution to this problem statement is found by solving the linear system in n^2+n+1 unknowns. The dimension of the linear system can be reduced as elaborated in [19].

Proposition 2 (Model 2: surlssvm2) *The solution to (21) is given by*

$$\begin{bmatrix} DKD^T + I_{n_c}/\gamma & DKR & 0 \\ RKD^T & RKR + I_n/\mu & \delta \\ 0^T & \delta^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ Rt \\ 0 \end{bmatrix}, \tag{22}$$

where α and β are the Lagrange multipliers related to the constraints, $t = [t_1\ \ldots\ t_n]^T$, $\delta = [\delta_1\ \ldots \delta_n]^T$ and $R \in \mathbb{R}^{n\times n}$ is the diagonal matrix $R = \text{diag}(\delta)$. The health function $u(x)$ can be evaluated at a new point x^* as

$$u(x^*) = (\hat{\alpha}^T D + \hat{\beta}^T) K_n(x^*) + \hat{b}\,, \tag{23}$$

where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{b}$ are the solution to (22).

Proof See appendix B.

Remark 1 *In case μ is taken to be 0 the model reduces to the* SURLSSVM *model as in equation (15).*

Remark 2 *In case γ is taken to be 0 and without censoring, the model reduces to the standard regression* LS-SVM *model.*

3.2 The proportional hazard assumption

As stated earlier, Cox' proportional hazard model is the most commonly used model in survival analysis. Since this model assumes that the hazard can be written as the product of a time dependent and a covariate dependent term, the hazard rate

$$\frac{\lambda(t,x_1)}{\lambda(t,x_2)} = \frac{\lambda_0(t)\exp(\beta^T x_1)}{\lambda_0(t)\exp(\beta^T x_2)} = \frac{\exp(\beta^T x_1)}{\exp(\beta^T x_2)} \tag{24}$$

is constant over time, hence the name proportional hazard model.

An advantage of the model proposed in this work is that it does not assume proportional hazards. The following reasoning illustrates this statement. The use of the least squares criterion in equation (21) suggests a normally distributed error term ϵ in the equation $t = u(x)+\epsilon$. The probability function of the normal distribution for an individual with covariates x is defined as

$$f(t,x) = \frac{1}{\sigma(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(\frac{t-\mu(x)}{\sigma(x)}\Big)^2\right), \tag{25}$$

where the mean and the standard deviation depend on the covariates x. The corresponding survival function is

$$S(t,x) = \int_t^\infty \frac{1}{\sigma(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(\frac{u-\mu(x)}{\sigma(x)}\Big)^2\right)du\,. \tag{26}$$

and the hazard is defined as

$$\begin{aligned}\lambda(t|x) &= \frac{\frac{1}{\sigma(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(\frac{t-\mu(x)}{\sigma(x)}\Big)^2\right)}{\int_t^\infty \frac{1}{\sigma(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(\frac{u-\mu(x)}{\sigma(x)}\Big)^2\right)du} \\ &= \frac{\exp\left(-\frac{1}{2}\Big(\frac{t-\mu(x)}{\sigma(x)}\Big)^2\right)}{\int_t^\infty \exp\left(-\frac{1}{2}\Big(\frac{u-\mu(x)}{\sigma(x)}\Big)^2\right)du}\,,\end{aligned} \tag{27}$$

which depends in general on t. The following example illustrates the possibility of non-proportional hazards.

Example 1 (Non-proportional hazards within the surlssvm model) *Let $\mu(x) = x$ and $\sigma(x) = 2$. Taking $x_1 = 4.5$ and $x_2 = 5.5$, the probability distribution functions are shifted versions of the normal distribution with zero mean and standard deviation equal to 2 (Figure 3, top left). The top right and bottom left figures in Figure 3 illustrate the corresponding survival and hazard functions.*

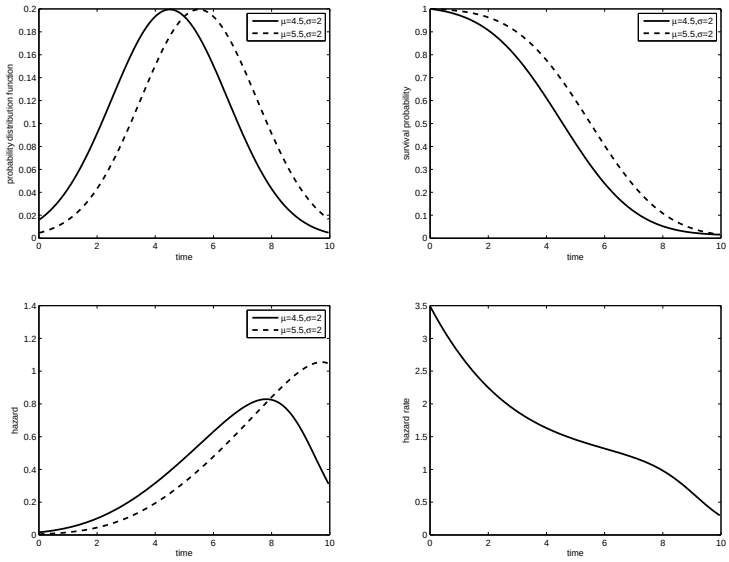


Figure 3: Illustration of the possibility of non-proportional hazards in the SURLSSVM model on an artificial example. Let $\mu(x) = x$ and $\sigma(x) = 2$. Taking $x_1 = 4.5$ and $x_2 = 5.5$, the probability distribution functions are shifted versions of the normal distribution with zero mean and standard deviation equal to 2, as shown in the top left figure. The top right and the bottom left figure show the survival and hazard function respectively. The hazard rate (bottom right figure) is non-constant over time, indicating non-proportional hazards.

As shown in the bottom right figure the hazard rate is non-constant over time, indicating non-proportional hazards.

3.3 Interpretability and Additive Models

Within medical applications, the predictive performance is not the only specification that a good model requires. The model's interpretability is important for understanding the underlying mechanisms of the disease. To increase the interpretability of black box models, the use of componentwise kernels has been proposed [25]:

$$K(x_i, x_j) = \sum_{p=1}^{P} K^{(p)}(x_i^{(p)}, x_j^{(p)}), \quad \forall\ x_i,\ x_j\,, \tag{28}$$

where $x^{(p)}$ represents the pth covariate of datapoint x and $K^{(p)}(x_i^{(p)}, x_j^{(p)})$ denotes a positive definite kernel function evaluated at the pth covariate of the points x_i and x_j.

