Learning Transformation Models for Ranking and Survival Analysis

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

V. Van Belle,
K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10 bus 2446, B-3001 Leuven (Belgium),
vanya.vanbelle@esat.kuleuven.be
K.Pelckmans,
Uppsala University, Department of Information Technology ITC building 2,
SE-751 05 Uppsala, Sweden, kristiaan.pelckmans@it.uu.se
J.A.K. Suykens,
K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10 bus 2446, B-3001 Leuven (Belgium),
johan.suykens@esat.kuleuven.be
S. Van Huffel,
K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10 bus 2446, B-3001 Leuven (Belgium),
sabine.vanhuffel@esat.kuleuven.be

Abstract

This paper studies the task of learning transformation models for ranking problems, ordinal regression and survival analysis. The present contribution describes a machine learning approach termed MINLIP. The key insight is to relate ranking criteria as the Area Under the Curve to monotone transformation functions. Consequently, the notion of a Lipschitz smoothness constant is found to be useful for complexity control for learning transformation models, much in a similar vein as the 'margin' is for Support Vector Machines for classification. The use of this model structure in the context of high dimensional data, as well as for estimating non-linear, and additive models based on primal-dual kernel machines, and for sparse models is indicated. Given n observations, the present method solves a quadratic program existing of $\mathcal{O}(n)$ constraints and $\mathcal{O}(n)$ unknowns, where most existing risk minimization approaches to ranking problems typically result in algorithms with $\mathcal{O}(n^2)$ constraints or unknowns. We specify the MINLIP method for three different cases: the first one concerns the preference learning problem. Secondly it is specified how to adapt the method to ordinal regression with a finite set of ordered outcomes. Finally, it is shown how the method can be used in the context of survival analysis where one models failure times, typically subject to censoring. The current approach is found to be particularly useful in this context as it can handle, in contrast with the standard statistical model for analyzing survival data, all types of censoring in a straightforward way, and because of the explicit relation with the Proportional Hazard and Accelerated Failure Time models. The advantage of the current method

is illustrated on different benchmark datasets, as well as for estimating a model for cancer survival based on different micro-array and clinical datasets.
keywords: Support Vector Machines, Preference Learning, Ranking Models, Ordinal Regression, Survival Analysis

1 Introduction

Methods based on ranking continue to challenge researchers in different scientific areas, see e.g. Clémençon et al. (2005); Herbrich, Graepel, and Obermayer (2000) and the references therein. Learning ranking functions offers a solution to different types of problems including ordinal regression, bipartite ranking and discounted cumulative gain ranking (DCG, see Clémençon and Vayatis, 2007), studied frequently in research on information retrieval. These cases distinguish themselves in the definition (of the cardinality k) of the output domain and the chosen loss function. This paper deals with the general problem where the output domain can be arbitrary (with possibly infinite members $k = \infty$), but possesses a natural ordering relation between the members. Examples in which $k = \infty$ are found in survival analysis and preference learning in cases where the number of classes is not known in advance.

Earlier approaches to learning preference functions reduce the ranking problem to pairwise classification problems. This reasoning was followed in Ailon and Mohri (2008); Fürnkranz and Hüllermeier (2003); Herbrich et al. (1998) and references therein. However, functions having high pairwise margins might still be bad approximations to real ranking problems. This is certainly the case in the (general) preference learning problem where possibly $k = \infty$: here a nonzero pairwise margin would need unnecessarily large parameters of the model. In this paper 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 1 summarizes the ideas exposed in this work. First we describe the class of transformation models which contains two different components. The first component of a transformation model consists of 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 literature depending on the problem setting, including a scoring, ranking, utility or health function. In this paper 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 now 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. For ranking and survival analysis one typically ignores the second phase, whereas in ordinal regression a prediction of the output level

is found by combining the first and the second components.

Transformation models are especially appropriate when considering data arising from a survival study. Survival analysis concerns data which represent a time-to-event, as e.g. a patient relapsing after surgery, or the time till a part of a mechanical device breaks down, see Kalbfleisch and Prentice (2002) for a broad survey of this field. The goal in survival analysis is often to relate time-to-event of an instance to a corresponding set of covariates. While practice and theoretical results here 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: as (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 (if at all, see Kalbfleisch and Prentice, 2002 and references).

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 nicely 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 Cox' proportional hazard and accelerated failure time (AFT) models, see e.g. Dabrowska and Doksum (1988); Koenker and Geling (2001); Cheng, Wei, and Ying (1997) and citations. In the following, 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). Thirdly, in studies of failure time data, *censoring* is omnipresent. Censoring prohibits that one observes the actual event of interest fully, but gives partial information on the outcome instead. The prototypical case is 'a patient hasn't suffered the event as yet, but may experience an event in the future', but many other examples are studied. We will see how the proposed approach can handle censored observations conveniently.

The computational merit of this paper is then how one can fit such a model efficiently to data. Therefore, we consider an appropriate class of utility functions, either linear functions, or kernel based models. Secondly, instead of restricting attention to a parameterized class of transformation functions, we let the transformation function of interest be unspecified as one does for partial likelihood approaches, see Kalbfleisch and Prentice (2002). Especially, we define the appropriate transformation function only on the observed samples, by inferring an appropriate set of ordinal relations between them. Then we observe that the Lipschitz smoothness constant associated to such a transformation function can also be evaluated based on the samples only.

Consequently, our fitting strategy called 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 we allow for misfit, we extend the model using slack-variables. It is then found that this problem can be solved as a convex Quadratic Program (QP), for which highly efficient software is readily available. In the case of utility functions which are kernel based models, we indicate how one can represent the solution as a sum of positive definite kernels, and the Lagrange dual problem again 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 zero parameters ('sparseness') suggesting structure in the data using an 1-norm regularization mechanism (see also Tibshirani, 1996), and using positivity constraints on the parameters. Interestingly, we find that the latter procedure produces more accurate models and has better recovery properties than the 1-norm approach.

Besides the conceptual and computational discussion, this paper 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 and 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 dataset 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 Cox regression 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 smoothness strategy as presented in Van Belle et al. (2009). This work extends considerably the results of this short paper. Most notably, this paper additionally elaborates on the case of survival analysis and a number of new case studies. The different application areas in which the proposed method can be applied are summarized in Table 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 paper is organized as follows. The following Section discusses in some detail the use of transformation models and its relation with ranking methods. Section 3 studies the estimator in a context of ranking. Section 4 specifies how MINLIP is to be used in a context of ordinal regression, where only k different output levels are possible. Section 5 discusses the use of MINLIP in the presence of censoring. In Section 6 experiments illustrate the use of the MINLIP method.

2 Transformation Models and Ranking Methods

In this paper we work in a stochastic context, so we denote random variables as capital letters, e.g. $X, Y, \ldots$, which follow an appropriate stochastic law $P_X, P_Y, \ldots$, abbreviated (generically) as P. Deterministic quantities as constants and functions are represented in lower case letters (e.g. $d, h, u, \ldots$). Matrices are denoted as boldface capital letters (e.g. $\mathbf{X}, \mathbf{D}, \ldots$). Ordered sets will

be denoted as $\{S_{(i)}\}$, indicating that $S_{(i)} \leq S_{(i+1)}$. Before the relation between transformation models and ranking methods can be explored, some terminology needs to be defined.

Definition 1 (Lipschitz smoothness) *A univariate function $h(Z)$ has a Lipschitz constant $L \geq 0$ if*

$$|h(Z) - h(Z')| \leq L|Z - Z'|, \; \forall Z, Z' \in \mathbb{R}\,. \tag{1}$$

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 $L < \infty$, 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*

$$Y = h(u(X) + \epsilon). \tag{2}$$

In the remainder of the paper, we will use Z to 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 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.

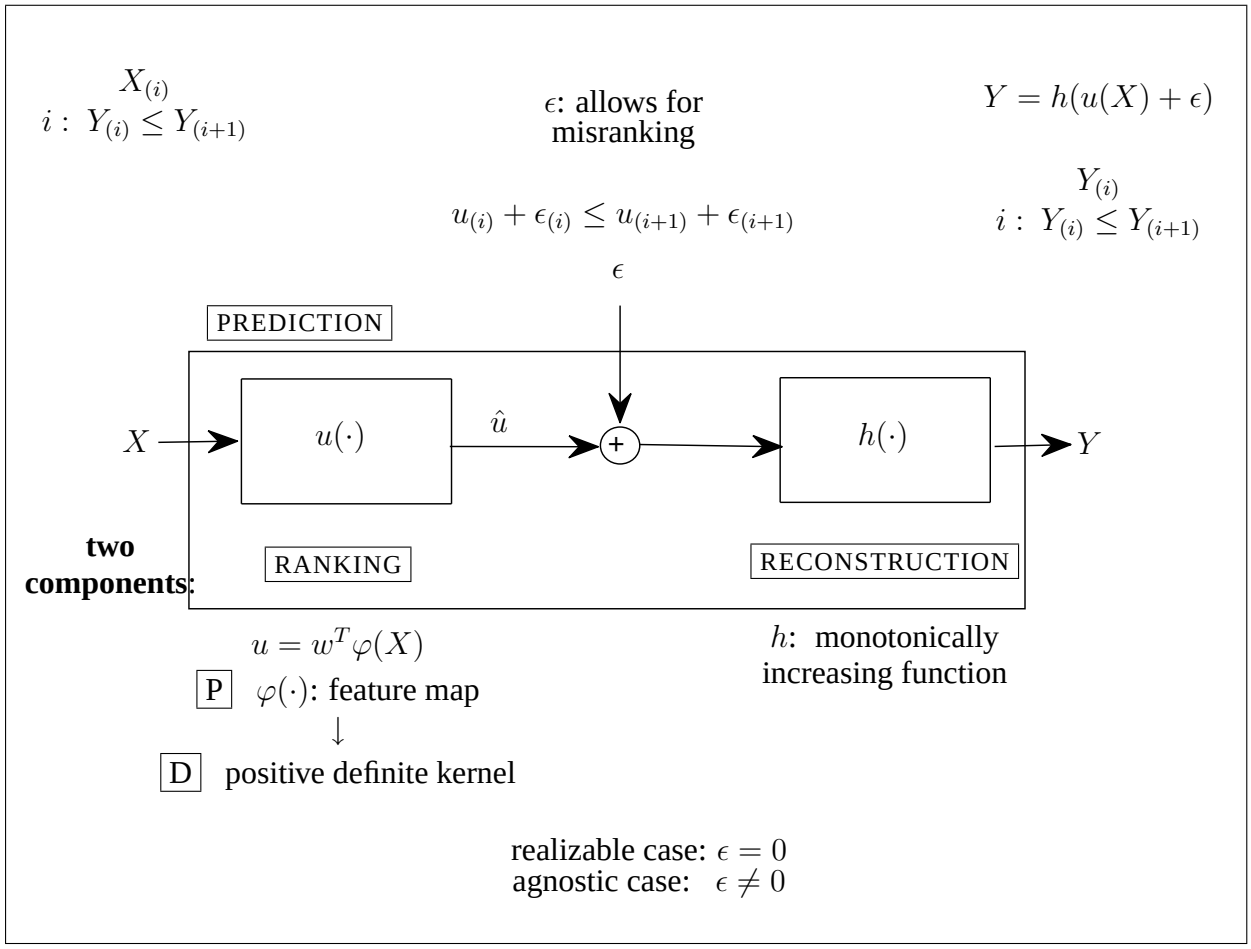


Figure 1: Overview: Transformation models consist of two components, the *utility function* u and a *transformation function* h. Given a dataset $\mathcal{D} = \{(X_{(i)}, Y_{(i)})\}_{i=1}^n$ where the instances are sorted such that $Y_{(i)} \leq Y_{(i+1)}$, a utility function $u(X) = w^T \varphi(X)$ 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 (nonzero) 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'). In practice (depending on the problem at hand) one is mostly interested in implementing the first step only.

Table 1: Overview of methods and applications proposed in the paper. 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 B, ...) 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 l^{th} interval $\hat{Y}_{il}$, ...).

Task	Subtasks	algorithm	necessary training data for algorithm	result of algorithm	equation	comment
ranking	ranking	MINLIP	$\{(X_i, Y_i)\}$	$\hat{u}(X)$	(22-23)	Use in combination with (28-29)* to obtain sparse models
ordinal regression	ranking	MINLIP	$\{(X_i, Y_i)\}, B$	$\hat{u}(X), \hat{v}$	(33-34)	Use in combination with (28-29)* to obtain sparse models
	reconstruction	comparison with thresholds v	$\{\hat{u}(X_i)\}, \hat{v}$	$\hat{Y}$	Figure 6	The goal is to predict class label
survival analysis	ranking	MINLIP	$\{(X_i, Y_i)\}$	$\hat{u}(X)$	(52-53)	Use in combination with (28-29)* to obtain sparse models
	reconstruction	monotonic regression after replication in consecutive time intervals	$\{(\hat{u}(X_i), Y_i)\}$	$\{\hat{Y}_{il}\}$		The goal is to predict hazard and/or survival function

*Equations (28-29) can only be used in combination with a linear kernel

Kalbfleisch and Prentice (2002) considered transformation models for failure time models. The transformation models discussed in Cheng, Wei, and Ying (1997); Dabrowska and Doksum (1988); Koenker and Geling (2001) differ from the above definition in the transformation function h. They define the model as $h^-(Y) = u(X) + \epsilon$, which is equivalent to (2) if $h^-(h(Z)) = h(h^-(Z)) = Z$ for all Z.

