Additive survival least squares support vector machines

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

Katholieke Universiteit Leuven, ESAT-SCD,
Kasteelpark Arenberg 10 bus 2446, B-3001 Leuven (Belgium),
{vanya.vanbelle, kristiaan.pelckmans, johan.suykens, sabine.vanhuffel}@esat.kuleuven.be

Abstract

This work studies a new survival modelling technique based on least squares support vector machines. 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 it 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. The proposed model was compared with different survival models from the literature on the clinical German Breast Cancer Study Group data and on the high dimensional Norway/Stanford Breast Cancer Dataset.

1 Introduction

Quantifying the risk a patient faces is of major relevance for deciding on a certain medical treatment. The acquisition of such risk scores 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.

Analysis of survival data is of interest within medical applications in general, and specifically within cancer studies. The oldest survival models are parametric models, in which an underlying model is assumed and the parameters are estimated. Examples are the Weibull model, with the exponential model as special case and the accelerated failure time model. For more details on parametric models, we refer to [1]. The most commonly used survival model is the semi-parametric proportional hazard model [2]. COX proposed

‡Corresponding author

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

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 only on time. 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.

Different methods allowing for non-linearities in the covariates and/or without the proportional hazard assumption have been proposed. Hastie and Tibshirani [3] introduced generalized additive models in which the linear form is replaced by a sum of smooth functions. Splines were used in [4, 5, 6, 7]. Kernel COX regression was proposed by [8], while functional ANOVA modelling was used in [9]. Neural networks were used to extend the COX model towards non-linear covariate effects by [10, 11, 12].

The application of Support Vector Machines (SVMs) within survival analysis was previously presented in [13, 14]. We advance this work and according to the methodology of LS-SVMs [15, 16], 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 [17] for classification and regression problems, where the solution is found by solving linear Karush-Kuhn-Tucker systems [18]. Non-linearity is introduced by using a feature map (transformation to a high dimensional space). Due to Mercer's theorem it is not necessary to specify the feature map explicitly, instead one makes use of a positive definite kernel [17].

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 datapoints into account. Interpretability is obtained by using componentwise kernels [19, 20].

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

2 Survival models

2.1 Survival data

Failure time data are 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 right, left and interval censored data. However, within this work we will restrict to right censored data. We denote a datapoint $i = 1, \ldots, n$ with covariates x_i as (x_i, t_i, δ_i). The observed

failure time t_i can be expressed as $t_i = \min(c_i, \tilde{t}_i)$, with $\tilde{t}_i$ the true failure time, c_i the censoring time and $\delta_i = I(\tilde{t}_i \leq c_i)$ an event indicator.

Let T be a random variable. One possible way to present survival data is to calculate the survival (distribution) function (SDF) $S(t)$ [1]:

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

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 The 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 [2] models the hazard function 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

$$L(\beta) = \prod_{j=1}^{k} \left(\frac{\exp(\beta^T x_j)}{\sum_{l \in \mathcal{R}_j} \exp(\beta^T x_l)} \right)\,, \tag{4}$$

is maximized. $\mathcal{R}_j$ represents all patients at risk at the jth failure time and k the number of distinct failure times.

2.3 Partial Logistic Artificial Neural Networks

Biganzoli *et al.* [12] 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

$$L(\beta) = \prod_{i=1}^{n} \prod_{l=1}^{l_i} \lambda_\beta(x_i, t_l)^{\delta_i(t_l)} (1 - \lambda_\beta(x_i, t_l))^{(1-\delta_i(t_l))}\,, \tag{5}$$

with $l = 1, \ldots, l_i$ 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. The parameter vector β is in this case composed of the weights (w_h, w_{ih} and w) and the bias terms (b_h and b) of the hidden and input layer. PLANN uses one input for each covariate and an additional input for the time interval. The multilayer perceptron structure is

$$\frac{\lambda_\beta(x, t_k)}{1 - \lambda_\beta(x, t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h g\Big(\sum_{i=1}^{N_i} w_{ih} x_i + w t_k + b_h\Big) + b\right), \tag{6}$$

where the indices i and h denote the input and hidden node layers respectively and $g(.)$ is the hidden layer activation function. The maximum likelihood $\max_\beta L(\beta)$ is used to estimate β.

2.4 Survival svm

The methodology using support vector machines (SVM) for the analysis of survival data, starts from a different perspective [13, 14]. 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

$$u(x) = w^T \varphi(x)\,, \tag{7}$$

where w indicates the unknown parameter vector and $\varphi(x)$ represents a transformation (feature map) of the covariates x. Mercer's condition [17] enables that the feature space can be left undefined when using positive definite kernels, the use of which is illustrated in appendix for the more theoretically interested reader.

The concordance index (c-index) as introduced by Harrell [21], measures the concordance between the utility function and the observed failure times using censored as well as uncensored observations. The empirical concordance between the event times t_i and a prognostic index u_i based on dataset $\mathcal{D} = \{(x_i, t_i, \delta_i)\}_{i=1}^n$ is defined as

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

where $\mathcal{I}$ is the indicator function and v_{ij} indicates whether the datapoints (x_i, t_i, δ_i) and (x_j, t_j, δ_j) are comparable

$$v_{ij} = \begin{cases} 1 & (t_i < t_j \text{ and } \delta_i = 1) \text{ or } (t_j < t_i \text{ and } \delta_j = 1) \\ 0 & \text{otherwise}\,. \end{cases} \tag{9}$$

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

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

In the remainder of this work, we will assume that $t_i < t_j$ for $i < j$. If this assumption is not valid, the data can be sorted first. Knowing that t_i is less than t_j, we want $u(x_i)$ to be less than $u(x_j)$ in order to extract the ranking. The survival SVM 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:

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

with $\gamma \geq 0$ a regularization constant and w defined as in equation (7).

