

KATHOLIEKE UNIVERSITEIT
LEUVEN

Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Engineering
Department of Electrical Engineering (ESAT)

Non-linear survival models and their applications within breast cancer prognosis

Vanya VAN BELLE

Dissertation presented in partial
fulfillment of the requirements for
the degree of Doctor
in Engineering

December 2010

Non-linear survival models and their applications within breast cancer prognosis

Vanya VAN BELLE

Jury:
Prof. dr. ir. H. Hens, chairman
Prof. dr. ir. S. Van Huffel, promotor
Prof. dr. ir. J.A.K. Suykens, promotor
Prof. dr. P. Neven
Prof. dr. D. Timmerman
Prof. dr. M. Diehl
Prof. dr. A.M. De Meyer
Prof. dr. P. Lisboa (L.J.M.U.)
dr. K. Pelckmans (U.U.)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

December 2010

© Katholieke Universiteit Leuven – Faculty of Electrical Engineering
Address, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2010/7515/132
ISBN 978-94-6018-293-8

Voorwoord

Je was acht toen je het nieuws te horen kreeg. Ik wist toen nog niet wat het betekende. Twee jaar later werd het duidelijk. Een ongekende pijn borrelde in me op, een pijn die nog erger moest zijn voor je mama, je papa en mensen die dichter bij je stonden dan ik. Langzaamaan groeide mijn boosheid gericht naar de dokters die je niet hadden kunnen helpen. Ze hadden hun werk niet goed gedaan. Mijn besluit stond dan ook vast. Ik zou dokter worden en ik zou de fouten die jouw dokters gemaakt hadden nooit maken. Mijn patienten zouden genezen!

Het duurde twee jaar voor het besef kwam dat er geen schuldigen waren aan jouw vroegtijdige heengaan. De dokters hadden gedaan wat ze konden, maar dat bleek niet voldoende. Ik wou niet langer dokter worden, niet geconfronteerd worden met het lijden van mensen dat ik niet kan verzachten. Ik zou ingenieur worden en dokters helpen de mensen te helpen, meewerken aan het ontwikkelen van nieuwe scanners die tumoren vroeger zouden detecteren, meewerken aan studies om de onderliggende mechanismen beter te leren kennen. Dat was het, dat zou ik doen.

Ondertussen zijn we vele jaren later. Ik ben ingenieur en heb het pad gevolgd dat ik op mijn twaalfde voor mezelf had uitgestippeld. Maar dit was me niet gelukt zonder de steun van mijn ouders, die me mijn weg hebben laten gaan, ook op die momenten dat ze dachten dat het anders beter was geweest.

Mijn onderzoek situeert zich op de rand van een heel aantal domeinen. Het eindresultaat is een geheel geworden dat de brug slaat tussen mathematische modellering en klinische toepasbaarheid. Uiteraard zou dit resultaat er niet geweest zijn zonder de nauwe samenwerking met specialisten uit de verschillende domeinen. Ik zou dan ook graag Sabine willen bedanken, die mij het domein van de biomedische ingenieurstechnieken heeft leren kennen. Ze heeft mij de kans gegeven om mij te verdiepen in de classificatie van hersentumoren in mijn laatste ingenieursjaar. Al snel bleek dat ze mij ook de kans wou geven om mezelf verder te ontwikkelen binnen het domein van wiskundige ingenieurstechnieken, toegepast op kankerprognose. De grote moeilijkheid lag hier in het opstellen van nieuwe wiskundige methodes voor een betere modellering van overlevingsanalyse. Het was niet gelukt om deze uitdaging aan te gaan zonder de hulp van Kristiaan en Johan

S., die altijd klaar stonden voor een discussie. Kristiaan wil ik ook extra bedanken om te zetelen in de jury. Ook Moritz wil ik bedanken voor zijn enthousiasme in het organiseren van talrijke OPTEC meetings, waar optimisatieproblemen met een allegaartje aan toepassingen werden besproken en geleid hebben tot een mooie samenwerking met professor S. Boyd. Doordat mijn onderzoek zich op de rand van verschillende onderzoeksdomeinen bevindt, was ik ook betrokken bij de begeleiding van verschillende thesissen in de statistiek. Hierbij discussieerde ik onder andere met professoren An Carbonez en Anne-Marie De Meyer over wat er al dan niet van studenten verwacht wordt. Ik wil professor De Meyer ook bedanken om te zetelen in mijn jury. I would also like to thank professor Paulo Lisboa. His research solves medical and financial questions by means of complex mathematical models. I enjoyed to contribute to several special issues on artificial intelligence in medicine organized by professor Lisboa and Alfredo Vellido. In addition, I would like to thank Paulo for accepting the invitation to be member of my jury.

Naast al deze mensen uit de meer technische wereld, wil ik ook nog alle contacten met de klinische wereld vermelden. Ik denk hierbij aan Patrick, Dirk, Tom, Cecilia, Emma, Aris, Olivier, Saskia, Isabelle, Cho, Inge, Anne-Sophie, en vele anderen met wie ik vele malen gediscussieerd heb over het opstellen van studies, statistische analyses en het interpreteren van de resultaten. I would also like to thank Anne-Lise Bϕrrensen-Dale, Bjϕrn Naume, Paula Murray, Gro Wiedswang and Vernon Harvey for the nice cooperation. In particular, I would like to thank professor Harvey for noting that we forgot to thank a very important group of people in our manuscript, namely the patients. Having said this: Thanks to all the patients who participated to the studies in which I was involved. Without the consent of patients, it would not be possible to improve daily clinical practice. Uiteraard wil ik ook alle collega's bedanken. Een hele dikke merci aan mijn vaste bureaugenootjes, Ben en Jan, en de andere collega's, Steven, Alex, Bogdan, Katrien, Maarten, Wouter, Diana, Mariya, Maria, Anca, Bori, Thijs, Anne-Sophie, Kris, Joachim, Ivan, Vladimir, Lieveke ... Naast mijn onderzoek was ik ook betrokken bij het vak Toegepaste Algebra in de eerste Bachelor Burgerlijk Ingenieur. Ik wil dan ook professor Joos Vandewalle bedanken voor de leuke ervaring die ik opgedaan heb tijdens het geven van oefenzittingen. Natuurlijk mogen ook de administratieve krachten niet ontbreken in dit bedankingslijstje. Lut, Ida, Ilse, John: heel erg bedankt om mij op tijd en stond van koffiekaarten en reisaanvragen te voorzien! Liesbeth en Maarten hun inzet om computerproblemen op te lossen heb ik ook zeer geapprecieerd! Uiteraard is ook een bedankje aan het IWT voor de financiële steun op zijn plaats.

Johan V.R., mama, Eddy en Nelly wil ik bedanken voor de steun en het inspringen waar ze maar konden tijdens de hectische tijd die samen gaat met het schrijven van een doctoraat.

Tot slot wil ik nog twee mensen in herinnering brengen zonder wie de applicatie van dit onderzoek niet binnen kankeronderzoek had gelegen en het onderwerp niet

overlevingsanalyse was geweest. Liesbeth en *mijnen dikke kameraad*, dit boekje is voor jullie!

Vanya

Leuven
December 2010

Abstract

Everybody will be confronted with the diagnosis of cancer during their life. If not personally, it might be a family member or a friend. The diagnosis of cancer creates a feeling of fear. Fear for the unknown, fear for the therapy, fear to die. Thanks to the knowledge of clinical doctors the patients can be informed on their disease and can be treated adequately. In order to decide on treatment options, estimation of risk of survival is crucial. Although chemotherapy might reduce the risk on metastatic tumors, the therapy in itself might be harmful. Deciding which patients might benefit from a specific type of therapy in a way that the benefit is higher than the negative effects of the therapy, is of major importance in clinical practice. These questions are answered by means of survival analysis. By studying patient variables and their relation with survival time (time to relapse, time to death, time to healing) an estimate of the patient specific risk can be calculated. In addition, this type of models allows to estimate the effect of different types of treatment. The information which is provided to clinicians in this way can help them making a more adequate treatment choice.

Survival data are most often analyzed by parametric and semi-parametric models. Although these models perform well in general, some drawbacks remain. The most popular model for survival analysis is the proportional hazards model (PH). The name of this model reveals that hazards for observations with different variable values are assumed to be proportional. However, this assumption is not always realistic. Secondly, the model assumes a linear parametric form of the variables.

To overcome these drawbacks, this research presents a new mathematical model for the analysis of survival data. The largest threshold in developing survival models is the occurrence of censoring in the data. Censoring indicates that not all survival times are observed exactly. Some patients might not experience the event under study during the study time, others might be lost to follow-up. To deal with this issue, the survival problem is reformulated as a ranking problem, where only pairs of observations for which the event order is known, are taken into account. This principle is combined with the methodology of support vector machines. However, the maximal margin principle is replaced by minimizing the Lipschitz constant.

This model is further explored by considering different adaptations, such as L_2 instead of L_1 norms and regression constraints instead of ranking constraints. This research revealed that the use of equality constraints is less appropriate in analyzing survival data since the information provided by censored data can not completely be included. In addition, it is shown that the inclusion of regression constraints improves the performance significantly.

A second achievement of this research is that the proposed method appears to be very promising when handling high dimensional data. The last decade, more and more research was done with regard to genomics, proteomics and other "omics", resulting in high dimensional data. The problem with conventional techniques is that the estimates of the effects of each variable becomes less reliable when more and more variables are included. The presented method overcomes this problem by solving the optimization problem in the dual space.

Although our experiments illustrate that the proposed model results in well performing models, the chance that this model will be used in clinical practice is very small. The problem with the clinical use of non-parametric models is that they are black-box models. As a result, clinicians can not be provided with information on how the model estimates a certain risk, making it hard for clinicians to trust the model. In addition, this type of models can not be used to search for new predictive and prognostic markers since one does not know which variables contribute to the result. These major drawbacks are solved in the interval coded scoring index (ICS). Although this new model starts from the model explained above, the results are highly interpretable and easily applicable in clinical practice. The basic idea of the ICS is that variables can be divided into a relatively small number of intervals, in which the effect on the risk of the event remains the same. In addition, it is assumed that the effects of all variables can be summed up to become the final risk. Classification and prognostic models developed by means of this new methodology can be presented as a questionnaire or by means of color bars, which makes application possible within software applications as well as on paper.

Korte inhoud

Iedereen wordt tijdens zijn leven wel eens geconfronteerd met de diagnose van kanker. Als het niet persoonlijk is, dan gaat het over een familielid of een vriend. De diagnose van kanker creëert bij de patiënt en zijn omgeving een gevoel van angst. Angst voor het onbekende, voor de therapie, angst om de strijd te verliezen. Dankzij de kennis van gespecialiseerde dokters kunnen patiënten geïnformeerd worden over hun ziekte en kunnen ze op een adequate manier behandeld worden. Opdat de juiste therapie gekozen zou worden, is het cruciaal dat dokters een schatting kunnen maken over de overlevingskansen van de patiënt. Hoewel chemotherapie de kans op metastasen kan verminderen, kan de therapie op zich schadelijke gevolgen hebben. Het is dan ook belangrijk dat de negatieve effecten van een specifieke behandeling afgewogen worden ten opzichte van de verwachte positieve effecten. Overlevingsanalyse is de wiskundige techniek die nodig is om dit soort vragen te beantwoorden. Binnen deze techniek worden gegevens van de patiënt en de link met de overlevingstijd (dit kan tijd tot herval, dood of genezing zijn) bestudeerd. Op deze manier is het mogelijk om een schatting te maken van het risico dat elke specifieke patiënt loopt. De informatie die op deze manier aan clinici geleverd kan worden, helpt bij het nemen van beslissingen over de meest geschikte therapie.

Gegevens over overlevingstijden worden meestal geanalyseerd aan de hand van parametrische of semi-parametrische modellen. Deze technieken leiden vaak tot goede performanties, maar toch zijn er ook een aantal nadelen aan verbonden. Het proportionele hazards model (PH) wordt het vaakst gebruikt voor overlevingsanalyse. Zoals de naam al aangeeft, veronderstelt dit model dat de hazard voor verschillende patiënten proportioneel is aan een ongekende basis hazard. Deze veronderstelling is echter niet steeds realistisch. Het model veronderstelt verder een lineaire parametrische vorm in de variabelen.

Dit onderzoek stelt een nieuw model voor dat bovenstaande nadelen niet heeft. Het grootste probleem bij het opstellen van wiskundige modellen voor overlevingsanalyse, bestaat in het inbrengen van gecensureerde data. Censurering duidt hier op het feit dat de exacte overlevingstijd niet gekend is voor een deel

van de observaties. Zo zullen een deel van de patiënten het event niet krijgen tijdens de studie, anderen zullen verhuizen of voortijdig stoppen met de studie. Om deze vorm van ontbrekende data in te brengen in het model, wordt het overlevingsprobleem geformuleerd als een ranking probleem. Hierbij worden enkel paren van patiënten in rekening gebracht voor wie geweten is wie het event eerst ervaart. Dit principe wordt verder gecombineerd met de methodologie van support vector machines. Het principe van de maximale marge wordt hierbij vervangen door het minimalizeren van de Lipschitz constante. Het aldus bekomen model wordt verder verfijnd door verschillende alternatieven te bestuderen. Zo wordt het verschil tussen een L_2 en L_1 norm bestudeerd. Er wordt ook nagegaan wat het verschil is tussen ranking- en regressiebeperkingen. Dit onderzoek heeft aangetoond dat gelijkheidsbeperkingen minder geschikt zijn voor de modellering van survival data aangezien de informatie die vervat zit in de gecensureerde datapunten niet volledig in rekening kan gebracht worden. Bijkomend toont dit onderzoek aan dat regressiebeperkingen de performantie significant verbeteren.

Een voordeel van het ontwikkelde model ten opzichte van de bestaande modellen is dat het veelbelovend lijkt voor hoog-dimensionale data. De laatste jaren is er meer en meer onderzoek verricht naar genomics, proteomics en andere "omics" die resulteren in hoog-dimensionale datasets. De standaard methodes voor overlevingsanalyse hebben problemen met hoog-dimensionale data omdat de schatting van de effecten minder betrouwbaar worden wanneer er meer coefficiënten geschat moeten worden dan er observaties zijn. De voorgestelde methode lost dit probleem op door het optimalisatieprobleem op te lossen in de duale ruimte.

De resultaten van dit onderzoek tonen aan dat het ontwikkelde model kan leiden tot modellen met een hoge performantie. Toch is de kans klein dat dit type van model gebruikt zal worden in de klinische praktijk. Het probleem is dat het hier om een zwarte-doos-model gaat en er dus geen informatie gegeven wordt over hoe het model tot zijn risicoschatting komt. Dit model zou dan ook niet gebruikt kunnen worden om nieuwe prognostische factoren te zoeken aangezien het niet geweten is welke gegevens leiden tot het bekomen resultaat. Deze belangrijke nadelen worden opgelost aan de hand van het interval gecodeerde score (ICS) systeem. De onderliggende idee van ICS is dat variabelen verdeeld kunnen worden in een klein aantal intervallen waarbinnen het effect op het risico gelijk blijft. Er wordt bijkomend verondersteld dat de effecten van alle variabelen opgeteld kunnen worden om het finale risico te bekomen. Modellen die aan de hand van ICS opgesteld worden zijn heel interpreteerbaar en klinisch makkelijk toepasbaar. Classificatie- en prognostische modellen die ontwikkeld zijn via ICS, kunnen voorgesteld worden aan de hand van vragenlijsten of kleurenbalken. Hierdoor is het mogelijk om de modellen toe te passen binnen software pakketten, maar ook om ze te gebruiken aan de hand van een eenvoudige papieren versie.

Nomenclature

List of symbols

$\underline{a}$	lower bound
$\overline{a}$	upper bound
$\alpha, \beta, \eta, \nu, \theta$	Lagrange multipliers
b	bias term
C	concordance
c_i	censoring time for observation i
d	dimension of the input space
δ_i	event indicator
ε	small constant
ϵ,ξ	slack variables
φ	feature map
γ, μ	regularization constant
h	transformation function
I	identity matrix
$\mathcal{I}$	indicator function
k	kernel function
K	kernel matrix
L	partial likelihood
$\mathcal{L}$	Lagrangian
ℓ	Lipschitz constant
λ	hazard function
n	number of observations
n_k	number of classes
R	risk
σ	kernel parameter of the Gaussian kernel
S	survival function
t_i	true failure time for observation i
τ	constant of the polynomial kernel
u	prognostic index or utility
v	vector of threshold values on the utility
w	coefficient vector

x_i	input/variable vector of observation i
χ	weighting factor
y_i	outcome for observation i

List of sets

$\mathcal{D}$	training data set
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^d$	vector of real numbers
$\mathbb{R}^{d\times n}$	matrix of real numbers
$\mathcal{R}$	risk set
$\mathcal{S}$	set of indices
$\mathcal{S}_{(\cdot)}$	set of ordered indices

List of basic operations

x^T	transpose of x
$\|a\|$	absolute value of a
$\|\|\cdot\|\|_p$	p-norm of $\cdot$
$\sum$	sum
$\prod$	product
$\in$	element of
$\|\mathcal{D}\|$	cardinality of $\mathcal{D}$

List of abbreviations

AFT	accelerated failure time model
AI	aromatase inhibitor
ANOVA	analysis of variance
ARD	automatic relevance determination
AUC	area under the ROC curve
CDS	clinical decision support
CDF	cumulative distribution function
CI	confidence interval
CSVR	support vector regression for censored data
DOR	diagnostic odds ratio
EPU	early pregnancy unit
ER	estrogen receptor
EvR	event rate
EXT%	percentage of observations classified into the most extreme risk groups
FISH	fluorescence in situ hybridisation
HER2	human epidermal growth factor receptor 2
HT	hormone therapy
ICS	interval coded score

IHC	immunohistochemistry
i.i.d.	independently and identically distributed
iNPI	improved Nottingham Prognostic Index
IPUVI	intra uterine pregnancy of uncertain viability
KKT	Karush-Kuhn-Tucker
LNPOS	number of positive lymph nodes
LNR	lymph node ratio
log	natural logarithm
LPI	log-odds prognostic index
LR^+	positive likelihood ratio
LR^-	negative likelihood ratio
LS-SVM	least-squares support vector machine
NPI	Nottingham Prognostic Index
ODI	ordinal discrimination index
PCR	PH model with principal component regression
PDI	polytomous discrimination index
PDF	probability density function
PH	proportional hazards model
PLANN	partial logistic artificial neural network
PLANNARD	PLANN with ARD
PLS	PH model with partial least-squares pre-processing
PR	progesterone receptor
QP	quadratic programming problem
R^2_{adj}	adjusted measure of explained variance
RANKSVMC	ranking SVM for censored data
RBF	radial basis function
RMI	risk of malignancy index
ROC	receiver operating characteristic curve
RT	radio therapy
RVM	relevance vector machine
SPCR	PH model with principal component regression used on a subset of the variables
SVM	support vector machine
TM	transformation model

List of data sets

`ABC`	Auckland breast cancer data set
`GBC`	German breast cancer study group data set
`DBCD`	Dutch breast cancer data set
`DLBCL`	diffuse large-B-cell lymphoma data set
`EPD`	early pregnancy unit data set
`IOTA`	International Ovarian Tumor Analysis Group data set
`LBC`	Leuven breast cancer data set

`LCR`	leukemia data set (end point: complete remission)
`LD`	leukemia data set (end point: death)
`MLC`	Maya lung cancer data set
`NSBCD`	Norway/Stanford breast cancer data set
`OBC`	Oslo breast cancer data set
`PC`	prostatic cancer data set
`VLC`	veteran lung cancer data set

Contents

Contents xiii

1 Introduction 1

1.1 An introduction to survival analysis 1

1.2 Breast cancer 4

1.2.1 Predictive and prognostic factors 5

1.2.2 Prognostic models 9

1.3 Aims of the thesis 12

1.4 Chapter-by-chapter overview 12

1.5 Other research 15

1.5.1 Clinical studies 15

1.5.2 Development of discriminatory measures for multi-class classification problems 23

2 A Methodological Primer 25

2.1 Terms and definitions 25

2.2 The history of survival models 26

2.2.1 Non-parametric estimation of the survival function 26

2.2.2 Survival distributions and parametric models 27

2.2.3 The proportional hazards model 29

2.2.4 Partial logistic artificial neural network 31

2.3 Machine learning methods 32
2.3.1 Support vector machines 32
2.3.2 Least-squares support vector machines 36
2.3.3 Model selection 37
2.4 Model evaluation 39
2.4.1 Discrimination 39
2.4.2 Calibration 45
2.5 Conclusion 46

3 Transformation models for ranking and survival analysis 47

3.1 Introduction 47
3.2 Transformation models and ranking methods 51
3.3 A convex approach to learning transformation models 56
3.3.1 The realizable case 56
3.3.2 The agnostic case 58
3.3.3 A non-linear extension using Mercer kernels 58
3.3.4 Toward sparse solutions using L_1 regularization and sparsity constraints 60
3.3.5 Comparison with other methods 61
3.4 Learning for ordinal regression 63
3.4.1 A modification to MINLIP 63
3.4.2 Prediction for ordinal regression 65
3.4.3 Difference with other methods 66
3.5 Transformation models for failure time data 68
3.5.1 Transformation models for survival analysis 68
3.5.2 A modification to MINLIP 69
3.5.3 Prediction with transformation models 70
3.6 Application studies 71

3.6.1 Artificial examples 72
3.6.2 Ordinal regression 74
3.6.3 Failure time data: micro-array studies 76
3.6.4 Failure time data: clinical data 78
3.7 Conclusions 82

4 Support vector methods for survival analysis: optimization problem formulation and comparison 85

4.1 Existing SVM based methodologies for survival analysis 85
4.1.1 Support vector regression for censored data 86
4.1.2 Support vector machines based on ranking constraints . . . 87
4.2 Proposed SVM formulations for survival analysis 90
4.3 Least-squares alternatives 93
4.4 Adaptations for high dimensional data 95
4.5 Results 96
4.5.1 Comparing kernel based methods 97
4.5.2 Comparison with other methods 105
4.6 Conclusions 107

5 Interval coded score index: a flexible model for interpretable clinical decision support 111

5.1 Introduction 111
5.2 Scoring systems for classifiers and prognostic models based on machine learning methods 115
5.2.1 Classification 115
5.2.2 Survival analysis 121
5.3 Real life applications 124
5.3.1 Diagnosis of malignancy of adnexal masses 125
5.3.2 Prediction of non-viable pregnancies 136

5.3.3 Breast cancer prognosis 140
5.4 Conclusions 153

6 Conclusions and future work **159**
6.1 Conclusion 159
6.2 Future work 161

A Data description **165**
A.1 Breast cancer data set from the Leuven University Hospitals 165
A.2 International Ovarian Tumor Analysis Study 168
A.3 The early pregnancy unit 168
A.4 Publicly available data sets 169
A.4.1 Clinical data 169
A.4.2 Micro-Array data 175

B Primal-dual formulations **177**
B.1 MINLIP for ranking problems 177
B.2 MINLIP for ordinal regression 179
B.3 MODEL 1 181
B.4 MODEL 2 182
B.5 MODEL 3 184

Bibliography **187**

Curriculum vitae **201**

Publication list **203**

Chapter 1

Introduction

This Chapter gives a short introduction to breast cancer and illustrates how treatment choices are made depending on predictive factors. The importance of defining prognostic factors and combining them into prognostic models is discussed. The most commonly used prognostic indices are summarized. A description of the goal of this thesis follows. In Section 1.4 an overview of the thesis is given. To keep this work concise, several studies were not included in this dissertation. However, they are discussed shortly in Section 1.5.

1.1 An introduction to survival analysis

The goal of survival analysis (see Parmar & Machin, 1995; Hosmer & Lemeshow, 1999; Therneau & Grambsch, 2000; Kalbfleisch & Prentice, 2002; Armitage & Colton, 2007) is to estimate the risk for a specified event at study to occur in function of the time and to identify characteristics on which the risk depends. The time from zero to the occurrence of the event is called the failure or survival time. Examples of survival studies are:

1. *Cancer prognosis*: In this field one is interested in linking tumor and patient characteristics to the observed survival time in order to be able to treat future patients in a better and more personalized way. Different settings are possible. In primary operable breast cancer patients for example, the main interest lies in finding prognostic factors, indicating the risk for each patient to experience a breast cancer related event. By the identification of prognostic markers, patients can be divided into prognostic risk groups, yielding optimized treatment choices. Two examples illustrate the advantage

of such an approach. First consider a patient with a very low risk to develop a breast cancer related event. In case this would not be known due to the fact that no prognostic markers are available, this patient would receive chemotherapy in order to reduce the risk. Due to the aggressiveness of the therapy, a patient might experience more side-effects than advantages since the therapy in itself can be harmful. Secondly, consider a patient at high risk of experiencing a breast cancer related event. Again, without knowledge on prognostic factors, this patient might be undertreated. Therefore, in order to treat patients in an optimal way, the identification of prognostic indices is of major importance. A second application is to test whether a new treatment is better than the standard of care.

2. *Electronic and mechanical components*: Survival models can be used in order to reduce the overhead cost due to too frequent check-ups of mechanical and electronic components (e.g. engine in cars, aircrafts, ...) and the cost due to missing a defect. Identifying the number of kilometers, cycles, flights, ..., can be undertaken without an increased risk on defects results in an optimal planning of defects, reducing the number of check-ups and the costs.

3. *Economics*: In economics, survival models are used to derive the risk that a client won't be able to pay off his depths. After the monetary crisis of 2008 banks are investigating increasingly in this type of risk modelling strategies. A second application is found in insurance companies, who try to minimize the risk on claims. Yet another example is to estimate the duration of unemployment.

4. *Criminology*: Survival analysis can be used to estimate the time until a released criminal commits a new crime.

5. *Sports*: In sports, survival analysis can be used to find the necessary characteristics to win the Tour the France or to run a marathon.

In most of these examples, the estimate of the failure time is not the primary interest. The goal of survival analysis is most often to identify risk factors such that a negative event can be avoided or a positive event can be favored. Note that although this type of analysis is called *survival* analysis, events are not restricted to death times. The meaning of survival in this sense is that the event under study did not occur.

A major difficulty in analyzing survival data is the occurrence of censoring (see Parmar & Machin, 1995; Hosmer & Lemeshow, 1999; Kalbfleisch & Prentice, 2002; Gijbels, 2010, for a broad introduction on censored data). Censoring is a term indicating the incomplete observation of the survival time. Most methods for survival analysis assume random or non-informative censoring. This assumption indicates that the reason why observations are not observed is not related with the

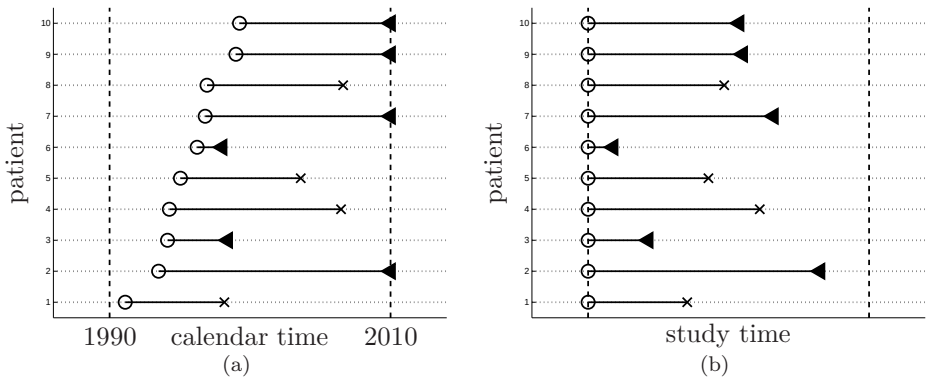


Figure 1.1: *Illustration of a typical study pattern: (a) in calendar time; (b) in study time. (a) From 1990 on, patients are gradually included after diagnosis of primary operable breast cancer and surgery. Patients 1, 4, 5 and 8 experience a breast cancer related event during the study. Hence, their failure times are known exactly and thus uncensored. Patients 2, 7, 9 and 10 survive until the end of the study. Their failure time is right censored. The remaining patients drop out during the study. Possible reasons are that the patient has moved, or had another event, independent of the event of interest in this study, making follow-up for this study impossible (e.g. the patient died in a car accident). In panel (b), the same study is visualized in the perspective of the study time instead of the calendar time. A circle indicates the inclusion time, a cross the event time and a left-pointing triangle a right censored failure time.*

risk for the event to happen. The most common censoring type is *right censoring*. This type of censoring occurs in cases where the event time is not known, but a lower bound on this time is known. One example is found in breast cancer studies. Consider a study to evaluate a new type of chemotherapy. Patients are included in the study from 1990 to 1995 and followed until 2010. Starting from 1990, patients diagnosed with primary operable breast cancer are randomized in this study from the moment of surgery and followed until 2010 or the occurrence of a breast cancer related event. Figure 1.1 illustrates a typical study pattern. Every patient is included at its own time of surgery (time zero of the study). Some patients experience the event under study and their survival time is uncensored. Other patients survive until the end of the study and have a right censored survival time. Some patients will be lost of follow-up or withdrawn during the study period. This happens when patients move and are no longer able to follow the study program or when another event happens, making observation of the event at study impossible (e.g. patients who die in car accidents).

A second censoring mechanism, is *left censoring*. In this case an upper bound of the failure time is known. A well known example of this type of censoring is

published by Wagner & Altmann (1973). In the seventies, scientists were interested to know at which time baboons come down from the trees to beverage. Therefore, the scientists came to the trees at a certain time of the day and observed the animals until halve of the troop was down from the trees. This time represented the survival time. However, at some days it happened that half of the troop was already down before the scientists arrived and the failure time was left censored. Other examples can be found in studies investigating at which age children can talk, count or walk. Some children will already have these abilities before the study starts and their survival time will be left censored.

The last censoring type that will be discussed here is *interval censoring*, in which case the failure time is not observed exactly, but an interval in which the event occurs is known (Lindsey & Ryan, 1998; Gómez *et al.*, 2009). Interval censoring typically occurs in clinical studies where patients are seen at regular check-up times. Although several methods are available to handle interval-censored data, most studies assume the failure time to be in the beginning, middle or end of the interval and use standard survival methods for uncensored and right censored data. In the remainder of this text, it will be assumed that the data only contain uncensored and right censored data, unless stated otherwise. More information on censoring can be found in Elandt-Johnson & Johnson (1980); Miller (1981); Andersen *et al.* (1993); Harrell (2001); Kalbfleisch & Prentice (2002).

1.2 Breast cancer

Breast cancer is the most frequent malignancy among Western women (35% of all cancers in Belgian women in 2005) (Ferlay *et al.*, 2007). In the year 2005, 9405 new breast cancer cases were registered in Belgium. Although treatment of breast cancer has evolved drastically the last decades (see for example Blamey *et al.*, 2007), it is still the leading cause of death by cancer in females (20.6% of all cancer deaths)[1]. One of the biggest challenges in managing breast cancer is the heterogeneity of the disease. Patients with the same stage may respond differently on therapies and may have different life expectancies. In order to provide patients with the best treatment, predictive and prognostic factors have been identified. In the remainder of this Section, a short introduction on these factors is given. This text does not aim to give an overview of different breast cancer types, subgroups, nor treatments. Only primary operable invasive breast cancer will be discussed, since this is the population which was most studied in this work.

[1]Numbers taken from *Cancer Incidence in Belgium, 2004-2005*, Belgian Cancer Registry, Brussels 2008.

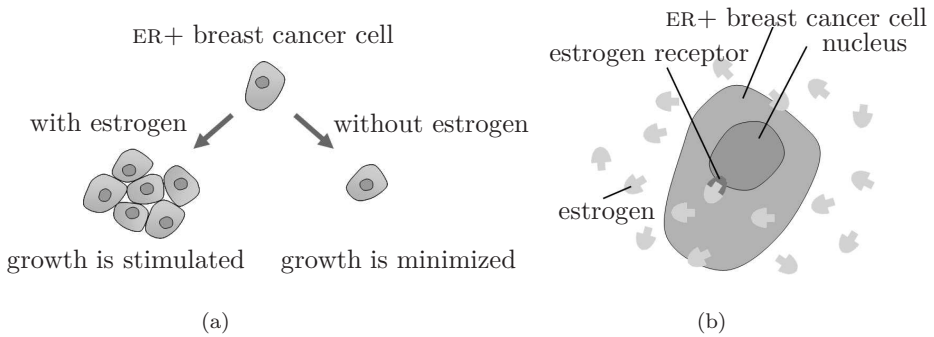


Figure 1.2: *(a) In the presence of estrogen, the estrogen receptor will bind with the estrogen hormone. (b) Together with environmental factors, this will induce the co-activator to bind with the formed complex and transcription will start. As a result the cells will get the signal to divide and the tumor grows. In the presence of tamoxifen (an anti-estrogen drug), tamoxifen metabolites will compete with estrogen to bind the estrogen receptor. This reduction in estrogen and estrogen receptor complexes will inhibit the increase in cell proliferation. Picture adapted from* `http://healthinfoispower.wordpress.com`.

1.2.1 Predictive and prognostic factors

Different biological markers have been found which contribute to the reactiveness to different therapies (predictive factors) and to the prognosis regarding relapse and/or mortality (prognostic factors). Cianfrocca & Goldstein (2004) defines a *predictive factor* as a measurement associated with response to a given therapy and a *prognostic factor* as a measurement available at the time of diagnosis or surgery that correlates with disease-free or overall survival. Two predictive factors are the estrogen (ER) and the human epidermal growth factor receptor 2 (HER2). When the tumor expresses one of these receptors, an appropriate therapy can be given. Without expression of such receptors, therapy is restricted to chemotherapy. Figure 1.2a illustrates the effect of a tumor with estrogen expression. When cells contain estrogen receptors, estrogen will bind with the estrogen receptor and induce cell proliferation after binding of the co-activator (see Figure 1.2b). Since ER+ tumors need estrogen to bind to the estrogen receptor in the nucleus, it was postulated that blocking the action of estrogen might treat this type of breast cancer. Anti-hormonal therapy (HT) exists in introducing drugs which bind with the estrogen receptor, inhibiting the co-activator to bind to the receptor-ligand complex and transcription to start (Shiau *et al.*, 1998). The progesterone receptor (PR) is also a predictive marker since the response to hormone therapy is better in tumors expressing PR (Bardou *et al.*, 2003). Over-expression of HER2, for example by HER2 gene amplification, leads to increased expression of the HER2

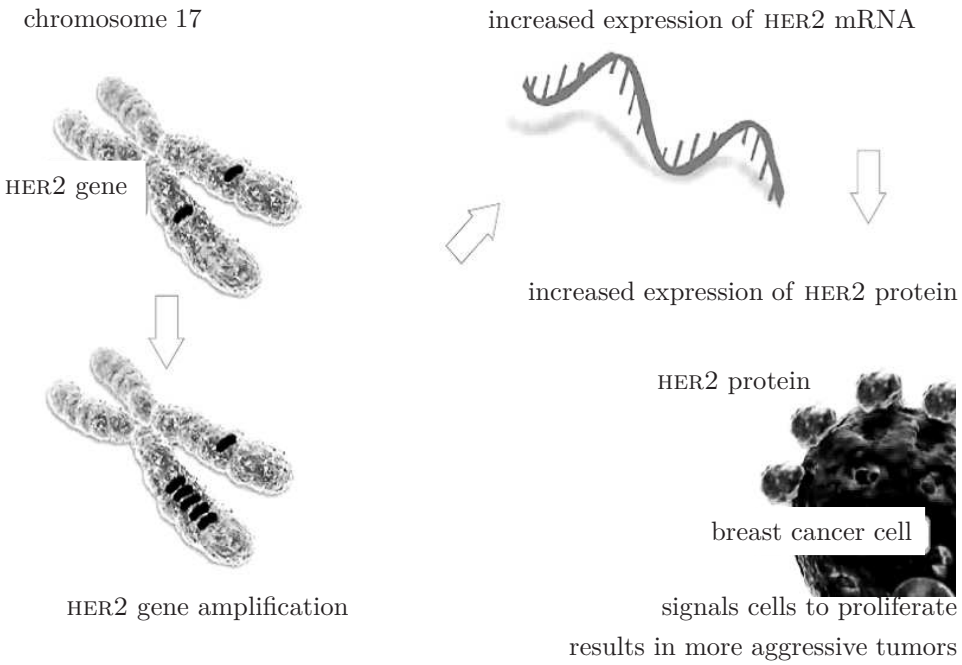


Figure 1.3: *Illustration of the effect of* HER2 *gene amplification in breast cancer.* HER2 *positive breast cancer cells have too many copies of the* HER2 *receptor on their cell membrane. The primary cause of this over-expression of* HER2 *in breast cancer is an increase in copy number of the* HER2 *gene on chromosome 17, which is called gene amplification.* HER2 *gene amplification leads to an increased expression of* HER2 *mRNA and an increased presence of* HER2 *receptors on the cell membrane. As a result, these cells grow faster and divide more than healthy cells.*

protein, which again signals cells to proliferate (see Figure 1.3).

In a similar perspective as for the hormonal treatment, a therapy targeting HER2+ tumors was developed. The different therapeutic options are visualized in a very simplistic way in Figure 1.4.

Where predictive factors are valuable in deciding which therapy will be efficient for a specific patient, prognostic factors are important in defining the risk for the patient to relapse after surgery. A prognostic factor is defined as a measurement which is available at the time of diagnosis or surgery, which is associated with disease relapse or mortality. Identification of prognostic factors leads to a better classification of patients into risk groups and thus more appropriate treatment decisions. Patients with aggressive tumors are more likely to develop new tumors than patients with indolent tumors, and therefore need more aggressive treatment. Additionally, patients with indolent tumors might avoid the risks of unnecessary

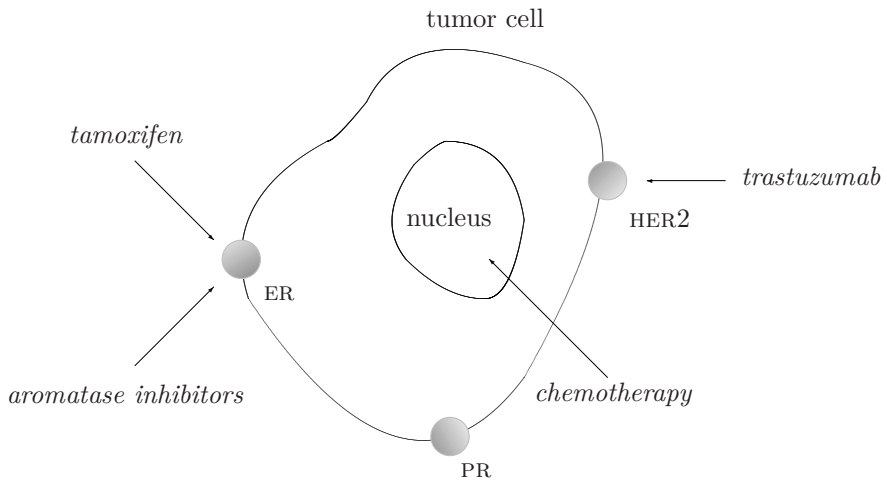


Figure 1.4: *Illustration of treatment options in primary operable breast cancer. The appropriate treatment for breast cancer depends on several characteristics of the tumor cells. In* ER *and/or* PR *positive tumors, the tumor can be targeted with hormone therapy, typically tamoxifen or aromatase inhibitors. In* HER2 *positive tumors, transcription of* HER2 *mRNA is blocked by binding of trastuzumab with the* HER2 *receptor. The nuclei of cells are targeted by chemotherapy in order to kill the tumor cells.*

treatment. The age at diagnosis, life style, race, location of the tumor, among others, are recognized as prognostic factors in breast cancer. However, the prognostic factors discussed below are the ones which are most often used and recognized as important prognostic factors.

A first prognostic factor is the size of the tumor (Carter *et al.*, 1989; Rampaul *et al.*, 2001; Verschraegen *et al.*, 2005). The larger the tumor at diagnosis, the worse the prognosis. Although tumor size was considered as an independent prognostic factor, its prognostic value diminishes in a multivariate model including the number of positive lymph nodes. Lymph node involvement indicates whether the tumor has spread into the lymph nodes in the armpit and is considered the most important prognostic factor for breast cancer prognosis (Fisher *et al.*, 1983). Figure 1.5 illustrates how tumor cells can invade the lymphatic system and follow the lymph drainage to spread towards the lymph nodes. The prognostic value of the presence of tumor cells within the lymph nodes comes from an increased risk of tumor cells travelling to other parts of the body through the lymphatic system,

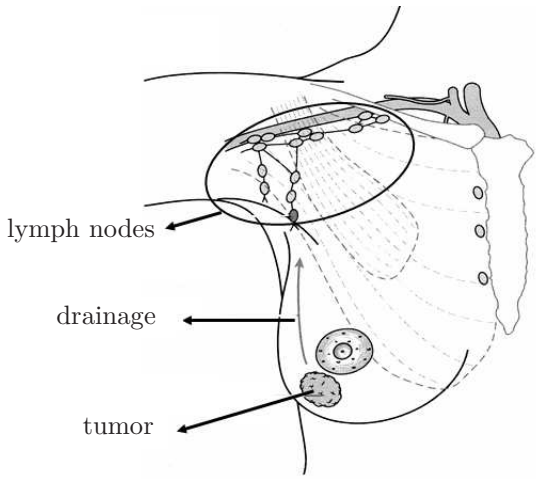


Figure 1.5: *Illustration of the spread of tumor cells towards the lymph nodes via the lymphatic system. Tumor cells are able to travel from the tumor to other parts of the body to form metastases. A marker for this possible spread of the tumor are lymph nodes in the armpit. Picture adapted from* `http://www.patientpictures.com`

resulting in metastases. Prognosis also worsens with an increasing number of positive nodes. The last decade, several authors have pointed out that the number of examined nodes can give additional information to the number of positive nodes (Vinh-Hung *et al.*, 2004, 2006; Woodward *et al.*, 2006; Vinh-Hung *et al.*, 2009; Van Belle *et al.*, 2009c; Danko *et al.*, 2010).

Another very important prognostic factor is the histological grade of the tumor as described in Elston & Ellis (1991). The tumor grade is a system classifying tumors according to how well tumor cells resemble normal cells. The grade is composed out of three factors: (i) the mitotic count, (ii) the nuclear pleomorphism and (iii) the tubule formation. Each of these contributions are given 1 to 3 points. Depending on the sum, the tumor is grade I (sum 3 to 5), grade II (sum 6 to 7) or grade III (sum 8 or 9). Although many papers report on the prognostic value of the histological grade (Rampaul *et al.*, 2001; Kakha *et al.*, 2008), it is argued that the mitotic count in itself is as prognostic as the grading system.

Apart from the prognostic factors discussed so far, the predictive factors are also prognostic factors. HER2 positive breast cancers are known to be more aggressive (Ménard *et al.*, 2001; Suen & Chow, 2006; Van Belle *et al.*, 2010d). In a comparison of patients with and without hormonal treatment, Bardou *et al.* (2003) showed that ER was a prognostic factor in patients without hormonal therapy. In the same

group of patients, PR was not significantly altering survival. However, in the group of patients treated with hormonal therapy, PR was an independent prognostic factor. More information on predictive and prognostic factors in breast cancer can be found in Fitzgibbons *et al.* (2000); Bundred (2001); Rampaul *et al.* (2001).

1.2.2 Prognostic models

Identification of several prognostic factors has triggered researchers to look for combinations of such factors which are able to better classify patients into risk groups than one single factor. Combinations of several prognostic factors are called prognostic models or prognostic indices.

pTNM stage

The pTNM stage (Singlatary *et al.*, 2002) is a prognostic index for breast cancer relapse and survival proposed by the American Joint Committee on Cancer (AJCC). It involves the size of the tumor (T), the number of positive lymph nodes (N) and the presence of metastatic deposits (M).

The Nottingham prognostic index

The prognostic model which is most often used in primary operable breast cancers is the Nottingham Prognostic Index (NPI) (Haybittle *et al.*, 1982; Galea *et al.*, 1992; Blamey *et al.*, 2007). The NPI combines tumor size, tumor grade and the number of positive nodes as follows:

$$\text{NPI} = 0.2 \times \text{tumor size (cm)} + \text{tumor grade} + \text{pN} \tag{1.1}$$

where pN is a categorized version of the number of positive lymph nodes. pN equals 1, 2 or 3, for no lymph node involvement, 1 to 3 positive nodes and more than 3 positive nodes, respectively. According to the value of the NPI, patients are classified into three different risk groups: (i) patients with an NPI less than 3.4 have a low risk for relapse, (ii) a score of 3.4 to 5.4 classifies patients into the intermediate risk group and (iii) a score larger than 5.4 categorizes patients in the high risk group. Figure 1.6 shows the survival curves according to the different NPI risk groups, for a cohort of patients treated at the University Hospitals Leuven, between January 2000 to June 2005 and follow-up until September 2009 (LBC). See Appendix A.1 for a detailed description of the data.

Collett *et al.* (1998); Hlupić *et al.* (2004); Suen & Chow (2006) indicated that the NPI might be improved by inclusion of steroid and HER2 receptors, respectively. In Van Belle *et al.* (2010d) it was shown that within the intermediate and high risk

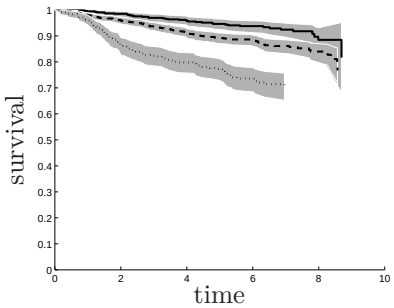


Figure 1.6: *Survival curves according to the* NPI *risk groups in the breast cancer data set of Leuven (see Appendix A.1): solid line, low risk group; dashed line, intermediate risk group; dotted line, high risk group. The grey areas represent the 95% confidence bands. The three survival curves are significantly different.*

NPI groups, patients with different expression of ER, PR or HER2 had significantly different survival characteristics (see Figure 1.7). The NPI risk groups classification can therefore be improved by inclusion of PR and HER2 (see Van Belle *et al.*, 2010d, for more information).

Adjuvant Online

Ravdin *et al.* (2001) proposed the use of a computer programme, called Adjuvant! Online[2], to obtain information on the effect of different therapies on disease free and overall survival. The treatment recommendations for patients with stage I or II breast cancer are based on the estimated risk of relapse and death obtained from clinicians, and the likely benefit of adjuvant therapy. The standard version incorporates the tumor stage and characteristics of the tumor, together with the expected effect of treatment according to randomized trials. The method is validated by Olivotto *et al.* (2005); Campbell *et al.* (2009).

Gene signatures

More recently, research is done in order to detect genes responsible to good or poor breast cancer prognosis (van 't Veer *et al.*, 2002; van de Vijver *et al.*, 2002; Wang *et al.*, 2005; Buyse *et al.*, 2006; Bueno-de-Mesquita *et al.*, 2007; Liu *et al.*, 2007; Martin *et al.*, 2008). Although some of these prognostic models are commercially available, their improved performance over clinical factors is still to be proven (Neven *et al.*, 2008).

[2] `http://www.adjuvantonline.com/index.jsp`

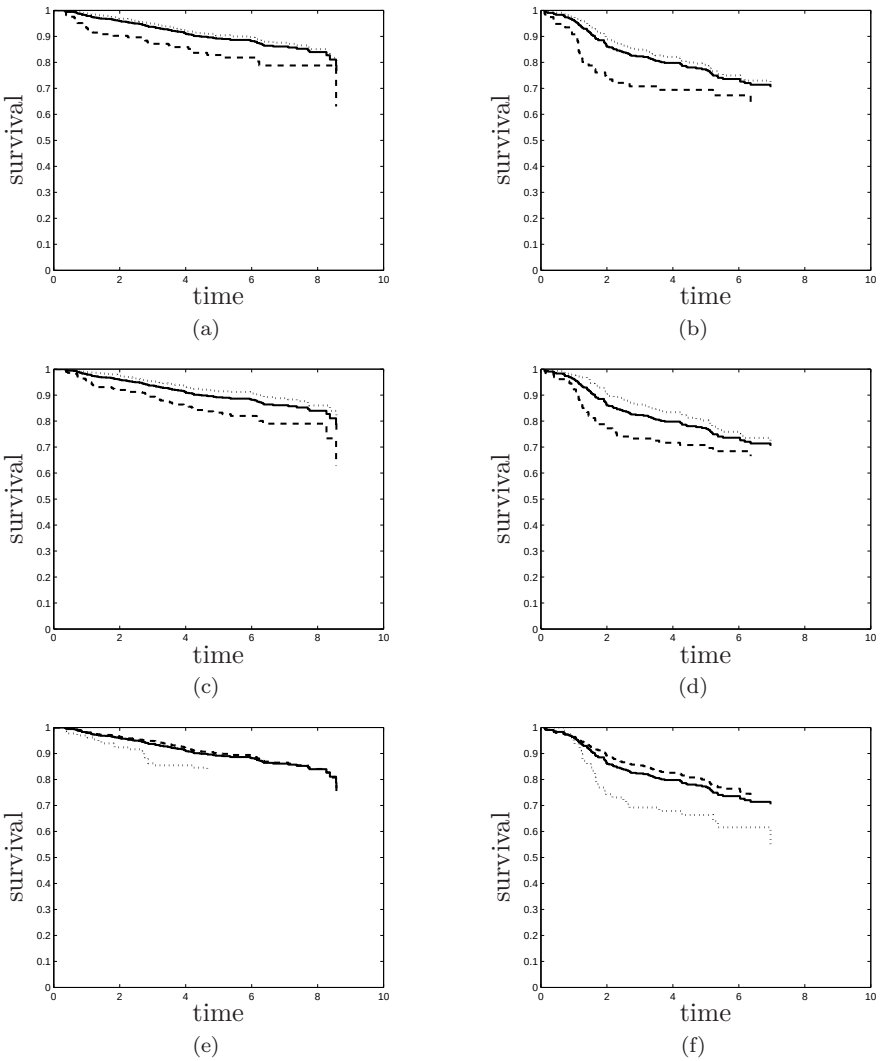


Figure 1.7: *Survival curves within the* NPI *intermediate (a-c-e) and high (b-d-f) risk groups, stratified for* ER *(a-b),* PR *(c-d) and* HER2 *(e-f), in the breast cancer data set of Leuven (see Appendix A.1 for a broad description of the data): solid line,* NPI *risk group; dashed line, negative receptor; dotted line, positive receptor. Within the* NPI *risk groups,* ER/PR *positivity improved disease free survival, while* ER/PR *negativity worsened disease free survival.* HER2 *positivity shortens survival, while* HER2 *negativity improves survival.*

1.3 Aims of the thesis

The aim of this dissertation is to develop a new mathematical framework for the analysis of survival data. The approach taken here differs from the classical statistical approach in the sense that this research will not try to estimate parameters of a predefined model. Figure 1.8 illustrates that the approach consists in reformulating the survival problem as a convex optimization problem. In contrast with standard survival models, the maximal (partial) likelihood (Cox, 1972; Kalbfleisch & Prentice, 2002), will not be used. Instead the methodology behind support vector machines (SVM) (see Vapnik, 1998; Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004) and least-squares support vector machines (LS-SVM) (Suykens & Vandewalle, 1999; Suykens *et al.*, 2002) will be extended to handle censored data. These methods are known for their good generalization on unseen data. However, as they produce black-box models (see Figure 1.9), they do not provide information on how a specific prediction is obtained.

Several points of attention are important in order to obtain the desired result. In first instance the survival problem is reformulated as a ranking problem, taking care of the censored nature of survival data. Although several papers handle survival problems as classification problems by considering one specific point in time (e.g. five years) and consider a group of survivors (low risk) versus a group of non-survivors (high risk) (Michiels *et al.*, 2005; Sanchez-Carbayo *et al.*, 2006; Gevaert *et al.*, 2006), this approach is not generally applicable (see Green & Symons, 1983; Callas *et al.*, 1998, and references). Another approach is to drop the censored cases. However, this would result in underestimated survival estimates. In order to incorporate all the available information, the concordance index (c-index) proposed by Harrell *et al.* (1984); Harrell (2001) will be optimized. In a next step, a more theoretical basis of the model is developed by linking the resulting method with the maximal margin principle lying on the basis of SVMs. The problem formulation is optimized by comparing different norms, regularization schemes and ranking versus regression constraints. Since SVM-based models are known to deal well with the curse of dimensionality, the developed models are applied to high-dimensional data. In addition, it is investigated whether sparsity in the number of variables can be obtained. Finally, a new approach bridging the gap between mathematical modelling and clinical application is proposed. This method combines the flexibility of sophisticated mathematical models with the interpretability and applicability of simple scoring systems.

1.4 Chapter-by-chapter overview

Figure 1.10 gives an overview of the different chapters in this dissertation. The central problem is to find a mathematical formulation for the survival problem.

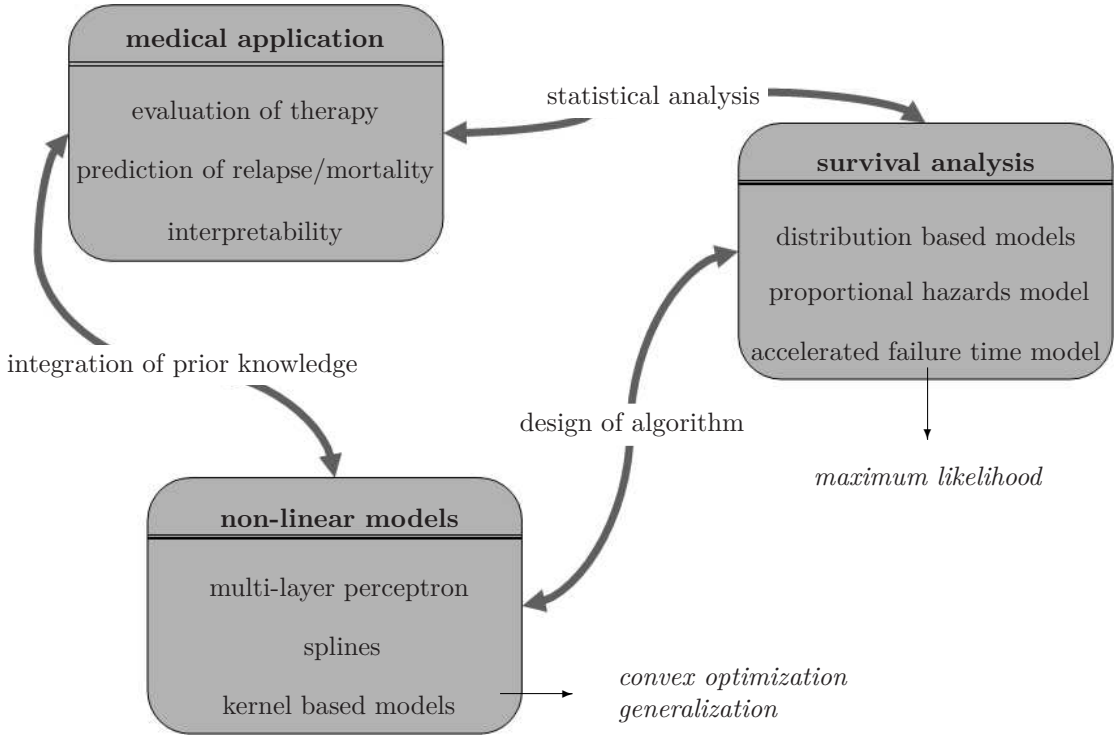
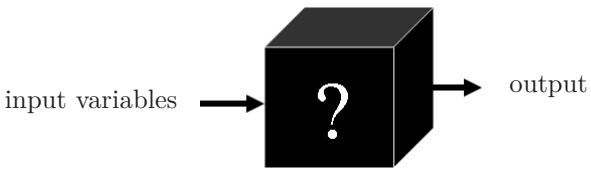
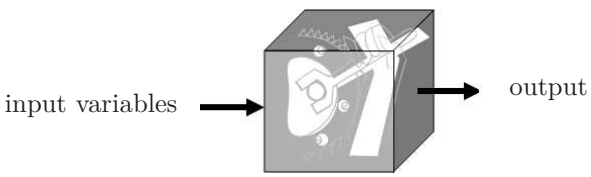


Figure 1.8: *Illustration of the aim of the thesis. A new prediction framework will be developed with attention for generalization properties. In contrast to the standard statistical approach, which assumes a (semi-)parametric model and estimates the parameters of this model, the model structure will not be fixed in advance. To reach this goal, support vector machines and kernel functions will form the basis of the new methodology. Within this modelling process, attention will be given to the clinical use of the resulting model.*



(a) black box



(b) opening the black box

Figure 1.9: *Standard machine learning methods generate black box models as in (a). The user knows which variables need to be provided and knows the result given by the model, but no information is given on how the output is constructed from the given variables. In order to make this type of models clinically useful, the black box needs to be illuminated to make the insights visible as in (b), and thus making the output interpretable.*

Chapter 1 summarizes the main properties of survival data. An illustration of the main problems in the analysis of survival data is given by using the example of breast cancer prognosis. An introduction to the most frequently used statistical survival models follows in **Chapter 2**. The methodological building blocks which will be used and adapted in order to obtain a new methodological framework for the analysis of survival data are discussed as well. The basics of support vector machines and the differences with least-squares support vector machines are explained. The advantages and drawbacks of these methods are shortly discussed. In addition, different methods for model selection and comparison are summarized. In **Chapter 3** a first survival method based on machine learning techniques is proposed and a theoretical basis is elaborated. The optimization problem is tackled in more detail in **Chapter 4**, where a comparison is made between different norms, equality versus inequality constraints and ranking versus regression constraints. The main contribution of the dissertation ends with bridging the gap between advanced mathematical models and clinical practice in **Chapter 5**. Starting from the models discussed in **Chapter 4**, a new methodology is developed in order to ensure that the generated models are easily interpretable and easily applicable for

clinicians. In addition, a visualization technique, making application of the models possible at the bedside of the patient as well as within software environments, is proposed. The dissertation ends with conclusions and ideas for future work (see **Chapter 6**).

1.5 Other research

1.5.1 Clinical studies

As stated in Section 1.3, the goal of this dissertation is to generate a new framework for the analysis of survival data. However, a major issue in the development of this methodology is the interpretability and applicability of the obtained model. This goal can only be met in close collaboration with clinicians. Thanks to participation in several clinical studies, it became clear how clinicians like to use models. It is only thanks to these collaborations that Chapter 5 of this thesis became possible. However, due to place restrictions, inclusion of these research topics in this thesis is infeasible. Therefore, a short overview of the different studies is given below.

Breast cancer studies

A first topic within the clinical studies is most related with the topic of the thesis and concerns clinical questions regarding breast cancer prognosis. The data set used in these studies contains information on breast cancer patients operated on in the Leuven University Hospitals between 1999 and 2005 (LBC). More information can be found in Appendix A.1.

The main questions raised in this topic are whether a group of patients with a good/bad prognosis can be identified in order to adapt clinical practice. A group of patients with an extremely good prognosis should not be treated with highly toxic drugs since the disadvantages of the treatment will lead to a worse survival. On the other hand, patients with a very bad prognosis should not be undertreated. In a first publication (Brouckaert *et al.*, 2009) the role of ER, PR and HER2 in defining prognostic groups is illustrated. This paper illustrates the significant role of these receptors for prognosis in primary operable breast cancers. It led to the question whether inclusion of ER, PR and HER2 in the NPI, which is the standard prognostic model in Belgium and many other countries, can improve the estimation of prognosis. This question was investigated in Van Belle *et al.* (2010d). In our cohort, inclusion of PR and HER2 status improved the NPI. The number of patients allocated to the intermediate risk group was significantly lower for the newly developed improved NPI (iNPI). The advantage is that more patients are allocated to a risk group for which the appropriate treatment options are clear

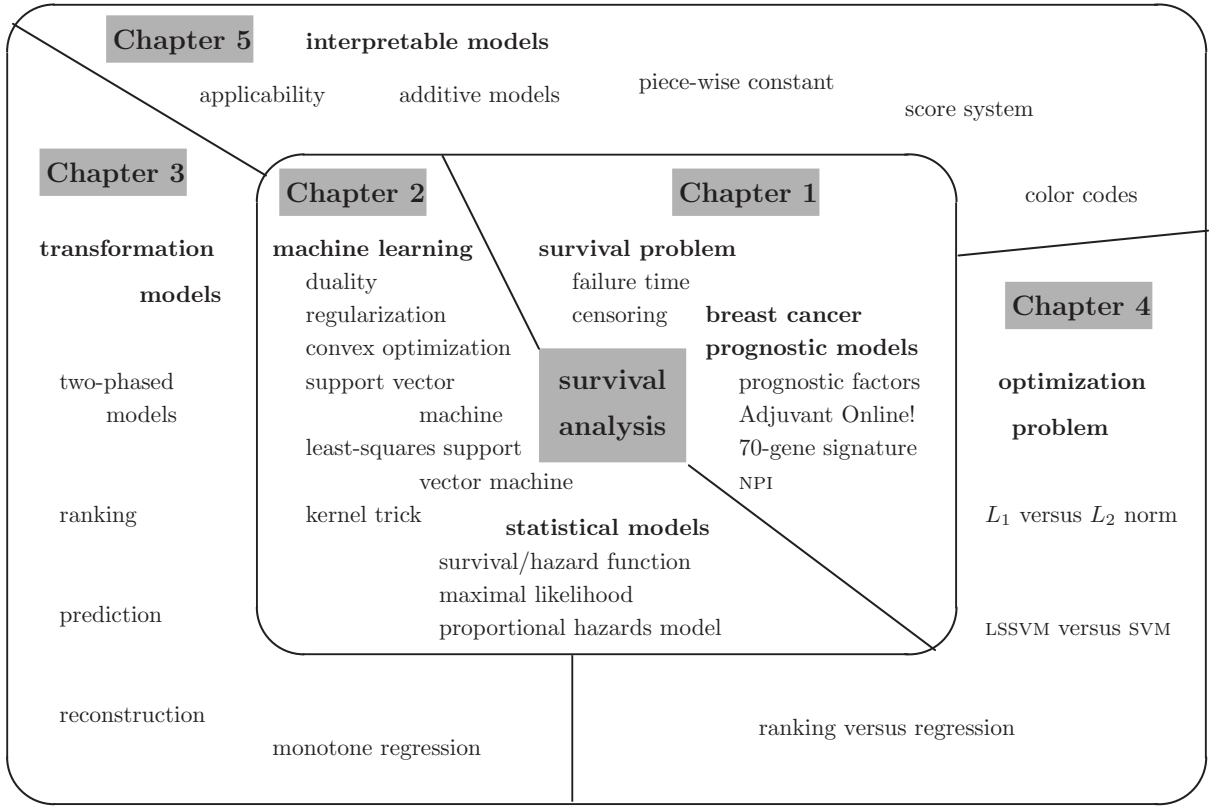


Figure 1.10: *Schematic representation of the dissertation.*

(Janes *et al.*, 2008; Pencina *et al.*, 2008). Additionally, the number of patients in the high risk group is reduced, resulting in a lower survival. The observed survival in the low risk groups for NPI and iNPI are not significantly different. These observations were confirmed on two external data sets. A first set involves 862 primary operable patients from the Breast Cancer Micrometastasis group in Oslo (Norway) diagnosed between 1995 and 1998 (OBC). The second data set contains 2805 patients from the Auckland Breast Cancer Registry (New Zealand) diagnosed between 2000 and 2005 (ABC). The analysis was done on the cases with complete information (676 and 1192 patients in Oslo and Auckland respectively). In second instance, the model and the evaluation measures were calculated on 100 multiple imputation data sets (Siddique & Harel, 2009) for all three cohorts. No differences between these and the complete case analysis were noted.