To relate transformation models with ranking functions, we reason as follows. To express the performance of a ranking function one can use Kendall's τ, area under the curve (AUC) or a related measure. In this paper we will work with the *concordance* of a function $u : \mathbb{R}^d \to \mathbb{R}$ respective to the outcome. 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:

$$\mathcal{C}(u) = P((u(X) - u(X'))(Y - Y') > 0). \tag{3}$$

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

$$\mathcal{C}_n(u) = \frac{2}{n(n-1)} \sum_{i<j} I[(u(X_i) - u(X_j))(Y_i - Y_j) > 0], \tag{4}$$

where the indicator function $I(z)$ equals 1 if $z > 0$, and equals zero otherwise. 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*

$$\mathcal{R}(u) = P((h(u(X)) - h(u(X')))(Y - Y') < 0). \tag{5}$$

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

$$\mathcal{R}(u) = P((u(X) - u(X'))(Y - Y') < 0). \tag{6}$$

Its empirical counterpart then becomes

$$\mathcal{R}_n(u) = 1 - \mathcal{C}_n(u)\,. \tag{7}$$

Empirical Risk Minimization (ERM) is then performed by solving

$$\hat{u} = \arg\min_{u \in \mathcal{U}} \mathcal{R}_n(u) = \arg\max_{u \in \mathcal{U}} \mathcal{C}_n(u), \tag{8}$$

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. $\ell(z) \leq \max(0, 1 - z)$ yielding rankSVM (Herbrich, Graepel, and Obermayer, 2000). The disadvantage of this solution is that it leads to $O(n^2)$ number of constraints or unknowns, often making it difficult to apply to real life problems. A solution to this problem is found in relating transformation models with equation (8): if a function $u : \mathbb{R}^d \to \mathbb{R}$ exists such that $\mathcal{C}_n(u) = 1$,

one describes implicitly a transformation function (see Figure 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 $y = h(u)$. Or more formally:

Lemma 1 (Existence of a Transformation Function) *Given a collection of pairs $\{(Z_{(i)}, Y_{(i)})\}_{i=1}^{n}$, enumerated such that $Y_{(i)} \leq Y_{(j)}$ if and only if $i \leq j$, and considering the conditions on the observations for $L < \infty$:*

$$0 \leq Y_{(i)} - Y_{(j)} \leq L\left(Z_{(i)} - Z_{(j)}\right) \;\; \forall\, i < j = 1, \ldots, n, \tag{9}$$

we state that:

1. *If one has for a finite value $L \geq 0$ that (9) 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 $L < \infty$ such that (9) holds.*

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

$$h_n(Z) = \frac{Z - Z_{\underline{z}(Z)}}{Z_{\overline{z}(Z)} - Z_{\underline{z}(Z)}}\left(Y_{\overline{z}(Z)} - Y_{\underline{z}(Z)}\right) + Y_{\underline{z}(Z)} \tag{10}$$

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 $h_n(Z') - h_n(Z) \leq L(Z' - Z)$. For notational convenience define $l = \underline{z}(Z)$, $u = \overline{z}(Z)$, $l' = \underline{z}(Z')$ and $u' = \overline{z}(Z')$, then

$$\begin{aligned} h_n(Z') - h_n(Z) &= \frac{Z' - Z_{l'}}{Z_{u'} - Z_{l'}}(Y_{u'} - Y_{l'}) + Y_{l'} - \frac{Z - Z_l}{Z_u - Z_l}(Y_u - Y_l) - Y_l \\ &\leq L(Z' - Z_{l'}) - L(Z - Z_l) + L(Z_{l'} - Z_l) \\ &= L(Z' - Z), \end{aligned} \tag{11}$$

where we use that $Y_{l'} - Y_l \leq L(Z_{l'} - Z_l)$.

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

$$(Y_j - Y_i) = (Z_j - Z_i)h'(Z) \leq L(Z_i - Z_j) \tag{12}$$

where $L = \sup_Z h'(Z)$.
Note that equation (9) implies that $\mathcal{C}_n(Z) = 1$. ■

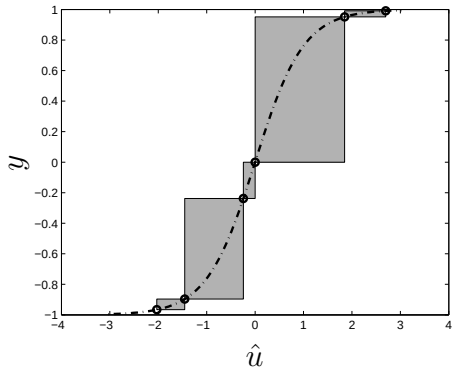


Figure 2: Relation between ranking and transformation models: if two variables u and y are perfectly concordant, they describe a monotonically increasing function $y = h(u)$. The dots represent u and outcome y for training points. In these observations the value of the function h is known exactly. To predict the $y-$value of the test observations, the function h needs to be approximated between the training points (grey area). All functions $\hat{h}$ which are onotonically increasing and lie within the grey zones are valid prediction rules.

3 MINLIP: a Convex Approach to Learning a Transformation Model

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 $\mathcal{C}(u) = 1$. Lemma 1 describes the existence of h, but since this transformation function is only known at the training points, it is not unique. Figure 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 $L < \infty$, such that $h(Z) - h(Z') \leq L(Z - Z')$ for all $Z \geq Z'$. Restricting attention to the observations $\{(X_{(i)}, Y_{(i)})\}_{i=1}^{n}$, one has the necessary conditions

$$h\left(u(X_{(i)})\right) - h\left(u(X_{(i-1)})\right) \leq L\left(u(X_{(i)}) - u(X_{(i-1)})\right)$$

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

$$u(X) = w^T X\,, \tag{13}$$

are considered. Extensions towards non-linear utility functions using Mercer kernels are handled in Subsection 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 $\alpha > 0$, one has $h(u(X)) = \tilde{h}(\tilde{u}(X))$ where $\tilde{h}(Z) \triangleq h(\alpha^{-1}Z)$ and $\tilde{u}(X) = \alpha u(X)$), 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

$$\begin{array}{ll} \min\limits_{v,L} & \frac{1}{2}L^2 \\ \text{s.t.} & \begin{cases} \|v\|_2 = 1 \\ Y_{(i)} - Y_{(i-1)} \leq L\left(v^T X_{(i)} - v^T X_{(i-1)}\right), \ \forall\, i = 2, \ldots, n. \end{cases} \end{array} \tag{14}$$

Substituting $w = Lv$ we get equivalently:

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2}w^T w \\ \text{s.t.} & Y_{(i)} - Y_{(i-1)} \leq w^T X_{(i)} - w^T X_{(i-1)}, \ \forall\, i = 2, \ldots, n\,, \end{array} \tag{15}$$

which goes along similar lines as the hard margin SVM (see e.g. Shawe-Taylor and Cristianini, 2004) and ranking SVM (Freund et al., 2004), 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. Observe that this problem has $n-1$ linear constraints. We will refer to this model as MINLIP.

Problem (15) can be compactly rewritten as

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2}w^T w \\ \text{s.t.} & \mathbf{D}\mathbf{X}w \geq \mathbf{D}\mathbf{Y}, \end{array} \tag{16}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a matrix with each row containing one observation, i.e. $\mathbf{X}_i = X_{(i)} \in \mathbb{R}^d$ and $\mathbf{Y} = [Y_{(1)} \cdots Y_{(n)}]^T$, a vector with the corresponding outcomes. The matrix $\mathbf{D} \in \{-1, 0, 1\}^{(n-1)\times n}$

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \ldots & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{17}$$

gives the first order differences of a vector, i.e. assuming no ties in the output, $\mathbf{D}_i\mathbf{Y} = Y_{(i+1)} - Y_{(i)}$ for all $i = 1, \ldots, n-1$, with $\mathbf{D}_i$ the i^{th} row of $\mathbf{D}$.

In the presence of ties, $Y_{(i+1)}$ is replaced by $Y_{(j)}$, with j the smallest output value with $Y_{(j)} > Y_{(i)}$. See Section 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*

$$Y = h_0(w_0^T X), \tag{18}$$

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

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

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 $L' > 0$ be defined as

$$\frac{1}{L'} = \frac{w_0^T(X - X')}{Y - Y'}, \tag{19}$$

where $L' = \infty$ if $w_0^T(X - X') = 0$. By construction we have that $L' \leq L_0$ and that the constant exists everywhere. The result of the MINLIP *model then becomes:*

$$\pounds = \max_{\|w\|_2=1} \min_{Y>Y'} \frac{w^T(X - X')}{Y - Y'} \quad = \quad \max_{\|w\|_2=1} \min_{Y>Y'} \frac{w^T(X - X')}{L' w_0^T(X - X')}\,. \tag{20}$$

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 $L' \approx L_0$.*

Proof: 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 (20) as

$$\max_{\|w\|_2=1} \min_{Y>Y'} \frac{w^T\overline{(X - X')}}{L' w_0^T\overline{(X - X')}}\,. \tag{21}$$

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 $\pounds = w^T(X - X')/(Y - Y')$ for which this value $\pounds$ is actually achieved. It is seen that the estimate w lies in the span of this set $\mathcal{S}$ as otherwise the maximum value could be increased. When we assume that the dataset contains enough observations (X, Y) such that $w_0^T(X - X') \approx 1$ and $L' \approx L_0$, 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 $L' \approx L_0$ (cfr. duality results in convex optimization), one should also have that $w \approx w_0$. ■

Formally, consistency of MINLIP in the asymptotic case under a sufficient condition of the data being non-degenerate is derived in Appendix A.

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:

$$Y = h(w^T X + \epsilon)\,, \tag{22}$$

where $u = w^T X$. The introduction of the error variable asks for an adaptation of the Lipschitz-based complexity control. As a loss function $\ell : \mathbb{R} \to \mathbb{R}$ we choose the absolute value loss $\ell(\epsilon) = |\epsilon|$ 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 $\ell(\epsilon) = \epsilon^2$; (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:

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \gamma\|\epsilon\|_1 \\ \text{s.t.} & \mathbf{D}(\mathbf{X}w + \epsilon) \geq \mathbf{DY}, \end{array} \tag{23}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ represents the errors, $\|\epsilon\|_1 = \sum_{i=1}^n |\epsilon_i|$ and $\gamma > 0$ is a regularization constant, making a trade-off between model complexity and error. This problem can again be solved as a convex quadratic program.

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, possibly infinite). A non-linear utility function can then be defined as

$$u(X) = w^T \varphi(X), \tag{24}$$

with $w \in \mathbb{R}^{d_\varphi}$ a vector of unknowns (possibly infinite dimensional). Take $\mathbf{\Phi} = [\varphi(X_{(1)}), \ldots, \varphi(X_{(n)})]^T \in \mathbb{R}^{n \times d_\varphi}$. The realizable learning problem can then be represented as:

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2}w^T w \\ \text{s.t.} & \mathbf{D\Phi}w \geq \mathbf{DY}, \end{array} \tag{25}$$

with the matrix $\mathbf{D}$ defined as before. The Lagrange dual problem becomes

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\alpha^T \mathbf{DKD}^T \alpha - \alpha^T \mathbf{DY} \\ \text{s.t.} & \alpha \geq 0_{n-1} \end{array} \tag{26}$$

where the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ contains the kernel evaluations such that $\mathbf{K}_{ij} = \varphi(X_i)^T \varphi(X_j)$ for all $i, j = 1, \ldots, n$. The estimated utility $\hat{u}$ can be evaluated at any point $X^* \in \mathbb{R}^d$ as

$$\hat{u}(X^*) = \hat{\alpha}^T \mathbf{DK}_n(X^*), \tag{27}$$

where $\mathbf{K}_n(X^*) = [K(X_1, X^*), \ldots, K(X_n, X^*)]^T \in \mathbb{R}^n$. The dual (Shawe-Taylor and Cristianini, 2004; Suykens et al., 2002; Vapnik, 1998) of the agnostic learning machine of Subsection 3.2 is obtained analogously:

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\alpha^T \mathbf{DKD}^T \alpha - \alpha^T \mathbf{DY} \\ \text{s.t.} & \begin{cases} -\gamma 1_n \leq \mathbf{D}^T \alpha \leq \gamma 1_n \\ \alpha \geq 0_{n-1}\,, \end{cases} \end{array} \tag{28}$$

with $\mathbf{K}$ as above and the resulting estimate can be evaluated as in (27) without computing explicitly $\hat{w}$ nor φ. We refer to Appendix B for a detailed derivation. Typical choices for kernel

functions are:

$$\begin{aligned} K(X, X_i) &= X_i^T X \text{ (linear kernel)} \\ K(X, X_i) &= (\tau + X_i^T X)^d \,,\ \tau \geq 0 \text{ (polynomial kernel of degree d)} \\ K(X, X_i) &= \exp\left(-\frac{||X - X_i||_2^2}{\sigma^2}\right) \text{ (RBF kernel)} \,. \end{aligned}$$

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

$$u(X) = \sum_{p=1}^{d} u^p(X^p) \,, \tag{29}$$

where X^p represents the p^{th} covariate of datapoint X. Using equation (27) this can be written as:

$$\hat{u}(X) = \sum_{p=1}^{d} \alpha^T \mathbf{D} \mathbf{K}^p(X^p) \tag{30}$$

$$= \alpha^T \mathbf{D} \sum_{p=1}^{d} \mathbf{K}^p(X^p) \,, \tag{31}$$

where the kernel matrix $\mathbf{K}^p \in \mathbb{R}^{n \times n}$ contains the kernel evaluations such that $\mathbf{K}_{ij}^p = \varphi(X_i^p)^T \varphi(X_j^p)$ for all $i,\ j = 1, \ldots, n$. As a result, componentwise kernels (Pelckmans et al., 2005a):

$$K(X, X_i) = \sum_{p=1}^{d} K^p(X^p, X_i^p) \,, \tag{32}$$

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.

3.4 Prediction with Transformation Models

Prediction of the outcome using transformation models is a two-step approach (see Figure 1). First, the utility $u(X)$ is estimated, giving an ordering relation between the observations. When interested in an outcome prediction, the transformation function h has to be estimated. The prediction step is a univariate regression problem, which can be solved using monotonic regression models. Remark that in the ranking setting, one is not interested in the estimation of the transformation function since the goal is to find the ranking. Estimation of the transformation function for ordinal regression and survival analysis will be illustrated later.

3.5 Toward Sparse Solutions using $\|w\|_1$ 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 primal case where $u(X) = w^T X$ for the linear case 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$_{L1}$:

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \|w\|_1 + \gamma\|\epsilon\|_1 \\ \text{s.t.} & \mathbf{D}(\mathbf{X}w + \epsilon) \geq \mathbf{DY}, \end{array} \tag{33}$$

where $\|w\|_1 = \sum_{p=1}^{d} |w_p|$. This linear programming problem (LP) can be solved efficiently with standard mathematical solvers. This formulation does not allow for a straightforward dual derivation.

