Statistical Methods for the Evaluation of Diagnostic Biomarkers in the Presence of Censoring

Leonidas E. Bantis

Ph.D. Thesis

University of the Aegean
School of Sciences

Leonidas E. Bantis, Samos, Greece, 2013
Statistical Methods for the Evaluation of Diagnostic Biomarkers in the Presence of Censoring.
Ph.D. Thesis, University of the Aegean, School of Sciences,
Dept. of Statistics and Actuarial Financial Mathematics.

Λεωνίδας Ε. Μπαντής, Σάμος, Ελλάδα, 2013
Στατιστικές Μέθοδοι Αξιολόγησης Διαγνωστικών Ελέγχων Παρουσία Λογοκρισίας.
Διδακτορική διατριβή, Πανεπιστήμιο Αιγαίου, Σχολή Θετικών Επιστημών,
Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών.
(Ελληνική περίληψη παρατίθεται στο τέλος της διατριβής)

Statistical Methods for the Evaluation of Diagnostic Biomarkers in the Presence of Censoring

Leonidas E. Bantis

Ph.D. Thesis

University of the Aegean
School of Sciences, Dept. of Statistics and Actuarial-Financial Mathematics,
Samos Island, Greece, 2013

SUPERVISOR:

John V. Tsimikas
Associate Professor, University of the Aegean,
Head of Dept. of Statistics and Actuarial-Financial Mathematics.

COMMITTEE:

Antzoulakos D.L. (Participant of the 7 member committe)
Associate Professor, University of Piraeus,
Dept. of Statistics and Insurance Science.

Georgiou S.D. (Participant of the 7 member committe)
Associate Professor, University of the Aegean,
Dept. of Statistics and Actuarial-Financial Mathematics.

Karagrigoriou A. (Participant of the 7 member committe)
Associate Professor, University of Cyprus,
Dept. of Mathematics and Statistics.

Koutras M.V. (Participant of the 7 member committe)
Professor, University of Piraeus,
Dept. of Statistics and Insurance Science.

Nakas C.T. (Participant of the 3 and 7 member committe)
Assistant Professor, University of Thessaly, Laboratory of Biometry.

Nicoleris T. (Participant of the 3 and 7 member committe)
Assistant Professor, University of the Aegean,
Dept. of Statistics and Actuarial-Financial Mathematics.

Tsimikas J.V. (Supervisor, participant of the 3 and 7 member committee)
Associate Professor, University of the Aegean,
Head of Dept. of Statistics and Actuarial-Financial Mathematics.

This thesis contains original work published or submitted for publication to international journals. The following papers are included:

1. Tsimikas J.V., Bantis L.E. and Georgiou S.D. (2012). Inference in Generalized Linear Regression Models with a Censored Covariate. *Computational Statistics and Data Analysis*. **56(6)** 1854-1868.

2. Bantis L.E., Tsimikas J.V., and Georgiou S.D. (2012). Survival Estimation through the Cumulative Hazard Function with Monotone Natural Cubic Splines. *Lifetime Data Analysis*. **18(3)** 364-396.

3. Bantis L.E., Tsimikas J.V., and Georgiou S.D. (2013). Smooth ROC Curves and Surfaces for Markers Subject to a Limit of Detection Using Monotone Natural Cubic Splines. *Biometrical Journal*. Accepted for publication.

4. Bantis L.E., Georgiou S.D., and Tsimikas J.V. A MATLAB routine for Estimating Survival Functions with Monotone Natural Cubic Splines through the Cumulative Hazard. *Submitted for publication*.

During writing this thesis, the following conference presentations/talks have been made by the author:

- Paper 1: 6th EMR-International Biometric Society (IBS) conference, Crete (2011).
- Paper 1: Greek Statistical Institute, 23rd pan-hellenic Statistics conference, Veria (2010). (*published in the conference proceedings*)
- Paper 2: Greek Statistical Institute, 24th pan-hellenic Statistics conference, Patra (2011). (*published in the conference proceedings*)
- Paper 3: Greek Statistical Institute, 25th pan-hellenic Statistics conference, Volos (2012). (*published in the conference proceedings*)
- Paper 4: Greek Statistical Institute, 25th pan-hellenic Statistics conference, Volos (2012), (*with poster*)
- Additional talk: Accuracy of a binary time dependent diagnostic marker, Greek Statistical Institute, 21st pan-hellenic Statistics conference, Samos Island, (2008). (*published in conference proceedings*)

Talks regarding Paper 1 have also been made at the department of Statistics and Actuarial-Financial Mathematics, University of the Aegean. (Dates: May 20th, 2009 and October 21st, 2009)

ACKNOWLEDGMENTS:

First, I would like to thank my family for their encouragement and their support to all my decisions and during my studies. I would like to express my gratitude to my supervisor Dr. John V. Tsimikas for his guidance and patience. He gave me the freedom to focus on points that interested me most and was always supportive and encouraging. I consider myself fortunate for having the opportunity to work with him. I would also like to thank Dr. Stelios D. Georgiou for his help and contribution during this research. I also thank the seven member committee for reading this thesis and providing useful comments. I would also like to thank Georgia for her companionship, patience and for letting me share all my worries with her. Finally, for the endless discussions we had, my thanks also go to all friends I met during my stay in Samos.

Contents

Chapter 1

Introduction

1.1 Biomarkers

When the disease status of a patient/subject is of interest it is crucial that high quality and accurate data are obtained to assist the clinicians in the decision making process. These decisions are of great importance since they may affect the quality of life as well as its prolonging when fatal diseases are involved. Biomarkers play a significant role in decision making and are considered as the key to early diagnosis which may in turn lead to complete cure of a disease or the limitation of its progress. It is desired that the development of new biomarkers will contribute to obtaining more predictive information regarding the disease status of a patient as well as to providing a better understanding concerning the biological mechanism of various diseases. But what is considered as a biomarker (or medical/diagnostic test or simply marker)? The Biomarkers and Surrogate Endpoint Working Group provides the following definition (see Biomarker Definitions Working Group (2002)): "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention". This Working Group also defines a further three type classification for biomarkers. Type 0 biomarkers involve measurements of the history of a disease through time that is expected to be affected by known clinical indicators. Type I biomarkers refer to the effect of a clinical intervention (i.e. a cure, or a drug). Type II biomarkers refer to Surrogate Endpoints markers, namely markers that are expected to predict clinical deterioration or amelioration based on epidemiologic, therapeutic, pathophysiologic or other scientific evidence.

Another group at Bayer Corporation state the following definition for a biomarker (see Colburn (2003)): "It is a measurable property that reflects the mechanism of action of the molecule based on its pharmacology, pathophysiology of the disease, or an interaction between the two. A biomarker may or may not correlate perfectly with clinical efficacy/toxicity but could be used for internal decision making within a pharmaceutical company".

As mentioned in Naylor (2004) the definitions and classifications of biomarkers are still being discussed and debated. However, this field is developing and as the discovery of new biomarkers moves forward the use of reliable techniques for evaluating their accuracy is a crucial matter. Diagnostic biomarkers not only provide important information for the patients' health status, but also they contribute to reducing the cost of diagnosis

and may help us understand the underlying mechanism of a disease.

The evaluation of diagnostic markers is a necessary procedure before a medical test is used. Early detection of a disease may imply its complete cure or a slower progression given that an effective treatment exists. An ideal medical test should be accurate and harmless. Accurate, meaning that it is able to correctly distinguish the healthy from the diseased subjects, and harmless, meaning that it causes no physical pain, emotional discomfort, psychological stress or side-effects. A disadvantage of medical tests is the cost. The medical cost may also depend on the accuracy of the medical test itself. For example, when a healthy subject is falsely categorized as "positive", then this may lead to unnecessary further tests increasing the cost as well the subject's discomfort. An inaccurate test may have a more serious impact when a diseased subject is falsely categorized as healthy. Some criteria to be considered before applying a medical test in practice have been discussed in the literature and a detailed overview is provided in Pepe (2003). The main criteria are that the disease under study should be serious, treatable, and that the test should be harmless to the subject and accurate.

In this thesis we will introduce new methodologies that contribute to the evaluation of a marker regarding mainly its diagnostic accuracy in the presence of censoring. Censoring is a phenomenon primarily met in survival analysis in which patients are monitored over time until they experience an event. This event is usually death when fatal diseases are involved. In many cases, due to practical, psychological or other reasons many patients decide to leave/quit the study (for example due to their relocation to another country or due to the end of a study when some patients are event-free). The time at which these patients have or will experience the event is unknown to the clinical researchers and the only information available is that they were event-free until a given time. The time points at which these subjects left are considered as censored time points, and in particular right censored since the event will occur or has occurred at a time greater that the censoring time. Two other common censoring schemes that occur in survival analysis are left and interval censoring. In the first case the event occurred at some time point prior to the time of entering the study and interval censoring involves cases where the event time lies in a closed interval defined by two time points. It might be the case that a marker's measurements change with time and are used to predict the future disease status of a subject. An example of such a biomarker is the Framingham risk score (FR-score), which is considered to be predictive of myocardial infarction and stroke (see Wilson et al. (1998) and Grundy et al. (1998)). The FR-score is a score system (different for men and women) that is computed based on factors such as age, blood pressure, diabetes mellitus, blood cholesterol, and high density lipoprotein cholesterol. It is used to predict the risk of a cardiovascular event of an individual within the next 10 years. For such biomarkers it is expected that measurements taken closer to the event will be more indicative of the disease or its progression so it is natural to assume that biomarker measurements are related to the time to event variable. Hence, the phenomenon of censoring naturally arises in the concept of evaluating such a biomarker.

Censoring can also occur on the biomarker value itself. There are cases where due to practical reasons or technological limitations marker measurements cannot be provided below or beyond some known limit, usually called limit of detection (LOD) which may vary from batch to batch. In such cases the biomarker is itself subject to left or right censoring respectively. In this thesis we examine settings where censoring may occur on

the time to event as well as settings where the biomarker measurements themselves are subject to censoring.

In the following sections of the Introduction we give quantities that refer to the classification problem in medical decision making. Marker measurements may be binary, ordinal, or continuous. With ordinal or continuous measurements an ROC curve (for the two class case) or an ROC surface (for the three class case) can be constructed for assessing the accuracy of the biomarker. We also briefly discuss the framework where a biomarker is time dependent. We end the first chapter by discussing in more detail the basic features of the proposed techniques that are to be presented in this thesis.

1.2 Classification probabilities and the ROC

Diagnostic testing is an imperfect procedure. Assuming that there are two groups for classification (healthy and diseased), a perfect diagnostic marker would perfectly discriminate the diseased from the healthy group. On the other extreme, the classification procedure of a non-informative biomarker would be equivalent to deciding the health status of a subject by tossing a fair coin. In practice, the discriminatory capability of most biomarkers falls between these two extremes.