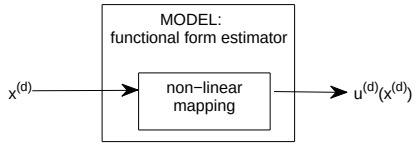


Figure 4: Estimation of the functional form. The SURLSSVM2 model can be used to estimate the functional form of a single covariate. The value of the dth covariate of all training samples is used as input and the output is the contribution $u^{(d)}$ of this covariate $x^{(d)}$ to the health function $u(x)$.

Dependence of the model's outcome on one specific covariate $x^{*(d)}$ of a sample with covariates x^* can be visualized by looking at the contribution $u^{(d)}$ of the variable $x^{(d)}$ on the health function $u(x)$, with $u^{(d)}(x^{*(d)}) = (\hat{\alpha}^T D + \hat{\beta}^T)K_n^{(d)}(x^{*(d)}) + \hat{b}$ in case of the SURLSSVM2 model (Figure 4). The prognostic index is constructed as a weighted sum of the componentwise contributions as

$$u(x) = \sum_{p=1}^{P} r_p u^{(p)}(x^{(p)}) \,, \tag{29}$$

with weights r_p determined by a second model which is linear (Figure 5).

3.4 Comparison of surlssvm with other models

Cox' model is a semiparametric model which assumes the existence of a baseline hazard. Covariates are only able to change this hazard in a proportional way. The effect of the covariates is modelled parametrically. The SURLSSVM2 model on the contrary, is a non-parametrical method estimating the ranking of the survival times. Advantages of the Cox model are its extendibility towards time-dependent covariates and competing risk [26]. It is not yet clear how these extensions can be made in the SURLSSVM2 model. Advantages of the latter model include interpretability and the possibility to handle non-proportional hazards. The solution also follows from solving a linear system and the model formulation is convex.

The PLANN model proposed by Biganzoli *et al.* [17] has the drawback that samples need to be replicated at all time intervals where they are at risk. This is not necessary for the Cox and SURLSSVM2 models. As in the case of the SURLSSVM2 model, there is no assumption of proportional hazards. Effects of covariates can be interpreted on the estimated hazard. An advantage of the method is its straightforward extension towards time-dependent covariates.

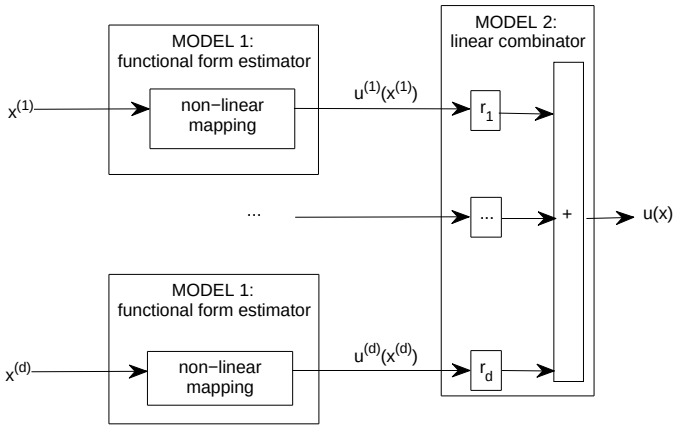


Figure 5: Construction of the prognostic index. The SURLSSVM2 model is used to estimate the functional form of each covariate $x^{(p)}$ with $p = 1, \ldots, d$. These contributions are combined in a weighted sum. The weights r_p are determined by a second model which is linear.

Current research demands for techniques capable of dealing with high dimensional data, as for example in the case of computational genetics. The Cox proportional hazard model as well as the PLANN model generate less accurate predictions of the coefficients for increasing dimensions. It is well known that kernel-based methods as SVM and LS-SVM are well suited for handling high dimensional data [22, 23].

4 Empirical Results

This section illustrates results on artificial and real life datasets. Tuning of the regularization and kernel parameter was performed as follows. The available data was divided into training and test set (2/3 versus 1/3) and stratified for the censoring variable δ. Three different kernels were used in the examples:

- linear kernel: $K(x, x^*) = x^T x^*$
- polynomial kernel of degree d: $K(x, x^*) = (\tau + x^T x^*)^d$, where $\tau \geq 0$ is a tuning parameter. d was chosen to be $d = 2$.
- RBF kernel: $K(x, x^*) = \exp(-||x - x^*||_2^2/\sigma^2)$, where σ is a tuning parameter.

Depending on the model and the kernel, parameters need to be tuned. The models SURSVM and SURLSSVM have one regularization parameter γ, whereas the SURLSSVM2 model has two tuning parameters γ and μ. A linear kernel needs no further tuning, but for the polynomial and the RBF kernel the value of the parameter τ and σ respectively needs to be optimized. Optimal values of all these parameters were found after 10-fold cross-validation on the training set. The search for the tuning parameter was done over a grid of 10 by 10 values, whereafter a fine-tuning step was performed. Tuning parameters resulting in the best cross-validation performance were selected. As a performance measure we used the c-index (see Definition 1) and a mean squared error adapted for censored data (see Definition 2).

4.1 The impact of the amount of censoring

In a first example a dataset on lung capacity [32] is used to illustrate the impact of the amount of censoring on test performance. The data contains information on 654 persons, aged 3-19 years. The task of this study is to predict the lung capacity, given the age and the height of each person. In this example we only considered non-smoking male persons. Two third of the data are used for training, the other third is used as a test set. In order to illustrate how the percentage of censored data involves the performance of the different methods, the values of the lung capacity are censored to different amounts. Figure 6 illustrates the effect for non-additive and additive models. The models SURSVM and SURLSSVM were tuned using the c-index. Both the c-index and the mse were used to tune the SURLSSVM2 model. A polynomial kernel was used for all kernel based models. All models perform well at high censoring performances. No difference in performance was noted between Cox' model and the additive models, the non-additive models based on (LS-)SVMs seem to be less reliable.