3 A Survival Least Squares Support Vector Machine

This section studies the use of least squares support vector machines [15, 16] for kernel-based modelling of failure time data. A model with utility function $u(x)$ 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 datapoints correctly with respect to their survival time. As in [13, 14] we base our method on maximizing the concordance index.

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 allows for misrankings. In the loss function the L_2-norm of the slack variables is included. Here we study a utility function $u(x) : \mathbb{R}^d \to \mathbb{R}$ where $u(x) = w^T \varphi(x)$ which is fit by the following constrained minimization:

$$\boxed{\text{SURLSSVM}} \qquad \min_{w,\xi} \qquad \tfrac{1}{2}w^T w + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2$$
$$\text{subject to} \quad w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij},\ \forall\ i,j = 1,\ldots,\ n\,, \tag{12}$$

where $\gamma > 0$ is a fixed regularization constant trading model complexity and reconstruction error. We refer to appendix A for more details. Remark that in contrast with the above equation a minus sign is used in the right hand side of the constraint in equation (11) and positive error variables ξ_{ij}. The same results are obtained when using $1 - \xi_{ij}^*$ with ξ_{ij}^* negative. Since a squared error is used in equation (12) the same result would be obtained for a right hand side equal to $1 - \xi_{ij}^*$ where ξ_{ij}^* would equal $-\xi_{ij}$.

3.2 Interpretability and Additive Models

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 question 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 mean squared error (mse) instead of the c-index. The standard definition of the mean squared error between observation t_i and the corresponding prediction $\hat{t}_i$

$$\text{mse}_{\text{standard}} = \frac{\sum_{i=1}^{n}(\hat{t}_i - t_i)^2}{n}\,, \tag{13}$$

where n is the number of datapoints in the set over which the measure is calculated. However, this measure can only be used for uncensored observations. In order to include information contained in right censored observations, the mean squared error is adapted as:

$$\text{mse} = \frac{\sum_{i=1}^{n} q_i(\hat{t}_i - t_i)^2}{n}\,, \tag{14}$$

with q_i defined as

$$q_i = \delta_i + (1 - \delta_i)\mathcal{I}(\hat{t}_i - t_i < 0)\,. \tag{15}$$

The weight q_i is 1 for uncensored datapoints. For right censored datapoints, the weight is 0 in case the prediction is later than the observed failure time (we do not have evidence that the prediction is wrong) and 1 otherwise. Note that this measure is closely related to the censored Least Absolute Deviation (CLAD), as used in [22].

Model (12) is then modified as follows:

$$\boxed{\text{SURLSSVM2}} \qquad \min_{w,\xi,b,\chi} \quad \tfrac{1}{2}w^T w + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 + \frac{1}{2}\mu \sum_{j=1}^{n} \chi_j^2$$

$$\text{subject to} \quad \begin{cases} w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij}, & \forall\, i,j = 1,\ldots,n \\ \delta_j t_j = \delta_j(w^T\varphi(x_j) + b) - \chi_j, & \forall\, j = 1,\ldots,n\,, \end{cases} \tag{16}$$

where $\gamma \geq 0$ and $\mu \geq 0$ are fixed regularization constants. The solution to this problem statement is found by solving a linear system in $n^2 + n + 1$ unknowns (see appendix B). The dimension of the linear system can be reduced as elaborated in [14].

Remark 1 *In case μ is taken to be 0 the model reduces to the* SURLSSVM *model as in equation (12).*

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

A drawback of SVM models in a clinical setting is that interpretation of the effect of a single variable to the outcome of the model is not possible (the outcome is produced as an unknown combination of the inputs). To increase the interpretability, the utility function $u(x)$ is modelled as the weighted sum of d utilities:

$$u(x) = \sum_{p=1}^{d} r_p u^{(p)}(x)\,, \tag{17}$$

with $u^{(p)}(x) = {w^{(p)}}^T \varphi^{(p)}(x)$ the utility of the pth covariate obtained by a linear combination of the features after transforming the pth covariate to a higher dimensional feature space by means of the pth feature map $\varphi^{(p)}$. The effect of each variable on the outcome can then be visualized when plotting the utility corresponding to the pth covariate in function of this covariate. The weights r_p are determined by a second model which is linear (Figure 1). This corresponds to using a componentwise kernel [20] as shown in appendix B.

3.3 Comparison of surlssvm with other models

The COX model is a semi-parametric 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 risks [1]. 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 follows from solving a linear system and the model formulation is convex.

The PLANN model proposed by Biganzoli *et al.* [12] has the drawback that datapoints 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.

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. Penalized versions of the partial likelihood to deal with this issue were proposed [23, 24, 25]. It is well known that kernel-based methods as SVM and LS-SVM are well suited for handling high dimensional data [16, 17].

4 Empirical Results

This Section illustrates results on clinical and micro-array 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 a: $K(x, x^*) = (\tau + x^T x^*)^a$, where $\tau \geq 0$ is a tuning parameter. a was chosen to be $a = 2$.
- 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 an exponential 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 and a mean squared error adapted for censored data.

4.1 Evaluation criteria

To compare the results of models, different measures are used. Firstly the concordance between the developed prognostic indices and the failure times are calculated as previously defined (see equation (8)). A common tool to examine performance of predictive tests is

the Receiver Operating Curve (ROC) and the area under this curve (AUC). Heagerty *et al.* proposed an extension of ROCs for survival data [26]. This time dependent ROC summarizes the discrimination potential of a marker by calculating AUCs for cumulative disease incidence by a certain time t. An AUC of 0.5 indicates that the test has no discriminative power, whereas an AUC of 1 indicates that the method is capable of perfectly discriminating between relapsed and non-relapsed patients.