The diagnostic accuracy of a binary biomarker is completely summarized by its *sensitivity* and *specificity*. The sensitivity is defined as the probability of getting a positive marker result, given that the subject has the disease, while the specificity is the probability of getting a negative marker result given that the subject is healthy. We denote with Y the outcome of the diagnostic marker, and with D the binary indicator (which equals to 1 if the disease is present and 0 otherwise) of the true health status of a subject. The sensitivity and specificity are defined respectively as

$$Sensitivity = Se = P(Y = 1|D = 1),$$
$$Specificity = Sp = P(Y = 0|D = 0).$$

We desire high values of both sensitivity and specificity. The marker result can be one of the following: A *true positive* is said to occur when the marker outcome is positive for a subject that suffers from the disease, a *true negative* occurs when the marker outcome is negative for a healthy subject, a *false positive* occurs when the marker outcome is positive for a healthy subject, and finally a *false negative* occurs when the outcome of the marker is negative for a subject that has the disease. The false positive rate (FPR) and the true positive rate (TPR) are defined as:

$$FPR = P(Y = 1|D = 0),$$
$$TPR = P(Y = 1|D = 1) = Se.$$

Note that the TPR is the sensitivity, and that the *false negative rate* equals to $1-TPR$. Thus, the two components of the misclassification probability are $1 - TPR$ and FPR. The impact of the misclassification errors referring to these two probabilities is generally different. A false positive error is likely to lead to further examination of the health status of the subject. This unnecessary procedure may increase the medical cost or cause discomfort to the subject. The impact of a false negative error is usually more serious. Based on a false negative marker value people may forego a suitable treatment which may lead to death.

1.2.1 ROC curves

The most popular tool for evaluating the discriminatory capability of a continuous (or ordinal) biomarker is the receiver operating characteristic (ROC) curve. ROC curves were initially used in signal detection (Green and Swets (1966)). Lusted (1971) indicated their potential for medical diagnostic testing in which a decision must be made regarding the presence or absence of a disease. For a thorough overview regarding the ROC curves we refer the reader to Zhou et al. (2002) and Pepe (2003).

There are many examples of continuous markers. Cancer and liver disease biomarkers measure the serum concentration. Other examples are temperature, blood pressure, serum cholesterol etc. Ordinal markers usually involve the subjectiveness of a medical expert. For example a radiologist reading a mammography may classify the subject as 'definitely yes', 'probably yes', 'probably no', 'definitely no' regarding the presence of a disease. Similar classification rules may be also applied regarding the severity of the stage of progression of a disease. Also in psychology, it is common to base classification on frequency, that is use a scale such as 'always', 'sometimes', 'never' etc.

In most applications higher marker measurements, Y, are regarded to be more indicative of the presence of the disease. However, when this is not the case one can simply work with $-Y$. In practice a threshold c, is used to dichotomize Y so as to define a decision rule. Conventionally if $Y \geq c$ then the marker value is regarded as 'positive', and otherwise as 'negative'. The choice of c depends on the tradeoff of the cost of a false negative and a false positive result. The ROC curve is a tool that considers all possible threshold values depicting all these tradeoffs.

Using a threshold c, the sensitivity and false positive rate can be written as:

$$TPR(c) = P(Y \geq c|D = 1)$$
$$FPR(c) = P(Y \geq c|D = 0).$$

The ROC curve is defined through:

$$ROC(c) = (FPR(c), TPR(c)),\ c \in (-\infty, \infty).$$

Hence, the ROC curve consists of all possible pairs of the false positive and true positive rate constructed by the threshold c. Since TPR and FPR are probabilities, the ROC is a non-decreasing curve lying in the positive unit square:

$$\{(t, ROC(t)),\ t \in (0, 1)\}.$$

An ROC curve is invariant to strictly increasing transformations of Y and it holds that $ROC(0) = 0$ and $ROC(1) = 1$. The ROC curve referring to a perfect medical test, that is a test that perfectly discriminates the healthy from the diseased, starts from the point (1,1), moves on the upper left corner of the positive unit square (point (0,1)) and drops down to point (0,0). The ROC curve referring to a useless biomarker coincides with the unit square's diagonal with endpoints (1,1) and (0,0). In practice most biomarkers yield an ROC between those two extremes (see also Figure 1.1).

An ROC curve can be also represented by the survivor function of Y for the diseased and the healthy group. If we denote by $S_0(y) = P(Y \geq y|D = 0)$ and by $S_1(y) =$

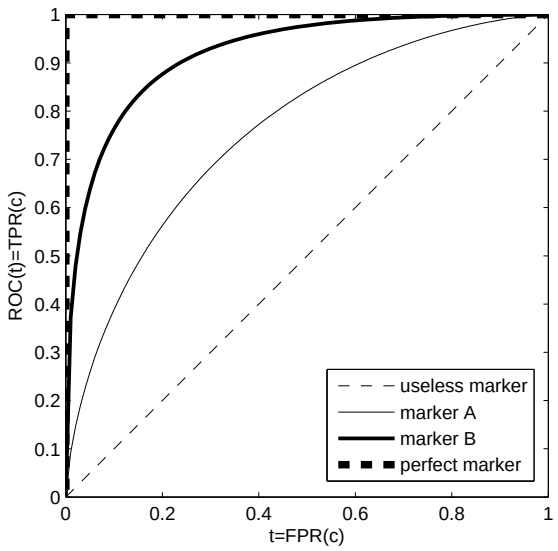

Figure 1.1: Example of four ROC curves referring to four hypothetical biomarkers. The diagonal line refers is an ROC curve that corresponds to a useless marker and the dashed thick line refers to an ROC curve of a perfect biomarker. The other two ROC curves indicate that marker B is better than marker A.

$P(Y \geq y|D = 1)$ the survivor functions corresponding to the heathy and the diseased respectively, it can be shown that (see result 4.2. in Pepe (2003)):

$$ROC(t) = S_1(S_0^{-1}(t)),\ t \in (0, 1). \tag{1.1}$$

Under the assumption of binormality, that is when both populations are assumed to be normally distributed, the ROC curve can be written in closed form. This is not always the case and the derivation of such closed form expressions depends on the complexity of the assumed parametric models for the underlying populations.

The two most popular approaches for estimating an ROC curve is the empirical (Lusted (1971) first adopted this for medical decision making) and the parametric one. Green and Swets (1966) first employed the Gaussian model to estimate an ROC curve. The empirical ROC curve involves estimating the TPR and FPR as follows:

$$\hat{TPR}(c) = \sum_{i=1}^{n_1} I(Y_{1i} \geq c)/n_1$$
$$\hat{FPR}(c) = \sum_{j=1}^{n_0} I(Y_{0j} \geq c)/n_0$$

where n_0 and n_1 are the number of healthy and diseased individuals, Y_{0i} and Y_{1j} are the marker values of the $i-$th and $j-$th subjects that belong to the healthy and

diseased group respectively. The empirical ROC curve is constructed by plotting $\hat{TPR}(c)$ versus $\hat{FPR}(c)$ for the whole spectrum of c. The parametric approach involves fitting a parametric model to each population and then simply plugging in the survival estimates in (1.1).

1.2.1.1 Area under the ROC curve

The most commonly used index of a biomarker's performance is the area under the ROC curve (AUC). Bamber (1975) was the first to focus on the AUC as a measure of accuracy of a biomarker. He first showed that the AUC equals to the U statistics of the Wilcoxon 2 sample nonparametric test. McClish (1989) employed parametric methods (binormal model) for the estimation of AUC and noted that it is a global measure of a biomarker's accuracy. The AUC is defined as:

$$AUC = \int_0^1 ROC(t)dt.$$

A biomarker that perfectly discriminates the two groups yields an AUC=1, while a useless marker yields an AUC=0.5. It can be shown that the AUC for a continuous marker equals to

$$AUC = P(Y_1 > Y_0).$$

This means that if we randomly select a pair of two individuals, one of each group, the probability that the marker will correctly classify them equals to the AUC. In the case of an ordinal marker where ties may be present the AUC can be shown to equal to (see Bamber (1975) and Hanley and McNeil (1982)):

$$AUC = P(Y_1 > Y_0) + 0.5P(Y_1 = Y_0).$$

Generally marker B is considered better than marker A if $AUC_B > AUC_A$ (see also Figure 1.1). However, in many cases one may be particularly interested in a specific range of FPR values. It might be the case that two ROC curves, that correspond to two different markers yield the same AUC without coinciding. One marker may be better than the other for a specific range of FPR values. In these cases a commonly used index is the partial area under the curve (pAUC) which is defined as

$$pAUC(t_1, t_2) = \int_{t_1}^{t_2} ROC(t)dt,\ 0 < t_1 < t_2 < 1.$$

The $pAUC$ can be interpreted as the average sensitivity for the considered range of specificities (see McClish (1989)).

For biomarkers subject to lower LODs, where measurements are undetected below some value replacement values are typically used in order to proceed to the calculation of the AUC. Nehls and Akland (1973) propose imputing the undetectable lower values with $d_L/2$, while the replacement value of $d_L/\sqrt{2}$ has also been proposed (see Hughes (2000) for an overview). However, these replacement values apply only to biomarkers that are allowed to yield non-negative scores and this is not always the case. Although most biomarkers yield positive scores, we often work with transformations that project the score values to the real line. Furthermore, Perkins et al. (2006) showed that even in the

case where the scores are non-negative any replacement value for the censored marker scores will cause bias in estimating the AUC. One can always proceed by maximum likelihood that takes into account the censored values after assuming parametric models for the marker measurements of the healthy and the diseased. However, parametric models make strict assumptions that may not always be justified by the available data. Non-parametric methods that would accommodate the censored nature of the measurements still remain unexplored. In this thesis we propose a spline based approach for the ROC estimation when the biomarker measurements are subject to a lower or an upper LOD.

1.2.1.2 ROC surfaces and the VUS

In a situation where a discrimination of three populations is needed an ROC surface is preferable and provides a natural generalization of the ROC curve in three dimensions (see Mossman (1999)). For example, mammogram readings are evaluated by radiologists who need to decide between between cancerous, benign growth or no nodules. In these situations, where interest focuses in discriminating simultaneously three populations an ROC surface is constructed, which in turn provides pairwise ROC curves for each pair of the disease status. For an overview see also Nakas and Yannoutsos (2004).

In the three class case where Y_3 tends to yield higher values than Y_2 which in turn tends to yield higher values than Y_1, there are three rates referring to a correct decision:

$$TPR_i = P(Y = i|D = i), i = 1, 2, 3$$

Since there are three populations for discrimination we consider that there are two ordered decision thresholds $c_1 < c_2$. Then the following decision rule applies: if $Y < c_1$ then classify as 'group 1', else if $c_1 < Y < c_2$ then classify as 'group 2', else classify as 'group 3'. As we vary the two thresholds to the support of all three underlying distributions, a three dimensional graph of an ROC surface can be constructed in the unit cube with axes TPR_1, TPR_2, and TPR_3. Hence the ROC surface is

$$ROC(c_1, c_2) = (TPR_1(c_1), TPR_2(c_1, c_2), TPR_3(c_2),\ c_1 < c_2).$$

If $Y_1 \sim F_1(\cdot)$, $Y_2 \sim F_2(\cdot)$, and $Y_3 \sim F_3(\cdot)$ the parametric representation of the corresponding ROC surface value is

$$ROC(TPR_1, TPR_3) = F_2(F_3^{-1}(1 - TPR_3)) - F_2(F_1^{-1}(TPR_1)). \tag{1.2}$$

As in the two class case, the two most popular ways to estimate an ROC surface is the parametric and the empirical one. Under a parametric assumption of each of the three underlying distributions one can construct the corresponding ROC surface based on (1.2). The empirical estimate of an ROC surface can be obtained in an analogous way as presented in the two class case. An appealing property of an ROC surface is that one can simultaneously assess the discriminatory capability of a marker referring to three groups without foregoing the pairwise analysis since the projections of the ROC surface on the sides of the unit cube are actually the pairwise ROC curves referring to each of the 3 couples of interest.