For additive models the impact of the censoring percentage on the test set performance is dependent on the model selection criterion (Figure 7). This figure illustrates the performance on 1000 bootstraps of the test sets after training the functional forms by means of the SURLSSVM2 model with a RBF kernel. Both the c-index and the mse were used as model selection criteria. The functional forms were then combined with two linear models: Cox' model and the SURLSSVM2 model with a linear kernel. Without censoring there was no evidence for a performance difference between models tuned with both model selection criteria. When increasing the amount of censoring, the difference in performance increases, models tuned with the modified mse performing better. An explanation for the drop in performance for higher censoring percentages when using the concordance index as a model selection criterion is found in the estimation of the functional form. Figure 8 shows the change in estimated functional form for different censoring percentages. When using the c-index, the functional form deviates more from the estimated functional form without censoring for increasing censoring percentages. When using the mse on the contrary, the estimated functional forms are less sensitive to the amount of censoring, leading to a more stable test set performance.

4.2 Importance of taking censored observations into account in the training step

Censored observations are more rule than exception in survival analysis. An important contribution of formulating models for the analysis of failure time data consists of adapting models to make it possible to incorporate censored observations. We illustrate the effect of taking censored samples into account in the training step on the same data as in the previous example. Figure 9 confirms that censored observations do hold extra information. We compared the performance of models which were trained by only incorporating uncensored observations with models which did include censored observations, for different censoring percentages. With 30% of censored observation, the observed performance was slightly better when including the censored observation in the training step. At higher censoring percentages, the improvement was significant.

4.3 High dimensional data

The increasing demand for handling high dimensional data asks for methods able to handle a small amount of training samples in a high dimensional space. In this example, an artificial dataset is created with a Weibull distributed failure time and normal distributed covariates. The distribution of the failure time depends on the covariates as:

$$f(t;a,b) = \begin{cases} \frac{a}{b}(\frac{t}{b})^{a-1}\exp(-(\frac{t}{b})^a) & t \leq 0 \\ 0 & t < 0\,, \end{cases} \tag{30}$$

were $a = 10^{\frac{\sum_{i=1}^{d} x^{(i)}}{d}}$ and $b = 10$. We created 90 training and test sets, with 30 percent of censoring, 100 covariates and a varying number of samples. Figure 10 shows the mean concordance index on the test sets after training on the training set. The standard Cox model is compared with the non-additive SURLSSVM2 model using a linear kernel. Table 1 tabulates p-values according to Wilcoxon's signed rank sum test for differences in performance. For datasets for which the number of samples is considerably larger than the dimension d, both models perform equally well. However, for smaller datasets with a high dimension the SURLSSVM2 model performs significantly better.

4.4 A breast cancer study

The last example shows the results of the proposed additive model compared with the SURLSSVM, SURSVM and Cox model on the German Breast Cancer Study Group data [33]. This dataset contains data on 686 women with breast cancer and can be found at http://www.blackwellpublishers.com/rss/Volumes/A162p1.htm. Available variables are: hormone treatment, age, menopausal status, tumor size, tumor grade, the number of lymph nodes involved, the progesterone receptor (fmol) and the estrogen receptor (fmol). Two third of the

Table 1: Significance of the difference in performance between Cox' model and the SURLSSVM2 model with a linear kernel on high dimensional, Weibull distributed data. The new method performs significantly better than Cox for datasets for which the number of samples is not considerably larger than the dimension. For datasets with considerably more samples than covariates, the performance of both models is comparable.

n	p-value signed rank sum test
100	<0.0001
120	<0.0001
150	<0.0001
200	<0.0001
300	0.0064
400	0.9599
500	0.5080

data were used for training, the rest for testing the models. 299 (43.6%) patients had a breast cancer related event within the study time, leaving all other patients with a right censored failure time.

Figure 11 summarizes results; a polynomial kernel with degree $d = 2$ was used for all kernel based models. Model selection was based on 10-fold cross-validation with the modified mse or the c-index as performance measure. Afterwards the model was trained on the whole training set. We used the mean squared error adapted for censored data in case of the SURLSSVM2 model and the c-index for the other (LS-)SVM based models. We compared results with the PLANN-ARD model. An L_1-regularized version of the Cox model [34] (Cox-$L1$) selected the number of positive lymph nodes, the progesterone receptor and grade as relevant features after 10-fold cross-validation maximizing the partial likelihood. The results obtained with models are compared with the result obtained with the Nottingham Prognostic Index (NPI) [1]. This index is used in clinical practise to classify breast cancer patients into risk groups according to relapse. Using non-additive methods, the SURLSSVM2 and the PLANN-ARD model perform the best. When using additive models, the performance is increased and interpretability is gained.

The estimated functional forms are shown in Figure 12. Functional forms were investigated using a linear, polynomial of the second degree and RBF kernel. On this dataset, the results were better with the polynomial kernel than with the linear and RBF kernels. We prefer to use the polynomial kernel because the estimated effects are not restricted to be linear (as with the linear kernel) and are not too flexible (as with the RBF kernel). The linear kernel suggests a lower survival for younger patients. The polynomial kernel also estimates a higher risk for younger patients but additionally indicates an increased risk for

older patients. The clinical literature already described this observation [35]. Concerning tumor size, the number of lymph nodes, progesterone and estrogen receptors, the estimated functional forms agree with what is known from clinical practise [36, 37, 38, 39].

To be able to compare the results of PLANN-ARD with the SURLSSVM2 model, the prognostic indices of the first model are averaged over time. Both the Cox-L_1 and the SURLSSVM2 model predict an increasing risk for breast cancer relapse in older patients, as shown in literature [35]. The standard Cox model is not able to discover this effect, as is the case for the SURLSSVM2 model with a linear kernel. We expect a decreasing survival time for larger tumors, which is confirmed by the SURLSSVM2 model and Cox, but the PLANN-ARD model does not show a clear effect of age. The predictions for the number of positive lymph nodes, progesterone and estrogen receptors are concordant for SURLSSVM2, PLANN-ARD and Cox. The first model shows a drop in the expected survival time for very high levels of progesterone and estrogen receptors, but since the number of datapoints in the training set at these levels, the prediction is not reliable in these regions.

Figure 13 shows the distribution of the difference in c-index on 1000 bootstraps of the test set between the method in the row and the method in the column. A paired Wilcoxon rank sum test indicated significantly better performance of all 6 methods compared to the classical NPI and Cox' model with L_1 regularization ($p<0.001$). The histograms show a slightly better performance of the NPI compared to Cox-L_1 but a worse performance compared to the other methods. The performance of the additive SURLSSVM2 model is better than that of other models.