To quantify to which extend the models are able to predict the order in which patients will relapse, a measure of explained variation is calculated. The difference between the observed and predicted order after using one of the models is compared with the difference between the observed and the predicted order without using any of the covariates. The explained variation is more formally defined as

$$R^2 = \frac{L_0(y, \hat{y}) - L(y, \hat{y})}{L_0(y, \hat{y})},$$

with y the observed and $\hat{y}$ the predicted event times, $L_0(y, \hat{y})$ the loss obtained without using any covariates and $L(y, \hat{y})$ the loss obtained with the model. In comparing different models, the model with the lowest loss $L(y, \hat{y})$ is considered to have the best ability to predict the outcome. Since we are not directly interested in point predictions, we calculate this measure using the ranks of the failure times and predictions instead of the absolute values. The measure is adapted for censoring by only incorporating an error for censored datapoints in cases where the prediction is less than the censoring time. More formally, the loss is defined as:

$$L(y, \hat{y}) = \sum_{i=1}^{n} |\min(\text{rank}(y_i), \text{rank}(c_i)) - \min(\text{rank}(\hat{y}_i), \text{rank}(c_i))|, \tag{18}$$

with $\text{rank}(y_i)$ the rank of observation y_i within the test set, c_i the censoring time and $\text{rank}(\hat{y}_i)$ the rank of the predicted observation within the test set.

4.2 Micro-array Data

This first example illustrates the use of the presented model on high dimensional micro-array data. The Norway/Stanford breast cancer data (NSBCD) [27] contains gene expression measurements from 115 women with breast cancer. The performance of the non-additive SURLSSVM and SURLSSVM2 models tuned with the c-index and a linear kernel, are compared with the performance of five models based on COX proportional hazard model. The principal components regression (PCR) uses principal component analysis to select a number of principal components, accounting for as much variation in the gene expression as possible and uses these as covariates for COX regression (see [28], chapter 3.4.4). The supervised PCR method (SPCR) [29, 30] first selects a subset of genes which are correlated with survival and applies PCR to this set of genes. As a third method the partial least squares regression applied to COX' model (PLS) is used [31]. Two penalized COX regression

models are also used: COX' model with L_1 regularization (COX_{L1}) [23, 24] and COX' model with L_2 regularization (COX_{L2}) [25].

The performances on 100 train-test-set randomizations are illustrated in Figure 2. There are no significant differences although three groups can be noted: COX_{L1} and COX_{L2} perform the least, followed by the SURLSSVM models. The other three methods perform slightly better. The performance of the SURLSSVM models can possibly be improved when combined with a feature selection method.

4.3 A Breast Cancer Study

The second example shows the results of the proposed additive model compared with other survival models on the German Breast Cancer Study Group (GBSG) data [32]. This dataset contains data on 686 women with breast cancer[1]. Available variables are hormone treatment, age, menopausal status, tumor size and grade, the number of positive lymph nodes, the progesterone and estrogen receptor (fmol). Two third of the 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 3 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 (SURLSSVM2) or the c-index (other (LS-)SVM based models) as model selection criterion. We compared results with the PLANN-ARD model, a COX model with L_1 regularization [24] and a COX model with the non-linear transformations obtained with the additive SURLSSVM2 model. In using the COX model the proportional hazard assumption was tested and no evidence was found to assume this assumption was violated.

The results are compared with the Nottingham Prognostic Index (NPI) [33], an index used in clinical practice to classify breast cancer patients into risk groups. 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. Figure 3(b) shows that the additive SURLSSVM2 model has the best discriminative potential over the whole time range. Table 1 shows that SURLSSVM2 models have the lowest loss and hence the highest measure of explained variation.

The estimated functional forms are shown in Figure 4. Functional forms were investigated using a linear, polynomial of the second degree and RBF kernel. 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 [34]. Concerning tumor size, the number of lymph nodes, progesterone and estrogen receptors, the estimated functional forms agree with what is known from clinical practice [35, 36, 37].

To be able to compare the results of PLANN-ARD with the SURLSSVM2 model, the

[1] `http://www.blackwellpublishers.com/rss/Volumes/A162p1.htm`

prognostic indices of the first model are averaged over time. Both the PLANN-ARD and the SURLSSVM2 model predict an increasing risk for breast cancer relapse in older patients, as shown in literature [34]. 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.

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 higher than the mean prognostic index, the other patients being high risk. For models using COX regression, low risk patients have an index lower than the mean index. High risk patients have an index higher than the mean index. Discrimination between risk groups is illustrated in Figure 5, where the Kaplan-Meier method [38] 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 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 a 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.

5 Conclusions

We introduced a least squares support vector model for censored data and compared its performance with other survival models on micro-array and clinical datasets. The performance of the tested models were comparable on the micro-array dataset. The performance on a clinical breast cancer dataset 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.