As an alternative to 1-norm regularization, we propose a second extension incorporating positivity constraints of the parameters, which we will denote as MINLIP$_p$:

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \gamma\|\epsilon\|_1 \\ \text{s.t.} & \begin{cases} \mathbf{D}(\mathbf{X}w + \epsilon) \geq \mathbf{DY} \\ w \geq 0_d\,. \end{cases} \end{array} \tag{34}$$

This can be useful, even if the problem of interest not strictly implies those exact requirements. As a preprocessing step one should calculate the correlation between each covariate $p = 1, \ldots, d$, and the outcome. Covariates X^p which are inversely related with the outcome should be transformed as $X^p := -X^p$. This QP problem can be solved as before.

Figure 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 5 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 has the best trade-off between selected and correctly selected features.

3.6 Comparison with other methods

An approach often seen within preference ranking problems is the reformulation of the ranking problem as a classification problem. Examples of this strategy can be found in Ailon and Mohri

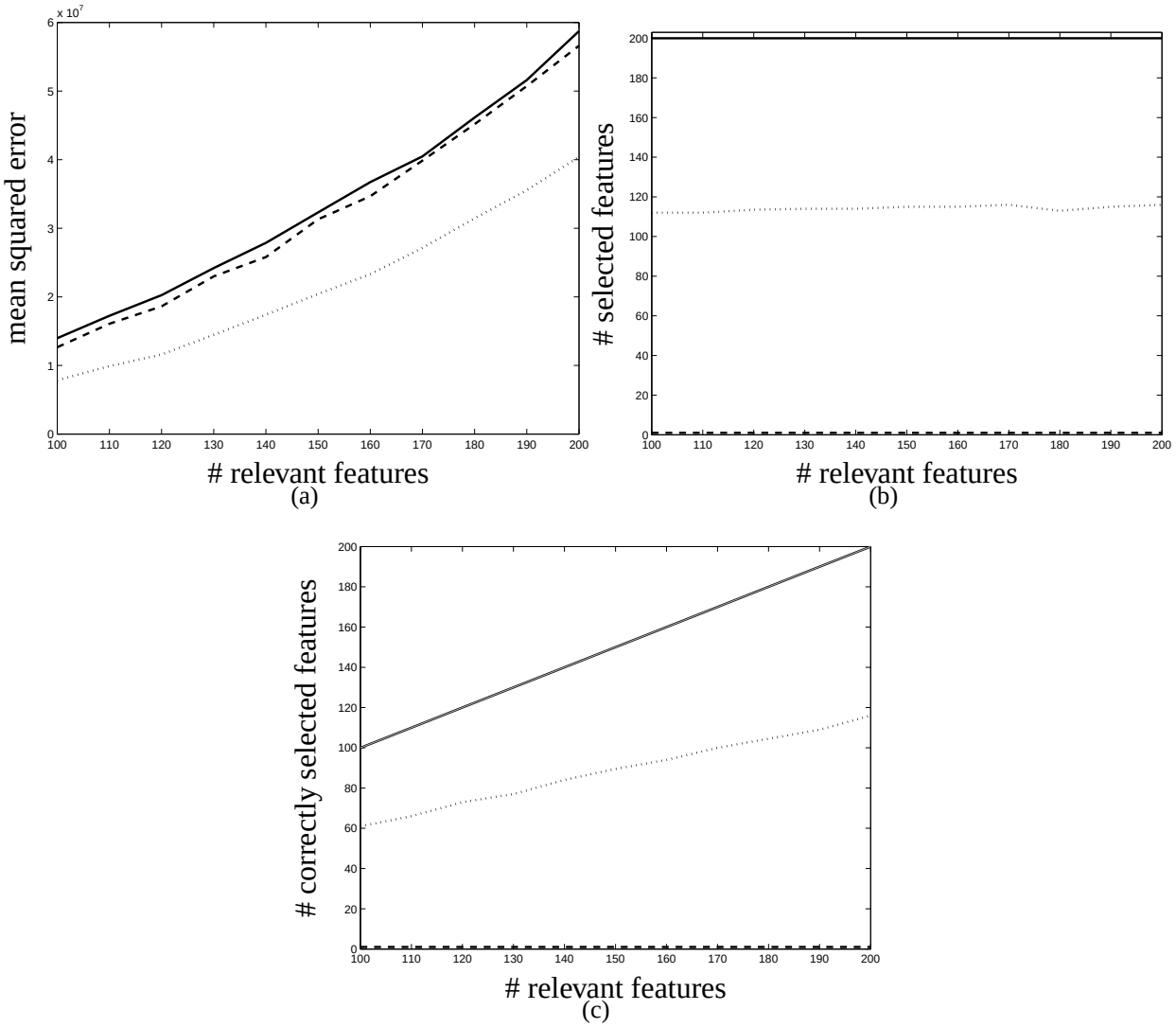


Figure 3: Performance and feature selection ability of MINLIP (solid) and two sparse alternatives (MINLIP$_{\mathrm{L1}}$ (dashed) and MINLIP$_{\mathrm{p}}$ (dotted) on an artificial dataset (n=100 for training, n_{test}=50 for testing). 200 $\mathcal{N}(5,1)$ distributed covariates were generated, a 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 form a standard normal distribution). The results are averaged over 100 datasets. (a) Median mean squared error on the test sets: MINLIP$_{\mathrm{p}}$ 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$_{\mathrm{p}}$ model. The MINLIP$_{\mathrm{L1}}$ model selects very few variables, but those which are selected are also relevant.

(2008); Fürnkranz and Hüllermeier (2003); Herbrich et al. (1998). 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, Graepel, and Obermayer, 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.

4 Learning for Ordinal Regression

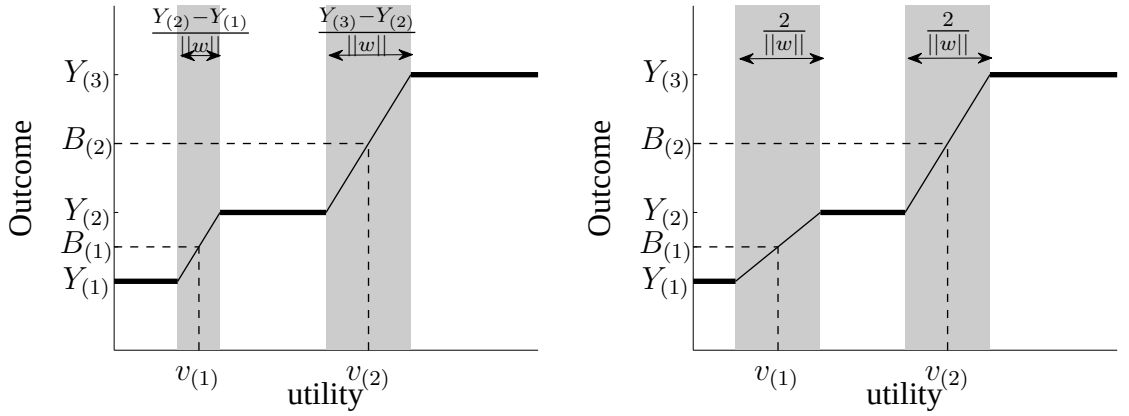
Consider now the situation where the output takes a finite number of values - say $k \in \mathbb{N}$ - and where the k different classes possess a natural ordering relation. In this case the outcome Y is an element of the finite ordered set $\{Y_{(1)}, \ldots, Y_{(k)}\}$.

4.1 A Modification to MINLIP

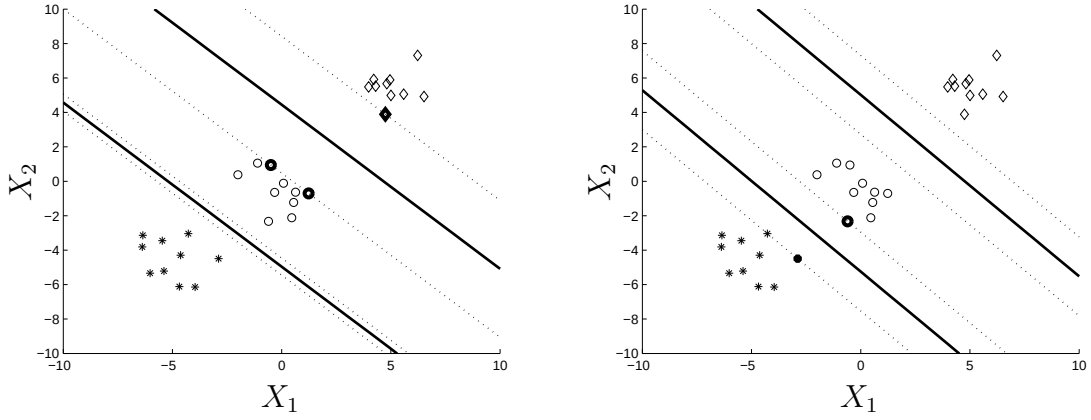
In Section 3.1 it is mentioned that comparisons are made between points (i) and (j) where $Y_{(j)}$ is the first ordered value bigger than $Y_{(i)}$. Applying this methodology in the ordinal setting would lead to as many comparisons with point (i) from class k_i as there are observations in class $k_i + 1$. To cope with this issue, we add dummy observations (X, B) in between two consecutive ordinal classes with levels $Y_{(i)} < Y_{(i+1)}$ such that $B_{(i)} = \frac{1}{2}(Y_{(i+1)} + Y_{(i)})$ (see Figure 4) and leaving their covariates and utility function 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 solution of this problem can be found implicitly by extending the $\mathbf{Y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ matrices as follows:

$$\bar{\mathbf{Y}} = \left[\begin{array}{c} \mathbf{Y} \\ \hline B_{(1)} \\ B_{(2)} \\ \cdots \\ B_{(k-1)} \end{array}\right] \quad \text{and} \quad \bar{\mathbf{X}} = \left[\begin{array}{c|c} \mathbf{X} & 0 \\ \hline 0 & I_{k-1} \end{array}\right], \tag{35}$$

where $\bar{\mathbf{X}} \in R^{(n+k-1)\times(d+k-1)}$ and $\bar{\mathbf{Y}} \in \mathbb{R}^{n+k-1}$ and I_{k-1} represents the identity matrix of dimension $k-1$. The problem is then formulated as in equation (23) after replacing $\mathbf{X}$ by $\bar{\mathbf{X}}$ and $\mathbf{Y}$ by $\bar{\mathbf{Y}}$ and results in the parameter vector $\bar{w} = [w; v]$.



(a) minimize Lipschitz constant (b) maximize margin



(c) minimize Lipschitz constant (d) maximize margin

Figure 4: Adaptation of the MINLIP algorithm to ordinal regression: (a) Inclusion of dummy data points with output values B_i intermediate to the observed output values and undefined covariates and utility v_i. 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 (diamond) 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 smoothness 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.

Using the above formulation, the thresholds v as well as the weights w are regularized. Since the motivation for this regularization scheme is not clear, one can formulate the problem explicitly as:

$$\begin{aligned} \min_{w,e,e^*,v} \quad & \|w\|_2 + \gamma \textstyle\sum_{i=1}^{n} 1_n^T (e + e^*) \\ \text{s.t.} \quad & \begin{cases} \mathbf{X}w - \mathbf{Q}v + e \geq \mathbf{Y} - \mathbf{Q}\mathbf{B} \\ -\mathbf{X}w + \mathbf{Q}^* v + e^* \geq -\mathbf{Y} + \mathbf{Q}^*\mathbf{B} \\ e \geq 0 \\ e^* \geq 0 \\ \mathbf{M}v \leq 0\,, \end{cases} \end{aligned} \tag{36}$$

with γ a positive regularization constant, $\mathbf{Q}$ and $\mathbf{Q}^* \in \mathbb{R}^{n \times (k-1)}$ matrices with all elements equal to zero except for positions $\{(i, k_i - 1)\}_{k_i=2}^{k}$ and $\{(i, k_i)\}_{k_i=1}^{k-1}$ respectively (where k_i represents the index of the output level of observation i), which contain ones. These positions correspond to the dummy data points with which one wishes to compare data points i. Vector $\mathbf{B} \in \mathbb{R}^{k-1}$ contains outcomes corresponding to the thresholds: $\mathbf{B} = [B_{(1)}, \cdots, B_{(k-1)}]^T$. Vector v contains all unknown utility function values for the dummy data points $v = [v_{(1)}, \cdots, v_{(k-1)}]^T$, and $\mathbf{M} \in \mathbb{R}^{(k-1)\times k}$ gives the first order differences of a vector and is defined as:

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \dots & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{37}$$

The Lagrange dual problem becomes

$$\begin{aligned} \min_{\alpha,\beta} \quad & \tfrac{1}{2}\alpha^T \mathbf{K}\alpha + \tfrac{1}{2}\beta^T \mathbf{K}\beta - \alpha^T \mathbf{K}\beta - \alpha^T(\mathbf{Y} - \mathbf{B}^T\mathbf{Q}) + \beta^T(\mathbf{Y} - \mathbf{B}^T\mathbf{Q}^*) \\ \text{s.t.} \quad & \begin{cases} 0_n \leq \alpha \leq \gamma 1_n \\ 0_n \leq \beta \leq \gamma 1_n \\ 0_{k-2} \leq \nu \\ \mathbf{Q}^T\alpha - \mathbf{Q}^{*T}\beta + \mathbf{M}^T\nu = 0_{k-1}\,, \end{cases} \end{aligned} \tag{38}$$

where 1_n and 0_n represent column vectors of size n with all elements equal to 1 and 0 respectively. Solving this explicit formulation is computationally less demanding and faster than solving the implicit problem formulation. We refer to Appendix C for a detailed derivation. The estimated $\hat{u}$ can be evaluated at any point $X^* \in \mathbb{R}^d$ as

$$\hat{u}(X^*) = (\hat{\alpha}^T - \hat{\beta}^T)\mathbf{K}_n(X^*)\,, \tag{39}$$

with $\mathbf{K}_n(X^*)$ defined as before.

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 5, the predictions can be easily obtained from the value of the utility function in comparison with the different threshold values.