An important application of survival models is the ability of the produced prognostic indices to discriminate patients according to risk of relapse. To illustrate this ability, we define low risk patients as patients with a prognostic index lower than the mean prognostic index, the other patients being high risk. For models using Cox regression, low risk patients have an index higher than the mean index. High risk patients have an index lower than the mean index. Discrimination between risk groups is illustrated in Figure 14, where the Kaplan-Meier method [40] is used to calculate survival curves for low and high risk patients. For all methods the difference between survival of both groups was significant according to the logrank test ($p<0.001$). The χ^2 statistics for this test are given in table 2. The survival curves obtained with the NPI as prognostic index are added to each plot for comparison. The L_1 regularized Cox model achieves a lower survival for the high risk patients than the NPI model. No difference is observed in survival of low risk patients. The PLANN-ARD and the SURLSSVM2 models obtain a small decrease and increase in survival for the high and low risk groups respectively as compared to the NPI. The SURSVM and the SURLSSVM models achieve an decreased survival within the low risk patients, but no difference in survival is noted for the high risk patients. The discriminating ability of both additive models are the best, with the SURLSSVM2 model being slightly better than the Cox model.

Table 2: χ^2 values of the logrank test, testing for a difference in survival between low and high risk groups.

method	χ^2
NPI	18.7253
Cox-$L1$	118.9082
PLANN-ARD	130.8947
SURSVM	120.9805
SURLSSVM	57.1949
SURLSSVM2	35.9437
additive model: Cox	179.5401
additive model: SURLSSVM2	91.7322

5 Conclusions

We introduced a least squares support vector model for censored data. The performance of this model was comparable to the clinically accepted Cox and the more flexible PLANN-ARD model. However, after using an additive kernel the performance of the presented model was increased. The estimated functional forms when using a linear kernel agreed on the estimations of Cox' model. The use of a polynomial kernel allows more flexible functional forms. The impact of the continuous covariates by this more flexible model was comparable with the estimates made by the PLANN-ARD model, but the first model has the advantage of being convex.

The ability of separating patients into risk cohorts was visualized by dividing patients into two separate risk groups according to the risk predicted by the different methods and the survival curves for each of these cohorts were estimated. The prognostic index obtained by the additive SURLSSVM2 model separated both groups the best.

Another advantage of the proposed model is its performance on high dimensional data. We showed that the performance of this model is better than the performance of the Cox model for datasets with a high dimension compared to the number of samples. In future work we will investigate the use of the SURLSSVM2 model for high dimensional data on real life survival data.

Acknowledgements

This research is supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006, IDO 05/010, IOF-KP06/11, IOF-SCORES4CHEM, several PhD and postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, G.0407.02, G.0360.05, G.0519.06, FWO-G.0321.06, G.0341.07, 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. S. Van Huffel is a full professor and J.A.K. Suykens is a

professor at the Katholieke Universiteit Leuven, Belgium.

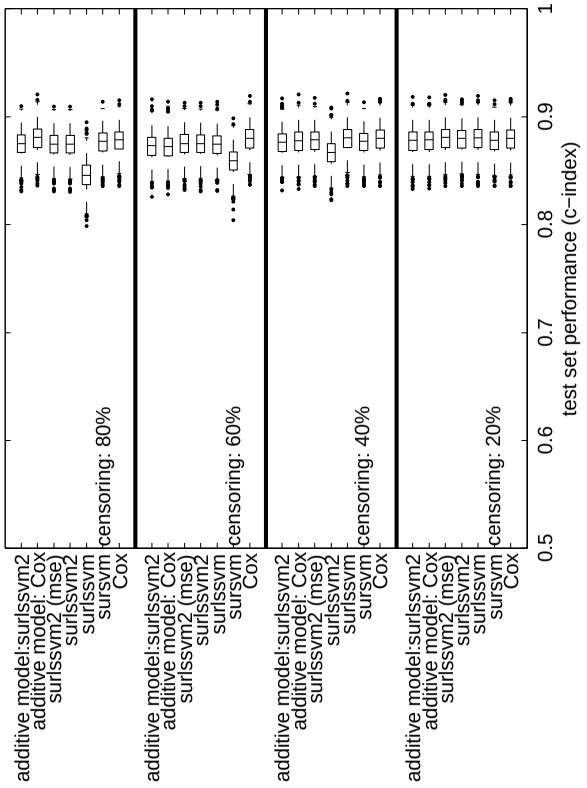


Figure 6: Illustration of the impact of the censoring percentage on test set performance on a dataset on lung capacity. The boxplots illustrate the performance on 1000 bootstraps of the test set. All models perform well at high censoring performances. No difference in performance was noted between Cox' model and the additive models, the non-additive models based on (LS-)SVMs seem to be less reliable.

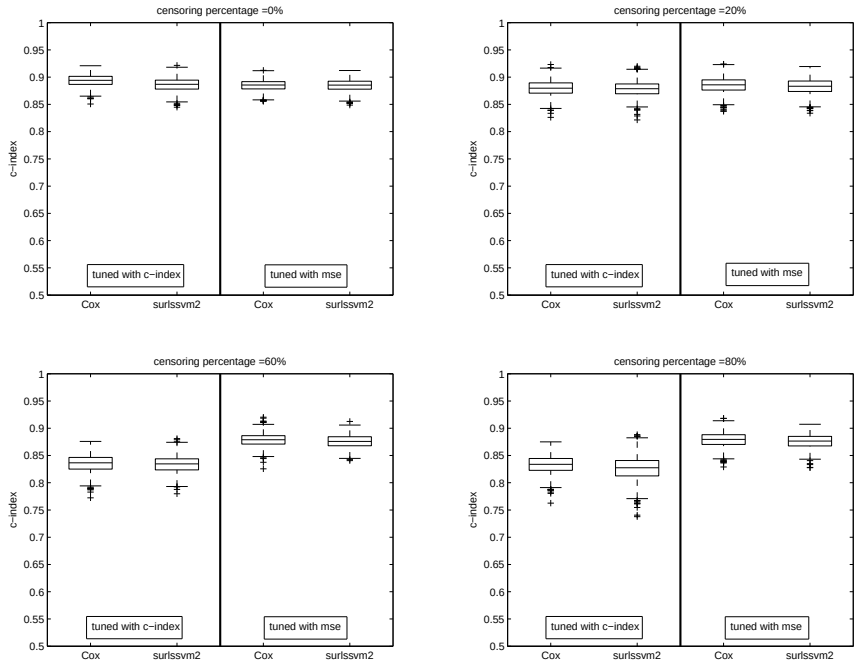
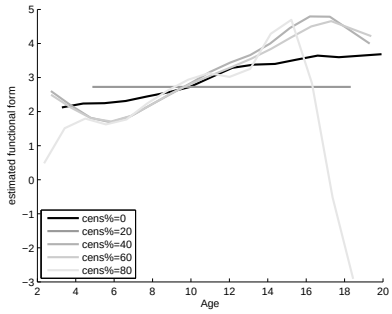
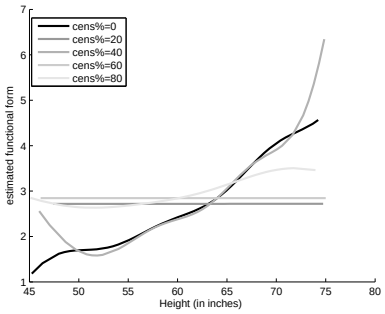