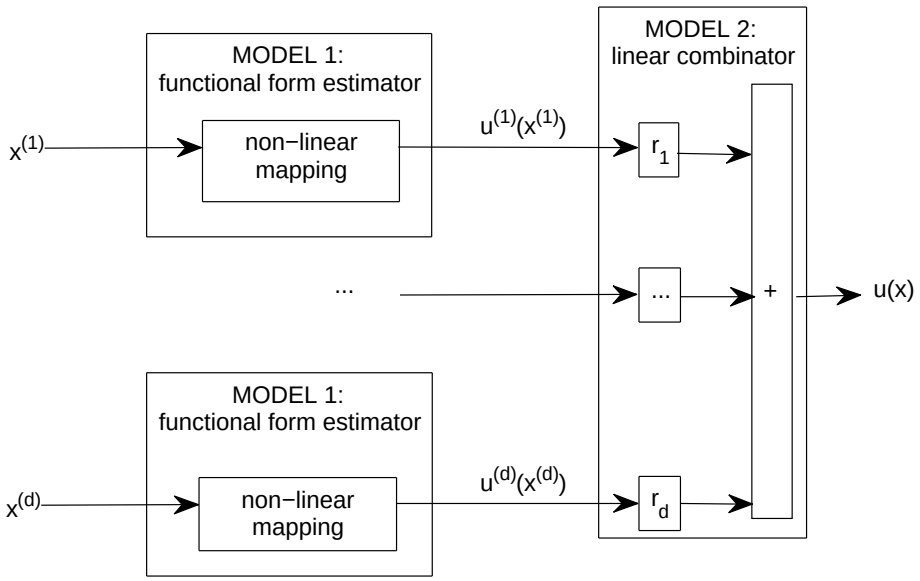
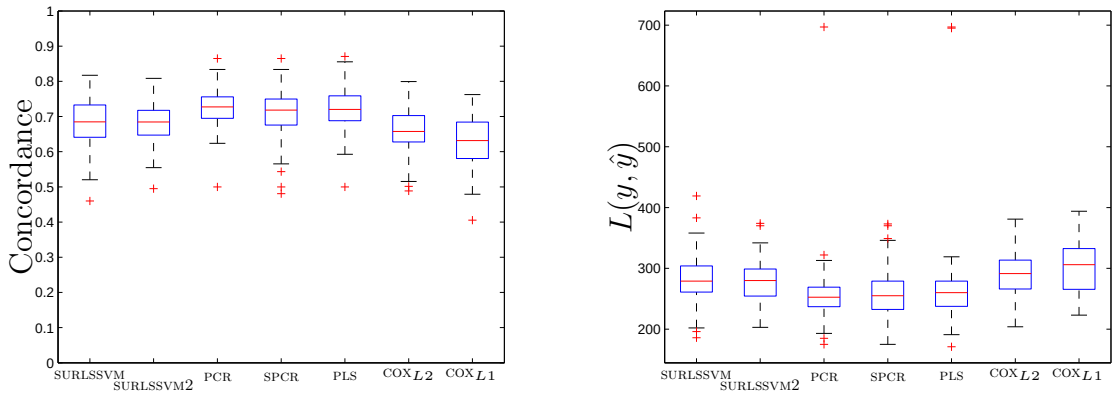


Figure 1: Construction of the prognostic index. The SURLSSVM2 model is used to estimate the functional form of each covariate $u^{(p)}(x^{(p)})$ with $p = 1, \ldots, d$. These contributions are than fed as inputs of a second model. This second model is a linear model combining the inputs as in a weighted sum to obtain the final prognostic index.

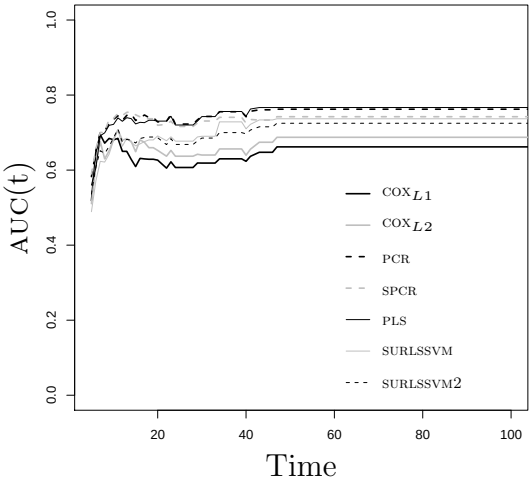
Table 1: Loss for the different survival models on the GBSG data. The loss is defined as the difference in ranking between the predicted and observed event times. For censored observations this difference was only added to the loss in cases where the prediction rank proceeded the observed rank. Both SURLSSVM2 models have the lowest loss and thus the highest measure of explained variation. These models are the best in predicting the order in which patients will relapse from breast cancer.

NPI	COX-L1	PLANN-ARD	SURSVM	SURLSSVM	SURLSSVM2	additive textsccox	additive SURLSSVM2
11350	10129	10771	11075	11764	9987	10453	10046



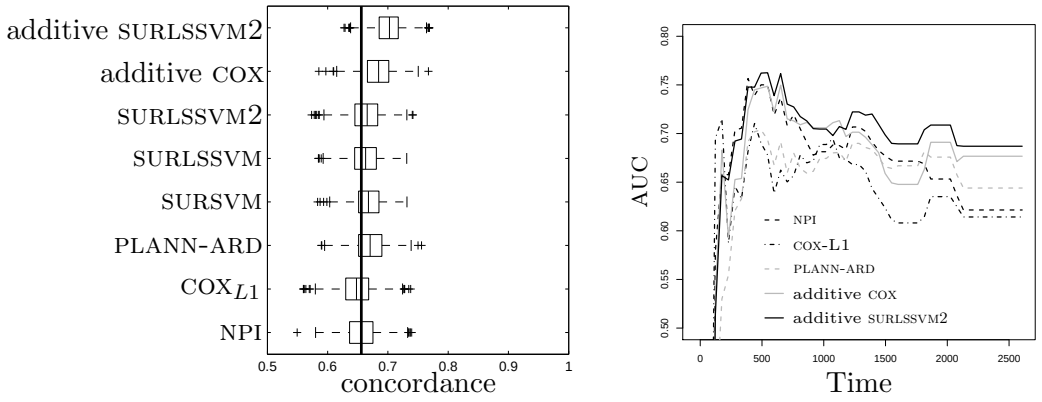
(a) concordance

(b) Loss



(c) time dependent AUC

Figure 2: Performance on 100 train-test-set randomizations of the Norway/Stanford breast cancer data set: (a) Concordance index, (b) Loss and (c) time dependent AUC. No significant differences are noted concerning concordance and time dependent AUC between the SURLSSVM, PCR, SPCR and PLS methods. The time dependent AUC and the loss for COX_{L1} and COX_{L2} are lower than for the other methods.