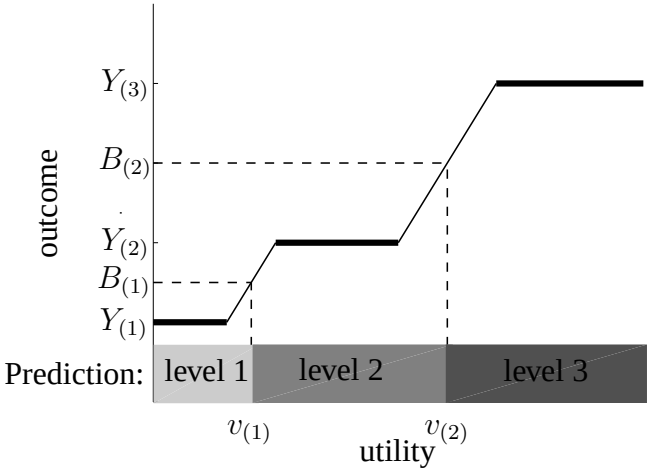


Figure 5: Prediction for ordinal regression. MINLIP for ordinal regression, including unknown thresholds, has the advantage to reduce the prediction step to a simple comparison between the utility of a new observation and the utility of the thresholds. If the utility has a value between threshold $j-1$ and j, the predicted outcome equals the j^{th} output level.

4.3 Difference with Other Methods

Chu and 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 Y_i has two slack variables: one relating to the threshold between classes k_i-1 and k_i and a second relating to the threshold between classes k_i and k_i+1. To ensure that the threshold between classes k_i-1 and k_i is smaller than the threshold between classes k_i and k_i+1, $k-1$ additional constraints are explicitly included. The problem can be written as:

$$\begin{array}{ll} \min\limits_{w,e,e^*,v} & \|w\|_2 + \gamma \sum_{i=1}^{n}(e_i + e_i^*) \\ \text{s.t.} & \begin{cases} w^T X_i - v_j + e_i \geq 1 & \forall\, i = 1,\ldots,n;\ j = \arg\max_j (T_i > v_j) \\ -w^T X_i + v_j + e_i^* \geq 1 & \forall\, i = 1,\ldots,n;\ j = \arg\min_j (T_i < v_j) \\ e_i \geq 0 & \forall\, i = 1,\ldots,n \\ e_i^* \geq 0 & \forall\, i = 1,\ldots,n \\ v_j \leq v_{j+1} & \forall\, j = 1,\ldots,k-2\,. \end{cases} \end{array} \tag{40}$$

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 datapoint. The problem can be formulated as follows:

$$\begin{array}{ll} \min\limits_{w,e,e^*,v} & \|w\|_2 + \gamma \sum_{i=1}^{n}(e_i + e_i^*) \\ \text{s.t.} & \begin{cases} w^T X_i - v_j + e_i \geq 1 & \forall\, i = 1,\ldots,n;\ \forall\, j:\ T_i > v_j \\ -w^T X_i + v_j + e_i^* \geq 1 & \forall\, i = 1,\ldots,n;\ \forall\, j:\ T_i < v_j \\ e_i \geq 0 & \forall\, i = 1,\ldots,n \\ e_i^* \geq 0 & \forall\, i = 1,\ldots,n\,. \end{cases} \end{array} \tag{41}$$

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 k_i and $k_i + 1$ is defined as $\frac{|Y_{(i+1)} - Y_{(i)}|}{||w||_2}$.

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

$$\begin{array}{ll} \min\limits_{w,e,e^*,b} & \|w\|_2 + \gamma \sum_{i=1}^{n}(e_i + e_i^*) \\ \text{s.t.} & \begin{cases} w^T X_i + b + e_i \geq 1 & \forall\, i \in \mathcal{C}_1 \\ -w^T X_i - b + e_i^* \geq 1 & \forall\, i \in \mathcal{C}_2 \\ e_i \geq 0 & \forall\, i \in \mathcal{C}_1 \\ e_i^* \geq 0 & \forall\, i \in \mathcal{C}_2\,. \end{cases} \end{array} \tag{42}$$

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

$$\begin{array}{ll} \min\limits_{w,e,e^*,v} & \|w\|_2 + \gamma \sum_{i=1}^{n}(e_i + e_i^*) \\ \text{s.t.} & \begin{cases} w^T X_i - v + e_i \geq 1 & \forall\, i \in \mathcal{C}_1 \\ -w^T X_i + v + e_i^* \geq 1 & \forall\, i \in \mathcal{C}_2 \\ e_i \geq 0 & \forall\, i \in \mathcal{C}_1 \\ e_i^* \geq 0 & \forall\, i \in \mathcal{C}_2\,, \end{cases} \end{array} \tag{43}$$

which equals the model in equation (42) 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:

$$\begin{array}{ll} \min\limits_{w,e,e^*,v} & \|w\|_2 + \gamma \sum_{i=1}^{n}(e_i + e_i^*) \\ \text{s.t.} & \begin{cases} w^T X_i - v + e_i \geq Y_i - B & \forall\, i \in \mathcal{C}_1 \\ -w^T X_i + v + e_i^* \geq B - Y_i & \forall\, i \in \mathcal{C}_2 \\ e_i \geq 0 & \forall\, i \in \mathcal{C}_1 \\ e_i^* \geq 0 & \forall\, i \in \mathcal{C}_2\,, \end{cases} \end{array} \tag{44}$$

where only one dummy observation (v, B) needs to be introduced. The difference between equations (42,43) and equation (44) lies in the right hand side of the two first inequalities, which

is a consequence of the used complexity control. Models (43) and (44) are equivalent up to the choice of the regularization constant.

Chu and 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.

5 Transformation Models for Failure time Data

We now turn our attention to the case where the data originate from a survival study, i.e. the dependent variable is essentially a time-to-failure and typically requires specific models and tools to capture its behavior. We will adopt a classical statistical setup, and will show how the techniques as described in Section 3 provide a powerful alternative to the classical statistical (semi-parametric) toolbox.

5.1 Survival Data

The observations are assumed to fit in the following statistical setup, see e.g. Kalbfleisch and Prentice (2002) for a more elaborate introduction. Let $T \in \mathbb{R}^+$ and $X \in \mathbb{R}^d$ be a random variable and random vector respectively, jointly following a probability law characterized by P as classical. The former variable T describes the *time to the event* of interest, and the random vector X taking values in $\mathbb{R}^d$ describes d covariates. Note that in this Section T has the same role as Y in the previous Sections. We assume that no ties will occur in the data in order to keep the explanations as simple as possible. We will consider predictive models where the covariates come in through a linear combination with weights $w \in \mathbb{R}^d$ as before, or $\mathcal{U} = \left\{u : \mathbb{R}^d \to \mathbb{R} : u(X) = w^T X,\ \forall\, X \in \mathbb{R}^d\right\}$. A key quantity in survival analysis is the *conditional survival function* $S(t|u(X)) : \mathbb{R}^+ \to [0,1]$ defined as

$$S(t|u(X)) = P\left(T > t \middle|\ u(X)\right), \tag{45}$$

denoting the probability of the event occurring past t given the value of the utility function $u(X) = w^T X$. A related quantity to the conditional survival function is the *conditional hazard function* $\lambda : \mathbb{R} \to \mathbb{R}^+$ defined as

$$\begin{aligned} \lambda(t|u(X)) &= \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t \middle| u(X), T \ge t\right)}{\Delta t} && (46) \\ &= \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t \middle| u(X)\right)}{S\left(t \middle|\ u(X)\right)}\,. && (47) \end{aligned}$$

If the derivative $s : \mathbb{R}^+ \to \mathbb{R}$ with $s(t|u(X)) = \dfrac{\partial S(t|u(X))}{\partial t}$ exists, one can write $\lambda(t|u(X)) = \frac{s(t|u(X))}{S(t|u(X)}$. The conditional hazard function reflects the instantaneous probability that the event will occur given that the subject already survived beyond time t. Finally, one can make the relation between the hazard λ and the survival function S even more explicit by introducing the *conditional cumulative hazard function* $\Lambda(t|u(X)) = \int_0^t \lambda(r|u(X))dr$ for $t \geq 0$ such that

$$\Lambda(t|u(X)) = -\ln\big(S(t \mid u(X))\big). \tag{48}$$

The following Subsection enumerates some commonly used (semi-)parametric methods for modelling the survival and hazard functions.

5.2 Transformation Models for Survival Analysis

The Transformation model (see Definition 2) encompasses a broad class of models, including the following classical survival models.

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

$$S(T, X) = [S_0(T)]^{\exp(-\beta^T X)}\,, \tag{49}$$

where $S_0(t) = \exp(-\Lambda_0(t))$ is called the baseline survival function. Taking $\ln(-\ln(\cdot))$ of both sides in (49) leads to

$$\begin{aligned} \ln(-\ln(S(T,X)) &= \ln(-\ln(S_0(T))) - \beta^T X \\ &= \ln(\Lambda_0(T)) - \beta^T X \\ \Rightarrow\ \epsilon &= g(T) - u(X) \\ \Rightarrow T &= h(u(X) + \epsilon)\,. \end{aligned} \tag{50}$$

Remark 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

$$\ln\left(\frac{F(t|X)}{1 - F(t|X)}\right) = \alpha(t) + \beta^T X\,, \tag{51}$$

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 (51) to

$$\begin{aligned} \ln\left(\frac{1 - S(T|X)}{S(T|X)}\right) &= \alpha(T) + \beta^T X \\ \Rightarrow\ \epsilon &= \alpha(T) + u(X) \\ \Rightarrow T &= h(-u(X) + \epsilon)\,. \end{aligned} \tag{52}$$

Remark that the last transition is only possible if $\alpha(T)$ is invertible.

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

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 and Doksum (1988); Koenker and Geling (2001); Cheng, Wei, and Ying (1997) and citations.

5.3 Censoring

A typical property of failure time data is the occurrence of censoring. A failure time is called censored when the exact time of failure is not observed. Despite this, censored times do provide relevant information. Define $\mathcal{T}_i = (T_i, \delta_i)$ with δ_i the censoring indicator, capturing all censoring information: $\delta = 0$ indicates the occurrence of an event at a known failure time (uncensored data point); right, left and interval censoring are indicated by $\delta = 1$, $\delta = 2$ and $\delta = 3$ respectively. Without censoring all possible pairs of datapoints $\{(\mathcal{T}_i, \mathcal{T}_j)\}_{i \neq j}$ can be used for comparison in equation (15). The presence of censoring leads to a lack of comparability between certain data points. Let $\Delta(\mathcal{T}_i, \mathcal{T}_j)$ be a comparability indicator, indicating whether the datapoints i and j are comparable:

$$\Delta(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 0 \text{ if } \mathcal{T}_i \text{ and } \mathcal{T}_j \text{ are not comparable} \\ 1 \text{ if } \mathcal{T}_i \text{ and } \mathcal{T}_j \text{ are comparable}\,. \end{cases} \tag{53}$$

This indicator is defined depending on the censoring types present in the data:

Right censoring occurs when the event of interest did not occur until the last follow-up time. This type of censoring typically occurs at the end of the study period. Although the exact failure time is not known in this case, the failure time is known to be later than the date of last follow-up. In case of right censoring the comparability indicator Δ takes the value 1 for two observations i and j when the observation with the earliest failure time is observed, and zero otherwise:

$$\Delta(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 1 & \text{if } (T_i < T_j \text{ and } \delta_i = 0) \text{ or } (T_j < T_i \text{ and } \delta_j = 0) \\ 0 & \text{otherwise.} \end{cases}$$

Left censoring deals with the case when the failure is known to have happened before a certain time. An example of left censoring arises in case a variable can only be measured when its value is above a certain level. For left censoring, two observations i and j are comparable when the observation with the highest failure time is non-censored and zero otherwise:

$$\Delta(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 1 & \text{if } (T_i < T_j \text{ and } \delta_j = 0) \text{ or } (T_j < T_i \text{ and } \delta_i = 0) \\ 0 & \text{otherwise.} \end{cases}$$

Interval censoring is a combination of the previous two censoring types. In this case the failure time is not known exactly, instead an interval including the failure time is indicated. This

type of censoring is often found in medical studies where the patients are subject to regular check up times (Finkelstein, 1986). Whether two observations are comparable or not in case of interval censoring depends on the censoring times $\underline{T}_i$ and $\overline{T}_i$ defining the failure interval for each observation i: $T_i \in [\underline{T}_i, \overline{T}_i]$. For uncensored observations, the failure interval reduces to one time, namely the failure time $T_i = \underline{T}_i = \overline{T}_i$. The comparability indicator is defined as:

$$\Delta(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 1 & \text{if } \overline{T}_i < \underline{T}_j \text{ or } \overline{T}_j < \underline{T}_i \\ 0 & \text{otherwise.} \end{cases}$$

In case the data consists of data points with different types of censoring, the comparability indicator is defined as follows. In the most general case, the failure time T_i is considered to be an element of the interval $[\underline{T}_i, \overline{T}_i]$. For right censored data points the right edge of the interval equals infinity, whereas for left censored observation the left edge of the interval equals zero. The comparability indicator is then defined as:

$$\Delta(\mathcal{T}_i, \mathcal{T}_j) = \begin{cases} 1 & \text{if } \overline{T}_i < \underline{T}_j \text{ or } \overline{T}_j < \underline{T}_i \\ 0 & \text{otherwise.} \end{cases}$$

More information on censoring can be found in Andersen et al. (1993); Elandt-Johnson and Johnson (1980); Harrell (2001); Kalbfleisch and Prentice (2002); Miller (1981).

Standard statistical methods for modelling survival data obtain parameter estimates by maximizing a (partial) likelihood with regard to these parameters. This likelihood depends on the ranking of the failure times. In the presence of right censoring, this ranking can uniquely be defined and estimates for the parameters can be obtained. However, in the presence of interval censoring, a unique ranking of the failure of all instances is not always possible. Peto (1972) and Satten (1996) among others, suggested extensions of the proportional hazard model where censoring is not restricted to right censoring. However, estimation of the parameters in these cases remain difficult. In the next section we will illustrate that MINLIP can be easily adapted for right, left, interval censoring and combined censoring schemes. However, we first need an appropriate measure of concordance equivalent to equation (3). Therefore, we resort to the concordance index as described by Harrell et al. (1984); Harrell (2001).