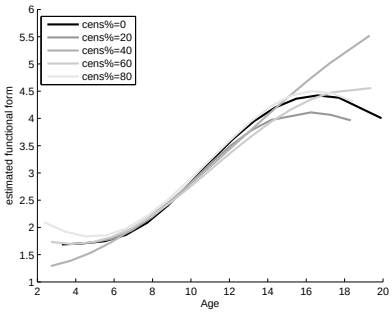
Figure 7: Additive model using Cox or SURLSSVM2 as linear combinator, illustrating the impact of the censoring percentage on test set performance for two different model selection criteria: c-index and modified mse. Without censoring, no difference in performance is seen between the two selection criteria. With an increasing amount of censoring, the performance after tuning with the c-index drops, while tuning with the modified mse leads to more stable results.



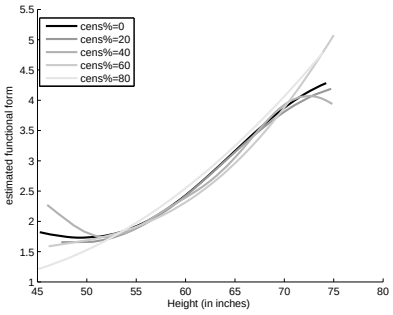
(a) c-index as model selection criterion



(b) c-index as model selection criterion



(c) mse as model selection criterion



(d) mse as model selection criterion

Figure 8: Impact of censoring percentage on the estimated functional form for two different model selection criteria: c-index and mse, on the lung capacity dataset. Results are illustrated after using the SURLSSVM2 model with a RBF kernel. With an increasing amount of censoring, the functional form after tuning with the c-index varies in comparison with the estimated functional form without censoring. In case the mse is used as a model selection criterion, the estimated functional form is more stable.

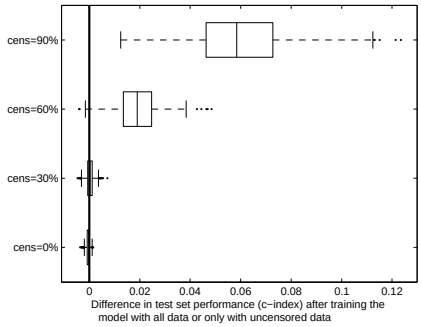


Figure 9: Impact of taking censored observations into account in the training step. We compared the performance of models which were trained by only incorporating uncensored observations, with models which did include censored observations, for different censoring percentages. The boxplots show the result of the difference in c-index between a model with and without censored observations on 1000 bootstraps of the test set. With 30% of censored observations, the observed performance was slightly better when including the censored observations. At higher censoring percentages, the improvement was significant.

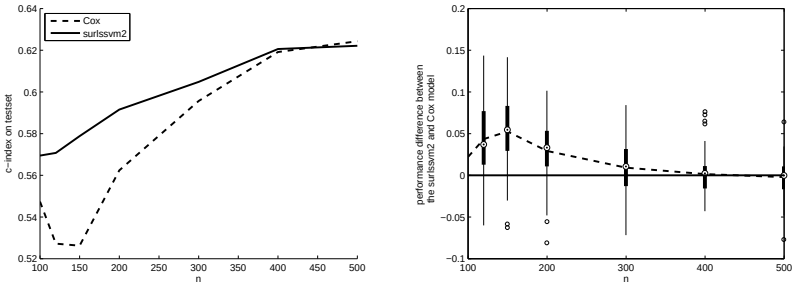


Figure 10: Performance of Cox' and the SURLSSVM2 model on high dimensional data. We created 90 training and test sets, with a Weibull distributed survival time depending on $d = 100$ covariates. The number of samples was varied between $n = 100$ and $n = 500$ and the censoring percentage was 30%. The curves in the left figure represent the mean test set performance. The right figure shows the spread of the performance difference between the SURLSSVM2 model and the Cox model on the test sets. For datasets for which the number of samples is considerably larger than the dimension, both models perform equally well. However, for smaller datasets with a high dimension the SURLSSVM2 model performs significantly better.

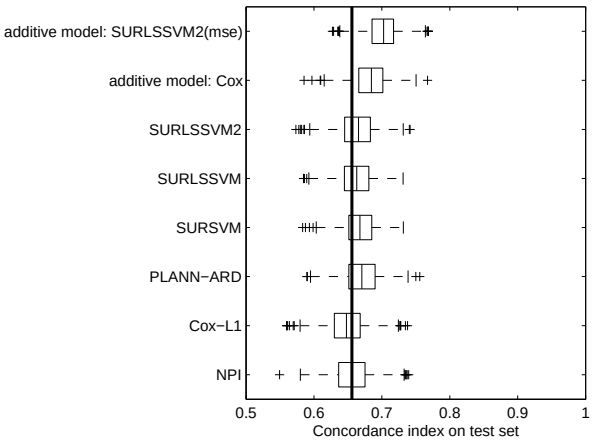
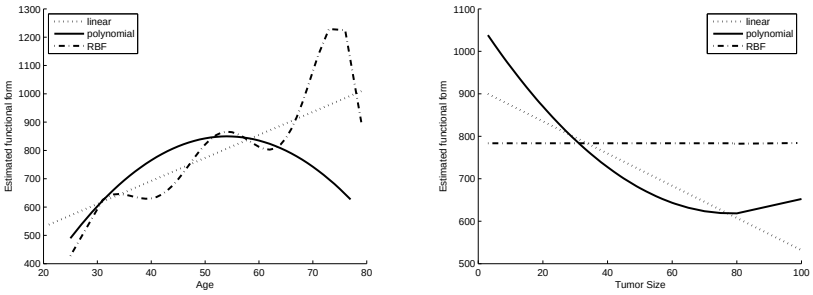
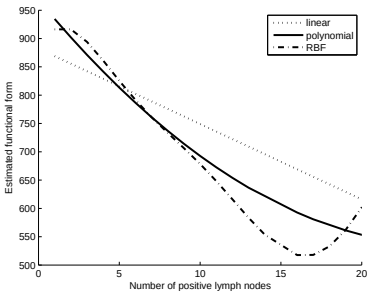