(a) (b)

Figure 3: (a) 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 $a = 2$. 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. (b) Time dependent AUC, summarizing the discrimination potential of a marker by calculating ROC curves for cumulative disease incidence by a certain time t. The additive SURLSSVM2 model has the best discriminative potential over the whole time range.

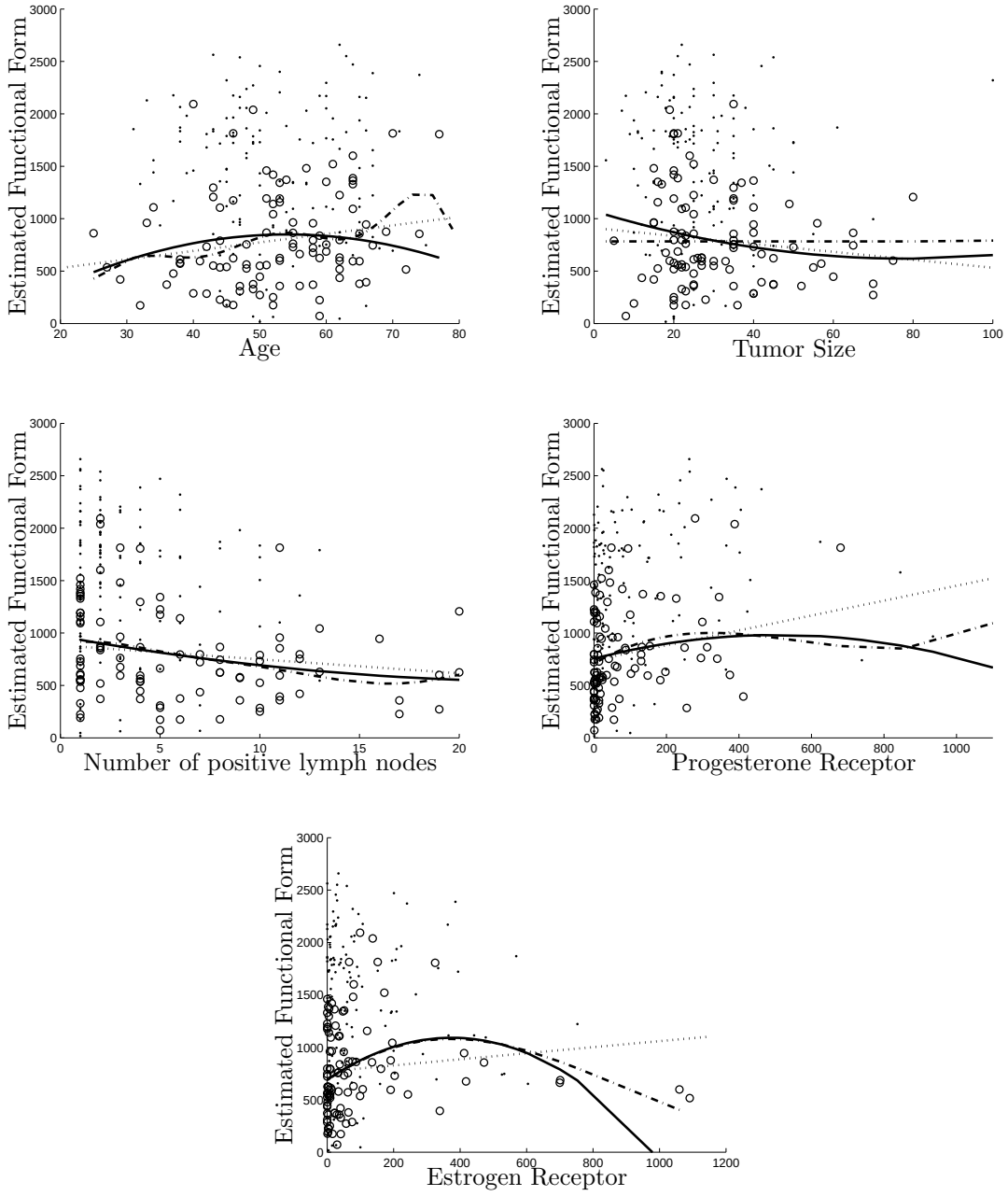
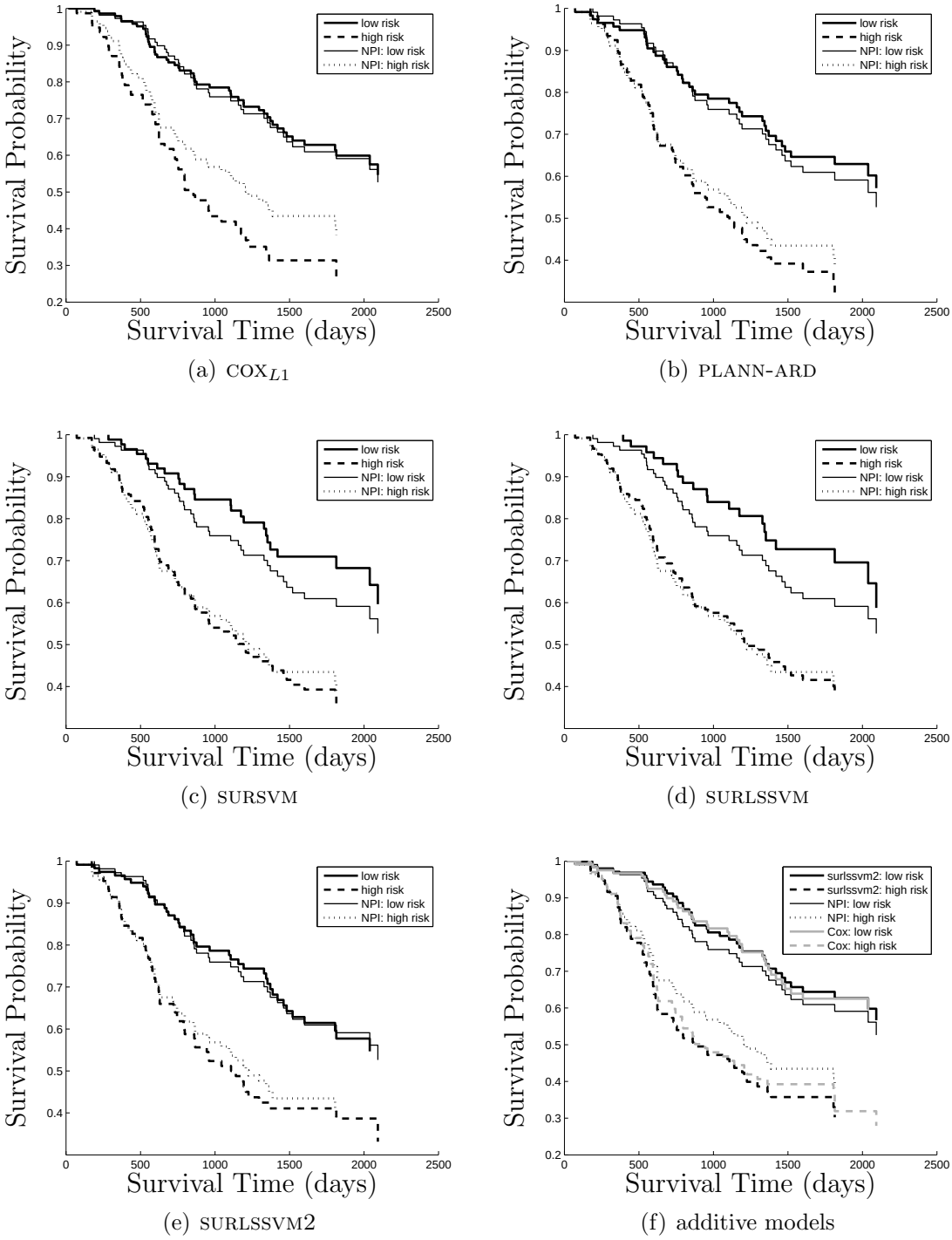