In a second study, it was confirmed that the lymph node ratio (LNR) proposed by Vinh-Hung *et al.* (2009) was a better prognostic factor than the number of positive lymph nodes in lymph node positive breast cancer (Van Belle *et al.*, 2009c). In addition, Vinh-Hung *et al.* (2004, 2006) proposed the log odds prognostic index (LPI), using the LNR in a prognostic model. We validated the LPI on 1838 patients of our series without sentinel nodes (Benson & Querci della Rovere, 2007; Mabry & Giuliano, 2007). In this study, we could not confirm the improved prognostic value of the LPI over the NPI. Although the LNR is a better prognostic factor than the number of positive lymph nodes, addition of this factor in a multivariate model did not show improved performance above the NPI. After updating (Steyerberg, 2009), it was concluded that the LPI can be of value, but updating the model equation might be necessary for new centers.

The last three studies investigate clinical questions within specific patient groups. A first study establishes the link between PR and the use of oral aromatase inhibitors (AI) on the prognosis in operable HER2 positive breast cancers. Out of our primary operable breast cancer series, 283 were recognized to be HER2 positive. However, only 220 patients were eligible for this study. Most non-eligible patients received trastuzumab, which was an exclusion criterium in this study. The goal of this study was to identify a group of HER2 positive breast cancer patients who are at low risk to relapse. The standard treatment for HER2 positive patients in the adjuvant as well as the metastatic setting, is trastuzumab. Results from several studies have shown that the use of trastuzumab significantly reduces recurrence and overall survival (Cobleigh *et al.*, 1999; Slamon *et al.*, 2001; Vogel *et al.*, 2002; Marty *et al.*, 2005; Piccart-Gebhart *et al.*, 2005; Romond *et al.*, 2005; Smith *et al.*, 2007). However, the significant effect of this drug is mainly seen during therapy and has less effect thereafter. So, HER2 positive tumors relapsing beyond the first one or two years after surgery or completion of chemotherapy do not seem to benefit from adjuvant trastuzumab. In addition, considering the substantial cost of trastuzumab (36000 euro for one year of treatment), identification of a subgroup of low-risk patients is highly relevant. Therefore, the potential prognostic value

of several clinico-pathological factors of patients before the adjuvant trastuzumab era, was investigated. From our series, we were able to identify 41 patients with a low risk to relapse. The low risk group constitutes of patients with a PR positive tumor, who received AI therapy, without positive lymph nodes. The results of this study were presented at the San Antonio Breast Cancer Symposium 2010 (Cho *et al.*, 2009). In order to validate the results, we are currently looking for collaborations with other centers.

As a second specific cohort, the invasive lobular breast cancers (ILA) were studied. ILAs differ from non-ILAs in many perspectives. ILAs are a heterogeneous group with a large variety in histologic subtypes and disease free survival. However, a prognostic model specific for ILAs is not available. In this study a model based on demographic and clinicopathological features specific for ILA tumors was developed. A retrospective cohort study of 380 consecutive patients treated between January 2000 and December 2006 for primary operable ILA, none E-cadherin positive, and all receiving local and systemic therapy, was available for analysis. Excluded from the analyses were 31 and 19 patients respectively because of neo-adjuvant therapy and multifocal or bilateral disease with a non-ILA tumor having a higher NPI than the ILA tumor. After a mean follow-up of 5.3 years, 37 patients (9.7%) experienced a breast cancer related event. 18% received adjuvant chemotherapy. In a univariate setting, variables considered as significant ($p<0.05$) were: node positivity, tumor size, grade (1-2 vs 3), mitotic count (1 vs 2-3), the amount of nuclear pleomorphism (1-2 vs 3) and subtype, classical ILA or not. The tubule formation was not considered as a variable since 98.4% of the patients had less than 10% of the tumor forming tubules. A multivariate Cox model revealed that node positivity, nuclear atypia and mitotic count are independent prognostic factors. Our prognostic model predicts a low risk for node negative ILAs. In node positive ILAs, the combination of nuclear atypia and mitotic index distinguished a medium and high risk group. It was concluded that risk groups can be defined without the complex definition of classical ILA. More information on this study can be found in Dierickx *et al.* (2010). Validation of the results on independent datasets remains necessary.

The last study involves patients with inoperable breast tumors. These patients will first receive neo-adjuvant treatment. In response to this treatment, the tumor should shrink or even disappear (complete remission). For some patients however, the tumor progresses, for others no change is noted. The main questions in this study were: (i) which factors predict pathologic complete response after neo-adjuvant chemotherapy for breast cancer and (ii) which factors predict disease free survival after neo-adjuvant chemotherapy for breast cancer. A study including 270 consecutive patients diagnosed with large size or locally advanced breast cancer showed that ER, HER2 and the clinical lymph node stage are independent predictive factors for pathologic complete remission. Additionally, PR positivity and pathologic complete response were shown to be significant prognostic factors

for disease free survival. Both parts of this study were presented at the European Breast Cancer Conference (Van Belle *et al.*, 2010a; Cho *et al.*, 2010). In order to validate the results, data from other centers are necessary.

The International Ovarian Tumor Analysis group

In cooperation with the International Ovarian Tumor Analysis group, three research questions, regarding classification between benign and malignant ovarian masses, were investigated. The data are discussed in detail in Appendix A.2. First, it was investigated whether intravenous contrast ultrasound examination is superior to grey-scale or power Doppler ultrasound for discrimination between benign and malignant adnexal masses with complex ultrasound morphology. In an international multicenter study, 134 patients with an ovarian mass with solid components or a multilocular cyst with more than 10 cyst locules, underwent a standardized transvaginal ultrasound examination followed by contrast examination using the contrast-tuned imaging technique and intravenous injection of the contrast medium SonoVue®. Time intensity curves were constructed, and peak intensity, area under the intensity curve, time to peak, sharpness and half wash-out time were calculated. The sensitivity and specificity with regard to malignancy were calculated and ROC curves were drawn for grey-scale, power Doppler, contrast variables and for subjective assessment. The gold standard was the histological diagnosis of the surgically removed tumors. After exclusions (surgical removal of the mass >3 months after the ultrasound examination, technical problems), 72 adnexal masses with solid components were used in our statistical analyses. The values for peak contrast signal intensity and area under the contrast signal intensity curve in malignant tumors were significantly higher than those in borderline and benign tumors, while those for the benign and borderline tumors were similar. The AUC of the best contrast variable with regard to diagnosing borderline or invasive malignancy (0.84) was larger than that of the best grey-scale (0.75) and power Doppler ultrasound variable (0.79) but smaller than that of subjective evaluation (0.93). Findings on ultrasound contrast examination differed between benign and malignant tumors but there was a substantial overlap in contrast findings between benign and borderline tumors. It appears that ultrasound contrast examination is not superior to conventional ultrasound techniques, which also have difficulty in distinguishing between benign and borderline tumors, but can easily differentiate invasive malignancies from other tumors. More information on this topic can be found in Testa *et al.* (2009).

In a second study the value of acoustic streaming to discriminate between endometriomas and other adnexal masses was investigated. Acoustic streaming is defined as the movement of particulate material within fluid due to energy transfer when an ultrasound wave is directed at it. After a study on 23 cystic masses, Edwards *et al.* (2003) proposed acoustic streaming as the first sonographic feature

that may be able to completely exclude endometrioma as a possible diagnosis for an adnexal cyst. The aim of our study was to determine the ability of acoustic streaming to discriminate between endometriomas and other adnexal masses in a large prospective multicenter study. We used data from 1938 patients with an adnexal mass included in Phase 2 of the International Ovarian Tumor Analysis study. Assessment of acoustic streaming was voluntary and was carried out only in lesions containing echogenic cyst fluid. Acoustic streaming was defined as movement of particles inside the cyst fluid during grey-scale and/or color Doppler examination provided that the probe had been held still for two seconds to ensure that the movement of the particles was not caused by movement of the probe or the patient. Only centers where acoustic streaming had been evaluated in >90% of cases were included. Sensitivity, specificity, positive and negative likelihood ratios (LR^+, LR^-), and positive and negative predictive values (PPV and NPV) of acoustic streaming with regard to endometrioma were calculated. 460 (24%) masses were excluded because they were examined in centers where <90% of the masses with echogenic cyst fluid had been evaluated for the presence of acoustic streaming. Acoustic streaming was evaluated in 633 of 646 lesions containing echogenic cyst fluid. It was present in 19 (9%) of 209 endometriomas and in 55 (13%) of 424 other lesions. This corresponds to a sensitivity of absent acoustic streaming with regard to endometrioma of 91% (190/209), a specificity of 13% (55/424), LR^+ of 1.04, LR^- of 0.69, PPV of 34% (190/559) and NPV of 74% (55/74). We therefore concluded that acoustic streaming cannot discriminate reliably between endometriomas and other adnexal lesions, and the presence of acoustic streaming does not exclude an endometrioma. This study is published in Van Holsbeke *et al.* (2010c).

In a last study the value of the ovarian crescent sign to discriminate between benign and malignant adnexal masses was validated. The ovarian crescent sign was first described by Hillaby *et al.* (2004). In their study including 100 patients, the ovarian crescent sign was present in only one out of 24 invasive tumors, in two out of 9 borderline tumors and in 56 out of 67 benign tumors. Therefore, Hillaby *et al.* (2004) proposed to use the ovarian crescent sign to exclude any type of malignancy (borderline and invasive tumors were considered malignant). The aim of this study was to determine the ability of the ovarian crescent sign to discriminate between benign and malignant adnexal masses in the hands of experienced ultrasound examiners in different ultrasound centers. The patients included were a subgroup of 1938 patients participating in the International Ovarian Tumor Analysis Phase 2 study. Within phase 2, the evaluation of the ovarian crescent sign was optional. Only patients from centers that had evaluated the ovarian crescent sign in $\geq$ 90% of their cases were included. The gold standard was the histological diagnosis of the adnexal mass. The ability of the ovarian crescent sign to discriminate between borderline or invasively malignant versus benign adnexal masses, as well as between invasively malignant versus other (benign and borderline) tumors, was determined and compared with the performance of subjective evaluation of ultrasound findings by the ultrasound examiner. In this cohort, the ovarian crescent sign was evaluated

in 1377 adnexal masses, 938 (68%) masses being benign, 86 (6%) borderline, 305 (22%) primary invasive and 48 (3%) metastases. The ovarian crescent sign was present in 398 (42%) of 938 benign masses, in 14 (16%) of 86 borderline tumors, in 18 (6%) of 305 primary invasive tumors and in two (4%) of 48 ovarian metastases. Hence, the sensitivity and specificity for absent ovarian crescent sign to identify a malignancy was 92% and 42%, respectively, and the positive and negative likelihood ratios (LR^+ and LR^-, respectively) were 1.60 and 0.18. Subjective impression performed significantly better than the OCS. Sensitivity and specificity were 90% and 92%, respectively, LR^+ was 11.0 and LR^- was 0.10. For discrimination between invasive versus benign or borderline tumors, the sensitivity for absent ovarian crescent sign was 94%, the specificity was 40%, the LR^+ was 1.58 and the LR^- was 0.14. In conclusion, this study confirms previous reports that the presence of the ovarian crescent sign decreases the likelihood of invasive malignancy in adnexal masses. However it is a poor discriminator between benign and malignant adnexal masses. This study is published in Van Holsbeke *et al.* (2010a).

The early pregnancy unit

This last clinical topic involves research on the behavior of early pregnancies. The data set is an observational cohort of 1882 pregnancies in 1767 women attending the Early Pregnancy and Gynaecology Ultrasound unit (EPU) of St. George's University in London with a positive pregnancy test from January to October 2006.

A first study investigated the possibility to determine the optimal gestational age at which to establish the location and viability of an early pregnancy using transvaginal ultrasonography. This study included 1442 women undergoing initial transvaginal ultrasonography at no more than 84 days gestation. Most excluded patients were not able to report gestational age. Logistic regression analysis was performed to determine the relationship between gestational age and the ability to confirm viability or non-viability. The commonest transvaginal ultrasonography finding prior to 35 days was a pregnancy of unknown location, from 35 to 41 days an early intrauterine pregnancy of uncertain viability and from 42 days a viable intrauterine pregnancy. Miscarriage could only be diagnosed on initial transvaginal sonography after 35 days. There was no difference between the ability to make a diagnosis for women with certain or uncertain dates ($p = 0.719$). The probability of confirming viability increased rapidly per day of gestation until 49 days and remained constant thereafter. Of the 29 ectopic pregnancies diagnosed, 72% presented prior to 49 days gestation and all of these women presented with pain, bleeding or a previous ectopic pregnancy. Is was therefore concluded that the ability to confirm viability or non-viability is significantly related to gestational age. In asymptomatic women with no previous ectopic pregnancy transvaginal

ultrasonography should be delayed until 49 days. Our data suggest that this would reduce the number of inconclusive scans, without an associated increase in morbidity from missed ectopic pregnancies. A more detailed description of this study is given in Bottomley *et al.* (2009).

A second study investigates the ability to predict ongoing viability in very early intrauterine pregnancies. The ultrasound findings of a normally developing pregnancy at 4-5 weeks gestation may be identical to those of an early failed or failing pregnancy. The finding of such an intrauterine pregnancy of uncertain viability (IPUVI) creates maternal anxiety and increases follow up visits. Our objective was to determine the incidence and natural history of IPUVI and analyze whether ongoing viability could be predicted from maternal demographics, symptoms or ultrasound variables. This cohort contained 493 women with an empty gestation sac <20mm, a gestation sac <25mm containing a yolk sac only or a fetal pole <6mm without fetal heart beat at transvaginal ultrasound. The main outcome measures were ongoing pregnancy viability at initial 7-14 day follow up and viability at the end of the first trimester. The incidence of IPUVI was 29.2%. 62.3% of these pregnancies were viable at initial follow up and 50.5% at the end of the first trimester. First trimester viability was associated with maternal age over 35 years (OR 0.79, 95% CI = 0.66-0.94), gestational age less than 42 days (OR 3.62, 95% CI = 1.94-6.76), the vaginal bleeding score (OR 0.35, 95% CI = 0.21-0.57), mean gestational sac diameter (OR 0.88, 95% CI = 0.81-0.97) and the presence of a yolk sac (OR 3.04, 95% CI = 1.58-5.84). A simple mathematical model was found to predict first trimester outcome with an area under the receiver operating characteristics curve of 0.83 in a training data set and 0.71 in a test data set.

The last study aimed at developing a scoring system to inform women in early pregnancy of their likelihood of having a viable pregnancy at the end of the first trimester. Exclusion criteria were: twin or triplet pregnancy, termination of pregnancy, not pregnant. Multiple demographic features, symptoms and ultrasound variables were tested for a significant influence on ongoing viability. Multivariable logistic analysis was used to determine the individual likelihood of first trimester viability. The clinicians were interested in three models: (i) a model only including demographics and symptoms, (ii) a model only including scan findings available at the initial scan, and (iii) a model considering both types of variables. The questions that should be answered from these models are: (i) Do demographic variables and symptoms provide enough information to obtain a reliable prediction of the viability of a pregnancy at 3 months of gestation?; (ii) Do scan findings obtained from an initial scan provide enough information to obtain a reliable prediction of the viability of the pregnancy at the end of the first term?; (iii) Do scan findings provide extra information above the information given by demographics and symptoms? This last question is very relevant in clinical practice, since it answers the question whether women should get an ultrasound scan early in pregnancy. The models were developed on a training data set and

converted into a points system (Sullivan *et al.*, 2004). As a result, the risk for a patient can be calculated by adding the points for each variable in the model. This model was tested on a separate test data set. A numeric scoring system using maternal age and amount of bleeding gave an area under the curve of 0.724 in the training set and 0.729 in the test set. Significant ultrasound variables alone (mean gestation sac diameter, mean yolk sac diameter and presence of fetal heart beat) gave an AUC of 0.873 (training) and 0.900 (test). Demographic and initial ultrasound variables together gave an AUC of 0.910 (training) and 0.938 (test). This system, if validated, is likely to be useful in managing parental expectations and guiding further first trimester assessment if the prediction of non-viability is high.

1.5.2 Development of discriminatory measures for multi-class classification problems

As a final side-path, the extension of two-class discrimination measures towards multi-class problems was investigated. This topic involves issues as the Receiver Operating Characteristic Curve (ROC) (Swets, 1988; Bradley, 1997; Pepe, 2003; Lasko *et al.*, 2005), the Area Under the ROC (AUC) (Hanley & McNeil, 1982) and the concordance index (Harrell *et al.*, 1984, 1988; Harrell, 2001). This research resulted in a paper at the International Joint Conference on Neural Networks (see Van Calster *et al.*, 2008). Currently two journal papers are under construction. A first paper gives an overview of different measures given in the literature and describes advantages and disadvantages for the use of these methods in multi-class problems. A second paper proposes a new measure of discrimination for polytomous outcomes (Polytomous Discrimination Index or PDI) and ordinal outcomes (Ordinal Discrimination Index or ODI).

Chapter 2

A Methodological Primer

This Chapter introduces the most common statistical methods for handling censored data and summarizes the advantages and disadvantages of these methods. Section 2.3 covers the basic properties of machine learning techniques, which lie at the basis of the methods developed in the following chapters. The next Section summarizes different model performance measures.

2.1 Terms and definitions

The distribution of survival or failure times is characterized in different ways. Survival can be represented by the probability density function (PDF)

$$f(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq y < t + \Delta t\right)}{\Delta t}\,, \tag{2.1}$$

expressing the probability of the event under study to occur over time. The cumulative distribution function (CDF)

$$F(t) = P(y \leq t)\,, \tag{2.2}$$

expresses the probability that the event occurs up to time t and equals

$$F(t) = 1 - S(t)\,, \tag{2.3}$$

with $S(t)$ the survival function. The survival function $S(t)$ is defined as the probability that the event did not occur before time t. Finally, the survival can be represented by the hazard function or hazard rate

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq y < t + \Delta t | y \geq t\right)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.4}$$

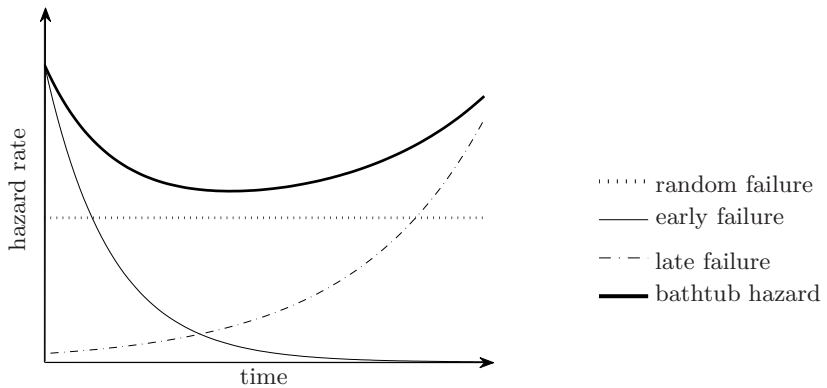


Figure 2.1: *Illustration of different types of hazard rates.*

The quantity $\lambda(t)\Delta t$ is the approximate probability for the event to occur within the interval $[t, t + \Delta t]$, given that the event did not occur before. A decreasing hazard rate corresponds to a high likelihood of very early failures. This type of hazard functions are encountered in certain electronic devices and in patients after transplantation. An increasing hazard is more common and corresponds to an ageing effect. The bathtub hazard is most often appropriate for human and product life, where the risk of death is large at and shortly after birth (infant diseases), then stabilizes and starts increasing when ageing effects start playing a role. Figure 2.1 illustrates the different types of hazard rates.

2.2 The history of survival models

This Section discusses different survival models. The literature describes survival distributions, parametric models and semi-parametric models. The advantages and disadvantages of the different models are shortly discussed.

2.2.1 Non-parametric estimation of the survival function

A non-parametric estimator of the survival function was proposed by Kaplan & Meier (1958). Consider a data set $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^{n}$, where x_i, y_i and δ_i represent the variable vector, a positive survival time and the event indicator for observation i, respectively. The event indicator is equal to 1 if an event

occurred, and zero if the subject is (right) censored at time y_i. The Kaplan-Meier estimator estimates the survival function based on the observed set of observations $\mathcal{D}$, without taking the variables x_i into account. Assume that k observations experience an event at k distinct event times $y_i,\ i = 1, \ldots, k)$, with $y_i < y_{i+1}, i = 1, \ldots, k-1$. The assumption of distinct failure times is also denoted by the assumption that there are no *ties* in the data set. The probability of all the events is

$$S(t) \quad = \quad P(t > y_1) \prod_{l=2}^{k} P(t > y_l | t > y_{l-1}) \,. \tag{2.5}$$

$P(t > y_l | t > y_{l-1})$ is approximated by $1 - \frac{d_l}{|\mathcal{R}_l|}$, where d_l represents the number of failures at time y_l and $\mathcal{R}_l$ represents the risk set. This set contains all observations which are at risk at time y_l, namely patients experiencing the event later than y_l and patients having a censoring time later than y_l. The Kaplan-Meier estimator of the survival function is then expressed as

$$\hat{S}(t) = \prod_{l:y_l \leq t} \left(1 - \frac{d_l}{|\mathcal{R}_l|}\right) . \tag{2.6}$$

Note that the Kaplan-Meier estimator only changes at failure times, remains one up to the first failure time and only goes to zero if the last observation is an event.

2.2.2 Survival distributions and parametric models

The statistical literature describes different models for the analysis of failure time data, an overview of which can be found in Kalbfleisch & Prentice (2002). The oldest models are fully parametric models specifically designed for data skewed near zero (failure times can not be negative). They assume a distribution on the failure times and fit this distribution to the observations. Examples of typical failure time distributions are the exponential distribution, the Weibull distribution, the log-normal distribution and the log-logistic distribution (Lawless, 2006). The Weibull distribution depends on two parameters b_1 (scale parameter) and b_2 (shape parameter) as follows

$$f(t) = b_1^{-b_2} b_2 t^{b_2 - 1} \exp\left(-\left(\frac{t}{b_1}\right)^{b_2}\right) . \tag{2.7}$$

Figure 2.2 illustrates the Weibull distribution and survival function for several values of the parameters. A shape parameter less than 1 is typical when events mainly occur at early times. The hazard rate decreases as time goes on. A shape parameter of one indicates a constant hazard rate and the Weibull distribution

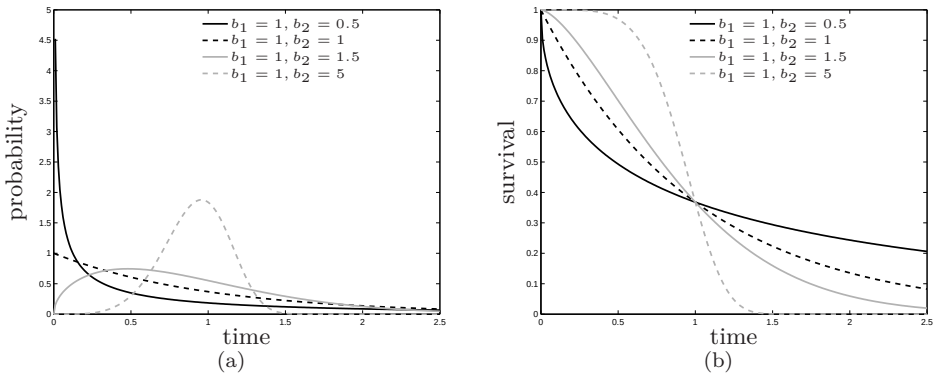


Figure 2.2: *Weibull* PDF *(a) and survival function (b) for different values of the parameters* b_1 *and* b_2*.*

reduces to an exponential distribution in this case. A hazard rate which is increasing over time corresponds to a shape parameter larger than 1 and is typical for aging effects.

The exponential distribution

$$f(t) = b_1 \exp(-b_1 t) \,, \tag{2.8}$$

is a special case of the Weibull distribution. Later on, parametric regression models were proposed. A first type of regression models is based on the distribution models, but instead of estimating single distribution parameters, the parameters are modelled as a function of the variables x. The Weibull regression model is defined as

$$S(t|x) = \exp\left(-(b_1(x)t)^{b_2(x)}\right) \,. \tag{2.9}$$

Often used parameterizations are $b_1(x) = \exp(w^T x)$, with w a coefficients vector, and $b_2(x) = b$. A more commonly used regression model is the accelerated failure time model, which models the logarithm of the failure time as a linear function of the variables

$$\log(t) = w^T x + b \,. \tag{2.10}$$

This model has an appealing interpretation. Let the value of the single input variable $x = 1$ for men and $x = 0$ for women. The value $\exp(w)$ is then interpreted as the estimated ratio of expected mean survival for both groups. More information on and examples of survival distributions and parametric survival models can be found in Kalbfleisch & Prentice (2002).

2.2.3 The proportional hazards model

The largest breakthrough in modelling survival data came in 1972 when Cox proposed his proportional hazards model (PH) (Cox, 1972). The PH is built on two assumptions: (i) proportional hazards and (ii) linear parametric form in the covariates. The PH is defined as

$$\lambda(x,t) = \lambda_0(t)\exp(w^T x),\ t>0\,, \tag{2.11}$$

with $\lambda_0(t)$ an unspecified baseline hazard function. The model assumes proportional hazards since

$$\frac{\lambda(x_i,t)}{\lambda(x_j,t)} = \frac{\lambda_0(t)\exp(w^T x_i)}{\lambda_0(t)\exp(w^T x_j)} = \frac{\exp(w^T x_i)}{\exp(w^T x_j)}\,. \tag{2.12}$$

As such, the covariates only have a proportional effect on the baseline hazard. The PH can be rewritten as

$$\log\left(\frac{\lambda(x,t)}{\lambda_0(t)}\right) = w^T x\,, \tag{2.13}$$

which reveals that the PH is a linear model in the log of the hazard ratio. One often refers to this as to the linearity assumption of the PH. The probability of surviving associated with an observation with covariates x_i is related to $u(x_i) = w^T x_i$, also called a prognostic index or utility. The PH is a semi-parametric model since the baseline hazard is left unspecified. The model parameters w are estimated in a way similarly to the maximal likelihood estimators in linear and logistic regression. There are two types of contributions to the likelihood. Individuals which are censored at y_i contribute $S(x_i,y_i)$ to the likelihood. Individuals who fail at y_i contribute $f(x_i,y_i) = S(x_i,y_i)\lambda(x_i,y_i)$. In the assumption of i.i.d. observations, the likelihood is then defined as

$$\begin{aligned} L(x) &= \prod_{i=1}^{n} (\lambda(x_i,y_i)S(x_i,y_i))^{\delta_i} S(x_i,y_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda(x_i,y_i)^{\delta_i} S(x_i,y_i) \\ &= \prod_{i=1}^{n} \left(\frac{\lambda(x_i,y_i)}{\sum_{j\in\mathcal{R}_i}\lambda(x_i,y_i)}\right)^{\delta_i} \left(\sum_{j\in\mathcal{R}_i}\lambda(x_i,y_i)\right)^{\delta_i} S(x_i,y_i)\,. \end{aligned} \tag{2.14}$$

Since

$$S(x_i,y_i) = S_0(y_i)^{\exp(w^T x_i)}\,, \tag{2.15}$$

with $S_0(y_i) = \exp(-\int_0^{y_i} \lambda_0(u)du)$ the baseline survival function. Inserting model (2.11) in (2.14) results in

$$L(x, \lambda_0, w) \quad = \quad L(w) \prod_{i=1}^{n} \left(\sum_{j \in \mathcal{R}_i} \lambda_0(y_i) \exp(w^T x_j) \right)^{\delta_i} S_0(y_i)^{\exp(w^T x_i)} \quad (2.16)$$

with

$$L(w) = \prod_{i=1}^{n} \left(\frac{\exp(w^T x_i)}{\sum_{j \in \mathcal{R}_i} \exp(w^T x_j)} \right)^{\delta_i}, \quad (2.17)$$

the partial likelihood. Cox (1972) argued that the partial likelihood contains almost all information on w, and the second term contains almost all information on $\lambda_0(t)$. The parameters w are inferred by maximizing this partial likelihood.

Although the PH is the most commonly used survival model, some drawbacks remain. The model is based on the assumption that hazards are proportional. A second drawback is the linear form in the variables. The last decade different models dealing with one or both of these issues have been proposed. The linearity assumption can be handled by generalization of the PH. In Huang *et al.* (2000) the linear form of the PH is replaced with a functional ANOVA model. ANOVA models have the advantage that they are more flexible than linear models and that they deal with the curse of dimensionality by restricting their components to a constant effect, the main effects (function involving one single variable) and low-order interactions (functions involving a few variables). Additive regression splines were used in the PH by Leblanc & Crowley (1999). Penalized smoothing splines were used in Eilers & Marx (1996); Hurvich *et al.* (1998). The PH model is rewritten as

$$\lambda(x, t) = \lambda_0(t) \exp(f(x)), \quad (2.18)$$

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

$$f(x) = \sum_{a=1}^{m} w_a B_a(x). \quad (2.19)$$

The penalized log partial likelihood function then becomes

$$\log(L_p(w)) = \log(L(w)) - \frac{\gamma}{2} \sum_{a=3}^{m} (w_a - 2w_{a-1} + w_{a-2})^2, \quad (2.20)$$

with γ a positive regularization constant. This approach will be denoted by $\text{PH}_{\text{pspline}}$.

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

$$\log(L_p(w)) = \log(L(w)) - \frac{\gamma}{2} w^T w \,, \tag{2.21}$$

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

2.2.4 Partial logistic artificial neural network

In order to overcome both assumptions of the PH, survival models based on artificial neural networks (Bishop, 1995; Ripley, 1996; Suykens *et al.*, 1996) for censored data were proposed (Faraggi & Simon, 1995; Biganzoli *et al.*, 1998, 2002; Lisboa *et al.*, 2003; Biganzoli *et al.*, 2006). Artificial neural networks are based on the functioning of neurons in the brain. A popular model for one neuron is the McCulloch-Pitts model (McCulloch & Pitts, 1943). The neuron is modelled as a static non-linear element, making a weighted sum of its input variables x^p with the weights w_p. Typically, a bias term is added to the model. The weighted sum is then fed through a non-linear activation function with saturation characteristics. To amplify the strength of the model, multiple layers of neurons are often combined to form a multi-layer perceptron (MLP) model. The survival model proposed by Biganzoli *et al.* (1998) is an MLP, generalizing logistic regression for grouped failure data. The model contains three layers: (i) an input layer containing a neuron for each input $p = 1, \ldots, d$, and one neuron for the time variable; (ii) a hidden layer containing n_h hidden neurons and (iii) an output layer with one output neuron. The model is trained as follows. First the time in which observations were followed is divided into time intervals $[t_{k-1}, t_k], k = 2, \ldots, n_k$. The goal is then to estimate the hazard within each of these intervals. Therefore, each data point is replicated for each time interval in which the outcome is known. The input of the model consists of two parts: the covariate vector and a time variable indicating the time interval. The output is one if the patient experienced the event under study within the considered time interval, zero otherwise. For discrete time studies, the output of the model will represent the predicted hazard within each time interval. The model equation is

$$\frac{\lambda(x, t_k)}{1 - \lambda(x, t_k)} = g_0 \left(\sum_{h=1}^{n_h} w_h g_h \left(\sum_{p=1}^{d} w_{ph} x^p + w_t t_k + b_k \right) + b_h \right) , \tag{2.22}$$

where $g_0(\cdot) = \exp(\cdot)$ and g_h is the logistic function.

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

$$P(w|\mathcal{D}, \alpha, \mathcal{H}) = \frac{P(\mathcal{D}|w, \alpha, \mathcal{H})P(w|\alpha, \mathcal{H})}{P(\mathcal{D}|\alpha, \mathcal{H})}\,. \tag{2.23}$$

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

2.3 Machine learning methods

The basic machine learning building blocks which are used throughout the dissertation to develop a new methodology for the analysis of survival data are explained in this Section. First support vector machines (SVM) for classification and regression purposes are introduced. It is then shown how the SVM approach is modified to obtain a least-squares SVM.

2.3.1 Support vector machines

Support vector machines for classification

Consider a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $y_i = -1$ for observations in class 1 and $y_i = 1$ for observations in class 2. Assume a linear model with the sign function as prediction function

$$\hat{y}_i = \text{sign}(w^T x_i + b)\,. \tag{2.24}$$

Consider a two-class classification problem with 2 input variables x_1 and x_2. Figure 2.3(a) shows that there exist several hyperplanes that separate the two classes.

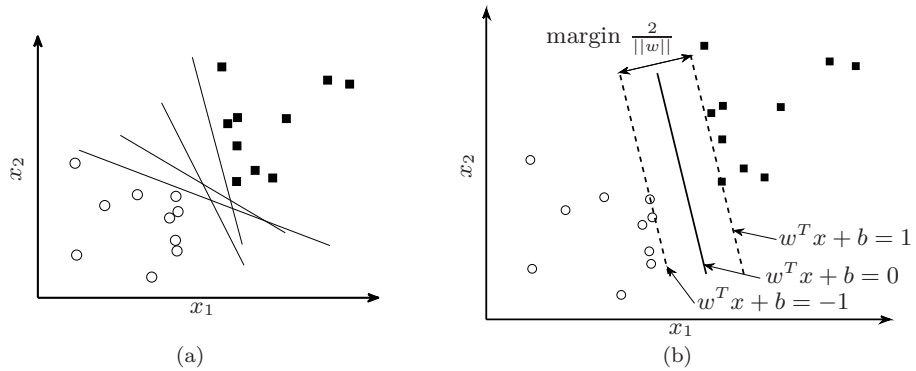


Figure 2.3: *Illustration of the maximal margin principle. The circles and squares represent observations belonging to the first and second class, respectively. (a) In the separable case, different hyperplanes are possible. (b) A unique hyperplane can be defined by maximizing the distance between the points from both classes which are closest to each other.*

In order to define a unique separating hyperplane, Vapnik (1998) proposed to maximize the distance between the two points of the two classes which are closest to each other (see Figure 2.3(b)). He rescaled the problem such that the points closest to the hyperplane satisfy the condition $|w^T x+b| = 1$. A good generalization is then obtained by maximizing the margin, or minimizing $||w||_2$.

The primal problem formulation for SVM classification is

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & \begin{cases} y_i \left(w^T \varphi(x_i) + b \right) \geq 1 - \epsilon_i, & \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{2.25}$$

In this notation, $\varphi(x)$ represents a feature map, projecting the inputs x into a higher dimensional space. The reasoning behind this procedure is that a non-linearly separable problem can become linearly separable in a higher dimensional space (see Figure 2.4). The slack variables ϵ_i are introduced to tolerate misclassifications. Cases for which $\epsilon_i \geq 1$, $0 < \epsilon_i < 1$ or $\epsilon_i = 0$, are misclassified, correctly classified but lying inside the margin and correctly classified and lying on or outside the margin, respectively. A trade-off between misclassifying training observations and model complexity is made by means of the positive regularization

constant γ. The dual problem is then formulated as

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \varphi(x_i)^T \varphi(x_j)\alpha_i\alpha_j - \sum_{i=1}^{n}\alpha_i \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq \gamma, \quad \forall\, i = 1,\ldots,n\,. \end{cases} \end{aligned} \tag{2.26}$$

The classifier then becomes

$$\hat{y}(x) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i \varphi(x_i)^T\varphi(x_j) + b\right). \tag{2.27}$$

The constant b is found by solving the KKT condition $\alpha_i\left(y_i\left(w^T\varphi(x_i)+b\right)-1+\epsilon_i\right) = 0,\ \forall\, i = 1,\ldots,n$. An advantage of this approach is that the feature map $\varphi(x)$ does not need to be constructed explicitly. Any continuous function $k(x, x_\star)$ for any points x and $x_\star$ satisfying Mercer's condition (Mercer, 1909) can be expressed as an inner product

$$k(x, x_\star) = \varphi(x)^T\varphi(x_\star)\,. \tag{2.28}$$

The inner product in equation (2.27) can thus be replaced by evaluating a positive semi-definite function (referred to as the kernel function). This procedure is often called the *kernel trick*. Often used kernels are

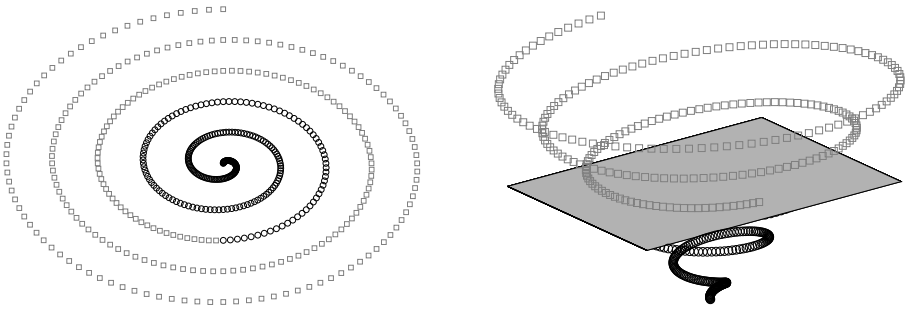
- the linear kernel: $k(x, x_\star) = x^T x_\star$,
- the polynomial kernel of degree a: $k(x, x_\star) = (\tau + x^T x_\star)^a$, with $\tau \geq 0$
- the RBF kernel: $k(x, x_\star) = \exp\left(-\frac{||x-x_\star||_2^2}{\sigma^2}\right)$.

More recently, a kernel for clinical data (Daemen & De Moor, 2009) was proposed as an additive kernel $k(x, x_\star) = \sum_{p=1}^{d} k^p(x, x_\star)$, where $k^p(\cdot,\cdot)$ depends on the type of the variable p. For continuous and ordinal variables, $k^p(\cdot,\cdot)$ is defined as

$$k^p(x^p, x_\star^p) = \frac{c - |x^p - x_\star^p|}{c}\,, \tag{2.29}$$

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

$$k^p(x^p, x_\star^p) = \begin{cases} 1 \text{ if } x^p = x_\star^p \\ 0 \text{ if } x^p \neq x_\star^p\,. \end{cases} \tag{2.30}$$



(a) (b)

Figure 2.4: *Illustration of a two class (circles: class 1; squares: class 2) non-linearly separable classification problem. (a) In the two dimensional input space, the two classes are not linearly separable. (b) After mapping the problem in a three dimensional space, the problem becomes linearly separable.*

Support vector machines for regression

Estimation of a function in the sense of Vapnik (1998) is done by means of the following primal formulation

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T \varphi(x_i) + b \geq y_i - \epsilon_i, & \forall\ i = 1, \ldots, n \\ -w^T \varphi(x_i) - b \geq -y_i - \epsilon_i^*, & \forall\ i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n \\ \epsilon_i^* \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{2.31}$$

The dual problem is found as

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j (\alpha_i - \alpha_i^*) \varphi(x_i)^T \varphi(x_j) (\alpha_j - \alpha_j^*) - \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 & \\ 0 \leq \alpha_i \leq \gamma, & \forall\ i = 1, \ldots, n \\ 0 \leq \alpha_i^* \leq \gamma, & \forall\ i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{2.32}$$

The dual representation of the model becomes

$$\hat{y}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i \varphi(x_i)^T \varphi(x) + b\,, \tag{2.33}$$

where $\varphi(x_i)^T\varphi(x)$ is again replaced by $k(x_i, x)$.

From the above derivations it is noted that the SVM formulations lead to a convex quadratic programming problem (QP). For positive (semi-)definite kernels, the solution is global. An advantage of SVMs is their sparseness. Many α_i values are equal to zero and the dual representation of both the classification and regression model can be found by summing over those observation for which $\alpha_i \neq 0$. These observations are called the support vectors.

2.3.2 Least-squares support vector machines

The SVM classifier was modified by Suykens & Vandewalle (1999) at two points: (i) the inequality constraints are replaced with equalities and (ii) a squared loss function is used. The least-squares SVM (LS-SVM) classification problem then becomes

$$\begin{array}{ll} \min\limits_{w,b,\epsilon} & \frac{1}{2}w^Tw + \frac{1}{2}\gamma\sum\limits_{i=1}^{n}\epsilon_i^2 \\ \text{subject to} & y_i\left(w^T\varphi(x_i) + b\right) = 1 - \epsilon_i, \quad \forall\, i = 1, \ldots, n\,. \end{array} \tag{2.34}$$

The classifier in the primal space then becomes

$$\hat{y} = \text{sign}(w^T\varphi(x) + b)\,. \tag{2.35}$$

Forming the Lagrangian of this problem and taking the derivatives towards the primal and dual variables yields the following system

$$\left[\begin{array}{cc} \Omega + \frac{I}{\gamma} & y \\ y^T & 0 \end{array}\right]\left[\begin{array}{c} \alpha \\ b \end{array}\right] = \left[\begin{array}{c} 1_n \\ 0 \end{array}\right], \tag{2.36}$$

where Ω is a matrix with elements $\Omega_{ij} = y_iy_j\varphi(x_i)^T\varphi(x_j) = y_iy_jk(x_i, x_j)$, where the kernel trick is applied as before, $y = [y_1, \ldots, y_n]^T$ is a vector with all class labels and $\alpha = [\alpha_1, \ldots, \alpha_n]^T$. The classifier in the dual space then becomes

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{n}\alpha_iy_ik(x_i, x) + b\right). \tag{2.37}$$

LS-SVMs can also be used for function estimation. The model formulation in primal space is

$$\begin{array}{ll} \min\limits_{w,b,\epsilon} & \frac{1}{2}w^Tw + \frac{1}{2}\gamma\sum\limits_{i=1}^{n}\epsilon_i^2 \\ \text{subject to} & w^T\varphi(x_i) + b = y_i - \epsilon_i, \quad \forall\, i = 1, \ldots, n\,, \end{array} \tag{2.38}$$

and the dual problem is found as

$$\left[\begin{array}{cc} \Omega + \frac{I}{\gamma} & 1_n \\ 1_n^T & 0 \end{array} \right] \left[\begin{array}{c} \alpha \\ b \end{array} \right] = \left[\begin{array}{c} y \\ 0 \end{array} \right], \tag{2.39}$$

where Ω is a matrix with elements $\Omega_{ij} = k(x_i, x_j)$. The model for function estimation then becomes

$$\hat{y} = \sum_{i=1}^{n} \alpha_i k(x_i, x) + b\,. \tag{2.40}$$

The advantage of LS-SVMs in comparison with SVMs is that the LS-SVM formulation leads to a linear KKT system, whereas the SVM formulation leads to a QP problem. It was shown that SVMs can be solved by iteratively solving adapted linear KKT systems, where each system has a form similar to one single KKT system in LS-SVMs (Navia-Vazquez *et al.*, 2001). As for SVMs, the solution is global. A drawback of LS-SVMs is the lack of sparseness since the Lagrange multipliers α_i are related to the errors as $\alpha_i = \gamma \epsilon_i$.

2.3.3 Model selection

Model selection schemes

An important issue in generating models which generalize well is the tuning of regularization (γ) and kernel (e.g. σ in the case of an RBF kernel) parameters. Different strategies are possible for model selection, but all of them are intended to reduce the dependance on the training data. The goal of model selection is to find those values of the parameters which result in a high performance on unseen data. It is therefore important not to tune these parameters on the data which are used to train the model. This would result in memorizing the training data and in general yields a bad performance on unseen data. Different model selection algorithms correspond to different manners of splitting the data set into a training and a validation part. The model is trained on the training set and the tuning parameters are found by looking for those parameters resulting in the best performance on the validation set. The easiest method for model selection is therefore to divide the training set randomly in two parts (typically 2/3 for training and 1/3 for validation) and calculate the performance of models trained on the first set and tested on the second set, for different values of the parameters. However, this approach is sensitive to the specific split of the data.

An alternative approach is to use n_k-fold cross validation. Here, the data are randomly split into n_k sets of equal size. In a first run, the model is trained on the last $n_k - 1$ subsets and tested on the first subset. In the second run, the model is

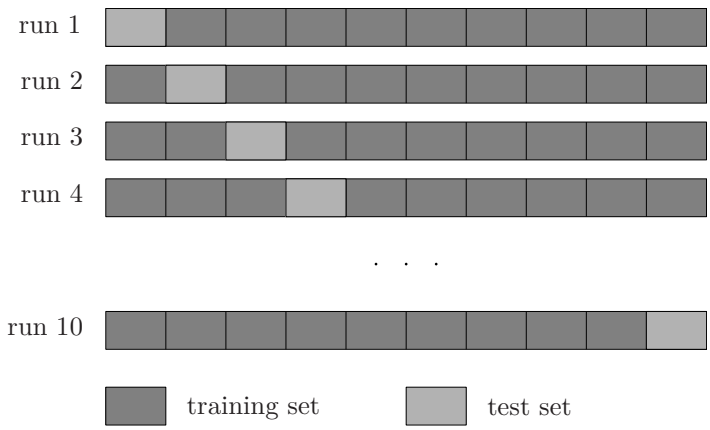


Figure 2.5: *Illustration of* 10*-fold cross validation. The data set is randomly divided into* 10 *subsets. The model is trained* 10 *times (*10 *runs). In the* k^{th} *run,* $k = 1, \ldots, 10$*, the* k^{th} *subset is left out during training. The model is then evaluated on the* k^{th} *subset.*

trained on all subsets except the second one and is tested on the second subset. In total n_k runs are performed (see Figure 2.5) and the overall performance for the specific setting of the parameters used to train the models is found by calculating the performance of the results obtained on all small validation sets. If n_k is chosen to equal n, the procedure is called leave-one-out cross validation.

Parameter tuning

One of the advantages of SVMs and LS-SVMs is that they result in a global solution. However, this is true given the regularization parameter γ and possible additional kernel parameters. Since the problem of optimizing the value of these parameters is not a convex problem, many local minima exist. This Subsection discusses two possible ways to tune the parameters. A first possibility is to use a grid search. Here, a grid of possible parameter values is defined. For each combination of parameter values, the model performance is calculated. To ensure that the global optimum can be found, the grid should be built with care. An exponential grid, for which the difference between two grid points is larger for larger values and smaller for smaller values is often used. In addition, when using an RBF kernel, the value for the kernel parameter should scale up with the dimension of the data. This method can be refined by using the optimal values obtained in a first grid, as starting point to build a second, finer, grid.

A second approach which is particularly useful for tuning multiple parameters, is

coupled simulated annealing (Xavier de Souza *et al.*, 2010). Coupled simulated annealing is a method for global optimization of continuous variables based on simulated annealing. In simulated annealing, one starts from a random initialization of the parameters (called state). A probe state is generated by adding a random variable, sampled from a distribution changing in time, to the current state. Early in the optimization process, the sampling distribution is wide and a wide range of parameter values is explored. Later on, the distribution narrows and the search becomes gradually more local. The new state is accepted when the cost for this state is lower than the cost for the current state. In addition, a higher cost is accepted with a probability depending on time. In the beginning, this probability is high in order to be able to escape from local optima. Later in the process, these movements towards states with higher costs become less and less probable.

2.4 Model evaluation

Models need to be evaluated on two points: (i) discrimination and (ii) calibration. Discrimination measures whether the model is able to discriminate between low and high risk groups, whereas calibration checks whether the predicted probability equals the observed probability. This Section starts with a discussion on both topics for prediction models, where the outcome for each observation is known to be 0 or 1. Afterwards it is discussed how these model evaluation methods can be extended towards survival data. In addition, typical performance measures for survival data are introduced.

2.4.1 Discrimination

After model training, a utility $u(x)$ depending on the input variables is defined. In logistic regression $u(x) = w^T x$, in SVMs or LS-SVMs $u(x) = w^T \varphi(x) + b$. In neural networks and machine learning, the utility is also called the latent variable. To obtain a prediction, a threshold v on the utility needs to be defined. Observations with a utility lower than v, have a predicted outcome of 0, whereas observations with a utility higher than v have a predicted outcome of 1. Depending on both the true and predicted outcome, four different numbers are defined:

- true positives (TP): number of observations with outcome equal to 1, and a predicted outcome equal to 1
- true negatives (TN): number of observations with a predicted outcome equal to 0, and a true outcome equal to 0

- false positives (FP): number of observations with outcome equal to 0 and predicted outcome equal to 1
- false negatives (FN): number of observations with outcome equal to 1 and predicted outcome equal to 0.

Two measures which are derived from the previous numbers are the true positive and true negative fraction. The true positive fraction or the sensitivity is defined as

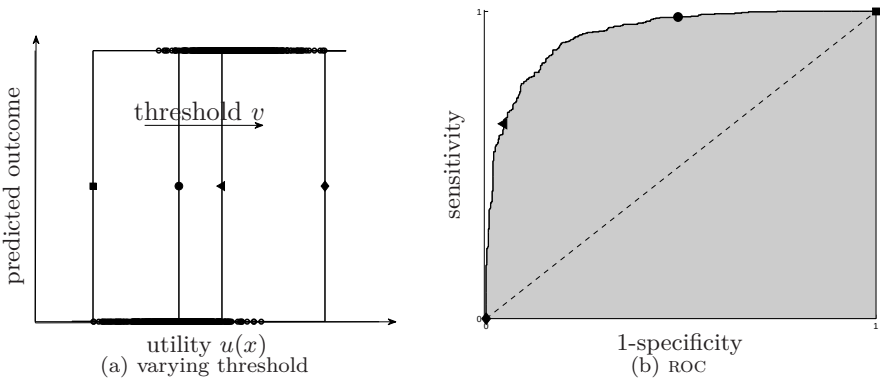
$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,, \tag{2.41}$$

and represents the capacity of a model to detect the presence of the event under study (outcome=1). The true negative fraction or the specificity represents the capacity of the model to detect the absence of the event under study and is defined as

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}\,. \tag{2.42}$$

The relation between sensitivity and specificity, depending on the threshold v is illustrated in the receiver operating characteristic (ROC) curve (Figure 2.6) (Swets, 1988; Bradley, 1997; Pepe, 2003; Lasko *et al.*, 2005). A model with a random performance corresponds to the dashed ROC curve. The ROC curve represented by the solid line corresponds to a model with a good discrimination ability. The area under this curve (grey area) summarizes the discrimination ability of the model is one number (Hanley & McNeil, 1982). This measure (AUC) lies in the unit interval. An AUC equal to 0.5 indicates random performance, whereas an AUC equal to 1 indicates perfect discrimination. The AUC corresponds to the fraction of concordant over comparable pairs of observations. Each pair of observations with one event (outcome = 1) and one non-event (outcome = 0) is comparable. A pair is concordant when the utility of the event is higher than the utility of the non-event.

Figure 2.7 illustrates how the AUC can be generalized for survival data by means of the concordance index (c-index) (Harrell *et al.*, 1984). The c-index is also calculated as the ratio of concordant over comparable pairs. However, due to censoring not all possible pairs of datapoints are comparable. The condition for two observations to be comparable is that the ranking of their event times is known. Restricting attention towards right censoring, this condition is fulfilled for a pair of events and a pair containing one event and one censored observation for which the event time of the former comes before the censoring time of the latter. In order for a comparable pair to be concordant, the failure order as predicted by the utility should equal the observed failure order. Note that the definition of *concordant* is model dependent. For models modelling the failure time an increasing utility will indicate an increasing failure time. For models modelling the hazard however,

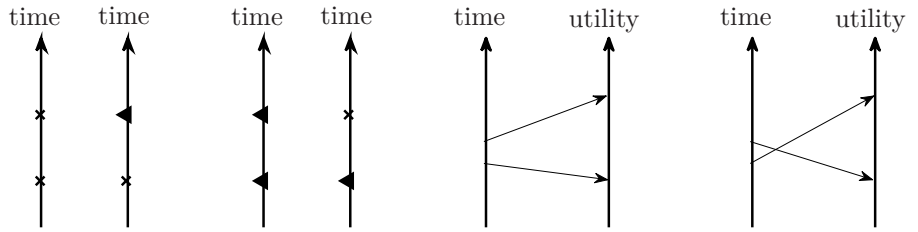


(a) varying threshold (b) ROC

Figure 2.6: *Illustration of the* ROC *curve. (a) By varying the threshold v on the utility $u(x)$, different values for the sensitivity and specificity are obtained. The true outcome is indicated by the circles. (b) The* ROC *curve is a visualization of these different values. A very low threshold (square) corresponds to a predicted outcome equal to 1 for nearly all observations. The sensitivity will be close to 1, but the specificity will be nearly zero. A very high threshold value (diamond) results in a predicted outcome of zero for nearly all observations, and the sensitivity will be close to zero, whereas the specificity will be close to 1. Better thresholds are found in between these extreme values (circle and triangle) and make a trade-off between sensitivity and specificity. The* AUC *(grey area in (b)) is a measure of the performance of the model. The dashed line illustrates the* ROC *curve for a random model.*

an increasing utility will correspond to a decreasing failure time. This difference needs to be taken into account when comparing different models.

Extending the ROC curve towards survival data is less straightforward. Akritas (1994) and Heagerty *et al.* (2000) proposed an extension of the ROC curve towards survival data by specifying a time-dependent sensitivity and specificity. The ROC curve at time t is the curve of the points sensitivity(v, t) and specificity(v, t) with v the value of the prognostic index under which one predicts no event and above which one predicts the event to occur. The time-dependent sensitivity is defined



(a) comparable (b) non comparable (c) concordant (d) non concordant

Figure 2.7: *Illustration of the generalization of the* AUC *towards to c-index. (a) Two observations are comparable when the order in which the event occurs is known. Two comparable pairs are illustrated: (i) two events (crosses), (ii) an event (cross) and a right censored observation (triangle) with the event time of the former less than the censoring time of the latter. In (b) two non comparable cases are illustrated. For comparable cases, one had to check whether the utility is concordant with the observed failure. In (c) two concordant observations are shown: their predicted event order equals the observed event order. In (d) two non concordant cases are shown.*

as:

$$\begin{aligned}
\text{sensitivity}(v,t) &= P\left(x > v | y \leq t\right), \qquad (2.43)\\
&= \frac{P\left(y \leq t | x > v\right) P(x > v)}{P(y \leq t)}\\
&= \frac{(1 - S\left(t | x > v\right))(1 - F_x(v))}{1 - S(t)}\\
&= \frac{1 - F_x(v) - S(t | x > v)\left(1 - F_x(v)\right)}{1 - S(t)}\\
&= \frac{1 - F_x(v) - S(v,t)}{1 - S(t)}.
\end{aligned}$$

In the above, $S(v,t)$ is defined as $S(v,t) = P(y > t|x > v)P(x > v) = S(t|x > v)(1 - F_x(v))$. The time-dependent specificity is defined analogously as:

$$\begin{aligned}
\text{specificity}(v,t) &= P\left(x \leq v | y > t\right) \\
&= \frac{P\left(y > t | x \leq v\right) P(x \leq v)}{P(y > t)} \\
&= \frac{S\left(t | x \leq v\right) P(x \leq v)}{P(y > t)} \\
&= \frac{\left[\frac{S(t) - S(t|x>v)(1-F_x(v))}{F_x(v)}\right] F_x(v)}{S(t)} \\
&= 1 - \frac{S(v,t)}{S(t)}\,.
\end{aligned} \tag{2.44}$$

Considering equations (2.43) and (2.44), defining a time dependent ROC curve reduces to defining an estimate for the bivariate survival function $S(v,t)$. Akritas (1994) proposed to use the following estimator:

$$\hat{S}(v,t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}(t|x = x_i)\mathcal{I}(x_i > v)\,, \tag{2.45}$$

where $\hat{S}(t|x = x_i)$ is a suitable estimator of the conditional survival function and $\mathcal{I}(z)$ the indicator function such that $\mathcal{I}(z) = 1$ if z is true. Define $\hat{S}(t|x = x_i)$ as a weighted Kaplan-Meier (Kaplan & Meier, 1958) estimator

$$\hat{S}(t|x = x_i) = \prod_{s \in \mathcal{Y}_n, s \leq t} \left(1 - \frac{\sum_j k(x_j, x_i)\mathcal{I}(y_j = s)\delta_j}{\sum_j k(x_j, x_i)\mathcal{I}(y_j \geq s)}\right)\,, \tag{2.46}$$

where $\mathcal{Y}_n$ represents all failure times in the data set $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$ and $k(\cdot,\cdot)$ is a kernel function. Akritas (1994) uses a zero-one nearest neighbor kernel

$$k(x, x_*) = \mathcal{I}(-\kappa < F_x(x) - F_x(x_*) < \kappa)\,, \tag{2.47}$$

with 2κ the percentage of observations that is included in each neighborhood. Following the above yields an ROC curve for each time point, for which an AUC can be calculated. The resulting AUCs can be plotted in function of time, yielding a curve representing the separability performance in function of time. The higher this curve the better the performance.

Table 2.1: *Illustration of the calculation of the logrank test. At every distinct failure time $l = 1, \ldots, k$, a 2×2 table of the observed events in function of the group label is calculated. $d_{0,l}$ and $d_{1,l}$ represent the number of events at time l in group 0 and 1, respectively. The number of observations at risk at this time point in each group, is indicated by $|\mathcal{R}_{0,l}|$ and $|\mathcal{R}_{1,l}|$.*

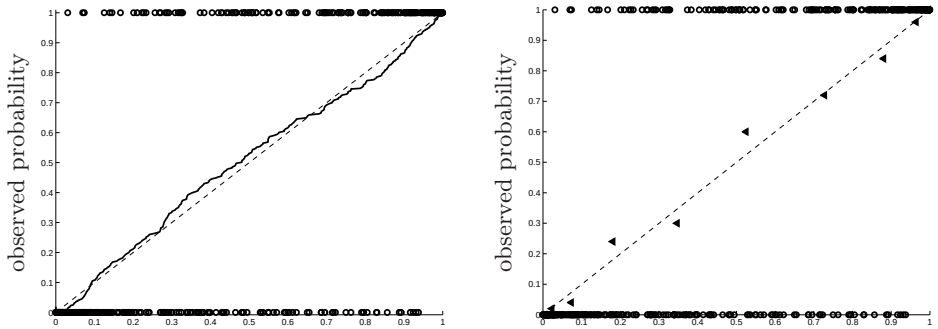
	observed event		
group	yes	no	total
1	$d_{1,l}$	$\lvert\mathcal{R}_{1,l}\rvert - d_{1,l}$	$\lvert\mathcal{R}_{1,l}\rvert$
0	$d_{0,l}$	$\lvert\mathcal{R}_{0,l}\rvert - d_{0,l}$	$\lvert\mathcal{R}_{0,l}\rvert$
total	d_l	$\lvert\mathcal{R}_l\rvert - d_l$	$\lvert\mathcal{R}_l\rvert$

In addition to measures related to the ROC curve, two more measures will be used throughout this dissertation. In clinical applications, one is often interested in defining risk groups. In this context it is relevant to test the ability of the obtained model to differentiate high and low risk observations. The observations are therefore divided into a low and high risk group according to whether their prognostic index is below or above the median value. The logrank test is used to test significant differences in survival between both groups. Instead of working with the resulting p-value, we will report on the χ^2-value of the test. The higher this value, the better the separating ability of the model. The logrank test (Mantel, 1966) is calculated as follows. The test is obtained by constructing a set of 2×2 tables at each distinct event time. At each distinct event time $l = 1, \ldots, k$ a table as in Table 2.1 can be obtained. The test is defined as

$$\text{logrank } \chi^2 = \frac{\left(\sum_{l=1}^{k}\left(d_{1,l} - \frac{|\mathcal{R}_{1,l}|d_l}{|\mathcal{R}_l|}\right)\right)^2}{\sum_{l=1}^{k}\frac{|\mathcal{R}_{0,l}||\mathcal{R}_{1,l}|d_l\,(|\mathcal{R}_l| - d_l)}{|\mathcal{R}_l|^2(|\mathcal{R}_l| - 1)}}, \tag{2.48}$$

where $d_{0,l}$ and $d_{1,l}$ represent the number of events at time l in group 0 and 1, respectively. The number of observations at risk at this time point in each group, is indicated by $|\mathcal{R}_{0,l}|$ and $|\mathcal{R}_{1,l}|$. The logrank χ^2 is compared with a χ^2 distribution with $k-1$ degrees of freedom. The resulting p-value is interpreted as the probability of observing a more extreme result than that seen in the data, under the hypothesis of equal survival distributions.

Finally, models will be compared by means of the hazard ratio. Since utilities can vary in values depending on the model they are derived from, all utilities will be normalized to the unit interval. These normalized utilities will then be fed to a PH, to obtain an estimate of the hazard ratio. Models with a higher hazard ratio are preferred.



(a) non-parametric technique (b) grouped observations

Figure 2.8: *Calibration plot comparing the predicted probability of an event with the observed probability estimated (a) in a non-parametric way, using a loess smoother (solid line) and (b) by grouped observations (triangles). Ideally, the smoothed line and grouped observations should coincide with the 45° line (dashed line). Observations are indicated by the circles.*

2.4.2 Calibration

A second measure of performance is calibration. Models who present the user with an estimate of the predicted probability of the event, can be checked for calibration. Calibration answers the question whether the predicted probability corresponds with the observed probability. Calibration plots have predicted probabilities on the x-axis and observed probabilities on the y-axis. A model is well calibrated when the resulting curve corresponds with the 45° line. For binary outcomes, probabilities are not observed directly and smoothing techniques (e.g. loess (locally estimated scatterplot smoothing) (Cleveland, 1979)) are used to obtain a calibration plot. A second approach is to divide the observations in deciles according to the predicted probability and calculate the mean predicted and observed probability in these groups. Figure 2.8 illustrates both approaches. For survival data, the calibration plots are considered at fixed time points (e.g. 5 years of follow-up). Grouped observations are used to calculate Kaplan-Meier survival curves in order to get an estimate of the observed survival, which is then compared with the mean predicted survival of the model. Note that the number of observations within each group should be large enough to calculate the Kaplan-Meier curves.

2.5 Conclusion

This Chapter started with a brief summary of survival distributions, a non-parametric survival estimator and survival models. The popular PH was introduced together with the partial likelihood, lying on the basis of the parameter estimation within this model. Both the linearity and proportional hazards assumption of the model were explained. The PLANN model of Biganzoli *et al.* (1998) and the PLANNARD model of Lisboa *et al.* (2003) were illustrated. The advantages of this model above the PH were discussed. In a second part of this Chapter, the basics of SVMs and LS-SVMs were explained. In a final part, different evaluation measures were discussed. In this Section, we started with measures encountered in standard classification and/or regression models and showed how they can be extended for censored survival data.

Chapter 3

Transformation models for ranking and survival analysis

This Chapter studies transformation models for ranking, ordinal regression and survival analysis. The link between standard statistical models and models based on support vector machines is found within the framework of transformation models. The relation between the AUC *and monotone transformation functions lies at the basis of this link. The maximal margin strategy of Vapnik is replaced by the notion of the Lipschitz continuity constant, controlling the model complexity. The presented model is illustrated for ranking problems, ordinal regression and survival analysis. Finally, it is shown how the model can be adapted to deal with high dimensional data. The results of this Chapter can be found in Van Belle* et al. *(2009).*

3.1 Introduction

Methods based on ranking continue to challenge researchers in different scientific areas, see e.g. Clémençon *et al.* (2005); Herbrich *et al.* (2000) and the references therein. Learning ranking functions offers a solution to different types of problems including ordinal regression and bipartite ranking, studied frequently in research on information retrieval. These cases distinguish themselves in the definition of the output domain and the chosen loss function. This Chapter deals with the general problem where the output domain can be arbitrary, but possesses a natural ordering relation between the members.