Definition 4 (Concordance Index) *The concordance index (c-index) is a measure of association between the predicted and observed failures in case of censored data. The c-index equals the ratio of concordant to comparable pairs of data points. Two observations i and j are comparable if their relative order in survival time is known. A pair of observations i and j is concordant if they are comparable and the observation with the lowest failure time also has the lowest score for the utility function $u(X)$. Formally, the observation based c-index of a model generating*

predictions $u(X_i)$ for data X_i from a dataset $\mathcal{D} = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ can be expressed as

$$\mathcal{C}_n(u) = \frac{\sum_{i \neq j} \Delta(\mathcal{T}_i, \mathcal{T}_j) I[(u(X_j) - u(X_i))(Y_j - Y_i) > 0]}{\sum_{i \neq j} \Delta(\mathcal{T}_i, \mathcal{T}_j)}. \tag{54}$$

This index is an estimate probability of concordance between predicted and observed survival, with c-index $= 0.5$ for random predictions and c-index $= 1$ for a perfectly concordant model. Without censoring, this definition is exactly equal to the concordance as defined in equation (3).

5.4 Modifications to MINLIP

This section describes how the standard MINLIP model can be extended towards failure time data including the handling of censored data. Therefore, equation (23) is adapted to include censored data. In particular, the matrix $\mathbf{D}$ needs to be changed in order to allow for pairs of data points not to be comparable. Let $\mathbf{R} \in \mathbb{R}^{(n-1)\times(n-1)}$ be defined as the diagonal matrix with $\mathbf{R}_{ii} = \Delta(Z_i, Z_{i+1}), \forall\, i = 1, \ldots, n-1$. The matrix $\mathbf{D}$, representing the datapoints to be compared, is adapted for censoring according to:

$$\mathbf{D}_c = \mathbf{R}\mathbf{D}\,, \tag{55}$$

where Δ is defined as in Section 5.3, resulting in multiple rows with only zero entries in the matrix $\mathbf{D}_c$. For computational convenience these rows can be left out. It is seen that issues concerning the type(s) of censoring in the data are easily dealt with by using the comparability indicator. In the remainder of this paper we will restrict our attention to right censored data.

The learning objective can now be formalized as

$$\begin{array}{ll} \min\limits_{w,e} & \frac{1}{2} w^T w + \gamma \|e\|_1 \\ \text{s.t.} & \mathbf{D}_c(\mathbf{\Phi} w + e) \geq \mathbf{D}_c \mathbf{T}, \end{array} \tag{56}$$

where $\|e\|_1 = \sum_{i=1}^n |e_i|$ and $\mathbf{T} = [T_{(1)}, T_{(2)}, \ldots, T_{(n)}]^T$ is a vector containing all failure times, censored or not. As in Section 3, the dual of this optimization problem becomes

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2} \alpha^T \mathbf{D}_c \mathbf{K} \mathbf{D}_c^T \alpha - \alpha^T \mathbf{D}_c \mathbf{T} \\ \text{s.t.} & \begin{cases} -\gamma 1_n \leq \mathbf{D}_c^T \alpha \leq \gamma 1_n \\ \alpha \geq 0_{n-1}\,, \end{cases} \end{array} \tag{57}$$

Given the solution $\hat{\alpha}$, the predicted utility can be calculated for a new point X^* as

$$u(X^*) = \hat{\alpha}^T \mathbf{D}_c \mathbf{K}_n(X^*), \tag{58}$$

with $\mathbf{K}_n(X^*) = [K(X^*, X_1) \;\ldots\; K(X^*, X_n)]^T \in \mathbb{R}^n$. Since the censoring mechanism can be handled by a proper choice of $\mathbf{D}_c$, it is not too difficult to extend the formulations of Subsection 3.5 as well.

5.5 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 hazard model estimates these functions, by assuming that a baseline hazard function exists; the covariates 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 $\{u_i, T_i, \delta_i\}_{i=1}^n$, the outcome in each time interval is defined as:

$$Y_{il} = \begin{cases} 0 \text{ if } T_i > t_l \\ 1 \text{ if } T_i \leq t_l \text{ and } \delta_i = 0\,. \end{cases} \tag{59}$$

Remark 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 $\hat{F}(u_i, l)$ is estimated. The survival function is found as $\hat{S}(u_i, l) = 1 - \hat{F}(u_i, l)$. The hazard function is then found as

$$\hat{\lambda}(u_i, l) = \frac{\hat{F}(u_i, l+1) - \hat{F}(u_i, l)}{t_{l+1} - t_l}, \forall\, l = 1, \ldots, k-1\,. \tag{60}$$

Remark the analogy with the partial logistic artificial neural network approach to the survival problem proposed by Biganzoli et al. (1998). However, since the latter can not be seen as a transformation model, data replication is necessary even when one is only interested in risk groups. Thanks to the two-step approach of transformation models, data replication can be avoided.

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 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 and Keerthi (2005) and GPOR as proposed in Chu and 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-validation concordance index was used as model selection criterion since the main interest lies in the ranking of the patients.

6.1 Artificial examples

This section illustrates the different steps needed to obtain the desired output for ranking, regression and survival problems, using artificial examples. Together with Table 1, this Subsection illustrates the different tasks considered in the paper.

6.1.1 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 10^{th} race , we want to predict the rank of the remaining 50. 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 (23) 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 6). Since one is only interested in ranking the cyclists, the value of the utility is irrelevant. Additionally, one is not interested in estimating the transformation function h.

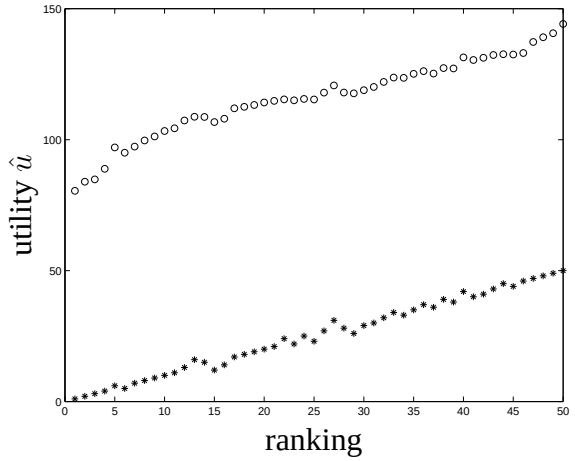


Figure 6: Artificial example illustrating the use of the MINLIP transformation model for the ranking setting. The estimated utility of the test observations are denoted by the circles. The value of the utility does not correspond to the ranking. However, the rank of the estimated utility (denoted by stars) are a good prediction of the observed rank.

6.1.2 Ordinal regression

In a second artificial example, consider the scenario in which one wishes to divide students into 3 groups: bad, average and good student. 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.

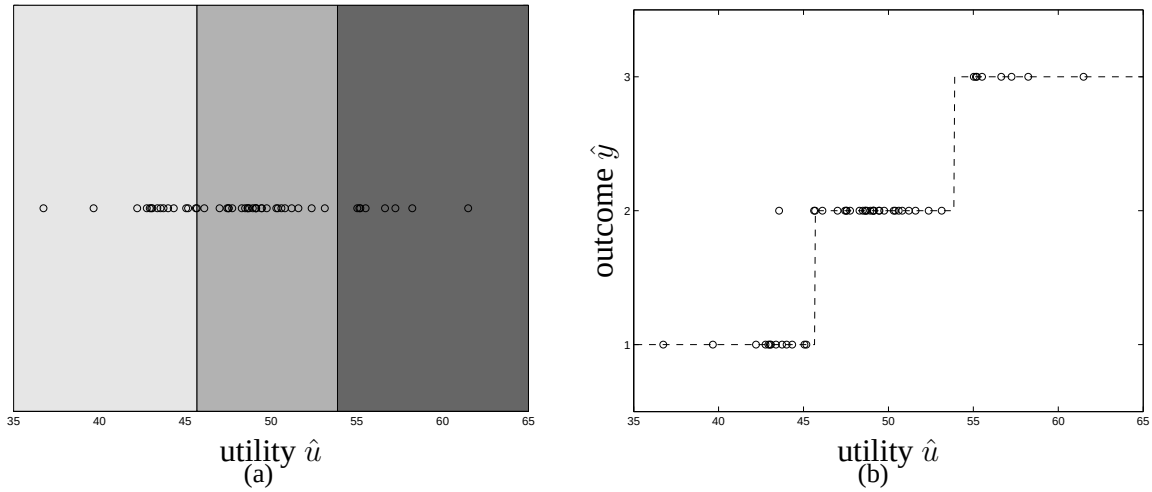


Figure 7: Artificial example illustrating the use of the MINLIP transformation model for ordinal regression. (a) The estimated utility of the test observations are denoted by the circles. The MINLIP model for ordinal regression results in an estimate of the utility function and threshold values. Students 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).

In a first step, all students are ranked according to the available data, namely their grades on 10 courses. Using equation (36) 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 utility, the MINLIP model for ordinal regression gives threshold values which can be used to predict the outcome of new observations (see Figure 7). Since, the MINLIP model generates these thresholds, no additional model is needed to obtain the transformation function.

6.1.3 Survival analysis

In a last artificial example, a randomized trial is simulated. Assume 150 patients are randomly divided into 2 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. Assuming that the age is irrelevant for the survival, the treatment will be the only important factor in predicting the patients' survival.

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 8 illustrates that the MINLIP model is able to divide the group of test patients into two groups with a significant different survival (p=0.03, logrank test). The first part of the transformation model obtained 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 5.5, the estimated survival curves for all patients are calculated (Figure 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.

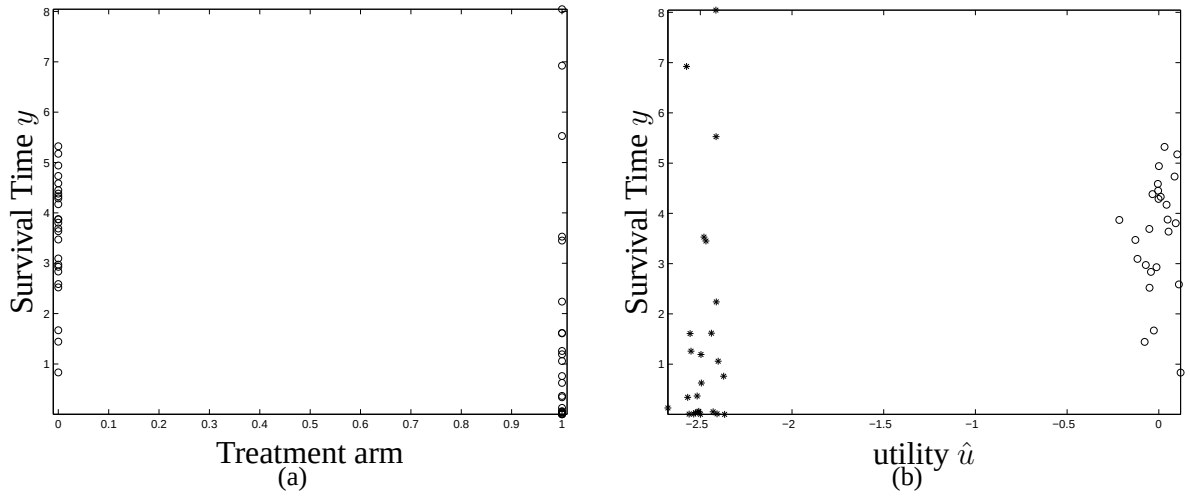


Figure 8: Artificial example illustrating the use of the MINLIP transformation model for survival analysis. (a) Survival time as a function of the treatment arm. Patients receiving the first treatment survive longer in general. The second treatment results in lower survival times. However, some patients 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).

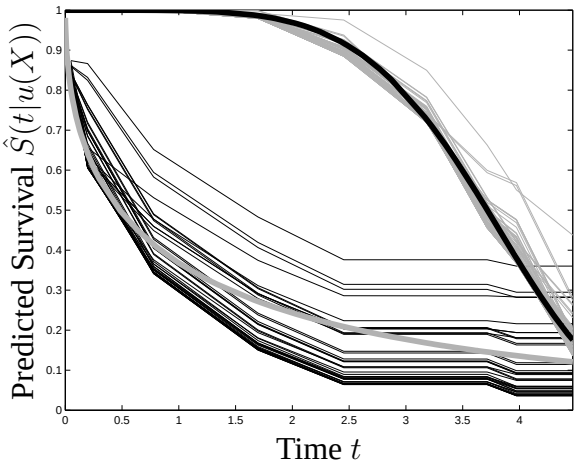


Figure 9: Artificial example illustrating the use of the MINLIP transformation model for survival analysis: illustration of the reconstruction step. For each patient, the survival curve is calculated using the method discussed in Section 5.5. Grey and black curves represent patients from the first and 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.

6.2 Ordinal Regression

At first 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 and Keerthi (2005) (see Table 2). Both of these methods optimize multiple thresholds to define parallel discriminant hyperplanes for the ordinal levels. The first method (EXC) explicitly imposes the ordering of the thresholds, whereas this is done only implicitly in the second method (IMC). Tuning of the Gaussian kernel parameter and the regularization 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. The 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.

6.3 Failure Time Data: Micro-array Studies

The MINLIP technique is derived from machine learning techniques as SVMs, techniques which are shown to be especially useful to handle high-dimensional datasets. We therefore test the

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

Table 2: Test results of MINLIP, EXC and IMC using a Gaussian kernel and GPOR. The targets of the datasets 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 rank sum test between the EXC, IMC, GPOR and the MINLIP (reference) model are indicated with * $(\mathrm{p} < 0.05)$, ** $(\mathrm{p} < 0.01)$ or *** $(\mathrm{p} < 0.001)$.

dataset	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*****
machine CPU	0.65±0.04	0.65±0.06	**0.64±0.04**	0.80±0.09***
auto MPG	**0.58±0.03**	**0.58±0.03**	/	0.68±0.08***
Boston housing	**0.57±0.04**	**0.57±0.04**	/	0.61±0.03***
dataset	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***	**0.10±0.00*****	0.16±0.02***
Wisconsin	29.56±2.18	28.80±1.17	**28.23±1.28**	34.52±4.27***
machine CPU	**29.41±4.27**	30.57±6.94	30.48±4.10	157.00±124.67***
auto MPG	2.08±0.16	**2.05±0.18**	/	4.13±1.13***
Boston housing	2.49±0.27	**2.33±0.25***	/	2.97±0.29***

performance of MINLIP on 3 micro-array datasets.