Figure 4: Estimated functional forms (SURLSSVM2 model) for continuous covariates in the Breast Cancer Study Group Data with a linear (dotted line), polynomial (degree $a = 2$) (solid line) and an RBF kernel (dashed line) using the additive SURLSSVM2 model. The circles represent the relapsed patients with their respectively covariate and failure time values. The dots represent the censored observations. The polynomial kernel is preferred 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. Within intervals where the amount of observations is satisfying, the results of all three kernels agree with the literature: an increasing size and number of involved lymph nodes decreases the survival time, whereas an increasing presence of both receptors increases the survival time.



(a) COX_{L1} (b) PLANN-ARD

(c) SURSVM (d) SURLSSVM

(e) SURLSSVM2 (f) additive models

Figure 5: Kaplan-Meier survival curves for low versus high risk groups according to the different methods. We define low risk patients as patients with a prognostic index higher than the mean prognostic index, the other patients being high risk. For models using COX regression, low risk patients have an index lower than the mean index. High risk patients have an index higher than the mean index. Both additive models have the best discriminating ability, the additive SURLSSVM2 model being slightly better than the additive COX model.

Acknowledgments

We kindly acknowledge the support and constructive remarks of E. Biganzoli and P. Boracchi. This research is supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006, IDO 05/010, IOF-KP06/11, IOF-SCORES4CHEM; Flemish Government: G.0407.02, G.0360.05, G.0519.06, G.0321.06, G.0341.07, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0302.07, McKnow-E, Eureka-Flite; Belgian Federal Science Policy Office: IUAP P6/04; EU: FP6-2002-LIFESCIHEALTH 503094, IST-2004-27214, FP6-MC-RTN-035801; Prodex-8 C90242; EU: ERNSI. V. Van Belle is supported by a grant from the IWT. K. Pelckmans is a postdoctoral researcher with FWO Flanders (A 4/5 SB 18605). S. Van Huffel is a full professor and J.A.K. Suykens is a professor at the Katholieke Universiteit Leuven, Belgium.

Supplementary material

A Matlab file for the SURLSSVM and SURLSSVM2 models can be downloaded from:
http://www.esat.kuleuven.be/sista/lssvmlab/survival/survival.zip.

This appendix presents the derivation of the SURLSSVM and SURLSSVM2 models. We start from the formulations of the SURLSSVM and SURLSSVM2 model as described in equations (12) and (16). We will formulate the Lagrangian and the Karush-Kuhn-Tucker optimality conditions.