Earlier approaches to learning preference functions reduce the ranking problem to pairwise classification problems. This reasoning was followed in Herbrich *et al.* (1998); Fürnkranz & Hüllermeier (2003); Ailon & Mohri (2008); and references therein. However, functions having high pairwise margins might still be bad approximations to real ranking problems. We address this issue by presenting a conceptual different approach: we adopt a smoothness condition on the ranking function to structure the space of ranking functions, and claim that this structure aligns in many applications better with the learning problems. This reasoning is motivated from relating a pairwise ranking criterion to a monotone *transformation* function. Besides empirical validation of this claim, we present formal relationships to other (statistical) models used for such tasks.

Figure 3.1 summarizes the ideas exposed in this Chapter. First we describe the class of transformation models which contains two different components. The first component of a transformation model is a function $u : \mathbb{R}^d \to \mathbb{R}$ mapping the covariates $x \in \mathbb{R}^d$ to a value in $\mathbb{R}$ such that the natural order on $\mathbb{R}$ induces the ranking (approximately). Different names for such a function are found in the literature depending on the problem setting, including a scoring, ranking, utility or health function. We will refer to this as to the utility function. The second component of the model maps this utility to an outcome in $\mathbb{R}$ by a transformation function $h : \mathbb{R} \to \mathbb{R}$. This is a univariate monotonically increasing function, basically capturing the scale of the output. The central observation is that when one knows the ordinal relations between instances, one can estimate a transformation function mapping the instances to their utility value $\hat{u}(x)$. Depending on the problem at hand one is interested in the results of the first or second component of the transformation model. This Chapter focuses on finding an optimal utility function. More information on capturing the transformation function is given in Chapter 5.

Transformation models are especially appropriate when considering data arising from a survival study. While practice and theoretical results continue to have a strong impact in most quantitative scientific areas, survival analysis has been studied only sporadically in a context of machine learning, and such studies are mostly found in the field of artificial neural networks, see e.g. Biganzoli *et al.* (1998); Kattan *et al.* (1997). However, we claim that there is a large potential for such studies. (i) The approach of classical likelihood based approaches have their intrinsic limitations, especially when a realistic underlying model cannot be assumed. A distribution-free approach is more appropriate here. (ii) A risk-based approach is often easier as one does not care about recovering the exact parameters describing the process of interest, but one is only interested in making good predictions, or exploring structure in the data. (iii) Computational issues for the classical statistical approach persist, and the question how to solve the estimation equations numerically is often approached in an ad hoc way (see Kalbfleisch & Prentice (2002) and references).

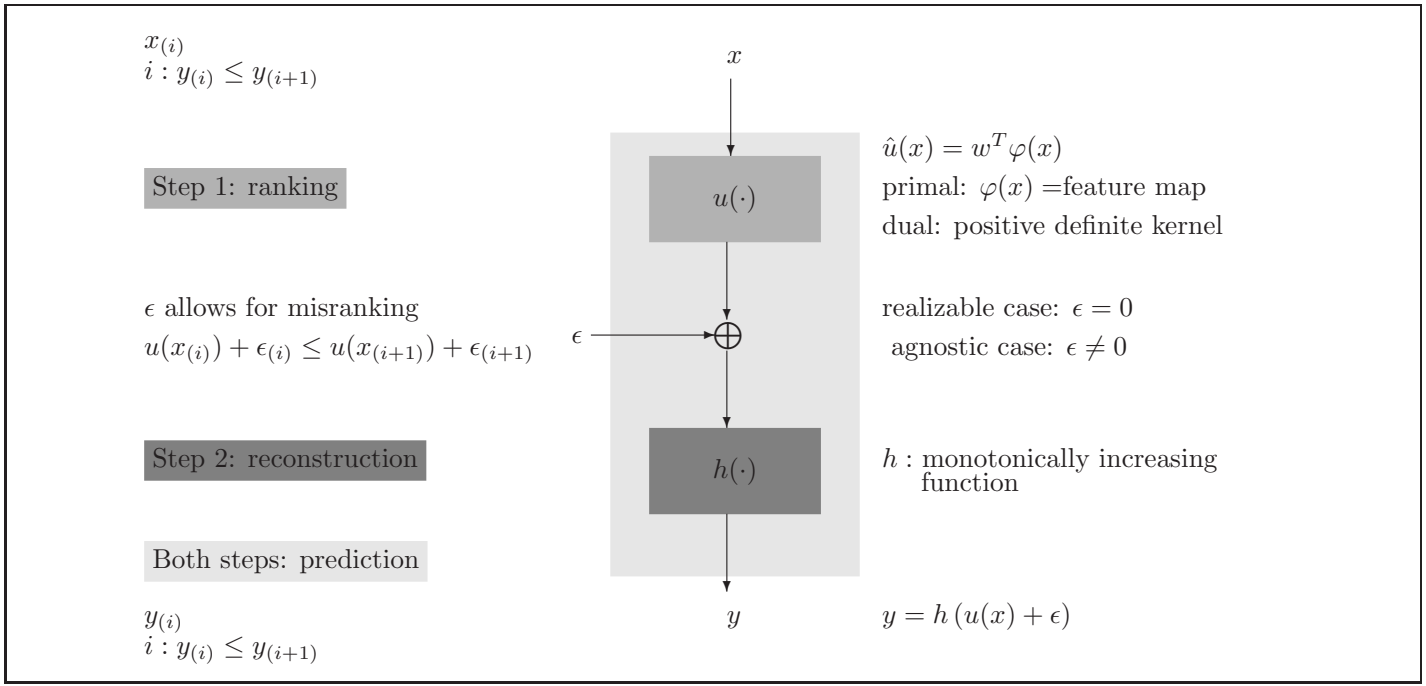


Figure 3.1: *Overview: Transformation models consist of two components, the* utility function *u and a* transformation function *h. Given a data set $\mathcal{D} = \{(x_{(i)}, y_{(i)})\}_{i=1}^n$ where the observations are sorted such that $y_{(i)} \leq y_{(i+1)}$, a utility function $u(x) = w^T\varphi(x)$ is trained such that the ranking on the evaluations of this function is representative for the ranking on the outcome. In the* realizable case *the ordering in utility will exactly coincide with the ordering in observed outcome $\{y_{(i)}\}_i$. In the* agnostic case *however, the ordering will only be exact up to appropriate error variables $\{\epsilon_i\}_i$. The modelling procedure will also be performed in two steps. The first step recovers u (*ranking*), while the second step is concerned with learning an explicit representation of the transformation function (*reconstruction*).*

We find that the class of transformation models is a powerful tool to model data arising from survival studies for different reasons. The first reason being that they separate the model for the time-scale (via the transformation function), and the qualitative characterization of an instance (via the utility function). We will furthermore argue on the close relationship with existing techniques as PH and accelerated failure time (AFT) models, see e.g. Dabrowska & Doksum (1988); Koenker & Geling (2001); Cheng *et al.* (1997) and citations. We will relate the transformation function to ranking criteria as Kendall's τ or area under the curve (AUC), hence outlining a unified framework to study survival models as used in a statistical context and machine learning techniques for learning ranking functions. This relation indicates how one may apply the method of structural risk minimization (SRM, see Vapnik (1998)) here. The immediate consequence is the possibility to apply learning theory, with the capabilities to explain good performances in modelling high-dimensional datasets as well as for non-linear models (see Vapnik (1998)). Finally, it will be illustrated how the proposed approach can handle censored observations conveniently.

This and the following Chapter consider the utility function, either linear functions, or kernel based models. Instead of restricting attention to a parameterized class of transformation functions the transformation function is left unspecified as in partial likelihood approaches (see Kalbfleisch & Prentice (2002)). The transformation function is only defined on the observed samples. This is done by inferring a set of ordinal relations between them. The Lipschitz continuity constant associated to such a transformation function can then be evaluated based on the samples only. Consequently, the presented fitting strategy (MINLIP) finds the maximally smooth (implicitly defined) transformation function fitting the data samples. This is the *realizable case* where we can make the assumption of the existence of such a transformation model. In case misfit is allowed, the model is extended using slack variables. The resulting problem can be solved as a convex QP, for which highly efficient software is readily available. In the case of utility functions which are kernel based models, the solution is found as a sum of positive definite kernels, and the Lagrange dual problem solves the corresponding problem as a convex QP. For the case linear utility functions are considered, we suggest two different ways how one can obtain sparseness using an 1-norm regularization mechanism (see also Tibshirani (1996)), and using positivity constraints on the parameters.

Besides the conceptual and computational discussion, this Chapter gives empirical evidence for the approach. We consider empirical studies of ordinal regression and survival analysis. Performance of MINLIP on ordinal regression is analyzed using the ordinal data compiled by Chu & Keerthi (2005). MINLIP is applied on two different survival studies. A first study involves micro-array datasets: two breast cancer datasets (Sørlie *et al.*, 2003; van Houwelingen *et al.*, 2006) and one data set concerning diffuse large B-cell carcinoma (Rosenwald *et al.*, 2002). In a last

study, concerning a clinical breast cancer survival study (Schumacher *et al.*, 1994), we investigate the estimation of non-linear covariate effects and compare results obtained with MINLIP with PH with penalized smoothing splines.

In Van Belle *et al.* (2007) we proposed a modification to standard SVMs to handle censored data. A computationally less demanding algorithm was presented in Van Belle *et al.* (2008). Starting from this latter model, we replaced the maximal margin strategy with the minimal Lipschitz continuity strategy as presented in Van Belle *et al.* (2009a). This Chapter extends the results of this short paper. The different application areas in which the proposed method can be applied are summarized in Table 3.1. In addition, it is stated how the model needs to be used and which equations need to be solved to obtain the solution.

This Chapter is organized as follows. The following Section discusses in some detail the use of transformation models and its relation with ranking methods. Section 3.3 studies the estimator in a context of ranking. Section 3.4 specifies how MINLIP is to be used in a context of ordinal regression, where only k different output levels are possible. Section 3.5 discusses the use of MINLIP in the presence of censoring. In Section 4.5 experiments illustrate the use of the MINLIP method.

3.2 Transformation models and ranking methods

This Section starts with exploring the relation between transformation models and ranking methods. However, some terminology needs to be defined before this exploration can start.

Definition 1 (Lipschitz continuity)**.** *A univariate function $h(z)$ has a Lipschitz constant $\ell \geq 0$ if*

$$|h(z) - h(z')| \leq \ell |z - z'|, \ \forall z, z' \in \mathbb{R}\,. \tag{3.1}$$

A *transformation model* is then defined as follows.

Definition 2 (Transformation model)**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be a* strictly *increasing function with Lipschitz constant $\ell < \infty$, and let $u : \mathbb{R}^d \to \mathbb{R}$ be a function of the covariates $x \in \mathbb{R}^d$. Let ϵ be a random variable (noise) independent of x, with cumulative distribution function $F_\epsilon(e) = P(\epsilon \leq e)$ for any $e \in \mathbb{R}$. Then a Noisy Transformation Model (*NTM*) takes the form*

$$y = h(u(x) + \epsilon). \tag{3.2}$$

Let z denote $u(x) + \epsilon$ for notational convenience. Now the problem is reduced to estimating a utility function $u : \mathbb{R}^d \to \mathbb{R}$ and a transformation function h

Table 3.1: *Overview of methods and applications proposed in this Chapter. Depending on the problem at hand, a different version of the proposed model needs to be applied. Depending on the subtasks, different training data (variables x_i, target value y_i, utility $u(x_i)$, dummy responses s, ...) need to be given to the algorithm to obtain the desired response (utility $u(x_i)$, transformation function $h(u(x))$, prediction $\hat{y}_i = h(u(x_i))$, threshold values $\hat{v}$, risk on event within the lth interval $\hat{y}_{il}$, ...).*

Task	Subtasks	algorithm	necessary training data for algorithm	result of algorithm	equation
ranking	ranking	MINLIP	$\{(x_i, y_i)\}$	$\hat{u}(x)$	(3.24)
ordinal regression	ranking	MINLIP	$\{(x_i, y_i)\}, s$	$\hat{u}(x), \hat{v}$	(3.33)
	reconstruction	comparison with thresholds v	$\{\hat{u}(x_i)\}, \hat{v}$	$\hat{y}$	Figure 3.5
survival analysis	ranking	MINLIP	$\{(x_i, y_i)\}$	$\hat{u}(x)$	(3.45)
	reconstruction	monotonic regression after replication in consecutive time intervals	$\{(\hat{u}(x_i), y_i)\}$	$\{\hat{y_{il}}\}$	

from a set of i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ without imposing any distributional (parametric) assumptions on the noise terms $\{\epsilon_i\}$. Note that without structural assumptions, the utility can not uniquely be defined. Later on, we will specify similar assumptions as in the maximal margin strategy of Vapnik when introducing support vector machines, to find a unique solution for the utility function.

The transformation models discussed in Cheng *et al.* (1997); Dabrowska & Doksum (1988); Koenker & Geling (2001) differ from the above definition in the transformation function h. They define the model as $g(y) = u(x) + \epsilon$, which is equivalent to (3.2) if h is invertable.

To express the performance of a ranking function one can use Kendall's τ, AUC or a related measure. We will work with the concordance of a function $u : \mathbb{R}^d \to \mathbb{R}$ respective to the outcome (see Chapter 2). The concordance is defined as the probability that the order in outcome of two i.i.d. observations (x, y) and (x', y') is preserved in the utility u:

$$\mathrm{C}(u) = P((u(x) - u(x'))(y - y') > 0). \tag{3.3}$$

Given a set of n i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$, the empirical concordance index is then calculated as

$$\mathrm{C}_n(u) = \frac{2}{n(n-1)} \sum_{i<j} \mathcal{I}[(u(x_i) - u(x_j))(y_i - y_j) > 0]\,. \tag{3.4}$$

Equivalently, the risk is defined as follows.

Definition 3 (Risk of (h, u))**.** *The risk associated with a monotonically increasing function penalizes discordant samples $h(u(x))$ and $h(u(x'))$ as*

$$\mathrm{R}(u) = P((h(u(x)) - h(u(x')))(y - y') \leq 0). \tag{3.5}$$

Or, since h is monotonically increasing, the risk is expressed as

$$\mathrm{R}(u) = P((u(x) - u(x'))(y - y') \leq 0). \tag{3.6}$$

Its empirical counterpart then becomes

$$\mathrm{R}_n(u) = 1 - \mathrm{C}_n(u)\,. \tag{3.7}$$

Empirical Risk Minimization (ERM) is then performed by solving

$$\hat{u} = \arg\min_{u \in \mathcal{U}} \mathrm{R}_n(u) = \arg\max_{u \in \mathcal{U}} \mathrm{C}_n(u), \tag{3.8}$$

where $\mathcal{U} \subset \{u : \mathbb{R}^d \to \mathbb{R}\}$ is an appropriate subset of ranking functions, see e.g. Clémençon *et al.* (2005) and citations. However, this approach results in

combinatorial optimization problems. One therefore majorizes the discontinuous indicator function by the Hinge loss (i.e. $\max(0, 1-z)$) yielding rankSVM (Herbrich *et al.*, 2000).

$$\boxed{\text{rankSVM}}$$

$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \sum_{j:y_i > y_j} \epsilon_{ij} \\ \text{subject to} \quad & \begin{cases} w^T(\varphi(x_i) - \varphi(x_j)) \geq 1 - \epsilon_{ij}, \\ \qquad \forall\, i = 1, \ldots, n;\ \forall\, j : y_i > y_j \\ \epsilon_{ij} \geq 0, \quad \forall\, i = 1, \ldots, n;\ \forall\, j : y_i > y_j\,. \end{cases} \end{aligned} \tag{3.9}$$

The disadvantage of this formulation is that it leads to $O(n^2)$ constraints or unknowns, often making it difficult to apply the method to real life problems. A solution to this problem is found in relating transformation models with equation (3.8). If a function $u : \mathbb{R}^d \to \mathbb{R}$ exists such that $\mathrm{C}_n(u) = 1$, one describes implicitly a transformation function (see Figure 3.2). If two variables u and y are perfectly concordant, then there exists a monotonically increasing function h such that $h(u)$ and y are perfectly concordant. Moreover, there exists such a function h, with Lipschitz constant ℓ, mapping u to y such that $y = h(u)$. This is stated more formally as follows.

Lemma 1 (Existence of a transformation function)**.** *Given a collection of pairs* $\{(z_{(i)}, y_{(i)})\}_{i=1}^{n}$*, sorted such that* $y_{(i)} \leq y_{(j)}$ *if and only if* $i \leq j$*, and considering the conditions on the observations for* $\ell < \infty$*:*

$$0 \leq y_{(i)} - y_{(j)} \leq \ell \left(z_{(i)} - z_{(j)}\right) \quad \forall\, i < j = 1, \ldots, n, \tag{3.10}$$

we state that:

1. *If one has for a finite value* $\ell \geq 0$ *that (3.10) holds, then there exists a monotonically increasing function* $h : \mathbb{R} \to \mathbb{R}$ *with Lipschitz constant* ℓ *interpolating the data points.*
2. *If for all admissible* $(z, y) \in \mathbb{R} \times \mathbb{R}$ *one has that* $y = h(z)$ *for an (unknown) continuous, (finite) differentiable and monotonically increasing function* $h : \mathbb{R} \to \mathbb{R}$*, then there is a value* $\ell < \infty$ *such that (3.10) holds.*

Proof. To prove 1, consider the linear interpolation function $h_n : \mathbb{R} \to \mathbb{R}$, defined as

$$h_n(z) = \frac{z - z_{\underline{z}(z)}}{z_{\overline{z}(z)} - z_{\underline{z}(z)}} \left(y_{\overline{z}(z)} - y_{\underline{z}(z)}\right) + y_{\underline{z}(z)} \tag{3.11}$$

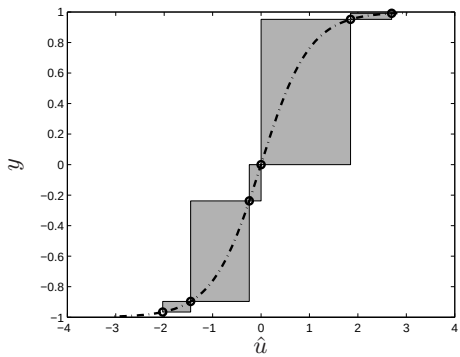


Figure 3.2: *Relation between ranking and transformation models: if two variables u and y are perfectly concordant, they describe a monotonically increasing function* $y = h(u)$. *The dots represent u and outcome y for training points. In these observations the value of the function h is known exactly. To predict the y−value of the test observations, the function h needs to be approximated between the training points. All functions* $\hat{h}$ *which are monotonically increasing and lie within the grey zones are valid prediction rules.*

where we define $\overline{z}(z) = \arg \min_{i \in \{1,\ldots,n\}} (z_i : z_i > z)$ and $\underline{z}(z) = \arg \max_{i \in \{1,\ldots,n\}} (z_i : z_i \leq z)$. Direct manipulation shows that this function is monotonically increasing and continuous. Now take $z < z' \in \mathbb{R}$, then we have to show that $h_n(z') - h_n(z) \leq \ell(z' - z)$. For notational convenience define $l = \underline{z}(z)$, $u = \overline{z}(z)$, $l' = \underline{z}(z')$ and $u' = \overline{z}(z')$, then

$$\begin{aligned} h_n(z') - h_n(z) &= \frac{z' - z_{l'}}{z_{u'} - z_{l'}}(y_{u'} - y_{l'}) + y_{l'} - \frac{z - z_l}{z_u - z_l}(y_u - y_l) - y_l \\ &\leq \ell(z' - z_{l'}) - \ell(z - z_l) + \ell(z_{l'} - z_l) \qquad (3.12) \\ &= \ell(z' - z), \end{aligned}$$

where we use that $y_{l'} - y_l \leq \ell(z_{l'} - z_l)$.

Item 2 is proven as follows. Let such an h exist, then the mean value theorem asserts that for any two samples (z_i, y_i) and (z_j, y_j) for which $z_i \leq z_j$, there exists a z within the interval $[z_i, z_j] \subset \mathbb{R}$ such that

$$(y_j - y_i) = (z_j - z_i)h'(z) \leq \ell(z_i - z_j) \qquad (3.13)$$

where $\ell = \sup_z h'(z)$.
Note that equation (3.10) implies that $\mathrm{C}_n(z) = 1$. □

3.3 A convex approach to learning transformation models

3.3.1 The realizable case

The realizable case refers to the situation where there exists a function $u(x)$ such that the ranking of $u(x)$ perfectly reflects the ranking of y. Otherwise stated, there exists a function $u(x)$ such that $\mathrm{C}_n(u) = 1$. Lemma 1 describes the existence of h, but since this transformation function is only known at the training points, it is not unique. Figure 3.2 illustrates that all monotonically increasing functions lying within the grey bounds satisfy the conditions. Therefore, the Lipschitz constant is used to control the complexity of the transformation function. Transformation functions with a smaller Lipschitz constant will be preferred. For notational convenience, we will assume no coinciding outcomes (ties). Let h be a monotonically increasing function with Lipschitz constant $\ell < \infty$, such that $h(z) - h(z') \leq \ell(z - z')$ for all $z \leq z'$. Restricting attention to the observations $\{(x_{(i)}, y_{(i)})\}_{i=1}^n$, one has the necessary conditions

$$h\left(u(x_{(i)})\right) - h\left(u(x_{(i-1)})\right) \leq \ell\left(u(x_{(i)}) - u(x_{(i-1)})\right) ,$$

for all $i = 2, \ldots, n$. Here, we assume that the data obey a noiseless transformation model ($\epsilon = 0$ in (3.2)). For now, linear utility functions defined as

$$u(x) = w^T x \,, \tag{3.14}$$

are considered. Extensions towards non-linear utility functions using Mercer kernels are handled in Subsection 3.3.3. Since the function $u(x) = w^T x$ can be arbitrary rescaled such that the corresponding transformation function has an arbitrary Lipschitz constant (i.e. for any $b > 0$, one has $h(u(x)) = \tilde{h}(\tilde{u}(x))$ where $\tilde{h}(z) \triangleq h(b^{-1}z)$ and $\tilde{u}(x) = bu(x)$), we fix the norm $w^T w$ and try to find $u(x) = v^T x$ with $v^T v = 1$. Hence learning a transformation model with minimal Lipschitz constant of h can be written as

$$\begin{array}{ll} \min\limits_{v,\ell} & \frac{1}{2}\ell^2 \\ \text{subject to} & \begin{cases} ||v||_2 = 1 \\ y_{(i)} - y_{(i-1)} \leq \ell\left(v^T x_{(i)} - v^T x_{(i-1)}\right), \quad \forall\, i = 2, \ldots, n. \end{cases} \end{array} \tag{3.15}$$

Substituting $w = \ell v$ we get equivalently:

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2} w^T w \\ \text{subject to} & y_{(i)} - y_{(i-1)} \leq w^T x_{(i)} - w^T x_{(i-1)}, \quad \forall\, i = 2, \ldots, n \,, \end{array} \tag{3.16}$$

which goes along similar lines as the hard margin SVM (see e.g. Shawe-Taylor & Cristianini (2004)) and ranking SVM (Herbrich *et al.*, 2000), where the threshold

value 1 is replaced by $y_{(i)} - y_{(i-1)}$. Note that an intercept term is not needed since differences in utility are used. This problem has $n-1$ linear constraints. We will refer to this model as MINLIP.

In the presence of ties, $y_{(i-1)}$ is replaced by $y_{(j)}$, with j the largest output value with $y_{(j)} < y_{(i)}$. See Section 3.4 for more details. Solving this problem as a convex QP can be done efficiently with standard mathematical solvers as implemented in MOSEK[1] or R-quadprog[2]. The following proposition states when the MINLIP model is valid.

Proposition 1 (Validity of MINLIP)**.** *Assume that* $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ *would obey the relation*

$$y = h_0(w_0^T x), \tag{3.17}$$

where we refer to the fixed but unknown vector $w_0 \in \mathbb{R}^d$ *as to the true parameters, and to the fixed but unknown monotonically increasing function* $h_0 : \mathbb{R} \to \mathbb{R}$ *as the true transformation function. Let for each couple* (x, y) *and* (x', y') *where* $y \neq y'$ *the constant* $\ell' > 0$ *be defined as*

$$\frac{1}{\ell'} = \frac{w_0^T(x - x')}{y - y'}, \tag{3.18}$$

where $\ell' = \infty$ *if* $w_0^T(x - x') = 0$*. By construction we have that* $\ell' \leq \ell_0$*, with* ℓ_0 *the true Lipschitz constant, and that the constant exists everywhere. The result of the* MINLIP *model then becomes:*

$$\hat{w} = \max_{||w||_2=1} \min_{y>y'} \frac{w^T(x - x')}{y - y'} \quad = \quad \max_{||w||_2=1} \min_{y>y'} \frac{w^T(x - x')}{\ell' w_0^T(x - x')}. \tag{3.19}$$

We then state that MINLIP *yields a good approximation of the parameter vector* w_0 *in the noiseless case as long as there are enough observations* (x, y) *such that* $w_0^T(x - x') \approx 1$ *and* $\ell' \approx \ell_0$*.*

Let the unit-length vector $\overline{(x - x')} \in \mathbb{R}^d$ be defined as $x - x' = \overline{(x - x')}|x - x'|_2$, then we can write (3.19) as

$$\hat{w} = \max_{||w||_2=1} \min_{y>y'} \frac{w^T\overline{(x - x')}}{\ell' w_0^T\overline{(x - x')}}. \tag{3.20}$$

Let us now focus attention on the set $\mathcal{S} = \{(x - x', y - y') : (x, y), (x', y') \in \mathcal{D} = \{x_i, y_i\}_{i=1}^n$, where $\ell' = w^T(x - x')/(y - y')$ for which this value ℓ' is actually achieved. It is seen that the estimate w lies in the span of this set $\mathcal{S}$ as otherwise the

[1] http://www.mosek.org

[2] http://cran.r-project.org/web/packages/quadprog/index.html

maximum value could be increased. When we assume that the data set contains enough observations (x, y) such that $w_0^T(x - x') \approx 1$ and $\ell' \approx \ell_0$, they will end up in the set $\mathcal{S}$, and as a result we have that $w^T w_0 \approx 1$. As the optimal solution is fully determined by the terms $w_0^T \overline{(x - x')} \approx 1$ and $\ell' \approx \ell_0$ (cfr. duality results in convex optimization), one should also have that $w \approx w_0$.

3.3.2 The agnostic case

In case it is impossible to find a utility function $u : \mathbb{R}^d \to \mathbb{R}$ extracting the ranking perfectly, a noisy transformation model is considered:

$$y = h(w^T x + \epsilon)\,, \tag{3.21}$$

where $u = w^T x$. The introduction of the error variable asks for an adaptation of the Lipschitz-based complexity control. We choose the absolute value loss as a loss function for three reasons. (i) It is known that this loss function is more robust to misspecification of the model and outliers than e.g. the squared loss. (ii) The use of the absolute value loss will result in sparse solutions with many error terms equal to zero. (iii) In binary classification this norm is well performing in SVMs. However, the choice of the loss remains arbitrary. Incorporation of the errors (slack variables) leads to the following model formulation:

$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} |\epsilon_i| \\ \text{subject to} \quad & w^T (x_{(i)} - x_{(i-1)}) \geq y_{(i)} - y_{(i-1)} - \epsilon_i + \epsilon_{i-1}, \\ & \qquad\qquad\qquad\qquad \forall\ i = 2, \ldots, n\,, \end{aligned} \tag{3.22}$$

where $\gamma > 0$ is a regularization constant, making a trade-off between model complexity and error. This problem can again be solved as a convex QP.

3.3.3 A non-linear extension using Mercer kernels

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^{d_\varphi}$ be a feature map mapping the data to a high dimensional feature space of dimension d_φ, which can be infinite. A non-linear utility function can then be defined as

$$u(x) = w^T \varphi(x), \tag{3.23}$$

with $w \in \mathbb{R}^{d_\varphi}$ a vector of unknowns. The realizable learning problem can then be represented as:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w^T w \\ \text{subject to} \quad & w^T (\varphi(x_{(i)}) - \varphi(x_{(i-1)})) \geq y_{(i)} - y_{(i-1)}, \quad \forall\ i = 2, \ldots, n\,. \end{aligned} \tag{3.24}$$

The Lagrange dual problem becomes

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,j=2}^{n} \alpha_i \alpha_j \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right)^T \left(\varphi(x_{(j)}) - \varphi(x_{(j-1)})\right) \\ & - \sum_{i=2}^{n} \alpha_i (y_{(i)} - y_{(i-1)}) \\ \text{subject to} \quad & \alpha_i \geq 0, \ \forall\ i = 2, \ldots, n\,. \end{aligned} \tag{3.25}$$

As in Chapter 2, the inner product $\varphi(x_{(i)})^T \varphi(x_{(j)})$ is replaced by evaluating a positive definite kernel function $k(x_{(i)}, x_{(j)}) = \varphi(x_{(i)})^T \varphi(x_{(j)})$. The estimated utility $\hat{u}$ can be evaluated at any point $x_\star \in \mathbb{R}^d$ as

$$\hat{u}(x_\star) = \sum_{i=2}^{n} \alpha_i \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right)^T \varphi(x_\star)\,. \tag{3.26}$$

The dual (Shawe-Taylor & Cristianini (2004); Suykens *et al.* (2002); Vapnik (1998)) of the agnostic learning machine of Subsection 3.3.2 is obtained analogously:

MINLIP

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,j=2}^{n} \alpha_i \alpha_j \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right)^T \left(\varphi(x_{(j)}) - \varphi(x_{(j-1)})\right) \\ & - \sum_{i=2}^{n} \alpha_i (y_{(i)} - y_{(i-1)}) \\ \text{subject to} \quad & \begin{cases} \alpha_i \geq 0, & \forall\ i = 2, \ldots, n, \\ -\gamma \leq \alpha_i - \alpha_{i+1} \leq \gamma, & \forall\ i = 2, \ldots, n-1, \\ -\gamma \leq \alpha_2 \leq \gamma, \\ -\gamma \leq \alpha_n \leq \gamma\,. \end{cases} \end{aligned} \tag{3.27}$$

A detailed derivation is given in Appendix B.1.

In cases where one is interested in an estimate of the individual covariate effects, one could use an additive utility function:

$$u(x) \quad = \quad \sum_{p=1}^{d} u^p(x^p)\,, \tag{3.28}$$

where x^p represents the pth covariate of observation x. Using equation (3.26), this can be written as:

$$\hat{u}(x) \quad = \quad \sum_{i=2}^{n} \alpha_i \sum_{p=1}^{d} \left(\varphi(x^p_{(i)}) - \varphi(x^p_{(i-1)})\right)^T \varphi(x^p))\,, \tag{3.29}$$

As a result, componentwise kernels (Pelckmans *et al.*, 2005c):

$$k(x, x_i) = \sum_{p=1}^{d} k^p(x^p, x_i^p)\,, \tag{3.30}$$

which can be seen as a special case of ANOVA kernels (Vapnik, 1998), can be used. The use of such componentwise kernels allows for interpreting the non-linear effects of the covariates.

3.3.4 Toward sparse solutions using L_1 regularization and sparsity constraints

This subsection describes some straightforward but powerful extensions to the above model. Specifically, we will be interested in the case where d is large compared to n. Consequently, we will be interested in computational methods which reveal the relevant input variables of use in the learned prediction rule. We will consider two mechanisms for obtaining sparse representations in the linear coefficients. We restrict ourselves to the linear case where $u(x) = w^T x$ and an unknown monotonically increasing function $h : \mathbb{R} \to \mathbb{R}$. In a first extension an L_1 penalty (Tibshirani, 1996) is used instead of the term $w^T w$. We shall refer to this model as MINLIP$_{L1}$:

$$\boxed{\text{MINLIP}_{L1}} \qquad \begin{array}{ll} \min\limits_{w,\epsilon} & \displaystyle\sum_{p=1}^{d} |w_p| + \gamma \sum_{i=1}^{n} |\epsilon_i| \\ \text{subject to} & w^T(x_{(i)} - x_{(i-1)}) \geq y_{(i)} - y_{(i-1)} - \epsilon_i + \epsilon_{i-1}, \\ & \qquad\qquad \forall\ i = 2, \ldots, n\,, \end{array} \tag{3.31}$$

where w_p represents the pth component of the vector w. This linear programming problem (LP) can be solved efficiently with standard mathematical solvers.

As an alternative to L_1-norm regularization, we propose a second extension incorporating positivity constraints of the parameters, which we will denote as MINLIP$_p$:

$$\boxed{\text{MINLIP}_{p}} \qquad \begin{array}{ll} \min\limits_{w,\epsilon} & \displaystyle\frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} |\epsilon_i| \\ \text{subject to} & \left\{ \begin{array}{ll} w^T(x_{(i)} - x_{(i-1)}) \geq y_{(i)} - y_{(i-1)} - \epsilon_i + \epsilon_{i-1}, & \\ \qquad\qquad \forall\ i = 2, \ldots, n, & \\ w_p \geq 0, & \forall\ p = 1, \ldots, d\,. \end{array} \right. \end{array} \tag{3.32}$$

This can be useful, even if the problem of interest does not strictly imposes those exact requirements. As a preprocessing step one should calculate the correlation between each input variable x^p with $p = 1, \ldots, d$, and the outcome. Variables x^p which are inversely related with the outcome should be transformed as $x^p := -x^p$.

Figure 3.3 illustrates the possible advantage of both sparse alternatives over the standard MINLIP formulation. We created 100 artificial datasets, each containing 150 observations with 200 covariates. 100 observations were used for training, the remaining for testing. A varying number of $d = 100, 110, \ldots, 200$ covariates were used to build the outcome, all other features being irrelevant. All covariates were drawn from a normal distribution with mean 0 and standard deviation 1. The outcome was obtained as a weighted sum of the relevant covariates, where the weights were drawn from a standard normal distribution. The test error of the MINLIP$_{L1}$ model was a bit lower than for the standard model. However, the MINLIP$_p$ model performed significantly better than both other models. Where the MINLIP model selects all variables (absolute value of the estimated weight $> 10^{-8}$), the MINLIP$_{L1}$ model selects too few variables. The MINLIP$_p$ model obtains the best trade-off between selected and correctly selected features.

3.3.5 Comparison with other methods

An approach which is often used within ranking problems is the reformulation of the ranking problem as a classification problem. Examples of this strategy can be found in Herbrich *et al.* (1998); Fürnkranz & Hüllermeier (2003); Ailon & Mohri (2008). However, transforming ranking to classification deflects attention from the underlying problem within ranking problems. In contrast with these methods, the MINLIP approach concentrates on the ranking problem by use of the transformation model.

Currently used ranking methods include rankSVM (Herbrich *et al.*, 2000) and RankBoost (Freund *et al.*, 2004). Although the method proposed here and rankSVM are both based on SVMs, two differences can be noted. (i) Firstly, the rankSVM uses all pairs of data points for training, which results in $\mathcal{O}(n^2)$ comparisons, where MINLIP has a complexity of $\mathcal{O}(n)$. This reduction in complexity makes the model more applicable to large datasets. (ii) Secondly, the complexity control, being the margin and the Lipschitz constant, is different in both methods. In rankSVM all margins are equal and the model is tuned to maximize this margin. In MINLIP the margins differ corresponding to the difference in the output levels.

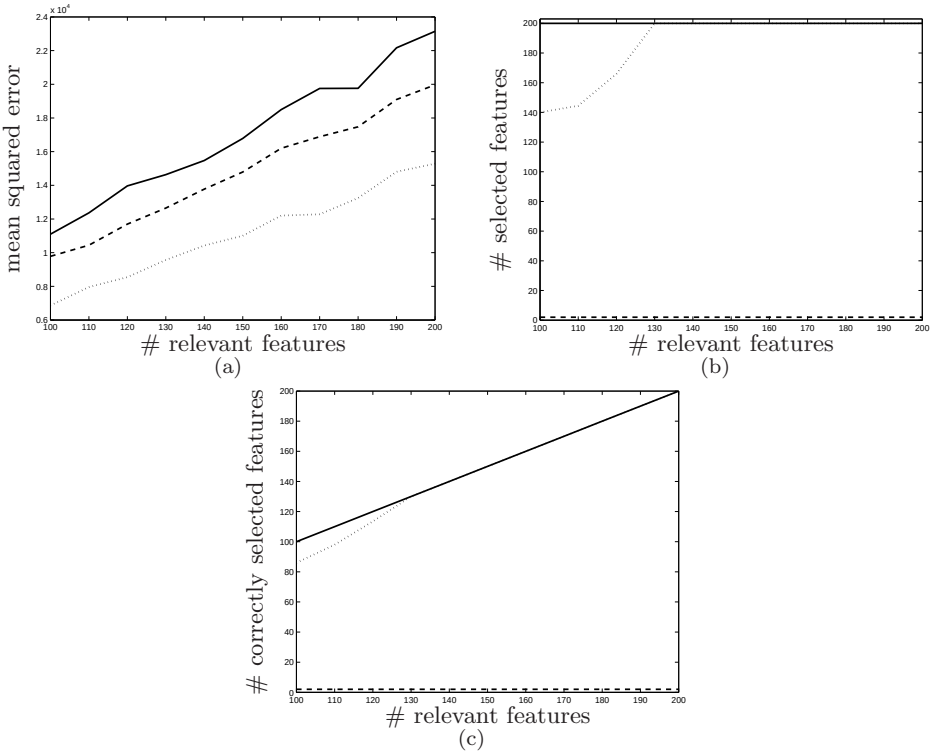


Figure 3.3: *Performance and feature selection ability of* MINLIP *(solid) and two sparse alternatives (*$\text{MINLIP}_{\text{L1}}$ *(dashed) and* MINLIP_{p} *(dotted)) on an artificial data set (n=100 for training,* n_{test}*=50 for testing). 200* $\mathcal{N}(0,1)$ *distributed covariates were generated, a varying number* $d = 100, 110, \ldots, 200$ *of which were used to generate the outcome (*$y = \sum_{p=1}^{d} w_p x^p$*, with w drawn from a standard normal distribution). The results are averaged over 100 datasets. (a) Median mean squared error on the test sets:* MINLIP_{p} *outperforms both other methods. (b-c) Number of selected (absolute value of estimated weight>* 10^{-8}*) and correctly selected variables versus number of relevant variables. The* MINLIP *method selects all variables, a lot of them not being relevant. The number of selected features is lower for the* MINLIP_{p} *model. The* $\text{MINLIP}_{\text{L1}}$ *model selects very few variables, but those which are selected are also relevant.*

3.4 Learning for ordinal regression

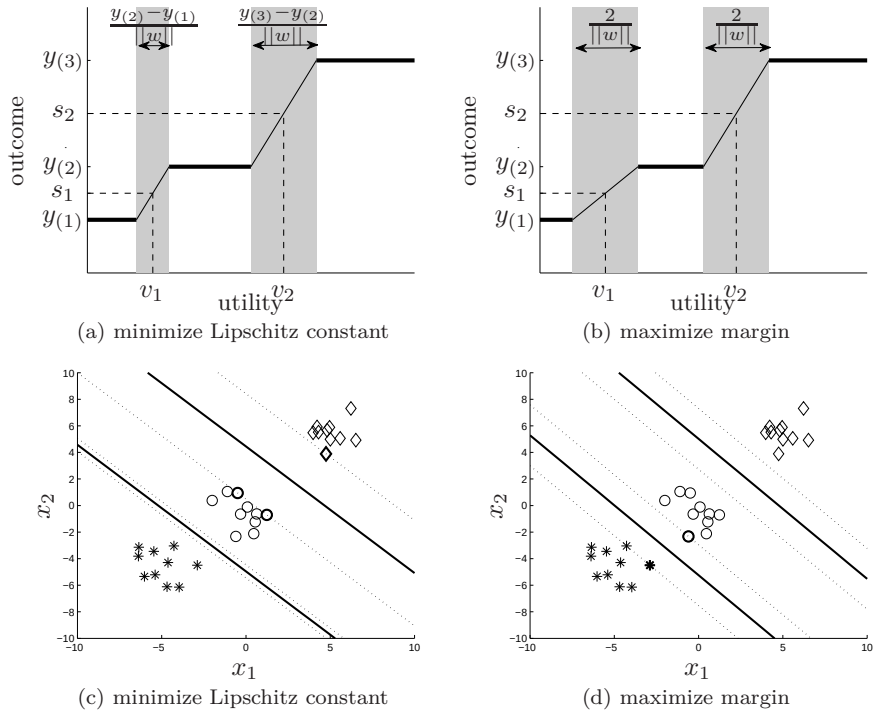
Consider now the situation where the output takes a finite number of $k \in \mathbb{N}$ values, where the k different classes possess a natural ordering relation. In this case the outcome y is an element of the finite ordered set $\{y_{(1)}, \ldots, y_{(k)}\}$.

3.4.1 A modification to MINLIP

In Section 3.3.1 it is mentioned that comparisons are made between points (i) and (j) where $y_{(j)}$ is the highest ordered value smaller than $y_{(i)}$. Applying this methodology in the ordinal setting would lead to as many comparisons with point (i) from class l as there are observations in class $l-1$. To cope with this issue, we add dummy observations with utility v_l and outcome s_l, with $l = 1, \ldots, k-1$. The outcome of a dummy observation s_l is located in between two consecutive ordinal classes with levels $y_{(l)}$ and $y_{(l+1)}$ and is defined as $s_l = \frac{1}{2}(y_{(l+1)} + y_{(l)})$ (see Figure 3.4). The input variables and utility are left unspecified. This implies that one has to compare each observation only twice, once with the dummy observation in between the previous and the current ordinal class and once with the dummy observation in between the current and the next class, restricting the number of constraints to $\mathcal{O}(n)$.

The problem is then formulated as:

$$
\begin{aligned}
\min_{w,\epsilon,\epsilon^*} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\
\text{subject to} \quad & \begin{cases}
w^T \varphi(x_i) - \sum_{l=1}^{k-1} \underline{q}_l(i) v_l \geq y_i - \sum_{l=1}^{k-1} \underline{q}_l(i) s_l - \epsilon_i, & \\
\qquad\qquad \forall\ i = 1, \ldots, n, & \\
-w^T \varphi(x_i) + \sum_{l=1}^{k-1} \overline{q}_l(i) v_l \geq -y_i + \sum_{l=1}^{k-1} \overline{q}_l(i) s_l - \epsilon_i^*, & \\
\qquad\qquad \forall\ i = 1, \ldots, n, & \\
\epsilon_i \geq 0, \qquad \forall\ i = 1, \ldots, n, & \\
\epsilon_i^* \geq 0, \qquad \forall\ i = 1, \ldots, n, & \\
v_{l+1} - v_l \geq 0, \qquad \forall\ l = 1, \ldots, k-2\,, &
\end{cases}
\end{aligned}
\tag{3.33}
$$



(a) minimize Lipschitz constant

(b) maximize margin

(c) minimize Lipschitz constant

(d) maximize margin

Figure 3.4: *Adaptation of the* MINLIP *algorithm to ordinal regression. (a) Inclusion of dummy data points with output values s_l intermediate to the observed output values and undefined covariates and utility v_l. All data points are compared with two dummy data points. (b) Comparison with the maximal margin strategy used in standard* SVM *where the margin is equal between all classes. (c-d) Example with 3 linearly separable cases with outcomes equal to 1 (stars), 2 (circles) and 10 (diamonds), respectively. The bold symbols represent the support vectors. In the maximal margin strategy there exists one margin, equal between every two successive classes, which results in a different Lipschitz constant. Using the* MINLIP *strategy, the Lipschitz continuity is optimized, resulting in margins which are proportional to the difference in the class labels. Support vectors of the latter method are therefore more likely to be observations of two classes for which the output labels differ the most.*

with γ a positive regularization constant as before and

$$\underline{q}_l(i) = \mathcal{I}\left(l = \max_m s_m \leq y_i\right), \forall\ i = 1, \ldots, n,$$

$$\overline{q}_l(i) = \mathcal{I}\left(l = \min_m s_m > y_i\right), \forall\ i = 1, \ldots, n\,.$$

The Lagrange dual problem becomes

$$\begin{aligned}
\min_{\alpha,\alpha^*,\theta} \quad & \tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\varphi(x_i)^T\varphi(x_j)(\alpha_j - \alpha_j^*) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \\
& + \sum_{i=1}^{n}\sum_{l=1}^{k-1}\left(\alpha_i \underline{q}_l(i) - \alpha_i^* \overline{q}_l(i)\right)s_l \\
\text{subject to} \quad & \begin{cases}
0 \leq \alpha_i \leq \gamma, & \forall\ i = 1, \ldots, n, \\
0 \leq \alpha_i^* \leq \gamma, & \forall\ i = 1, \ldots, n, \\
\sum_{i=1}^{n}\left(\alpha_i \underline{q}_l(i) - \alpha_i^* \overline{q}_l(i)\right) - \theta_{l-1} + \theta_l = 0, & \\
& \forall\ l = 2, \ldots, k-2, \\
\sum_{i=1}^{n}\left(\alpha_i \underline{q}_l(i) - \alpha_i^* \overline{q}_l(i)\right) - \theta_{l-1}, & \forall\ l = k-2, \\
\sum_{i=1}^{n}\left(\alpha_i \underline{q}_l(i) - \alpha_i^* \overline{q}_l(i)\right) + \theta_l = 0, & \forall\ l = 1\,,
\end{cases}
\end{aligned} \tag{3.34}$$

with α_i, α_i^* and θ_l the Lagrange multipliers corresponding to the first, second and last constraint in equation (3.33). We refer to Appendix B.2 for a detailed derivation. The utility can be evaluated at any point $x_\star \in \mathbb{R}^d$ as

$$u(x_\star) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x_i, x_\star)\,. \tag{3.35}$$

3.4.2 Prediction for ordinal regression

A clear advantage of the approach which includes unknown thresholds is that the prediction step becomes very simple. As illustrated in Figure 3.5, the predictions can be easily obtained from the value of the utility function in comparison with the different threshold values.

3.4.3 Difference with other methods

Chu & Keerthi (2005) proposed two SVM based models for ordinal regression. Both methods introduce $k-1$ thresholds with k the number of ordinal levels in the data. As with SVM classifiers, the margin between two ordinal levels is set to $\frac{2}{||w||_2}$. In their first method (EXC) a data point x_i belonging to class l has two slack variables: one relating to the threshold between classes $l-1$ and l and a second relating to the threshold between classes l and $l+1$. To ensure that the threshold between classes $l-1$ and l is smaller than the threshold between classes l and $l+1$, $k-1$ additional constraints are explicitly included. The problem can be written as:

EXC

$$\begin{aligned}
&\min_{w,\epsilon,\epsilon^*,v} && \tfrac{1}{2}w^T w + \gamma \sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*) \\
&\text{subject to} && \begin{cases}
w^T\varphi(x_i) - v_l + \epsilon_i \geq 1, & \\
\qquad\qquad \forall\ i = 1,\ldots,n;\ l = \arg\max_m(y_i > v_m) & \\
-w^T\varphi(x_i) + v_l + \epsilon_i^* \geq 1, & \\
\qquad\qquad \forall\ i = 1,\ldots,n;\ l = \arg\min_m(y_i < v_m) & \\
\epsilon_i \geq 0 \qquad \forall\ i = 1,\ldots,n & \\
\epsilon_i^* \geq 0 \qquad \forall\ i = 1,\ldots,n & \\
v_l \leq v_{l+1} \quad \forall\ l = 1,\ldots,k-2\,. &
\end{cases}
\end{aligned} \tag{3.36}$$

In their second approach (IMC) the constraints on the thresholds are added implicitly by adding $k-1$ slack variables, one for each threshold, for every observation. The problem can be formulated as follows:

IMC

$$\begin{aligned}
&\min_{w,\epsilon,\epsilon^*,v} && \tfrac{1}{2}w^T w + \gamma \sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*) \\
&\text{subject to} && \begin{cases}
w^T\varphi(x_i) - v_l + \epsilon_i \geq 1 & \forall\ i = 1,\ldots,n;\ \forall\ l:\ y_i > v_l \\
-w^T\varphi(x_i) + v_l + \epsilon_i^* \geq 1 & \forall\ i = 1,\ldots,n;\ \forall\ l:\ y_i < v_l \\
\epsilon_i \geq 0 & \forall\ i = 1,\ldots,n \\
\epsilon_i^* \geq 0 & \forall\ i = 1,\ldots,n\,.
\end{cases}
\end{aligned} \tag{3.37}$$

In our method, we adopt the approach of the EXC method concerning slack variables, the method differing in the definition of the margin. Instead of defining an equal margin at every border, the margin between classes l and $l+1$ depends on the difference in outcome level.

Note the similarity between these models and the standard SVMs (Vapnik, 1998) in the binary classification problem (with two classes $\mathcal{C}_1$ and $\mathcal{C}_2$):

$$\begin{aligned} \min_{w,\epsilon,\epsilon^*,b} \quad & \tfrac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T \varphi(x_i) + b + \epsilon_i \geq 1 & \forall\ i \in \mathcal{C}_1 \\ -w^T \varphi(x_i) - b + \epsilon_i^* \geq 1 & \forall\ i \in \mathcal{C}_2 \\ \epsilon_i \geq 0 & \forall\ i \in \mathcal{C}_1 \\ \epsilon_i^* \geq 0 & \forall\ i \in \mathcal{C}_2\,. \end{cases} \end{aligned} \tag{3.38}$$

In case $k = 2$, both EXC and IMC reduce to the model:

$$\begin{aligned} \min_{w,\epsilon,\epsilon^*,v} \quad & \tfrac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T \varphi(x_i) - v + \epsilon_i \geq 1 & \forall\ i \in \mathcal{C}_1 \\ -w^T \varphi(x_i) + v + \epsilon_i^* \geq 1 & \forall\ i \in \mathcal{C}_2 \\ \epsilon_i \geq 0 & \forall\ i \in \mathcal{C}_1 \\ \epsilon_i^* \geq 0 & \forall\ i \in \mathcal{C}_2\,, \end{cases} \end{aligned} \tag{3.39}$$

which equals the model in equation (3.38) when the threshold v (note that there is only one threshold in this case) is considered as the constant term. The MINLIP model reduces to:

$$\begin{aligned} \min_{w,\epsilon,\epsilon^*,v} \quad & \tfrac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T \varphi(x_i) - v + \epsilon_i \geq y_i - s & \forall\ i \in \mathcal{C}_1 \\ -w^T \varphi(x_i) + v + \epsilon_i^* \geq s - y_i & \forall\ i \in \mathcal{C}_2 \\ \epsilon_i \geq 0 & \forall\ i \in \mathcal{C}_1 \\ \epsilon_i^* \geq 0 & \forall\ i \in \mathcal{C}_2\,, \end{cases} \end{aligned} \tag{3.40}$$

where only one dummy observation (v, s) needs to be introduced. The difference between equations (3.38,3.39) and equation (3.40) lies in the right hand side of the two first inequalities, which is a consequence of the used complexity control.

Chu & Ghahramani (2005) proposed a probabilistic approach to ordinal regression in Gaussian processes (GPOR). They impose a Gaussian process prior distribution on the utility function (called latent function in their work) and employ an appropriate likelihood function for ordinal variables. Experiments will compare our methods with the Bayesian inference technique of MacKay (1992), using the Laplacian approximation to implement model adaptation. The GPOR approach differs from ours since it uses a Bayesian framework.

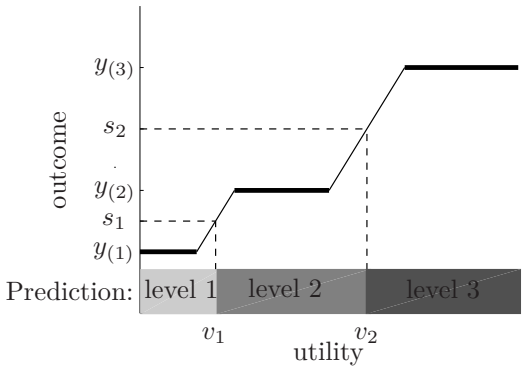


Figure 3.5: *Prediction for ordinal regression.* MINLIP *for ordinal regression, including unknown thresholds, has the advantage to reduce the prediction step to a simple comparison between the utility of a new observation and the utility of the thresholds. If the utility has a value between threshold* $l-1$ *and* l*, the predicted outcome equals the* l^{th} *output level.*

3.5 Transformation models for failure time data

We now turn our attention to the case where the data originate from a survival study. We will adopt a classical statistical setup, and show how the techniques as described in Section 3.3 provide a powerful alternative to the classical statistical models.

3.5.1 Transformation models for survival analysis

The TM (see Definition 3.2) encompasses a broad class of models, including the following classical survival models.

1. Cox' *proportional hazards model* is recovered when one defines $g = h^{-1}$ (if it exists) as $g(z) = \ln(-\ln(z))$. Under the PH, the value of the survival function at time t is

$$S(t,x) = [S_0(t)]^{\exp(w^T x)}\,, \tag{3.41}$$

where $S_0(t) = \exp(-\Lambda_0(t))$ is called the baseline survival function. Taking $\ln(-\ln(\cdot))$ of both sides in (3.41) leads to

$$\begin{aligned} \ln(-\ln(S(t,x)) &= \ln(-\ln(S_0(t))) + w^T x \qquad (3.42) \\ &= \ln(\Lambda_0(t)) + w^T x \\ \Rightarrow \epsilon &= g(t) + u(x) \\ \Rightarrow t &= h(-u(x) + \epsilon)\,. \end{aligned}$$

Note that the last transition is only possible if $g(t)$ is invertible. The noise terms are i.i.d. observations from the extreme value distribution $F_\epsilon(z) = 1 - \exp(-\exp(z))$.

2. The *proportional odds model* is defined as

$$\ln\left(\frac{F(t|x)}{1-F(t|x)}\right) = \alpha(t) + w^T x\,, \qquad (3.43)$$

with $F(t|x)$ the conditional cumulative distribution function and $\alpha(t)$ a monotonically increasing function. In general the survival function equals $S(t) = 1 - F(t)$, leading together with equation (3.43) to

$$\begin{aligned} \ln\left(\frac{1-S(t|x)}{S(t|x)}\right) &= \alpha(t) + w^T x \qquad (3.44) \\ \Rightarrow \epsilon &= \alpha(t) + u(x) \\ \Rightarrow t &= h(-u(x) + \epsilon)\,. \end{aligned}$$

Note that the last transition is only possible if $\alpha(t)$ is invertible.

3. The *accelerated failure time* is given when $g(z) = \ln(z)$.

For an extended discussion on the use of the class of transformation models and specific parameterizations of the functions h or g, see e.g. Dabrowska & Doksum (1988); Koenker & Geling (2001); Cheng *et al.* (1997) and citations.

3.5.2 A modification to MINLIP

This Section describes how the standard MINLIP model can be extended towards failure time data including the handling of censored data. Due to censoring,

the observations $(x_{(i)}, y_{(i)}, \delta_{(i)})$ and $(x_{(i-1)}, y_{(i-1)}, \delta_{(i-1)})$ will not always be comparable. The model is therefore adapted as follows.

$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i + \epsilon_{\bar{j}(i)}, & \\ & \forall\ i = 1, \ldots, n, \\ \epsilon_i \geq 0, & \forall\ i = 2, \ldots, n\,, \end{cases} \end{aligned} \tag{3.45}$$

with

$$\begin{aligned} \bar{j}(i) = \quad & \arg\max_{j} \quad j \\ & \text{subject to} \quad \begin{cases} (x_{(i)}, y_{(i)}, \delta_{(i)}) \text{ and } (x_{(j)}, y_{(j)}, \delta_{(j)}) \text{ comparable} \\ y_j < y_i\,. \end{cases} \end{aligned} \tag{3.46}$$

The dual problem is then found as

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,k=2}^{n} \alpha_i \alpha_k \left(\varphi(x_{(i)}) - \varphi(x_{\bar{j}(i)})\right)^T \left(\varphi(x_{(k)}) - \varphi(x_{\bar{l}(k)})\right) \\ & - \sum_{i=2}^{n} \alpha_i (y_{(i)} - y_{\bar{j}(i)}) \\ \text{subject to} \quad & \begin{cases} \alpha_i \geq 0, & \forall\ i = 2, \ldots, n, \\ -\gamma \leq \alpha_i - \alpha_{\bar{j}(i)} \leq \gamma, & \forall\ i = 2, \ldots, n, \\ -\gamma \leq \alpha_2 \leq \gamma, & \\ -\gamma \leq \alpha_n \leq \gamma\,. & \end{cases} \end{aligned} \tag{3.47}$$

The predicted utility for a new point $x_\star$ is then found as

$$u(x_\star) = \sum_{i=2}^{n} \alpha_i \left(\varphi(x_{(i)}) - \varphi(x_{(\bar{j}(i))})\right)^T \varphi(x_\star). \tag{3.48}$$

Since the censoring is handled by a proper choice of $\bar{j}(i)$, the formulas in Subsection 3.3.4 can be extended in a similar way.

3.5.3 Prediction with transformation models

The prediction step in survival analysis, refers to the estimation of survival and hazard functions rather than the estimation of the failure time itself. The

proportional hazards model estimates these functions, by assuming that a baseline hazard function exists; the input variables changing the hazard only proportionally. The baseline hazard function is estimated using the Breslow estimator of the cumulative baseline hazard (Breslow, 1974).

In our setting, the cumulative distribution function (CDF), can be estimated, after estimation of the utility, as follows. The time axis is divided in k equidistant time intervals $[t_{l-1}, t_l], \forall\ l = 2, \ldots, k$. For each observation in the set $\{(u_i, y_i, \delta_i)\}_{i=1}^n$, the outcome in each time interval is defined as:

$$y_{il} = \begin{cases} 0 \text{ if } y_i > t_l \\ 1 \text{ if } y_i \leq t_l \text{ and } \delta_i = 1\,. \end{cases} \tag{3.49}$$

Note that censored observations are not considered at times later than the censoring time. Using a monotone least squares support vector regression model (Pelckmans *et al.* (2005b)) with a Gaussian kernel (or another monotonic regression model) the utility and the time interval number l as inputs and y_{il} as output, the CDF $\hat{F}(u_i, l),\ \forall\ l = 1, \ldots, k-1$, is estimated. The survival function is found as $\hat{S}(u_i, l) = 1 - \hat{F}(u_i, l)$. The hazard function is then estimated as

$$\hat{\lambda}(u_i, l) = \frac{\hat{F}(u_i, l+1) - \hat{F}(u_i, l)}{t_{l+1} - t_l}, \forall\ l = 1, \ldots, k-1\,. \tag{3.50}$$

We refer to Chapter 5 for more information on the reconstruction step for survival data.

3.6 Application studies

This final Section describes some experiments to illustrate the use of the presented method. In a first Subsection, 3 artificial examples will illustrate how transformation models are used within the ranking, ordinal regression and survival setting (see also Table 4.1). A first real life application illustrates the use of MINLIP for ordinal data. We use 6 benchmark datasets and compare the performance of MINLIP with EXC and IMC as proposed in Chu & Keerthi (2005) and GPOR as proposed in Chu & Ghahramani (2005). The last two examples concern survival data, one with micro-array data (data also used in Bøvelstad *et al.* (2007)) and one with clinical data (Schumacher *et al.*, 1994).

Unless stated otherwise, 10-fold cross-validation was used for model selection. For every kernel and regularization parameter to be tuned a grid of values was searched and the combination of parameter values yielding the lowest cross-validation error or highest cross-validation performance was selected. In the first example the mean absolute error between predicted and observed output levels was used as model selection criterion since prediction is relevant in this case. For both survival

examples, the cross-validated concordance index was used as model selection criterion since the main interest lies in the ranking of the patients.

3.6.1 Artificial examples

This section illustrates the different steps needed to obtain the desired output for ranking, ordinal regression and survival problems, using artificial examples. Together with Table 4.1, this Subsection illustrates the different tasks considered in this Chapter.

Ranking

In this first example, we consider the ranks of 150 cyclists in 9 different races. Using the ranks of 100 out of these cyclists in a 10th race, we want to predict the rank of the remaining 50 cyclists. Additional information includes: age, weight and condition score. The outcome is defined as the ranking given by a weighted sum of the ranking in the 9 races, age, weight and condition score. Weights are drawn from a uniform distribution on the unit interval. The ranking on the previous races are numbers from 1 to 100, all other variables are drawn from a standard normal distribution.

A first step in all transformation models is to estimate the utility function u. Using equation (3.22) with a linear kernel and 5-fold cross validation with the concordance index (ranking criterion) as a model selection criterion, a concordance index of 0.98 on the test set was obtained. The predicted ranking corresponds very well with the observed ranking (see Figure 3.6). Since one is only interested in ranking the cyclists, the absolute value of the utility is irrelevant. Additionally, one is not interested in estimating the transformation function h.

Ordinal regression

In a second artificial example, consider the scenario in which one wishes to divide students into 3 groups: bad, average and good students. For this task, the grades on 10 courses for 150 students are available. The outcome depends on the average grade. A bad, average and good student has an average grade below 55%, between 55% and 65% and above 65% respectively. The results on 100 students will be used for training, the remaining students will be used for testing.

In a first step, all students are ranked according to the available data, namely their grades on 10 courses. Using equation (3.33) with the concordance index as model selection criterion and a linear kernel, a concordance of 0.99 between the utility and the outcome on the test set is obtained. In addition to an estimate of the

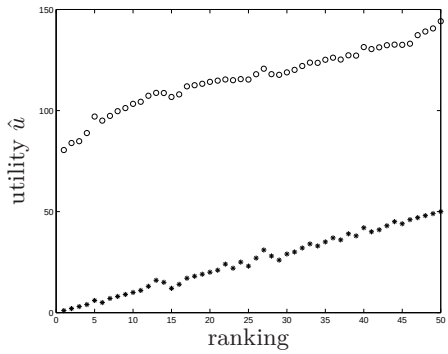


Figure 3.6: *Artificial example illustrating the use of the* MINLIP *transformation model for the ranking setting. The estimated utility of the test observations are denoted by the circles. The value of the utility does not correspond to the ranking. However, the rank of the estimated utility (denoted by stars) is a good prediction of the observed rank.*

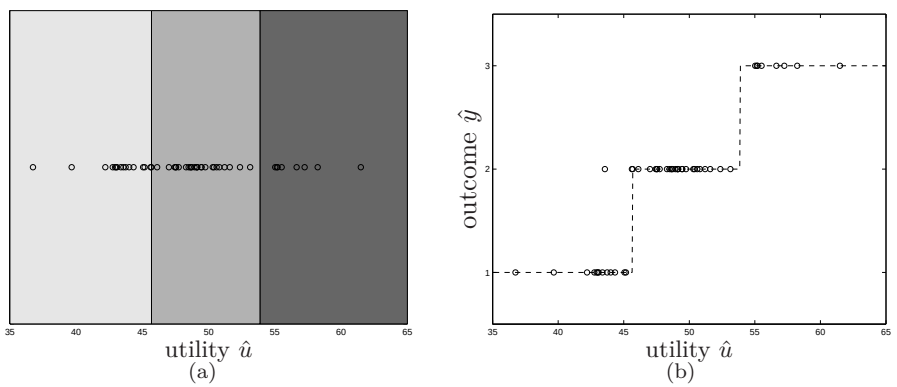


Figure 3.7: *Artificial example illustrating the use of the* MINLIP *transformation model for ordinal regression. (a) The estimated utility of the test observations are denoted by the circles. The* MINLIP *model for ordinal regression results in an estimate of the utility function and threshold values. Students with the lowest utility (less than the first threshold) are predicted to be bad students (light grey). Students with a utility between both thresholds (medium grey) are estimated to be average students and students with a utility higher than the second threshold (dark grey) are predicted to be good students. (b) Illustration of the transformation function h (dashed line).*

utility, the MINLIP model for ordinal regression gives threshold values which can be used to predict the outcome of new observations (see Figure 3.7). Since, the MINLIP model generates these thresholds, no additional model is needed to obtain the transformation function.