In this example we compare the performance of model (56) (MINLIP) and linear extensions as in equations (33) ($\mathrm{MINLIP}_{\mathrm{L1}}$) and (34) ($\mathrm{MINLIP}_{\mathrm{p}}$) 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 (Cox, 1972): ridge regression ($\mathrm{Cox}_{\mathrm{L2}}$) and $L1$ regularization ($\mathrm{Cox}_{\mathrm{L1}}$). 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 Cox regression model (see Hastie, Tibshirani, and Friedman, 2001, chapter 3.4.4). The SPCR (Bair and Tibshirani, 2004; Bair et al., 2006) method 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 and Næs, 1989). The application of PLS to the Cox model 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.

Three different micro-array datasets are used in this experiment:

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

The Norway/Stanford breast cancer data (NSBCD) from Sørlie et al. (2003) contains gene expression measurements from 115 women with breast cancer. 549 intrinsic genes introduced in Sørlie et al. (2003) were used. Missing values were previously imputed using 10-nearest neighbor imputation (Bøvelstad et al., 2007). 38 (33%) patients experienced an event.

Figure 10 summarizes performances $\mathcal{C}_n^u$ 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 concordance on the 100 test sets. In the right panels of Figure 10 the time dependent receiver operator characteristics (TDROC) (Heagerty, Lumley, and Pepe, 2000) are shown. The TDROC 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 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 a significantly higher χ^2 was obtained for the former model.

6.4 Failure Time Data: Cancer Study

In this last example, we investigate the ability of the MINLIP model to estimate how the different covariates influence the survival time. We use the German Breast Cancer Study Group data[4] (Schumacher et al., 1994), containing information on 686 patients and 8 variables. Available variables are: hormonal treatment, age, menopausal status, tumor size, tumor grade, the number of positive lymph nodes, the progesterone receptor (fmol) and the estrogen receptor (fmol). 299 (43.6%) patients had a breast cancer related event within the study period, leaving all other patients with a right censored failure time. The dataset was randomly divided in training and test set (2/3 versus 1/3).

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

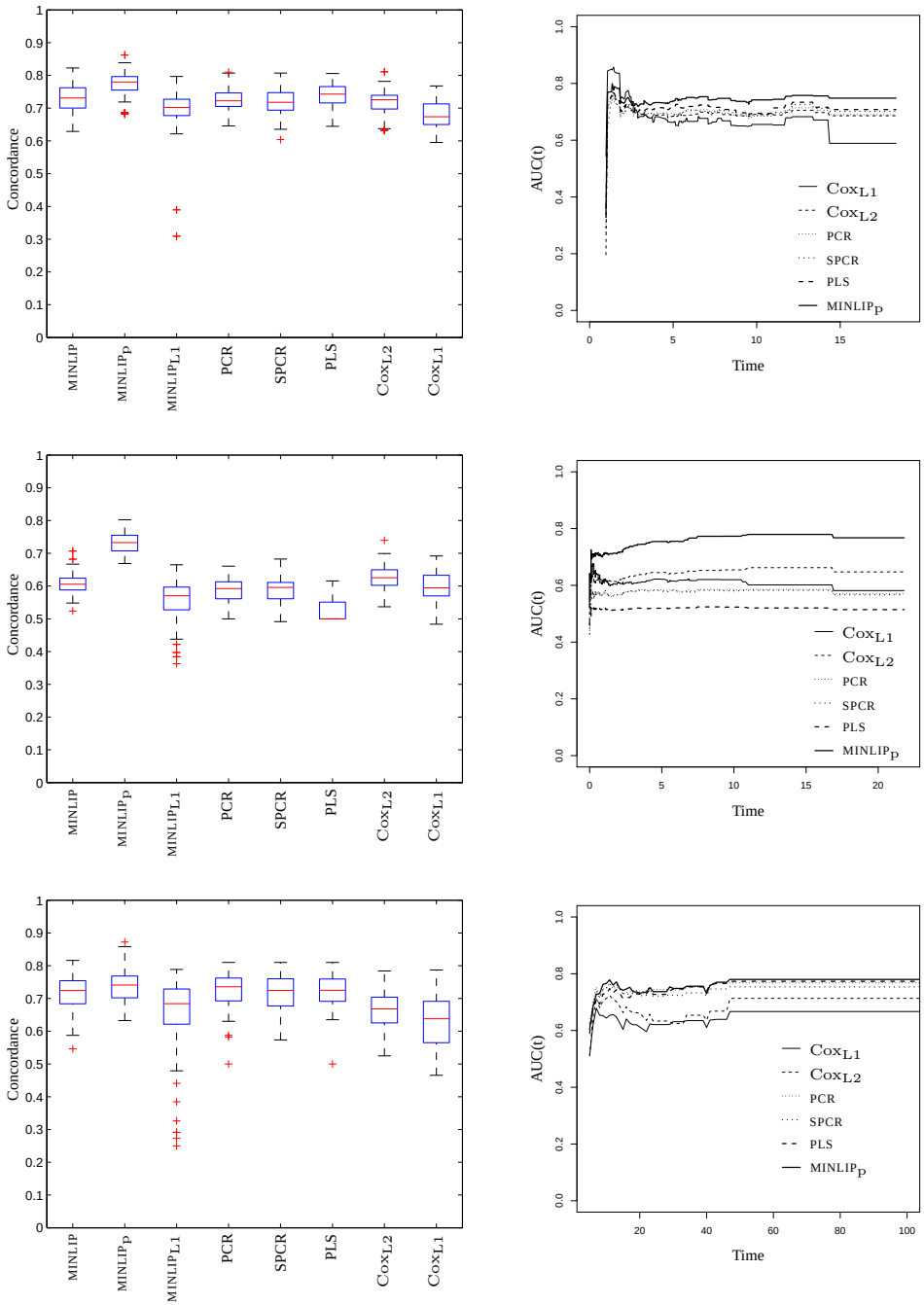


Figure 10: Concordance (left) and time dependent receiver operating characteristic curve (TDROC) (right) on the test set for three micro-array survival data sets (top: DBCD, middle: DLBCL, bottom: NSBCD). The MINLIP$_p$ model obtains a significantly higher performance than all other models. The TDROC for the MINLIP$_p$ model is higher than for the other models indicating that this model is better in distinguishing patients with low and high risk for relapse.

Since medical data are typically not highly non-linear, we use a componentwise polynomial kernel

$$K(X, Z) = \sum_{p=1}^{d} (\tau + X^{pT} Z^p)^2, \ \tau \geq 0 \,, \tag{61}$$

with d the number of variables and X^p the p^{th} covariate, to model non-linearities. Model selection is done by 10-fold cross-validation with the concordance index as model selection criterion.

We compare our results with Cox' proportional hazard model. However, the standard Cox model (Cox, 1972) assumes linearity in the covariates, 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 in the effects of the covariates on the hazard, the functional forms of the covariates were estimated using penalized smoothing splines (Eilers and Marx, 1996; Hurvich, Simonoff, and Tsai, 1998). In this method, a comparative small set of basis functions is fit and a likelihood penalizing the integrated second derivatives (see Therneau and Grambsch, 2000, Section 5.5) is used to estimate the coefficients. Akaike's information criterion (AIC=log likelihood - degrees of freedom) is used to select the degrees of freedom for each term.

Figures 11 and 12 show the estimated covariate effects for Cox regression with penalized splines and MINLIP, respectively. Remark that in Figure 11 the estimates are inversely related with the survival time, whereas in Figure 12 the estimates are related with the survival time itself. Cox' model predicts a decreasing risk for relapse for older patients, up to the age of 40, whereafter the risk increases slowly; for tumors up to 20mm 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 of the covariate effects agree with what is 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 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. Cox' model with penalized smoothing splines obtains a concordance on the test set equal to 0.6715, while the MINLIP model obtains a performance of 0.6857.

Figure 13 illustrates the ability of the models to generate prognostic indices. In clinical practice one is interested in groups of patients with low/high risk for the event to occur. Therefore the median value of the model output is used as a threshold to divide the test set into two groups: one group including patients with an estimated risk lower than the average and a second group with an estimated risk higher than the average. Kaplan-Meier curves and 95%-confidence intervals are plotted in Figure 13. The logrank test χ^2 value is 20.4281 and 29.6984 for Cox and MINLIP respectively. The latter method results in a better split between low and high risk patients.

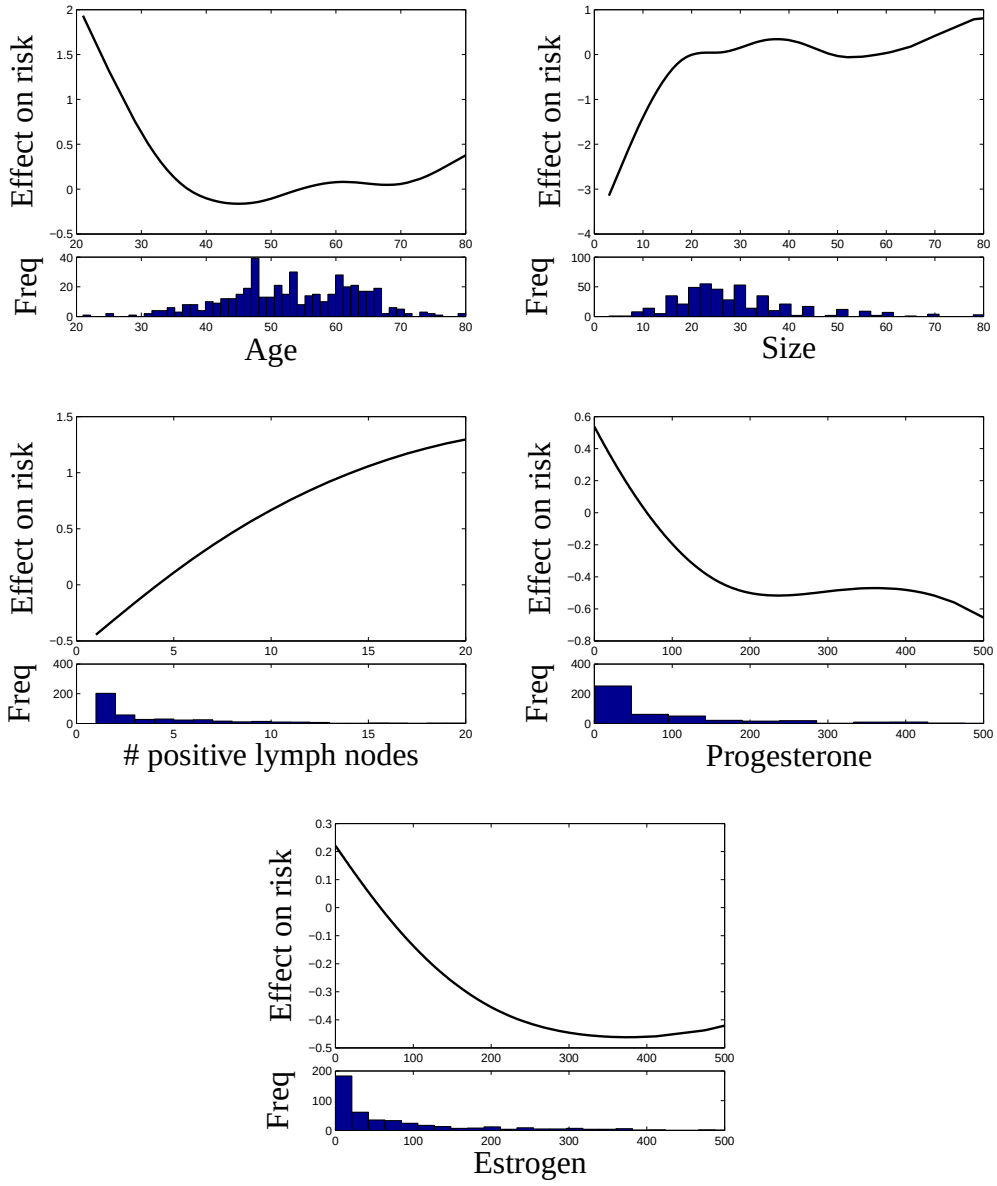


Figure 11: Estimation of the covariate effects on the risk of relapse (remark the difference with Figure 12) with smoothing splines within Cox' proportional hazard model and histograms of the variables. The estimated effects are inversely related with the survival time. The model estimates a lower chance for relapse for older patients up to the age of 40, whereafter the risk increases again, albeit slowly. The chance for relapse increases for larger tumors until a size of 20mm, whereafter the chance 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.