The summary index of interest when constructing an ROC surface is the volume under the surface (VUS) and is defined as

$$VUS = \int_0^1 \int_0^1 F_2(F_3^{-1}(1 - TPR_3)) - F_2(F_1^{-1}(TPR_1))dTPR_1 dTPR_3.$$

In the case of a continuous biomarker it can be shown (Dreiseitl et al. (2000)) that the VUS equals to:

$$VUS = P(Y_1 < Y_2 < Y_2),$$

and when ties are present

$$VUS = P(Y_1 < Y_2 < Y_3) + 0.5P(Y_1 = Y_2 < Y_3) + 0.5P(Y_1 < Y_2 = Y_3) + \frac{1}{6}P(Y_1 = Y_2 = Y_3).$$

If for two biomarkers, A and B, $VUS_B > VUS_A$, then biomarker B is considered to discriminate better between the three groups compared to marker A. A perfect marker would yield $VUS = 1$. The ROC surface referring to a useless marker would be a triangular surface with edges at (1,0,0), (0,1,0), (0,0,1) yielding $VUS = 1/6$ (see also Figure 1.2).

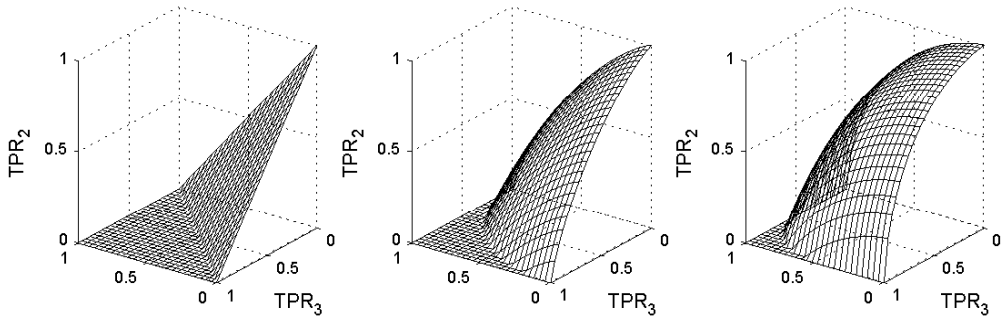

Figure 1.2: Example of three hypothetical ROC surfaces. Left: Y_1, Y_2, Y_3 follow the standard normal distribution yielding an ROC surface for a useless marker, Middle: $Y_1 \sim N(0,1), Y_2 \sim N(0.5,1), Y_3 \sim N(1,1)$, Right: $Y_1 \sim N(0,1), Y_2 \sim N(1,1), Y_3 \sim N(2,1)$

As in the two class case, more research is needed in developing non parametric estimates for the ROC surface when the measurements are subject to a lower or an upper limit of detection. This is an issue that will be explored in this thesis.

1.3 Time dependent biomarkers

1.3.1 Binary case

When two populations are under study, it may be the case that the diagnostic accuracy of a marker may depend on time and thus the need of defining time dependent sensitivity and specificity as well as the development of time dependent ROC curves arises. Recently, such methodologies were explored by Heagerty et al. (2000) as well as Cai et al. (2006) in the context of survival studies where the biomarker's outcome is supposed to

depend on the time to event. Interest also grows in modeling such a biomarker so as to understand how its results vary over time and derive valuable information regarding its (time dependent) accuracy.

In the case of a binary marker, one can model the sensitivity and FPR by using generalized linear models. The available data would be of the form $\{Y_i, T_i, \Delta_i, Z_i\}$, $i = 1, \ldots, n$, where Δ_i is the event indicator, that is $\Delta_i = I(X_i < C_i)$, where X_i is the time to event variable and $T_i = min(X_i, C_i)$. With C_i we denote the censoring variable. Z_i refers to other fully observed covariate(s) that may be available. The models assumed would be of the form

$$TPR = P(Y_i = 1|X = x, Z_i) = g_1(x, Z_i; \boldsymbol{\psi}_1), \quad 0 \leq t \leq \tau \tag{1.3}$$

$$FPR = P(Y_i = 1|X > \tau, Z_i) = g_0(Z_i; \boldsymbol{\psi}_0) \tag{1.4}$$

where g_0 and g_1 are known link functions that relate a linear predictor to the FPR and TPR respectively, $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ are the unknown parameter row vectors of the generalized linear models assumed for FPR and TPR. Note that FPR does not depend on time. If we denote all unknown parameters with the vector $\boldsymbol{\psi} = [\boldsymbol{\psi_0} \quad \boldsymbol{\psi_1}]$, then the corresponding likelihood would be

$$\prod_{i=1}^{n} p_i(\boldsymbol{\psi})^{Y_i}(1 - p_i(\boldsymbol{\psi}))^{1-Y_i} \tag{1.5}$$

where $p_i(\boldsymbol{\psi}) = P(Y_i = 1|T_i, \Delta_i, Z_i)$ which can be shown to be equal to (Cai et al. (2006)):

$$p_i(\boldsymbol{\psi}) = \begin{cases} TPR_{T_i,Z_i} & \text{if } T_i \leq \tau, \Delta_i = 1 \\ \frac{-\int_{T_i}^{\tau} TPR_{t,Z_i} dS_{Z_i}(x) + FPR_{\tau,Z_i} S_{Z_i}(\tau)}{S_{Z_i}(T_i)} & \text{if } T_i \leq \tau, \Delta_i = 0 \\ FPR_{Z_i,\tau} & \text{if } T_i > \tau, \end{cases}$$

where the first and third group can be considered as cases and controls respectively. Individuals of the second group have been censored before τ and hence have unknown case/control status. A consistent estimator of $\boldsymbol{\psi}$ can be obtained by using only data that fall into the first and third group (see also Leisenring et al. (1997)). The survival function $S_{Z_i}(\cdot)$ can be consistently estimated by assuming a proportional hazards model. In the case when all data are exactly observed and measurements are repeatedly taken for each subject over time, the so called ALR (alternating logistic regression) models have been proposed (see Carey et al. (1993)). Handling a binary response that also depends on a censored covariate when the data are longitudinal in nature is also a case that falls in the class of generalized linear models and in need of further exploration.

1.3.2 Continuous case

Under the same notion it may be reasonable to assume that marker values of diseased individuals may be a function of time, whereas healthy individuals exhibit marker values that are independent of time. Under such a setting there is a time lag between the time of the measurement and the occurrence of the event. One should take under consideration

the time lag since measurements made closer to the time of the event tend to be higher. Examples are the Framingham risk score (FR-score) and gene expression profiles of tumor tissue that are used to predict survival in cancer patients. Note that the time to event might be subject to right censoring, that is $T = min(X, C)$ where X is the time to event random variable and C is the censoring variable. The definition of sensitivity and FPR for a continuous marker must be extended in a way that time dependency as well as available covariates can be accommodated. One such extension is discussed in Cai et al. (2006):

$$TPR_{Z_i,t}(y) = P(Y \geq y|X = x, Z_i), \quad 0 \leq x \leq \tau \tag{1.6}$$

$$FPR_{Z_i,\tau}(y) = P(Y \geq y|X > \tau, Z_i) \tag{1.7}$$

where Y is the marker, Z_i is the covariate, and τ is some known 'distant' time point. Individuals who experience the event before τ are considered as diseased while individuals who survive beyond τ the control (healthy) group.

In a more general setting we assume that individuals are measured repeatedly at various times s_{ij}. The data would then be of the form $\{Y_{ij}, T_{ij}, \Delta_i, Z_i\}$, $i = 1, \ldots, n$, where Y_{ik} is the marker measurement at s_{ij} and $T_{ij} = T_i - s_{ij}$ is the time lag between the time of the event or censoring and the time of the measurement. When censoring is not involved the use of popular techniques such as the Generalized Estimating Equation method or the well known mixed models could be used (see Fitzmaurice et al. (2004) for an in depth overview of both techniques). However, methods that could accommodate both the longitudinal nature of the data as well as the presence of a censored covariate are not available in the class of generalized linear models. In this thesis we evaluate the sensitivity and FPR after modeling the marker process using generalized linear models. We explore both the simple setting when one measurement per subject is available as well as the more general longitudinal setting.

1.4 Thesis Overview

In this thesis we initially explore parameter estimation in the class of the generalized linear regression models when one covariate is censored. As previously mentioned, this is the typical situation that one may have to deal with when modeling a time dependent biomarker, with the time to event being the censored covariate. More specifically in Chapter 2 we propose a method for such a setting that is based on an estimating function approach. This method need not assume a parametric form for the distribution of the response given the regressors and is computationally simple. In the linear regression case the proposed approach implies the use of mean imputation of the censored regressor. We use flexible parametric models for the distribution of the covariate. When survival time is considered as the covariate subject to censoring, we use the generalized gamma distribution, since it is considered as a platform distribution covering a wide variety of hazard rate shapes. We further robustify our method by considering models of nonparametric nature typically used in survival analysis such as the logspline for the censored covariate. For models involving additional, fully observed, covariates we employ the generalized gamma accelerated failure time regression model. In this setting

no parametric family assumption for the extra covariates is needed. The proposed approach is broader than likelihood based multiple imputation techniques. Moreover, even in cases with a known parametric form for the response distribution, our method can be considered a feasible alternative to likelihood based estimation due to its computational simplicity which allows use of standard software. In cases where a parametric model is not justified by the data at hand we develop a spline based approach that can provide an on parametric alternative. This spline approach involves convex optimization and convergence is guaranteed unlike other likelihood based methods that assume complex parametric models. We conduct simulation studies for continuous, binary and count data to evaluate the performance of the proposed method and to compare the estimates to standard ones.

In Chapter 3 we consider the generalization of the previous approach to the longitudinal framework where the data are taken repeatedly over time. Our approach focuses on population based characteristics and is different from the joint modeling approach typically used in such situations (see Rizopoulos (2011)). The main advantage of our approach is that it does not require any assumptions regarding the parametric form of the distribution of the marker measurements or the censored covariate. We do not assume in any stage of our approach a model that incorporates random effects. We instead use a working correlation matrix to accommodate the within subjects' correlation of marker measurements. For the survival function of the censored time to event covariate we employ our monotone natural cubic spline model and the accommodation of other baseline covariates is done through semiparametric models.

In Chapter 4 we discuss and apply the above methodologies to the construction of time dependent ROC curves and the evaluation of the biomarkers through AUC over time. We consider the definition of 1.6 for the time dependent sensitivity and specificity since we explore fitting a broken-line model where a distant time point is used to separate the healthy from the diseased group. We also consider another definition for the time dependent sensitivity and specificity introduced by Heagerty et al. (2000)) for an application where longitudinal data are involved.