A surlssvm model

Starting from the model specification as defined in (12)

$$\begin{aligned} \min_{w,\xi} \quad & \tfrac{1}{2} w^T w + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 \\ \text{subject to} \quad & w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij},\ \forall\, i,j = 1,\ldots,\ n\,, \end{aligned} \tag{19}$$

the Lagrangian is formulated as

$$\mathcal{L}(w,\xi;\alpha) \;=\; \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}(w^T\varphi(x_j) - w^T\varphi(x_i) - 1 - \xi_{ij}) \tag{20}$$

where α_{ij} are the Lagrange multipliers. The optimality conditions [18] yield

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i,j=1}^{n} \alpha_{ij}(\varphi(x_j) - \varphi(x_i)) \\ \frac{\partial \mathcal{L}}{\partial \xi_{ij}} = 0 \rightarrow \alpha_{ij} = \gamma v_{ij} \xi_{ij} \\ \frac{\partial \mathcal{L}}{\partial \alpha_{ij}} = 0 \rightarrow w^T(\varphi(x_j) - \varphi(x_i)) - 1 - \xi_{ij} = 0\,. \end{cases} \tag{21}$$

Elimination of w and ξ_{ij} leads to

$$\left[\gamma D K D^T + I_{n^2}\right]\alpha = \gamma 1_{n^2/2}\,, \tag{22}$$

where $D \in \{-1, 0, 1\}^{(n^2/2 \times n)}$ is defined as

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & -1 & \dots & 0 & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & \dots & 1 & 0 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix}, \tag{23}$$

and $K_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ denote the elements of the kernel matrix. The model's outcome for a new datapoint x^* is defined as

$$\begin{aligned} \hat{y}^* &= w^T \varphi(x^*) && (24) \\ &= \sum_{ij} \alpha_{ij}(\varphi(x_j) - \varphi(x_i))^T \varphi(x^*) && (25) \\ &= \alpha^T D K_n(x^*)\,, && (26) \end{aligned}$$

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

B surlssvm2 model

The formulation of the second model was given in (16) as

$$\begin{aligned} \min_{w,\xi,b,\chi} \quad & \tfrac{1}{2} w^T w + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\xi_{ij}^2 + \frac{1}{2}\mu \sum_{j=1}^{n} \chi_j^2 \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_j) - w^T\varphi(x_i) = 1 + \xi_{ij}, & \forall\ i,j = 1,\dots,n \\ \delta_j t_j = \delta_j(w^T\varphi(x_j) + b) - \chi_j, & \forall\ j = 1,\dots,n\,. \end{cases} \end{aligned} \tag{27}$$

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

$$
\begin{aligned}
\mathcal{L}(w,\xi,\chi;\alpha,\beta) &= \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{i<j} v_{ij}\xi_i^2 + \frac{1}{2}\mu\sum_{j=1}^{n}\chi_j^2 \\
&\quad - \sum_{i,j=1}^{n}\alpha_{ij}[w^T(\varphi(x_j)-\varphi(x_i)) - 1 - \xi_{ij}] \\
&\quad - \sum_{j=1}^{n}\beta_j[\delta_j(-t_j + w^T\varphi(x_j) + b) - \chi_j]\,.
\end{aligned} \tag{28}
$$

The Karush-Kuhn-Tucker conditions [18] for optimality become

$$
\begin{cases}
\frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i,j=1}^{n}\alpha_{ij}(\varphi(x_j)-\varphi(x_i)) + \sum_{j=1}^{n}\beta_j\delta_j\varphi(x_j) \\
\frac{\partial\mathcal{L}}{\partial\xi_{ij}} = 0 \rightarrow \alpha_{ij} = -\gamma v_{ij}\xi_{ij} \\
\frac{\partial\mathcal{L}}{\partial\chi_j} = 0 \rightarrow \beta_j = -\mu\chi_j \\
\frac{\partial\mathcal{L}}{\partial b} = 0 \rightarrow \sum_{j=1}^{n}\beta_j\delta_j = 0 \\
\frac{\partial\mathcal{L}}{\partial\alpha_{ij}} = 0 \rightarrow w^T(\varphi(x_j)-\varphi(x_i)) - 1 - \xi_{ij} = 0 \\
\frac{\partial\mathcal{L}}{\partial\beta_j} = 0 \rightarrow \delta_j(-t_j + w^T\varphi(x_j) + b) - \chi_j = 0
\end{cases} \tag{29}
$$

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

$$
\begin{bmatrix} DKD^T + I_{n_c}/\gamma & DKR & 0 \\ RKD^T & RKR + I_n/\mu & \delta \\ 0 & \delta^T & 0 \end{bmatrix}
\begin{bmatrix} \alpha \\ \beta \\ b \end{bmatrix} =
\begin{bmatrix} 1 \\ Rt \\ 0 \end{bmatrix}. \tag{30}
$$

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

$$
\begin{aligned}
\hat{y}^* &= w^T\varphi(x^*) + b && (31)\\
&= \Big(\sum_{ij}\alpha_{ij}(\varphi(x_j)-\varphi(x_i)) + \sum_{j=1}^{n}\beta_j\delta_j\varphi(x_j)\Big)^T\varphi(x^*) + b && (32)\\
&= (\alpha^T D + \beta^T R)K_n(x^*) + b\,. && (33)
\end{aligned}
$$

In cases where interpretability is pursued $u(x)$ is calculated as a sum of utility functions $u(x) = \sum_{p=1}^{d} u^{(p)}(x^{(p)})$. A prediction can than be made as

$$\hat{y}^* = \sum_{p=1}^{d} w_{\varphi^{(p)}}^T \varphi^{(p)}(x^*) + b \tag{34}$$

$$= \sum_{p=1}^{d} \Big(\sum_{ij} \alpha_{ij}^{(p)} (\varphi^{(p)}(x_j) - \varphi^{(p)}(x_i)) + \sum_{j=1}^{n} \beta_j \delta_j \varphi^{(p)}(x_j) \Big)^T \varphi^{(p)}(x^*) + b \tag{35}$$

$$= (\alpha^T D + \beta^T R) \sum_{p=1}^{d} K_n^{(p)}(x^*) + b\,, \tag{36}$$

with $K_n^{(p)}(x^*, x) = [K^{(p)}(x_1, x^*) \ \ldots \ K^{(p)}(x_n, x^*)]^T$ and $K^{(p)}(x, z) = \varphi^{(p)}(x)^T \varphi^{(p)}(z)$.