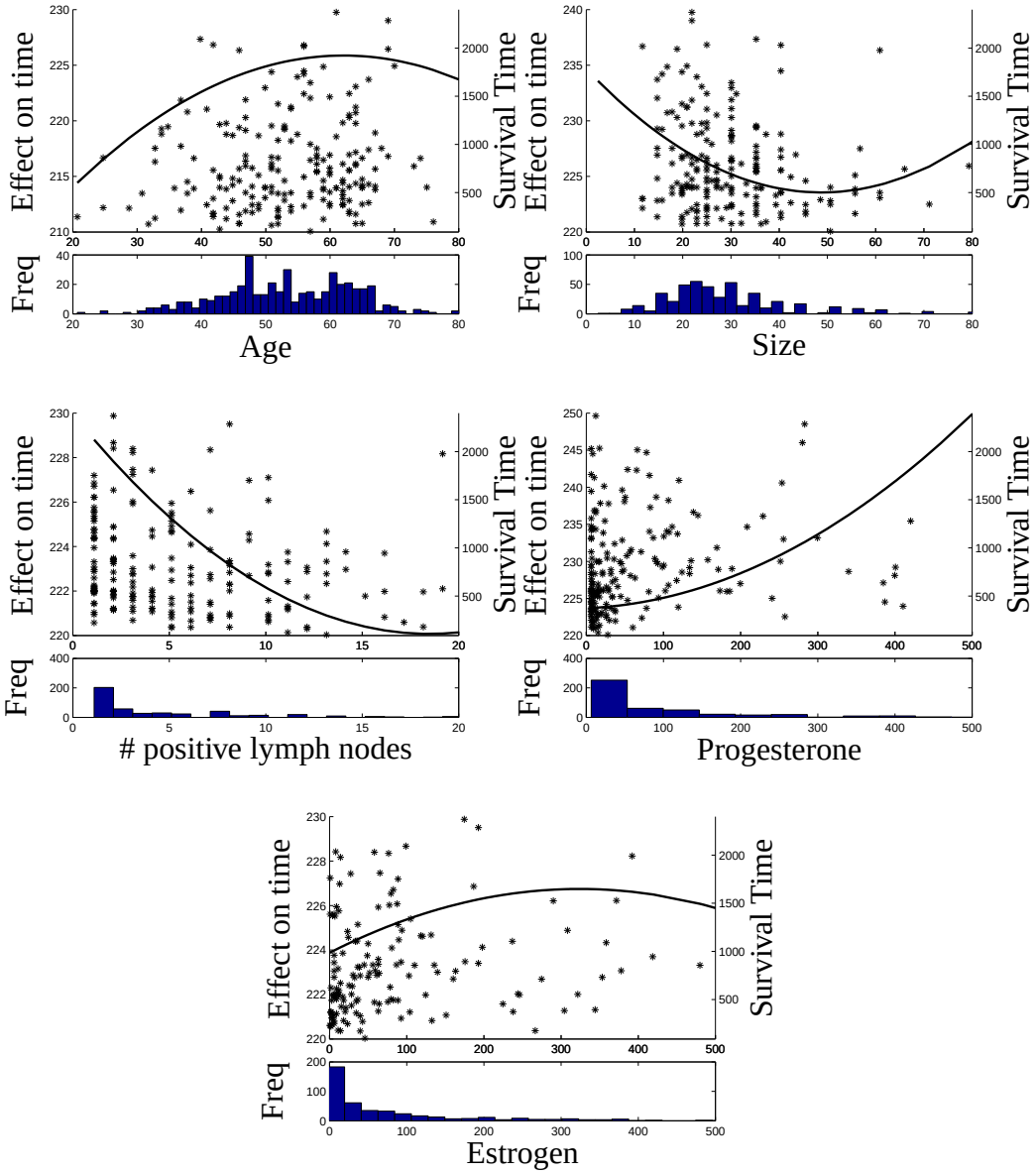
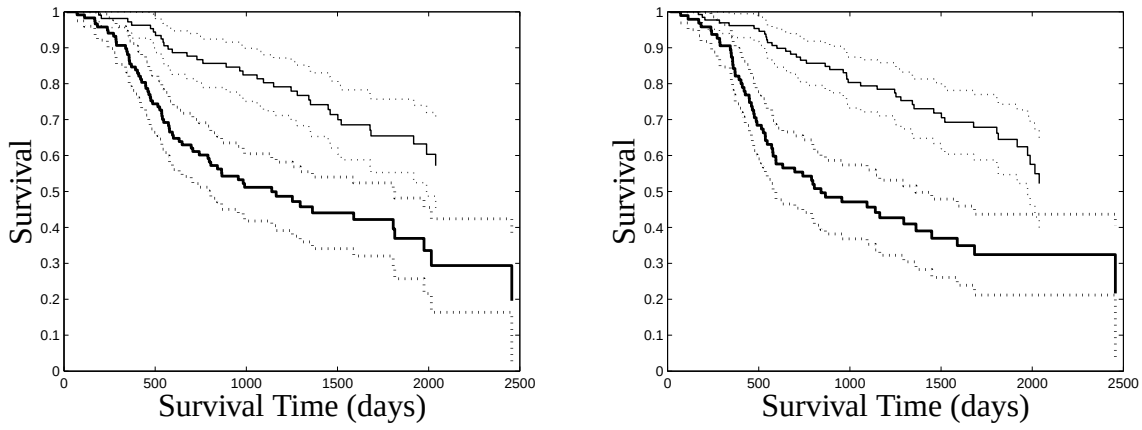


Figure 12: Estimation of covariate effects on survival time (remark the difference with Figure 11) with the MINLIP model ($\mathcal{C}_n(u)$ was used for model selection) and histograms of the variables. The stars indicate the observed failure times for breast cancer related events. The estimated covariate 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 13: The use of Cox' 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 larger value of the log rank test statistic.

7 Conclusions

This paper 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 curve and others. The derivations go along the same lines as used for support vector machines, except for replacing the notion of (pairwise) margin with a Lipschitz smoothness condition on the transformation function. The presented learner finds a (non-linear) non-parametric transformation model by solving a convex Quadratic Program. Extensions towards tasks where transformation models provide only a (good) approximation (agnostic case), 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.

Acknowledgments

This research is supported by Research Council KUL: GOA AMBioRICS, GOA MaNet, CoE EF/05/006, IDO 05/010 IOF-KP06/11, IOF-SCORES4CHEM, several PhD, postdoc end fellow grants; Flemish Government: FWO: PhD and postdoc grants, G.0407.02, G.0360.05, G.0519.06, FWO-G.0321.06, G.0341.07, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0302.07; IWT: PhD Grants, McKnow-E, Eureka-Flite; Belgian Federal Science Policy Office: IUAP P6/04

(DYSCO, 'Dynamical systems, control and optimization', 2007-2011); EU: FP6-2002-LIFESCI-HEALTH 503094, IST-2004-27214, FP6-MC-RTN-035801; Prodex-8 C90242; EU: ERNSI; the European Research Council under the Seventh Framework Program and within Advanced Grant no. 247035 "Systems and Signals Tools for Estimation and Analysis of Mathematical Models in Endocrinology and Neurology". V. Van Belle is supported by a grant from the IWT. K. Pelckmans is an associated professor/researcher ('forskarassistent') at Uppsala University, Sweden at the department of Information Technology, division of SysCon. S. Van Huffel is a full professor and J.A.K. Suykens is a professor at the Katholieke Universiteit Leuven, Belgium.

A Consistency and identifiability

This first Appendix deals with the issues of consistency and identifiability of the proposed method. We study the question under what conditions MINLIP is consistent, i.e. if we have enough datapoints at our disposal, would the estimate $\hat{w}$ converge to the desired parameter vector?

Assume that any observation $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ would obey the relation

$$Y = h_0(w_0^T X), \tag{62}$$

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. We will focus on estimating the vector of parameters w_0, recovery of h_0 may be done in a second stage. Note that assuming that a finite value L_0 exists, together with an observation (X, Y) where this Lipschitz constant is met, is sufficient for consistency. We will fix $\|w_0\|_2 = 1$ to avoid the obvious identifiability issue, namely that for any strictly positive constant $\alpha > 0$, the system $Y = h_{0,\alpha}(\alpha w_0^T X)$ is not distinguishable from (62) when $h_{0,\alpha}(Z) \triangleq h_0(\frac{Z}{\alpha})$ for any $Z \in \mathbb{R}$.

Let the set of all (possibly an infinite number) observations $\{(X, Y)\} \subset \mathbb{R}^d \times \mathbb{R}$ obeying the system (62) be denoted as $\mathcal{D}$. We consider that this set is an ϵ-non-degenerate set, which is defined as follows

Definition 5 (An ϵ-non-degenerate Set) *Let $\epsilon > 0$ be any (arbitrarily small) constant. We say that a set $\mathcal{D} = \{(X, Y)\} \subseteq \mathbb{R}^d \times \mathbb{R}$ is ϵ-non-degenarate if for any observation $(X, Y) \in \mathcal{D}$, and for any vector $v \in \mathbb{R}^d$, one has an observation $(X', Y') \in \mathcal{D}$ different from (X, Y) such that $\|X - X'\|_2 \leq \epsilon$ so that*

$$v^T(X - X') \geq 0. \tag{63}$$

This requirement can be relaxed as it only has to hold for the point (X, Y) where the Lipschitz condition is met. In addition, we assume that h_0 is (L_0, a)-Lipschitz on this set $\mathcal{D} \subseteq \mathbb{R}$.

Definition 6 (h_0 is (L_0, a)-Lipschitz on $\mathcal{D}' \subseteq \mathbb{R}$) *The monotonically increasing function h_0 is said to be (L_0, a)-Lipschitz on $\mathcal{D} \subseteq \mathbb{R}$ if (1) h_0 is Lipschitz smooth with Lipschitz constant L_0 for all pairs Z, Z^*, with $Z \geq Z^*$:*

$$h_0(Z) - h_0(Z^*) \leq L_0 \left(Z - Z^*\right), \tag{64}$$

and (2) there exists a pair $Z, Z' \in \mathcal{D}$ with $Z > Z'$ where the Lipschitz constant is met:

$$h_0(Z) - h_0(Z') = L_0 \, (Z - Z') \, , \tag{65}$$

and (3) one has for any $\epsilon > 0$ and $Z'' \in \mathcal{D}'$ where $0 \leq Z - Z'' \leq \epsilon$ that

$$\frac{L_0}{1 + a\epsilon}(Z - Z'') \leq h_0(Z) - h_0(Z''), \tag{66}$$

with $a \geq 0$.

Hence a denotes how 'smooth' the constant L_0 decays in a neighborhood of Z where the actual Lipschitz constraint is met (that is, a smaller a indicates higher smoothness) (see Figure 14). In particular, a value $a \to 0$ (arbitrarily small) implies that the function h_0 is linear with slope L_0. Note that this definition does not require that the function $\dfrac{\partial h_0(Z)}{\partial Z}$ exists for any $Z \in \mathbb{R}$. This definition implies the inequality

$$\frac{1}{L_0} = \frac{Z - Z'}{h_0(Z) - h_0(Z')} \leq \frac{Z - Z''}{h_0(Z) - h_0(Z'')} \leq \frac{Z - Z'}{h_0(Z) - h_0(Z')} + \frac{a\epsilon}{L_0} \, . \tag{67}$$

The first inequality holds due to the Lipschitz smoothness of h_0. The second inequality follows from (66).

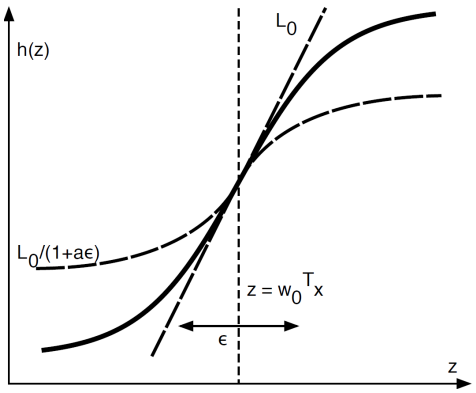


Figure 14: *Illustration of a function $h : \mathbb{R} \to \mathbb{R}$ (solid curved line) which is (L_0, a)-Lipschitz monotone according to Definition 6. The Lipschitz inequality is tight around Z indicated by the dotted line, while in the ϵ-neighborhood of Z the actual Lipschitz constant decays as slowly as $\frac{L_0}{1+a\epsilon}$.*

We now state that (w_0, h_0) can be recovered successfully ('identified') if $\mathcal{D}$ and h_0 were such that Definition 5 and 6 hold. We consider the MINLIP estimator based on an ϵ-non-degenerate set $\mathcal{D}$ which is defined as

$$w_\epsilon = \arg \max_{\|w\|_2 = 1} \inf_{(X,Y),(X',Y') \in \mathcal{D} : Y > Y'} \frac{w^T (X - X')}{Y - Y'} \, , \tag{68}$$

or equivalently (up to a scaling)

$$w_\epsilon \propto \quad \arg\min_{w} \quad \frac{1}{2} w^T w \qquad (69)$$
$$\text{s.t.} \qquad (Y - Y') \leq w^T (X - X') \quad \forall\, (X, Y), (X', Y') \in \mathcal{D} : Y > Y'.$$

Specifically $w_\epsilon = \frac{\hat{w}}{\|\hat{w}\|_2}$ where $\hat{w}$ is the optimizer of (69). If $\mathcal{D}$ contains a finite number of elements this problem can be solved efficiently as a convex Quadratic Program (QP) using standard solvers. This estimator would return the desired result $w_\epsilon = w_0$ if enough observations were given. This is stated more formally as follows.

Lemma 2 (Identifiability) *Let $\epsilon > 0$ be any (arbitrarily small) constant. Given a model (h_0, w_0) governing the observations in $\mathcal{D}$. Assume that (i) the set $\mathcal{D}$ is $(a\epsilon)$-non-degenerate as in Definition 5; (ii) the function h_0 is (L_0, a)-Lipschitz monotone on the set $\mathcal{D}' = \{Z = w_0^T X \in \mathbb{R} : (X, Y) \in \mathcal{D}\}$, as in Definition 6. Then one has for all $w \in \mathbb{R}^d$ where $\|w\|_2 = 1$ that*

$$\frac{1}{L_0} \leq \inf_{(X,Y),(X',Y') \in \mathcal{D} : Y > Y'} \frac{w^T (X - X')}{Y - Y'}, \qquad (70)$$

with equality if $w = w_0$.

Proof: Let $(X, Y), (X', Y') \in \mathcal{D}$ be such that $Y \neq Y'$ and the Lipschitz constant is achieved, or

$$\frac{1}{L_0} = \frac{w_0^T (X - X')}{h_0(w_0^T X) - h_0(w_0^T X')} = \frac{w_0^T (X - X')}{Y - Y'}. \qquad (71)$$

Such an observation exists assuming that h_0 is (L_0, a)-Lipschitz on $\mathcal{D}'$. We prove that for a bad estimation w of w_0 ($w^T w_0 < 1 - \epsilon$), one can always find an observation $(X^*, Y^*) \in \mathcal{D}$ such that $\frac{w^T (X - X^*)}{Y - Y^*}$ is strictly lower than $\frac{1}{L_0}$. This implies that when w deviates a fraction ϵ from w_0, the objective in (69) can never achieve the maximum value as would be the case when $w = w_0$. This implies consistency of the MINLIP estimator.