In Chapter 5 we study the construction of a smooth ROC curve (or surface in the case of three populations) when there is a lower or upper limit of detection. Again, we use spline models that incorporate monotonicity constraints for the cumulative hazard function of the marker distribution. The proposed technique is computationally stable and can accommodate other covariates.

In Chapter 6 we state some issues for further research and mention cases where our approaches may apply outside the field of biostatistics.

In the end of each chapter some technical notes regarding the proved results are provided when necessary. In the Appendix we provide some additional simulations for completeness of the results presented in the main body of the thesis. An algorithm for a new survival estimation approach, along with the corresponding package/software description built with `MATLAB 2011a` are also provided. An updated version of the software will soon be available by the author's current website:

`www.leobantis.net23.net`

or available upon request via e-mail:

lbantis@aegean.gr or leobantis@gmail.com.

All graphs, simulations and computational development of the proposed methods were done using MATLAB. Some of the competitive approaches were simulated in R. The SAS was also used for applying traditional methods during data analysis.

Chapter 2

Generalized Linear Models with a Censored Covariate

As discussed in the introduction time dependent ROC analysis involves modeling the marker values as a function of time, and generalized models may need to be employed. However, in such settings the time to event covariate may be subject to censoring. Parameter estimation and statistical inference in models where the response variable is subject to censoring has been thoroughly studied in the past with the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model being the most celebrated models. Less attention has been paid to the situation where the covariate is subject to censoring though. Gomez et al. (2003) consider a regression model with an interval-censored covariate and develop an algorithm for the nonparametric maximum likelihood estimation of the regression coefficients. In their setting no distributional form is assumed for the covariate. However, this is not the case for the response distribution. Pawitan and Self (1993) consider a repeated marker measurement setting and use Weibull regression models for infection and disease occurrence times that are subject to censoring. They also present arguments in favor of constructing models that consider modeling the disease marker process given the time variable (in their case, time to AIDS).

Another area where censored regressors are encountered frequently is econometrics where the covariate includes *unlimited* top and bottom categories (Rigobon and Stoker (2007)) or are categorized in groups (Hsiao (1983)). For instance observed household income would have a *top coded* response. In survival studies, which is our focus, the time to event variable may play the role of a covariate, and it is obvious that random or *bound* censoring may be present in such a setting. Note, that in survival studies bound censoring (or top coding) occurs due to the end of study (cutpoint). This type of censoring is also known as a ceiling effect in many scientific disciplines. Left censoring is somewhat less common in survival studies, but occurs frequently in econometrics and other scientific fields where it is typically referred to as bottom coding or the floor effect.

A related problem that has received attention in the past deals with surrogate predictors. Some strategies based on approximate quasi likelihood techniques, including Regression Calibration, are discussed in Carroll and Stefanski (1990). Prentice (1982) and Armstrong (1985) discuss the regression calibration technique under the frameworks of proportional hazards and generalized linear models respectively. Regression calibration in failure time regression models when some covariate values may be missing

or mismeasured is addressed in Wang et al. (1997). More recently Wang and Pepe (2000) discussed the use of Expected Estimating Equations in problems with measurement error in the covariate.

When dealing with a censored covariate the simplest approach is to discard the censored data and perform the analysis only with the observed data. This is called a Complete Case (CC) analysis. The CC method provides consistent parameter estimates under noninformative censoring. Obviously the CC method suffers from low efficiency which can be dramatic when heavy censoring is involved. In the case where both the distribution of the response given the covariates and the covariate distribution are assumed to lie within known parametric families, one can estimate the parameters via maximum likelihood. This is the approach taken by Austin and Hoch (2004) in the simple fully parametric setting where a ceiling effect is present on the covariates and where the joint distribution of the response and covariates is a multivariate normal distribution. However, when dealing with distributions other than the normal, computational issues arise. The method proposed in this chapter is both computationally simple and can be used without assumptions about the response distribution given the covariates. In many cases a parametric model, regarding the censored covariates, may be justified. Consider, for example, a situation where a new time dependent biomarker is to be evaluated. It is not uncommon to have historical data that allow us to use a parametric model for the distribution of the time to event covariate. Moreover, other fully observed covariates of interest may be utilized via an accelerated failure time (AFT) model. A flexible parametric model is the AFT generalized gamma regression model, which is considered as a platform for parametric analysis for survival data (see Cox et al. (2007)).

This chapter is organized as follows: In Section 2.1 we discuss the simple linear regression problem with the covariate being censored. In Section 2.2 we present our method, based on Estimating Equation theory that deals with the generalized linear model. We give details regarding the three most common settings, a continuous, a binary and a count response. We show how our method can be extended to accommodate other observed covariates, via the use of an AFT parametric model and in section 2.3 discuss the use of the Generalized Gamma distribution for the censored covariate. We also present some simulations for this case. We apply our method on real data from a randomized placebo controlled trial of the drug D-penicillamine (DPCA) for treatment of primary biliary cirrhosis (PBC) conducted at the Mayo Clinic (Fleming and Harrington (1990)). We continue with Section 2.4 where we explore a new spline approach (named as HCNS) that is used to model the censored covariate. Our spline approach is evaluated via simulations and compared to other non-parametric approaches. We discuss how this spline approach can be used in the context of a GLM with a censored covariate as well as in the case where additional (fully observed) covariates are present. Finally we present some additional simulations to evaluate the non-parametric spline technique.

2.1 Simple Linear Regression with a Censored Covariate

Consider the case of the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, ..., n \tag{2.1}$$

with the covariate subject to random right censoring, Type I censoring, or both. The data for the i-th subject consist of (Y_i, T_i, Δ_i), where Y_i is the response $T_i = min(X_i, C_i)$, C_i being the censoring variable and Δ_i is the indicator variable that informs us whether the i-th subject's covariate value is censored ($\Delta_i = 0$) or observed ($\Delta_i = 1$).

We assume that $E(\epsilon_i|X_i) = 0$. If we want to apply the Complete Case (CC) analysis, then we use only data with $\Delta_i = 1$. For the regression model to be well specified and for the CC to yield unbiased estimators, we assume that the mean of ϵ_i does not vary with Δ_i, that is $E(\epsilon_i|\Delta_i) = 0$. Given the Δ_i's, a parametric approach is to assume that both $f_{Y|X}(y_i|x_i)$ (the conditional distribution of Y_i given $X_i = x_i$) and $f_X(x_i)$ (the marginal distribution of X_i) lie within known parametric families. Given that the censoring and event times are independent we have

$$f_{Y|T,\Delta=1}(y_i|t_i, \delta_i = 1) \quad = \quad \frac{f_Y(y_i)}{f_X(t_i)S_C(t_i)} \int_{t_i}^{\infty} f_{X,C|Y}(t_i, c|y_i)dc,$$

where $S_C(t) = P(C > t)$, and y_i, t_i, δ_i are the realizations of the random variables Y_i, T_i, Δ_i respectively. Furthermore, if we assume conditional independence of censoring and event times given the response value $Y = y$ we obtain

$$f_{Y|T,\Delta=1}(y_i|t_i, \delta_i = 1) \quad = \quad \frac{f_{Y|X}(y_i|t_i)}{S_C(t_i)} \int_{t_i}^{\infty} f_{C|Y}(c|y_i)dc.$$

The contribution of an event ($\Delta_i = 1$) to the likelihood is

$$f_{Y,T,\Delta=1}(y_i, t_i, \delta_i = 1) \quad = \quad S_{C|Y}(t_i|y_i)f_X(t_i)f_{Y|X}(y_i|t_i).$$

Similarly, for observations with a censored covariate value we have

$$f_{Y,T,\Delta=0}(y_i, t_i, \delta_i = 0) \quad = \quad f_{C|Y}(t_i|y_i)f_{Y|X>t}(y_i|x_i > t_i)S_X(t_i).$$

For the case of the simple linear regression model in (2.1), let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\tau^2 = \text{var}(\epsilon_i)$ and define $\boldsymbol{\theta}$ to be the parameter vector of the distribution of the covariate. Assuming that the distribution of the censoring variable given the marker depends on a parameter vector, $\boldsymbol{\lambda}$, which is not a function of the parameters of interest, we maximize the likelihood $L(y_i, t_i, \delta_i; \boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ which is proportional to

$$\prod_{i=1}^{n} \left[\left\{ f_{Y|X}(y_i|t_i; \boldsymbol{\beta}, \tau) f_X(t_i; \boldsymbol{\theta}) \right\}^{\delta_i} \left\{ \int_{t_i}^{\infty} f_{Y|X}(y_i|x; \boldsymbol{\beta}, \tau) f_X(x; \boldsymbol{\theta}) dx \right\}^{1-\delta_i} \right]. \tag{2.2}$$

A perfectly reasonable assumption in most settings would be to simply assume joint independence of C, X and ϵ which also implies independence of C and Y. These assumptions result in the same likelihood.

Computational issues may arise when evaluating of the integral in (2.2) for subjects with a censored covariate ($\delta_i = 0$). The case where both X and $Y|X$ are assumed to follow normal distributions leads to a multivariate normal distribution for (X, Y), with the X variable subject to censoring. In this case the integral is computationally tractable and likelihood inference becomes feasible as investigated in Austin and Hoch (2004). However, when dealing with survival data we expect the covariate distribution to be positive supported. The case of a simple normal linear regression with a censored

exponentially distributed covariate is computationally relatively simple. In this case it can be shown that

$$\begin{aligned}
\int_t^\infty f_{Y|X}(y|x;\boldsymbol{\beta},\tau)f_X(x;\boldsymbol{\theta})dx &= \frac{\theta}{|\beta_1|}\frac{\phi\left(\frac{y-\beta_0}{\tau}\right)}{\phi\left(\frac{y-\beta_0-\tau^2\theta\beta_1^{-1}}{\tau}\right)}\bar{\Phi}\left(\frac{t\beta_1-(y-\beta_0-\tau^2\theta\beta_1^{-1})}{sign(\beta_1)\tau}\right)\\
f_{Y|X>t}(y) &= \frac{\theta e^{\theta t}}{|\beta_1|}\frac{\phi\left(\frac{y-\beta_0}{\tau}\right)}{\phi\left(\frac{y-\beta_0-\tau^2\theta\beta_1^{-1}}{\tau}\right)}\bar{\Phi}\left(\frac{t\beta_1-(y-\beta_0-\tau^2\theta\beta_1^{-1})}{sign(\beta_1)\tau}\right)\\
f_Y(y) &= \frac{\theta}{|\beta_1|}\frac{\phi\left(\frac{y-\beta_0}{\tau}\right)}{\phi\left(\frac{y-\beta_0-\tau^2\theta\beta_1^{-1}}{\tau}\right)}\Phi\left(\frac{y-\beta_0-\tau^2\theta\beta_1^{-1}}{sign(\beta_1)\tau}\right)\\
f_{X|Y}(x) &= \frac{|\beta_1|\tau^{-1}\phi\left(\frac{x-\beta_1^{-1}(y-\beta_0-\tau^2\theta\beta_1^{-1})}{\tau|\beta_1^{-1}|}\right)}{\Phi\left(\frac{y-\beta_0-\tau^2\theta\beta_1^{-1}}{sign(\beta_1)\tau}\right)}.
\end{aligned}$$