Figure 11: Concordance index on 1000 bootstraps of the test set of the German Breast Cancer Study Group dataset. The results for all kernel based methods were obtained with a polynomial kernel with degree $d = 2$. Comparing all non-additive models, PLANN-ARD and the SURLSSVM2 models perform best. Using additive models results in a better test set performance, and we gain interpretability of the covariates effects.

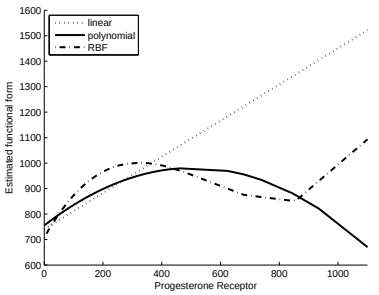


(a) Age

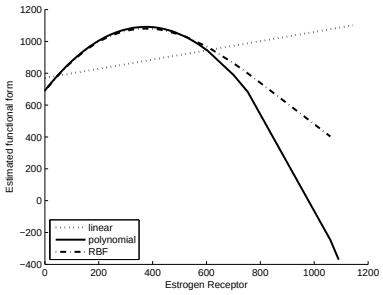
(b) Tumor Size



(c) Number of positive lymph nodes



(d) Progesterone Receptor



(e) Estrogen Receptor

Figure 12: Estimated functional forms (SURLSSVM2 model) for continuous covariates in the Breast Cancer Study Group Data with a linear, polynomial (degree $d = 2$) and a RBF kernel. The polynomial kernel is taken because the estimated effects are not restricted to be linear (as with the linear kernel) and are not too flexible (as with the RBF kernel). The linear kernel suggests a lower survival for younger patients. The polynomial kernel estimates a higher risk for younger patients but also indicates an increase in risk for older patients. Concerning tumor size, the number of lymph nodes, progesterone and estrogen receptors, the estimated functional forms agree with what is known from clinical practise.

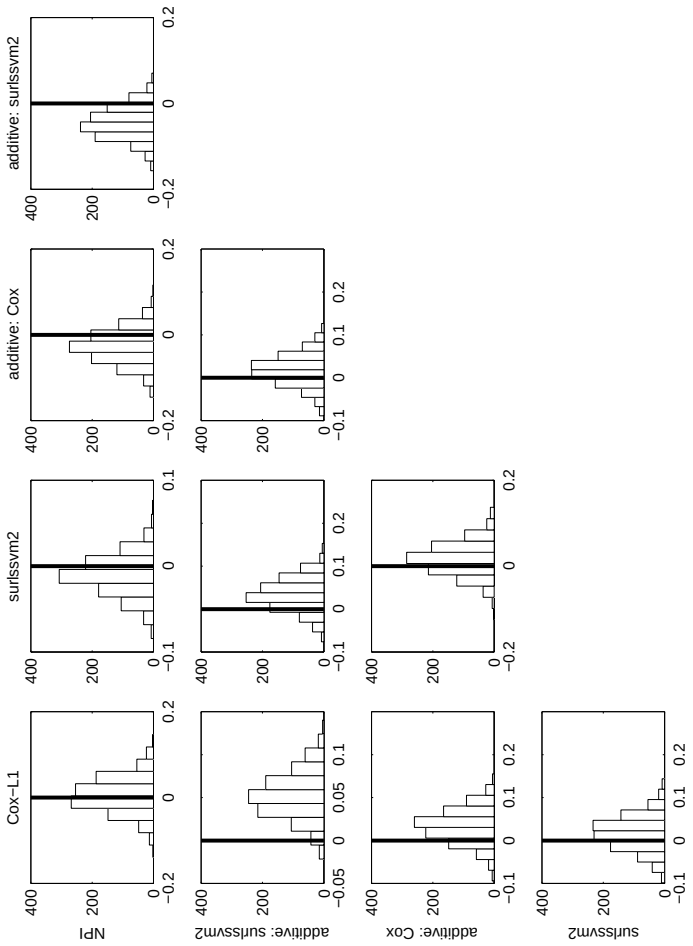
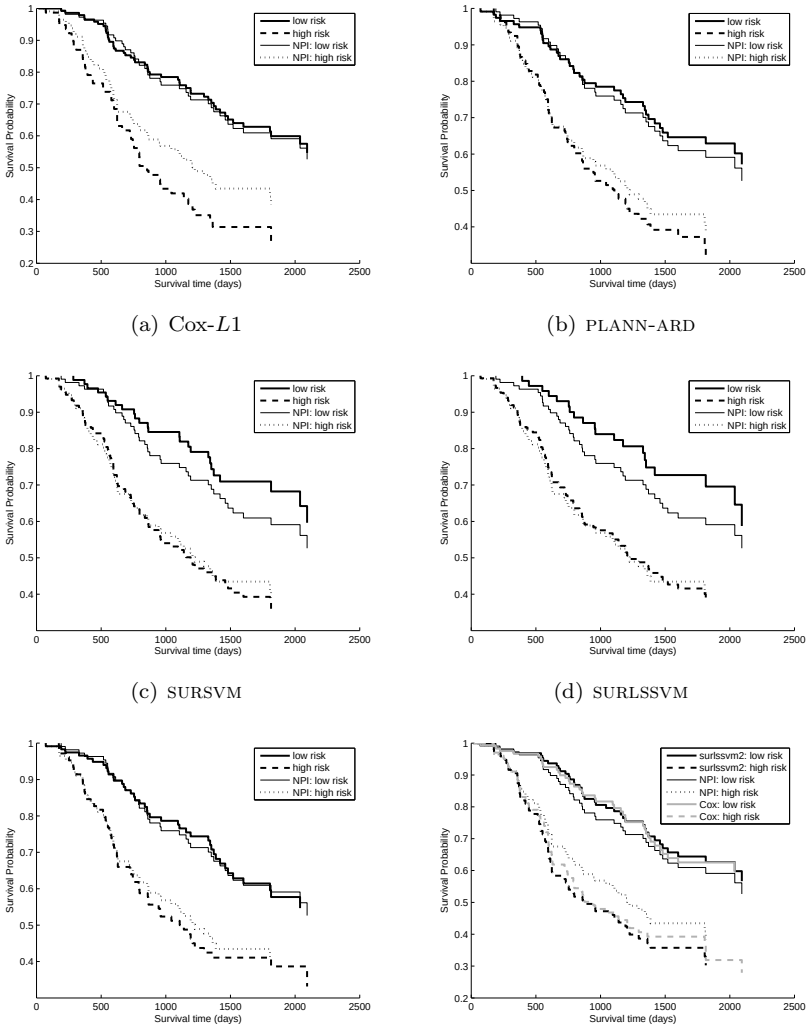