Survival analysis

In a last artificial example, a randomized trial is simulated. Assume 150 patients are randomly divided into two treatment groups. Additionally, the age (drawn from a standard normal distribution) of the patients is known. The survival time is known for 100 patients. For the first treatment arm, the survival time has a Weibull distribution with parameters 1 and 0.5. For the second treatment arm, the survival time is Weibull distributed with parameters 4 and 5. Using the information on age, treatment arm and survival on 100 patients, one would like to predict the survival for the remaining 50 patients. Note that the survival characteristics do not depend on the age of the patients.

As with the two previous examples, the MINLIP model is used to estimate the utility of the patients. Using a linear kernel and 5-fold cross validation in comparison with the concordance index as model selection criterion, a c-index of 0.70 is obtained on the test set. Figure 3.8 illustrates that the MINLIP model is able to divide the group of test patients into two groups with a significantly different survival (p=0.03, logrank test). The first part of the transformation model obtains a nice result. However, in survival analysis, additional information can be provided when performing the second part of the transformation model, namely estimating the transformation function. Applying the method as explained in Section 3.5.3, the estimated survival curves for all patients are calculated (Figure 3.9). One clearly notices two distinct survival groups. The grey and black survival curves correspond to patients in the first and second treatment arm, respectively. The true survival function for the first and second treatment are illustrated in thick black and grey lines, respectively. The differences in predicted survival within each treatment arm are due to age differences.

3.6.2 Ordinal regression

To illustrate the performance of MINLIP for ordinal regression, 6 regression datasets[3] were converted to ordinal regression datasets as follows. The datasets were divided into 20 folds with 10 equal-frequency bins. The output value for each bin was set to the average output within the bin. The performance of the MINLIP model was compared with two methods described in Chu & Keerthi (2005) (see Table 3.2). Tuning of the Gaussian kernel parameter and the regularization

[3]Data are available at `http://www.liacc.up.pu/~ltorgo/Regression/DataSets.html`

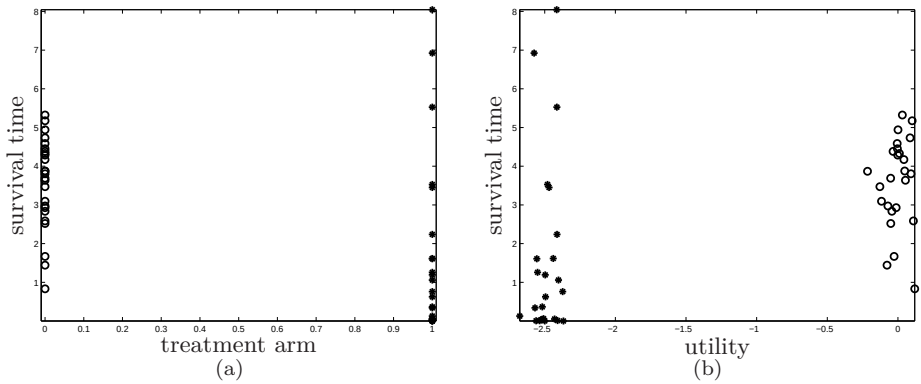


Figure 3.8: *Artificial example illustrating the use of the* MINLIP *model for survival analysis. (a) Survival time as a function of the treatment arm. Patients receiving the first treatment in general survive longer. The second treatment results in lower survival times. However, some patients have extremely large survival times. (b) Survival time versus estimated utility. The circles and the stars denote the first and second treatment arm, respectively. The utility is able to group the patients according to the relevant variable treatment (see the clear separation between circles and stars).*

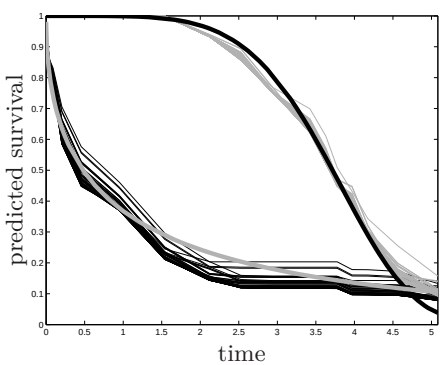


Figure 3.9: *Artificial example illustrating the use of the* MINLIP *model for survival analysis: illustration of the reconstruction step. For each patient, the survival curve is calculated using the method discussed in Section 3.5.3. Grey and black curves represent patients from the first and second treatment arm, respectively. The true survival curve for the first and second treatment, are illustrated in thick black and grey lines. One clearly notices two distinct survival groups, corresponding to the treatment groups.*

parameter was performed with 10-fold cross-validation on an exponential grid using mean absolute error as model selection criterion. After an initial search, a finer search was performed in the neighborhood of the initial optimum. Results of the GPOR method are reported for comparison. GPOR has a lower mean zero-one error on small datasets. For larger datasets and for mean absolute errors, it performs less. The IMC method has the disadvantage that large QP problems need to be solved for growing training samples, requiring more computational time. The MINLIP method makes a nice trade-off between computational time and performance.

3.6.3 Failure time data: micro-array studies

The MINLIP technique is derived from machine learning techniques as SVMs, techniques which are known to be especially useful to handle high-dimensional data sets. We therefore test the performance of MINLIP on 3 micro-array data sets.

In this example we compare the performance of model (3.45) (MINLIP) and linear extensions as in equations (3.31) ($\text{MINLIP}_{\text{L1}}$) and (3.32) ($\text{MINLIP}_{\text{p}}$) with 5 other methods which are discussed and implemented by Bøvelstad *et al.* (2007): principal components regression (PCR), supervised principal components regression (SPCR), partial least squares regression (PLS) and two penalized Cox regression models: ridge regression (PH_{l2}) and L_1 regularization (PH_{l1}). The PCR method uses principal component analysis to select n_λ principal components which account for as much variation in the gene expression profiles as possible. The selected principal components are then used as covariates in a PH (see Hastie *et al.* (2001)). The SPCR method (Bair & Tibshirani (2004); Bair *et al.* (2006)) first selects a subset of genes which are correlated with survival by using univariate selection and then applies PCR to this subset. The standard PLS method performs regression of the outcome using n_λ components which are a linear combination of the original covariates (Martens & Næs, 1989). The application of PLS to PH is not straightforward since the PLS algorithm assumes a linear relation between outcome and covariates. See Nygård *et al.* (2008) for a detailed description of the method.

The three micro-array data sets used in this experiment are discussed in Appendix A. Figure 3.10 summarizes the c-index on all methods for 100 different randomizations between training and test sets (2/3 training, 1/3 test). The MINLIP_p model obtains the highest c-index. In the right panels of Figure 3.10 the area underneath the TDROCs (Heagerty *et al.*, 2000) are shown. This curve for the MINLIP_p model is the highest for all three datasets. The Wilcoxon signed rank sum test was used to test for significant differences in concordance index. The MINLIP_p model was significantly better than all other models. In a second test, the performance of separating low and high risk patients was tested. The threshold between low and high risk patients was defined as the median value

Table 3.2: *Test results for* MINLIP, EXC, IMC *using a Gaussian kernel and* GPOR. *The targets of the data sets were discretized by 10 equal-frequency bins. The output value for each bin is set to the average output within the bin. The results are averaged over 20 trials. The best performing model is indicated in bold. Significant differences as calculated by Wilcoxon's signed ranksum test between the* EXC, IMC, GPOR *and the* MINLIP *(reference) model are indicated with* [1] ($p < 0.05$), [2] ($p < 0.01$) *or* [3] ($p < 0.001$).

data set	mean zero-one error			
	minlip	exc	imc	gpor
pyrimidines	0.74±0.07	0.79±0.08	0.75±0.09	**0.73±0.09**
triazines	**0.86±0.05**	**0.86±0.04**	0.87±0.04	**0.86±0.03**
Wisconsin	0.89±0.02	0.89±0.03	0.89±0.02	**0.85±0.03**[3]
machine CPU	0.65±0.04	0.65±0.06	**0.64±0.04**	0.80±0.09[3]
auto MPG	**0.58±0.03**	**0.58±0.03**	/	0.68±0.08[3]
Boston housing	**0.57±0.04**	**0.57±0.04**	/	0.61±0.03[3]
data set	mean absolute error			
	minlip	exc	imc	gpor
pyrimidines	**0.05±0.01**	0.06±0.01	**0.05±0.01**	0.06±0.01
triazines	0.12±0.06	**0.10±0.01**[1]	**0.10±0.00**[3]	0.16±0.02[3]
Wisconsin	29.56±2.18	28.80±1.17	**28.23±1.28**	34.52±4.27[3]
machine CPU	**29.41±4.27**	30.57±6.94	30.48±4.10	157.00±124.67[3]
auto MPG	2.08±0.16	**2.05±0.18**	/	4.13±1.13[3]
Boston housing	2.49±0.27	**2.33±0.25**[1]	/	2.97±0.29[3]

of predictions at the training set. The logrank test was then used to test the difference in survival between both groups. The χ^2 value of this test obtained with the MINLIP$_p$ model was compared with the other models and significantly higher χ^2-values were obtained with the former model.

3.6.4 Failure time data: clinical data

In a last example the GBC (see Appendix A.4.1) is used to investigate the ability of the MINLIP model to estimate how the different covariates influence the survival time. The data set was randomly divided in a training and test set (2/3 versus 1/3). Since medical data are typically not highly non-linear, a componentwise polynomial kernel

$$k(x,z) = \sum_{p=1}^{d} (\tau + x^{pT} z^p)^2, \ \tau \geq 0\,, \tag{3.51}$$

was used to model non-linearities. Model selection is done by 10-fold cross-validation with the c-index as model selection criterion.

The results obtained with MINLIP were compared with PH. However, the standard PH assumes a linear form in the variables, implying that for a continuous variable as age for example, the risk ratio between two patients aged 45 and 50 is the same as the risk ratio between two patients aged 75 and 80. To allow for non-linearities, the functional forms of the covariates were estimated using penalized smoothing splines (PH$_{\text{pspline}}$) (Eilers & Marx, 1996; Hurvich *et al.*, 1998). In this method, a comparative small set of basis functions is fit and a likelihood penalizing the integrated second derivatives (see Therneau & Grambsch (2000)) is used to estimate the coefficients. Akaike's information criterion (Akaike, 1974) is used to select the degrees of freedom for each term.

Figures 3.11 and 3.12 show the estimated effects of the variables for PH$_{\text{pspline}}$ and MINLIP, respectively. Note that in Figure 3.11 the estimates are inversely related with the survival time, whereas in Figure 3.12 the estimates are related with the survival time itself. PH predicts a decreasing risk for relapse for older patients, up to the age of 40, whereafter the risk increases slowly; for tumors up to 20 mm the risk for relapse increases with size, with a threshold effect for larger tumors; the number of positive lymph nodes is inversely related with survival and larger values for the progesterone and estrogen receptors are related with longer survival. All conclusions correspond to effects known from literature (Fisher *et al.* (1983); Lamy *et al.* (2002); Pichon *et al.* (1980); Verschraegen *et al.* (2005)). The MINLIP model estimates a higher survival time for older patients, up to the age of 65, whereafter

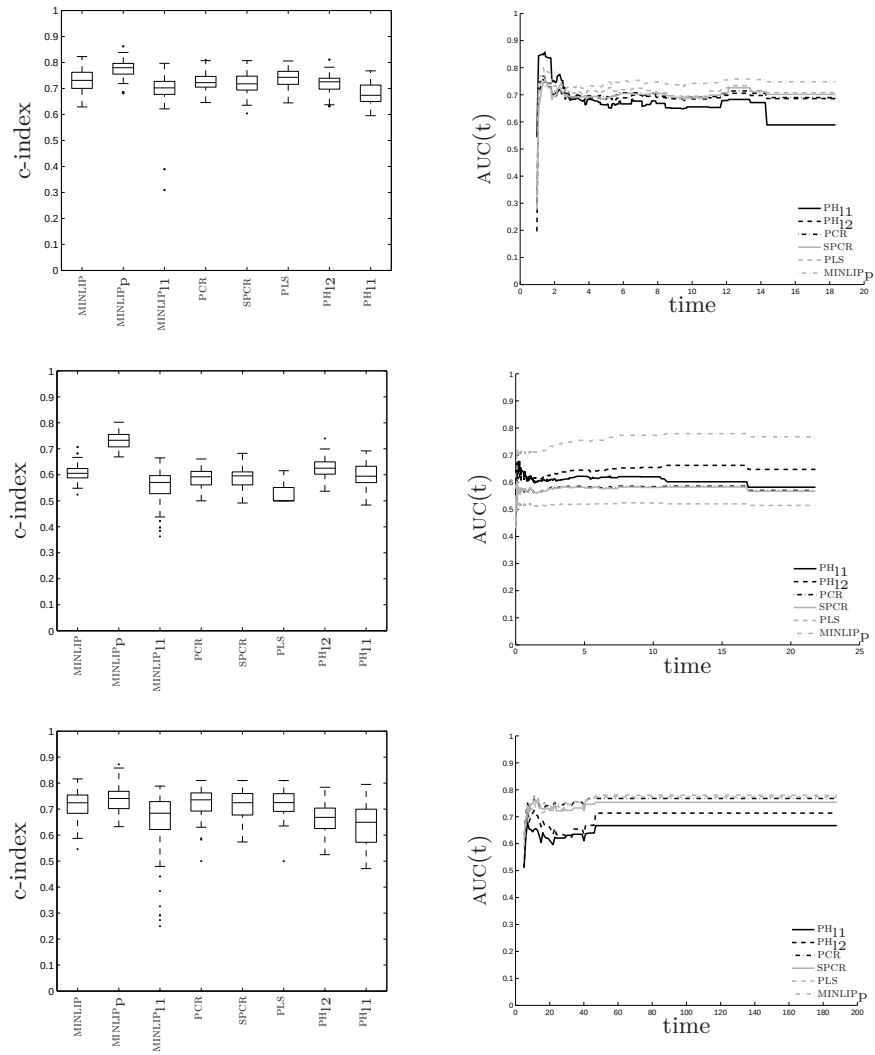


Figure 3.10: *Concordance (left) and time dependent* AUC *(right) on the test set for three micro-array survival data sets (top:* `DBCD`*, middle:* `DLBCL`*, bottom:* `NSBCD`*). The* MINLIP$_p$ *model obtains a significantly higher performance than all other models. The* TDROC *for the* MINLIP$_p$ *model is higher than for the other models.*

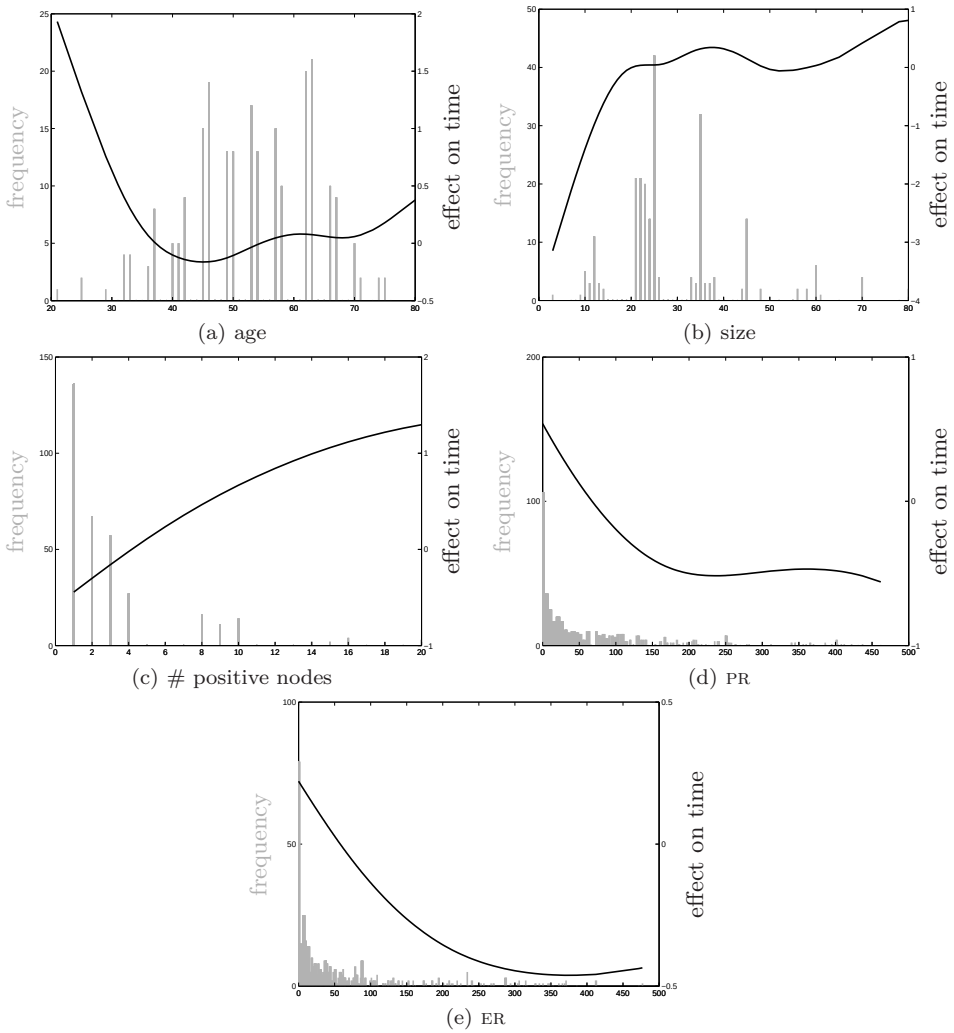


Figure 3.11: *Estimation of the effect of the variables on the risk of relapse with smoothing splines within* PH. *The grey vertical lines indicate the number of patients in the training set for each variable value. The estimated effects are inversely related with the survival time. The model estimates a lower risk for relapse for older patients up to the age of 40, whereafter the risk slowly increases. The risk for relapse increases for larger tumors until a size of 20 mm, whereafter the risk remains fairly constant. For common values of the number of positive lymph nodes and receptors, the risk increases for larger/lower values respectively. Conclusions drawn by the model agree with what is known from literature.*

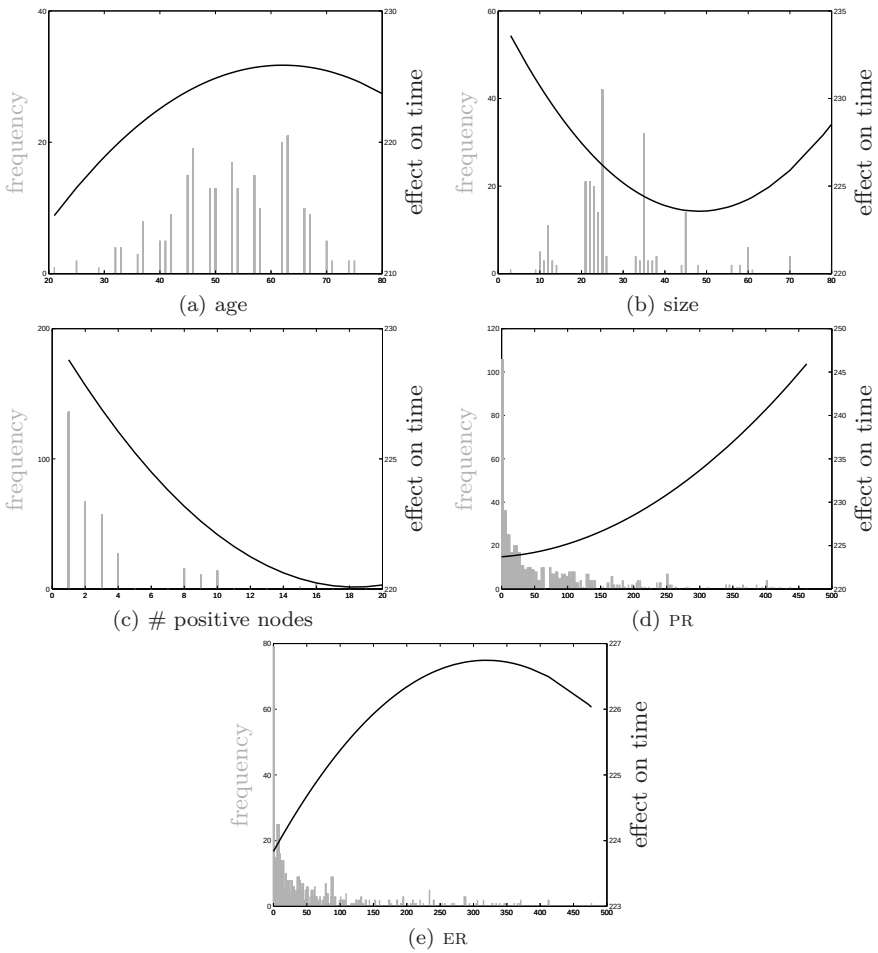
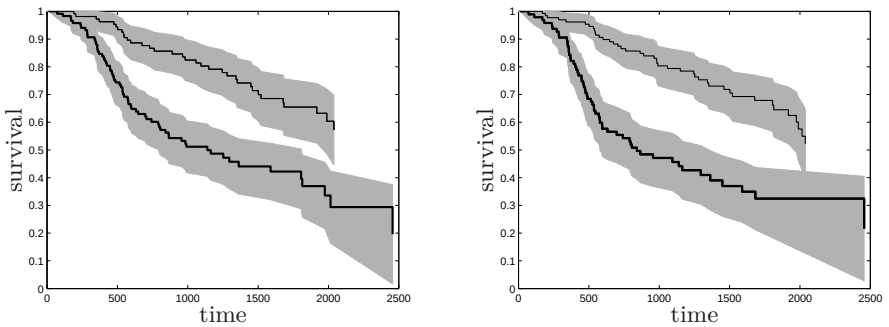


Figure 3.12: *Estimation of effect of the variables on survival time with the* MINLIP *model (the c-index was used for model selection). The grey vertical lines indicate the number of patients in the training set for each variable value. The estimated covariate effects are directly related with the survival time. The* MINLIP *model estimates the covariate effects as follows: the estimated survival time increases with age until the age of 65, whereafter the survival time drops slightly. The larger the tumor, the higher the number of positive lymph nodes, the lower the expression of the receptors, the lower the estimated survival time is. Conclusions drawn by the model agree with what is known from literature.*



(a) Cox (b) MINLIP

Figure 3.13: *The use of* $\text{PH}_{\text{pspline}}$ *and* MINLIP *model as a prognostic index. The output of both models is used to divide the test set into two groups, one with high and one with low risk for relapse. The threshold between both groups is defined as the median value of the model's output. Kaplan-Meier curves and 95% confidence intervals are shown for each group. The spread in the survival curves is broader for the* MINLIP *model, which is confirmed by a slightly larger value of the logrank test statistic.*

the survival time drops again. According to this model, a larger tumor, a higher number of positive lymph nodes and a lower progesterone and estrogen receptor level result in lower survival times and thus a higher risk for relapse. $\text{PH}_{\text{pspline}}$ obtains a c-index on the test set equal to 0.67, while the MINLIP model obtains a performance of 0.69.

Figure 3.13 illustrates the ability of the models to generate prognostic indices. Kaplan-Meier curves and 95%-confidence intervals are calculated for both groups. The logrank test χ^2 value is 20.43 and 29.70 for $\text{PH}_{\text{pspline}}$ and MINLIP respectively. The latter method results in a slightly better split between low and high risk patients.

3.7 Conclusions

This Chapter studied a machine learning approach for finding transformation models. Such models are found useful in a context of ordinal regression and survival analysis, and relate directly to commonly used risk measures as the area under the ROC curve. The derivations go along the same lines as used for support vector machines, except for replacing the notion of margin with a Lipschitz continuity condition on the transformation function. The presented learner finds

a possible non-linear, non-parametric transformation model by solving a convex QP. Extensions towards tasks where transformation models only provide an approximation, ordinal regression and survival analysis are given. Experiments on ordinal regression and survival analysis, on both clinical and high dimensional data sets, illustrate the use of the proposed method.

Chapter 4

Support vector methods for survival analysis: optimization problem formulation and comparison

The previous Chapter proposed one possible formulation of kernel based methods for the analysis of survival data. However, it was not explored whether the presented method was the best possible one. In this Chapter different optimization problems based on minimizing the Lipschitz constant, for dealing with censored data are formulated and compared. A comparison will be made between L_1 and L_2 norm regularization, between ranking and regression constraints and between different possibilities for reducing the number of comparisons between pairs from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. The results of this Chapter resulted in two manuscripts: Van Belle et al. *(2009b, 2010c).*

4.1 Existing SVM based methodologies for survival analysis

When analyzing survival data, one is interested in the time between a certain starting point and the occurrence of a predefined event. A first approach to model survival data is therefore to use regression models to model the time of occurrence of the event. However, this approach can not deal with censored data and only

incorporating observations for which the exact failure time is known would lead to underestimated survival times. Different proposals on how to model the survival problem by means of support vector machines found in the literature are elaborated below.

4.1.1 Support vector regression for censored data

In standard support vector regression (Vapnik, 1998, see also Chapter 2 of this work) the prediction is defined as a linear combination of a transformation of the variables $\varphi(x)$ plus a constant $b : \hat{y} = w^T\varphi(x) + b$. To obtain this prediction, the coefficients w and the constant b need to be calculated. In order to obtain good generalizability properties, the coefficients are kept small. SVMs are formulated as optimization problems, where a loss function should be minimized subject to certain constraints. The constraints include that the prediction $\hat{y}$ should be larger than $y-\epsilon$, with $\epsilon > 0$. Similarly $-\hat{y}$ should be larger than $-y-\epsilon^*$, with $\epsilon^* > 0$. The loss function penalizes large values of ϵ and ϵ^* such that the resulting predictions $\hat{y}$ will be close to the observed values y. Formally, the problem is formulated as:

$$\begin{aligned} \min_{w,b,\epsilon,\epsilon^*} \quad & \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall\ i = 1,\ldots,n \\ -w^T\varphi(x_i) - b \geq -y_i - \epsilon_i^*, & \forall\ i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall\ i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall\ i = 1,\ldots,n\,. \end{cases} \end{aligned} \tag{4.1}$$

The problem of using support vector regression for censored data lies in the uncertainty about the outcomes y. The earliest approaches to censored data using regression strategies either omitted the censored observations, resulting in underestimated failure times, or treated the censored observations as non-events, resulting in biased models. In order to use all the available information Shivaswamy *et al.* (2007) proposed a support vector regression approach to censored data (CSVR). For uncensored observations, the same constraints account as in the standard SVM regression model. For right censored observations, it is known that the failure did not occur until the censoring time. The first constraint in (4.1) is therefore still valid. However, the second constraint is too restrictive for right censored observations. The support vector regression model for censored data as

proposed by Shivaswamy can therefore be formulated as:

$$\boxed{\text{CSVR}}$$

$$\begin{aligned} \min_{w,b,\epsilon,\epsilon^*} \quad & \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall\ i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \epsilon_i^*, & \forall\ i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall\ i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall\ i = 1,\ldots,n\,. \end{cases} \end{aligned} \tag{4.2}$$

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

$$\hat{u}(x_\star) = \sum_i (\alpha_i - \delta_i\alpha_i^*)k(x_i, x_\star) + b\,, \tag{4.3}$$

with α_i and α_i^* the Lagrange multipliers.

A second proposal to SVM regression for censored data was made by Khan & Zubek (2008) (SVMc). The difference between both models lies in the applied penalty or loss for incorrect prognostic indices (see Figure 4.1), which can be interpreted as predicted failure times for regression approaches. Note that $\varepsilon = 0$ was assumed in the model equations. The CSVR method penalizes incorrect predictions the same whether the prediction was higher or lower than the observed failure time and penalizes incorrect predictions for right censored data only if the prediction is lower than the observed censoring time. The reasoning being that it is only known that the failure time is higher than the observed outcome. In addition, the penalty for wrong predictions is the same, whether it consists of a prediction for censored or observed failure times. On the contrary, SVMc applies different penalties for the four possible cases: (i) penalty γ for events with predicted survival less than the observed survival; (ii) penalty γ^* for events with predicted survival larger than the observed survival, with $\gamma^* > \gamma$; (iii) penalty γ_c for right censored data with predicted survival less than the observed censoring time; and (iv) penalty γ_c^* for right censored data with predicted survival larger than the observed censoring time, with $\gamma_c^* < \gamma_c$. Additionally, this model provides different $\epsilon-$losses for events and right censored data and for predictions higher and lower than the observed time. Therefore, the major drawback of the latter method is the large number of hyperparameters.

4.1.2 Support vector machines based on ranking constraints

When not using regression models, survival problems are often translated into classification problems answering the question whether the patient survives a

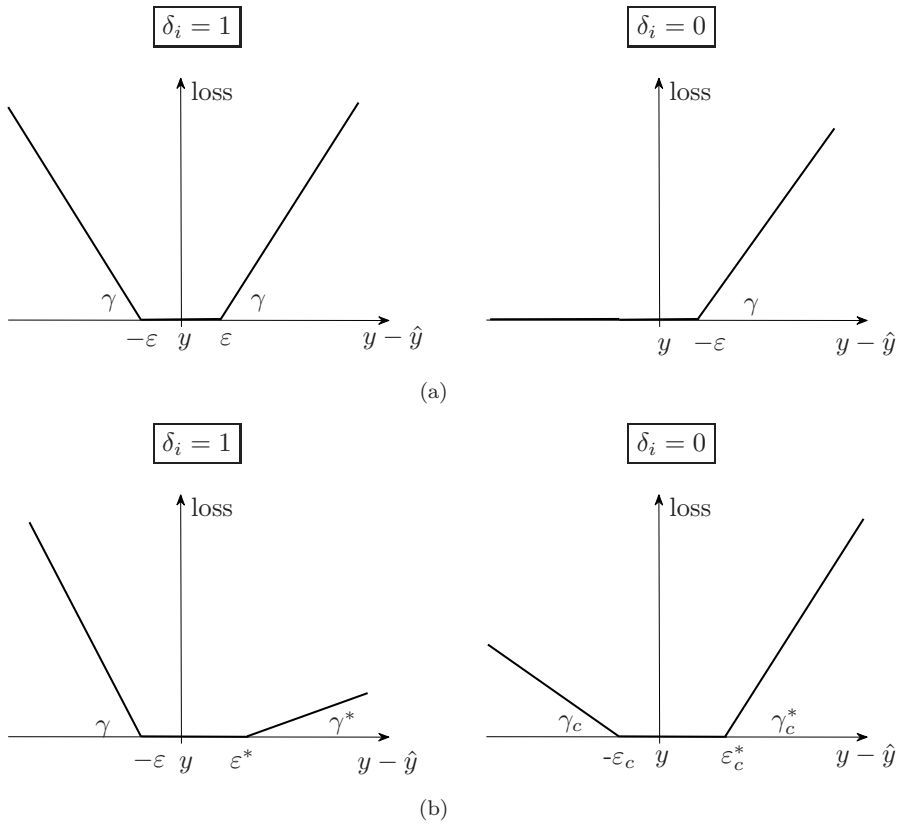


Figure 4.1: *Loss functions used in (a)* CSVR *as defined by Shivaswamy* et al. *(2007) and (b)* SVR*c as defined by Khan & Zubek (2008). Both methods differ in the way the loss function is calculated. A major disadvantage of the* SVR*c method is the necessity to define four hyperparameters.*

certain predefined time (e.g. surviving 5 years after surgery). However, to be able to include as many events as possible, this predefined time should be taken as late as possible. On the contrary, one would like an early time, in order to retain as many patients as possible, since all patients censored before that time will be lost for the analysis. A second problem in this approach is that the validity of the method decreases when more and more patients are censored earlier. To overcome the problems described above, it was proposed to formulate the survival problem as a ranking problem in Van Belle *et al.* (2007) and Evers & Messow (2008). A computational simplification was proposed in Van Belle *et al.* (2008). The idea behind formulating the survival problem as a ranking problem is that in clinical applications one is often interested in defining risk groups. One is not primarily

interested in a prediction of the survival time, but in whether the patient has a high or low risk for the event to occur, such that appropriate treatment can be given. To obtain this goal, an SVM ranking method is used, similar to the rankSVM model for ranking or preference learning (Herbrich *et al.*, 2000). The method proposed by Van Belle *et al.* (2007) and Evers & Messow (2008) involves a regularization as usual and a penalization for each comparable pair of observations for which the order in prognostic index differs from the observed order. The model is then formulated as

RANKSVMC

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \gamma \sum\limits_{i=1}^{n} \sum\limits_{j\,:\,y_i > y_j} \epsilon_{ij} \\ \text{subject to} & \begin{cases} w^T(\varphi(x_{(j)}) - \varphi(x_{(i)})) \geq 1 - \epsilon_{ij}, \\ \qquad\qquad \forall\ i = 1, \ldots, n;\ \forall\ j > i \\ \epsilon_{ij} \geq 0, \quad \forall\ i = 1, \ldots, n;\ \forall\ j > i\,. \end{cases} \end{array} \tag{4.4}$$

In the presence of censoring, only comparisons between comparable pairs (x_i, y_i, δ_i) and (x_j, y_j, δ_j) are included. Note that addition of a constant term b is not useful here since only differences in prognostic index are considered. The prognostic index for a new point $x_\star$ is found as

$$\hat{u}(x_\star) = \sum_{i=1}^{n} \sum_{j\,>\,i} \alpha_{ij} \left(\varphi(x_{(j)}) - \varphi(x_{(i)})\right)^T \varphi(x_\star)\,, \tag{4.5}$$

with α_{ij} the Lagrange multipliers.

A drawback of the method is that a QP of $\mathcal{O}(n^2)$ constraints needs to be solved. To overcome this problem, a computationally simplified approach solving a QP of $\mathcal{O}(n)$ constraints was proposed in Van Belle *et al.* (2008). The reduction was found by comparing each data point with its nearest comparable neighbor instead of comparing with all comparable data points. We will refer to this simplified model as RANKSVMC.

An important result on how to relate standard statistical survival models like the PH model and the accelerated failure time model with SVM-based survival models, is found in transformation models (see Chapter 3). The MINLIP model presented in the previous Chapter is similar to the RANKSVMC model, the difference lying in the right hand side of the first set of constraints in equation (4.4), where 1 is replaced by $y_{(i)} - y_{\bar{j}(i)}$ and the error term. We will consider the same error definition as in the RANKSVMC model in the remainder of this work. The model we will start

from in this Chapter is then formulated as

$$\boxed{\text{MODEL 1}}$$
$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \\ & \forall\ i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{4.6}$$

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

$$\hat{u}(x_\star) = \sum_{i=1}^{n} \alpha_i (\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T \varphi(x_\star)\,, \tag{4.7}$$

with α_i the Lagrange multipliers. We refer the interested reader to Appendix B.3 for a detailed deviation of the model.

4.2 Proposed SVM formulations for survival analysis

In this Section we propose two different adaptations to MODEL 1. In first instance, the proposal to reduce the complexity of the ranking formulation of Van Belle *et al.* (2008) will be used. Afterwards, it will be investigated whether the performance of the model can be improved by replacing this arbitrary choice of comparisons with unknown dummy observations. We will assume $u = w^T \varphi(x) + b$, unless stated otherwise.

In survival analysis the primary question to be answered is not to predict the failure time as accurately as possible, but to define risk profiles to the observations. This is exactly what is done in the PH: a prognostic index $u(x)$ is estimated and the link function between this predictor and the true survival time is left unspecified. Kernel based models for survival analysis also ignore $h(\cdot)$, but additionally let the estimation of the prognostic index free, resulting in more flexible models.

Instead of trying to estimate the survival time, a prognostic index is sought that estimates the order in which objects/persons relapse. In its most general form, the model can be formulated as

$$\begin{aligned} \min_{w,\epsilon} \quad & \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & \begin{cases} u(x_i) - u(x_{\bar{j}(i)}) + \epsilon_i \geq \rho_i, & \forall\ i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n\,, \end{cases} \end{aligned} \tag{4.8}$$

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

Two more problems need to be solved: (i) how to define ρ_i and (ii) how to include regularization. It is seen that the RANKSVMC model is a special case of the above formulation taking $\rho_i = 1$ and regularizing using the maximal margin principle (Vapnik, 1998). The methods that we propose here, go further along the work of Van Belle *et al.* (2009, which is also presented in Chapter 3), using the Lipschitz constant instead of the maximal margin for regularization purposes. Doing so solves the question as how to define ρ, since this methodology takes $\rho_i = y_i - y_{\bar{j}(i)}$.

Since MODEL 1 is built by minimizing the empirical risk on misrankings, the value of the prognostic model only has a relative importance. When one is interested in the prognostic model for which the prognostic index can be interpreted as times, one might include regression constraints as in Shivaswamy *et al.* (2007) and Khan & Zubek (2008). MODEL 1 is therefore modified as described below:

MODEL 2

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

$$\text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, \\ \qquad\qquad\qquad\qquad \forall\, i = 1,\ldots,n \\ w^T\varphi(x_i) + b \geq y_i - \xi_i, \quad \forall\, i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, \\ \qquad\qquad\qquad\qquad \forall\, i = 1,\ldots,n \\ \epsilon_i \geq 0, \qquad\qquad \forall\, i = 1,\ldots,n \\ \xi_i \geq 0, \qquad\qquad \forall\, i = 1,\ldots,n \\ \xi_i^* \geq 0, \qquad\qquad \forall\, i = 1,\ldots,n\,. \end{cases} \tag{4.9}$$

The first set of constraints is the same as in MODEL 1 and is the ranking constraint optimizing the concordance index. The second set of constraints indicates that the prognostic index should be larger than the observed failure time. The third set indicates that the prognostic index u should be smaller than y. Multiplication with the censoring indicator δ_i makes sure that this constraint is not taken into account for right censored data. Taking the second and third set of constraints together, the value of the prognostic index u is targeted at the failure time. However, for right censored data, the third constraint doesn't need to be fulfilled. The prognostic index for a new observation $x_\star$ is then found as

$$\hat{u}(x_\star) = \sum_{i=1}^{n}\left(\alpha_i(\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T + (\beta_i - \delta_i\beta_i^*)\varphi(x_i)^T\right)\varphi(x_\star) + b\,, \tag{4.10}$$

with α_i, β_i and β_i^* the Lagrange multipliers. See Appendix B.4 for more information.

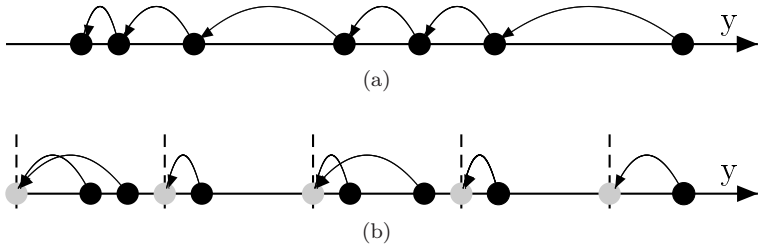


Figure 4.2: *Illustration of the difference between* MODEL 2 *(a) and* MODEL 3 *(b). In* MODEL 2 *each observation i (black solid dots) is compared (indicated by the arrows) with one observation with a failure time the closest to, but smaller than, the failure time of observation i. Since this is a rather arbitrary choice,* MODEL 3 *includes dummy observations with failure times distributed equidistantly over the time axis (indicated by the dashed vertical lines and grey circles). Each observation i is then compared with 2 dummy observations: (i) the dummy observation with the highest failure time lower than y_i (indicated by the arrows); and (ii) the dummy observation with the lowest failure time higher than y_i (not shown).*

Both MODEL 1 and MODEL 2 restrict the number of comparisons or ranking constraints to $\mathcal{O}(n)$ by comparing each observation with its closest (in observed time) comparable point. Since this choice is rather arbitrary, we question whether the performance of the model can be improved by generating k dummy observations with unknown utility v_l (also called threshold values) and known failure time s_l, with $l = 1, \ldots, k$. The s-values are spread uniformly over the observed time interval. The minimum s value equals zero since this is the lowest possible outcome, the largest s-value is higher than the highest observed outcome. All survival and censoring times lie therefore within the interval $[s_1, s_k]$. The inclusion of the thresholds is an elegant way to make an non-arbitrary choice of comparisons. The difference between this approach and the approach of MODEL 2 is illustrated in Figure 4.2). Assume no censoring for now. In MODEL 2 a comparison will be made between each observation i and the observation with the highest failure time for which this failure time is smaller than y_i. In MODEL 3, these comparisons are made less arbitrary by comparing each observation with a dummy observation.

Modifying MODEL 2 by including these thresholds leads to the following problem formulation

MODEL 3

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

$$\text{subject to} \quad \begin{cases} w^T\varphi(x_i) + b - \ell_i^T v \geq y_i - \ell_i^T s - \epsilon_i, & \\ \qquad\qquad \forall\, i = 1,\ldots,n & \\ -\delta_i(w^T\varphi(x_i) + b - r_i^T(v-s) - y_i) \geq -\epsilon_i^*, & \\ \qquad\qquad \forall\, i = 1,\ldots,n & \\ w^T\varphi(x_i) + b \geq y_i - \xi_i, & \forall\, i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\, i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall\, i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall\, i = 1,\ldots,n \\ \xi_i \geq 0, & \forall\, i = 1,\ldots,n \\ \xi_i^* \geq 0 & \forall\, i = 1,\ldots,n \\ v_l - v_{l-1} \geq 0, & \forall\, l = 2,\ldots,k\,. \end{cases} \tag{4.11}$$

In the above problem formulation ℓ_i and r_i represent vectors of length k containing all zero values, except for the value $\bar{l} = \arg\max_l l$, subject to $s_l < y_i$ for which $\ell_i(\bar{l})$ equals 1 and $\bar{l} = \arg\min_l l$ subject to $z_l > y_i$ for which $r_i(\bar{l})$ equals 1. The first and second set of constraints are similar to the ones in equation (4.9), replacing the arbitrary comparisons with the nearest neighbor by comparisons with the nearest thresholds. The last set of equations makes sure that the thresholds are increasing. The prognostic index $u(x_\star)$ for a new observation $x_\star$ can then be found as

$$\hat{u}(x_\star) = \sum_{i=1}^{n}(\alpha_i + \beta_i - \delta_i(\alpha_i^* + \beta_i^*))\varphi(x_i)^T\varphi(x_\star) + b\,, \tag{4.12}$$

where $\alpha_i, \alpha_i^*, \beta_i$ and β_i^* are the Lagrange multipliers. We refer to Appendix B.5 for more details.

4.3 Least-squares alternatives

The standard formulation of SVMs by Vapnik (1998) has been adapted towards LS-SVMs by Suykens & Vandewalle (1999); Suykens *et al.* (2002). Due to equality constraints and the use of a least-squares error in the cost function, the solutions of LS-SVM classification and regression problems are found by solving a linear system of KKT equations, instead of a QP in the SVM setting. In this section, the SVM approach for survival analysis is adapted in the same way as SVM for classification

was adapted to obtain the LS-SVM formulation for classification. Adapting MODEL 1 in this way results in the following model.

MODEL 4

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum\limits_{i=1}^{n} \epsilon_i^2 \\ \text{subject to} & w^T(\varphi(x_i) - \varphi(x_{\overline{j}(i)})) = y_i - y_{\overline{j}(i)} - \epsilon_i, \quad \forall\, i = 1, \ldots, n\,. \end{array} \tag{4.13}$$

The Lagrangian of (4.13) is

$$\begin{aligned} \mathcal{L}(w, \epsilon; \alpha) &= \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum_{i=1}^{n} \epsilon_i^2 \\ &\quad - \sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{\overline{j}(i)})) - y_i + y_{\overline{j}(i)} + \epsilon_i \right)\,. \end{aligned} \tag{4.14}$$

The KKT conditions for optimality are now found by derivation of the Lagrangian towards primal and dual variables:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum\limits_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\overline{j}(i)}) \right) \\ \frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \epsilon_i = \dfrac{\alpha_i}{\gamma} \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow w^T(\varphi(x_i) - \varphi(x_{\overline{j}(i)})) - y_i + y_{\overline{j}(i)} + \epsilon_i = 0\,. \end{cases} \tag{4.15}$$

After elimination of the primal variables, the solution is found as

$$\begin{aligned} \sum_{k=1}^{n} \alpha_k \left(\varphi(x_k) - \varphi(x_{\overline{l}(k)}) \right) \left(\varphi(x_i) - \varphi(x_{\overline{j}(i)}) \right) - y_i + y_{\overline{j}(i)} + \frac{\alpha_i}{\gamma}, \\ \forall\, i = 1, \ldots, n\,. \end{aligned} \tag{4.16}$$

The prognostic index $u(x_\star)$ for a new observation $x_\star$ is calculated as

$$u(x_\star) = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\overline{j}(i)}) \right) \varphi(x_\star)\,. \tag{4.17}$$

In analogy to MODEL 2 we introduce MODEL 5 including regression as well as ranking constraints. The model formulation is as follows:

$$\boxed{\text{MODEL 5}}$$

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum\limits_{i=1}^{n} \epsilon_i^2 + \frac{1}{2}\mu \sum\limits_{i=1}^{n} \xi_i^2 \\ \text{subject to} & \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\overline{j}(i)})) = y_i - y_{\overline{j}(i)} - \epsilon_i, \\ \qquad\qquad\qquad\qquad\qquad\qquad \forall\ i = 1, \ldots, n\,, \\ \delta_i \left(w^T \varphi(x_i) + b\right) = \delta_i y_i - \xi_i, \quad \forall\ i = 1, \ldots, n\,. \end{cases} \end{array} \tag{4.18}$$

The first set of constraints is the ranking constraint. The second set of constraints is targeting the value of the prognostic index at the failure time. For right censored observations, the errors are put to zero and thus not taken into account in the cost function. The solution is obtained from

$$\begin{cases} \sum\limits_{k=1}^{n} \left[\alpha_k \left(\varphi(x_k) - \varphi(x_{\overline{l}(k)})\right) + \beta_k \delta_k \varphi(x_k)\right] \left(\varphi(x_i) - \varphi(x_{\overline{j}(i)})\right) \\ \qquad\qquad -y_i + y_{\overline{j}(i)} + \frac{\alpha_i}{\gamma}, \qquad\qquad \forall\ i = 1, \ldots, n, \\ \sum\limits_{k=1}^{n} \left[\alpha_k \left(\varphi(x_k) - \varphi(x_{\overline{l}(k)})\right) + \beta_k \delta_k \varphi(x_k)\right] \delta_i \varphi(x_i) + \delta_i b + \dfrac{\beta_i}{\mu} \\ \qquad\qquad\qquad\qquad\qquad\qquad \forall\ i = 1, \ldots, n\,. \end{cases} \tag{4.19}$$

The prognostic index $u(x_\star)$ for a new observation $x_\star$ is calculated as

$$u(x_\star) = \sum_{i=1}^{n} \left[\alpha_i(\varphi(x_i) - \varphi(x_{\overline{j}(i)})) + \beta_i \delta_i \varphi(x_i)\right] \varphi(x_\star) + b\,. \tag{4.20}$$

4.4 Adaptations for high dimensional data

When dealing with high dimensional data, the risk of overfitting the model on the training data increases. Therefore, methods used to reduce this risk should be used. In Chapter 2, several penalization techniques were discussed. In Chapter 3, three additional methods based on PH, but suitable for high dimensional data were discussed.

Feature selection is included in MODEL 2 by constraining the weights w to positive weights (see also Chapter 3) and will be denoted as MODEL 2P.

MODEL 2P

$$\begin{aligned} \min_{w,\epsilon_i,\xi_i,\xi_i^*} \quad & \frac{1}{2}w^T w + \gamma \sum_{i=2}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} w^T x_i - w^T x_{\overline{j}(i)} \geq y_i - y_{\overline{j}(i)} - \epsilon_i & \forall\, i = 1, \ldots, n \\ w^T x_i + \xi_i \geq y_i & \forall\, i = 1, \ldots, n \\ -\delta_i w^T x_i + \xi_i^* \geq -\delta_i y_i & \forall\, i = 1, \ldots, n \\ w_p \geq 0 & \forall\, p = 1, \ldots, d \\ \epsilon_i \geq 0 & \forall\, i = 1, \ldots, n \\ \xi_i \geq 0 & \forall\, i = 1, \ldots, n \\ \xi_i^* \geq 0 & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{4.21}$$

Note the link with the nonnegative garotte estimator (Breiman, 1995; Yuan & Lin, 2007).

Since the above formulation can not be extended to non-linear kernels, we propose a two-step approach to include feature selection in this case. In a first step the model as described in (4.9) is solved. Using a non-linear but additive kernel $k(x_i, x_j) = \sum_{p=1}^{d} k^p(x_i^p, x_j^p)$ (e.g. the clinical kernel (Daemen & De Moor, 2009; Van Belle *et al.*, 2010b)), the estimated non-linear transformation of each covariate $x_\star^p$ can be estimated as

$$\begin{aligned} \tilde{x}_\star^p &= u^p(x_\star^p) \\ &= \sum_{i=2}^{n} \alpha_i \left[k^p(x_i^p, x_\star^p) - k^p(x_{i-1}^p, x_\star^p) \right] + \sum_{i=1}^{n} (\beta_i - \delta_i \beta_i^*)\, k^p(x_i^p, x_\star^p)\,. \end{aligned} \tag{4.22}$$

The second step performs feature selection by considering $\tilde{x}_\star^p$ as the covariates and applying model (4.21) with a linear kernel. We will refer to this model using a clinical kernel as MODEL $2\text{P}_{\text{clinical}}$.

4.5 Results

This Section compares the methods for survival analysis which are discussed in this Chapter on different types of data. First, the properties of the different kernel based methods are compared. The performances of these methods using a linear and a clinical kernel are compared on 6 clinical data sets concerning survival in cancer patients (see Appendix A for a description of the data sets). Next, the best

kernel based method is compared with other methods found in the literature. In a last comparison, methods able to deal with high dimensional data are compared on 3 micro-array data sets concerning survival of cancer patients. Table 4.1 gives an overview of the different kernel based survival methods discussed in this Chapter. It is indicated whether the methods use regression or ranking constraints and how many parameters need tuning in case a linear kernel is used. For methods using ranking constraints, it is mentioned how pairs for comparison are formed. A summary on whether the methods are able to include non-linear forms in the covariates, whether they are able to perform covariate selection, to handle high dimensional data, whether the hazard is estimated directly and which software was used for each method, is given in Table 4.2.

4.5.1 Comparing kernel based methods

The different optimization problems considered in this Chapter are compared with RANKSVMC and CSVR on 50 randomizations between training (two thirds of the data) and test set (remaining third of the data) on 6 clinical datasets, using a linear and a clinical (Daemen & De Moor, 2009) kernel. Tables 4.3 and 4.4 summarize the results. Statistically significant differences between MODEL 2 and the other models were calculated using the Wilcoxon signed rank sum test. It is seen that both MODEL 1, MODEL 4 and RANKSVMC are performing significantly less than the other models, independently of the used kernel. Addition of the regression constraints improves the performance significantly. No significant differences are seen between the arbitrary choice of comparisons and the approach with dummy observations. Comparing SVM with LS-SVM based approaches, the former perform better. This can be explained by the fact that the use of equality constraints necessarily imply that the information provided by the censored data (which is by nature an inequality) can not be incorporated in the regression constraints. In order to merge the advantage of LS-SVM based approaches of solving a system of linear equations and obtaining a higher performance, iteratively reweighted versions could offer a solution.

To investigate the difference in performance between models with and without regression constraints, 1000 bootstrap samples from one training set of the `VLC` data set were taken. Using a linear kernel, the estimated weights for each covariate and the resulting concordance index on the test set were calculated. Figures 4.3 and 4.4 show the distribution of the estimated weights of the Karnofsky performance score and the concordance index, respectively. Both models only taking ranking constraints into account result in high uncertainties about the weight estimates and therefore high differences in concordance index occur for slightly different training sets. Inclusion of regression constraints results in more stable estimates of weights and concordance index.

Table 4.1: *Overview of different kernel based survival methods and their properties, in case a linear kernel is used.*

	ranking constraint	regression constraint	# tuning parameters	formation of pairs
RANKSVMC	✓		1	nearest neighbor
CSVR		✓	1	
SVRc		✓	4	
MODEL 1	✓		1	nearest neighbor
MODEL 2	✓	✓	2	nearest neighbor
MODEL 3	✓	✓	2	dummy observation
MODEL 4	✓		1	nearest neighbor
MODEL 5	✓	✓	2	nearest neighbor

Table 4.2: *Summary of the different methods discussed in this Chapter. It is indicated whether the methods are able to model non-linearities, to perform covariate selection, to handle high dimensional data, whether or not the hazard is estimated, and which software program was used in the experiments.*

	reference	non-linear	covariate selection	high dimensional	hazard	software
RANKSVMC	Evers & Messow (2008) Van Belle *et al.* (2007, 2008)	✓		✓		mosek, matlab
CSVR	Shivaswamy *et al.* (2007)	✓		✓		mosek, matlab
SVRc	Khan & Zubek (2008)	✓		✓		mosek, matlab
MODEL 1	Van Belle *et al.* (2009)	✓		✓		mosek, matlab
MODEL 2	Van Belle *et al.* (2009b)	✓		✓		mosek, matlab
MODEL 2P	Van Belle *et al.* (2010c)	✓	✓	✓		mosek, matlab
MODEL 3	Van Belle *et al.* (2009b)	✓		✓		mosek, matlab
MODEL 4		✓		✓		mosek, matlab
MODEL 5		✓		✓		mosek, matlab
$\text{PH}_{\text{linear}}$	Cox (1972)				✓	matlab
PH_{l1}	Goeman (2010)		✓	✓	✓	R
PH_{l2}	Bøvelstad *et al.* (2007)			✓	✓	matlab
$\text{PH}_{\text{pspline}}$	Eilers & Marx (1996)	✓			✓	R
PLANNARD	Lisboa *et al.* (2003)	✓	✓		✓	matlab
PCR	Hastie *et al.* (2001)			✓	✓	matlab
SPCR	Bair *et al.* (2006)			✓	✓	matlab
PLS	Nygård *et al.* (2008)			✓	✓	matlab

Table 4.3: *Comparison of the different kernel based survival* SVM *models on 6 cancer data sets with clinical variables, using a linear kernel. The median performance and median absolute deviance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between* MODEL 2 *(indicated in grey) and the other models was tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly.*

data set	method	c-index	logrank χ^2	hazard ratio	# sv
VLC	MODEL 1	0.61±0.04[3]	1.38±1.07[3]	3.97±3.14[3]	84.5±6.5[3]
	MODEL 2	**0.69±0.02**	5.27±2.35	**10.5±3.97**	91±0.0
	MODEL 3	**0.69±0.02**	4.73±2.66	9±3.47	91±0.0
	MODEL 4	0.56±0.06[3]	1.14±0.96[3]	1.86±2.96[3]	91±0.0[1]
	MODEL 5	0.67±0.04[2]	4.31±2.74	8.45±5.24	91±0.0[2]
	RANKSVMC	0.62±0.05[3]	3.02±2.32[3]	5.03±2.87[3]	**78±3.0**[3]
	CSVR	**0.69±0.02**	**5.75±3.23**	9.5±4.56	89±1.0[3]
LCR	MODEL 1	0.55±0.05[3]	1.08±0.88[1]	2.4±1.27[3]	**61±7.0**[3]
	MODEL 2	**0.60±0.04**	1.87±1.21	**5.06±2.96**	86±0.0
	MODEL 3	**0.60±0.04**	**2.04±1.61**	4.77±2.45	84.5±1.5[3]
	MODEL 4	0.57±0.04[2]	1.06±1.01	2.49±1.14[3]	82±1.0[3]
	MODEL 5	0.57±0.04[1]	1.26±0.96[1]	2.68±1.23[3]	85±1.0[3]
	RANKSVMC	0.51±0.04[3]	0.51±0.44[3]	1.43±0.87[3]	72.5±2.5[3]
	CSVR	0.59±0.03	2.02±1.13	4.46±1.97	80±2.0[3]
LD	MODEL 1	0.62±0.04[3]	2.19±1.85[3]	4.45±2.61[3]	83.5±2.5[3]
	MODEL 2	0.68±0.02	7.53±3.38	13.80±6.14	86±0.0
	MODEL 3	0.63±0.04	3.07±2.51	7.68±4.56	86±0.0
	MODEL 4	0.61±0.07[3]	2.56±2.13[3]	5.08±3.68[3]	86±0.0[1]
	MODEL 5	0.66±0.03	**8.58±3.95**	13.20±7.07	86±0.0[2]
	RANKSVMC	0.62±0.06[3]	2.55±2.42[3]	4.68±3.04[3]	**80±4.0**[3]
	CSVR	**0.69±0.03**	6.70±2.99	**14.21±8.8**	81±5.0[3]
MLC	MODEL 1	0.60±0.03[2]	2.27±1.82	3.22±1.4	**88.5±11.5**[3]
	MODEL 2	**0.62±0.03**	3.32±2.13	**4.12±1.97**	110.5±0.5
	MODEL 3	**0.62±0.03**	2.72±2.01	3.64±2.0	111±0.0
	MODEL 4	0.54±0.05[3]	1.08±0.95[3]	2.18±1.08[3]	111±0.0[3]
	MODEL 5	**0.62±0.03**	**3.77±2.28**	4.09±1.93	111±0.0[3]
	RANKSVMC	0.59±0.04[1]	3.06±1.95	3.2±1.46	99±3.0[3]
	CSVR	0.61±0.03	2.27±1.49	3.48±1.60	95±3.0[3]
PC	MODEL 1	0.75±0.03	10.05±3.82	148.86±86.60	**64±18.0**[3]
	MODEL 2	0.76±0.02	11.86±4.10	231.79±119.05	322±0.0
	MODEL 3	**0.77±0.02**	**12.90±3.46**	**269.10±164.71**	322±0.0
	MODEL 4	0.72±0.04[3]	7.25±4.39[2]	57.71±36.06[3]	307±3.5[3]
	MODEL 5	0.71±0.03[3]	7.21±4.04[1]	71.71±36.38[3]	311±3.0[3]

Continued on next page

Table 4.3 – continued from previous page

data set	method	c-index	logrank χ^2	hazard ratio	# sv
	RANKSVMC	0.75±0.03	12.56±3.82	166.1±91.72	255.5±6.0[3]
	CSVR	0.76±0.02	10.15±3.13	198.45±111.21	137.5±6.5[2]
GBC	MODEL 1	0.64±0.03[3]	12.03±6.56[1]	36.02±26.14[2]	399.5±25.5[3]
	MODEL 2	**0.67±0.02**	16.45±3.81	70.14±48.90	457±0.0
	MODEL 3	**0.67±0.02**	17.74±5.51	**112.65±75.22**	457±0.0[2]
	MODEL 4	0.58±0.03[3]	4.11±3.24[3]	6.06±4.27[3]	445.5±1.5[3]
	MODEL 5	0.65±0.01[1]	15.36±4.74	29.10±11.73[3]	451±2.0[3]
	RANKSVMC	0.62±0.03[3]	9.99±5.33[3]	21.36±13.35[3]	399.5±7.5[3]
	CSVR	**0.67±0.02**	**20.17±4.08**	67.29±37.38	**366±37.0**[3]

[1], $p < 0.05$;[2], $p < 0.01$;[3], $p < 0.001$ (Wilcoxon signed rank sum test)

Table 4.4: *Comparison of the different kernel based survival methods on 6 cancer datasets with clinical variables, using a clinical kernel. The median performance and median absolute deviance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between* MODEL 2 *(indicated in grey) and the other models was tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly.*

data set	method	c-index	logrank χ^2	hazard ratio	# sv
VLC	MODEL 1	0.58±0.04[3]	2.01±1.71[3]	3.26±1.90[3]	91±0.0
	MODEL 2	0.69±0.02	6.27±2.77	9.78±3.93	91±0.0
	MODEL 3	0.69±0.02	6.85±3.01	9.85±4.40	90.5±0.5
	MODEL 4	0.53±0.04[3]	0.96±0.89[3]	1.48±0.72[3]	91±0.0[3]
	MODEL 5	**0.70±0.04**	7.59±4.34	9.88±6.41	91±0.0[3]
	RANKSVMC	0.57±0.06[3]	1.60±1.33[3]	2.70±1.40[3]	**77±4.0**[3]
	CSVR	0.69±0.02	**9.88±4.29**[1]	**11.12±4.32**	90±1.0[3]
LCR	MODEL 1	0.54±0.04[3]	1.04±0.92	1.98±0.81[2]	80±2.0[3]
	MODEL 2	0.59±0.03	1.08±1.00	3.56±1.90	83±3.0
	MODEL 3	0.58±0.02	1.17±1.02	2.56±1.00	**76.5±3.5**[3]
	MODEL 4	0.54±0.04[3]	0.86±0.76	1.65±0.95[3]	82±1.0
	MODEL 5	0.56±0.03[2]	0.8±0.70	2.08±0.84[2]	85±1.0[2]
	RANKSVMC	0.52±0.04[3]	0.37±0.30	1.38±0.73[3]	78±6.0[3]
	CSVR	**0.60±0.02**	**1.94±1.27**[1]	**4.04±1.90**	80.5±3.5[3]
LD	MODEL 1	0.66±0.05[2]	3.60±2.51[3]	5.97±2.97[3]	86±0.0
	MODEL 2	**0.69±0.03**	8.50±4.18	17.23±8.32	86±0.0
	MODEL 3	**0.69±0.03**	6.05±3.24	17.47±8.02	86±0.0
	MODEL 4	0.62±0.05[3]	2.14±1.43[3]	4.14±3.03[3]	86±0.0

Continued on next page

Table 4.4 – continued from previous page

data set	method	c-index	logrank χ^2	hazard ratio	# sv
	MODEL 5	0.66±0.04[1]	7.01±5.05	**19.52±10.3**	86±0.0[1]
	RANKSVMC	0.65±0.06[3]	2.68±2.16[3]	5.12±3.17[3]	**76±4.0**[3]
	CSVR	**0.69±0.02**	**10.70±4.25**	13.1±5.74[3]	82±4.0[3]
MLC	MODEL 1	0.61±0.03	3.13±2.16	3.48±1.64	106.5±3.5
	MODEL 2	0.61±0.03	2.53±1.86	3.28±1.41	111±0.0
	MODEL 3	0.61±0.04	2.77±1.84	2.96±1.23	111±0.0
	MODEL 4	0.57±0.04[3]	1.09±1.04	2.03±1.07[2]	111±0.0[2]
	MODEL 5	0.56±0.03[3]	1.00±0.93[2]	1.81±0.85[3]	111±0.0[3]
	RANKSVMC	0.61±0.03	2.55±2.10	3.30±1.65	**95±2.0**[3]
	CSVR	**0.62±0.03**	**3.70±1.94**[1]	**4.58±2.17**[1]	96±4.0[3]
PC	MODEL 1	0.76±0.02[1]	11.22±4.24[1]	108.39±52.18[1]	**88.5±21**[3]
	MODEL 2	**0.78±0.02**	**14.56±3.50**	148.08±60.05	112±6.5
	MODEL 3	**0.78±0.02**	14.52±3.68	144.72±64.11	89.5±6.0[3]
	MODEL 4	0.72±0.03[3]	8.33±4.63[3]	47.42±23.46[3]	307±3.5[3]
	MODEL 5	0.73±0.03[3]	10.42±4.41[1]	63.02±35.19[3]	311±3.0[3]
	RANKSVMC	0.76±0.03	11.89±3.37	111.26±53.54	264±7.0
	CSVR	**0.78±0.02**	13.15±3.91	**150.66±84.56**	160±23.5[1]
GBC	MODEL 1	0.63±0.04[3]	9.34±5.18[3]	10.1±5.06[3]	443±4.0[3]
	MODEL 2	**0.68±0.02**	**22.62±5.59**	**24.23±8.42**	457±0.0
	MODEL 3	**0.68±0.02**	22.42±6.17	24.20±6.95	456.5±0.5[3]
	MODEL 4	0.61±0.04[3]	8.51±6.62[3]	5.69±3.61[3]	445.5±1.5[3]
	MODEL 5	0.66±0.02[2]	15.7±6.02[2]	13.35±5.03[3]	451±2.0[3]
	RANKSVMC	0.62±0.04[3]	8.93±5.42[3]	11.20±5.70[3]	**404±9.5**[3]
	CSVR	**0.68±0.02**	19.51±5.36[1]	18.99±8.25	453±4.0[3]

[1], $p < 0.05$;[2], $p < 0.01$;[3], $p < 0.001$ (Wilcoxon signed rank sum test)

To investigate whether the improvement of MODEL 2 over MODEL 1 is dependent on the censoring percentage, artificial data are created in order to mimic the clinical setting. 100 datasets are constructed with 10 variables, 100 training and 500 test observations. All variables were normally distributed with zero mean and variance equal to 1. The correlations between the variables were zero except for the first and second covariate, the third and fourth, the fifth and sixth, and the seventh and eight, having a correlation coefficient of 0.7, 0.3, -0.7 and -0.3, respectively. The vector of regression coefficients w was taken to be $w = [-0.1, 1, 0.3, 0.4, 0.2, 0.03, 0.02, -0.01, -0.02, 0.01]^T$. The failure time was exponentially distributed with parameter equal to $w^T x$, with x the covariate vector. In a first example, the failure time was randomly censored (mimicking administrative censoring); in a second example the censoring time was sampled from the same distribution as the survival time $F_y(x)$, the spread of the distribution

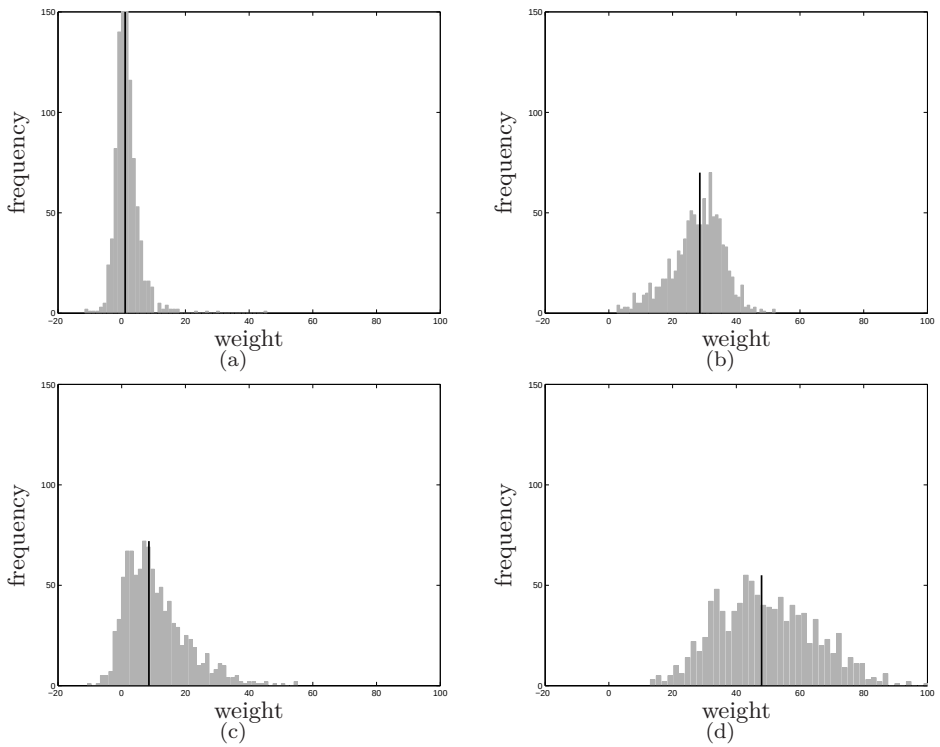


Figure 4.3: *Distribution of the estimated weights (using a linear kernel) of the Karnofsky score on 1000 bootstraps of one particular training set of the* VLC *data set: (a)* MODEL 1*, (b)* MODEL 2*, (c)* MODEL 4*, (d)* MODEL 5*.* MODEL 1 *and* MODEL 4 *have a broad distribution without a clear estimate (estimates with different signs). Addition of the regression constraints in* MODEL 2 *and* MODEL 5 *results in an undoubted positive relation between the Karnofsky score and the survival time ($w > 0$).*

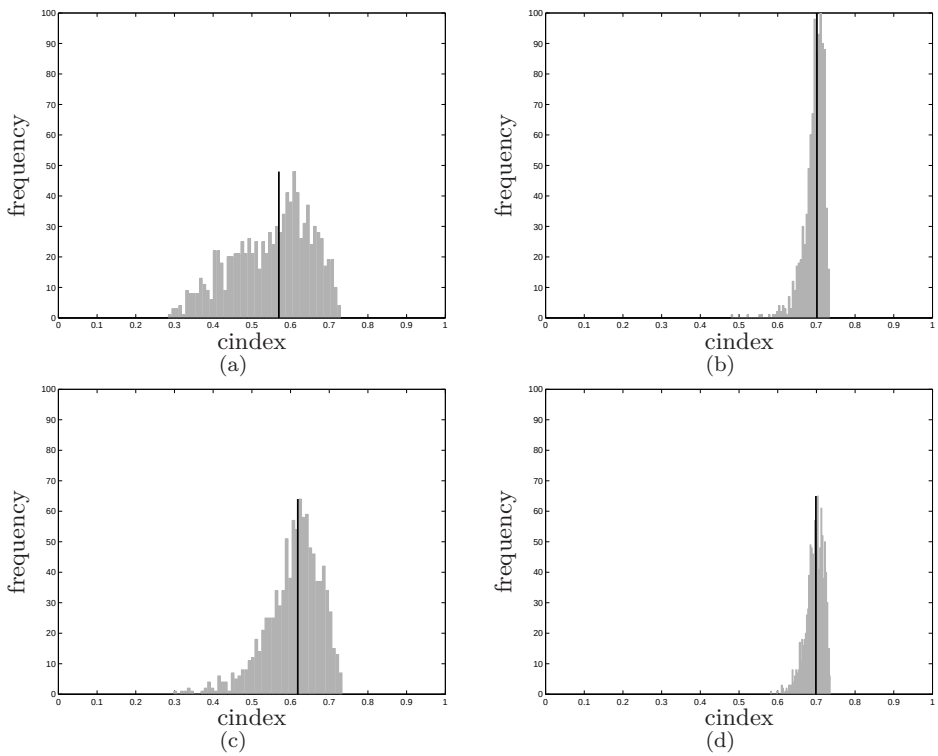


Figure 4.4: *Distribution of the concordance index on test on 1000 bootstraps of one particular training set of the* `VLC` *data set: (a)* MODEL 1*, (b)* MODEL 2*, (c)* MODEL 4*, (d)* MODEL 5*.* MODEL 1 *and* MODEL 4 *have a broad distribution from which one can not decide whether the estimated prognostic index is related or inversely related to the failure time. Addition of the regression constraints in* MODEL 2 *and* MODEL 5 *results into a narrowing of this distribution. This results in an undoubted positive relation between the prognostic index and the survival time (c-index*> 0.5*)*

was changed to create different censoring percentages ($G_y(x) = aF_y(x)$). Figure 4.5 illustrates the results. For both the random and informative censoring, the largest performance differences between both models are noted for low censoring percentages. However, MODEL 2 performs better than the MODEL 1 for all censoring percentages.