We note that in this setting $f(x|y)$ is simply a truncated Normal distribution, truncated at 0, i.e. $X|Y=y \sim TN_{[0,\infty)}(\beta_1^{-1}(y-\beta_0-\tau^2\theta\beta_1^{-1}),(\tau|\beta_1|^{-1})^2)$. In cases where the covariate distribution is more complex, such as the Weibull or the Generalized Gamma, it is harder to derive the likelihood function explicitly. Thus, maximization problems naturally arise. An alternative to the direct maximization of the likelihood is to apply multiple imputation to the censored covariate values based on the conditional distribution $f_{X|X>t,Y}(x|x > t,y)$. With the use of multiple imputation one avoids possible numerical difficulties associated with the maximization of the likelihood. However, use of multiple imputation in practice depends on the complexity of the form of $f_{X|X>t,Y}(x|x > t,y)$. In the setting of an exponentially distributed covariate and a simple normal linear regression model for $f_{Y|X}$, one can show that the distribution $f_{X|X>t,Y}(x|x > t,y)$ is a truncated normal, truncated at t, i.e. $TN_{[t,\infty)}(\beta_1^{-1}(y-\beta_0-\tau^2\theta\beta_1^{-1}),(\tau|\beta_1|^{-1})^2)$ and multiple imputation is straightforward. We note that both maximum likelihood and multiple imputation require the distribution of the covariate given the response.

Another approach would be to forego the normality assumption for the error term in (2.1) and perform a weighted least squares regression. In this case the regression model would be $E(Y_i|T_i,\Delta_i)=\beta_0+\beta_1E(X_i|T_i,\Delta_i)$. The expected value $E(X_i|T_i,\Delta_i)$ is equivalent to mean imputation of the censored values of the covariate. Note that the mean imputation is performed only by the information given from T and Δ. As weights one can select the inverse of $Var(Y_i|T_i,\Delta_i)$, i.e. $w_i=(\tau^2)^{-1}$ for data with the covariate being uncensored, and $w_i=(\beta_1^2Var(X|T,\Delta)+\tau^2)^{-1}$ for the ones with censoring present. The estimator produced is $\hat{\boldsymbol{\beta}}=(\mathbf{X'WX})^{-1}\mathbf{X'Wy}$, with $\mathbf{W}=\text{diag}(w_1,\ldots,w_n)$ and $\mathbf{X}$ the design matrix with ones in the first column and the $i-$th element of the second column equal to $E(X_i|T_i,\Delta_i)$. This estimator is dependent on the slope β_1 and hence an iterative procedure is needed. Moreover, an estimator of the first two moments $E(X_i|T_i,\Delta_i)$ and $E(X_i^2|T_i,\Delta_i)$ is required which leads to the use of a parametric model on the covariate. In the next section we propose a method for parameter estimation in generalized linear models (GLMs). The above estimator turns out to be a limiting case of the proposed method.

2.2 Parameter Estimation when a Single Covariate is Censored

We are interested in estimating the parameters of a GLM when the covariate suffers from censoring. We consider the generalized linear model specified by

$$\begin{aligned} E(Y_i|X_i = x_i, \mathbf{z}_i) &= \mu_i \\ g(\mu_i) &= [\mathbf{x}_i^{'}, \mathbf{z}_i^{'}]'[\beta_0, \beta_1, \boldsymbol{\beta}_2^{'}]^{'} \\ Var(Y_i|X_i = x_i, \mathbf{z}_i) &= \tau^2 v(\mu_i);\ i = 1, \ldots, n, \end{aligned} \tag{2.3}$$

where $g(\cdot)$ is the link function, β_0 is the intercept, β_1 is the coefficient of the censored covariate, $\boldsymbol{\beta}_2$ is the vector of coefficients corresponding to the fully observed covariates, the vector $\mathbf{x}_i = [1, x_i]'$ refers to the covariate values that may be censored, the vector $\mathbf{z}_i^{'} = [z_{i1}, \ldots, z_{i,p-1}]$ refers to the fully covariates values, $v(\mu_i)$ is the variance function and τ^2 is the dispersion parameter (McCullogh and Searle (2001)). The model (2.3) in vector form can be written as

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu},\ \ \mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},\ \ \mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \tau^2\mathbf{V}(\boldsymbol{\mu}), \tag{2.4}$$

where $\mathbf{Y} = [y_1, \ldots, y_n]'$, $\mathbf{X} = [(\mathbf{x}_1^{'}, \mathbf{z}_1^{'}), \ldots, (\mathbf{x}_n^{'}, \mathbf{z}_n^{'})]'$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \boldsymbol{\beta}_2^{'}]^{'}$, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]'$, $\mathbf{g}(\boldsymbol{\mu}) = [g(\mu_1), \ldots, g(\mu_n)]'$ and $\mathbf{V}(\boldsymbol{\mu}) = \mathrm{diag}(v(\mu_1), ..., v(\mu_n))$. The data is

$$\{\mathbf{Y}, \mathbf{T}, \boldsymbol{\Delta}, \mathbf{Z}\} = \begin{pmatrix} y_1 & t_1 & \delta_1 & z_{1,1} & \ldots & z_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ y_n & t_n & \delta_n & z_{n,1} & \ldots & z_{n,p-1} \end{pmatrix}, \tag{2.5}$$

where $t_i = min(x_i, c_i)$, c_i is the censoring time for the $i-$th subject and $\delta_i = I_{(x_i<c_i)}$, where $I_{(A)}$ denoted the indicator function of the event A.

2.2.1 Optimal estimating functions in the case of a single covariate

In order to develop our method we recall the notation and the general theory regarding optimal estimating functions and O_F-optimality, as introduced by Heyde (1997). Let $\{\mathbf{D}_n, n \leq N\}$ be a sample of discrete or continuous data. Let also the class $\mathcal{G}$ of zero mean, square integrable estimating functions $\mathbf{G}_N = \mathbf{G}_N(\{\mathbf{D}_n, n \leq N\}, \boldsymbol{\beta})$ of dimension p with $E\mathbf{G}_N(\boldsymbol{\beta}) = 0$ and for which the p-dimensional matrices $E\dot{\mathbf{G}}_N$ and $E\mathbf{G}_N\mathbf{G}_N'$ are nonsingular. The dot denotes the derivative with respect to $\boldsymbol{\beta}$, that is $E\dot{\mathbf{G}}_N = EdG_{N,i}(\boldsymbol{\beta})/d\beta_j$. If we consider $\mathcal{H} \subseteq \mathcal{G}$, then optimality within $\mathcal{H}$ is acquired if the covariance matrix of the standardized estimating functions $\mathbf{G}_N^{(s)} = -(E\dot{\mathbf{G}}_N)'(E\mathbf{G}_N\mathbf{G}_N')\mathbf{G}_N$ is maximized. Equivalently, if the score function $(\mathbf{U}_N)$ exists, an optimal estimating function within $\mathcal{H}$ is one with minimum dispersion distance from $\mathbf{U}_N$, or alternatively, one with maximum correlation with the generally unknown score function. O_F-optimality is achieved by choosing the estimating function that maximizes the information criterion

$$E(\mathbf{G}_N^{(s)}\mathbf{G}_N^{(s)'}) = (E\dot{\mathbf{G}}_N)'(E\mathbf{G}_N\mathbf{G}_N')^{-1}(E\dot{\mathbf{G}}_N) \tag{2.6}$$

which is a generalization of the Fisher information. It can be shown that $\mathbf{G}_N^* \in \mathcal{H}$ is an O_F-optimal estimating function within $\mathcal{H}$ if $(E\dot{\mathbf{G}}_N)^{-1}E(\mathbf{G}_N\mathbf{G}_N^{*'})$ is a constant matrix for all $\mathbf{G}_N \in \mathcal{H}$.

In the following theorem we present a plausible family of estimating functions that can be considered when a censored covariate is present in a generalized linear model and derive the O_F-optimal one. We assume a parametric model for the covariate, X, such that $V(Y) = E(Var(Y|X)) + Var(E(Y|X)) < \infty$.

Theorem 1. *Consider the generalized linear model in (2.4). Assume our data, $\{\mathbf{Y}, \mathbf{T}, \mathbf{\Delta}\}$, are the first three columns as in (2.5) with no fully observed covariates. Let the class $\mathcal{H}$ of square integrable estimating functions $\mathbf{G} = \mathbf{G}_N(\boldsymbol{\beta})$:*

$$\mathcal{H} : \{\mathbf{G} = \mathbf{A}(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}^c(\boldsymbol{\beta})) \ , \ \mathbf{A}(\boldsymbol{\beta}) \ is \ \mathbf{T}, \mathbf{\Delta} \ measurable\} . \tag{2.7}$$

where the $i-th$ component of $\boldsymbol{\mu}^c$ is $\mu_i^c = E_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}$. Assume that $E\left\{\mathbf{A}(\boldsymbol{\beta})\dot{\boldsymbol{\mu}}^c\right\}$ and $E\left\{\mathbf{A}(\boldsymbol{\beta})\mathbf{W}^{-1}\mathbf{A}'(\boldsymbol{\beta})\right\}$ are nonsingular, where $\mathbf{W}^{-1} = diag\{Var(Y_i|T_i, \Delta_i)\}$. If the unconditional variance of the response is finite then the O_F optimal estimating function $(\mathbf{G}^*)$ for estimating $\boldsymbol{\beta}$ within $\mathcal{H}$ is given by

$$\sum_{i=1}^{n}\left[\frac{y_i - E_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}}{\tau^2 E_{X_i|T_i,\Delta_i}\left\{v(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))\right\} + Var_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}} E_{X_i|T_i,\Delta_i}\left\{\frac{d(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\right\}\right] \tag{2.8}$$

Proof. The proof is given in the technical notes section of this chapter. □

When no censoring of the covariate occurs then (2.8) reduces to the well known estimating function used in generalized linear models presented in McCullogh and Searle (2001). Other computationally simpler alternatives to the O_F-optimal estimating function may exist in the family (2.7). For example, here we propose the use of an unweighted method, defined by the estimating function $G^{Un} = \dot{\boldsymbol{\mu}}^{\boldsymbol{c}'}(\mathbf{Y} - \boldsymbol{\mu}^{\boldsymbol{c}})$. Thus the unweighted estimator $\hat{\boldsymbol{\beta}}^{Un}$ is the solution of

$$\sum_{i=1}^{n}\left[\left\{y_i - E_{X_i|T_i,\Delta_i}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))\right\}\left\{E_{X_i|T_i,\Delta_i}\left(\frac{d(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\right)\right\}\right] = \mathbf{0}. \tag{2.9}$$

Of course the computationally simplest method is to perform Complete Case analysis which is a limiting case of the family we considered, obtained by setting the weights of observations with a censored covariate to zero.