Figure 13: Histograms of the difference in performance on 1000 bootstraps of the test set of the German Breast Cancer Study Group data between the method in the row and the method in the column. Cox-L_1 performs the worst and the additive SURLSSVM2 model the best.



(a) Cox-$L1$ (b) PLANN-ARD

(c) SURSVM (d) SURLSSVM

(e) SURLSSVM2 (f) additive models

Figure 14: Kaplan-Meier survival curves for low versus high risk groups according to the different methods. The barrier between low and high risk patients lies at the mean of the prognostic index. Both additive models have the best discriminating ability, the additive SURLSSVM2 model being slightly better than the additive Cox model.

A surlssvm

Starting from the model specification as defined in (15)

$$\begin{array}{ll} \min\limits_{w,\xi} & \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} v_{ij}\xi_{ij}^2 \\ \text{s.t.} & w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij},\ \forall\ i,j = \ 1,\ldots,\ n\,. \end{array} \tag{31}$$

The Lagrangian is formulated as

$$\mathcal{L}(w,\xi;\alpha) \quad = \quad \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}(w^T\varphi(x_j) - w^T\varphi(x_i) - 1 - \xi_{ij}) \tag{32}$$

where α_{ij} are the Lagrange multipliers. The optimality conditions [41] yield

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i,j=1}^{n}\alpha_{ij}(\varphi(x_j) - \varphi(x_i)) \\ \frac{\partial\mathcal{L}}{\partial\xi_{ij}} = 0 \rightarrow \alpha_{ij} = \gamma v_{ij}\xi_{ij} \\ \frac{\partial\mathcal{L}}{\partial\alpha_{ij}} = 0 \rightarrow w^T(\varphi(x_j) - \varphi(x_i)) - 1 - \xi_{ij} = 0\,. \end{cases} \tag{33}$$

Elimination of w and ξ_{ij} leads to

$$\left[\gamma DKD^T + I_{n^2}\right]\alpha = \gamma 1_{n^2}\,. \tag{34}$$

The model's outcome for a new datapoint x^* is defined as

$$\begin{aligned} \hat{y}^* &= w^T\varphi(x^*) && (35)\\ &= \sum_{ij}\alpha_{ij}(\varphi(x_j) - \varphi(x_i))^T\varphi(x^*) && (36)\\ &= \alpha^T DK_n(x^*)\,, && (37) \end{aligned}$$

with $K_n(x^*) = [K(x_1,x^*)\ \ldots\ K(x_n,x^*)]^T$.

B surlssvm2

The primal formulation of the second model was given in (21) as

$$\begin{array}{ll} \min\limits_{w,\xi,b,\chi} & \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} v_{ij}\xi_{ij}^2 + \frac{1}{2}\mu\sum\limits_{j=1}^{n}\chi_j^2 \\ \text{s.t.} & \begin{cases} w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij}, & \forall\ i,j = 1,\ldots,n \\ \delta_j t_j = \delta_j(w^T\varphi(x_j) + b) - \chi_j, & \forall\ j = 1,\ldots,n\,. \end{cases} \end{array} \tag{38}$$

Introduce the Lagrange multipliers α_{ij} and β_j and formulate the Lagrangian

$$\begin{aligned}\mathcal{L}(w,\xi,\chi;\alpha,\beta) &= \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{i<j} v_{ij}\xi_i^2 + \frac{1}{2}\mu\sum_{j=1}^{n}\chi_j^2 \\ &\quad - \sum_{i,j=1}^{n}\alpha_{ij}[w^T(\varphi(x_j)-\varphi(x_i)) - 1 - \xi_{ij}] \\ &\quad - \sum_{j=1}^{n}\beta_j[\delta_j(-t_j + w^T\varphi(x_j) + b) - \chi_j]\,.\end{aligned} \tag{39}$$

The Karush-Kuhn-Tucker conditions [41] for optimality become

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w} = 0 \to w = \sum_{i,j=1}^{n}\alpha_{ij}(\varphi(x_j)-\varphi(x_i)) + \sum_{j=1}^{n}\beta_j\delta_j\varphi(x_j) \\ \frac{\partial\mathcal{L}}{\partial\xi_{ij}} = 0 \to \alpha_{ij} = -\gamma v_{ij}\xi_{ij} \\ \frac{\partial\mathcal{L}}{\partial\chi_j} = 0 \to \beta_j = -\mu\chi_j \\ \frac{\partial\mathcal{L}}{\partial b} = 0 \to \sum_{j=1}^{n}\beta_j\delta_j = 0 \\ \frac{\partial\mathcal{L}}{\partial\alpha_{ij}} = 0 \to w^T(\varphi(x_j)-\varphi(x_i)) - 1 - \xi_{ij} = 0 \\ \frac{\partial\mathcal{L}}{\partial\beta_j} = 0 \to \delta_j(-t_j + w^T\varphi(x_j) + b) - \chi_j = 0 \end{cases} \tag{40}$$

After eliminating w, ξ and χ, the system of linear equations becomes

$$\begin{bmatrix} DKD^T + I_{n_c}/\gamma & DKR & 0 \\ RKD^T & RKR + I_n/\mu & \delta \\ 0 & \delta^T & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ Rt \\ 0 \end{bmatrix}. \tag{41}$$

The model's prediction for a new datapoint x^* can be calculated as

$$\hat{y}^* = w^T\varphi(x^*) + b \tag{42}$$

$$= \Big(\sum_{ij}\alpha_{ij}(\varphi(x_j)-\varphi(x_i)) + \sum_{j=1}^{n}\beta_j\delta_j\varphi(x_j)\Big)^T\varphi(x^*) + b \tag{43}$$

$$= (\alpha^T D + \beta^T)K_n(x^*) + b\,. \tag{44}$$

References

[1] Galea M., Blamey R., Elston C., and Ellis I. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment* 1992, 22:207–219.