4.5.2 Comparison with other methods

This Subsection compares the results obtained with MODEL 2 with other methods found in the literature. In a first example six clinical data sets are used for comparison. A second example illustrates the use on high dimensional data sets.

Clinical data

MODEL 2 and MODEL 2P are tested using two different kernels: the linear kernel (MODEL 2_{linear} and MODEL $2P_{\text{linear}}$) and the clinical kernel (MODEL 2_{clinical} and MODEL $2P_{\text{clinical}}$). The PH model is tested linearly in the covariates (PH_{linear}), with ridge regression (PH_{l2}), with a LASSO penalty (PH_{l1}) and with penalized smoothing splines (PH_{pspline}). Table 4.5 illustrates the results. None of the methods performs overall better or worse than the other methods.

Table 4.5: *Comparison of test performances of different survival models on clinical data sets. Median and median average deviance on 50 random train-test set splits are given. Statistically significant differences between* MODEL 2P *(reference model: indicated in grey) and the other methods are denoted by:* $^1, p < 0.05; ^2, p < 0.01; ^3, p < 0.001$. *The best results are indicated in bold.*

data set	method	c-index	logrank χ^2	hazard ratio
VLC	MODEL 2_{linear}	0.68±0.02^3	4.19±3.17^3	7.07±2.75^3
	MODEL 2_{clinical}	0.70±0.02	7.84±2.85	10.24±4.36^1
	MODEL $2P_{\text{linear}}$	**0.71±0.02**	**8.13±3.71**	12.75±4.65
	MODEL $2P_{\text{clinical}}$	0.69±0.02^1	6.84±2.93	7.80±3.20^3
	PH_{linear}	0.68±0.03^3	5.45±1.97^1	10.61±2.99^2
	PH_{l2}	0.70±0.02	6.17±0.98	**31.26±7.47**^{1}
	PH_{l1}	0.70±0.02^1	7.04±2.85^2	11.41±4.47^2
	PH_{pspline}	0.70±0.03	8.11±3.29^3	14.29±6.76^2
	PLANNARD	**0.71±0.03**	6.76±2.82	11.18±6.25
LCR	MODEL 2_{linear}	0.59±0.02	1.50±1.30	3.57±1.54
	MODEL 2_{clinical}	0.58±0.04	1.05±0.96	2.79±1.36
	MODEL $2P_{\text{linear}}$	0.56±0.11	1.37±1.08	**4.30±3.19**

Continued on next page

Table 4.5 – continued from previous page

data set	method	c-index	logrank χ^2	hazard ratio
	MODEL $2P_{clinical}$	0.59±0.03	1.14±0.97	3.10±1.43
	PH_{linear}	**0.60±0.02**	1.60±1.18	4.05±1.34
	PH_{l2}	**0.60±0.02**	**1.80±1.58**	3.66±1.04
	PH_{l1}	0.57±0.03[1]	1.11±0.89	3.22±1.37
	$PH_{pspline}$	0.56±0.03[1]	1.12±0.91	3.42±1.69
	PLANNARD	0.56±0.03	0.72±0.68	2.22±1.16
LD	MODEL 2_{linear}	0.67±0.03[3]	5.94±3.31[3]	16.04±9.27
	MODEL $2_{clinical}$	0.69±0.03[1]	9.06±3.99[1]	20.98±9.90
	MODEL $2P_{linear}$	**0.72±0.03**	**12.54±3.90**	18.45±8.43
	MODEL $2P_{clinical}$	0.70±0.03	7.52±4.06[3]	**29.39±13.20**[1]
	PH_{linear}	0.69±0.03[2]	9.73±4.23	22.21±13.02
	PH_{l2}	0.68±0.02	7.13±3.48	13.9±7.00
	PH_{l1}	0.69±0.03[1]	6.97±3.93	21.13±12.06[1]
	$PH_{pspline}$	0.67±0.03	5.1±3.56	14.96±8.80
	PLANNARD	0.69±0.03[2]	8.07±3.86[1]	11.95±7.20[1]
MLC	MODEL 2_{linear}	0.62±0.03	2.56±1.64	4.28±2.21
	MODEL $2_{clinical}$	0.61±0.03	2.51±2.00	3.53±1.27[2]
	MODEL $2P_{linear}$	**0.63±0.03**	3.78±2.37	5.75±2.66
	MODEL $2P_{clinical}$	0.60±0.03[2]	1.64±1.03[1]	3.06±1.20[3]
	PH_{linear}	0.61±0.03[1]	2.89±1.95	**6.04±2.67**
	PH_{l2}	0.62±0.03	4.10±2.61	3.97±1.74
	PH_{l1}	0.60±0.03[3]	1.46±1.32[2]	3.55±2.16
	$PH_{pspline}$	0.56±0.03[3]	1.06±0.92[2]	2.41±1.29[3]
	PLANNARD	**0.63±0.02**	**4.46±2.20**	3.81±1.50
PC	MODEL 2_{linear}	0.76±0.02	11.27±3.34	239.91±123.05
	MODEL $2_{clinical}$	**0.78±0.02**	**13.80±4.49**[1]	149.23±79.55[1]
	MODEL $2P_{linear}$	0.77±0.02	10.48±3.22	250.09±131.39
	MODEL $2P_{clinical}$	**0.78±0.02**	13.61±3.91[2]	144.33±73.35[3]
	PH_{linear}	0.77±0.02	11.39±3.46	**266.26±150.48**
	PH_{l2}	0.76±0.02	11.19±3.44	228.04±107.23
	PH_{l1}	0.76±0.02	10.49±3.27	247.56±128.35
	$PH_{pspline}$	**0.78±0.02**	12.61±4.26	207.16±88.44
	PLANNARD	0.76±0.02	10.87±3.80	49.44±31.14[3]
GBC	MODEL 2_{linear}	0.67±0.02	16.97±4.84	79.38±49.57
	MODEL $2_{clinical}$	**0.68±0.02**	20.93±4.64[2]	22.21±8.04[3]
	MODEL $2P_{linear}$	0.67±0.01	17.64±2.61	73.57±37.03
	MODEL $2P_{clinical}$	**0.68±0.02**	**24.14±7.56**[2]	20.27±6.95[2]
	PH_{linear}	0.67±0.02	18.11±4.83	102.78±61.38
	PH_{l2}	**0.68±0.02**	20.16±4.2	48.19±18.49[2]
	PH_{l1}	0.67±0.02	15.20±3.99	**193.77±139.29**[2]

Continued on next page

Table 4.5 – continued from previous page

data set	method	c-index	logrank χ^2	hazard ratio
	$\text{PH}_{\text{pspline}}$	**0.68±0.02**	21.99±6.69[2]	158.69±107.79
	PLANNARD	0.67±0.02	15.00±5.11	19.63±9.28[3]

High dimensional data

Since this example concerns high dimensional data, MODEL 2 and MODEL 2P will be compared with PH models adapted for the high dimensionality in these data sets. Table 4.6 indicates that the clinical kernel performs better than the linear one. However, when including a feature selection algorithm, the results of the linear kernel become significantly better. MODEL 2P also outperforms all other tested methods. In addition to a better performance, MODEL 2P with a linear kernel results in a sparser model than the other methods. Due to the fact that the PH models approach dimensionality reduction by composing new features as a function of the measured covariates, nearly all gene expressions need to be obtained for any test case (see e.g. Park *et al.* (2002); Nguyen & Rocke (2002)). MODEL 2P is able to obtain a high performance based on less gene expressions.

4.6 Conclusions

This Chapter compared different support vector methods for survival analysis. Three questions were raised: (i) are ranking constraints sufficient to obtain good prognostic indices or is the inclusion of extra regression constraints significantly improving the model's performance?; (ii) Can the estimation of the prognostic index be improved by pinning the comparisons to dummy observations instead of arbitrarily choosing real observations to compare with; and (iii) Is there a performance difference when using LS-SVM based models versus SVM based models? The experiments revealed a significant improvement when including regression constraints in addition to ranking constraints. Inclusion of the thresholds did not improve the performance. Since a model with comparisons with the closest neighbor is simpler, we prefer to use the latter model. Changing the inequality constraints to equality constraints and using the 2-norm of the error results in a system of linear equations instead of a QP problem. However, the resulting models perform less than the SVM based models. The changes made in order to obtain a system of linear equations, lead to a symmetric loss function for both uncensored and censored observations. However, for censored observations a symmetric loss is less appropriate. MODEL 2 is therefore preferred above the other proposed models.

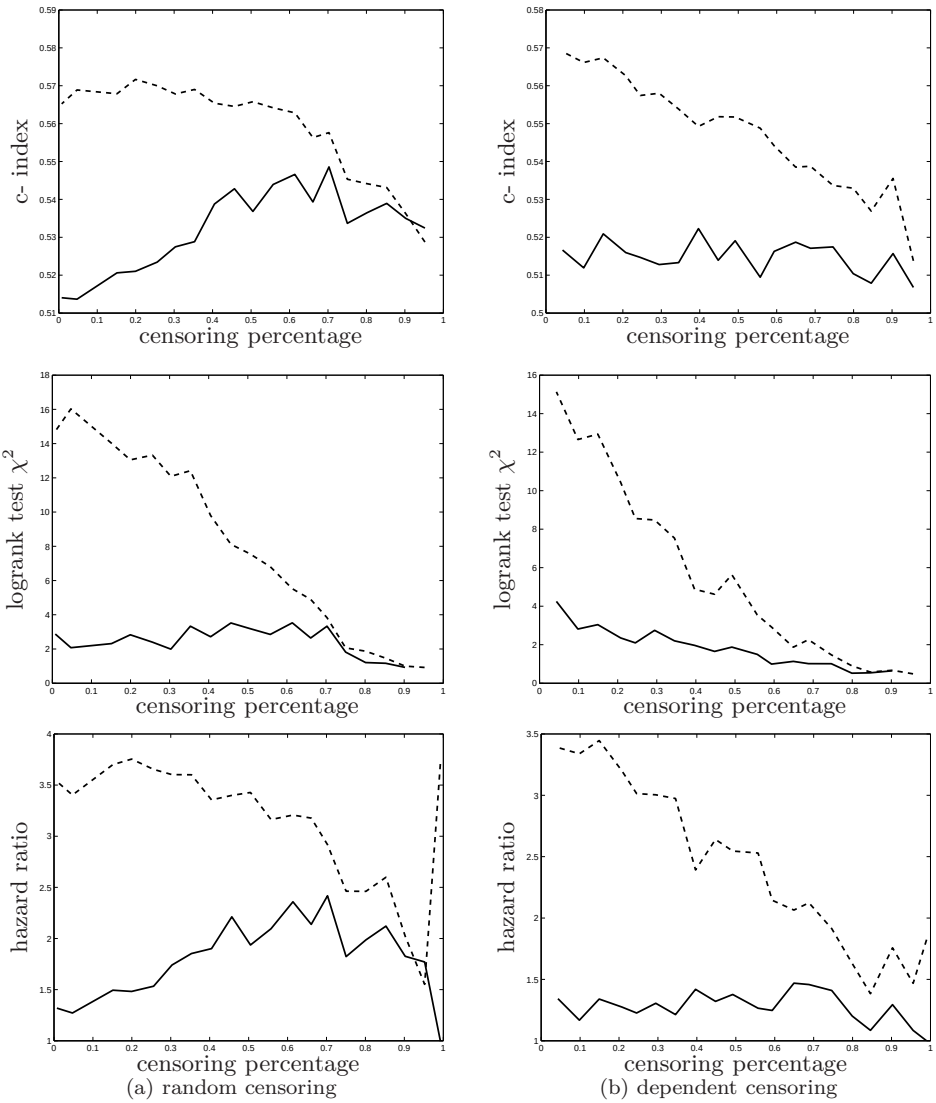


Figure 4.5: *Comparison between* MODEL 1 *(solid line) and* MODEL 2 *(dashed line) for (a) random and (b) informative censoring in a 10-dimensional artificial example. All variables were normally distributed with zero mean and variance equal to 1. All variables were uncorrelated except for 4 pairs of variables with correlations of 0.7, 0.3, -0.7 and -0.3. The coefficient vector w was taken to be fixed. The failure time was sampled from an exponential distribution with a parameter equal to $w^T x$, x being the variable vector. The percentage of censored observations was varied from 0 to 100%. The largest performance differences are noted for low censoring percentages. However,* MODEL 2 *seems to perform better for all censoring percentages, both for random and informative censoring.*

Table 4.6: *Test performance on three micro-array datasets.* MODEL 2 *and* MODEL 2P *are compared with five* PH *models with different regularization mechanisms to deal with high dimensional data. Median and median average deviance on 50 random train-test set splits are given. Statistically significant differences between* MODEL 2P *(reference model: indicated in grey) and the other methods are denoted by:* $^1, p < 0.05; ^2, p < 0.01; ^3, p < 0.001$. *The best results are indicated in bold.*

data set	method	c-index	logrank χ^2	hazard ratio	# weights
NSBCD	MODEL 2_{linear}	0.60±0.08^3	0.85±0.79^3	4.35±3.70^3	
	MODEL 2_{clinical}	0.71±0.03^3	4.46±2.31	16.90±9.46^3	
	MODEL $2P_{\text{linear}}$	**0.75±0.04**	**5.40±2.84**	**75.65±55.35**	171±16
	MODEL $2P_{\text{clinical}}$	0.68±0.04^3	3.72±2.10	16.49±10.18^3	288±183^3
	PCR	0.72±0.03^2	4.67±2.51	24.23±14.84^3	549±0^3
	SPCR	0.71±0.03^3	4.88±2.66	13.55±7.46^3	137±136^2
	PLS	0.73±0.04^1	4.00±2.09^1	21.39±11.58^3	549±0^3
	PH_{l2}	0.66±0.05^3	1.91±1.45^3	11.76±6.97^3	549±0^3
	PH_{l1}	0.71±0.04^3	4.73±1.89	17.97±8.21^3	**119.5±14**^{3}
DLBCL	MODEL 2_{linear}	0.61±0.03^3	4.14±2.41^3	6.22±3.21^3	
	MODEL 2_{clinical}	0.63±0.03^3	5.97±3.42^3	7.45±3.15^3	
	MODEL $2P_{\text{linear}}$	**0.76±0.02**	**22.47±5.72**	**224.78±143.10**	**2871±66**
	MODEL $2P_{\text{clinical}}$	0.62±0.02^3	5.71±2.50^3	7.53±3.02^3	7027±266^3
	PCR	0.60±0.02^3	2.88±1.83^3	4.66±1.90^3	7399±0^3
	SPCR	0.59±0.03^3	2.08±1.84^3	4.49±2.09^3	7399±0^3
	PLS	0.58±0.03^3	1.65±1.50^3	3.49±1.63^3	7399±0^3
	PH_{l2}	0.64±0.03^3	6.04±2.64^3	8.71±4.27^3	7399±0^3
	PH_{l1}	0.63±0.03^3	5.41±2.85^3	7.57±3.83^3	3960.5±2477
DBCD	MODEL 2_{linear}	0.64±0.02^3	3.49±1.48^3	7.74±2.62^3	
	MODEL 2_{clinical}	0.75±0.03^3	13.30±4.00^3	33.61±15.88^3	
	MODEL $2P_{\text{linear}}$	**0.82±0.02**	**18.66±4.97**	**360.47±233.84**	973±19
	MODEL $2P_{\text{clinical}}$	0.75±0.03^3	12.57±4.18^3	33.78±17.00^3	4232±271^3
	PCR	0.73±0.02^3	9.63±2.68^3	28.10±16.96^3	4919±0^3
	SPCR	0.73±0.03^3	10.31±3.02^3	24.34±12.09^3	**860±848**
	PLS	0.74±0.02^3	11.23±3.97^3	10.21±3.34^3	4919±0^3
	PH_{l2}	0.71±0.03^3	10.7±3.41^3	23.06±12.38^3	4919±0^3
	PH_{l1}	0.71±0.06^3	11.14±6.51^3	15.58±13.37^3	1786±896

For high dimensional data sets, inclusion of positivity constraints worked better than methods proposed in the literature in our experiments.

Comparison of MODEL 2 with PH revealed no significant differences in performance. The advantage of MODEL 2 above PH lies in the easy extension towards non-linear models without the need to check non-linearities in the variables before modelling. The PH model is most often used in clinical applications thanks to its easy applicability and interpretability. The disadvantages are that it assumes a linear parametric form of the covariates and proportional hazards. Both assumptions can be relaxed, the first e.g. by using regression splines, the second e.g. by including time dependent variables.

The PLANN model is a multi-layer perceptron model incorporating non-linearities in the covariates as well as interactions. The main disadvantage of such models is the difficulty to interpret them. Clinicians are generally interested in the contribution of each covariate to the estimated risk. Due to the complex architecture of multi-layer perceptrons, this contribution is more difficult to define. Inclusion of automatic relevance determination in the PLANN framework is a step in the right direction, although two problems remain. The first disadvantage of PLANN and PLANNARD is that they reduce the survival problem to time dependent classification problems. Therefore, all patients need to be replicated at each time at which they are at risk. This replication leads to an exponential increase of the problem to solve for an increasing number of data points. The second problem occurs when dealing with high dimensional data. For these data sets, the parameter space for multi-layer perceptrons becomes very large, with an increasing risk of overfitting the training data. Thanks to regularization mechanisms the number of effective parameters can be much lower than the number of weights. Nevertheless, for high dimensional data, training remains difficult and/or time-consuming. Finally, multi-layer perceptrons are known to be non-convex and the optimal solution will only be locally optimal.

MODEL 2 does not assume a linear form of the covariates, as the PLANN model, and has the additional advantage that data points do not need to be replicated. This is the result of approaching the survival problem as a combined ranking-regression approach instead of a time-dependent classification problem. The disadvantage of this model is that explicit estimation of the hazard is not incorporated in the model. The cumulative hazard can be estimated after categorizing patients into risk groups according to the estimated risk using the Nelson-Aalen estimator. The most important advantage of this model is the applicability and performance in high-dimensional problems. Additionally, since the survival problem is formulated in a convex way, the obtained solution is a global optimum.

In conclusion, one can state that the tested models have a similar performance on clinical data sets. However, on high dimensional data, MODEL 2P outperforms other models.

Chapter 5

Interval coded score index: a flexible model for interpretable clinical decision support

This Chapter aims at bridging the gap between advanced mathematical models and clinical practice. Although advanced methods can result in well performing diagnostic and prognostic models, their clinical application is not straightforward. In general, these methods yield a prediction without giving information on how this prediction is obtained. Rule extraction methods have been proposed to solve this issue. However, the resulting rules often lack clinical relevance. The method proposed in this Chapter starts from advanced mathematical models, but constrains the model to be interpretable and easily applicable in a clinical setting. This goal is obtained by restricting the models to additive models with stepwise functional forms. A prediction or prognosis is then obtained by summing points corresponding to the value of each input. Part of the results of this Chapter are summarized in Van Belle et al. *(2010).*

5.1 Introduction

Since the invention of the computer, people have acknowledged the possible advantages computers might offer in clinical decision making. In 1960 Ledley & Lusted (1960) summarized ideas on how computers could be used for medical data processing. Clinical decision support (CDS) systems have changed extensively since

these days. Thanks to increased computation power and the collection and storage of medical data, mathematical models designed for decision making are based on increasingly complex formulations. As a result, the models obtained perform better than in the early days. However, the price to be paid is the lower interpretability of the model by the user. The earliest CDS systems were problem-specific flowcharts designed by clinicians and encoded for use by a computer (Musen *et al.*, 2006), whereas later CDS systems are based on logistic regression models (Hosmer & Lemeshow, 2000), artificial neural networks (Bishop, 1995; Ripley, 1996; Suykens *et al.*, 1996), support vector machines (Vapnik, 1998; Suykens *et al.*, 2002), among others. A broad review on the use of these methods within cancer research is found in Cruz & Wishart (2006).

Although the possible use of CDS was mentioned 50 years ago, CDS systems are still not widely applied nor accepted in clinical practice (Miller, 1994; Wyatt & Altman, 1995; Bates *et al.*, 2003; Wears & Berg, 2005; Kong *et al.*, 2008). Potential features to make CDS systems more acceptable for clinical practice have been proposed by several authors (Percell, 2005; Kawamoto *et al.*, 2005; Osheroff, 2009). In Kawamoto *et al.* (2005) four features which are critical for CDS systems are identified: (i) the CDS system should be provided to clinicians in an automatic way, without interfering in the work flow, (ii) it should provide decision support at the time and location of decision making, (iii) it should provide a recommendation and (iv) it should be implemented on a computer. However, there is no real consensus on the properties of CDS systems. While some suggest they should be implemented on a computer, others state that prediction models should be taken to the bedside. Kattan proposed the use of nomograms (d' Ocagne, 1899) as a CDS system (Smaletz *et al.*, 2002; Kattan, 2003b,a; Van Zee *et al.*, 2003; Karakiewicz *et al.*, 2005; Kattan, 2005; Specht *et al.*, 2005; Chun *et al.*, 2007a,b; Shariat *et al.*, 2008). Nomograms were originally intended as an aid for engineers in solving mathematical equations, but an alternative nomogram is more recently used to calculate patient specific probability estimates. For this purpose, a nomogram is defined as a graphical tool representing a regression model, such that the user is able to calculate the patient specific probability without the use of a calculator or computer, given certain variables (see Figure 5.1 for an example). However, applications of nomograms in clinical practice are rare and mostly used within software applications. The problem of implementing CDS systems becomes even harder when using more advanced models. Musen *et al.* (2006) summarized this issue as: "There is no way that an observer can directly understand why an artificial neural network might reach a particular decision".

The main difficulties lie in the interpretability and applicability of the designed models. Much of the before-mentioned methods use unknown non-linear transformations of the input variables to form a prediction. Although the resulting models in general improve the performance in comparison with standard statistical methods (as concluded in Evers & Messow (2008)), the impact of the different

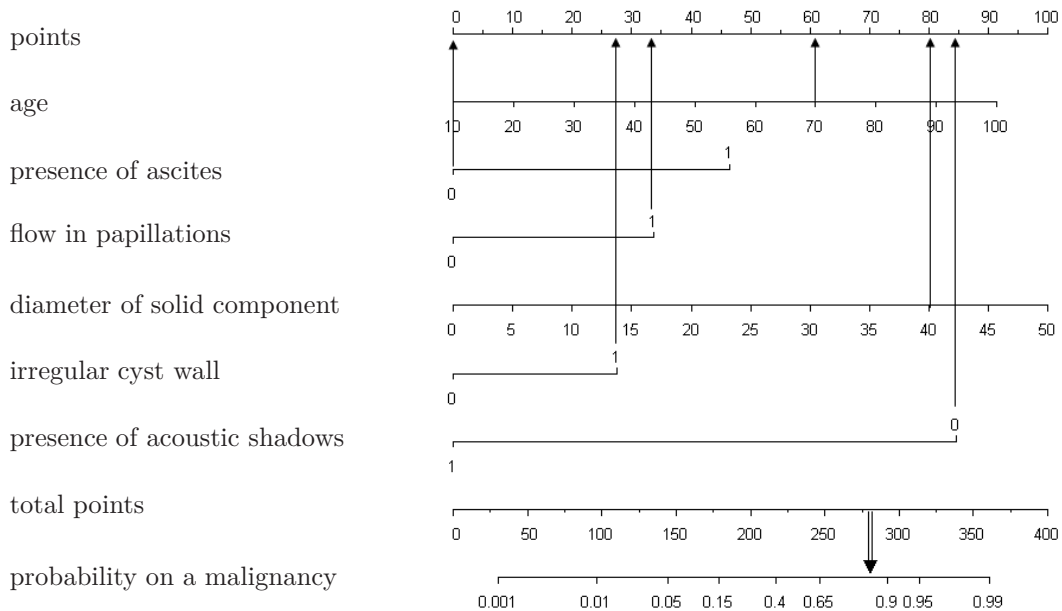


Figure 5.1: *Nomogram for the diagnosis of ovarian cancer derived from a logistic regression model (LR2 in Timmerman* et al. *(2005)). For each particular patient, the clinician has to draw vertical lines from the value of each variable to the point axis (upward arrows). The obtained points need to be added to get the total number of points for the patient. The probability of a malignant mass is found by drawing a vertical line from the total number of points to the probability axis (downward arrow).*

input variables on the outcome is unknown. Rule extraction methods for support vector classifiers (Chen *et al.*, 2007; Barakat & Diederich, 2005; Barakat & Bradley, 2007; Chaves *et al.*, 2005; Martens *et al.*, 2007; Núñez *et al.*, 2006; Fung *et al.*, 2005; Zhang *et al.*, 2005) have been developed to get an idea on how the model obtains its predictions. The previous citations are discussed and compared in Barakat & Bradley (2010). All the mentioned works start with training a SVM with labelled examples. In a second step rules are to be extracted. The different methods differ in the techniques used to build rule extraction models. This involves the use of the training data together with techniques such as clustering and decision trees. Although this approach is a step in the direction of applicable models, its main advantage remains the explanation of the decision boundary formed by the classifier. The resulting rules can still be rather complex (e.g. if four conditions are satisfied then malignant, etc.) and are still not widely accepted by clinicians.

This Chapter proposes a different approach, much in the line of the points system for logistic regression and proportional hazards regression models presented in Sullivan *et al.* (2004). As in the latter work, the goal is to create a scoring system or points system in order to make complex statistical models useful for clinicians. Instead of trying to extract rules from a previously built model, a model for which the underlying structure can be interpreted as a weighted sum of yes/no questions is generated. The strategy is as follows. All input variables $p = 1, \ldots, d$ are represented by $k_p + 1$ different binary variables. Those $k_p + 1$ variables are formed by dividing the measurable range of the variable p into $k_p + 1$ consecutive intervals. Consider the demographical variable age of the patient. For a study of breast cancer, a reasonable range for this variable is 25 to 95 years. A clinician would be interested to know whether an increase of 5 years in age has an effect on the probability of the tumor to be malignant. The input variable age is therefore represented by 15 binary variables answering the questions whether the age of the patient lies between 25 and 30 years, between 30 and 35 years, etc. The effect of the original input variable (functional form) is then modelled as the sum of the effects of the binary variables. The result is that a (non-linear) transformation of each input variable is found, such as within support vector classifiers with componentwise kernels (Pelckmans *et al.*, 2005c). However, the difference lies in the type of transformation, which will be a step function in the proposed method. This method is a special type of generalized additive models (Hastie & Tibshirani, 1986), for which the functional forms of the variables are restricted to piecewise constant functions (step functions or B-spline constant functions (de Boor, 1978; Hastie *et al.*, 2001)). By applying an L_1-norm regularization scheme (Tibshirani, 1996; Boyd & Vandenberghe, 2004), the number of steps in this function is minimized. This regularization can be used in different problems with sparsity conditions (see for example Pelckmans *et al.* (2005a); Candès *et al.* (2008)). Additionally, this regularization will perform feature selection. Although we present this method for support vector machines, its applicability is not restricted to this type of models.

5.2 Scoring systems for classifiers and prognostic models based on machine learning methods

This Section elaborates how simple scoring systems can be obtained starting from advanced mathematical methods. In the first Subsection, the SVM classifier (Vapnik, 1998) is adapted to obtain a simple scoring system. The next Subsection describes the proposed approach starting from MODEL 2, which was discussed in the previous Chapter.

5.2.1 Classification

The mathematical model proposed by Vapnik (1998) for linear classification is formulated as follows (see Chapter 2):

$$\begin{array}{ll} \min\limits_{w,b,\epsilon} & \frac{1}{2}w^T w + \gamma \sum\limits_{i=1}^{n} \epsilon_i \\ \text{subject to} & \begin{cases} y_i \left(w^T x + b\right) \geq 1 - \epsilon_i, & \forall\, i = 1, \ldots, n\,, \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n\,. \end{cases} \end{array} \tag{5.1}$$

In short, one tries to find the coefficients w_p with $p = 1, \ldots, d$ and a constant b, such that the model $\hat{y}_i = w^T x_i + b$ indicates what the most likely outcome for observation i is. In practice, one predicts a benign result for $\hat{y}_i < 0$ and a malignant one for $\hat{y}_i > 0$. This model is generalized towards non-linear classifiers by means of a feature map $\varphi(x)$, representing an unknown transformation function of the variables x. This model was explained in more detail in Section 2.3.1.

Defining the model

The method presented here involves three modifications with regard to equation (5.1). The first two modifications involve the type of transformation $\varphi(x)$ of the variables. A major problem for the application of models built using (5.1) in clinical settings is that the transformations can involve all variables. To increase the interpretability, we will only consider additive models. This involves that only transformations of the form

$$\varphi(x) = \sum_{p=1}^{d} \varphi_p(x^p) \tag{5.2}$$

will be allowed. If for the problem at hand one is interested in interactions, or one knows from literature that an interaction between two variables exists, one

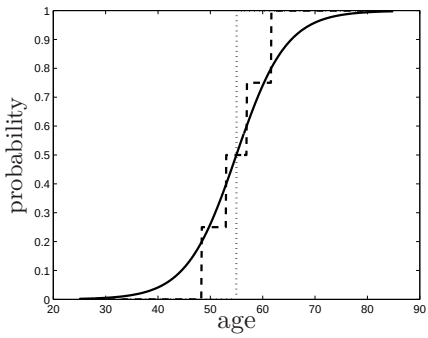


Figure 5.2: *Illustration of the allowed type of transformation of the input variables. Consider the age of female patients and the question: What is the probability for a particular woman to be post menopausal, given her age? Assume the black solid curve to be the true function. In the proposed model, only step functions (e.g. dashed and dotted curves) will be allowed. Depending on the problem a small increase in the probability will be relevant (such as for the dashed curve) or not (such as for the dotted curve). The model itself will find out how many steps are relevant, where these steps should be located and how big the steps should be.*

can include interactions concerning two variables. The second modification in the transformation function involves that we demand the transformation to be a step function. Figure 5.2 illustrates the principle. Consider the simple problem where one wishes to estimate the probability that a woman is post menopausal, given her age. Assume the black curve to be the true function. In the proposed model, only step functions (e.g. dashed and dotted curves) will be allowed. Depending on the problem a small increase in the probability will be relevant (such as for the dashed curve) or not (such as for the dotted curve). The model itself will find out how many steps are relevant, where these steps should be located and how big the steps should be. The presented model is thus a generalized additive model (Hastie & Tibshirani, 1990) with functional forms closely related to constant B-splines (de Boor, 1978; Hastie *et al.*, 2001). A third adaptation involves the sparsity of the model, meaning that we want only a small number of variables to be included in the model, and only a few steps will be allowed in the step function for each variable. This feature is obtained by minimizing the total variation of the coefficient vector w (see Rudin *et al.*, 1992, and references therein).

Inclusion of all above-mentioned adaptations leads to the following model:

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \sum_{p=1}^{d}\sum_{l=1}^{k_p+1} |w_{pl} - w_{p,l-1}| + \gamma \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & \begin{cases} y_i \left(\sum_{p=1}^{d}\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}(\tau_{p,l-1} \leq x_i^p < \tau_{p,l}) + b \right) \geq 1 - \epsilon_i, & \\ & \forall\ i = 1, \ldots, n\,, \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{5.3}$$

In the above formulation, k_p thresholds $\tau_{p,l}$, $l = 1, \ldots, k_p$ are included for variable p, such that $k_p + 1$ binary indicators $\mathcal{I}(\tau_{p,l-1} \leq x_i^p \leq \tau_{p,l})$, with $\tau_{p,0} = -\infty$ and $\tau_{p,k_p+1} = \infty$, represent the answers on $k_p + 1$ yes/no questions. Considering the example of the post menopausal status, these questions will be: is the woman between 25 and 30 years old, between 30 and 35, ..., older than 85? The weights $w_{p,l}$ represent the effect of being between 25 and 30 years old on the risk of being post menopausal. For categorical variables, the questions become whether the value of the variable is equal to the first level, the second level, etc. For binary variables, the question becomes whether the value of the variable is 1. Note the link with compressed sensing (Candès & Wakin, 2008). To achieve a very simple model including only a few steps in the transformation of some of the variables, the absolute value of the difference between the coefficients is minimized (first line in (5.3)) (Rudin *et al.*, 1992). Note that in equation (5.3) $y_i = -1$ for observations of the first class (non-events) and $y_i = 1$ for observations of the second class (events). Finally, the score for an observation with variables x_* will be found as

$$\text{score} = \sum_{p=1}^{d} \left(\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}(\tau_{p,l-1} \leq x_*^p < \tau_{p,l}) \right) + b\,. \tag{5.4}$$

Iteratively reweighting the model

A problem with the proposed method is that it can still lead to a solution including small, and for clinicians probably irrelevant steps. This problem is illustrated in Figure 5.3 on the menopausal problem stated before. Algorithm 1 will be used to reduce the number of small steps (see Candès *et al.* (2008) for more information). After solving model (5.3), weighting factors $\chi_{p,l}$ for differences in coefficients $w_{p,l} - w_{p,l-1}$ are calculated. The weighted model (see equation (5.5)) is iteratively solved until the estimated coefficients in two successive iterations differ by less than a predefined constant (here 10^{-8}).

Algorithm 1 Iteratively reweighted L_1 minimization for classification models

1: Determine the optimal tuning parameter γ, e.g. by 10-fold cross-validation.
2: Solve equation (5.3).
3: **repeat**
4: Let $w_{\text{old}} = w$
5: Find w, b from

$$\min_{w,b,\epsilon} \quad \sum_{p=1}^{d} \sum_{l=1}^{k_p+1} \chi_{p,l} |w_{pl} - w_{p,l-1}| + \gamma \sum_{i=1}^{n} \epsilon_i$$

$$\text{subject to} \quad \begin{cases} \hat{y}_i = \sum_{p=1}^{d} \left(\sum_{l=1}^{k_p+1} w_{p,l} \mathcal{I}(\tau_{p,l-1} \leq x_i^p < \tau_{p,l}) \right) + b, & \\ & \forall\ i = 1, \ldots, n\,, \\ y_i \hat{y}_i \geq 1 - \epsilon_i, & \forall\ i = 1, \ldots, n\,, \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \tag{5.5}$$

6: Let $\chi_{p,l} = f(w_{p,l} - w_{p,l-1}),\ \forall\ p = 1, \ldots, d$ and $l = 1, \ldots, k_p + 1$.
7: **until** $|w - w_{\text{old}}| < 10^{-8}$

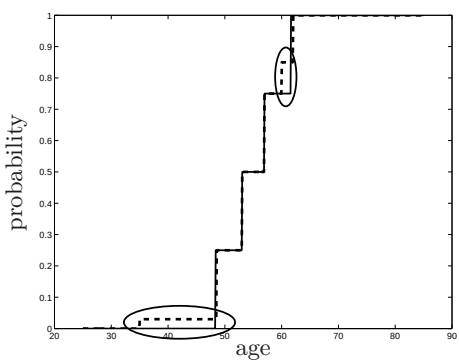


Figure 5.3: *Illustration of unwanted small steps which can be reduced by using reweighted L_1 minimization. The steps in the dashed step function indicated by the circles are unwanted since there is no clinical interpretation for such effects. Using reweighted L_1 minimization, the solid step function can be obtained.*

The $\chi_{p,l}$ should be chosen such that small differences in the coefficients w get a large penalization whereas large differences only get a small penalization. Doing so will result in less but larger steps.

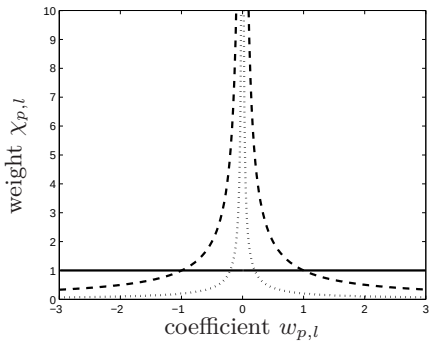


Figure 5.4: *Different weighting schemes. Black curve: no weighting. Dashed curve:* $\chi_{p,l} = \frac{1}{\varepsilon+|w_{p,l}-w_{p,l-1}|}$, $\forall\ l = 1,\ldots,k_p+1$. *Dotted curve:* $\chi_{p,l} = \frac{1}{\varepsilon+5|w_{pl}-w_{pl-1}|}$, $\forall\ p = 1,\ldots,d,\ \forall\ l = 1,\ldots,k_p+1$.

Different weighting schemes can be used depending on the desired result. We will consider the following schemes:

$$\chi_{p,l} = \frac{1}{\varepsilon + a|w_{pl} - w_{pl-1}|},,\ \forall\ p = 1,\ldots,d,\ \forall\ l = 1,\ldots,k_p+1\,, \tag{5.6}$$

where $\varepsilon = 0.0005$ and a can be optimized for the problem at hand. The dependence of the weighting scheme on the value of a is illustrated in Figure 5.4. The black curve represents the model as formulated in equation (5.3), without any weighting. The dashed and dotted weighting functions represent weighting schemes with $a = 1$ and $a = 5$, respectively. The difference between both schemes lies in the suppression of the weights. The first scheme (dashed curve), will suppress much more coefficients and will therefore result in a sparser model. Depending on the problem, it will be necessary to introduce different weighting schemes in order to obtain the desired result. The optimal weighting function can be chosen on the training set. The goal is to look for a weighting function which improves the sparsity of the obtained model without reducing the performance on the training (or validation) set.

Although the result is easy to interpret, it is not yet easy to use. We therefore propose to normalize the coefficients $w_{p,l}$ such that the smallest non-zero absolute value of the coefficients becomes 1. All other normalized coefficients are rounded to the nearest integer. The final score for a new observation $x_\star$ is then found as

$$\text{score} = \sum_{p=1}^{d}\left(\sum_{l=1}^{k_p+1}\tilde{w}_{p,l}\mathcal{I}(\tau_{p,l-1} \le x_\star^p < \tau_{p,l})\right)\,, \tag{5.7}$$

with $\tilde{w}_{p,l} = \text{round}\left(\frac{w_{p,l}}{\min\limits_{|w_{p,l}|>0} w_{p,l}}\right)$. This score is then used to make a prediction by comparing the score with b.

Prediction of the probability of the event

The model as described up to now is a scoring system, but it does not present the clinician with a probability of the event at study (second step in TMS). In order to obtain the probability of the outcome (e.g. a malignant mass) monotone least-squares SVM regression, based on monotone kernel regression as presented by Pelckmans *et al.* (2005b), is used. Additional constraints, above the monotonicity constraints, indicate that the prediction is restricted to the interval $[0, 1]$ such that it can be interpreted as a probability. The primal model formulation is

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} \epsilon_i^2 \\ \text{subject to} \quad & \begin{cases} w^T \varphi(x_i) + b = y_i + \epsilon_i, & \forall\, i = 1, \ldots, n\,, \\ w^T (\varphi(x_{(i)}) - \varphi(x_{(i-1)})) \geq 0, & \forall\, i = 2, \ldots, n\,, \\ w^T \varphi(x_i) + b \geq 0, & \forall\, i = 1, \ldots, n\,, \\ -w^T \varphi(x_i) - b \geq -1, & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{5.8}$$

In the above formulation x_i is the score as defined in equation (5.7), y_i equals 0 for non-events and 1 for events and $x_{(i)}$ is sorted such that $i > j$ if $x_i > x_j$. Building the Lagrangian and solving the KKT equations yields the dual problem

$$\begin{aligned} \min_{\alpha,\beta,\eta,\eta^*} \quad & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\alpha_i \alpha_j + \eta_i \eta_j + \eta_i^* \eta_j^*\right) \varphi(x_i) \varphi(x_j) \\ & + \frac{1}{2} \sum_{i=2}^{n} \sum_{j=2}^{n} \beta_i \beta_j \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right) \left(\varphi(x_{(j)}) - \varphi(x_{(j-1)})\right) \\ & + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=2}^{n} \left(\alpha_i + \eta_i + \eta_i^*\right) \beta_j \varphi(x_i) \left(\varphi(x_{(j)}) - \varphi(x_{(j-1)})\right) \\ & + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \left(\eta_j - \eta_j^*\right) \varphi(x_i) \varphi(x_j) \\ & - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_i \eta_j^* \varphi(x_i) \varphi(x_j) - \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \eta_i^* \\ \text{subject to} \quad & \begin{cases} \beta_i \geq 0 & \forall\, i = 2, \ldots, n\,, \\ \eta_i \geq 0 & \forall\, i = 1, \ldots, n\,, \\ \eta_i^* \geq 0 & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{5.9}$$

The predicted probability of the event for a new point $x_\star$ is found as

$$\hat{y}_\star = \left(\sum_{i=1}^{n} (\alpha_i + \eta_i - \eta_i^*) \varphi(x_i) + \sum_{i=2}^{n} \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)}) \right) \right) \varphi(x_\star) \,. \tag{5.10}$$

Visualization of the model

In order to facilitate the clinical application of the resulting models, the model is represented in two ways. A first possible representation is a questionnaire. A second possibility is to represent the model by means of color bars. For each input variable a color bar is visualized, the colors of which correspond to the coefficients $w_{p,l}$. The clinician can find the final score by adding the points for each color bar. A final color bar links the final score with the estimated probability. Section 5.3 gives two illustrations of the results. In the remainder of this work, the developed methodology will be denoted as the Interval Coded Score (ICS) Index.

5.2.2 Survival analysis

Defining the model

To develop an ICS model for prognostic problems, we start from MODEL 2 as described in Chapter 4, since we will use clinical data. The model is adapted in the same way as the SVM classification formulation of Vapnik was adapted and becomes:

$$\begin{aligned}
\min_{w,b,\epsilon,\xi,\xi^*} \quad & \sum_{p=1}^{d} \sum_{l=1}^{k_p+1} |w_{p,l} - w_{p,l-1}| + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi + \xi^*) \\
\text{subject to} \quad & \begin{cases}
\hat{y}_i = \displaystyle\sum_{p=1}^{d} \left(\sum_{l=1}^{k_p+1} w_{p,l} \mathcal{I}(\tau_{p,l-1} \le x_i^p < \tau_{p,l}) \right) + b, & \\
\qquad\qquad \forall\, i = 1, \ldots, n\,, & \\
\hat{y}_i - \hat{y}_{i-1} + \epsilon_i \ge y_i - y_{i-1}, \qquad \forall\, i = 2, \ldots, n\,, & \\
\hat{y}_i \ge y_i - \xi_i, \qquad \forall\, i = 1, \ldots, n\,, & \\
-\delta_i \hat{y}_i \ge -\delta_i y_i - \xi_i^*, \qquad \forall\, i = 1, \ldots, n\,, & \\
\xi_i \ge 0, \qquad \forall\, i = 1, \ldots, n\,, & \\
\xi_i^* \ge 0, \qquad \forall\, i = 1, \ldots, n\,, & \\
\epsilon_i \ge 0, \qquad \forall\, i = 1, \ldots, n\,. &
\end{cases}
\end{aligned} \tag{5.11}$$

As in the classification case, the score is found as

$$\text{score} = \sum_{p=1}^{d} \left(\sum_{l=1}^{k_p+1} w_{p,l} \mathcal{I}(\tau_{p,l-1} \leq x_\star^p < \tau_{p,l}) \right) . \tag{5.12}$$

Iteratively reweighting the model

To reduce the number of steps in each functional form of the input variables, an iterative reweighted L_1 minimization is performed in the survival setting as well. The algorithm is summarized in Algorithm 2. Again, the coefficients $w_{p,l}$ corresponding to each variable, are not limited to integers. The normalized coefficients are therefore found as

$$\tilde{w}_{p,l} = \text{round} \left(\frac{w_{p,l}}{\min\limits_{|w_{p,l}|>0} w_{p,l}} \right) , \tag{5.13}$$

and the final score is calculated by summing all points as

$$\text{score} = \sum_{p=1}^{d} \left(\sum_{l=1}^{k_p+1} \tilde{w}_{p,l} \mathcal{I}(\tau_{p,l-1} \leq x_\star^p < \tau_{p,l}) \right) . \tag{5.14}$$

Prediction of the survival function

Once the scores are calculated, a survival function needs to be estimated. Preferably, one survival curve is estimated for each possible score. However, estimation of this function will only be reliable when enough observations (with events) have the same score. The ICS survival model is therefore used a second time, using the scores as input and the failure times as output. The obtained step function will now denote which scores correspond to the same survival and can therefore be taken together when estimating the survival curves. Subsection 5.3.3 gives an illustration of this approach.

Assume that for each score enough observations are provided in order to obtain an estimate of the survival function. Instead of estimating the survival function, the CDF is estimated by means of monotone LS-SVMs. However, in order to apply this approach for censored data, some preprocessing of the data is required. Starting from a data set $\mathcal{D} = \{(x_i, y_i, \delta_i)\}_{i=1}^n$, an augmented data set $\mathcal{D}_{\text{aug}} = \{\mathcal{D}_i\}_{i=1}^n$ is created. Each data set $\mathcal{D}_i = \{(x_i, y_{i,k})\}_{k=1}^{n_t}$ represents a replication of observation i within n_t consecutive time intervals $k = 1, \ldots, n_t$. The outcome $y_{i,k}$ is zero when the event did not occur before the end of the kth time interval. For events,

Algorithm 2 Iterative reweighted L_1 minimization for survival models

1: Determine the optimal tuning parameters γ and μ, e.g. by 10-fold cross-validation.
2: Solve equation (5.11).
3: **repeat**
4: Let $w_{\text{old}} = w$.
5: Find w, b from

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

$$\text{subject to} \quad \begin{cases} \hat{y}_i = \sum_{p=1}^{d}\left(\sum_{l=1}^{k_p+1} w_{pl}\mathcal{I}(\tau_{p,l-1} \le x_{ip} < \tau_{p,l})\right) + b, \\ \qquad\qquad \forall\, i = 1,\dots,n\,, \\ \hat{y}_i - \hat{y}_{i-1} + \epsilon_i \ge y_i - y_{i-1}, \quad \forall\, i = 2,\dots,n\,, \\ \hat{y}_i \ge y_i - \xi_i, \quad \forall\, i = 1,\dots,n\,, \\ -\delta_i\hat{y}_i \ge -\delta_i y_i - \xi_i^*, \quad \forall\, i = 1,\dots,n\,, \\ \xi_i \ge 0, \quad \forall\, i = 1,\dots,n\,, \\ \xi_i^* \ge 0, \quad \forall\, i = 1,\dots,n\,, \\ \epsilon_i \ge 0, \quad \forall\, i = 1,\dots,n\,. \end{cases} \tag{5.15}$$

6: Let $\chi_{p,l} = f(w_{p,l} - w_{p,l-1}),\ \forall\, p = 1,\dots,d$ and $l = 1,\dots,k_p$.
7: **until** $|w - w_{\text{old}}| < 10^{-8}$

$y_{i,k} = 1$ for all intervals ending after the observed failure time. For censored data, the observations are only replicated within the intervals in which they are known to be at risk. The primal model formulation for the estimation of the CDF then becomes

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

$$\text{subject to} \quad \begin{cases} w^T\varphi(x_{i,k}) + b = y_{i,k} + \epsilon_i, \\ \qquad\qquad \forall\, i = 1,\dots,n; \forall\, k = 1,\dots,n_t\,, \\ w^T(\varphi(x_{i,k}) - \varphi(x_{i,k-1})) \ge 0, \\ \qquad\qquad \forall\, i = 1,\dots,n;\ \forall\, k = 2,\dots,n_t\,, \\ w^T\varphi(x_{i,k}) + b \ge 0, \quad \forall\, i = 1,\dots,n;\ \forall\, k = 1,\dots,n_t\,, \\ -w^T\varphi(x_{i,k}) - b \ge -1, \\ \qquad\qquad \forall\, i = 1,\dots,n;\ \forall\, k = 1,\dots,n_t\,, \end{cases} \tag{5.16}$$

with $x_{i,k}$ the augmented input $x_{i,k} = [x_i^T, k]^T$. In the above formulation x_i is the score as defined in equation (5.14). Building the Lagrangian and solving the KKT equations yields the dual problem

$$
\begin{aligned}
\min_{\alpha,\beta,\eta,\eta^*} \quad & \frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=1}^{n_t}\left(\alpha_{ik}\alpha_{jl}+\eta_{ik}\eta_{jl}+\eta^*_{ik}\eta^*_{jl}\right)\varphi(x_i,k)\varphi(x_j,l) \\
& +\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=2}^{n_t}\beta_{ik}\beta_{jl}\left(\varphi(x_i,k)\right)\left(\varphi(x_j,l)-\varphi(x_j,l-1)\right) \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=2}^{n_t}\beta_{ik}\beta_{jl}\left(\varphi(x_i,k-1)\right)\left(\varphi(x_j,l)-\varphi(x_j,l-1)\right) \\
& +\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=1}^{n_t}\sum_{l=2}^{n_t}\alpha_{ik}\beta_{jl}\varphi(x_i,k)\left(\varphi(x_j,l)-\varphi(x_j,l-1)\right) \\
& +\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=1}^{n_t}\alpha_{ik}\varphi(x_i,k)\left(\eta_{jl}\varphi(x_j,l)-\eta^*_{jl}\varphi(x_j,l)\right) \\
& +\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=2}^{n_t}\sum_{l=1}^{n_t}\beta_{ik}\left(\varphi(x_i,k)\right)\left(\eta_{jl}-\eta^*_{jl}\right)\varphi(x_j,l) \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=2}^{n_t}\sum_{l=1}^{n_t}\beta_{ik}\left(\varphi(x_i,k-1)\right)\left(\eta_{jl}-\eta^*_{jl}\right)\varphi(x_j,l) \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=1}^{n_t}\eta_{ik}\eta^*_{jl}\varphi(x_{ik},k)\varphi(x_{jl},l) \\
& -\sum_{i=1}^{n}\sum_{k=1}^{n_t}\alpha_{ik}y_{ik}+\sum_{i=1}^{n}\sum_{k=1}^{n_t}\eta^*_{ik} \\
\text{subject to} \quad & \begin{cases} \beta_{ik}\geq 0 & \forall\ i=1,\ldots,n;\ \forall\ k=2,\ldots,n_t\,, \\ \eta_{ik}\geq 0 & \forall\ i=1,\ldots,n;\ \forall\ k=1,\ldots,n_t\,, \\ \eta^*_{ik}\geq 0 & \forall\ i=1,\ldots,n;\ \forall\ k=1,\ldots,n_t\,. \end{cases}
\end{aligned}
\tag{5.17}
$$

The predicted probability of surviving until the kth time interval for a new point $x_\star$ is found as

$$
\hat{y}_{\star,k}=\sum_{i=1}^{n}\left[(\alpha_{ik}+\eta_{ik}-\eta^*_{ik})\varphi(x_i,k)+(\varphi(x_i,k)-\varphi(x_i,k-1))\right]\varphi(x_\star,k)\,. \tag{5.18}
$$

5.3 Real life applications

This Section illustrates the proposed method on three real life applications. First two classification problems are tackled. A first illustration concerns the prediction of the malignancy of ovarian tumors (Timmerman *et al.*, 2000). A second

illustration involves the prediction of non-viable pregnancies. In a second Section, the survival of primary operable breast cancer patients treated in the University Hospitals of Leuven is predicted. For all examples, 10-fold cross-validation and coupled simulated annealing (Xavier de Souza *et al.*, 2010) was used to tune the regularization parameters γ and μ.

5.3.1 Diagnosis of malignancy of adnexal masses

Data

The data set used in this experiment is discussed in Appendix A.2. For this experiment, the data obtained in phase 1, 1b and the data obtained in phase 2 by the centers from phase 1b will be used as training data (2512 patients). The data from the new centers in phase 2 will be used as independent test set (997 patients). Since the maximum diameter of the lesion and the maximum diameter of the solid component are strongly correlated, the maximum diameter of the lesion together with the ratio of the diameter of the solid component and the lesion will be used.

Development of the model

The procedure to build the model is illustrated in Algorithm 3. First 10-fold cross-validation with the c-index as model selection criterium is used to find the optimal value of the regularization constant in equation (5.3). Given the outcome of this model, a prediction of the probability of a malignant mass is found by means of a monotonic regression function, linking the output of the model to the observed malignancy. The best level of weighting a in the iteratively reweighted model, is selected by 5-fold cross-validation, using the c-index, R^2_{adj}, the likelihood, the number of necessary steps in the functional forms and the number of selected variables as performance and cost measures. These measures are illustrated as a function of the weighting a in Figure 5.5. By changing the weight level a, a trade-off is made between the number of variables and the performance of the model. It is noted that by inclusion of 9 instead of 25 variables, the performance remains nearly unaffected. From this a weight level $a = 0.5$, corresponding to more or less 9 selected variables is chosen. In order to obtain a more robust set of variables, 100 bootstrap samples from the original training set are taken and the model is trained on each of these. The distribution of the selected variables is illustrated in Figure 5.6. The set of variables which will be further considered include those which are selected in more than half of the bootstrap samples (indicated by the dashed vertical line). This set includes: the age of the patient, the maximal diameter of the lesion, the ratio of the maximal diameter of the solid component and the maximal diameter of the lesion, the number of papillary projections, whether there is blood flow present in the papillary structures, whether the internal cyst wall is irregular,

whether there are acoustic shadows present, the color score of the intratumoral blood flow, the presence of ascites and the amount of free fluid in the pouch of Douglas. Using this selected set of variables, the optimal tuning parameters were again tuned by means of cross-validation.

Results

The resulting model is visualized in Figure 5.7 (color bar representation) and summarized in Table 5.1 (table based representation). Looking at Figure 5.7, the clinician knows which questions need to be answered in order to obtain an estimate of the probability of a malignancy. A first question would be: What is the age of the patient? Depending on whether the patient is younger than 40, between 40 and 60 or older than 60, the patient receives zero, one or three points, respectively. A second question involves the ratio of the diameter of the solid component and the diameter of the lesion. Depending on whether the ratio is less than 0.25, between 0.25 and 0.55 or larger than 0.55, zero, four or seven points are given, respectively. The clinician needs to obtain points corresponding to all questions illustrated in Figure 5.7. For a good visual interpretation, the number of points is encoded in a color scheme, where dark red indicates a high and a blue color indicates a low probability of malignancy. The final score is found by adding the points corresponding to all answers. Figure 5.8 illustrates the predicted probability of a malignant tumor in function of the score. The AUC of the training set is 0.95 (95% CI=0.94-0.95). The AUC of the test set is 0.96 (95% CI=0.94-0.97). The R^2_{adj} equals 0.72 on the training and 0.67 on the test set. The model underestimates the probability of malignancy on the test set (Figure 5.9). The ratio of average predicted probability to the observed probability equals 0.78 in the test set, confirming the underestimation of the probability of a malignant tumor. A decision cut-off was taken to obtain the best specificity for a sensitivity of at least 90%. This resulted in predicting a malignant tumor for a score of 10 or more. This cut-off resulted in a sensitivity and specificity of respectively 94% and 76% on the training set and of 93% and 86% on the test set. The positive and negative likelihood ratio on the test set were 6.53 and 0.078, respectively, corresponding to a diagnostic odds ratio of 84. The complete ROC curve is illustrated in Figure 5.10.