Under some natural conditions the asymptotic behavior of $\hat{\boldsymbol{\beta}}$, in the family defined by (2.7), is derived using the results of Yuan and Jennrich (1998). They present three general conditions for the existence, consistency and asymptotic normality of estimating function estimators. 1. $\frac{1}{n}\mathbf{G}^* \to 0$ with probability one, 2. There is a neighborhood of $\boldsymbol{\beta}$ on which with probability one all $\mathbf{G}^*$ are continuously differentiable and $\frac{1}{n}\dot{\mathbf{G}}^*$ converges uniformly to a nonstochastic limit which is nonsingular at $\boldsymbol{\beta}$, 3. $\frac{1}{\sqrt{n}}\mathbf{G}^* \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$. Assumptions 1 and 2 are for the existence and consistency of the estimator and assumption 3 for its asymptotic normality.

In our case the proposed estimating function can be written as

$$\mathbf{G}^* = \sum_{i=1}^{n}\left\{w_i(y_i - \mu_i^c)\left(\frac{d\mu_i^c}{d\beta_0}, \frac{d\mu_i^c}{d\beta_1}, \ldots, \frac{d\mu_i^c}{d\beta_{p-1}}\right)'\right\} = \sum_{i=1}^{n}\left\{w_i(y_i - \mu_i^c)\dot{\boldsymbol{\mu}}_i^{c\prime}\right\} = \sum_{i=1}^{n}\mathbf{q}_i.$$

The expected value of the derivative of the estimating function is $E(\dot{\mathbf{G}}^*) = -E(\dot{\boldsymbol{\mu}}^{c\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^c)$ with (s,t)-th element:

$$E_{T_i,\Delta_i}\left(\sum_{i=1}^{n} w_i \frac{d\mu_i^c}{d\beta_s}\frac{d\mu_i^c}{d\beta_t}\right) = E_{T_i,\Delta_i}\left[\sum_{i=1}^{n} \frac{E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}(\mathbf{x}_i'\boldsymbol{\beta})}{d\beta_s}\right\} E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}(\mathbf{x}_i'\boldsymbol{\beta})}{d\beta_t}\right\}}{\tau^2 E_{X_i|T_i,\Delta}\left\{v(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))\right\} + Var_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}}\right].$$

The derivative is $\dot{\mathbf{G}}^* = \sum_{i=1}^n \dot{\mathbf{q}}_i$ where the (s,t)-th element of the $\dot{\mathbf{q}}_i$ equals to

$$\frac{d}{d\beta_s}\left(w_i(y_i - \mu_i^c)\frac{d\mu_i^c}{d\beta_t}\right).$$

Since the estimating equations are unbiased, the Strong Law of Large Numbers (SLLN) implies that $n^{-1}\mathbf{G}^* \overset{a.s.}{\to} \mathbf{0}$ and the first assumption is satisfied. The third assumption follows from the central limit theorem $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{q}_i \to N(\mathbf{0}, Var(\mathbf{q}_i))$, where $Var(\mathbf{q}_i) = E(\dot{\boldsymbol{\mu}}_i^{c\prime} w_i \dot{\boldsymbol{\mu}}_i^c)$. For the second assumption it can be shown that $\frac{1}{n}\dot{\mathbf{G}}^* = \frac{1}{n}\sum_{i=1}^n \dot{\mathbf{q}}_i$ converges to $-E(\dot{\boldsymbol{\mu}}_i^{c\prime} w_i \dot{\boldsymbol{\mu}}_i^c)$. If we further assume that the convergence is uniform, then the estimate $\hat{\boldsymbol{\beta}}$ exists, is consistent and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance $\left\{E(\dot{\boldsymbol{\mu}}^{c\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^c)\right\}^{-1}$. An estimate of the covariance matrix is given by $(\dot{\boldsymbol{\mu}}^{c\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^c)^{-1}$. The Fisher scoring iterative algorithm may be applied for solving the estimating equations, where for the $m+1$ iteration we have $\hat{\boldsymbol{\beta}}_{(m+1)} = \hat{\boldsymbol{\beta}}_{(m)} + \left(\dot{\boldsymbol{\mu}}^{c\prime}_{(m)}\mathbf{W}_{(m)}\dot{\boldsymbol{\mu}}^c{}_{(m)}\right)^{-1}\mathbf{A}^*_{(m)}(\mathbf{Y} - \boldsymbol{\mu}^c_{(m)})$.

Similarly and under the same assumptions, the estimator $\hat{\boldsymbol{\beta}}^{Un}$ of the Unweighted estimating function exists, is consistent and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance $\left\{E(\dot{\boldsymbol{\mu}}^{c\prime}\dot{\boldsymbol{\mu}}^c)\right\}^{-1} E(\dot{\boldsymbol{\mu}}^{c\prime}\mathbf{W}^{-1}\dot{\boldsymbol{\mu}}^c)\left\{E(\dot{\boldsymbol{\mu}}^{c\prime}\dot{\boldsymbol{\mu}}^c)\right\}^{-1}$ where the (s,t)-th element of the matrix $E(\dot{\boldsymbol{\mu}}^{c\prime}\dot{\boldsymbol{\mu}}^c)$ is $\sum_{i=1}^n\left[E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}(\mathbf{x}_i'\boldsymbol{\beta})}{d\beta_s}\right\} E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}(\mathbf{x}_i'\boldsymbol{\beta})}{d\beta_t}\right\}\right]$.

The previous discussion assumes that the dispersion parameter, τ, and the parameters of the distribution of the censored covariate, X, are known. In practice we estimate the dispersion parameter using the CC analysis by the usual moment estimator (McCullagh and Nelder, 1983). The parameters of the distribution of X are estimated by maximizing the likelihood based solely on the observations of the censored covariate which essentially defines additional estimating equations. Hence, strictly speaking the asymptotic variance has to be adjusted for estimating these unknown parameters. Given the high degree of complexity and difficulty in calculating this asymptotic variance we propose inference based on the bootstrap. We evaluated the performance of confidence intervals based on the unadjusted asymptotic variance and the bootstrap via simulation which we present in Section 2.3.3.

We note here that our method is applicable even when we can assume a parametric model for the conditional distribution of $Y|X$. As discussed in Section 2.1 its main advantage in this case is that it can be considerably easier to compute our estimates compared to the direct maximum likelihood approach or the methods of multiple imputation. In the parametric setting, knowledge of the marginal distribution of X and the conditional $Y|X$ (and hence the joint distribution of (X,Y)) allows the computation of the conditional distribution of $X|Y$. As for example in survival studies, with X being a survival time and Y a biomarker value, it is the predictive distribution, $X|Y$, that may

ultimately be of interest. We can view $Y|X$ as the biological or diagnostic model and X as the marginal lifetime distribution. A parametric model of $Y|X$ is not needed if one is simply interested in assessing the diagnostic accuracy of a time dependent biomarker (see Cai et al. (2006) and Heagerty et al. (2000)). However, given the parametric forms of the distributions of X and $Y|X$, our methods allow a computationally convenient way of obtaining estimates of the parameters in the predictive model $X|Y$ in prospective survival studies. The complete treatment of predictive inference based on this approach is beyond the scope of this study.

2.2.2 Examples

2.2.2.1 Continuous data with identity link function

In this case we have

$$\mu_i^c(\boldsymbol{\beta}) = E(Y_i|T_i, \Delta_i) = \begin{cases} \beta_0 + \beta_1 t_i, & \text{if } \Delta_i = 1 \\ \beta_0 + \beta_1 E(X_i|X_i > t_i), & \text{if } \Delta_i = 0 \end{cases}$$

and

$$w_i^{-1} = Var(Y_i|T_i, \Delta_i) = \begin{cases} \tau^2 & \text{if } \Delta_i = 1 \\ \beta_1^2 Var(X_i|X_i > t_i) + \tau^2, & \text{if } \Delta_i = 0. \end{cases}$$

The matrix $\dot{\boldsymbol{\mu}}^c$ will be an n by 2 matrix with its first column equal to $\mathbf{1}_n$, and the $i-$th element of the second column equal to t_i when the covariate is observed and $E(X_i|X_i > t_i)$ when censoring occurs. This simply amounts to mean imputation of the censored covariate values followed by computation of the appropriate weights. An estimate of the variance τ^2 can be derived from the CC analysis. In the linear case our method is applicable provided that $E(X_i^2) < \infty$.

2.2.2.2 Binary data with the logit link function

In this case

$$\mu_i^c(\boldsymbol{\beta}) = E(Y_i|T_i, \Delta_i) = \begin{cases} \frac{exp(\beta_0+\beta_1 t_i)}{1+exp(\beta_0+\beta_1 t_i)}, & \text{if } \Delta_i = 1 \\ E\left(\frac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)}|X_i > t_i\right), & \text{if } \Delta_i = 0 \end{cases}$$

and it can be shown that

$$w_i^{-1} = Var(Y_i|T_i, \Delta_i) = \mu_i^c(\boldsymbol{\beta})(1 - \mu_i^c(\boldsymbol{\beta})).$$

The matrix $\dot{\boldsymbol{\mu}}^c$ is an n by 2 matrix with the $i-$th element of the first column equal to $\frac{exp(\beta_0+\beta_1 t_i)}{\{1+exp(\beta_0+\beta_1 t_i)\}^2}$ when the covariate is observed and $E\left(\frac{exp(\beta_0+\beta_1 X_i)}{\{1+exp(\beta_0+\beta_1 X_i)\}^2}|X_i > t_i\right)$ when the covariate is censored. The $i-$th element of the second column equals to $\frac{t_i exp(\beta_0+\beta_1 t_i)}{\{1+exp(\beta_0+\beta_1 t_i)\}^2}$ when the covariate is observed and $E\left(\frac{X_i exp(\beta_0+\beta_1 X_i)}{\{1+exp(\beta_0+\beta_1 X_i)\}^2}|X_i > t_i\right)$ when the covariate is censored. Since $Var(Y) < \infty$, our method is always applicable. Of course one should be aware of the well known problems that occur in logistic regression such as total separation or the 'only successes' or 'only failures' scenarios.

2.2.2.3 Count data with the log link function

We have

$$\mu_i^c = E(Y_i|T_i,\Delta_i) = \begin{cases} exp(\beta_0+\beta_1 t_i), & \text{if } \Delta_i = 1 \\ E\left\{exp(\beta_0+\beta_1 X_i)|X_i > t_i\right\}, & \text{if } \Delta_i = 0. \end{cases}$$

Assuming that the mean to variance relationship is $Var(Y_i|T_i,\Delta_i = 1) = \tau^2\mu_i(\boldsymbol{\beta}) = \tau^2 E(Y_i|X_i = x_i)$, as in the Wedderburn (1974) quasi likelihood setting, we get

$$w_i^{-1} = Var(Y_i|T_i,\Delta_i) = \begin{cases} \tau^2\mu_i(\boldsymbol{\beta}), & \text{if } \Delta_i = 1 \\ \tau^2 E\left\{\mu_i(\boldsymbol{\beta})|X_i > t_i\right\} + Var\left\{\mu_i(\boldsymbol{\beta})|X_i > t_i\right\}, & \text{if } \Delta_i = 0. \end{cases}$$

In this case the $i-$th row of $\dot{\boldsymbol{\mu}}^c$ equals to $[exp(\beta_0+\beta_1 t_i), t_i exp(\beta_0+\beta_1 t_i)]$ when the covariate is observed and $[E\left\{exp(\beta_0+\beta_1 X_i)|X_i > t_i\right\}, E\left\{X_i exp(\beta_0+\beta_1 X_i)|X_i > t_i\right\}]$ when censoring occurs. Our method is applicable when $Var\left\{exp(\beta_0+\beta_1 X_i)\right\} < \infty$ or equivalently $E\left\{exp(2\beta_1 x)\right\} = M_X(2\beta_1) < \infty$, where $M_X(.)$ is the moment generating function (m.g.f) of the distribution of X. For example, if X follows the exponential distribution with mean θ, then we require that $\beta_1 < (2\theta)^{-1}$.