At first, by the (L_0, a)-Lipschitz condition on h_0, one has for all $(X'', Y'') \in \mathcal{D}$ where $|w_0^T (X - X'')| \leq \epsilon$ and $Y \neq Y''$ that (as in inequality (67)),

$$\frac{w_0^T (X - X')}{Y - Y'} \geq \frac{w_0^T (X - X'')}{Y - Y''} - \frac{a\epsilon}{L_0}. \qquad (72)$$

According to the Cauchy-Schwarz' inequality, the condition $|w_0^T (X - X'')| \leq \epsilon$ is fulfilled for $||X - X''||_2 \leq \epsilon$. Secondly, for any $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ and $w_0^T w < 1 - a\epsilon$, one has by the orthogonal decomposition of a vector that

$$\begin{aligned} w_0 - w &= \frac{w_0 w_0^T}{||w_0||_2} (w_0 - w) + v && (73) \\ &= w_0 (w_0^T w_0 - w_0^T w) + v && (74) \\ &= w_0 a\epsilon^+ + v\,, && (75) \end{aligned}$$

with v the orthogonal complement of the projection of $w_0 - w$ on w_0 and $\epsilon^+ > \epsilon$. It follows for any $(X''', Y''') \in \mathcal{D}$ where $Y \neq Y'''$ that

$$\frac{(w_0 - w)^T(X - X''')}{Y - Y'''} = \frac{a\epsilon^+ w_0^T(X - X''')}{Y - Y'''} + \frac{v^T(X - X''')}{Y - Y'''}\,. \tag{76}$$

Hence by assumption of the set $\mathcal{D}$ being $(a\epsilon)$-non-degenerate, there exists for any $w \in \mathbb{R}^d$ (and thus for any $v \in \mathbb{R}^d$) an observation $(X^*, Y^*) \in \mathcal{D}$ with $\|X - X^*\|_2 \leq a\epsilon$, $Y \neq Y^*$ such that

$$\frac{(w_0 - w)^T(X - X^*)}{Y - Y^*} = a\epsilon^+ \frac{w_0^T(X - X^*)}{Y - Y^*} > a\epsilon \frac{w_0^T(X - X^*)}{Y - Y^*} \geq \frac{a\epsilon}{L_0}. \tag{77}$$

From (72) and (77) it then follows that

$$\frac{1}{L_0} = \frac{w_0^T(X - X')}{Y - Y'} \geq \frac{w_0^T(X - X^*)}{Y - Y^*} - \frac{a\epsilon}{L_0} > \frac{w^T(X - X^*)}{Y - Y^*}. \tag{78}$$

Hence, for all $w \in \mathbb{R}^d$ for which $\|w\|_2 = 1$ and $w_0^T w < 1 - a\epsilon$, there are two observations $(X, Y), (X^*, Y^*) \in \mathcal{D}$ such that

$$\frac{1}{L_0} > \frac{w^T(X - X^*)}{Y - Y^*}, \tag{79}$$

proving the result. Equality as in (71) is reached for $w = w_0$. ■

B MINLIP for ranking problems

The formal derivation of the MINLIP method is given in this Appendix. We start with the problem formulation as denoted in equation (23):

$$\begin{array}{ll} \min\limits_{w,\epsilon} & \frac{1}{2}w^T w + \gamma\|\epsilon\|_1 \\ \text{s.t.} & \mathbf{D}(\mathbf{\Phi}w + \epsilon) \geq \mathbf{D}\mathbf{Y}\,, \end{array} \tag{80}$$

with $\mathbf{\Phi} = [\varphi(X_1), \cdots, \varphi(X_n)]^T$. Take $\epsilon = e^+ + e^-$ and suppose $e^+ \geqslant 0$ and $e^- \geqslant 0$. The problem can than be formulated as

$$\begin{array}{ll} \min\limits_{w,e^+,e^-} & \frac{1}{2}w^T w + \gamma 1_n^T(e^+ + e^-) \\ \text{s.t.} & \begin{cases} \mathbf{D}(\mathbf{\Phi}w + (e^+ - e^-)) \geq \mathbf{D}\mathbf{Y}, \\ e^+ \geqslant 0, \\ e^- \geqslant 0. \end{cases} \end{array} \tag{81}$$

The Lagrangian becomes

$$\mathcal{L}(w, e^+, e^-; \alpha, \beta_+, \beta_-) = \frac{1}{2}w^T w + \gamma 1_n^T(e^+ + e^-) - \beta_+^T e^+ - \beta_-^T e^- - \alpha^T \mathbf{D}(\mathbf{\Phi}w + e^+ - e^- - \mathbf{Y}) \tag{82}$$

with Lagrange multipliers $\alpha,\ \beta^+, \beta^- \geqslant 0$. The conditions for optimality (Karush-Kuhn-Tucker (KKT) conditions) become

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = (\Phi)^T \mathbf{D}^T \alpha \\ \frac{\partial \mathcal{L}}{\partial e^+} = 0 \rightarrow \gamma = \mathbf{D}^T \alpha + \beta^+ \\ \frac{\partial \mathcal{L}}{\partial e^-} = 0 \rightarrow \gamma = -\mathbf{D}^T \alpha + \beta^- \\ \mathrm{diag}(\alpha)\mathbf{D}(\mathbf{\Phi} w + e^+ - e^- - \mathbf{Y}) = 0 \\ \mathrm{diag}(\beta^+)e^+ = 0 \\ \mathrm{diag}(\beta^-)e^- = 0 \\ \alpha \geqslant 0 \\ \beta^+ \geqslant 0 \\ \beta^- \geqslant 0\,, \end{cases} \tag{83}$$

where $\mathrm{diag}(a)$ indicates a diagonal matrix with the elements of the vector a on the main diagonal. Now from Slater's condition one could exchange $\min_{w,e^+,e^-} \max_\alpha$ with $\max_\alpha \min_{w,e^+,e^-}$. Solving for w, e^+ and e^- gives the dual problem

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\alpha^T \mathbf{D}\mathbf{K}\mathbf{D}^T \alpha - \alpha^T \mathbf{D}\mathbf{Y} \\ \text{s.t.} & \begin{cases} -\gamma 1_n \leq \mathbf{D}^T \alpha \leq \gamma 1_n \\ \alpha \geq 0_{n-1} \end{cases} \end{array} \tag{84}$$

and from the first condition of (83) and the model specification $u(X) = w^T \varphi(X)$ one could write the solution for a new point X^* as

$$\hat{u}(X^*) = \mathbf{K}_n^* \mathbf{D}^T \hat{\alpha}\,, \tag{85}$$

with $\mathbf{K}_n^* \in \mathbb{R}^n$ and $\mathbf{K}_n^* = [K(X^*, X_1)\ \ldots\ \ K(X^*, X_n)]^T$.

C 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 v 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 threshold are increasing values of the utility function. More formally the problem is formulated as in equation (36), now using the kernel based version:

$$\begin{aligned}
\min_{w,e,e^*,v} \quad & \|w\|_2 + \gamma 1_n^T (e + e^*) \\
\text{s.t.} \quad & \begin{cases} \mathbf{\Phi} w - \mathbf{Q} v + e \geq \mathbf{Y} - \mathbf{QB} \\ -\mathbf{\Phi} w + \mathbf{Q}^* v + e^* \geq -\mathbf{Y} + \mathbf{Q}^* \mathbf{B} \\ e \geq 0 \\ e^* \geq 0 \\ \mathbf{M} v \leq 0 \,. \end{cases}
\end{aligned} \tag{86}$$

As in Appendix B we build the Lagrangian

$$\begin{aligned}
\mathcal{L}(w, e^+, e^-; \alpha, \beta, \eta, \eta_*, \nu) = \quad & \tfrac{1}{2} w^T w + \gamma 1_n^T (e + e^*) - \alpha^T (\mathbf{\Phi} w - \mathbf{Q} v + e - \mathbf{Y} + \mathbf{QB}) \\
& - \beta^T (-\mathbf{\Phi} w + \mathbf{Q}^* v + e^* + \mathbf{Y} - \mathbf{Q}^* \mathbf{B}) - \eta^T e - {\eta^*}^T e^* \\
& + \nu^T \mathbf{M} v \,,
\end{aligned} \tag{87}$$

and derive the set of optimality conditions

$$\begin{cases}
\frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \mathbf{\Phi}^T (\alpha - \beta) \\
\frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \gamma 1_n = \alpha + \eta \\
\frac{\partial \mathcal{L}}{\partial e^*} = 0 \rightarrow \gamma = \beta + \eta^* \\
\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow \alpha^T \mathbf{Q} - \beta^T \mathbf{Q}^* + \nu^T \mathbf{M} v = 0 \\
\mathrm{diag}(\alpha) D(\mathbf{\Phi} w - \mathbf{Q} v + e - \mathbf{Y} + \mathbf{QB}) = 0 \\
\mathrm{diag}(\beta)(-\mathbf{\Phi} w + \mathbf{Q}^* v + e^* + \mathbf{Y} - \mathbf{Q}^* \mathbf{B}) = 0 \\
\mathrm{diag}(\eta) e = 0 \\
\mathrm{diag}(\eta^*) e^* = 0 \\
\mathrm{diag}(\nu) \mathbf{M} v = 0 \\
\alpha \geqslant 0 \\
\beta \geqslant 0 \\
\eta \geqslant 0 \\
\eta^* \geqslant 0 \\
\nu \geqslant 0 \,.
\end{cases} \tag{88}$$

The dual problem formulation is than found as

$$\begin{aligned}
\min_{\alpha,\beta} \quad & \tfrac{1}{2} \alpha^T \mathbf{K} \alpha + \tfrac{1}{2} \beta^T \mathbf{K} \beta - \alpha^T \mathbf{K} \beta - \alpha^T (\mathbf{Y} - \mathbf{B}^T \mathbf{Q}) + \beta^T (\mathbf{Y} - \mathbf{B}^T \mathbf{Q}^*) \\
\text{s.t.} \quad & \begin{cases} 0_n \leq \alpha \leq \gamma 1_n \\ 0_n \leq \beta \leq \gamma 1_n \\ 0_{k-2} \leq \nu \\ \mathbf{Q}^T \alpha - {\mathbf{Q}^*}^T \beta + \mathbf{M}^T \nu = 0_{k-1} \,. \end{cases}
\end{aligned} \tag{89}$$

References

S. Clémençon, G. Lugosi, N. Vayatis, P. Aurer, and R. Meir. Ranking and scoring using empirical risk minimization. In *Proceedings of the 18th Annual Conference in Learning Theory (COLT)*, Bertinoro (Italy), June 2005.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In Smola, Bartlett, Schoelkopf, and Schuurmans, editors, *Advances in Large Margin Classifiers*, 2000.

S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

N. Ailon and M. Mohri. An efficient reduction of ranking to classification. In *COLT*, pages 87–98. Omnipress, 2008.

J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In *Proceedings of the European Conference on Machine Learning 2003, Cavtat-Dubrovnik*, pages 145–156. Springer-Verlag, 2003.

R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer. Learning preference relations for information retrieval. In *Proceedings of the Fifteenth Conference of the American Association for Artificial Intelligence*, pages 1–4, 1998.

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

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

M.W. Kattan, H. Ishida, P.T. Scardino, and J.R. Beck. *Applying a neural network to prostate cancer survival data*, chapter in Intelligent Data Analysis in Medicine and Pharmacology, pages 295–306. (N. Lavrac, R. Keravnou, B. Zupan, ed.), Kluwer, Boston, 1997.

D.M. Dabrowska and K.A. Doksum. Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics*, 15(1):1–23, 1988.

R. Koenker and O. Geling. Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96:458–468, 2001.

S.C. Cheng, L.J. Wei, and Z. Ying. Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association*, 92(437):227–235, 1997.

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

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

W. Chu and S.S. Keerthi. New approaches to support vector ordinal regression. In *In ICML 2005: Proceedings of the 22nd international conference on Machine Learning*, pages 145–152, 2005.

T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C.M. Perou, P.E. Lønning, P.O. Brown, A. Børresen-Dale, and D. Botstein. 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.

H.C. van Houwelingen, T. Bruinsma, A.A.M. Hart, L.J. van't Veer, and L.F.A. Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in Medicine*, 25(18): 3201–3216, 2006.

A. Rosenwald, G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, R.B. Smeland, J.M. Giltnane, E.M. Hurt, H. Zhao, L. Averett, and L. Yang. 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.

M. Schumacher, G. Basert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L.A. Neumann, and H.F. Rauschecker. 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.

V. Van Belle, K. Pelckmans, J.A.K. Suykens, and S. Van Huffel. Support vector machines for survival analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pages 1–8, Plymouth (UK), July 2007.

V. Van Belle, K. Pelckmans, J.A.K. Suykens, and S. Van Huffel. Survival SVM: a practical scalable algorithm. In *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN2008)*, pages 89–94, Bruges (Belgium), April 2008.

V. Van Belle, K. Pelckmans, J.A.K. Suykens, and S. Van Huffel. MINLIP: Efficient learning of transformation models. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN2009)*, pages 60–69, Limassol (Cyprus), September 2009.

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

Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(6):933–969, 2004.

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

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

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

D.J.C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computing*, 4(3):448–472, 1992.

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

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

R.C. Elandt-Johnson and N.L. Johnson. *Survival models and data anlysis*. John Wiley & Sons, Inc., 1980.

F. Harrell. *Regression Modeling Strategies. With applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

R.G. Miller. *Survival Anlysis*. John Wiley & Sons, 1981.

R. Peto. Discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society, Series B*, 34: 205–207, 1972.

G.A. Satten. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83:355–370, 1996.

F. Harrell, K. Klee, R. Califf, D. Pryor, and R. Rosati. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, 1984.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

K. Pelckmans, M. Espinoza, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Primal-dual monotone kernel regression. *Neural Processing Letters*, 22(2):171–182, 2005b.

H.M.M. Bøvelstad, S. Nygård, H.L.L. Størvold, M. Aldrin, O. Borgan, A. Frigessi, and O.C.C. Lingjærde. Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007.

D.R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of Statistical Learning.* Springer-Verlag, 2001.

E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2(4):511–522, April 2004.

R. Bair, T. Hastie, P. Debashis, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2006.

H. Martens and T. Næs. *Multivariate Calibration*. New York: John Wiley & Sons Inc., 1989.

S. Nygård, O. Borgan, O. Lingjærde, and H. Størvold. Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, 14(2):179–195, 2008.

M.J. van de Vijver, L.J. van't Veer, and H. Dai. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009, 2002.

P. Heagerty, T. Lumley, and M. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

P.H. Eilers and B.D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121, 1996.

C.M. Hurvich, J.S. Simonoff, and C.L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B*, 60:371–293, 1998.

T.M. Therneau and P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2 edition, 2000.

B. Fisher, M. Bauer, L. Wickerham, C.K. Redmong, and E.R. Fisher. 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.

P.J. Lamy, P. Pujol, S. Thezenas, A. Kramar, P. Rouanet, F. Guilleux, and J. Grenier. Progesterone receptor quantification as a strong prognostic determinant in postmenopausal breast cancer women under tamoxifen therapy. *Breast cancer reasearch and treatment*, 76(1):65–71, 2002.

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

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