Comparison with other models

The data set containing the patients from the first phase (see Appendix A.2) has been used before to develop logistic regression models (Timmerman *et al.*, 2005) (LR1 with 12 variables and a simplified version: LR2), Bayesian kernel-based methods (Van Calster *et al.*, 2007a) (LS-SVM with a linear, RBF and additive RBF kernel; and relevance vector machines with a linear, RBF and additive RBF kernel),

Algorithm 3 Procedure for model building.

1: Determine the optimal tuning parameter γ by 10-fold cross-validation for the unweighted model including all variables:

$$\min_{w,b,\epsilon} \quad \sum_{p=1}^{d}\sum_{l=1}^{k_p+1}|w_{p,l}-w_{p,l-1}| + \gamma\sum_{i=1}^{n}\epsilon_i$$
$$\text{subject to} \quad \begin{cases} y_i\left(\sum_{p=1}^{d}\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}(\tau_{p,l-1}\le x_i^p\le\tau_{p,l})+b\right)\ge 1-\epsilon_i, \\ \qquad\qquad \forall\ i=1,\ldots,n\,, \\ \epsilon_i\ge 0, \qquad\qquad \forall\ i=1,\ldots,n\,. \end{cases}$$

2: Determine the optimal tuning parameter for the monotonic least squares support vector regression model as discussed in Section 5.2.1.

3: Determine the optimal weighting in order to obtain a high performance with as less variables as possible as follows.

4: **for** $a=[a_{\min},a_{\max}]$ **do**

5: Let $\chi_{p,l}=1,\forall\ p=1,\ldots,d; l=1,\ldots,k_p$, w=NULL;

6: **repeat**

7: Put $w_{\text{old}}=w$

8: Find w,b from

$$\min_{w,b,\epsilon} \quad \sum_{p=1}^{d}\sum_{l=1}^{k_p+1}\chi_{p,l}|w_{p,l}-w_{p,l-1}| + \gamma\sum_{i=1}^{n}\epsilon_i$$
$$\text{subject to} \quad \begin{cases} y_i\left(\sum_{p=1}^{d}\sum_{l=1}^{k_p+1} w_{p,l}\mathcal{I}(\tau_{p,l-1}\le x_i^p\le\tau_{p,l})+b\right)\ge 1-\epsilon_i, \\ \qquad\qquad \forall\ i=1,\ldots,n\,, \\ \epsilon_i\ge 0, \qquad\qquad \forall\ i=1,\ldots,n\,. \end{cases}$$

9: Let $\chi_{p,l}=\frac{1}{0.0005+a|w_{p,l}-w_{p,l-1}|},\ \forall\ p=1,\ldots,d$ and $l=1,\ldots,k_p+1$.

10: **until** $|w-w_{\text{old}}|<10^{-8}$

11: Calculate performance (concordance index, R^2_{adj}, likelihood) on 5-fold cross-validation

12: **end for**

13: Select the level of weighting making a trade-off between a reasonable number of variables and an acceptable performance.

14: Make B=100 bootstrap samples selecting $\pm\ n_s$ variables to get a robust selection of n_s variables.

15: Determine the optimal regularization parameter for a model with the selected variables and train the model. Round the obtained coefficients.

16: Determine the optimal regularization constant for the monotone least-squares support vector regression and train the model to obtain a prediction of the probability of the event in function of the score.

17: Test the model.

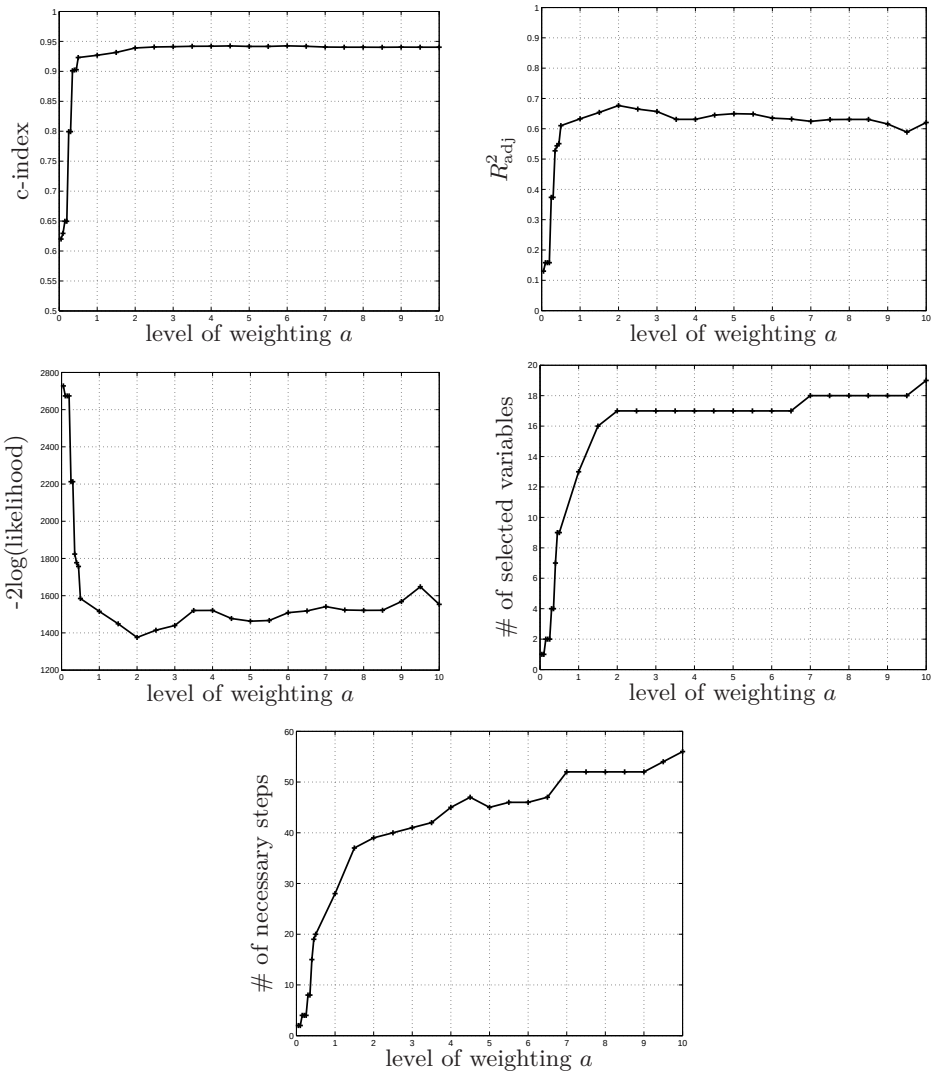


Figure 5.5: *Performance and number of variables and number of steps as a function of the level of weighting. When decreasing the level of weighting a, the number of selected variables drops and so does the number of steps in the functional forms of the variables. Initially, the performance is not affected by the weighting. When applying an even lower weight level, less and less variables are selected. From a weight level of 2 all performance measures start to drop. From a weight level of 0.5 on, a large performance reduction is noted. It is noted that by inclusion of 9 instead of 25 variables, the performance remains nearly unaffected.*

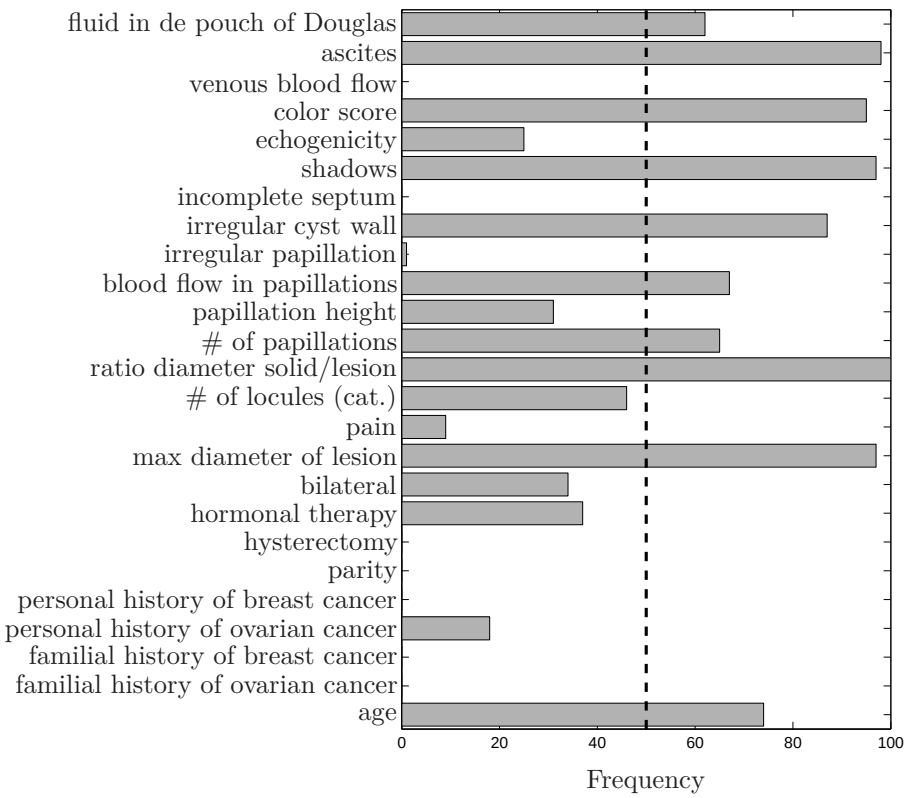


Figure 5.6: *Distribution of the selected variables over 100 bootstrap samples of the training set, using a weight level of 0.5. The dashed vertical line illustrates the variables which are selected in more than half of the bootstrap samples.*

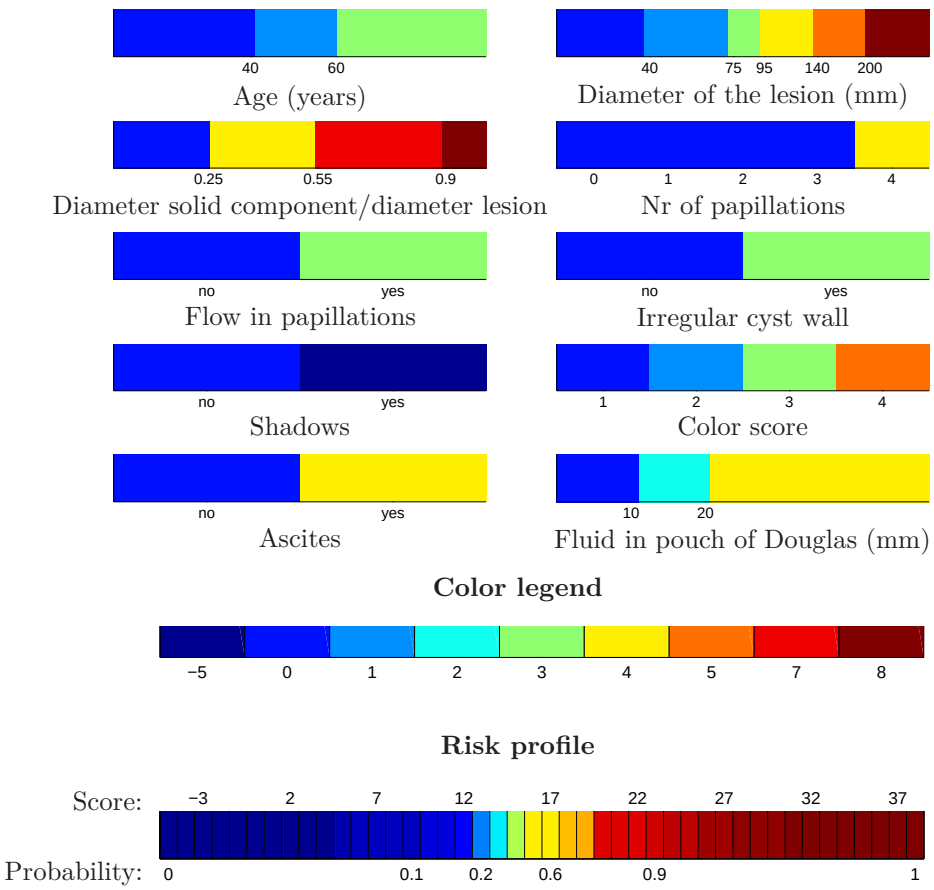


Figure 5.7: *Color based representation of the* ICS *index applied to an ovarian cancer diagnostic study. Given this chart, the clinician knows which questions need to be answered in order to obtain a prediction of the probability of malignancy. A patient aged 40 to 60 years receives 1 point; younger or older patients receive 0 and 3 points, respectively. A lesion of less than 40 mm, between 40 and 75 mm, 75 and 95, 95 and 140, 140 and 200 and larger than 200 results in 0, 1, 3, 4, 5 and 8 points, respectively. For each of the represented bars, the corresponding points are obtained. The total score is obtained by summing all points. The color bar at the bottom represents the predicted probability associated with the final score.*

Table 5.1: *Table based representation of the scoring system for* `IOTA`*. If the answer on the question is yes, the points at the right of the question, need to be added to the score. The bottom part of the table is used to link the obtained score with an estimate of the probability of a malignant mass (prob).*

variable	question	# points
Age		
	$40 \leq$ age < 60	1
	age ≥ 60	3
Maximal diameter of the lesion		
	$40 \leq$ max diameter of the lesion < 75	1
	$75 \leq$ max diameter of the lesion < 95	3
	$95 \leq$ max diameter of the lesion < 140	4
	$140 \leq$ max diameter of the lesion < 200	5
	$200 \leq$ max diameter of the lesion	8
Ratio of the diameter of the solid component and the lesion		
	$0.25 \leq$ ratio solid component/lesion < 0.55	4
	$0.55 \leq$ ratio solid component/lesion < 0.90	7
	$0.90 \leq$ ratio solid component/lesion	8
Number of papillations		
	number of papillations equal to 4	4
Blood flow in papillations		
	presence of papillary blood flow	3
Internal cyst wall		
	irregular internal cyst wall	3
Acoustic shadows		
	presence of acoustic shadows	-5
Color score		
	colorscore equal to 2	1
	colorscore equal to 3	3
	colorscore equal to 4	5
Ascites		
	presence of ascites	4
Free fluid in the pouch of Douglas		
	$10 \leq$ amount of free fluid < 20	2
	$20 \leq$ amount of free fluid	4

score	prob	score	prob	score	prob
≤ 1	< 0.001	13	0.24	24 to 25	0.93
2 to 4	0.01	14	0.36	26 to 27	0.96
5	0.04	15	0.54	28 to 30	0.98
6 to 9	0.06	16 to 17	0.63	≥ 31	> 0.99
10 to 11	0.10	18 to 19	0.69		
12	0.12	20 to 23	0.91		

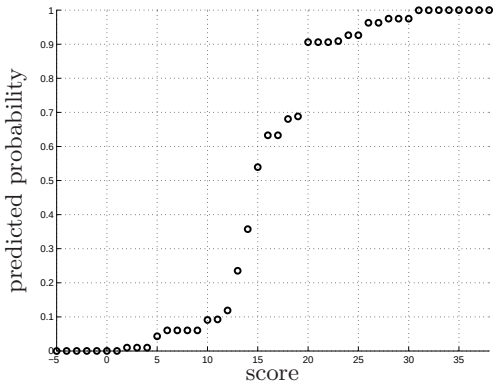


Figure 5.8: *Predicted probability as a function of the final score for the prediction of the malignancy of adnexal masses. Up to a score of 12, the predicted probability of a malignant tumor is less than 10%. A score of 20 or higher results in a predicted probability of a malignant tumor of nearly 90% or higher.*

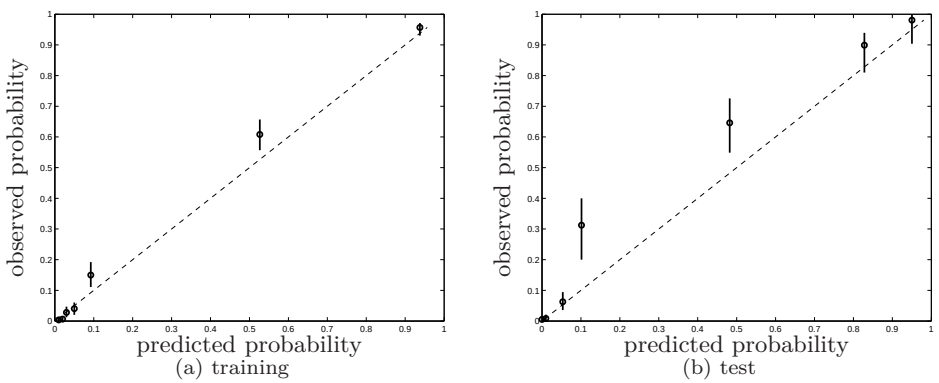


Figure 5.9: *Calibration on the training (a) and test (b) set for the prediction of the malignancy of adnexal masses. For each possible value of the predicted probability (some values were taken together in order to obtain at least 10% of the patients in each group), the observed probability of malignancies is calculated (dots). A 95% confidence interval on the percentage of the observed malignancies is illustrated by means of the vertical lines. The model underestimates the probability of malignancy on the test set.*

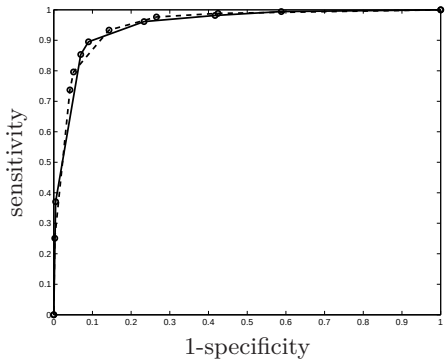


Figure 5.10: ROC *curve of the* ICS *model for* `IOTA` *on training (solid line) and test (dashed line) set.*

Bayesian neural networks with ARD (Van Calster *et al.*, 2007b) (MLPa, MPLb and Perc), simple rules (Timmerman *et al.*, 2008) and a scoring system (Ameye *et al.*, 2009). These models were temporally validated on the data from phase 1b (Van Holsbeke *et al.*, 2009) and/or on the patients from the second phase (Timmerman *et al.*, 2010b). External validation was performed on the same test set as used in this experiment (Timmerman *et al.*, 2010b; Van Holsbeke *et al.*, 2010b; Timmerman *et al.*, 2010a). Another model which is often used in clinical practice is the risk of malignancy index as proposed by Jacobs *et al.* (1990). The variables included in these models are compared with the variables selected by the ICS model (see Table 5.2). Except for the RMI, there is a large overlap between the selected variables in the different models. Only the amount of free fluid, which is a variable in the proposed model, was never used before. Table 5.3 summarizes the AUC of these models on the test set of 997 patients from centers which were not used to build any of the models. Additionally, the sensitivity, specificity, LR^+, LR^- and DOR at the cut-off level as proposed in the original manuscripts are calculated. The proposed method performs as well as the best models described in the literature. However, note that the model is trained on 2509 patients whereas the other models are trained only on the data of the first phase (1066 patients) or fewer. The model utilizes less variables than most of the other models. In contrast to the neural networks and LS-SVM and RVM models, the proposed model is highly interpretable and applicable.

Table 5.2: *Comparison of the variables included in several models for the prediction of the malignancy of adnexal masses found in the literature and the* ICS *model built in this Chapter. Three different kernel functions are used:* [1]*, linear;* [2] RBF*;* [3]*, additive* RBF.

variable	RMI	LR1	LR2	MLPa	MLPb	Perc	LS-SVM[1]	LS-SVM[2]	LS-SVM[3]	rvm[1]	rvm[2]	rvm[3]	ICS
Locularity	✓												
Evidence of solid component	✓												
Evidence of metastatses	✓												
Bilateral Lesions	✓												
Menopausal status	✓												
CA 125	✓												
Ascites	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Personal history of ovarian cancer		✓		✓	✓	✓	✓	✓	✓	✓	✓		
Hormonal therapy		✓		✓			✓	✓	✓	✓	✓	✓	
Maximal diameter of the lesion		✓		✓	✓	✓							✓
Pain during examination		✓			✓								
Maximal diameter of the solid component		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nr of papillations				✓									✓
Blood flow within papillary structures		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Irregular internal cyst wall		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Shadows		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Echogenicity		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Color score		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓
Amount of free fluid													✓
Suspected to be of ovarian origin							✓	✓	✓	✓	✓	✓	
Maximal diameter of ovary							✓	✓	✓	✓	✓	✓	
total number of variables	7	12	6	11	10	10	12	12	12	12	12	11	10

Table 5.3: *Summary of performance measure for previously built models for the prediction of the malignancy of adnexal masses and the* ICS *model built in this Chapter. The measures are calculated on the test set of 997 patients from external centers. Except for the* AUC*, all measures are calculated at the cut-off indicated in the original paper.*

model	AUC	95% CI on AUC	Sensitivity	Specificity	LR^+	LR^-	DOR	95% CI on DOR
RMI*	0.911	0.880-0.935	67.5	**94.6**	**12.6**	0.343	37	24-57
LR1*	0.956	0.939-0.968	92.2	86.5	6.84	0.091	75	46-125
LR2*	0.949	0.930-0.963	91.8	85.6	6.36	0.096	66	40-108
MLPa*	0.942	0.924-0.956	86.3	87.9	7.11	0.156	46	30-69
MLPb*	0.952	0.936-0.965	91.0	86.5	6.75	0.104	65	40-104
Perc*	0.953	0.935-0.965	90.6	88.9	8.20	0.106	77	48-125
LS-SVM linear*	0.950	0.932-0.964	87.1	91.2	9.94	0.142	70	45-109
LS-SVM RBF*	0.954	0.935-0.967	89.4	89.9	8.85	0.118	75	47-120
LS-SVM RBF_{add}*	0.941	0.920-0.956	87.1	88.7	7.83	0.128	61	39-96
RVM linear*	0.949	0.931-0.962	90.6	88.7	8.00	0.106	75	47-122
RVM RBF*	0.951	0.933-0.965	90.6	87.7	7.39	0.107	69	43-111
RVM RBF_{add}*	0.942	0.921-0.957	88.6	88.7	7.83	0.128	61	39-96
ICS	**0.958**	0.943-0.969	**93.3**	85.7	6.53	**0.078**	**84**	51-150

* Taken from Van Holsbeke *et al.* (2010b)

5.3.2 Prediction of non-viable pregnancies

Data

This experiment uses data on 1435 women with an early pregnancy presented at a single ultrasound unit. The data set is described in detail in Appendix A.3. The model is trained on 955 patients and tested on the remaining 480 patients.

Development of the model

The procedure of model building is the same as presented in Algorithm 3. However, since it is expected that the effect of the diameter of the gestational sac differs for fetuses with or without a visualized heart beat, an interaction effect was manually added as an additional variable.

Results

The ICS model for the prediction of non-viable pregnancies is illustrated in Figure 5.11 and Table 5.4. The use of the model is similar as in the previous application, except for the interaction effect. In the presence of a fetal heart beat, the color bar in the fourth row of the Figure should be used. In absence of a heart beat, the color bar in the third row should be used. Comparing both color bars, it is seen that the effect of a large gestational sac diameter varies largely depending on whether a fetal heart beat was seen or not. For a fetus with a heart beat, a large gestational sac diameter is a good sign and the probability of a non-viable pregnancy decreases with an increasing gestational sac diameter. Without a fetal heart beat, a small gestational sac is better. In this case, it might be too early to observe a heart beat. In contrast, when the gestational sac is already large, the pregnancy is older and the absence of a heart beat is more likely to correspond to miscarriage or a non-viable pregnancy. Figure 5.12 illustrates the estimated probability of a non-viable pregnancy as a function of the obtained score. The AUC on the training set is 0.90 (95% CI=0.88-0.92) and on the test set 0.92 (95% CI=0.90-0.94) (see Figure 5.14). The R^2_{adj} equals 0.64 and 0.62 on training and test set, respectively. The model has a ratio of average predicted probability to the observed probability of 0.98 on the test set and is thus well calibrated (see Figure 5.13). Little research (Falco *et al.*, 1996; Choong *et al.*, 2003) has been done with respect to prediction of pregnancy viability. However, comparison with results reported in these studies is difficult since the study population (ultimately viable embryo or fetus) is different from ours.

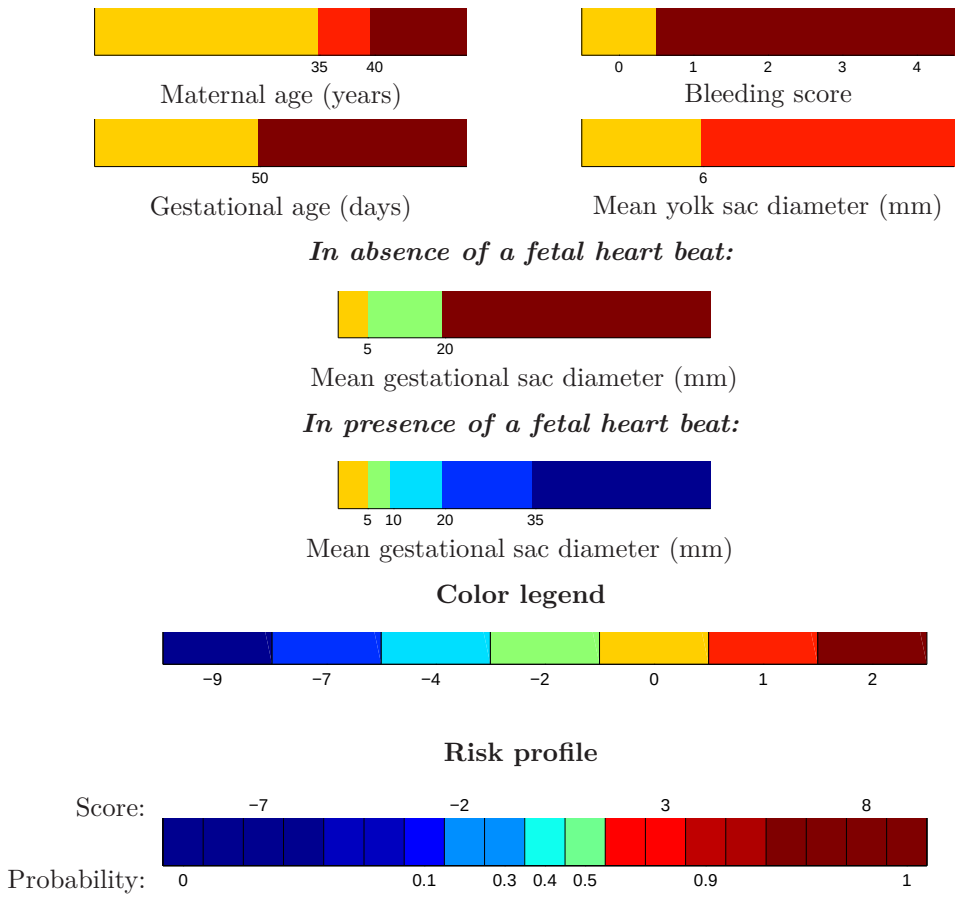


Figure 5.11: *Color based representation of the* ICS *model applied on* EPD. *Given this chart, the clinician knows which questions need to be answered in order to obtain a prediction of the probability of a non-viable pregnancy. The color bar at the bottom represents the predicted probability associated with the final score, obtained by summing the contributions of all variables.*

Table 5.4: *Construction of the scoring system for the diagnosis of a non-viable pregnancy at the end of the first trimester. If the answer to the question is yes, the points at the right are added to the score. The bottom part of the table is used to link the obtained score with an estimate of the probability of a non-viable pregnancy (prob).*

variable	question	# points
Age (years)		
	$35 \leq$ age < 40	1
	$40 \leq$ age	2
Bleeding score		
	bleeding score ≥ 1	2
Gestational days		
	$50 \leq$ gestational dates	2
Mean yolk sac diameter (mm)		
	$6 \leq$ mean yolk sac diameter	1
In absence of a fetal heart rate:		
Mean gestational sac diameter (mm)		
	$5 \leq$ mean gestational sac diameter < 20	-2
	$20 \leq$ mean gestational sac diameter	2
In presence of a fetal heart rate:		
Mean gestational sac diameter (mm)		
	$5 \leq$ mean gestational sac diameter < 20	-2
	$10 \leq$ mean gestational sac diameter < 20	-4
	$20 \leq$ mean gestational sac diameter < 35	-7
	$35 \leq$ mean gestational sac diameter	-9

score	prob	score	prob	score	prob
$\leq$ -6	< 0.001	-2 to -1	0.25	4	0.93
-5	0.05	0	0.38	5	0.94
-4	0.06	1	0.47	≥ 6	> 0.99
-3	0.11	2 to 3	0.87		

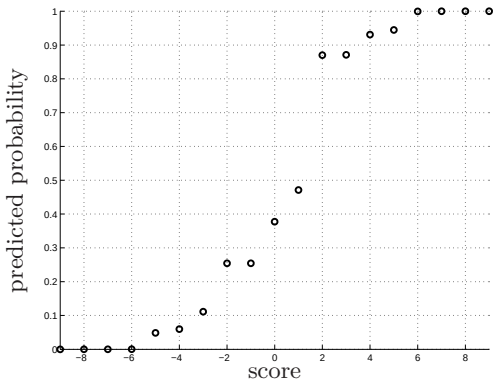
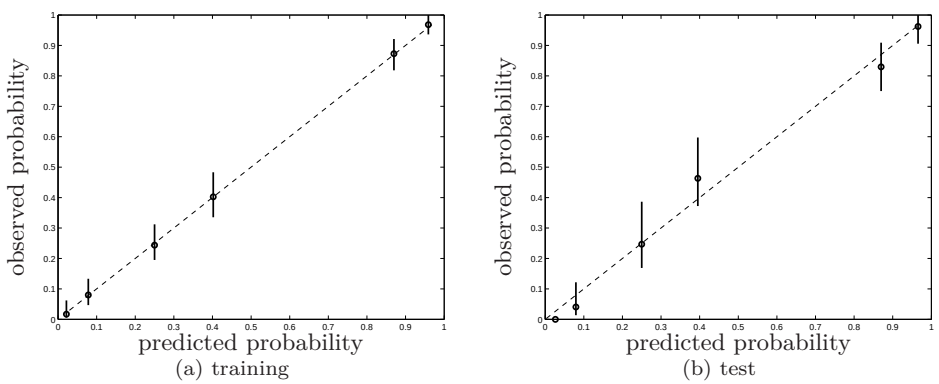


Figure 5.12: *Predicted probability of a non-viable pregnancy as a function of the final score. From a score of 2, the predicted probability of an non-viable pregnancy is higher than 85%. A score of -3 or lower results in a predicted probability of less than 15%.*



(a) training (b) test

Figure 5.13: *Calibration on the test set for the prediction of non-viable pregnancies. For each possible value of the predicted probability (some values were taken together in order to obtain at least 10% of the patients in each group), the observed percentage of non-viable pregnancies is calculated (dots). A 95% confidence interval on the percentage of the observed non-viable pregnancies is illustrated by means of the vertical lines. The model is well calibrated.*

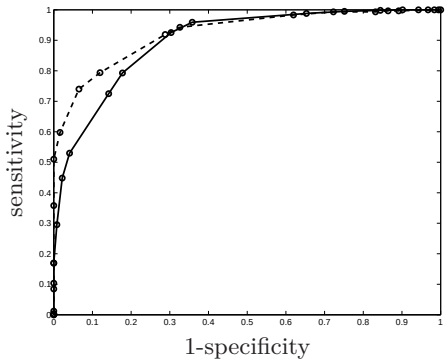


Figure 5.14: ROC *curve of the* ICS *model on* EPD *on training (solid line) and test (dashed line) set.*

5.3.3 Breast cancer prognosis

Data

A last experiment involves the prognosis of primary operable breast cancer patients. The ICS model is trained on the breast cancer data set of the University Hospitals Leuven described in Appendix A.1. Only patients with complete information for age at diagnosis, tumor size, number of positive lymph nodes (LNPOS), PR, HER2 and grade were used. Since the value of ER and PR are highly correlated, ER was not considered in this analysis. The scoring system is then validated on two external validation sets, one from the Breast Cancer Micrometastasis group in Oslo (Norway) (OBC) and another from the Auckland Breast Cancer Registry (New Zealand) (ABC). More information on these data sets can be found in Appendix A.1. After inspection of the training data, the variable size and LNPOS were truncated. Since only a small part of the tumors are larger than 100 mm, larger tumor sizes were restricted to 100 mm. The number of positive lymph nodes was restricted to 20. Figure 5.15 illustrates the Kaplan-Meier survival curves for the three sets. Since patients in the Leuven cohort (solid line) have a better survival than patients in the test cohorts, it can not be expected that the predicted survival curves will calibrate well on the test sets without updating the model (Steyerberg, 2009).

Model building

The model building procedure is similar to the one in the previous Subsection. After training the model to obtain a prognostic index, the survival function is estimated using the procedure elaborated in Section 5.2.2. However, due to memory problems, the survival curves were estimated in groups of patients, based

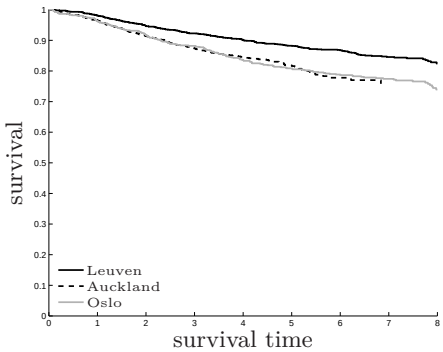


Figure 5.15: *Kaplan-Meier curves for the Leuven training set (black solid line) and the Auckland (dashed line) and Oslo test cohorts (grey solid line). The survival for both test cohorts is lower than in the training cohort. Therefore, it can not be expected that the predicted survival chances will calibrate well on the test sets.*

on the ICS index value. Simultaneous estimation of the survival curves for all patients for long follow-up or large datasets, will be developed in future work (see Chapter 6).

Results

The obtained model is illustrated in Figure 5.16 and Table 5.5. The model uses the number of positive lymph nodes, PR, HER2 and the grade of the tumor. In order to estimate survival functions, the patient groups need to be sufficiently large. An additional problem is that it is not known how many groups with different survival properties should be taken. The ICS model is therefore used with the ICS index as a single variable to obtain the best cut-off values to define risk groups (see also Section 5.2.2). Five different risk groups are recognized. The predicted survival for the different risk groups are illustrated in Figure 5.17 and in the risk profile bars in Figure 5.16. The predicted survival curves closely correspond with the observed survival curves, calculated by means of the Kaplan-Meier estimator. Figure 5.17 also illustrates that an increasing ICS index corresponds to an increasing survival time.

Figure 5.18 illustrates the calibration of the ICS index at (a) 2 and (b) 5 years of follow-up. The model is well calibrated on the training set (`LBC`) and on `ABC`, with exception of the third risk group, for which survival is overestimated at 5 years of follow-up. For the third set (`OBC`) the survival is overall overestimated.

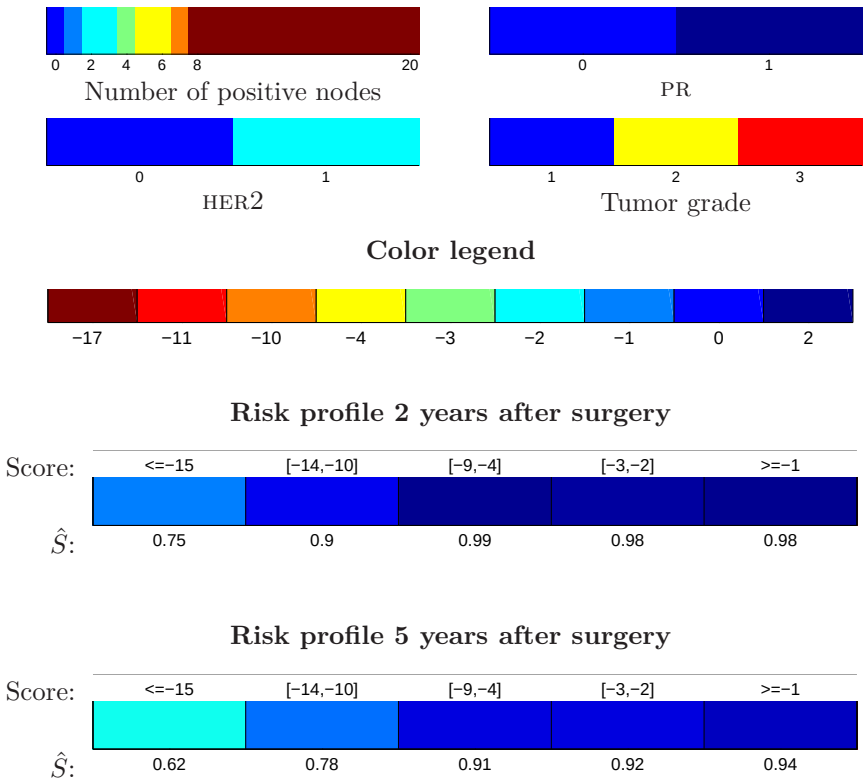


Figure 5.16: *Visualization of the model on* `LBC`*. Given this chart, the clinician knows which variables need to be collected in order to obtain an estimate of the patient's prognosis. For each of the represented bars, the corresponding points need to be calculated. The total score is obtained by addition of all these points. Patients of whom the characteristics fall within the dark blue areas have a good prognosis. Patients of whom the characteristics fall within the light blue regions, have a bad prognosis.*

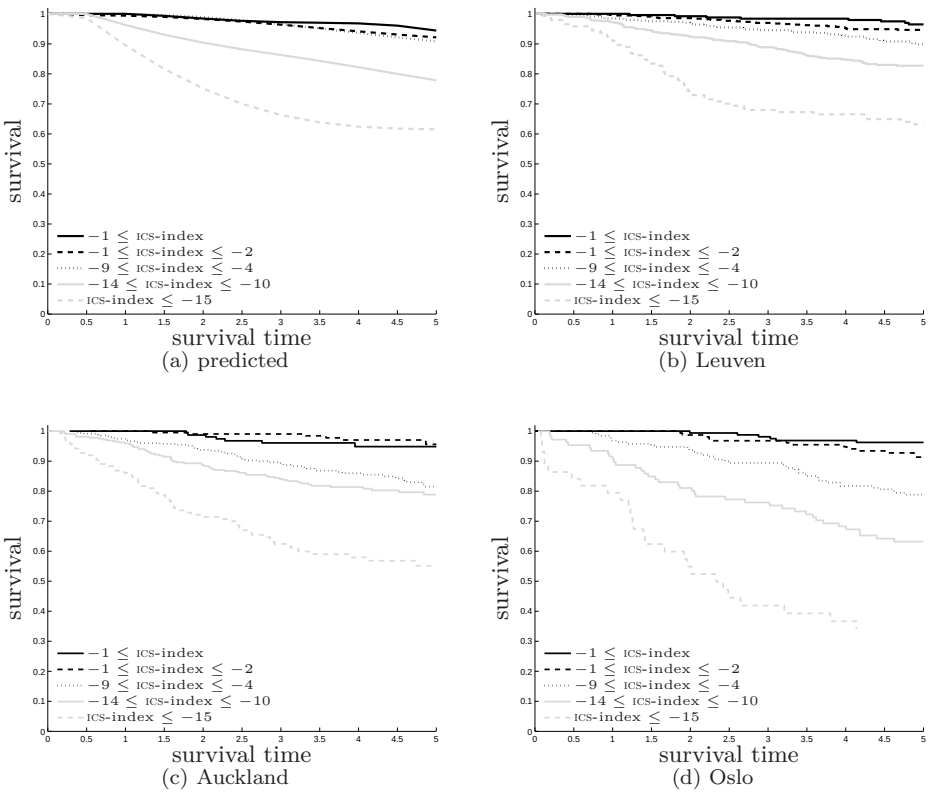


Figure 5.17: *Predicted and observed survival curves for primary operable breast cancer patients. (a) Estimated survival for the different possible scores. For an increasing score, the predicted survival increases. The Kaplan-Meier survival curves for* `LBC` *(b),* `ABC` *(c) and* `OBC` *(d) are shown as well. An increasing score corresponds to a better observed prognosis.*

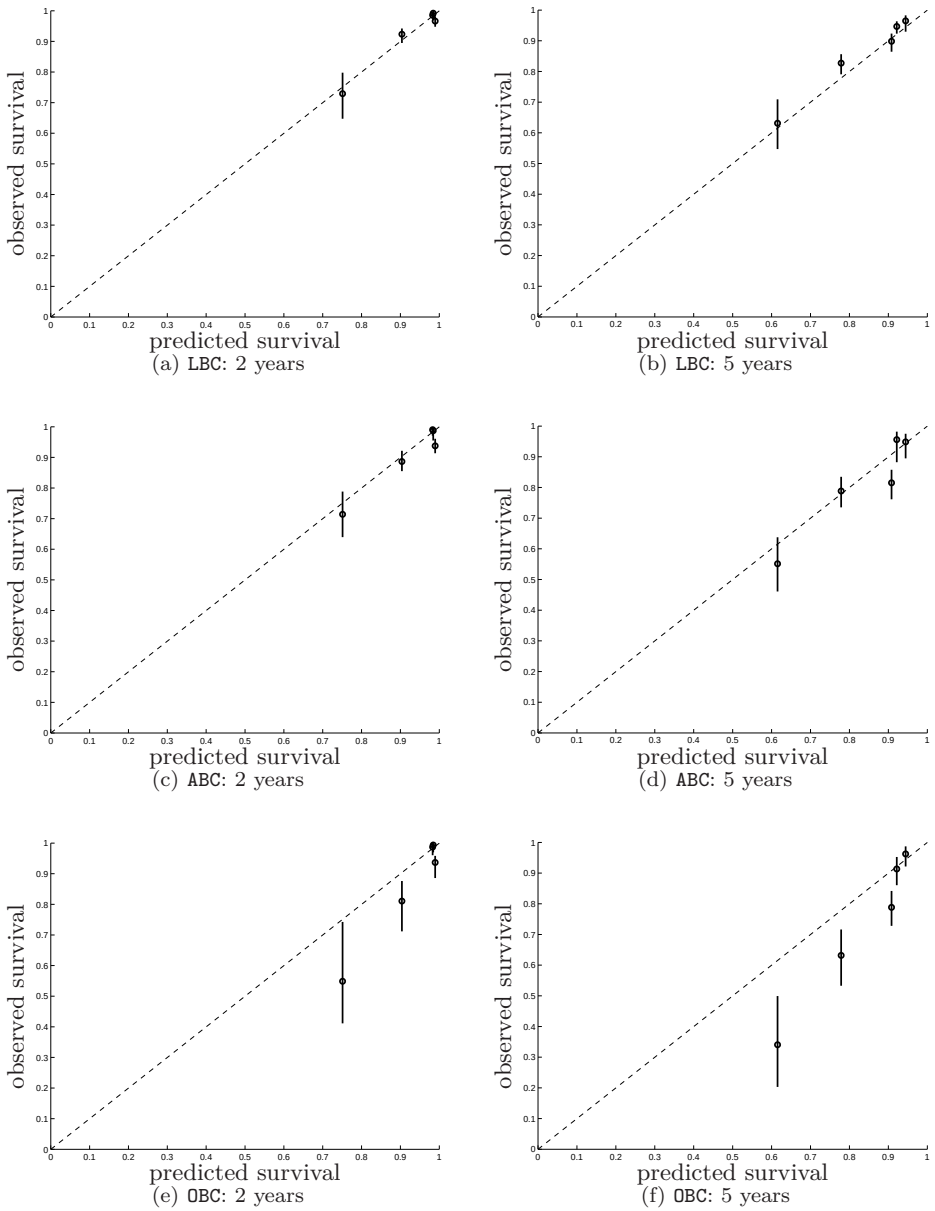


Figure 5.18: *Calibration of the* ICS *index at 2 and 5 years of follow-up. The model is well calibrated on the training set (`LBC`) and on `ABC`, with exception of the third risk group, for which survival is overestimated at 5 years of follow-up. For the third set (`OBC`) the survival is overall overestimated.*

Table 5.5: *Table based representation of the scoring system for the prognosis of primary operable breast cancer patients. If the answer on the question is yes, the points at the right of the question, need to be added to the score. The estimated survival probability is given at 2 and at 5 years after surgery (prob).*

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

risk profile 2 years after surgery

score	prob	score	prob	score	prob
≤ -15	0.75	-9 to -4	0.99	≥ -1	0.98
-14 to -10	0.90	-3 to -2	0.98		

risk profile 5 years after surgery

score	prob	score	prob	score	prob
≤ -15	0.62	-9 to -4	0.91	≥ -1	0.94
-14 to -10	0.78	-3 to -2	0.92		

Use of the ICS model to define risk groups in existing prognostic models

As stated before, clinicians like to divide patients into risk groups, preferably a low, intermediate and high risk group. Therefore, the NPI and iNPI models are defined together with two cut-offs. However, there is no theoretical foundation for the choice of three groups. The methods able to define optimal cut-offs which are found in the literature need the number of preferred clusters or are unsupervised (e.g. clustering). The proposed ICS model is able to automatically find the number of different risk groups together with the best cut-offs to define these groups.

Application of this approach on the NPI yields 3 risk groups with cut-offs at 2.6 and 4.4. The resulting groups on the iNPI are defined by the cut-offs 3.5, 4.3 and 6.3. The corresponding risk profiles are given in Figure 5.19. The NPI is the least

attractive model. It can only divide patients into three risk groups, for which the predicted survival at two and five years differ less than for both other models. The iNPI obtains the largest estimated survival difference between the most extreme risk groups. Figures 5.20 and 5.21 illustrate the calibration of the NPI and iNPI. The NPI does not achieve a wide spread in predicted nor observed survival within the risk groups. Figure 5.22 shows the Kaplan-Meier survival curves according to the risk groups defined by NPI, iNPI and ICS. The NPI is able to define a very good prognostic group, but does not find a risk group with a very poor prognosis. The iNPI on the contrary does find a very poor prognostic group, but fails to find a very good prognostic group. The ICS model finds both a very good and a very bad prognostic group.

Comparison with other models

In order to be able to compare the ICS model with the NPI and iNPI as they were proposed, the ICS model is divided into tree different risk groups. The optimal thresholds were defined as the pair with the highest c-index obtained with the 0.632+ bootstrap method (Efron & Tibshirani, 1997) on 1000 bootstrap samples. As a result, patients with an ICS value less than -14 are classified into the high risk group, a value from -13 to -10 corresponds to the intermediate risk group and a value of -9 or higher corresponds to the low risk group.

Table 5.6 summarizes the results of the NPI, iNPI and of the ICS-index on the training set and both test sets. Considering the c-index, no difference between the models is seen, except between NPI and both other models on the training set. From a clinical point of view, the performance of the model as a classification tool to divide patients into risk groups, is more important than the concordance on the continuous model. Therefore, risk stratification measures are also assessed (Janes *et al.*, 2008). The models are evaluated by means of the difference in event rate in the low and high risk group (EvR difference) and by the percentage of patients which are classified into the high and low risk group (EXT%). The iNPI classifies a significantly higher percentage of patients into the most extreme risk groups than the NPI. The EXT% of the ICS index is significantly higher than for both other models. No significant differences in EvRs were noted. However, due to the merging of three risk groups of the ICS index, the survival within this group was artificially reduced (see Figure 5.23).

Figure 5.23 illustrates the survival curves according to the defined risk groups. For all of the three models, the survival curves for the different risk groups are significantly different. Due to the fact that the iNPI and the ICS classify more patients into the low risk group, the corresponding survival is lower than for the NPI low risk group. Due to the merging of three separate ICS risk groups into one single low risk group, this group is large (see Table 5.7) and the corresponding

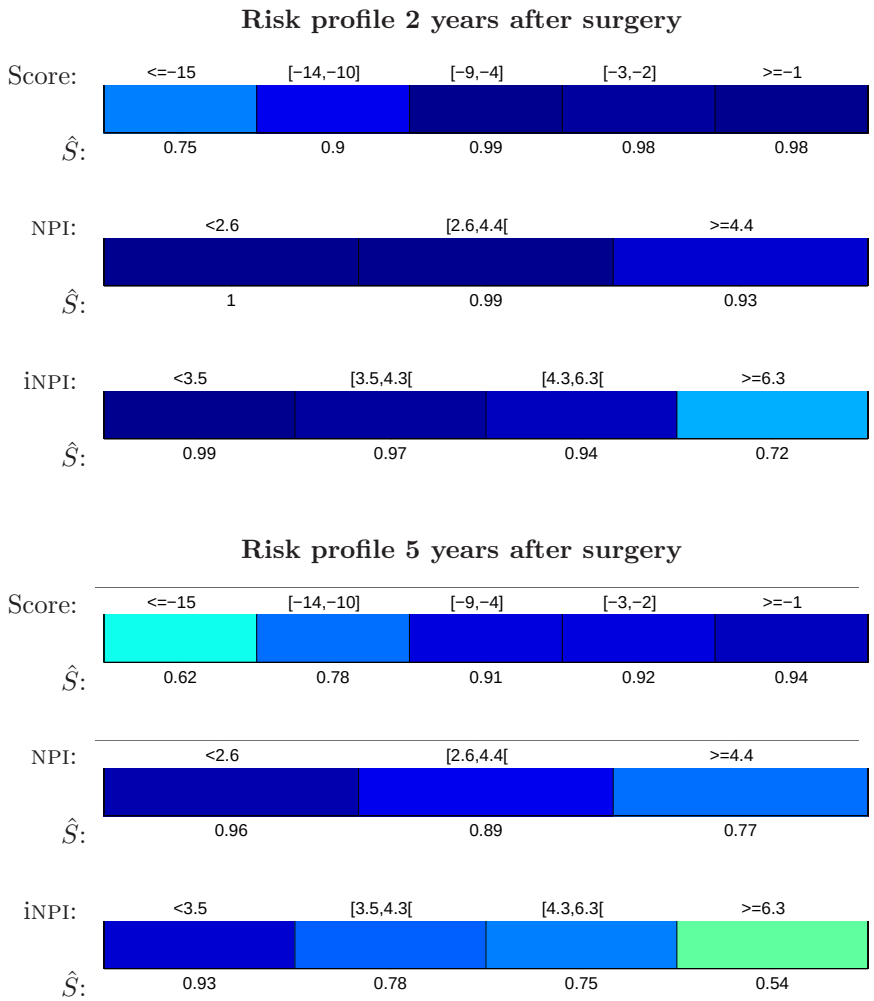


Figure 5.19: *Summary of the risk profiles according to the* ICS *model for three prognostic models: the points system (*ICS*), the* NPI *and the i*NPI*. Within the* NPI *only three different risk groups were observed. The spread in survival between the most extreme risk groups is lower than in the other two models. For the i*NPI *four different risk groups were found, with a large spread in predicted survival between the different risk groups.*

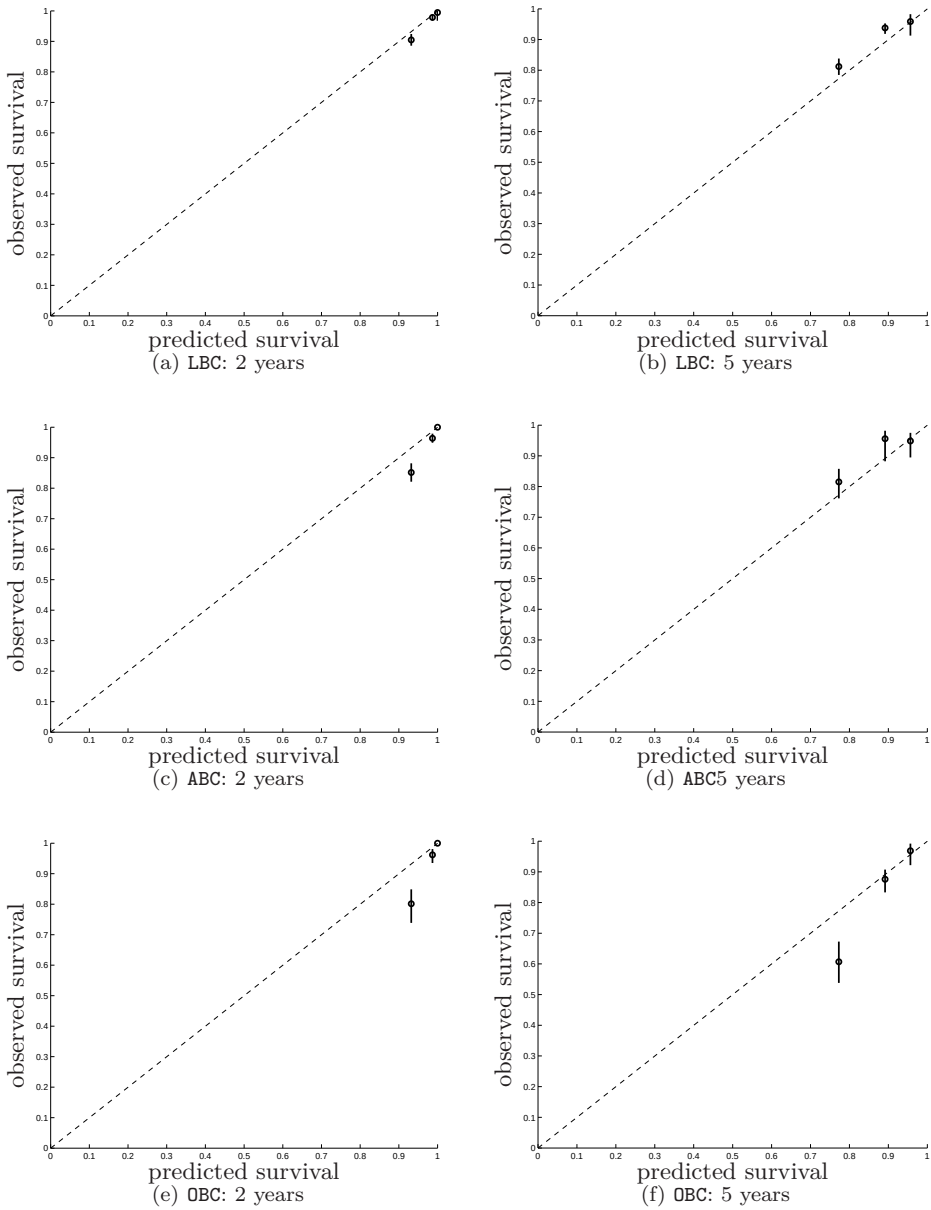


Figure 5.20: *Calibration of the* NPI *at 2 and 5 years of follow-up. The model is well calibrated on the training set but performs worse on both test sets.*

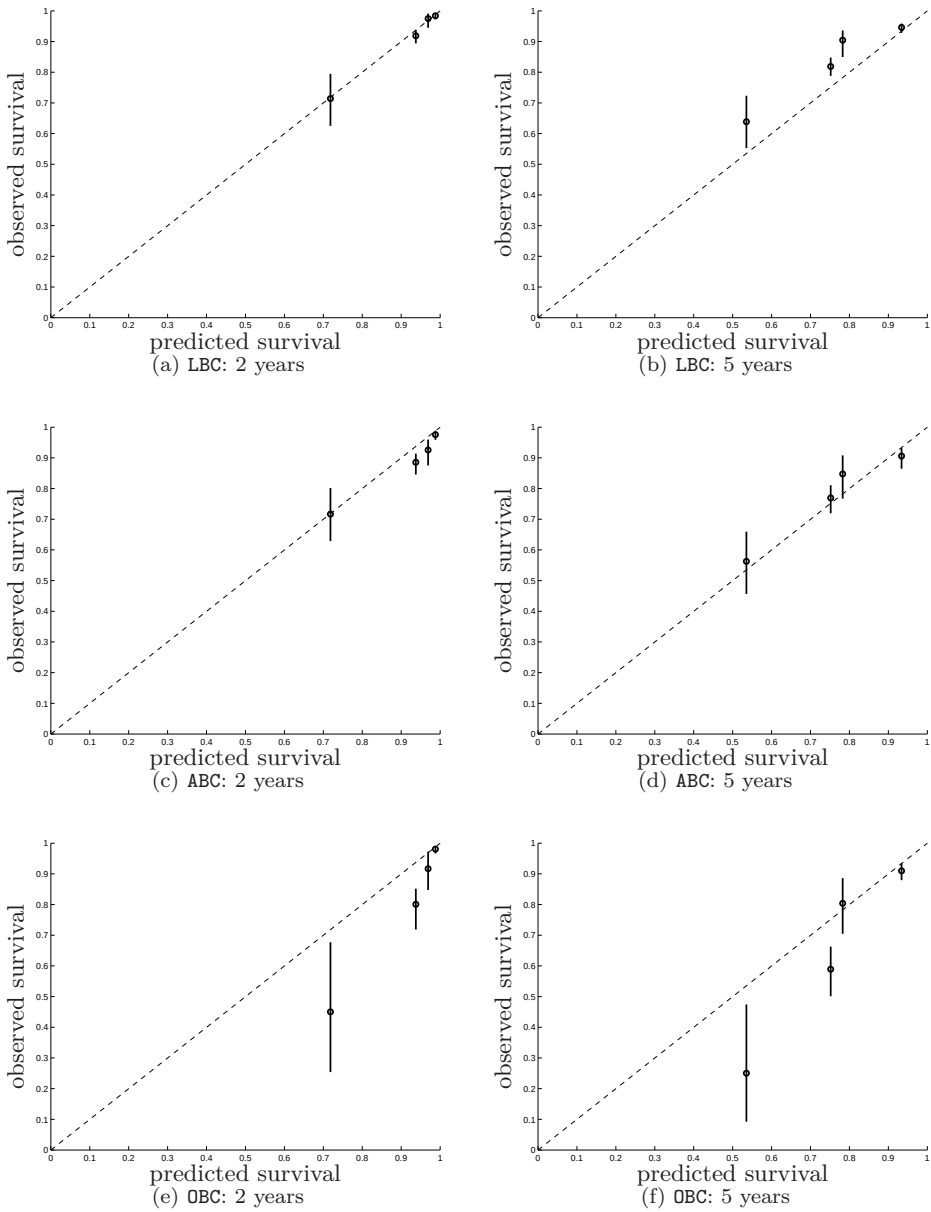


Figure 5.21: *Calibration of the i*NPI *at 2 and 5 years of follow-up. The model is well calibrated on* `LBC` *and* `ABC`*. The survival is overestimated for* `OBC`*.*

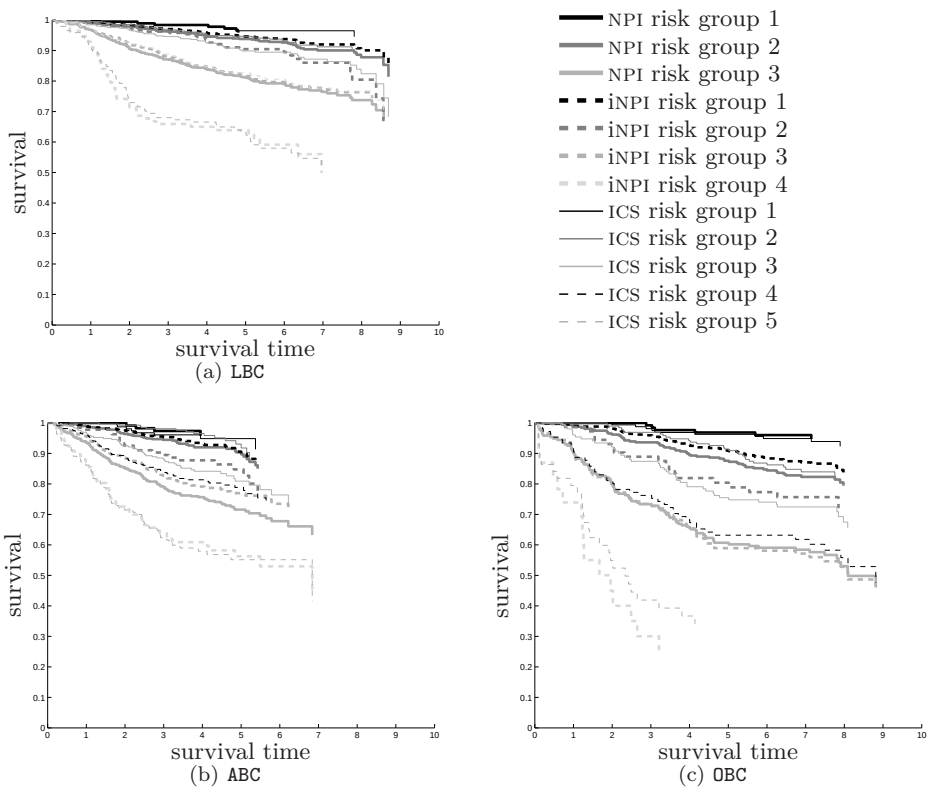


Figure 5.22: *Kaplan-Meier survival curves according to the risk groups defined by* NPI, *i*NPI *and* ICS. *The* NPI *is able to define a risk group with a very high survival during the whole follow-up time of the tested data. This group is not represented in the i*NPI *risk classification. However, the latter model is able to define a risk group with a very poor prognosis, which is not represented in the* NPI *risk classification. The* ICS *model finds both a very good and a very poor prognostic group.*

Table 5.6: *Comparison of the* NPI, *i*NPI *and* ICS*index on* `LBC`, `ABC` *and* `OBC`. *For each measure, the best performing model is indicated in bold. Statistical significant differences are indicated in bold as well. All three models are equally concordant with the failure time. The i*NPI *classifies a significantly higher percentage of patients into the most extreme risk groups than the* NPI. *The* ICS *index has an* EXT *% which is significantly larger than for both other models. No significant differences were seen in the test sets concerning c-index and event rate differences.*

				95% CI on difference		
measure	NPI	iNPI	ICS	iNPI-NPI	ICS-NPI	ICS-iNPI
LBC						
c-index	0.68	0.70	**0.70**	**0.0007 to 0.0282**	**0.0091 to 0.0448**	-0.0124 to 0.0219
EvR difference	0.04	0.05	**0.08**	**0.0011 to 0.0178**	**0.0231 to 0.0668**	**0.0159 to 0.0559**
EXT %	48.88	62.40	**73.95**	**11.02 to 15.65**	**22.15 to 27.73**	**9.30 to 13.99**
ABC						
c-index	**0.72**	**0.72**	**0.72**	-0.0120 to 0.0154	-0.0151 to 0.0235	-0.0099 to 0.0184
EvRdifference	0.08	0.09	**0.11**	-0.0098 to 0.0174	-0.0330 to 0.0512	-0.0061 to 0.0483
EXT %	51.55	61.11	**72.93**	**7.01 to 12.38**	**18.11 to 25.10**	**9.05 to 14.99**
OBC						
c-index	0.74	0.75	**0.75**	-0.0062 to 0.0282	-0.0273 to 0.0161	-0.0092 to 0.0221
EvRdifference	0.12	0.13	**0.16**	-0.0204 to 0.0660	-0.0140 to 0.1457	-0.0446 to 0.1319
EXT %	58.14	69.08	**84.02**	**7.40 to 15.05**	**22.04 to 30.18**	**11.39 to 18.34**

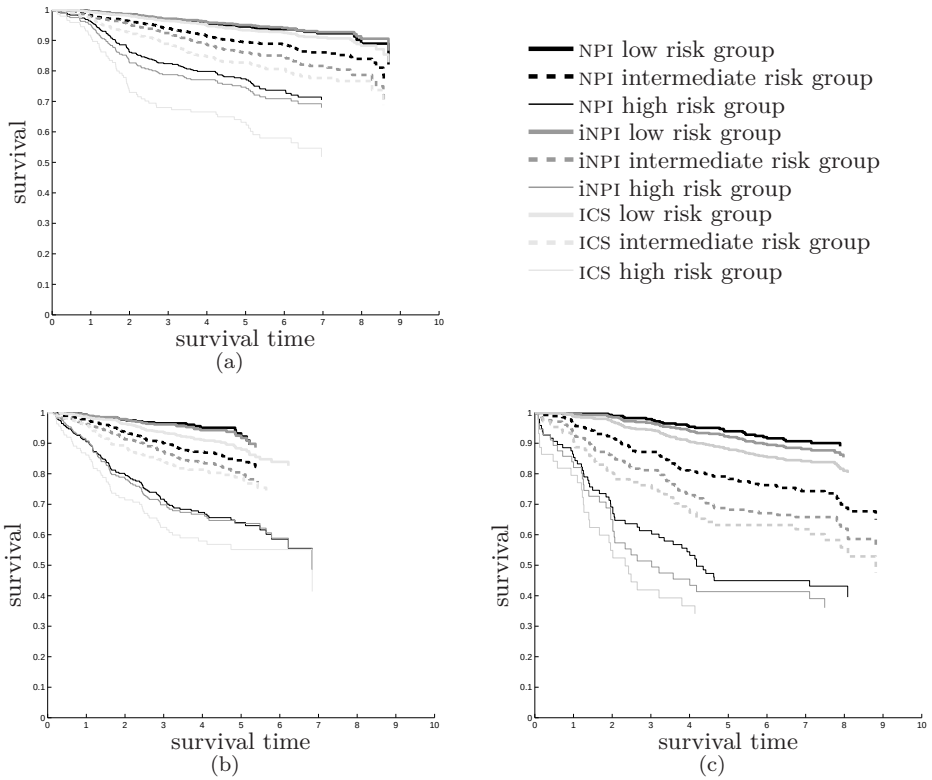


Figure 5.23: *Kaplan-Meier survival curves according to the risk groups of the* NPI, *i*NPI *and* ICS *models: (a)* `LBC`*, (b)* `ABC`*, (c)* `OBC`*. For all three models, the survival curves for the different risk groups are significantly different. Additionally, all low risk groups cluster together. The same accounts for the intermediate groups. The largest disagreement between the models is found in the high risk group. The* ICS *model finds a high risk group with a very poor prognosis, without significantly reducing the survival of the intermediate and low risk groups.*

survival is lower than in the NPI and iNPI low risk groups.