2.2.3 Accommodating other observed covariates

The method can be extended to accommodate additional fully observed covariates. We propose the use of accelerated failure time (AFT) models to account for information that the additional covariates may carry about the censored regressor.

For the $i-$th subject, denote the additional covariate values by $\mathbf{z}_i^{'} = [z_{i1},\ldots,z_{i,p-1}]$. Assume a generalized linear model that relates Y_i to x_i and $\mathbf{z}_i$ and is of the form (2.4). Assume further the AFT model with unknown vector of coefficients $\boldsymbol{\xi}$

$$log(X_i|z_{i1},...,z_{i,p-1}) = [1,\mathbf{z}_i^{'}]\boldsymbol{\xi} + \sigma u_i \tag{2.10}$$

where u_i follows some distribution that depends on a parameter vector $\boldsymbol{\kappa}$. For simplicity, consider the case where in addition to the censored regressor, one covariate z is fully observed ($p = 2$). We define $\boldsymbol{\zeta} = [\xi_0,\xi_1,\boldsymbol{\kappa}]$ and denote as $\boldsymbol{\eta}$ the parameter vector of the distribution of the covariate z. Then, under similar assumptions that we made for (2.2), the likelihood $\prod_{i=1}^n L(y_i,t_i,\delta_i,\boldsymbol{z}_i;\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\eta},\tau^2)$, is proportional to

$$\prod_{i=1}^n \left[\left\{f_{Y|X,Z}(y_i|t_i,z_i;\boldsymbol{\beta},\boldsymbol{\zeta},\tau^2) f_{X|Z}(t_i|z_i;\boldsymbol{\zeta})\right\}^{\delta_i}\left\{f_{Y|X>t,Z}(y_i|x_i>t_i,z_i;\boldsymbol{\beta},\boldsymbol{\zeta},\tau^2) f_{X>t|Z}(x_i|z_i;\boldsymbol{\zeta})\right\}^{1-\delta_i}\right].$$

The maximization of the above likelihood is a difficult task due to the presence of censoring. Here, we assume that the additional, uncensored, covariates are fixed at their observed values. In this setting and for the case of one additional fully observed covariate the O_F-optimal estimating function will be of the form $\mathbf{G}^* = \dot{\boldsymbol{\mu}}^{c'}\mathbf{W}(\mathbf{Y}-\boldsymbol{\mu}^{\boldsymbol{c}})$ where the $i-$th element of $\boldsymbol{\mu}^c$ is $\boldsymbol{\mu}_i^c = E_{X_i|T_i,\Delta_i}\left\{g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})\right\}$ and $\mathbf{W}^{-1} = diag\left\{Var(Y_i|T_i,\Delta_i,Z_i)\right\}$. Thus, the optimal estimating function can be written as

$$\mathbf{G}^* = \sum_{i=1}^n \left[\frac{\left\{y_i - E_{X_i|T_i,\Delta_i,Z_i}(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))\right\}\left\{E_{X_i|T_i,\Delta_i,Z_i}\left(\frac{d(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\right)\right\}}{\tau^2 E_{X_i|T_i,\Delta_i,Z_i}\left\{v(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))\right\} + Var_{X_i|T_i,\Delta_i,Z_i}\left\{g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})\right\}}\right]. \tag{2.11}$$

An estimate of the asymptotic variance matrix is provided by $(\dot{\boldsymbol{\mu}}^{\boldsymbol{c}\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^{\boldsymbol{c}})^{-1}$ where the s,t-th element of $(\dot{\boldsymbol{\mu}}^{\boldsymbol{c}\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^{\boldsymbol{c}})$ is $\sum_{i=1}^{n}\left[\frac{E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})}{d\beta_t}\right\}E_{X_i|T_i,\Delta_i}\left\{\frac{dg^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})}{d\beta_s}\right\}}{\tau^2 E_{X_i|T_i,\Delta}\left\{v(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))\right\}+Var_{X_i|T_i,\Delta_i}\left\{g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})\right\}}\right]$.

The extension of our method to accommodate additional fully observed covariates, that will be considered fixed ($p > 2$), is straightforward. Also, one can consider only a subset of the additional covariates in the AFT model in equation (2.10). As in the case of a single covariate, τ is estimated via the CC analysis, and the parameter vector $\boldsymbol{\xi}$ and σ via maximum likelihood.

2.3 Parametric model for the censored covariate.

2.3.1 Case I: A single covariate

In many instances, historical data from previous studies suggest the use of a specific parametric model for the censored covariate. In the absence of such information, we propose the use of a flexible parametric model. Here, we explore the use of the Generalized Gamma distribution which is considered as a platform for parametric analysis for survival data. The generalized gamma is a parametric family which was initially introduced by Stacy (1966). It includes most of the commonly used distributions in survival analysis (e.g. exponential, Weibull, gamma, log-normal,...) either as special or limiting cases. Its hazard function can be monotonically increasing, decreasing, arc-shaped or U-shaped (bathtub). For a detailed study regarding the generalized gamma family see Cox et al. (2007). Its density can be written as (see Lee and Wang (2003))

$$f(x;\alpha,\lambda,\gamma)=\frac{|\alpha|}{\Gamma(\gamma)}\gamma^{\gamma}\lambda^{\alpha\gamma}x^{\alpha\gamma-1}exp\left\{-\gamma(\lambda x)^{\alpha}\right\},$$

where $\alpha \neq 0$ and $\gamma > 0$ are the shape parameters, and $\lambda > 0$ is the scale parameter. When $\alpha = 0$ the limiting case of the lognormal distribution is obtained. When $\gamma = 1$ or $\alpha = 1$ we obtain the Weibull or the Gamma distribution respectively. Here, we denote the Generalized Gamma distribution by GG(α,λ,γ). The survival function is

$$S(x;\alpha,\lambda,\gamma)=\begin{cases} I\left\{\gamma(\lambda x)^{\alpha},\gamma\right\}, & \text{if } \alpha<0 \\ 1-I\left\{\gamma(\lambda x)^{\alpha},\gamma\right\}, & \text{if } \alpha>0. \end{cases}$$

where $I\{\cdot,\cdot\}$ is the incomplete gamma function. The r-th moment is given by:

$$E(X^r)=\begin{cases} (\lambda\gamma^{\frac{1}{\alpha}})^{-r}\frac{\Gamma\left\{(\alpha\gamma+r)\alpha^{-1}\right\}}{\Gamma(\gamma)}, & \text{if } \frac{r}{\alpha}>-\gamma \\ \infty, & otherwise. \end{cases}$$

The hazard rate of the GG(α,λ,γ) distribution takes the form of the four most common types, i.e. increasing, decreasing, arc-shaped and bathtub-shaped. The shapes of the hazard can be described using the parameters α, γ as follows

1. $\{(\alpha,\gamma):\alpha \geq max(1,1/\gamma)\}$ the hazards are increasing
2. $\{(\alpha,\gamma):1<\alpha<1/\gamma\}$ the hazards are bathtub-shaped

3. $\{(\alpha,\gamma) : (1/\gamma < \alpha < 1) \text{ or } (\alpha \leq 0)\}$ the hazards are arc-shaped

4. $\{(\alpha,\gamma) : 0 < \alpha \leq min(1, 1/\gamma)\}$ decreasing hazard rate.

A graphical representation is given in Figure 2.1. A similar graph is provided by Cox et al. (2007) for another parameterization of this distribution.

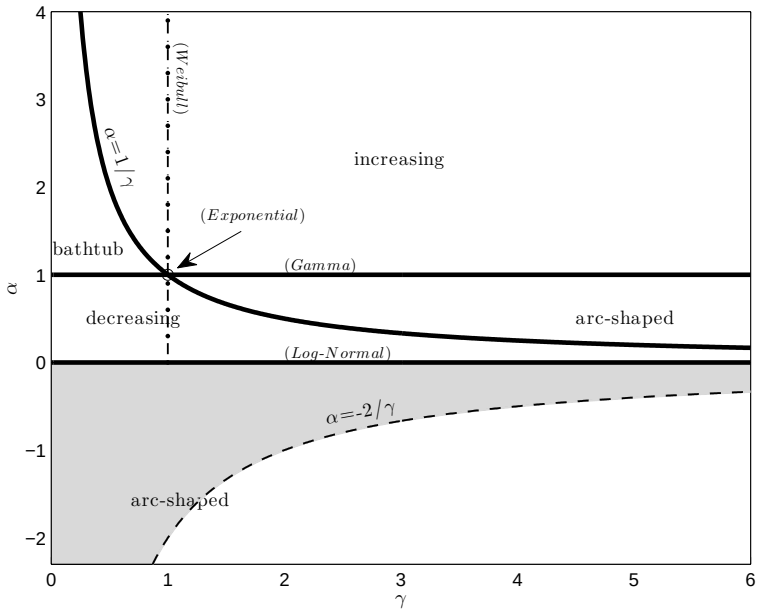

Figure 2.1: The various forms of the hazard rate of the GG(α,λ,γ) distribution depend on the parameters α and γ. The three solid curves define five regions that include the most common forms of the hazard rate. When the identity link function is used, then the method is inapplicable in the shaded region.

The computational issues of fitting the generalized gamma distribution to data are still under study when either the sample size is relatively small or the proportion of censored data is very high (or both).

To illustrate the applicability of the proposed method using the GG(α,λ,γ) distribution for the covariate we focus on the three most widely used generalized linear models, discussed in Section 3.3. We observe that (i) in the linear case the constraint $E(X_i^2) < \infty$ implies that $\alpha > 0$ or $\alpha < -2/\gamma$, (ii) in the binary case, no problems arise since the response variance is always finite and (iii) in the case of count data with the log link function, the constraint $M_X(2\beta_1) < \infty$ implies that our method can be used

(I) in the region $\{(\alpha,\gamma) : \alpha = 1, \gamma < 1\}$ for slope values $\beta_1 \leq \frac{\gamma\lambda}{2}$ (Gamma distribution)

(II) in $\{(\alpha,\gamma) : \alpha = 1, \gamma \geq 1\}$ for slope values $\beta_1 < \frac{\gamma\lambda}{2}$ (Gamma distribution)

(III) in $\{(\alpha,\gamma) : \alpha < 1, \gamma > 0\}$ for slope values $\beta_1 \leq 0$

(IV) in $\{(\alpha,\gamma) : \alpha > 1, \gamma \geq 0\}$ for all slope values.