[2] Boracchi P., Biganzoli E., and Marubini E. Joint modelling of cause-specific hazard functions with cubic splines: an application to a large series of breast cancer patients. *Computational Statistics & Data Analysis* 2003, 42(1-2):243–262.

[3] Sleeper L.A. and Harrington D.P. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the Americam Statistical Association* 1990, 85(412):941–949.

[4] Lisboa P., Wong H., Harris P., and Swindell R. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine* 2003, 28(1):1–25.

[5] Boracchi P., Biganzoli E., and Marubini E. Modelling cause-specific hazards with radial basis function artificial neural networks: application to 2233 breast cancer patients. *Statistics in Medicine* 2001, 20(24):3677–3694.

[6] Cox D.R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* 1972, 34:197–220.

[7] Hastie T. and Tibshirani R. *Generalized additive models.* Chapman and Hall, 1990.

[8] O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* 1988, 9:531–542.

[9] Harrell F.E., Lee K.L., and Pollock B.G. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute* 1988, 80(15):1198–1202.

[10] Gray R.J. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992, 87:942–951.

[11] Leblanc M. and Crowley J. Adaptive regression splines in the Cox model. *Biometrics* 1999, 55(1):204–213.

[12] Li H. and Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics* 2005, 21(10):2403–2409.

[13] Li H. and Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. In *Pacific Symposium on Biocomputing* 2003:65–76.

[14] Huang J.Z., Kooperberg C., Stone C.J., and Truang Y.K. Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics* 2000, 28(4):961–999.

[15] Faraggi D. and Simon R. A neural network model for survival data. *Statistics in Medicine* 1995, 14:73–82.

[16] Bakker B., Heskes T., Nejit J., and Kappen B. Improving Cox survival analysis with a neural-bayesian approach. *Statistics in Medicine* 2004, 23:2989–3012.

[17] Biganzoli E., Boracchi P. , Mariani L., and Marubini, E. Feedforward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine* 1998, 17(10):1169–1186.

[18] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Support Vector Machines for Survival Analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), 2007.

[19] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Survival SVM: a Practical Scalable Algorithm. In *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN08)* 2008, 89–94.

[20] Herbrich R., Graepel T., and Obermayer K. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers* 2000, 115–132.

[21] Suykens J.A.K. and Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters* 1999, 9(3):293–300.

[22] Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., and Vandewalle J. *Least Squares Support Vector Machines.* World Scientific, Singapore, 2002.

[23] Vapnik V. *Statistical Learning Theory.* Wiley and Sons, 1998.

[24] Gunn S.R. and Kandola J.S. Structural modelling with sparse kernels. *Machine Learning* 2002, 115–136.

[25] Pelckmans K., Goethals I., De Brabanter J., Suykens J.A.K., and De Moor B. *Componentwise Least Squares Support Vector Machines*, chapter Support Vector Machines: Theory and Applications, (Wang L., ed.), Springer 2005, 77–98.

[26] Kalbfleisch J.D. and Prentice R.L. *The Statistical Analysis of Failure Time Data 2002.* Wiley series in probability and statistics.

[27] Wagner S.S. and Altmann S.A. What time do the baboons come down from the trees? (an estimation problem). *Biometrics* 1973, 29:623–635.

[28] Finkelstein D.M. A proportional hazards model for interval-censored failure tine data. *Biometrics* 1986, 42:845–854.

[29] De Gruttola V.G. and Lagakos S.W. Analysis of double-censored survival data, with application to aids. *Biometrics* 1993, 49(1):13-22.

[30] Harrell F. Jr., Klee K., Califf R., Pryor D., and Rosati R. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine* 1984, 95:634–635.

[31] James L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 1984, 25(3):303–325.

[32] Kahn M. An exhalent problem for teaching statistics. *The Journal of Statistical Education* 2005, 13(2).

[33] Schumacher M., Basert G., Bojar H., Huebner K., Olschewski M., Sauerbrei W., Schmoor C., Beyerle C., Neumann R.L.A., and Rauschecker H.F. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* 1994, 12(10):2086-2093.

[34] Park M.Y. and Hastie T. L1 regularization path algorithm for generalized linear models. *J. R. Statist. Soc. B* 2007, 19:659–677.

[35] Brouckaert O. and Camerlynck E. and Van Belle V. and Van Huffel S. and Pintens S. and Amant F. and Leunen K. and Smeets A. and Berteloot P. and Van Limbergen E. and Weltens C. and Van den Bogaert W. and Paridaens R. and Vergote I. and Christiaens M.R. and Wildiers H. and Neven P. Tumor biology and prognosis of primary operable breast cancers (BC) by age. *Accepted for publication in Proc. of the 31st Annual San Antonio Breast Cancer Symposium* 2008, San Antonio, Texas (USA).

[36] Verschraegen C., Vihn-Hung C., Cserni G., Gordon R., Royce M.E., Vlastos G., Tai P., and Storme G. Modeling the effect of tumor size in early breast cancer. *Annals of Surgery* 2005, 241(2):309–318.

[37] Fisher B., Bauer M., Wickerham L., Redmong C.K., and Fisher E.R. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer* 1983, 52:1551-1557.

[38] Pichon M.F., Pallud C., Brunet M., and Milgrom E. Relationship of presence of progesterone receptors to prognosis in early breast cancer. *Cancer Research* 1980, 40:3357–3360.

[39] Lamy P.J., Pujol P., Thezenas S., Kramar A., Rouanet P., Guilleux F., and Grenier J. Progesterone receptor quantification as a strong prognostic determinant in postmenopausal breast cancer women under tamoxifen therapy. *Breast cancer reasearch and treatment* 2002, 76:65–71.

[40] Kaplan E.L. and Meier P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 1958, 53:457–481.

[41] Boyd S. and Vandenberghe L. *Convex optimization.* Cambridge University Press, 2004.