The survival curves for cross-classification of two of the three compared models are given in Figures 5.24 (`LBC`), 5.25 (`ABC`) and 5.26 (`OBC`). The survival curves of the patients who are reclassified by using the iNPI instead of the NPI, resemble more closely those for patients in the same iNPI group than those for patients in the same NPI group (more information on this comparison can be found in Van Belle *et al.* (2010d)). The same accounts when comparing the NPI with the ICS model. Comparison of the iNPI with the ICS model gives less obvious results. Patients who are allocated to different risk groups according to the iNPI and the ICS model

Table 5.7: *Classification table according to the* NPI*,* iNPI *and the* ICS *model for the three breast cancer data sets. The* NPI *classifies most patients in the intermediate risk group. The* ICS *model classifies most patients in the low risk group.*

	risk group		
	low	intermediate	high
data set	n (%)	n (%)	n (%)
LBC			
NPI	526 (27.35)	983 (51.12)	414 (21.53)
iNPI	880 (45.76)	723 (37.60)	320 (16.64)
ICS	1277 (66.41)	501 (26.05)	145 (7.54)
ABC			
NPI	330 (27.66)	578 (48.45)	285 (23.89)
iNPI	489 (40.99)	464 (38.89)	240 (20.12)
ICS	733 (61.44)	323 (27.07)	137 (11.48)
OBC			
NPI	297 (43.93)	283 (41.86)	96 (14.20)
iNPI	411 (60.80)	209 (30.92)	56 (8.23)
ICS	524 (76.38)	108 (15.98)	44 (6.51)

have a survival which is intermediate to the survival of their iNPI group and the survival of their ICS group.

Comparing the performance of these models should be done with care. The NPI is trained on untreated patients and can thus be used to make a choice on the appropriate treatment for a specific patient. The proposed model, as well as the iNPI, is trained on patients receiving the standard of care. Therefore, the model cannot be used to make decisions on whether a specific patient would benefit from a certain therapy. In order to be able to use the model for treatment decisions, a data set from a randomized trial should be used to train the model.

5.4 Conclusions

This Chapter proposed a new modelling approach in order to combine advanced mathematical models and clinical interpretability and applicability. The specific combination of (iteratively reweighted) L_1 norm regularization, the SVM classifier or MODEL 2 as basis and a visualization of the results by means of color bars results in generalizable, interpretable and clinically applicable models. The model was illustrated on three clinical problems. First, the model was used to predict malignancy of adnexal masses. The proposed model showed to perform as well as logistic regression models, perceptrons, Bayesian nets, artificial neural nets and least-squares support vector machines. Secondly, a well performing model for

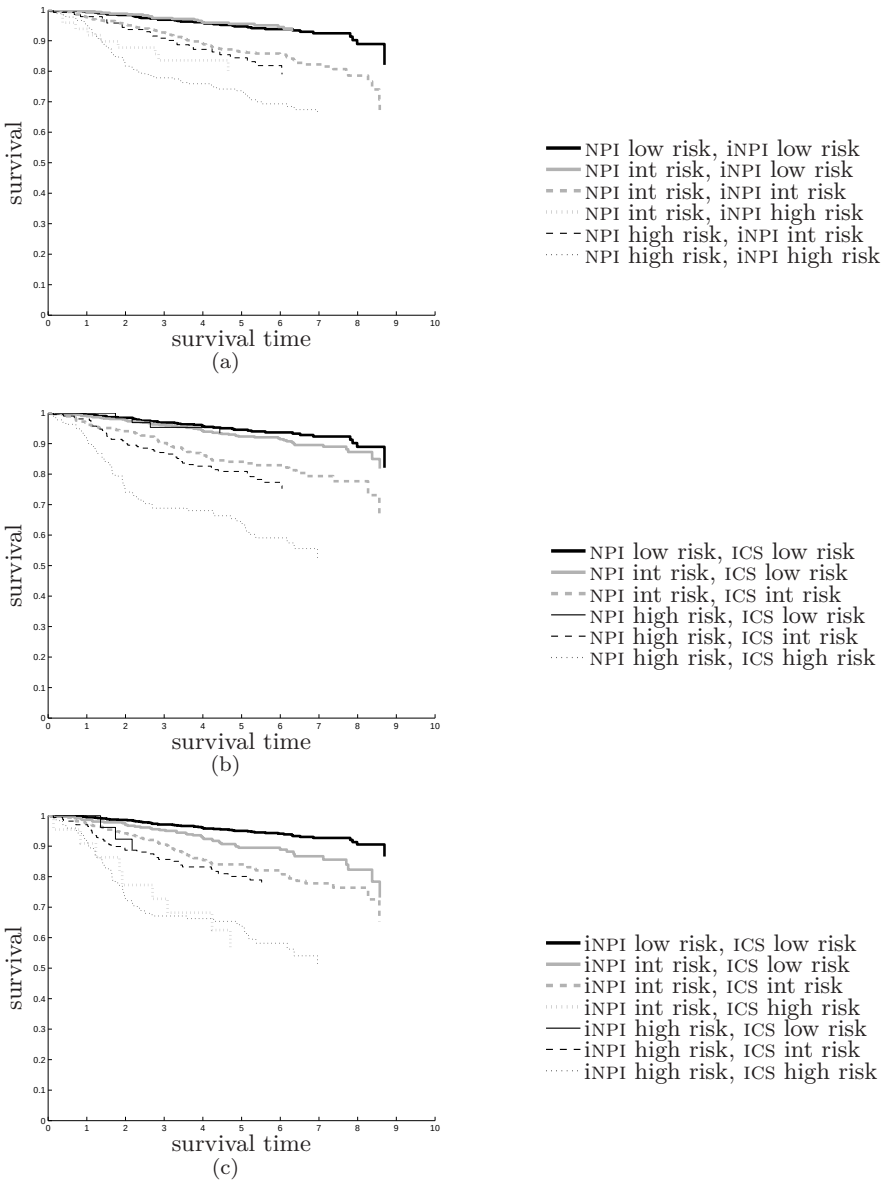


Figure 5.24: *Kaplan-Meier survival curves according to the cross-classification of two models on LBC: (a)* NPI *and i*NPI*, (b)* NPI *and* ICS*, (c) i*NPI *and* ICS*. Only groups containing 40 patients or more are represented.*

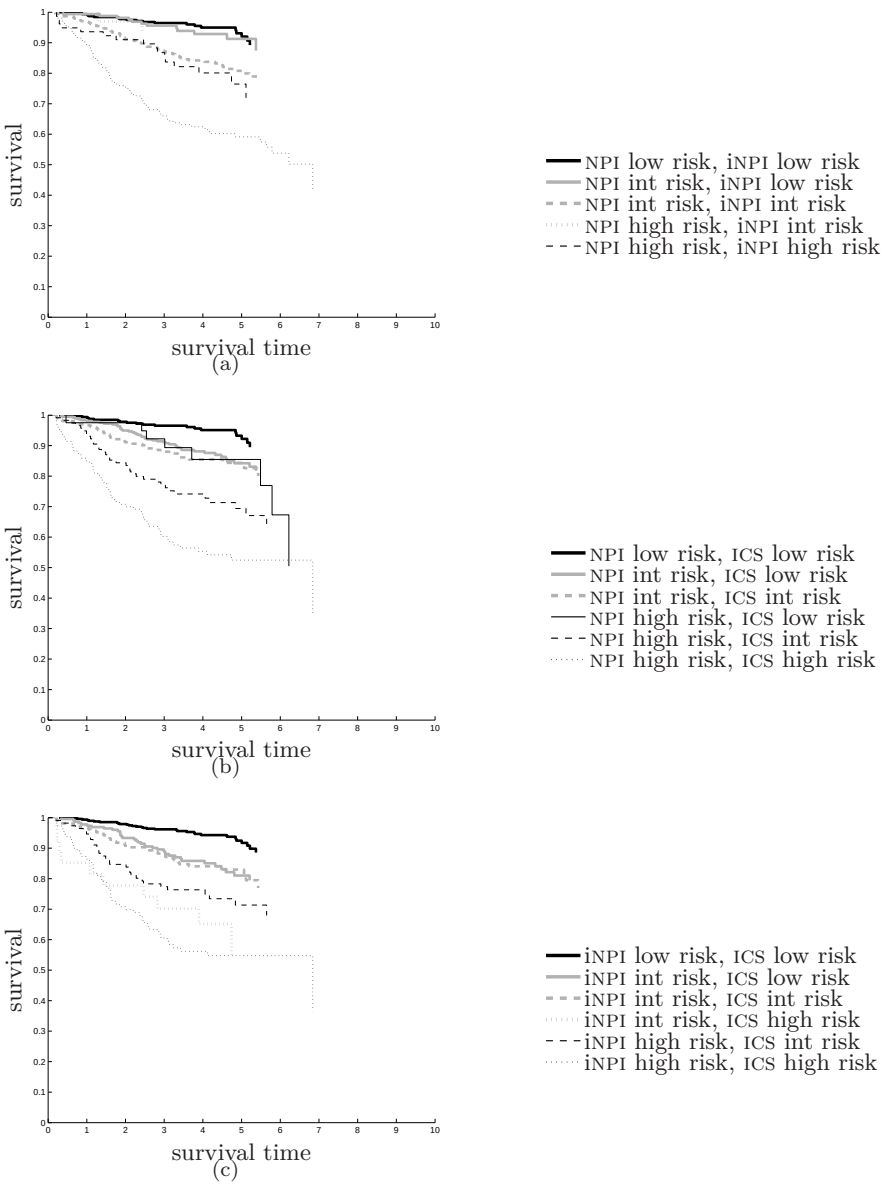


Figure 5.25: *Kaplan-Meier survival curves according to the cross-classification of two models on* ABC*: (a)* NPI *and i*NPI*, (b)* NPI *and* ICS*, (c) i*NPI *and* ICS*. Only groups containing 40 patients or more are represented.*

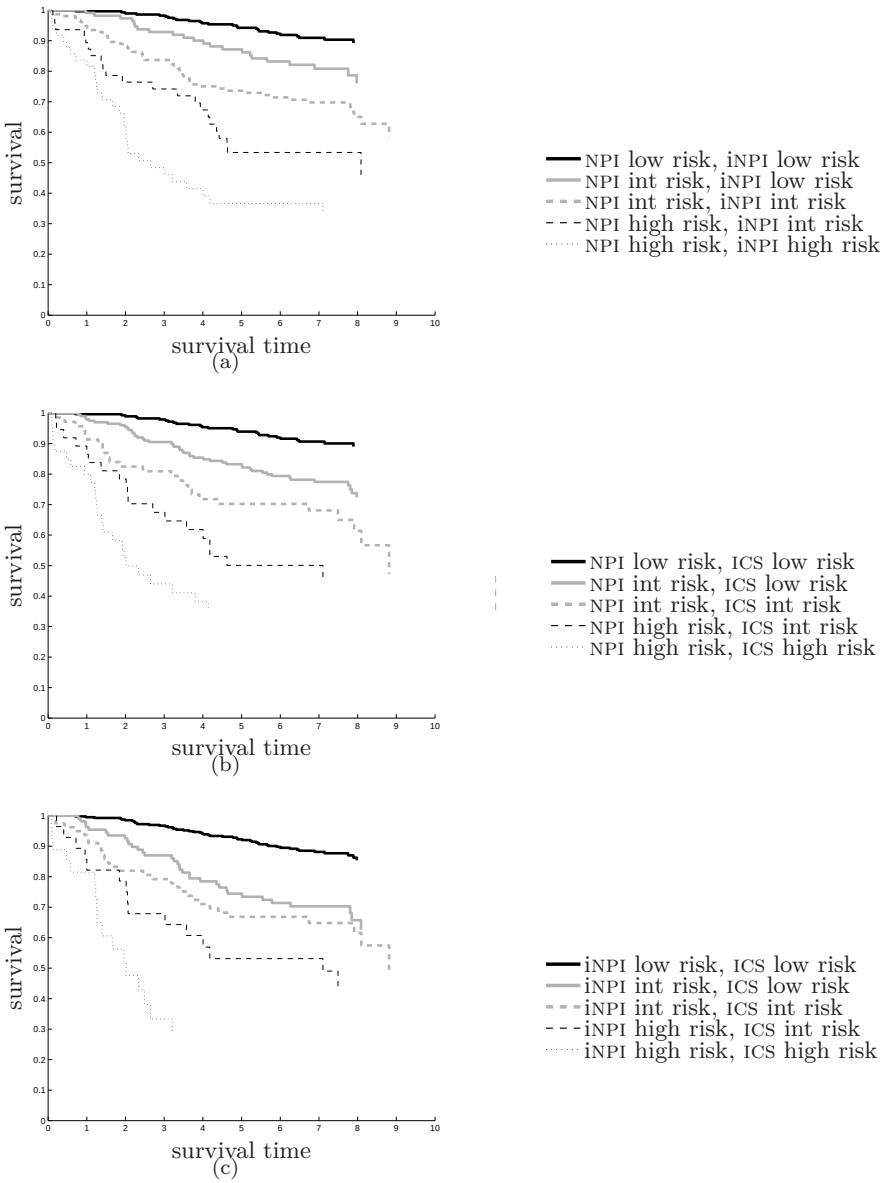


Figure 5.26: *Kaplan-Meier survival curves according to the cross-classification of two models on* `OBC`*: (a)* NPI *and i*NPI*, (b)* NPI *and* ICS*, (c) i*NPI *and* ICS*. Only groups containing 40 patients or more are represented.*

the prediction of non-viable pregnancies was found. In a last study, a cohort of 1923 primary, operable breast cancers was used to train a prognostic model. The performance was compared with the standard NPI and the recently proposed iNPI. Validation on two independent test sets, showed a significant improvement over the NPI, and a performance comparable to or slightly better than the iNPI.

Chapter 6

Conclusions and future work

This Chapter summarizes the main contributions of this dissertation. In a second Section different opportunities for future work are discussed.

6.1 Conclusion

In order to develop a new framework for the analysis of survival data, we collaborated with a team of medical doctors. As a result it was noted that clinicians are interested in dividing patients into different risk groups. We therefore started to formulate the survival problem as a ranking problem using the principles behind support vector machines. An optimal ranking function or utility was found by defining that the difference between the utility of a pair of observations (for which the outcome value of the second is smaller than the outcome value of the first) should be larger than the difference between the corresponding outcome values. The maximal margin principle was replaced with a smoothness condition on the ranking function. In first instance, this model was successfully applied to ranking problems. Extending the model towards ordinal regression problems, revealed that ranking instances by means of the estimated utility is only the first step. In order to get a prediction of the outcome level, a second step (reconstruction) is necessary. The link between the presented model and transformation models found in the literature was described in Chapter 3. It was proposed to adapt the ranking model by including dummy observations with an outcome in between subsequent outcome levels. As a result, the reconstruction step reduced to comparing the utility of a test observation with the utility of the dummy observations. An extra difficulty was introduced by considering possibly censored survival data. Due to the censoring, not all pairs of observations can be used. The solution to this

problem was found by comparing each observation with one single observation to which the former is comparable. Therefore, the model can be used in the presence of different types of censoring.

In Chapter 4, the proposed method is compared with different alternatives. Firstly, the model formulations of survival methods based on support vector machines which can be found in the literature were compared with our model. Secondly, different modifications to our model were proposed. The difference between the use of ranking and/or regression constraints, an L_1 or an L_2 norm, the use of equality versus inequality constraints and the comparison of each observation with one single (but arbitrarily chosen) comparable observation versus the comparison with dummy variables was discussed. The comparison of the methods based on SVMs revealed that inclusion of regression constraints significantly improved the performance. Additionally, it was shown that LS-SVM based models (with an L_2 norm and equality constraints), which have the advantage that the solution is found by solving a system of linear equations instead of a QP, was less appropriate for modelling censored data due to the fact that the information provided by the censored data can not be included in the regression constraints. It was therefore advised to use a model with ranking and regression constraints, using an L_1 norm loss function and Lipschitz continuity to control model complexity. Comparing this model with (semi-) parametric models on 6 clinical data sets concerning survival of cancer patients, did not reveal significant differences. In order to obtain sparse models, the coefficients of our model were, after preprocessing of the data, constraint to be positive. This adaptation resulted in a significantly better performance on high dimensional data as compared to proportional hazards models adapted for high dimensional data. It was concluded that SVM based models are preferred when non-linearities are present in the data. Thanks to the kernel trick, non-linearities are found automatically, in contrast with the PH, where a manual search is necessary. SVM based models also relax the proportional hazards assumption towards non-crossing survival function. When this assumption is not realistic, the PLANN or PLANNARD models should be used. For high dimensional data, the SVM approach performs the best.

In Chapter 5 it was noted that although mathematicians, statisticians and engineers continue to develop new, well performing models, clinical decision support systems based on these systems are rarely used in a clinical setting. Several reasons for this fact were found. Most importantly, clinicians are not encouraged to use these black-box models, since they are not user-friendly and difficult to interpret. Although rule extraction methods have been proposed to improve the applicability of the models, the resulting rules are often not clinically interpretable. In order to solve this gap between developers of mathematical models and users which are not familiar with the mathematical background of these models, a new mathematical framework was proposed. Instead of extracting rules from a previously built model, the problem was formulated in such a way that the

resulting model is a set of binary (yes/no) questions. In addition, two possible representations of the resulting models were proposed. A first representation was a collection (table) of yes/no questions related to each variable. Depending on which question has a positive answer, the patients receives a different number of points. The points for all variables were summed up to obtain the final score. The probability of the event corresponding to this score was found in a second part of table. The second representation used color bars instead of questions. Each variable was represented by a color bar, of which the colors corresponded to the points. Again, the final score was found by summing up all points. To obtain the estimated probability of the event corresponding to the score, a final color bar visualized the link between score and risk profile. Both representations allowed for an easy application of the model without interfering in the clinical work flow. Due to the simplicity of the result, the model can be used at the bed side as well as within software packages. An advantage of the model is that it can be used in a large variety of applications. It can be used to help clinicians with diagnostic and prognostic problems. In addition, the model can be used to inform patients on their risk profile and aid in making decisions regarding the appropriate treatment.

6.2 Future work

Although this research solved several problems, methodological challenges as well as challenges concerning the implementation of mathematical models within clinical practice remain. Several subjects for future work therefore involve:

- In the second stage of the survival model, the probability of the event under study needs to be estimated. At this point, the probability is estimated within groups of patients with a similar risk profile. However, an estimate of the probability for each specific patient is what clinicians and patients are interested in. A first problem in the probability estimator we proposed is that it did not incorporate assumptions made at the first stage of the transformation model. By optimizing the concordance index, it is assumed that the concordance index is independent of time. This assumption corresponds with non-crossing survival curves. This assumption should be incorporated in the probability estimation procedure. Additionally, the probability estimator should be adapted in order to deal with larger data sets. A possible solution could be found in formulating a fixed size version.
- In order for MODEL 2 to be applicable in a wide range of survival problems, the model needs to be extended towards time-dependent covariates and competing risks.
- An alternative for MODEL 2 developed in this thesis, could be found in an SVM version of the PLANN model. The advantage of this model would be

that the hazard could be estimated in one step, without assumptions on the hazards. A disadvantage of this model is that all observations need to be replicated at each time instance in which their survival status is known. To overcome this problem, one could use convex optimization techniques suitable for large scale problems.

- Since clinicians are not only interested in the estimate of the model, but also in the certainty of the results, addition of prediction intervals to the estimated functional forms of the different variables would be of high value. Although this might seem an easy problem, which can be solved by means of bootstrapping, finding the correct bootstrap is a difficult task. Looking at the research done on finding prediction intervals for regression problems, one might be tempted to use a wild bootstrap. However, the wild bootstrap needs the residuals defined as the difference between the true outcome and the estimated outcome. Due to the censoring in survival data, this residual cannot be calculated for censored observations. The solution lies in sampling from the censoring distribution, estimated on the whole data, in addition to sampling from the patient specific PDF.

- In order to find computational less demanding algorithms, weighted LS-SVM approaches could be investigated. This approach can possibly solve several issues stated above: (i) it leads to computationally simpler problems, (ii) the model can then be adapted for large data sets by means of a fixed-size version, (iii) it should be investigated whether confidence bounds can be found without bootstrapping as in the LS-SVM classification case, where the classifier is a linear smoother.

- Another possible research item for the future involves extending the ICS to other situations. First the model should be extended towards multi-class classification problems. For this goal, one can start from the multi-class SVM formulation. An additional problem is found in using high-dimensional data sets. Is the ICS model as it is implemented now able to deal with this type of data, or are adaptations necessary? Once this question is answered, the question whether the ICS model for survival can deal with high-dimensional data pops up.

- The ICS model was applied to three different problems in this dissertation, but thanks to the high interpretability and the straightforward applicability within the clinical work flow, ICS can be applied to a wide range of clinical problems. Possible application fields are diagnosing malignancy of tumors, detection of abnormal heart beats, estimation of cardiovascular risk, prognosis of cancer patients depending on the received treatment, prediction of therapeutic effects, informed decision making, among others. Additionally, it could be investigated whether ICS could be of use in other fields where interpretability is an issue. Possible applications can be found in economics and sociology.

- Since the number of high dimensional data sets is increasing, it would be interesting to investigate the use of ICS to high dimensional data sets. The method could be used to select a number of genes involved in different types of cancer. However, in order to validate whether the selected genes are indeed relevant, a collaboration with experts in genetics is necessary.
- The visualization methods proposed in this dissertation could be used to explain other mathematical models to a public which is not familiar with the mathematics.

Aside from these research points, there is one major point that needs attention. Well performing clinical decision support systems can only be developed and applied correctly when model developers and clinicians interact with each other. It is therefore of major importance that interdisciplinary research gets more attention than it currently receives.

Appendix A

Data description

A.1 Breast cancer data set from the Leuven University Hospitals

This first data set (Brouckaert *et al.*, 2009) contains information on 2064 patients with a primary operable breast cancer, treated at the Multidisciplinary Breast Center in the University Hospitals, Leuven and will be denoted as LBC. P. Neven and H. Wildiers were responsible for the collection of the data. All patients were diagnosed and operated on from January, 2000 to June, 2005. Patients receiving neo-adjuvant therapy were not included. For patients with bilateral breast cancer, the tumor characteristics of the lesion with the highest Nottingham prognostic index (Galea *et al.*, 1992) were retained. Information on the age at diagnosis, the location of the tumor, the size of the tumor, the number of positive lymph nodes (LNPOS), the progesterone receptor status (PR), the estrogen receptor status (ER), the human epidermal growth factor receptor 2 status (HER2), the grade of the tumor and whether the patient received chemo (CT), radio (RT) and/or hormonal (HT) therapy were available. Out of this set, 1923 (93.17%) patients had complete information. Most missings occur in the HER2 value for patients operated on in the early 2000. Patients with a positive progesterone receptor, but a negative ER were excluded (15 patients). The size of the tumor was evaluated by the pathologist and the largest diameter from the histopathology report was chosen. The grade of the tumor was performed according to the Ellis and Elson grading system (Elston & Ellis, 1991). For steroid receptors, any nuclear staining in invasive tumor cells was considered positive, both using Allred- or H-score. HER2 immunostaining was scored according to the standardized HercepTest scoring system in which HER2 expression is scored on a 0 to 3+ scale. Score 0 represents absence of staining whereas score 3+ indicates strong membranous

HER2 staining. In the case of intermediate scoring (score 2+), underlying HER2 gene copy number was investigated by fluorescence in situ hybridization or FISH (PathVision; Vysis, Downers Grove, IL, USA) for the presence of HER2 gene amplification. When the latter was present in score 2+ cases, the HER2 status considered positive. Surgical treatment consisted in wide local excision plus axillary dissection followed by whole breast radiotherapy plus a boost on the tumor bed, or modified radical mastectomy when breast-conserving surgery was not indicated. The sentinel node biopsy became standard treatment for cT1N0 patients in June 2003. Since then, completion of axillary dissection was carried out only in cases of metastatic sentinel node. Chest wall RT following mastectomy was given to patients with T3 or T4 tumors, with positive lymph nodes or with positive tumor section margins. Irradiation to the internal mammary chain was performed only in cases of axillary lymph node involvement or medial tumor sites. This was implemented first in the EORTC trial and subsequently as routine in this subset of patients. Patients received systemic adjuvant therapy that was considered appropriate for the time of treatment (rule based) and several patients participated in clinical studies with endocrine or chemotherapeutic modalities. All other patients were treated according to international guidelines and consensus conferences (rule based). Hormonal therapy was prescribed if ER and/or PR expression were present. The follow-up status of the patients was updated using the electronic patient files (last update: September 2009). If patients did not have follow-up information, the treating general physician was contacted to obtain this information. Loco-regional and metastatic relapse are considered as breast cancer related events. All other patients are censored at the last date of follow-up.

Models that use this data set as a training set, need to be validated on independent test sets. Two external data sets were available for this purpose. A first data set contains information on 862 consecutive patients from the Breast Cancer Micrometastasis group in Oslo (Norway) (denoted by `OBC`) diagnosed between 1995 and 1998. 676 patients had complete information (78%). The second data set contains 2805 patients from the Auckland Breast Cancer Registry (New Zealand) (denoted by `ABC`) diagnosed between January 2000 and December 2005. Complete information was available in 1192 patients (42%). In all data sets, HER2 is most often missing (4.0% in Leuven, 8.5% in Oslo, 57.5% in Auckland). In Leuven, low grade tumors more often had HER2 missing. In Auckland, HER2 is more missing in low grade tumors and in older patients, but the decision whether or not to test HER2 also depended on the laboratory. This may explain the large amount of missing values. In `OBC`, the missingness did not depend on grade or age. Descriptive statistics of the complete cases of all three sets are found in Table A.1.

Table A.1: *Descriptive statistics of the complete case datasets. In complete cases the variables size, # positive nodes,* PR, HER2 *and grade are available.*

Variable	variable level	Leuven		Oslo		Auckland	
Age (median, IQR)		57	48-66	58	50-67	57	48-67
	Missing	-		3		-	
Size (cm) (median, IQR)		2.0	1.5-3.5	1.7	1.1-2.5	1.8	1.2-2.8
	Missing	8		86		-	
# positive nodes (n, %)		0	0-1	0	0-1	0	0-1
	Missing	30		31		7	
ER (n, %)	Negative	252	13.08	120	17.75	320	26.85
	Positive	1675	86.92	545	80.62	872	73.15
	Missing	6		46		21	
PR (n, %)	Negative	472	24.49	152	22.49	459	38.51
	Positive	1455	75.51	524	77.51	733	61.49
	Missing	13		46		21	
HER2 (n, %)	Negative	1697	88.06	631	93.34	1016	85.23
	Positive	230	11.94	45	6.66	176	14.77
	Missing	82		73		1588	
Grade (n, %)	1	264	13.70	171	25.30	169	14.18
	2	889	46.13	327	48.37	555	46.56
	3	774	40.17	178	26.33	468	39.26
	Missing	6		19		36	
NPI risk group (n, %)	Low	601	31.19	297	43.93	333	27.94
	Intermediate	935	48.52	283	41.86	579	48.57
	High	391	20.29	96	14.20	280	23.49
	Missing	38		118		43	
EvR		0.0252		0.0377		0.0423	

A.2 International Ovarian Tumor Analysis Study

The data set collected by the International Ovarian Tumor Analysis Group (`IOTA`) contains information on 3511 patients with at least one overt persistent adnexal mass. The data were recruited in three phases. In phase 1, 1066 patients were recruited from 9 centers in 5 different countries from 1999 to 2002. In phase 1b, three of these centers included 507 extra patients in the period 2002-2005. In phase 2, 1938 patients were recruited from 2005 to 2007. In total, patients treated in 21 centers in 9 countries are included (see Table A.2). The principal investigator (D. Timmerman, J.P. Bernard, A. Czekierdowski, E. Epstein, E. Ferrazzi, D. Fischerov, D. Franchi, R. Fruscio, S. Greggi, S. Guerriero, J. Zhang, D. Jurkovic, F.P.G. Leone, A.A. Lissoni, H. Muggah, D. Paladini, A. Rossi, L. Savelli, A.C. Testa, D. Trio, C. Van Holsbeke) from one of the participating centers had to assess the mass. In case of multiple masses, the data from the most complex mass was included. The variables considered in the experiments are summarized in Table A.3. For two patients, the maximal diameter of the solid component was larger than the maximal diameter of the lesion. Since this is impossible, these two patients were not included in the analysis. The primary endpoint is to diagnose the malignancy (including borderline, primary invasive and metastatic tumor) of the tumor. Out of the 3509 included patients, 951 (27%) were malignant and 2558 (73%) were benign.

A.3 The early pregnancy unit

This data set (`EPD`) is an observational cohort of all women attending the Early Pregnancy and Gynaecology Ultrasound unit (EPU) of St. George's University in London from January to October 2006. T. Bourne and C. Bottomley were responsible for the data collection. Inclusion criteria were a positive pregnancy test and a gestational age of less than 12 weeks. Patients presenting with postpartum or complication from termination of pregnancy or recently diagnosed early pregnancy loss were not considered. Exclusion criteria were: twin or triplet pregnancies, patients who appeared not to be pregnant and patients who underwent termination of pregnancy. Patients presenting with more than one pregnancy over the study period were included only once (first presenting pregnancy). A total of 1435 patients remained for analysis. The data was randomly divided in a training and test set, containing 955 and 480 patients, respectively. The outcome was defined as viable when a live fetus had been seen at the routine 11-14 week scan, performed after the initial scan. The outcome was non-viable whenever the pregnancy had resulted in a miscarriage, an ectopic pregnancy or a failed pregnancy of unknown location. Out of the 1435 included pregnancies, 550 (38%) were non-viable and 885 (62%) were viable. The variables considered in the experiments are summarized

in Table A.4. Missings were imputed using distance-aided selection of donors (Siddique & Harel, 2009).

A.4 Publicly available data sets

A.4.1 Clinical data

This work makes use of 5 publicly available cancer data sets containing clinical information. The data sets can be found on the world-wide-web (http://lib.stat.cmu.edu/datasets (accessed: 29 August 2008) and the R-package http://cran.r-project.org/web/packages/survival/index.html (accessed: 20 November 2007)).

Leukemia data set

The first data set concerns the prediction of complete remission (`LCR`) and death (`LD`) of 129 patients with leukemia (Emerson & Banks, 1994). Information on the treatment, sex, age, performance score (Karnofsky score), white blood cells, platelets, hemoglobin and whether the patient received a kidney transplant was given.

Veteran lung cancer data set

A second data set concerns the veteran's administration lung cancer trial (`VLC`) (Prentice, 1974; Kalbfleisch & Prentice, 2002). Only 9 out of the 137 men with advanced inoperable lung cancer were alive at the end of the study. Patients were randomized into a standard or test chemotherapy. Information on the performance score, age, therapy, histology of the tumor, treatment and time between the diagnosis and the randomization was available.

Prostate cancer data set

The third data set concerns 506 patients with prostatic cancer (`PC`) (Byar & Green, 1980). The variables used for this experiment are: treatment, age, weight index, performance index, history of cardiovascular disease, size of the tumor, a combined index of stage of histologic grade and serum hemoglobin. The analysis was performed on the 483 patients with complete information. 125 (26%) patients died due to prostatic cancer during the study. All other patients were right censored at the date of last follow-up.

Table A.2: *Description of the centers included in* `IOTA`

Center code	City	Country	centertype	# patients included in		
				phase 1	phase 1b	phase 2
GBE	Genk	Belgium	regional			200
LBE	Leuven	Belgium	tertiary oncologic	263	287	252
OCA	Ontario	Canada	regional			12
BCH	Beijing	China	tertiary oncologic			73
PCR	Prague	Czech Republic	tertiary oncologic			90
MFR	Maurepas	France	regional	64		
BFR	Paris	France	regional	80		
BIT	Bologna	Italy	tertiary not oncologic			135
CIT	Milano	Italy	tertiary oncologic			94
MIT	Milano	Italy	tertiary not oncologic	87		50
VIT	Milano	Italy	regional			21
OIT	Monza	Italy	tertiary oncologic	46		251
GIT	Napoli	Italy	tertiary oncologic			9
NIT	Napoli	Italy	tertiary not oncologic	31		64
RIT	Roma	Italy	tertiary oncologic	126	96	122
SIT	Sardinia	Italy	regional			154
UDI	Udinese	Italy	tertiary oncologic regional			17
LPO	Lublin	Poland	tertiary oncologic			154
LSW	Lund	Sweden	tertiary oncologic			38
MSW	Malmö	Sweden	regional	315	124	137
KUK	London	UK	tertiary oncologic	54		65

Table A.3: *Description of the* `IOTA` *variables considered in this work*

variable	description	unit/level	type
age	age of patient	years	continuous
famhistovca	family members with history of ovarian cancer	binary (0 no; 1 yes)	binary
famhistbrca	nr of family members with history of breast cancer	count	ordinal
pershistovca	personal history of ovarian cancer	binary (0 no; 1 yes)	binary
pershistbrca	personal history of breast cancer	binary (0 no; 1 yes)	binary
Parity	parity	count	ordinal
hysterectomy	hysterectomy	binary (0 no; 1 yes)	binary
hormtherapy	current use of hormonal therapy	binary (0 no; 1 yes)	binary
bilateral	masses on both sides	binary (0 no; 1 yes)	binary
lesdmax	maximum diameter of lesion	mm	continuous
pain	pelvic pain during examination	binary (0 no; 1 yes)	binary
nrloculescat	number of locules	0 zero	ordinal
		1 one	
		2 two	
		3 three	
		4 four	
		5 five to ten	
		6 more than ten	
soldmax	max. diameter of largest solid component	mm	continuous
papnr	nr of papillary structures	0 zero	ordinal
		1 one	
		2 two	
		3 three	
		4 more than three	
papheight	height of largest papillary structure	mm	continuous

Continued on next page

Table A.3 – continued from previous page

variable	description	unit/level	type*
papflow	presence of blood flow within papillary structures	binary (0 no; 1 yes)	binary
papsmooth	largest papillary structure is irregular	binary (0 no; 1 yes)	binary
wallreg	irregular internal cyst wall	binary (0 no; 1 yes)	binary
incomplseptum	presence of an incomplete septum	binary (0 no; 1 yes)	binary
shadows	presence of acoustic shadows	binary (0 no; 1 yes)	binary
echogenicity	echogenicity of mass	1 anechoic	categorical
		2 homogeneous low-level	
		3 ground glass	
		4 hemorrhagic	
		5 mixed	
		6 no cyst fluid	
colscore	color score of intratumoural blood flow	1 no blood flow	ordinal
		2 minimal	
		3 moderate	
		4 very strong	
venous	venous blood flow only	binary (0 no; 1 yes)	binary
ascites	presence of ascites	binary (0 no; 1 yes)	binary
fluid	amount of fluid in Pouch of Douglas	mm	continuous

Table A.4: *Description of the variables of the* `EPD` *considered in this work*

variable	description	unit/level	type
age	maternal age	years	continuous
pbac	bleeding score	0: no bleeding	ordinal
		1: light bleeding	
		2: moderate bleeding	
		3: heavy bleeding	
		4: bleeding with cloths	
gravida	number of pregnancies (including the present pregnancy)		ordinal
nodel24	number of deliveries after 24 weeks		ordinal
notop	number of termination of pregnancies		ordinal
noemis	number of early miscarraiges		ordinal
nopul	previous pregnancies of unknown location	0: no	binary
		1: yes	
nolmis	previous late miscarriages	0: no	binary
		1: yes	
noect	previous ectopic pregnancies	0: no	binary
		1: yes	
gestdates	gestational age: days since start of last menstrual period	days	continuous
vas	pain score	cm	continuous
pain	is the patient in pain? (unspecified)	0: no	binary
		1: yes	
meangsd	mean gestational sac diameter	mm	continuous
meanysd	mean yolk sac diameter	mm	continuous
fhrseen	is a fetal heart rate seen?	0: no	binary
		1: yes	
fetus	is a fetus seen?	0: no	binary

Continued on next page

Table A.4 – continued from previous page

variable	description	unit/level	type
		1: yes	
crl	crown-rump-length	mm	continuous
regdates	did the patient report a regular cycle of 26 to 30 days?	0: no	binary
		1: yes	

Mayo lung cancer data set

The Mayo clinic lung cancer data (MLC) (Therneau & Grambsch, 2000) contains information on 167 patients with advanced lung cancer. The data set contains information on age, sex, two performance scores estimated by the physician and one estimated by the patient, the number of calories consumed and the weight loss in the last six months. During the study period, 72% of the patients died.

German breast cancer study group data set

A last data set contains information on 720 breast cancer patients (the German breast cancer study) (GBC) (Schumacher *et al.*, 1994; Sauerbrei & Royston, 1999). Information was available on hormonal therapy, menopausal status, age, grade, tumor size, number of positive lymph nodes, progesterone and estrogen receptors. The experiment is based on the 686 cases with complete information. During the study period, 299 patients (44%) experienced a breast cancer related event.

A.4.2 Micro-Array data

The Norway/Stanford breast cancer data

The Norway/Stanford breast cancer data (NSBCD) from Sørlie *et al.* (2003) contains gene expression measurements from 115 women with breast cancer. Missing values were previously imputed using 10-nearest neighbor imputation (Bøvelstad *et al.*, 2007). 38 (33%) patients experienced an event within the study period.

The Dutch Breast Cancer Data

The Dutch Breast Cancer Data (DBCD) from van Houwelingen *et al.* (2006) is a subset of the data from van de Vijver *et al.* (2002) and contains information on 4919 gene expression levels of a consecutive series of 295 women with breast cancer from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. All 295 tumors were primary invasive breast carcinoma less than 5 cm in diameter. The women were 52 years or younger. The diagnosis was made between 1984 and 1995 and there was no previous history of cancer, except non-melanoma skin cancer. In 79 (26.78%) patients distant metastases were noted within the study period. The median follow-up was 6.7 years (range 0.05-18.3).

The diffuse large-B-cell lymphoma data

The diffuse large-B-cell lymphoma data (`DLBCL`) from Rosenwald *et al.* (2002) contains data on 240 patients with diffuse large-B-cell lymphoma. The data consists of 7399 gene expression measurements. The median follow-up time was 2.8 years and 58% of the patients died during the study period.

Appendix B

Primal-dual formulations

B.1 MINLIP for ranking problems

This appendix provides the formal derivation of the MINLIP method. We start with the problem formulation as denoted in equation (3.22):

$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} |\epsilon_i| \\ \text{subject to} \quad & w^T (x_{(i)} - x_{(i-1)}) \geq y_{(i)} - y_{(i-1)} - \epsilon_i + \epsilon_{i-1}, \\ & \qquad\qquad \forall\ i = 2, \ldots, n\,. \end{aligned} \tag{B.1}$$

Let $\epsilon_i = \epsilon_i^+ - \epsilon_i^-$ and suppose $\epsilon^+ \geqslant 0$ and $\epsilon^- \geqslant 0$. The problem can than be formulated as

$$\begin{aligned} \min_{w,\epsilon^+,\epsilon^-} \quad & \tfrac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i^+ + \epsilon_i^-) \\ \text{subject to} \quad & \begin{cases} w^T (x_{(i)} - x_{(i-1)}) \geq y_{(i)} - y_{(i-1)} - \epsilon_i^+ + \epsilon_i^-, & \\ \quad + \epsilon_{i-1}^+ - \epsilon_{i-1}^- & \forall\ i = 2, \ldots, n\,, \\ \epsilon_i^+ \geq 0, & \forall\ i = 1, \ldots, n\,, \\ \epsilon_i^- \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{B.2}$$

The Lagrangian becomes

$$\begin{aligned}\mathcal{L}(w,\epsilon^+,\epsilon^-;\alpha,\beta,\beta^*) &= \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}(\epsilon_i^+ + \epsilon_i^-) - \sum_{i=1}^{n}\beta_i\epsilon_i^+ - \sum_{i=1}^{n}\beta_i^{*T}\epsilon^- \\ &- \sum_{i=2}^{n}\alpha_i\left(w^T(x_{(i)} - x_{(i-1)}) - y_{(i)} + y_{(i-1)} + \epsilon_i^+ - \epsilon_i^- - \epsilon_{i-1}^+ + \epsilon_{i-1}^-\right),\end{aligned} \tag{B.3}$$

with Lagrange multipliers $\alpha_i,\ \beta_i,\ \beta_i^* \geq 0$. The conditions for optimality (Karush-Kuhn-Tucker (KKT) conditions) become

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{1=2}^{n}\alpha_i\left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right) \\ \frac{\partial\mathcal{L}}{\partial\epsilon^+} = 0 \rightarrow \gamma = \begin{cases} \alpha_i - \alpha_{i+1} - \beta_i, & \forall\ i = 2,\ldots,n-1\,, \\ -\alpha_2 + \beta_1 \\ \alpha_n + \beta_n \end{cases} \\ \frac{\partial\mathcal{L}}{\partial\epsilon^-} = 0 \rightarrow \gamma = \begin{cases} -\alpha_i + \alpha_{i+1} + \beta_i^*, & \forall\ i = 2,\ldots,n-1, \\ \alpha_2 + \beta_1^* \\ -\alpha_n + \beta_n^* \end{cases} \\ \alpha_i\left(w^T(x_{(i)} - x_{(i-1)}) - y_{(i)} + y_{(i-1)} + \epsilon_i^+ - \epsilon_i^- - \epsilon_{i-1}^+ + \epsilon_{i-1}^-\right) = 0, \\ \qquad\qquad \forall\ i = 2,\ldots,n\,, \\ \beta_i\epsilon_i^+ = 0, \qquad \forall\ i = 1,\ldots,n\,, \\ \beta_i^*\epsilon_i^- = 0, \qquad \forall\ i = 1,\ldots,n\,, \\ \alpha_i \geq 0, \qquad \forall\ i = 2,\ldots,n\,, \\ \beta_i \geq 0, \qquad \forall\ i = 1,\ldots,n\,, \\ \beta_i^* \geq 0, \qquad \forall\ i = 1,\ldots,n\,. \end{cases} \tag{B.4}$$

From Slater's condition one can exchange $\min_{w,\epsilon^+,\epsilon^-}\max_\alpha$ with $\max_\alpha\min_{w,\epsilon^+,\epsilon^-}$. Solving for w, ϵ^+ and ϵ^- gives the dual problem

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=2}^{n}\alpha_i\alpha_j\left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right)^T\left(\varphi(x_{(j)}) - \varphi(x_{(j-1)})\right) \\ & - \sum_{i=2}^{n}\alpha_i(y_{(i)} - y_{(i-1)}) \\ \text{subject to} \quad & \begin{cases} \alpha_i \geq 0, & \forall\ i = 2,\ldots,n, \\ -\gamma \leq \alpha_i - \alpha_{i+1} \leq \gamma, & \forall\ i = 2,\ldots,n-1, \\ -\gamma \leq \alpha_2 \leq \gamma, \\ -\gamma \leq \alpha_n \leq \gamma\,. \end{cases} \end{aligned} \tag{B.5}$$

From the first condition of (B.4) and the model specification $u(x) = w^T\varphi(x)$ one could write the solution for a new point $x_\star$ as

$$\hat{u}(x_\star) = \sum_{i=2}^{n} \alpha_i \left(\varphi(x_{(i)}) - \varphi(x_{(i-1)})\right)^T \varphi(x_\star)\,. \tag{B.6}$$

B.2 MINLIP for ordinal regression

In this appendix the derivation of the MINLIP method for ordinal regression is exposed. In the ordinal regression case unknown thresholds s are introduced corresponding to an outcome intermediate between two successive outcome levels. The model is built by indicating that the difference between the utility of a certain observation x_i and the largest threshold lower than the outcome of that observation y_i should be larger than the difference between y_i and the outcome corresponding to the before-mentioned threshold. Analogously, the difference between the smallest threshold higher than y_i should be larger than the difference between the outcome corresponding to that threshold and y_i. As an extra constraint we impose that successive thresholds are increasing values of the utility function. The problem was formulated in equation (3.33) as:

$$\begin{aligned}
\min_{w,\epsilon,\epsilon^*} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\
\text{subject to} \quad & \begin{cases}
w^T\varphi(x_i) - \sum_{l=1}^{k-1} \underline{q}_l(i) v_l \geq y_i - \sum_{l=1}^{k-1} \underline{q}_l(i) s_l - \epsilon_i, \\
\qquad\qquad \forall\ i = 1, \ldots, n, \\
-w^T\varphi(x_i) + \sum_{l=1}^{k-1} \overline{q}_l(i) v_l \geq -y_i + \sum_{l=1}^{k-1} \overline{q}_l(i) s_l - \epsilon_i^*, \\
\qquad\qquad \forall\ i = 1, \ldots, n, \\
\epsilon_i \geq 0, \qquad \forall\ i = 1, \ldots, n, \\
\epsilon_i^* \geq 0, \qquad \forall\ i = 1, \ldots, n, \\
v_{l+1} - v_l \geq 0, \qquad \forall\ l = 1, \ldots, k-2\,.
\end{cases}
\end{aligned} \tag{B.7}$$

The Lagrangian of this problem formulation is

$$\begin{aligned}\mathcal{L}(w,\epsilon,\epsilon^*;\alpha,\alpha^*,\beta,\beta^*,\theta) &= \frac{1}{2}w^T w+\gamma\sum_{i=1}^{n}(\epsilon_i+\epsilon_i^*)-\sum_{i=1}^{n}\beta_i\epsilon_i\\ &-\sum_{i=1}^{n}\alpha_i\left(w^T\varphi(x_i)-\sum_{l=1}^{k-1}\underline{q}_l(i)v_l-y_i+\sum_{l=1}^{k-1}\underline{q}_l(i)s_l+\epsilon_i\right)\\ &-\sum_{i=1}^{n}\alpha_i^*\left(-w^T\varphi(x_i)+\sum_{l=1}^{k-1}\overline{q}_l(i)v_l+y_i-\sum_{l=1}^{k-1}\overline{q}_l(i)s_l+\epsilon_i^*\right)\\ &-\sum_{i=1}^{n}\beta_i^*\epsilon_i^*-\sum_{l=1}^{k-2}\theta_l(v_{l+1}-v_l)\,.\end{aligned} \tag{B.8}$$

The set of optimality conditions becomes

$$\begin{cases}\frac{\partial\mathcal{L}}{\partial w}=0\rightarrow w=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\,\varphi(x_i)\\ \frac{\partial\mathcal{L}}{\partial\epsilon_i}=0\rightarrow\gamma=\alpha_i-\beta_i\\ \frac{\partial\mathcal{L}}{\partial\epsilon_i^*}=0\rightarrow\gamma=\alpha_i^*-\beta_i^*\\ \frac{\partial\mathcal{L}}{\partial v_l}=0\rightarrow\begin{cases}\sum_{i=1}^{n}\alpha_i\underline{q}_l(i)\sum_{i=1}^{n}\alpha_i^*\overline{q}_l(i)-\theta_{l-1}+\theta_l=0, & \forall\ l=2,\ldots,k-2,\\ \sum_{i=1}^{n}\alpha_i\underline{q}_l(i)\sum_{i=1}^{n}\alpha_i^*\overline{q}_l(i)-\theta_{l-1}=0, & \forall\ l=k-2,\\ \sum_{i=1}^{n}\alpha_i\underline{q}_l(i)\sum_{i=1}^{n}\alpha_i^*\overline{q}_l(i)+\theta_l=0, & \forall\ l=1,\end{cases}\\ \alpha_i\left(w^T\varphi(x_i)-\sum_{l=1}^{k-1}\underline{q}_l(i)v_l-y_i+\sum_{l=1}^{k-1}\underline{q}_l(i)s_l+\epsilon_i\right)=0, & \forall\ i=1,\ldots,n\\ \alpha_i^*\left(-w^T\varphi(x_i)+\sum_{l=1}^{k-1}\overline{q}_l(i)v_l+y_i-\sum_{l=1}^{k-1}\overline{q}_l(i)s_l+\epsilon_i^*\right)=0, & \forall\ i=1,\ldots,n\\ \beta_i\epsilon_i=0, & \forall\ i=1,\ldots,n\,,\\ \beta_i^*\epsilon_i^*=0, & \forall\ i=1,\ldots,n\,,\\ \theta_l\,(v_{l+1}-v_l)=0, & \forall\ l=1,\ldots,k-2\,.\end{cases}\end{cases} \tag{B.9}$$

The dual problem formulation is than found as

$$\begin{aligned}
\min_{\alpha,\alpha^*,\theta} \quad & \tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)\varphi(x_i)^T\varphi(x_j)(\alpha_j-\alpha_j^*)-\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)y_i \\
& +\sum_{i=1}^{n}\sum_{i=1}^{n}\sum_{l=1}^{k-1}\left(\alpha_i\underline{q}_l(i)-\alpha_i^*\overline{q}_l(i)\right)s_l \\
\text{subject to} \quad & \begin{cases}
0\le\alpha_i\le\gamma, & \forall\ i=1,\dots,n,\\
0\le\alpha_i^*\le\gamma, & \forall\ i=1,\dots,n,\\
\sum_{i=1}^{n}\left(\alpha_i\underline{q}_l(i)-\alpha_i^*\overline{q}_l(i)\right)-\theta_{l-1}+\theta_l=0, & \\
& \forall\ l=2,\dots,k-2\,,\\
\sum_{i=1}^{n}\left(\alpha_i\underline{q}_l(i)-\alpha_i^*\overline{q}_l(i)\right)-\theta_{l-1}, & \forall\ l=k-2\,,\\
\sum_{i=1}^{n}\left(\alpha_i\underline{q}_l(i)-\alpha_i^*\overline{q}_l(i)\right)+\theta_l=0, & \forall\ l=1\,.
\end{cases}
\end{aligned} \tag{B.10}$$

B.3 MODEL 1

The derivation of MODEL 1 is given in this Appendix. The problem was formulated as

$$\begin{aligned}
&\boxed{\text{MODEL 1}}\\
\min_{w,\epsilon} \quad & \frac{1}{2}w^Tw+\gamma\sum_{i=1}^{n}\epsilon_i \\
\text{subject to} \quad & \begin{cases}
w^T(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))\ge y_i-y_{\bar{j}(i)}-\epsilon_i, & \\
& \forall\ i=1,\dots,n\,,\\
\epsilon_i\ge 0, & \forall\ i=1,\dots,n\,.
\end{cases}
\end{aligned} \tag{B.11}$$

Formulating the Lagrangian of (B.11)

$$\begin{aligned}
\mathcal{L}(w,\epsilon;\alpha,\beta) \quad = \quad & \frac{1}{2}w^Tw+\gamma\sum_{i=1}^{n}\epsilon_i-\sum_{i=1}^{n}\alpha_iw^T(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))\\
& -\sum_{i=1}^{n}\alpha_i\left(-y_i+y_{\bar{j}(i)}+\epsilon_i\right)-\sum_{i=1}^{n}\beta_i\epsilon_i\,,
\end{aligned} \tag{B.12}$$

and solving the KKTequations

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\bar{j}(i)}) \right), & \\ \frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \gamma = \alpha_i + \beta_i, & \forall\, i = 1, \ldots, n\,, \\ \alpha_i \geq 0, & \forall\, i = 1, \ldots, n\,, \\ \beta_i \geq 0, & \forall\, i = 1, \ldots, n\,, \\ \sum_{i=1}^{n} \alpha_i \left(w^T (\varphi(x_i) - \varphi(x_{\bar{j}(i)})) - y_i + y_{\bar{j}(i)} + \epsilon_i \right) = 0, & \\ \sum_{i=1}^{n} \beta_i \epsilon_i = 0\,, & \end{cases} \tag{B.13}$$

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

$$\begin{array}{ll} \min\limits_{\alpha_i} & \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k (\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T (\varphi(x_k) - \varphi(x_{\bar{l}(k)})) \\ & - \sum_{i=1}^{n} \alpha_i (y_i - y_{\bar{j}(i)}) \\ \text{subject to} & 0 \geq \alpha_i \geq \gamma, \forall\, i = 1, \ldots, n\,. \end{array} \tag{B.14}$$

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

$$\hat{u}(x_\star) = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\bar{j}(i)}) \right)^T \varphi(x_\star)\,. \tag{B.15}$$

B.4 MODEL 2

MODEL 2 was defined as

MODEL 2

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

$$\text{subject to} \quad \begin{cases} w^T (\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \\ & \forall\, i = 1, \ldots, n\,, \\ w^T \varphi(x_i) + b \geq y_i - \xi_i, & \forall\, i = 1, \ldots, n\,, \\ -\delta_i (w^T \varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\, i = 1, \ldots, n\,, \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n\,, \\ \xi_i \geq 0, & \forall\, i = 1, \ldots, n\,, \\ \xi_i^* \geq 0, & \forall\, i = 1, \ldots, n\,. \end{cases} \tag{B.16}$$

The Lagrangian becomes

$$\begin{aligned}\mathcal{L}(w,\epsilon,b;\alpha,\beta) &= \frac{1}{2}w^T w+\gamma\sum_{i=1}^{n}\epsilon_i+\mu\sum_{i=1}^{n}(\xi_i+\xi_i^*)\\ &\quad-\sum_{i=1}^{n}\alpha_i\left(w^T(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))-y_i+y_{\bar{j}(i)}+\epsilon_i\right)\\ &\quad-\sum_{i=1}^{n}\beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right)\\ &\quad-\sum_{i=1}^{n}\beta_i^*\left(-\delta_i(w^T\varphi(x_i)+b-y_i)+\xi_i^*\right)\\ &\quad-\sum_{i=1}^{n}\eta_i\epsilon_i-\sum_{i=1}^{n}\nu_i\xi_i-\sum_{i=1}^{n}\nu_i^*\xi_i^*\,.\end{aligned}\tag{B.17}$$

The KKT conditions for optimality are

$$\begin{cases}\frac{\partial\mathcal{L}}{\partial w}=0\rightarrow w=\sum_{i=1}^{n}\alpha_i\left(\varphi(x_i)-\varphi(x_{\bar{j}(i)})\right)+\sum_{i=1}^{n}(\beta_i-\delta_i\beta_i^*)\varphi(x_i),\\ \frac{\partial\mathcal{L}}{\partial b}=0\rightarrow\sum_{i=1}^{n}(-\beta_i+\delta_i\beta_i^*)=0,\\ \frac{\partial\mathcal{L}}{\partial\epsilon_i}=0\rightarrow\gamma=\alpha_i+\eta_i, & \forall\ i=1,\ldots,n\,,\\ \frac{\partial\mathcal{L}}{\partial\xi_i}=0\rightarrow\mu=\beta_i+\nu_i, & \forall\ i=1,\ldots,n\,,\\ \frac{\partial\mathcal{L}}{\partial\epsilon_i}=0\rightarrow\gamma=\beta_i^*+\nu_i^*, & \forall\ i=1,\ldots,n\,,\\ \alpha_i,\beta_i,\beta_i^*,\eta_i,\xi_i,\xi_i^*\geq 0 & \forall\ i=1,\ldots,n\,,\\ \sum_{i=1}^{n}\alpha_i\left(w^T(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))-y_i+y_{\bar{j}(i)}+\epsilon_i\right)=0,\\ \sum_{i=1}^{n}\beta_i\epsilon_i=0,\\ \sum_{i=1}^{n}\beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right)=0,\\ \sum_{i=1}^{n}\beta_i^*\left(-\delta_i(w^T\varphi(x_i)+b-y_i)+\xi_i^*\right)=0,\\ \sum_{i=1}^{n}\eta_i\epsilon_i=0,\\ \sum_{i=1}^{n}\nu_i\xi_i=0,\\ \sum_{i=1}^{n}\nu_i^*\xi_i^*=0\,.\end{cases}\tag{B.18}$$

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

$$\begin{aligned}\min_{\alpha_i,\beta,\beta^*} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))^T(\varphi(x_k)-\varphi(x_{\bar{l}(k)})) \\ & +\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i\beta_k+\delta_i\delta_k\beta_i^*\beta_k^*)\,\varphi(x_i)^T\varphi(x_k) \\ & +\sum_{i=1}^{n}\sum_{k=1}^{n}(\beta_i-\delta_k\beta_k^*)\,\alpha_k\varphi(x_i)^T(\varphi(x_k)-\varphi(x_{\bar{l}(k)})) \\ & -\sum_{i=1}^{n}\sum_{k=1}^{n}\beta_i^*\delta_i\beta_i\varphi(x_i)^T\varphi(x_k)-\sum_{i=1}\alpha_i(y_i-y_{i-1}) \\ & -\textstyle\sum_{i=1}^{n}\beta_i y_i+\sum_{i=1}^{n}\delta_i\beta_i^* y_i \\ \text{subject to} \quad & \begin{cases} 0\geq\alpha_i\geq\gamma, & \forall\ i=1,\ldots,n\,, \\ 0\geq\beta_i\geq\mu, & \forall\ i=1,\ldots,n\,, \\ 0\geq\beta_i^*\geq\mu, & \forall\ i=1,\ldots,n\,. \end{cases}\end{aligned} \tag{B.19}$$

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

$$\hat{u}(x_\star)=\sum_{i=1}^{n}\Big(\alpha_i(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))+(\beta_i-\delta_i\beta_i^*)\varphi(x_i)\Big)^T\varphi_p(x_\star)+b\,. \tag{B.20}$$

B.5 MODEL 3

The model formulation when including dummy observations to compare with within MODEL 2 was as follows:

MODEL 3

$$\begin{aligned}\min_{w,\epsilon,\epsilon^*,\xi,\xi^*,b,v} \quad & \frac{1}{2}w^Tw+\gamma\sum_{i=1}^{n}(\epsilon_i+\epsilon_i^*)+\mu\sum_{i=1}^{n}(\xi_i+\xi_i^*) \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_i)+b-\ell_i^Tv\geq y_i-\ell_i^Ts-\epsilon_i, & \\ \qquad\qquad \forall\ i=1,\ldots,n\,, & \\ -\delta_i(w^T\varphi(x_i)+b-r_i^T(v-s)-y_i)\geq-\epsilon_i^*, & \\ \qquad\qquad \forall\ i=1,\ldots,n\,, & \\ w^T\varphi(x_i)+b\geq y_i-\xi_i, & \forall\ i=1,\ldots,n\,, \\ -\delta_i(w^T\varphi(x_i)+b)\geq-\delta_iy_i-\xi_i^*, & \forall\ i=1,\ldots,n\,, \\ \epsilon_i\geq0, & \forall\ i=1,\ldots,n\,, \\ \epsilon_i^*\geq0, & \forall\ i=1,\ldots,n\,, \\ \xi_i\geq0, & \forall\ i=1,\ldots,n\,, \\ \xi_i^*\geq0 & \forall\ i=1,\ldots,n\,, \\ v_l-v_{l-1}\geq0, & \forall\ l=1,\ldots,k\,. \end{cases}\end{aligned} \tag{B.21}$$

The Lagrangian becomes

$$
\begin{aligned}
&\mathcal{L}(w,\epsilon,\epsilon^*,\xi,\xi^*,v,b;\alpha,\alpha^*,\beta,\beta^*,\nu,\nu^*,\eta,\eta^*,\theta) = \tfrac{1}{2}w^T w \\
&+\gamma\sum_{i=1}^{n}(\epsilon_i+\epsilon_i^*) + \mu\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{l=1}^{k}\theta_l\,(v_l - v_{l-1}) \\
&-\sum_{i=1}^{n}\alpha_i\left(w^T\varphi(x_i)+b-\ell_i^T(v-s)-y_i+\epsilon_i\right) \\
&-\sum_{i=1}^{n}\alpha_i^*\left(-\delta_i\left(w^T\varphi(x_i)+b-\mathrm{r}_i(v-s)-y_i\right)+\epsilon_i^*\right) \\
&-\sum_{i=1}^{n}\beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right) \\
&-\sum_{i=1}^{n}\beta_i^*\left(-\delta_i\left(w^T\varphi(x_i)+b-y_i\right)+\xi_i^*\right) \\
&-\sum_{i=1}^{n}\eta_i\epsilon_i - \sum_{i=1}^{n}\eta_i^*\epsilon_i^* - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\nu_i^*\xi_i^*\,.
\end{aligned}
\tag{B.22}
$$

The KKT conditions for optimality are

$$
\left\{
\begin{array}{ll}
\frac{\partial\mathcal{L}}{\partial w}=0\rightarrow w=\sum_{i=1}^{n}(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*))\varphi(x_i), & \\
\frac{\partial\mathcal{L}}{\partial b}=0\rightarrow \sum_{i=1}^{n}(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*)), & \\
\frac{\partial\mathcal{L}}{\partial\epsilon_i}=0\rightarrow \gamma=\alpha_i+\eta_i, & \forall\ i=1,\ldots,n\,, \\
\frac{\partial\mathcal{L}}{\partial\epsilon_i^*}=0\rightarrow \gamma=\alpha_i^*+\eta_i^*, & \forall\ i=1,\ldots,n\,, \\
\frac{\partial\mathcal{L}}{\partial\xi_i}=0\rightarrow \mu=\beta_i+\nu_i, & \forall\ i=1,\ldots,n\,, \\
\frac{\partial\mathcal{L}}{\partial\xi_i^*}=0\rightarrow \mu=\beta_i^*+\nu_i^*, & \forall\ i=1,\ldots,n\,, \\
\frac{\partial\mathcal{L}}{\partial v_l}=0\rightarrow \sum_{i=1}^{n}(\alpha_i\ell_{i,l}-\delta_i\alpha_i^*r_{i,l}-\theta_l+\theta_{l-1})=0, & \forall\ l=1,\ldots,k\,, \\
\alpha_i,\alpha_i^*,\beta_i,\beta_i^*,\eta_i,\eta_i^*,\xi_i,\xi_i^*\geq 0, & \forall\ i=1,\ldots,n\,, \\
\theta_l\geq 0, & \forall\ l=1,\ldots,k\,, \\
\sum_{i=1}^{n}\alpha_i\left(w^T\varphi(x_i)+b-\ell_i^T(v-z)-y_i+\epsilon_i\right)=0, & \\
\sum_{i=1}^{n}\alpha_i^*\left(-\delta_i(w^T\varphi(x_i)+b-r_i^T(v-z)-y_i)+\epsilon_i^*\right)=0, & \\
\sum_{i=1}^{n}\beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right)=0, & \\
\sum_{i=1}^{n}\beta_i^*\left(-\delta_i w^T\varphi(x_i)-b+\delta_i y_i+\xi_i^*\right)=0, & \\
\sum_{i=1}^{n}\eta_i\epsilon_i=0, & \\
\sum_{i=1}^{n}\eta_i^*\epsilon_i^*=0, & \\
\sum_{i=1}^{n}\nu_i\xi_i=0, & \\
\sum_{i=1}^{n}\nu_i^*\xi_i^*=0, & \\
\sum_{l=1}^{k}\theta_l(v_l-v_{l-1})=0\,, &
\end{array}
\right.
\tag{B.23}
$$

with $\ell_{i,l}$ and $r_{i,l}$ the lth element of the vectors ℓ_i and r_i respectively.The solution is found as

$$
\begin{aligned}
\min_{\alpha,\alpha^*,\beta,\beta^*,\theta} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i\alpha_j+\beta_i\beta_j\right)\varphi(x_i)^T\varphi(x_j) \\
& \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\left(\alpha_i^*\alpha_j^*+\beta_i^*\beta_j^*\right)\varphi(x_i)^T\varphi(x_j) \\
& +\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i\beta_j-\delta_j((\alpha_i+\beta_i)(\alpha_j^*+\beta_j^*))\right)\varphi(x_i)^T\varphi(x_j) \\
& -\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\alpha_i^*\beta_j^*\varphi(x_i)^T\varphi(x_j)-\sum_{i=1}^{n}\left(\alpha_i-\delta_i\alpha_i^*\right)y_i \\
& +\sum_{i=1}^{n}(\ell_i-\mathrm{r}_i)^T z-\sum_{i=1}^{n}\beta_i y_i+\sum_{i=1}^{n}\delta_i\beta_i^* y_i \\
\text{subject to} \quad & \begin{cases}
0\geq\alpha_i\geq\gamma, & \forall\ i=1,\ldots,n\,, \\
0\geq\alpha_i^*\geq\gamma, & \forall\ i=1,\ldots,n\,, \\
0\geq\beta_i\geq\mu, & \forall\ i=1,\ldots,n\,, \\
0\geq\beta_i^*\geq\mu, & \forall\ i=1,\ldots,n\,, \\
\displaystyle\sum_{i=1}^{n}\left(\alpha_i\ell_{i,l}-\delta_i\alpha_i^* r_{i,l}\right)-\theta_l+\theta_{l-1}, & \forall\ l=1,\ldots,k\,.
\end{cases}
\end{aligned}
\tag{B.24}
$$

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

$$
\hat{u}(x_\star)=\sum_{i=1}^{n}(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*))\varphi(x_i)^T\varphi(x_\star)+b\,. \tag{B.25}
$$

Bibliography

Ailon, N. & Mohri, M. An Efficient Reduction of Ranking to Classification. In *20th Annual Conference on Learning Theory*, pages 87–98. Omnipress (2008).