References

[1] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* Wiley: New York, 2002.

[2] Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* 1972; **34**(2):187–220.

[3] Hastie T, Tibshirani R. *Generalized additive models.* Chapman and Hall: New York, 1990.

[4] O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* 1988; **9**:531–542. DOI:10.1137/0909035.

[5] Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute* 1988; **80**(15):1198–1202. DOI:10.1093/jnci/80.15.1198

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

[7] Li H, Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics* 2005; **21**(10):2403–2409. DOI:10.1093/bioinformatics/bti324

[8] Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing* 2003; **8**:65–76. DOI:10.1142/9789812776303_0007

[9] Huang JZ, Kooperberg C, Stone CJ, Truang YK. Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics* 2000; **28**(4):961–999. DOI:10.1214/aos/1015956704

[10] Faraggi D, Simon R. A neural network model for survival data. *Statistics in Medicine* 1995; **14**(1):73–82. DOI:10.1002/sim.4780140108

[11] Bakker B, Heskes T, Neijt J, Kappen B. Improving Cox survival analysis with a neural-bayesian approach. *Statistics in Medicine* 2004; **23**(19):2989–3012. DOI:10.1002/sim.1904

[12] Biganzoli E, Boracchi P, Mariani L, 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. DOI:10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO;2-D

[13] Van Belle V, Pelckmans K, Suykens JAK, Van Huffel S. Support Vector Machines for Survival Analysis. *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK) 2007; 1–8.

[14] Van Belle V, Pelckmans K, Suykens JAK, Van Huffel S. Survival SVM: a Practical Scalable Algorithm. *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN2008)*, Bruges (Belgium) 2008; 89–94.

[15] Suykens JAK and Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters* 1999; **9**(3):293–300. DOI:10.1023/A:1018628609742

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

[17] Vapnik V. *Statistical Learning Theory.* Wiley: New York, 1998.

[18] Boyd S, Vandenberghe L. *Convex optimization.* Cambridge University Press, 2004. DOI:10.2277/0521833787

[19] Gunn SR, Kandola JS. Structural modelling with sparse kernels. *Machine Learning* 2002; **48**:137–163. DOI:10.1023/A:1013903804720

[20] Pelckmans K, Goethals I, De Brabanter J, Suykens JAK, De Moor B. *Componentwise Least Squares Support Vector Machines*, chapter in Support Vector Machines: Theory and Applications, (Wang L, ed.), Springer: Berlin, 2005; 77–98. DOI:10.1007/10984697_3

[21] Harrell F Jr, Klee K, Califf R, Pryor D, Rosati R. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**(2):143–152. DOI:10.1002/sim.4780030207

[22] Powell JL. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 1984; **25**(3):303–325.

[23] Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine* 1997; **16**(4):385–395.DOI:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

[24] Park MY, Hastie T. L1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: series B* 2007; **19**(4):659–677. DOI:10.1111/j.1467-9868.2007.00607.x

[25] Verweij PJM, van houwelingen HC. Penalized likelihood in Cox regression. *Statistics in Medicine* 1994; **13**(23-24):2427–2436. DOI:10.1002/sim.4780132307

[26] Heagerty P, Lumley T, Pepe M. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**(2):337–344. DOI:10.1111/j.0006-341X.2000.00337.x

[27] Sørlie T, Tibshirani R, Parker J, Hastie T, Marron J S, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale A, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**(14):8418–8423. DOI:10.1073/pnas.0932692100

[28] Hastie T, Tibshirani R, Friedman J. *The elements of Statistical Learning.* Springer: New York, 2009. DOI:10.1007/b94608

[29] Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology* 2004; **2**(4):e108+. DOI:10.1371/journal.pbio.0020108

[30] Bair E, Hastie T, Debashis P, Tibshirani R. Prediction by supervised principal components. *Journal of the American Statistical Association* 2006; **101**:119–137. DOI:10.1198/016214505000000628

[31] Nygård S, Borgan O, Lingjærde O, Størvold H. Partial least squares cox regression for genome-wide data. *Lifetime Data Anal* 2008; **14**(2):179–195. DOI:10.1007/s10985-007-9076-7

[32] Schumacher M, Basert G, Bojar H, Huebner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RLA, and Rauschecker HF. 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.

[33] Galea M, Blamey R, Elston C, and Ellis I. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment* 1992; **22**(3):207–219. DOI:10.1007/BF01840834

[34] Brouckaert O, Camerlynck E, Van Belle V, Van Huffel S, Pintens S, Amant F, Leunen K, Smeets A, Berteloot P, Van Limbergen E, Weltens C, Van den Bogaert W, Paridaens R, Vergote I, Christiaens MR, Wildiers H, Neven P. Biology and prognosis by age of primary operable breast cancer. *Cancer Research* 2009; **69**(2):181.

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

[36] Fisher B, Bauer M, Wickerham L, Redmong CK, and Fisher ER. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer* 1983; **52**(9):1551-1557. DOI: 10.1002/1097-0142(19831101)52:9<1551::AID-CNCR2820520902>3.0.CO;2-3

[37] Lamy PJ, Pujol P, Thezenas S, Kramar A, Rouanet P, Guilleux F, 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**(1):65–71. DOI:10.1023/A:1020228620173

[38] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**(282):457–481.