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723 (1974).

Akritas, M. G. Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring. *Annals of Statistics*, **22**(3), 1299–1327 (1994).

Ameye, L., Valentin, L., Testa, A. C., Van Holsbeke, C., Domali, E., Van Huffel, S., Vergote, I., Bourne, T., & Timmerman, D. A scoring system to differentiate malignant from benign masses in specific ultrasound-based subgroups of adnexal tumors. *Ultrasound in Obstetrics & Gynecology*, **33**(1), 92–101 (2009).

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. *Statistical Models based on Counting Processes*. Springer-Verlag, New York (1993).

Armitage, P. & Colton, T. *Survival and event history analysis*. Wiley Reference Series in Biostatistics. Wiley (2007).

Bair, E., Hastie, T., Debashis, P., & Tibshirani, R. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, **101**, 119–137 (2006).

Bair, E. & Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLoS Biology*, **2**(4), 511–522 (2004).

Barakat, N. & Diederich, J. Eclectic rule-extraction from support vector machines. *International Journal of Computational Intelligence*, **2**(1), 59–62 (2005).

Barakat, N. H. & Bradley, A. P. Rule Extraction from Support Vector Machines: A Sequential Covering Approach. *IEEE Transactions on Knowledge and Data Engineering*, **19**(6), 729–741 (2007).

—. Rule-extraction from support vector machines: A review. *Neurocomputing* (2010). In Press.

Bardou, V. J., Arpino, G., Elledge, R. M., Osborne, C. K., & Clark, G. M. Progesterone Receptor Status Significantly Improves Outcome Prediction Over Estrogen Receptor Status Alone for Adjuvant Endocrine Therapy in Two Large Breast Cancer Databases. *Journal of Clinical Oncology*, **21**(10), 1973–1979 (2003).

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., Spurr, C., Khorasani, R., Tanasijevic, M., & Middleton, B. Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality. *Journal of the American Medical Informatics Association*, **10**(6), 523–530 (2003).

Benson, J. R. & Querci della Rovere, G. Management of the axilla in women with breast cancer. *The Lancet Oncology*, **8**(4), 331 – 348 (2007).

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*, **17**(10), 1169–1186 (1998).

Biganzoli, E., Boracchi, P., & Marubini, E. A general framework for neural network models on censored survival data. *Neural Networks*, **15**, 209–218 (2002).

Biganzoli, E., Boracchiand F Ambrogi, P., & Marubini, E. Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial Intelligence in Medicine*, **37**(2), 119–130 (2006).

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press (1995).

Blamey, R. W., Ellis, I. O., Lee, S. E. P. A. H. S., Macmillan, R. D., Morgan, D. A. L., Robertson, J. F. R., Mitchell, M. J., Ball, G. R., Haybittle, J. L., & Elston, C. W. Survival of invasive breast cancer according to the Nottinghham Prognostic Index in cases diagnosed in 1990-1999. *European Journal of Cancer*, **43**, 1548–1555 (2007).

Bøvelstad, H. M. M., Nygård, S., Størvold, H. L. L., Aldrin, M., Borgan, O., Frigessi, A., & Lingjærde, O. C. C. Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**(16), 2080–2087 (2007).

Bottomley, C., Van Belle, V., Mukri, F., Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. The optimal timing of an ultrasound scan to assess the location and viability of an early pregnancy. *Human Reproduction*, **24**(8), 1811–1817 (2009).

Boyd, S. & Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge (2004).

Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), 1145–1159 (1997).

Breiman, L. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**(4), 373–384 (1995).

Breslow, N. Covariance Analysis of Censored Survival Data. *Biometrics*, **30**(1), 89–99 (1974).

Brouckaert, O., Pintens, S., Van Belle, V., Van Huffel, S., Camerlynck, E., Amant, F., Leunen, K., Smeets, A., Berteloot, P., Van Limbergen, E., Decock, J., Hendrickx, W., Weltens, C., Van den Bogaert W., Vanden Bempt I., Drijkoningen M., Paridaens R., Wildiers H., Vergote I., Christiaens, M. R., & Neven, P. Short-term outcome of primary operated early breast cancer by hormone and HER-2 receptors. *Breast Cancer Research and Treatment*, **115**(2), 349–358 (2009).

Bueno-de-Mesquita, J. M., van Harten, W. H., Retel, V. P., van't Veer, L., van Dam, F. S. A. M., Karsenberg, K., Douma, K. F. L., van Tinteren, H., Peterset, J. L., Wesseling, J., Wu, T. S., Atsma, D., Rutgers, E. J. T., Brink, G., Floore, A. N., Glas, A. M., Roumen, R. M. H., Bellot, F. E., van Krimpen, C., Rodenhuis, S., van de Vijver, M. J., & Linn, S. C. Use if 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). *The Lancet Oncology*, **8**, 1079–1087 (2007).

Bundred, N. J. Prognostic and predictive factors in breast cancer. *Cancer Treatment Reviews*, **27**, 137–142 (2001).

Buyse, M., Loi, S., van 't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., Bergh, M. D. J., Liderau, R., Ellis, P., Haaris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., & Piccart, M. J. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*, **98**(17), 1183–1192 (2006).

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

Callas, P. W., Pastides, H., & Hosmer, D. W. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American Journal of Industrial Medicine*, **33**(1), 33–47 (1998).

Campbell, H. E., Taylor, M. A., Harris, A. L., & Gray, A. M. An investigation into the performance of the Adjuvant! Online prognostic programme in early breast cancer for a cohort of patients in the United Kingdom. *British Hournal of Cancer*, **101**, 1074–1084 (2009).

Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, **25**(2), 21–30 (2008).

Candès, E. J., Wakin, M. B., & Boyd, S. Enhancing sparsity by reweighted L_1 minimization. *Journal of Fourier Analysis and Applications*, **14**(5-6), 877–905 (2008).

Carter, C. L., Allen, C., & Henson, D. E. Relation of tumor size, lymph node status, and survival in 24.740 breast cancer cases. *Cancer*, **63**, 181–187 (1989).

Chaves, A. C., Vellasco, M., & Tanscheit, R. Fuzzy rule extraction from support vector machines. In *Proceedings of the Fifth International Conference on Hybrid Intelligent Systems*, pages 386–393 (2005).

Chen, Z., Li, J., & Wei, L. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artificial Intelligence in Medicine*, **41**, 161–175 (2007).

Cheng, S. C., Wei, L. J., & Ying, Z. Predicting Survival Probabilities with Semiparametric Transformation Models. *Journal of the American Statistical Association*, **92**(437), 227–235 (1997).

Cho, H., Van Belle, V., Vandorpe, T., Wildiers, H., Leunen, K., Amant, F., Vergote, I., Berteloot, P., Smeets, A., Limbergen, E. V., Weltens, C., Paridaens, R., Huffel, S. V., Christiaens, M. R., & Neven, P. Prognostic significance of nodal and PgR status on early relapse in operable HER-2 positive breast cancer from the pre-trastuzumab era. *Cancer Research*, **69**(24), 837 (2009).

Cho, H., Van Belle, V., Wildiers, H., Paridaens, R., Amant, F., Van Limbergen, E., Moerman, P., Smeets, A., Vergote, I., & Neven, P. Factors predicting a pathological complete response following neo-adjuvant chemotherapy for breast cancer. *European Journal of Cancer*, **8**(3), 66 (2010).

Choong, S., Rombauts, L., Ugoni, A., & Meagher, S. Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions. *Ultrasound in Obstetrics and Gynecology*, **22**(6), 571–577 (2003).

Chu, W. & Ghahramani, Z. Gaussian Processes for Ordinal Regression. *Journal of Machine Learning Research*, **6**, 1019–1041 (2005).

Chu, W. & Keerthi, S. S. New approaches to support vector ordinal regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 145–152 (2005).

Chun, F. K., Graefen, M., Briganti, A., Gallina, A., Hopp, J., Kattan, M. W., Huland, H., & Karakiewicz, P. I. Initial Biopsy Outcome Prediction–Head-to-Head Comparison of a Logistic Regression-Based Nomogram versus Artificial Neural Network. *European Urology*, **51**(5), 1236–1243 (2007a).

Chun, F. K., Karakiewicz, P. I., Briganti, A., Walz, J., Kattan, M. W., Huland, H., & Graefen, M. A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. *BJU International*, **99**(4), 794–800 (2007b).

Cianfrocca, M. & Goldstein, L. J. Prognostic and Predictive Factors in Early-Stage Breast Cancer. *The Oncologist*, **9**(6), 606–616 (2004).

Clémençon, S., Lugosi, G., Vayatis, N., Aurer, P., & Meir, R. Ranking and scoring using empirical risk minimization. In *Proceedings of the 18th Annual Conference in Learning Theory*, pages 1–15. Bertinoro (Italy) (2005).

Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *American Statistical Association*, **74**(386), 829–836 (1979).

Cobleigh, M. A., Vogel, C. L., Tripathy, D., Robert, N. J., Scholl, S., Fehrenbacher, L., Wolter, J. M., Paton, V., Shak, S., Lieberman, G., & Slamon, D. J. Multinational Study of the Efficacy and Safety of Humanized Anti-HER2 Monoclonal Antibody in Women Who Have HER2-Overexpressing Metastatic Breast Cancer That Has Progressed After Chemotherapy for Metastatic Disease. *Journal of Clinical Oncology*, **17**(9), 2639–2648 (1999).

Collett, K., Skjærven, R., & Mæhle, B. O. The prognostic contribution of estrogen and progesterone receptor status to a modified version of the Nottingham Prognostic Index. *Breast Cancer Research and Treatment*, **48**(1), 1–9 (1998).

Cox, D. R. Regression Models and Life-Tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **34**(2), 187–220 (1972).

Cox, D. R. Partial Likelihood. *Biometrika*, **62**(2), 269–276 (1975).

Cruz, J. A. & Wishart, D. S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, **2**, 59–77 (2006).

d' Ocagne, M. *Traité de Nomographie*. Gauthier-Villars, Paris (1899).

Dabrowska, D. M. & Doksum, K. A. Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, **15**(1), 1–23 (1988).

Daemen, A. & De Moor, B. Development of a kernel function for clinical data. In *Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 5913–5917. IEEE, Piscataway (2009).

Danko, M. E., Bennet, K. L., Zhai, J., Marks, J. R., & Olson, J. A. Improved staging in node-positive breast cancer patients using lymph node ratio: restuls in 1788 patients with long-term follow-up. *Journal fo the American College of Surgeons*, **210**(5), 797–805 (2010).

de Boor, C. *A Practical Guide to Splines*. Springer, Berlin (1978).

Dierickx, I., Van Belle, V., Vandorpe, T., Wildiers, H., Leunen, K., Peeters, S., Moerman, P., Deraedt, K., Christiaens, M. R., & Neven, P. A prognostic model for breast cancer-related events in primary operated invasive lobular breast cancers from one center. *European Journal of Cancer*, **8**(3), 94 (2010).

Edwards, A., Clarke, L., Piessens, S., Graham, E., & Shekleton, P. Acoustic streaming: a new technique for assessing adnexal cysts. *Ultrasound in Obstetrics and Gynecology*, **22**(1), 74–78 (2003).

Efron, B. & Tibshirani, R. Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*, **92**(438), 548–560 (1997).

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

Elandt-Johnson, R. C. & Johnson, N. L. *Survival models and data analysis*. John Wiley & Sons, Inc. (1980).

Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, **19**(5), 403–410 (1991).

Emerson, S. S. & Banks, P. L. C. *Case Studies in Biometry*, chapter Interpretation of a leukemia trial stopped early, pages 275–299. Wiley-Interscience, New York (1994).

Evers, L. & Messow, C. M. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, **24**(14), 1632–1638 (2008).

Falco, P., Milano, V., Pilu, G., David, C., Grisolia, G., Rizzo, N., & Bovicelli, L. Sonography of pregnancies with first-trimester bleeding and a viable embryo: a study of prognostic indicators by logistic regression analysis. *Ultrasound in Obstetrics and Gynecology*, **7**(3), 165–169 (1996).

Faraggi, D. & Simon, R. A neural network model for survival data. *Statistics in Medicine*, **14**, 73–82 (1995).

Ferlay, J., Autier, P., Boniol, M., Heanue, M., Colombet, M., & Boyle, P. Estimates of the cancer incidence and mortality in Europe in 2006. *Annals of Oncology*, **18**(3), 581–592 (2007).

Fisher, B., Bauer, M., Wickerham, L., Redmong, C. K., & Fisher, E. R. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer*, **52**(9), 1551–1557 (1983).

Fitzgibbons, P. L., Page, D. L., Weaver, D., Thor, A. D., Allred, D. C., Clark, G. M., Ruby, S. G., O'Malley, F., Simpson, J. F., Connolly, J. L., Hayes, D. F., Edge, S. B., Lichter, A., & Schnitt, S. J. Prognostic factors in breast cancer. *Archives of Pathology and Laboratory Medicine*, **124**, 966–978 (2000).

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, **4**(6), 933–969 (2004).

Fung, G., Sandilya, S., & Rao, R. B. Rule extraction from linear support vector machines. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data mining*, pages 32–40. ACM, New York, NY, USA (2005).

Fürnkranz, J. & Hüllermeier, E. Pairwise preference learning and ranking. In *Proceedings of the European Conference on Machine Learning.*, pages 145–156. Springer-Verlag, Cavtat-Dubrovnik, Croatia (2003).

Galea, M., Blamey, R., Elston, C., & Ellis, I. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment*, **22**(3), 207–219 (1992).

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., & De Moor, B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, **22**(14), e184–190 (2006).

Gijbels, I. Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 178–188 (2010).

Goeman, J. J. L1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal*, **52**(1), 70–84 (2010).

Gómez, G., Calle, M. L., Oller, R., & Langohr, K. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, **9**(4), 259–297 (2009).

Green, M. S. & Symons, M. J. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases*, **36**(10), 715 – 723 (1983).

Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1), 29–36 (1982).

Harrell, F. *Regression Modeling Strategies. With applications to linear models, logistic regression, and survival analysis*. Springer (2001).

Harrell, F., Klee, K., Califf, R., Pryor, D., & Rosati, R. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, **3**(2), 143–152 (1984).

Harrell, F., Lee, K. L., & Pollock, B. G. Regression models in clinical studies: Determining Relationships between predictors and response. *Journal of the National Cancer Institute*, **80** (1988).

Hastie, T. & Tibshirani, R. Generalized Additive Models. *Statistical Science*, **1**, 297–318 (1986).

—. *Generalized additive models*. Chapman and Hall (1990).

Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning*. Springer-Verlag (2001).

Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholson, R. I., & Griffiths, K. A prognostic index in primary breast cancer. *British Journal of Cancer*, **45**(3), 361–366 (1982).

Heagerty, P., Lumley, T., & Pepe, M. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, **56**(2), 337–344 (2000).

Herbrich, R., Graepel, T., Bollmann-Sdorra, P., & Obermayer, K. Learning Preference Relations for Information Retrieval. *Americal Association for Artificial Intelligence*, pages 1–4 (1998).

Herbrich, R., Graepel, T., & Obermayer, K. Large Margin Rank Boundaries for Ordinal Regression. *Advances in Large Margin Classifiers*, pages 115–132 (2000).

Hillaby, K., Aslam, N., Salim, R., Lawrence, A., Raju, K. S., & Jurkovic, D. The value of detection of normal ovarian tissue (the ovarian crescent sign) in the differential diagnosis of adnexal masses. *Ultrasound in Obstetrics and Gynecology*, **23**(1), 63–67 (2004).

Hlupić, L., Jakić-Razumović, J., Božikov, J., Ćorić, M., Belev, B., & Vrbanec, D. Prognostic value of different factors in breast carcinoma. *Tumori*, **90**, 112–119 (2004).

Hosmer, D. W. & Lemeshow, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Interscience, New York, NY, USA (1999).

—. *Applied Logistic Regression*. Wiley-Interscience, New York, NY, USA, second edition (2000).

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

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, **60**, 271–293 (1998).

Jacobs, I., Oram, D., Fairbanks, J., Turner, J., Frost, C., & Grudzinskas, J. G. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *BJOG: An International Journal of Obstetrics and Gynaecology*, **97**(10), 922–929 (1990).

Janes, H., Pepe, M. S., & Gu, W. Assessing the Value of Risk Predictions by Using Risk Stratification Tables. *Annals of Internal Medicine*, **149**(10), 751–760 (2008).

Kakha, E. A., El-Sayed, M. E., Lee, A. H. S., Elston, C. W., Grainge, M. J., Hodi, Z., Blamey, R. W., & Ellis, I. O. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *Journal of Clinical Oncology*, **26**(19), 3153–3458 (2008).

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

Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481 (1958).

Karakiewicz, P. I., Benayoun, S., Kattan, M. W., Perrotte, P., Scardino, L. V. P. T., Cagiannos, I., Heinzer, H., Tanguay, S., Aprikian, A. G., Huland, H., & Graefen, M. Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. *Journal of Urology*, **173**(6), 1930–1934 (2005).

Kattan, M. W. Comparison of Cox regression with other methods for determining prediction models and nomograms. *Journal of Urology*, **170**(6), S6–S10 (2003a).

—. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. *Current Opinion in Urology*, **12**(2), 111–116 (2003b).

—. When and how to use informatics tools in caring for urologic patients. *Nature Clinical Practice Urology*, **2**, 183–190 (2005).

Kattan, M. W., Ishida, H., Scardino, P. T., & Beck, J. R. *Applying a neural network to prostate cancer survival data*, chapter Intelligent Data Analysis in Medicine and Pharmacology, pages 295–306. Kluwer, Boston (1997).

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Britisch Medical Journal*, **330**, 765–773 (2005).

Khan, F. M. & Zubek, V. B. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In Giannotti, F., Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., & Wu, X., editors, *Proceedings of the eight IEEE International Conference on Data Mining*. IEEE computer society, California (2008).

Koenker, R. & Geling, O. Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. *Journal of the American Statistical Association*, **96**, 458–468 (2001).

Kong, G., Xu, D., & Yang, J. Clinical decision support systems: a review of knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, **1**(2), 159–167 (2008).

Lamy, P. J., 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 Research and Treatment*, **76**(1), 65–71 (2002).

Lasko, T. A., abd K H Zou, J. G. B., & Ohno-Machado, L. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, **38** (2005).

Lawless, J. F. *Survival and event history analysis*, chapter Parametric models in survival analysis, pages 345–355. Wiley reference series in biostatistics. Wiley, West Sussex, England (2006).

Leblanc, M. & Crowley, J. Adaptive regression splines in the Cox model. *Biometrics*, **55**(1), 204–213 (1999).

Ledley, R. S. & Lusted, L. B. Computers in Medical Data. *Operations Research*, **8**(3), 299–310 (1960).

Lindsey, J. C. & Ryan, L. M. Tutorial in biostatistics: Methods for interval-censored data. *Statistics in Medicine*, **17**, 219–238 (1998).

Lisboa, P., Wong, H., Harris, P., & 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*, **28**(1), 1–25 (2003).

Liu, R., Wang, X., Chen, G. Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., & Clarke, M. F. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *The New England Journal of Medicine*, **356**(3), 217–226 (2007).

Mabry, H. & Giuliano, A. E. Sentinel node mapping for breast cancer: progress to date and prospects for the future. *Surgical Oncology Clinics of North America*, **16**(1), 55–70 (2007).

MacKay, D. J. C. A practical Bayesian framework for backpropagation networks. *Neural Computation*, **4**(3), 448–472 (1992).

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**(3), 163–170 (1966).

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, **183**(3), 1466 – 1476 (2007).

Martens, H. & Næs, T. *Multivariate Calibration*. New York: John Wiley & Sons Inc. (1989).

Martin, K. J., Patrick, D. R., Bissel, M. J., & Fournier, M. V. Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome acreoos independen datasets. *PLoS ONE*, **3**(8), e2994 (2008).

Marty, M., Cognetti, F., Maraninchi, D., Snyder, R., Mauriac, L., Tubiana-Hulin, M., Chan, S., Grimes, D., Antón, A., Lluch, A., Kennedy, J., O'Byrne, K., Conte, P., Green, M., Ward, C., Mayne, K., & Extra, J. M. Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 study group. *Journal of Clinical Oncology*, **23**(23), 4265–4274 (2005).

McCulloch, W. S. & Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133 (1943).

Ménard, S., Fortis, S., Castiglioni, F., Agresti, R., & Balsari, A. HER2 as a prognostic factor in breast cancer. *Oncology*, **61**(S1), 67–72 (2001).

Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, **209**, 415–446 (1909).

Michiels, S., Koscielny, S., & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, **365**(9458), 488 – 492 (2005).

Miller, R. *Survival Anlysis*. John Wiley & Sons (1981).

Miller, R. Medical diagnostic decision support systems-past, present and future. *Journal of the American Medical Informatics Association*, **1**, 8–27 (1994).

Musen, M. A., Shahar, Y., & Shortliffe, E. H. *Biomedial Informatics: computer applications in health care and biomedicine*, chapter Clinical decision making, pages 698–736. Springer, New York (2006).

Navia-Vazquez, A., Perez-Cruz, F., Artes-Rodriguez, A., & Figueiras-Vidal, A. R. Weighted least squares training of support vector classifiers leading to compact and adaptive schemes. *IEEE Transactions on Neural Networks*, **12**(5), 1047 –1059 (2001).

Neven, P., Van Belle, V., Brouckaert, O., Pintens, S., Paridaens, R., Christiaens, M.-R., Deraedt, K., & Moerman, P. Are gene signatures better than traditional clinical factors? *The Lancet Oncology*, **9**, 197–198 (2008).

Nguyen, D. V. & Rocke, D. M. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**(12), 1625–1632 (2002).

Núñez, H., Angulo, C., & Català, S. Rule-based learning systems for support vector machines. *Neural Processing Letters*, **24**, 1–19 (2006).

Nygård, S., Borgan, O., Lingjærde, O., & Størvold, H. Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, **14**(2), 179–195 (2008).

Olivotto, I. A., Bajdik, C. D., Ravdin, P. M., Speers, C. H., Coldman, A. J., Norris, B. D., Davis, G. J., Chia, S., & Gelmon, K. A. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *Journal of Clinical Oncology*, **23**(12), 2716–2725 (2005).

Osheroff, J. A. *Improving medication use and outcomes with clinical decision support: a step-by-step guide*. Healthcare Information and Management Systems Society, Chicago, IL (2009).

Park, P. J., Tian, L., & Kohane, I. S. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**(suppl1), S120–127 (2002).

Parmar, K. B. & Machin, D. *Survival Analysis: A Practical Approach*. Wiley (1995).

Pelckmans, K., De Brabanter, J., Suykens, J. A. K., & De Moor, B. Convex Clustering Shrinkage. In *Workshop on Statistics and Optimization of Clustering Workshop (PASCAL)*. London, UK (2005a).

Pelckmans, K., Espinoza, M., De Brabanter, J., Suykens, J. A. K., & De Moor, B. Primal-Dual Monotone Kernel Regression. *Neural Processing Letters*, **22**(2), 171–182 (2005b).

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

Pencina, M., d'Agostino, R., Sr, d'Agostino, R., Jr, & Vasan, R. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, **27**, 157–172 (2008).

Pepe, M. S. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, New York (2003).

Percell, G. P. What makes a good clinical decision support system. *Britisch Medical Journal*, **330**, 740–741 (2005).

Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C., Cameron, D., Dowsett, M., Barrios, C. H., STeger, G., Hung, C. S., Andersson, M., Inbar, M., Lichinitser, M., Láng, I., Nitz, U., Iwata, H., Thomssen, C., Lohrisch, C., Suter, T., Rüschoff, J., Suto, T., Greatorex, V., Ward, C., Straehle, C., McFadden, E., Dolci, M. S., & Gelber, R. D. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England Journal of Medicine*, **353**(16), 1659–1672 (2005).

Pichon, M. F., Pallud, C., Brunet, M., & Milgrom, E. Relationship of presence of progesterone receptors to prognosis in early breast cancer. *Cancer Research*, **40**, 3357–3360 (1980).

Prentice, R. L. A Log Gamma Model and Its Maximum Likelihood Estimation. *Biometrika*, **61**(3), 539–544 (1974).

Rampaul, R. S., Pinder, S. E., Elston, C. W., & Ellis, I. O. Prognostic and predictive factors in primary breast cancer and their role in patient management: The Nottingham Breast Team. *European Journal of Surgical Oncology*, **27**, 229–238 (2001).

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., & Parker, H. L. Computer program to assist in making decision about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, **19**(4), 980–991 (2001).

Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge (1996).

Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Jr, C. E. G., Davidson, N. E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P. A., Swain, S. M., Pisansky, T. M., Fehrenbacher, L., Kutteh, L. A., Vogel, V. G., Vissher, D. W., Yothers, G., Jenkins, R. B., Brown, A. M., Dakhil, S. R., Mamounas, E. P., Lingle, W. L., Klein, P. M., Ingle, J. N., & Wolmark, N. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *The New England Journal of Medicine*, **353**(16), 1673–1684 (2005).

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, R. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., & Yang, L. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *The New England Journal of Medicine*, **346**(25), 1937–1947 (2002).

Rudin, L. I., Osher, S., & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, **60**(1-4), 259–268 (1992).

Sanchez-Carbayo, M., Socci, N. D., Lozano, J., Saint, F., & Cordon-Cardo, C. Defining Molecular Profiles of Poor Outcome in Patients With Invasive Bladder Cancer Using Oligonucleotide Microarrays. *Journal of Clinical Oncology*, **24**(5), 778–789 (2006).

Sauerbrei, W. & Royston, P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society Series A*, **162**(1), 71–94 (1999).

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

Schumacher, M., Basert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. L. A., & 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*, **12** (1994).

Shariat, S. F., Karakiewicz, P. I., Suardi, N., & Kattan, M. W. Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. *Clinical Cancer Research*, **14**, 4400–4407 (2008).

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

Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A., & Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antogonism of this interaction by tamoxifen. *Cell*, **95**, 927–937 (1998).

Shivaswamy, P. K., Chu, W., & Jansche, M. A Support Vector Approach to Censored Targets. In *Proceedings of the Seventh IEEE International Conference on Data Mining*, pages 655–660. IEEE Computer Society, California (2007).

Siddique, J. & Harel, O. MIDAS: A SAS Macro for Multiple Imputation Using Distance-Aided Selection of Donors. *Journal of Statistical Software*, **29**(9), 1–18 (2009).

Singlatary, S. E., Allred, C., Eshley, P., Bassett, L. W., Berry, D., Bland, K. I., Borgen, P. I., Clark, G., Edge, S. B., Hayes, D. F., Hughes, L. L., Hutter, R. V. P., Morrow, M., Page, D. L., Recht, A., Theriault, R. L., Thor, A., Weaver, D. L., Wieand, H. S., & Greene, F. L. Revision of the American Joint Committe on Cancer Staging System for Breast Cancer. *Journal of Clinical Oncology*, **20**(17), 3628–3636 (2002).

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., & Norton, L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *The New England Journal of Medicine*, **344**(11), 783–792 (2001).

Smaletz, O., Scher, H. I., Small, E. J., Verbel, D. A., McMillan, A., Regan, K., Kelly, W. K., & Kattan, M. W. Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. *Journal of Clinical Oncology*, **20**(19), 3972–3982 (2002).

Smith, I., Procter, M., Gelber, R. D., Guillaume, S., Feuereislova, A., Dowsett, M., Goldhirsch, A., Untch, M., Mariani, G., Baselga, J., Kaufman, M., Cameron, D., Bell, R., Bergh, J., Coleman, R., Wardley, A., Harbeck, N., Lopez, R. I., Mallmann, P., Gelmon, J., Wilcken, N., Wist, E., Sánchez Rovira, P., & Piccart-Gebhart, M. J. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *The Lancet*, **369**(9555), 29–36 (2007).

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., 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*, **100**(14), 8418–8423 (2003).

Specht, M. C., Kattan, M. W., Gonen, M., Fey, J., & Van Zee, K. J. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. *Annals of Surgical Oncology*, **12**(8), 654–659 (2005).

Steyerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer, New York (2009).

Suen, D. & Chow, L. W. Prognostic contribution of the HER-2 oncogene overexpression to the Nottingham Prognostic Index in breast cancer. *Biomedecine and Pharmacotherapy*, **60**(6), 293 – 297 (2006).

Sullivan, L. M., Massaro, J. M., & D'Agostino, R. B. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Statistics in Medicine*, **23**(10), 1631–1660 (2004).

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. *Least Squares Support Vector Machines*. World Scientific, Singapore (2002).

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

Suykens, J. A. K., Vandewalle, J., & De Moor, B. *Artificial Neural Networks for Modelling and control of Non-Linear systems*. Kluwer Academic Publishers, Boston (1996).

Swets, J. A. Measuring the accuracy of diagnostic systems. *Science*, **240** (1988).

Testa, A. C., Timmerman, D., Van Belle, V., Fruscella, E., Holsbeke, C. V., Savelli, L., Ferrazzi, E., Leone, F. P. G., Marret, H., Tranquart, F., Exacoustos, C., Nazzaro, G., Bokor, D., Magri, F., Van Huffel, S., Ferrandina, G., & Valentin, L. Intravenous contrast ultrasound examination using contrast-tuned imaging (CnTI) and the contrast medium SonoVue for discrimination between benign and malignant adnexal masses with solid components. *Ultrasound in Obstetrics and Gynecology*, **34**(6), 699–710 (2009).

Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2 edition (2000).

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, series B*, **58**(1), 267–288 (1996).

—. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**(4), 267–288 (1997).

Timmerman, D., Ameye, L., Fischerove, D., Epstein, E., Melis, G. B., Guerriero, S., Van Holsbeke, C., Savelli, L., Fruscio, R., Lissoni, A. A., Testa, A. C., Veldman, J., Vergote, I., Van Huffel, S., Bourne, T., & Valentin, L. Simple ultrasound rules to distinguish between benign and malignant adnexal masses prior to surgery: a prospective validation by the IOTA group. Technical report, University Hospitals Leuven, Department of Obstetrics and Gynaecology (2010a).

Timmerman, D., Testa, A. C., Bourne, T., Ameye, L., Jurkovic, D., Van Holsbeke, C., Paladini, D., Van Calster, B., Vergote, I., Van Huffel, S., & Valentin, L. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound in Obstetrics & Gynecology*, **31**(6), 681–690 (2008).

Timmerman, D., Testa, A. C., Bourne, T., Ferrazzi, E., Ameye, L., Konstantinovic, M. L., Van Calster, B., Collins, W. P., Vergote, I., Van Huffel, S., & Valentin, L. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian tumor Analysis Group. *Journal of Clinical Oncology*, **23**(32), 8794–8801 (2005).

Timmerman, D., Valentin, L., Bourne, T. H., Collins, W. P., Verrelst, H., & Vergote, I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound in Obstetrics and Gynecology*, **16**(5), 500–505 (2000).

Timmerman, D., Van Calster, B., Testa, A., Guerriero, S., Fischerova, D., Lissoni, A., Van Holsbeke, C., Fruscio, R., Czekierdowski, A., Jurkovic, D., Savelli, L., Vergote, I., Bourne, T., Van Huffel, S., & Valentin, L. Ovarian cancer prediction in adnexal masses using ultrasound based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound in Obstetrics & Gynecology*, **36**(2), 226–234 (2010b).

Van Belle, V., Cho, H., Wildiers, H., Van Huffel, S., Paridaens, R., Weltens, C., Berteloot, P., Janssen, H., Christiaens, M. R., & Neven, P. Factors predicting disease free survival following neoadjuvant chemotherapy for breast cancer. *European Journal of Cancer*, **8**(3), 101 (2010a).

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

—. Survival SVM: a Practical Scalable Algorithm. In Verleysen, M., editor, *Proceedings of the European Symposium on Artificial Neural Networks*, pages 89–94. d-side, Evere (2008).

—. Learning Transformation Models for Ranking and Survival Analysis. Technical report, 09-45, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2009). Submitted for publication.

—. MINLIP: Efficient Learning of Transformation Models. In Alippi, C., Polycarpou, M. M., Panayiotou, C., & Ellinas, G., editors, *Proceedings of the International Conference on Artificial Neural Networks*, pages 60–69. Springer-Verlag Berlin, Limassol (Cyprus) (2009a).

—. On the use of a clinical kernel in survival analysis. In Verleysen, M., editor, *Proceedings of the European Symposium on Artificial Neural Networks*, pages 451–456 (2010b).

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Technical Report 09-235, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2009b). Submitted for publication.

—. Improved Performance on High-Dimensional Survival Data using Survival-SVM. Technical report, 10-59, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2010c). Accepted for publication in Bioinformatics.

Van Belle, V., Van Calster, B., Brouckaert, O., Vanden Bempt, I., Pintens, S., Paridaens, R., Amant, F., Leunen, K., Smeets, A., Drijkoningen, R., Wildiers, H., Christiaens, M. R., Vergote, I., Van Huffel, S., & Neven, P. Qualitative assessment of the progesterone receptor and HER-2 improve the Nottingham Prognostic Index for short term breast cancer prognosis. *Journal of Clinical Oncology*, **28**(27), 4129–4134 (2010d).

Van Belle, V., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., Neven, P., Van Huffel, S., Suykens, J. A. K., & Boyd, S. Bridging clinical decision making and flexible mathematical modelling: the Interval Coded Score index. Technical report, 10-170, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2010). Submitted for publication.

Van Belle, V., Van Calster, B., Wildiers, H., Van Huffel, S., & Neven, P. Lymph node ratio is a better prognostic predictor for disease free survival than the number of positive lymph nodes in Node-Positive Breast Cancer. *Journal of Clinical Oncology*, **27**(30), 150–151 (2009c).

Van Calster, B., Timmerman, D., Lu, C., Suykens, J. A. K., Valentin, L., Van Holsbeke, C., Amant, F., Vergote, I., & Van Huffel, S. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. *Ultrasound in Obstetrics & Gynecology*, **29**(5), 496–504 (2007a).

Van Calster, B., Timmerman, D., Nabney, I., Valentin, L., Testa, A., Van Holsbeke, C., Vergote, I., & Van Huffel, S. Using Bayesian neural networks with ARD input selection to detect malignant adnexal masses prior to surgery. *Neural Computing & Applications*, **17**, 489–500 (2007b).

Van Calster, B., Van Belle, V., Condous, G., Bourne, T., Timmerman, D., & Van Huffel, S. Multi-class AUC metrics and weighted alternatives. In *Proceedings of the 21st International Joint Conference on Neural Networks*, pages 1391–1397. Hong Kong, China (2008).

van de Vijver, M. J., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Pattish, M., Atsma, D., Witteveen, A., Glas, A., Delaheye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., & Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, **347**(25), 1999–2009 (2002).

Van Holsbeke, C., Van Belle, V., Leone, F. P. G., Guerriero, S., Paladini, D., Melis, G. B., Greggi, S., Fischerova, D., De Jonge, E., Neven, P., Bourne, T., Valentin, L., Van Huffel, S., & Timmerman, D. Prospective external validation of the ovarian crescent sign as a single ultrasound parameter to distinguish between benign and malignant adnexal pathology. *Ultrasound in Obstetrics and Gynecology*, **36**(1), 81–87 (2010a).

Van Holsbeke, C., Van Calster, B., Bourne, T., Melis, G. B., Testa, A. C., Guerriero, S., Fruscio, R., Lissoni, A. A., Czerkierdowski, A., Savelli, L., Van Huffel, S., Valentin, L., & Timmerman, D. External validation of diagnostic models to predict the risk of malignancy in adnexal masses. Technical report, 10-171, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2010b).

Van Holsbeke, C., Van Calster, B., Testa, A. C., Domali, E., Lu, C., Van Huffel, S., Valentin, L., & Timmerman, D. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results form the international ovarian tumor analysis study. *Clinical Cancer Research*, **15**(2), 684–691 (2009).

Van Holsbeke, C., Zhang, J., Van Belle, V., Paladini, D., Guerriero, S., Czekierdowski, A., Muggah, H., Ombelet, W., Jurkovic, D., Testa, A. C., Valentin, L., Van Huffel, S., Bourne, T., & Timmerman, D. Acoustic streaming cannot discriminate reliably between endometriomas and other types of adnexal lesion: a multicenter study of 633 adnexal masses. *Ultrasound in Obstetrics and Gynecology*, **35**(3), 349–353 (2010c).

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., & Wessels, L. F. A. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, **25**(18), 3201–3216 (2006).

van 't Veer, L., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536 (2002).

Van Zee, K. J., Manasseh, D. M., Bevilacqua, J. L., Boolbol, S. K., Fey, J. V., Tan, L. K., Borgen, P. I., Cody, H. S., & Kattan, M. W. A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. *Annals of Surgical Oncology*, **10**(10), 1140–1151 (2003).

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

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

Vinh-Hung, V., Tai, P., Storme, G., Cserni, G., Wallace, A. M., Ridder, M. D., Vlastos, G., Promish, D., & Royce, M. A ratio-based nodal prognostic index for breast cancer. *Breast Cancer Research and Treatment*, **100**, S220–S221 (2006).

Vinh-Hung, V., Verkooijen, H. M., Fioretta, G., Neyroud-Caspar, I., Rapiti, E., Vlastos, G., Deglise, C., Usel, M., Lutz, J., & Bouchardy, C. Lymph node ratio as an alternative to pN staging in node-positive breast cancer. *Journal of Clinical Oncology*, **27** (2009).

Vinh-Hung, V., Verschraegen, C., Promish, D. I., Cserni, G., de Steene, J. V., Tai, P., Vlastos, G., Voordeckers, M., Storme, G., & Royce, M. Ratios of involved nodes in early breast cancer. *Breast Cancer Research*, **6**, R680–R688 (2004).

Vogel, C. L., Cobleigh, M. A., Tripathy, D., Gutheil, J. C., Harris, L. N., Fehrenbacher, L., Slamon, D. J., Murphy, M., Novotny, W. F., Burchmore, M., Shak, S., Stewart, S. J., & Press, M. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *Journal of Clinical Oncology*, **20**(3), 719–726 (2002).

Wagner, S. S. & Altmann, S. A. What time do the baboons come down from the trees? (An estimation problem). *Biometrics*, **29**, 623–635 (1973).

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, A. M. L., Atkins, D., & Foekens, J. A. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, **365**, 671–679 (2005).

Wears, R. L. & Berg, M. Computer technology and clinical work: still waiting for Godot. *Journal of the Americal Medical Association*, **293**(10), 1261–1263 (2005).

Woodward, W. A., Vinh-Hung, V., Ueno, N. T., Cheng, Y. C., Royce, M., Tai, P., Vlastos, G., Wallace, A. M., Hortobagyi, G. N., & Nieto, Y. Prognostic value of nodal ratios in node-positive breast cancer. *Journal of Clinical Oncology*, **24**(18), 2910–2916 (2006).

Wyatt, J. C. & Altman, D. G. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, **311**, 1539–1541 (1995).

Xavier de Souza, S., Suykens, J. A. K., Vandewalle, J., & Bolle, D. Coupled Simulated Annealing. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, **40**(2), 320–335 (2010).

Yuan, M. & Lin, L. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 143–161 (2007).

Zhang, Y., Su, H., Jia, T., & Chu, J. Rule extraction from trained support vector machines. In *Proceedings of the Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference PAKDD2005*, pages 61–70 (2005).

Curriculum vitae

Vanya Van Belle was born in Brussels on October 27, 1983. In 2006, she obtained the master of science for electrical engineering in electronics, option biomedical techniques, magna cum laude from the Katholieke universiteit Leuven, Belgium. In October 2006, she started her PhD at the Department of Electrical Engineering (ESAT) of the Katholieke Universiteit Leuven under the supervision of prof. dr. ir. Sabine Van Huffel and prof. dr. ir. Johan A.K. Suykens. She received a grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Her research interests are kernel based methods, survival analysis, development of classifiers and their application within clinical practice. The goal of her research is to integrate advanced mathematical methods in a clinical environment.

Publication list

Papers in international journals

Published and accepted papers

Van Belle, V., Van Calster, B., Wildiers, H., Van Huffel, S., & Neven, P. Lymph node ratio is a better prognostic predictor for disease free survival than the number of positive lymph nodes in Node-Positive Breast Cancer. *Journal of Clinical Oncology*, **27**(30), 150–151 (2009). (IF 17.157)

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. Additive survival least squares support vector machines. *Statistics in Medicine*, **29**(2), 296 – 308 (2010). (IF 2.111)

Van Belle, V., Van Calster, B., Brouckaert, O., Vanden Bempt, I., Pintens, S., Paridaens, R., Amant, F., Leunen, K., Smeets, A., Drijkoningen, R., Wildiers, H., Christiaens, M. R., Vergote, I., Van Huffel, S., & Neven, P. Qualitative assessment of the progesterone receptor and HER-2 improve the Nottingham Prognostic Index for short term breast cancer prognosis. *Journal of Clinical Oncology*, **28**(27), 4129–4134 (2010). (IF 17.157)

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. Improved Performance on High-Dimensional Survival Data using Survival-SVM. Technical report, 10-59, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2010). Accepted for publication in *Bioinformatics*. (IF 4.926)

Neven, P., **Van Belle, V.**, Brouckaert, O., Pintens, S., Paridaens, R., Christiaens, M.-R., Deraedt, K., & Moerman, P. Are gene signatures better than traditional clinical factors? *The Lancet Oncology*, **9**, 197–198 (2008). (IF 13.283)

Decock, J., Hendrickx, W., Vanleeuw, U., **Van Belle, V.**, Van Huffel, S., Christiaens, M.R., Ye, S., & Paridaens, R. Plasma MMP1 and MMP8 expression in breast cancer : protective role of MMP8 against lymph node metastatis, *Britisch Medical Journal Cancer*, **8**, pp. 77–85 (2008). (IF 3.087)

Neven, P., Brouckaert, O., **Van Belle, V.**, Vanden Bempt, I., Hendrickx, W., Cho, H., Deraedt, K., Van Calster, B., Van Huffel, S., Moerman, P., Amant, F., Leunen, K., Smeets, A., Wildiers, H., Paridaens, R., Vergote, I., & Christiaens, M.R. In early-stage breast cancer, the estrogen receptor interacts with correlation between human epidermal growth factor receptor 2 status and age at diagnosis, tumor grade, and lymph node involvement. *Journal of Clinical Oncology*, **26**(10), 1768–1769 (2008). (IF 17.157)

Neven, P., Van Calster, B., Vanden Bempt, I., Van Huffel, S., **Van Belle, V.**, Hendrickx, W., Decock, J., Wildiers, H., Paridaens, R., Amant, F., Leunen, K., Berteloot, P., Timmerman, D., Van Limbergen, E., Weltens, C., Van den Bogaert, W., Smeets, A., Vergote, I., Christiaens, M.R., & Drijkoningen, M. Age interacts with the expression of steroid and HER-2 receptors in operable invasive breast cancer. *Breast Cancer Research and Treatment*, **110**(1), 153–159 (2008). (IF 5.684)

Brouckaert, O., Pintens, S., **Van Belle, V.**, Van Huffel, S., Camerlynck, E., Amant, F., Leunen, K., Smeets, A., Berteloot, P., Van Limbergen, E., Decock, J., Hendrickx, W., Weltens, C., Van den Bogaert W., Vanden Bempt I., Drijkoningen M., Paridaens R., Wildiers H., Vergote I., Christiaens, M. R., & Neven, P. Short-term outcome of primary operated early breast cancer by hormone and HER-2 receptors. *Breast Cancer Research and Treatment*, **115**(2), 349–358 (2009). (IF 5.684)

Pintens, S., Neven, P., Drijkoningen, M., **Van Belle, V.**, Christiaens, M.R., Smeets, A., Wildiers, H., & Vanden Bempt, I. Triple negative breast cancer: a study from the point of view of basal CK5/6 and HER-1. *Journal of Clinical Pathology*, **62**(7), 624–628 (2009). (IF 2.342)

Bottomley, C., **Van Belle, V.**, Mukri, F., Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. The optimal timing of an ultrasound scan to assess the location and viability of an early pregnancy. *Human Reproduction*, **24**(8), 1811–1817 (2009). (IF 3.773)

Testa, A. C., Timmerman, D., **Van Belle, V.**, Fruscella, E., Holsbeke, C. V., Savelli, L., Ferrazzi, E., Leone, F. P. G., Marret, H., Tranquart, F., Exacoustos, C., Nazzaro, G., Bokor, D., Magri, F., Van Huffel, S., Ferrandina, G., & Valentin, L. Intravenous contrast ultrasound examination using contrast-tuned imaging (CnTI) and the contrast medium SonoVue for discrimination between benign and malignant adnexal masses with solid components. *Ultrasound in Obstetrics and Gynecology*, **34**(6), 699–710 (2009). (IF 2.690)

Dieudonne, A.S., **Van Belle, V.**, Neven, P., Association between CYP2D6 polymorphisms and breast cancer outcomes. *Journal of the American Medical Association*, **303**(6), 516–517 (2010). (IF 31.718)

Van Holsbeke, C., Zhang, J., **Van Belle, V.**, Paladini, D., Guerriero, S.,

Czekierdowski, A., Muggah, H., Ombelet, W., Jurkovic, D., Testa, A. C., Valentin, L., Van Huffel, S., Bourne, T., & Timmerman, D. Acoustic streaming cannot discriminate reliably between endometriomas and other types of adnexal lesion: a multicenter study of 633 adnexal masses. *Ultrasound in Obstetrics and Gynecology*, **35**(3), 349–353 (2010). (IF 2.690)

Van Holsbeke, C., **Van Belle, V.**, Leone, F. P. G., Guerriero, S., Paladini, D., Melis, G. B., Greggi, S., Fischerova, D., De Jonge, E., Neven, P., Bourne, T., Valentin, L., Van Huffel, S., & Timmerman, D. Prospective external validation of the ovarian crescent sign as a single ultrasound parameter to distinguish between benign and malignant adnexal pathology. *Ultrasound in Obstetrics and Gynecology*, **36**(1), 81–87 (2010). (IF 2.690)

Papers accepted for publication conditioned on minor revisions

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. Learning Transformation Models for Ranking and Survival Analysis. Technical report, 09-45, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2009). Submitted to *Journal of Machine Learning Research*. (IF 2.789)

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. K. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Technical Report, 09-235, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2009). Submitted to *Artificial Intelligence in Medicine*. (IF 1.645)

Van Belle, V., Decock, J., Hendrickx, W., Brouckaert, O., Pintens, S., Moerman, P., Wildiers, H., Paridaens, R., Christiaens, M.R., Van Huffel, S., & Neven P. Short term prognostic index for breast cancer: NPI or Lpi. Submitted to *Pathology Research International*.

Submitted papers

Van Belle, V., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., Neven, P., Van Huffel, S., Suykens, J. A. K., & Boyd, S. Bridging clinical decision making and flexible mathematical modelling: the Interval Coded Score index. Technical report, 10-170, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2010). Submitted to *Science Translational Medicine*.

Papers in preparation

Van Belle, V.*, Cho, H.*, Wildiers, H., Van Huffel, S., Paridaens, R., Weltens, C., Berteloot, P., Janssen, H., Christiaens, M. R., & Neven, P. Factors predicting disease free survival following neoadjuvant chemotherapy for breast cancer.

Cho, H.*, **Van Belle, V.***, Vandorpe, T., Wildiers, H., Janssen, H., Leunen, K., Amant, F., Vergote, I., Berteloot, P., Smeets, A., Peeters, S., Van Limbergen, E., Weltens, C., Paridaens, R., Van Huffel, S., Christiaens, M.R., & Neven, P. The progesterone receptor and use of an oral aromatase inhibitor are prognostic for 5 year relapse in operable HER-2 positive breast cancers from the pre-trastuzumab era.

Bottomley, C., **Van Belle, V.**, Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. Prediction of pregnancy viability by means of scoring systems: towards a 'risk of miscarriage' index.

Bottomley, C., **Van Belle, V.**, Mukri, F., Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. Determination of outcome in very early intrauterine pregnancies of uncertain viability.

Van Calster, B., **Van Belle, V.**, Vergouwe, Y., Timmerman, D., Van Huffel, S., & Steyerberg, E. Extending the concordance probability to nominal and ordinal outcomes.

Van Calster, B., Vergouwe, Y., **Van Belle, V.**, Looman, C., Timmerman, D., Van Huffel, S., & Steyerberg, E. How to assess the discriminative ability of prediction models for more than two outcome levels?

Papers in international conference proceedings

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

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. Survival SVM: a Practical Scalable Algorithm. In Verleysen, M., editor, *Proceedings of the 16th European Symposium on Artificial Neural Networks*, pages 89–94 (2008).

Van Belle, V., Pelckmans, K., Suykens, J.A.K., & Van Huffel, S. Feature Selection in Survival Least Squares Support Vector Machines with Maximal Variation Constraints. In *Proceedings of the 10th International Work-Conference on Artificial Neural Networks*, pages 65–72 (2009).

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. MINLIP: Efficient Learning of Transformation Models. In Alippi, C., Polycarpou, M. M.,

Panayiotou, C., & Ellinas, G., editors, *Proceedings of the International Conference on Artificial Neural Networks*, pages 60–69 (2009).

Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. On the use of a clinical kernel in survival analysis. In Verleysen, M., editor, *Proceedings of the European Symposium on Artificial Neural Networks*, pages 451–456 (2010).

Van Calster, B., **Van Belle, V.**, Condous, G., Bourne, T., Timmerman, D., & Van Huffel, S. Multi-class AUC metrics and weighted alternatives. In *Proceedings of the 21st International Joint Conference on Neural Networks*, pages 1391–1397 (2008).

Papers presented at national conferences

Van Belle, V., Poullet, JB., Monleon, D., Celda, B., Martinez-Bisbal, M. & Van Huffel, S. Classification of brain tumors by means of high resolution magic angle spinning spectra. In *Proceedings of the Belgian Day on Biomedical Engineering IEEE/EMBS Symposium*, pages 199–202 (2006).

Internal reports

Pelckmans, K., **Van Belle, V.**, Van Huffel, S., Suykens, J.A.K. Learning Transformation Models for Failure Time Data. Internal Report 08-16, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.

Pelckmans, K., **Van Belle, V.**, Van Huffel, S., Suykens, J.A.K. Learning Lipschitz-smooth Utility Functions for Transformation Models in Ranking Problems. Internal Report 08-201, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.

Abstracts in international journals

Van Belle, V., Van Calster, B., Van Huffel, S., Brouckaert, O., Vanden Bempt, I., Pintens, S., Paridaens, R., Christiaens, M.R., & Neven, P. Improved NPI for breast cancer prognosis. *European Journal of Cancer*, 6(7), 190 (2008).

Van Belle, V., Brouckaert, O., Van Huffel, S., Schlichting, E., Synnestvedt, M., Naume, B., Christiaens, M.R., & Neven P. Improving NPI for breast cancer prognosis by including PR and HER-2 expression: own data and external validation set. *Cancer Research*, 69(S), 128 (2009).

Van Belle, V.*, Cho, H.*, Wildiers, H., Van Huffel, S., Paridaens, R.,

Weltens, C., Berteloot, P., Janssen, H., Christiaens, M. R., & Neven, P. Factors predicting disease free survival following neoadjuvant chemotherapy for breast cancer. *European Journal of Cancer*, **8**(3), 101 (2010).

Cho, H.*, **Van Belle, V.***, Vandorpe, T., Wildiers, H., Leunen, K., Amant, F., Vergote, I., Berteloot, P., Smeets, A., Limbergen, E. V., Weltens, C., Paridaens, R., Huffel, S. V., Christiaens, M. R., & Neven, P. Prognostic significance of nodal and PgR status on early relapse in operable HER-2 positive breast cancer from the pre-trastuzumab era. *Cancer Research*, **69**(24), 837 (2009).

Cho, H.*, **Van Belle, V.***, Wildiers, H., Paridaens, R., Amant, F., Van Limbergen, E., Moerman, P., Smeets, A., Vergote, I., & Neven, P. Factors predicting a pathological complete response following neo-adjuvant chemotherapy for breast cancer. *European Journal of Cancer*, **8**(3), 66 (2010).

Bottomley, C., **Van Belle, V.**, Mukri, F., Kirk, E.J., Papageorghiou, A.T., Van Huffel, S., Timmerman, D., & Bourne T. OC69: Ability to make a diagnosis at first early pregnancy ultrasound assessment according to gestational age. *Ultrasound in Obstetrics and Gynecology*, 30(4), 388 (2007).

Bottomley, C., **Van Belle, V.**, Mukri, F., Van Huffel, S., Timmerman, D., & Bourne, T. OC48: Development of a model to predict ability of ultrasound to make a diagnosis at first assessment in the unselected early pregnancy population. *Ultrasound in Obstetrics and Gynecology*, 30(4), 382 (2007).

Pintens, S., Vanden Bempt, I., Drijkoningen, M., **Van Belle, V.**, Brouckaert, O., Christians, M.R., Neven, P., & De Wolf-Peeters, C. uPAR gene amplification in breast cancer tissue: a rare event. *Breast Cancer Research and Treatment*, 106(S1), 82 (2007).

Pintens, S., Vanden Bempt, I., Drijkoningen, M., **Van Belle, V.**, Brouckaert, O., Christiaens, M.R., Neven, P., & De Wolf-Peeters, C. HER-1 overexpression is found only in oestrogen receptor negative breast cancer and is rarely associated with HER-1 gene amplification. *Breast Cancer Research and Treatment*, 106(S1), 81 (2007).

Brouckaert, O., Pintens, S., **Van Belle, V.**, Van Huffel, S., Amant, F., Leunen, K., Smeets, A., Berteloot, P., Van Limbergen, E., Decock, J., Hendrickx, W., Weltens, C., Van den Bogaert, W., Vanden, Bempt I., Drijkoningen, M., Paridaens, R., Wildiers, H., Vergote, I., Christiaens, M.R., & Neven, P. Stratifying primary operable breast cancers by steroid and HER-2 expression status predicts short term disease free survival. *Breast Cancer Research and Treatment*, 106(S1), 247 (2007).

Brouckaert, O., Pintens, S., **Van Belle, V.**, Van Huffel, S., Amant, F., Leunen, K., Smeets, A., Berteloot, P., Van Limbergen, E., Decock, J., Hendrickx, W., Vanden Bempt, I., Drijkoningen, M., Paridaens, R., Wildiers, H., Vergote, I., Christiaens, M.R., & Neven, P. Disease free survival (dfs) according to the joint steroid- and her-

2 receptor status in breast cancer. *European Society of Gynaecological Oncology* (2007).

Decock, J., Hendrickx, W., Vanleeuw, U., **Van Belle, V.**, Van Huffel, S., Christiaens, M.R., Edwards, D., Ye, S., & Paridaens R. Plasma MMP1, MMP8 and MMP13 expression in breast cancer: protective role of MMP8 against lymph node metastasis. *Cancerdegradome* (2007).

Decock, J., Hendrickx, W., **Van Belle, V.**, Drijkoningen, M., Brouckaert, O., Pintens, S., Van Huffel, S., Paridaens, R., Wildiers, H., Amant, F., Leunen, K., Smeets, A., Christiaens, M.R., & Neven P. Short term prognostic index for breast cancer: NPI or Lpi. *Breast Cancer Research and Treatment*, 106(S), 55–56 (2007).

Bottomley, C., **Van Belle, V.**, Mukri, F., Van Huffel, S., Timmerman, D., & Bourne, T. Development of a model to predict ability of ultrasound to make a diagnosis at first assessment in the unselected early pregnancy population. *BJOG: An International Journal of Obstetrics and Gynaecology*, 115 (1), 118–126 (2008).

Bottomley, C., **Van Belle, V.**, Mukri, F., Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. Determination of outcome in very early intrauterine pregnancies of uncertain viability. *Ultrasound in Obstetrics and Gynecology*, 32 (3), 279 (2008).

Van Holsbeke, C., Guerriero, S., **Van Belle, V.**, Czekierdowski, A., Ombelet, W., Valentin, L., Fischerova, D., Paladini, D., Jingzhang, Z., Jurkovic, D., Bourne, T., & Timmerman, D. Presence of acoustic streaming in different types of adnexal masses. *Ultrasound in Obstetrics and Gynecology*, 32 (3), 296 (2008).

Testa, A., Timmerman, D., Fruscella, E., Van Holsbeke, C., Savelli, L., Ferrazzi, E., Leone, F.P.G., Marret, H., Exacousos, C., De Placido, G., Bokor, D., Ferrandina, G., **Van Belle, V.**, Valentin, L. Ultrasound contrast examination to discriminate benign and malignant adnexal masses. *Ultrasound in Obstetrics and Gynecology*, 32(3), 297 (2008).

Bottomley, C., **Van Belle, V.**, Mukri, F., Kirk, E., Van Huffel, S., Timmerman, D., & Bourne, T. Optimal timing of transvaginal ultrasound to confirm early pregnancy viability. *Ultrasound in Obstetrics and Gynecology*, 32(3), 349 (2008).

Neven, P., **Van Belle, V.**, Vanden Bempt, I., Deraedt, K., Moerman, P., Paridaens, R., Leunen, K., Smeets, A., Wildiers, H., & Christiaens, M.R. The oestrogen receptor interacts with the correlation between HER-2 over-expression and age at diagnosis, tumor grade and lymph node involvement in operable breast cancers: a single center experience. *European Journal of Cancer*, 6(7), 85 (2008).

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 de Bogaert, W., Paridaens, R., Moerman, P., Vergote, I., Christiaens, M.R., Wildiers, H., & Neven P. Biology and prognosis by age of primary operable breast cancer. *Cancer Research*, 69(S), 181 (2009).

Dierickx, I., **Van Belle, V.**, Vandorpe, T., Wildiers, H., Leunen, K., Peeters, S., Moerman, P., Deraedt, K., Christiaens, M. R., & Neven, P. A prognostic model for breast cancer-related events in primary operated invasive lobular breast cancers from one center. *European Journal of Cancer*, **8**(3), 94 (2010).

Abstracts presented at international conferences

Van Calster, B., **Van Belle, V.**, Timmerman, D., & Van Huffel, S. An AUC-based measure to evaluate the performance of polytomous diagnostic models with an application on ovarian tumor diagnosis. Abstract presented at the 30th Annual Conference of the International Society for Clinical Biostatistics, Prague, Czech Republic, August 2009.

Van Calster, B.,Vergouwe, Y., **Van Belle, V.**, Steyerberg, E., Bourne, T., Timmerman, D., & Van Huffel, S. Performance evaluation for prediction models with three outcomes: a discrimination index applied for the prediction of pregnancy location. Accepted for presentation at International Biometric Conference, Floripa 2010.

Invited Talks

XXVI congreso nocional de oncologia (22-25/10/2008, Leon-Mexico) Coloquio internatcional de medicina translacional: la clinica como sustento de la investigación basica: Los indices pronosticos en oncologia. *Corrected NPI for prognosis in Breast Cancer.*

Arenberg Doctoral School of Science, Engineering & Technology
Faculty Engineering
Department of Electrical Engineering (ESAT)
SCD/SISTA - BIOMED
Kasteelpark Arenberg 10 box 2446, B-3001 Leuven