Thus, for count data with the log link, the method can be used if the slope $\beta_1 \leq 0$ regardless the shape of the hazard rate. When the hazard is bathtub shaped the method is always applicable. For decreasing hazard rates the only case where the method works for positive slopes is when X follows the Gamma distribution. For increasing hazard rates the method works for all positive slopes except when X follows the Gamma distribution. In any other case one should refer to the restrictions above.

2.3.2 Case II: Additional fully observed covariates

Assume an AFT model of the form (2.10) with unknown vector of coefficients $\boldsymbol{\xi}$ where u_i follows the log−generalized gamma distribution with density

$$f(u_i) = \frac{1}{\Gamma(\delta^{-2})}|\delta| \left\{exp(\delta u_i)/\delta^2\right\}^{1/\delta^2} exp\left\{-exp(\delta u_i)/\delta^2\right\}, \quad \text{if } \delta \neq 0.$$

When $\delta = 0$, $f(u_i)$ is taken to be a standard normal density. The AFT model implies that $X_i|z_{i1}, ..., z_{i,p-1}$ follows the GG(α,λ_i,γ) with $\lambda_i = exp(-[1, \mathbf{z}_i^{'}]\boldsymbol{\xi}),\ \alpha = \frac{\delta}{\sigma},\ \gamma = \frac{1}{\delta^2}$.

$$f(u_i) = \frac{1}{\Gamma(\delta^{-2})}|\delta| \left\{exp(\delta u_i)/\delta^2\right\}^{1/\delta^2} exp\left\{-exp(\delta u_i)/\delta^2\right\}, \quad \text{if } \delta \neq 0.$$

When $\delta = 0$, $f(u_i)$ is taken to be a standard normal density. The AFT model implies that $X_i|z_{i1}, ..., z_{i,p-1}$ follows the GG(α,λ_i,γ) with $\lambda_i = exp(-[1, \mathbf{z}_i^{'}]\boldsymbol{\xi}),\ \alpha = \frac{\delta}{\sigma},\ \gamma = \frac{1}{\delta^2}$.

The required moment condition for continuous data is $E(X_i^2|z_i) < \infty, \forall i$. Observe that in the generalized gamma AFT regression model the covariate z_i affects only the parameter λ_i which plays no role in the existence of the moments of X_i. Thus, if the moment condition is satisfied by one value of z then it will be satisfied for all values of z.

In the case of count data, we require that $M_{X_i|z_i}(2\beta_1) < \infty, \forall i$. Note that according to the discussion in Section 2.3.1 the method is applicable, irrespective of the covariate values, (i) for all β_1 when $\alpha > 1$ and $\gamma \geq 0$, (ii) for $\beta_1 \leq 0$ when $\alpha < 1$ and $\gamma > 0$. When $\alpha = 1$ then the method may or may not work for $\beta_1 > 0$, depending on the observed values of z in our data set.

2.3.3 Simulation Studies

We conducted Monte Carlo simulations for each of the three examples in Sections 2.2.2.1, 2.2.2.2, 2.2.2.3 as well as for the scenario of a linear model with one additional regressor (section 2.3.3.2).

In all simulations with a single covariate, X was generated from an Exponential distribution with mean 3. The censoring variable, C, was equal to $min(C^*, cutpoint)$ where C^* had an exponential distribution. The mean of C^* and the cutpoint were chosen so that half of the censorings were expected to occur due to the cutpoint and the other half due to random censoring. Two scenarios, one with 30% and one with 70% of expected total censoring, were considered. Simulations were conducted for sample sizes of $n = 300$ and $n = 100$ in the linear case and of $n = 300$ in the cases of binary and count data. We used the proposed method assuming the following distributions for the censored covariate: (i) the Exponential, (ii) the Weibull, and (iii) the generalized gamma. We denote the methods by QS(Exp), QS(Weib), QS(GG) respectively. Since the Exponential belongs to the family of the Weibull distributions, which in turn belongs to the family of the

GG distributions we were able to assess the loss in efficiency when moving to a broader family of distributions. We also performed the Unweighted method (denoted by Un(Exp), Un(Weib), Un(GG)) in the linear model. We did not apply the Unweighted method in the cases of binary and count data since very slow convergence of the Fisher Scoring algorithm was observed. We compared results of our methods to those obtained by the Complete Case analysis (CC).

Due to the known computational problems regarding the fitting of the GG distribution to the data (see also Gomes et al. (2008) for a detailed overview), especially when censoring is present, we did not consider the QS(GG) or the Un(GG) method for the scenarios of $n = 100$. For sample sizes of $n = 300$ and 70% censoring, we ended up discarding approximately 0.03% of the repetitions. In all tables presenting the simulation results the columns titled as, 'Bias', 'SE', and 'MSE' contain the simulation based estimates of the Bias, Standard Error and Mean Squared Error of the estimates. The columns titled as 'Width' and 'Cover' contain the observed average width and coverage of the asymptotic 95% confidence intervals. We note again that the asymptotic covariance used is an estimate not adjusted for the estimation of the dispersion parameter and the parameters of the censored covariate. For some cases where the performance of our confidence intervals was extremely poor, we also show the corresponding results for the bootstrap based coverage.

2.3.3.1 Simple Linear Model

We first simulated from the simple normal linear regression $Y = 5 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0, 1)$. Simulations were conducted for the values of β_1 corresponding to correlation values of $\rho = 0.2, 0.3, 0.5, 0.8$. The results are presented in Table 2.1. The proposed methods outperform the CC analysis for all values of ρ considered with more dramatic differences in MSE occurring the lower the ρ is. We note here that most often in practice ρ is less than 0.5 and for those situations our simulations indicate a clear superiority of our methods. As expected the use of the Exponential or the Weibull distribution yielded even better results since fewer parameters needed to be estimated.

When ρ is large the asymptotic confidence intervals may have poor coverage, caused by the estimation of the asymptotic SE. We performed some additional simulations (not presented here) in order to explore the coverage properties when the dispersion parameter, τ, is set to its true value. We observed no differences compared to the cases where the use of the CC estimate of τ was used. However, when the parameter of the distribution of the covariate X (here the Exponential) was set to its true value (and τ was estimated by the CC) we noticed significant improvement in terms of MSE as well as nice coverage properties. Thus we conclude that the poor coverage observed in some scenarios is primarily due to the estimation of the parameters of the distribution of the covariate. However, when the percentile bootstrap technique was employed, the corresponding coverage was satisfactory. This is highlighted in the footnotes of Table 2.1, as well as in Tables A2 and A4 of the Appendix A. We see that when the correlation is very high and the GG is assumed as the covariate distribution the coverage of the confidence intervals that use the estimated asymptotic variance is unacceptable (lower than 30%). In contrast, the confidence intervals based on the resampling technique attains coverage close to the nominal one. We note that in order to apply the bootstrap technique we resample pairs of values of the response and the covariate. The dispersion parameter as well as

the parameters of the distribution of the covariate are re-estimated in each bootstrap iteration. Thus, it is not surprising that inference based on the bootstrap considerably outperforms inference based on the estimated asymptotic covariance matrix.

We also present results for the maximum likelihood method (ML), assuming the true form of the distribution of our data. As noted in Section 2.1, estimation of the MLE in this special setting of a simple normal linear regression with an exponentially distributed covariate is computationally feasible. Note that in the case of 30% total censoring the results of the maximum likelihood method and the proposed method are relatively close. Minor differences in MSE occur between the QS and the Unweighted method and one might be tempted to choose the latter due to computational simplicity.

In order to investigate the robustness, we also considered a t distribution with 4 d.f. for the error term. In these simulations the ML method falsely assumed a $N(0,1)$ distribution for the error term. In such scenarios, although the ML method proved fairly robust to misspesification of the error term, it exhibited a higher MSE for lower values of ρ. The results are presented in Appendix A (Tables A2 and A3).

We also performed simulations for small $(n = 100)$ and large $(n = 1000)$ sample sizes. Since computational problems occur when fitting the GG distribution for small sample sizes, we did not perform simulations based on the GG for $n = 100$. All the results of simulations regarding sample size $n = 100$ are presented in Appendix A (Tables A1 and A3). We further considered one simulation study with sample size equal to 1000. The results are presented in Appendix A (Table A4) and the conclusions are similar to the case with $n = 300$. In this scenario the proposed methods continue to outperform the CC and yield minor differences in terms of MSE compared to the likelihood approach (where the correct model for the response and the covariate is assumed). Finally we note that when the GG distribution is used as the censored covariate distribution, our estimates and the CC estimates are close, in terms of MSE, when ρ is large. The CC actually outperformed our estimates for $\rho = 0.8$, $n = 1000$, and 70% censoring.

2.3.3.2 Linear Model with an additional fixed covariate

We conducted an additional simulation considering the case of two covariates, one being censored and the other fully observed. The true values of the parameters for this simulation, were close to the corresponding estimates of the application presented in Section 2.5 (n=318). We performed the method based on the optimal estimating function using the Weibull (QS(Weib)) as well as the GG AFT model (QS(GG)). In the first scenario, we generated the time variable from the Weibull AFT model $log(X|z) = 6.5038 - 0.0263z + 0.7828u$ where u follows the extreme value distribution. The model of interest is $Y = \beta_0 + \beta_1 X + \beta_2 z + \epsilon$, where $\epsilon \sim N(0, 0.9^2)$, and the true values of β_0, β_1 and β_2 were set at $3.4, -0.007$ and -0.036 respectively. The censoring variable follows a $Weib(93, 2.6)$, thus the expected censoring was approximately 70%. The values of age (z) were held fixed at the observed values of the data in the application. We compared our results to the ones provided by the CC analysis and observed that the proposed method (QS) outperformed the CC analysis in terms of MSE. The O_F optimality based method using the Weibull AFT regression model yielded smaller bias and standard error for all three coefficients. The results are presented in Table 2.2.

Table 2.1: Simulation results for 1000 repetitions for the linear case ($n = 300$). Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from $N(0, 1)$.

			β_0					β_1				
Cens.	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
		$Likelihood$	-0.0199	0.0719	0.0056	- - -	- - -	0.0064	0.0153	0.0003	- - -	- - -
		CC	-0.0025	0.1152	0.0133	0.4521	0.9530	0.0016	0.0558	0.0031	0.2197	0.9450
		$QS(Exp)$	-0.0005	0.0894	0.0080	0.3537	0.9610	-0.0001	0.0232	0.0005	0.0918	0.9550
	0.2	$Un(Exp)$	-0.0005	0.0893	0.0080	0.3537	0.9690	-0.0001	0.0232	0.0005	0.0918	0.9540
		$QS(Weib)$	-0.0008	0.0897	0.0080	0.3542	0.9590	0.0000	0.0235	0.0006	0.0920	0.9580
		$Un(Weib)$	-0.0008	0.0897	0.0080	0.3542	0.9570	0.0000	0.0235	0.0006	0.0920	0.9580
		$QS(GG)$	-0.0000	0.0900	0.0081	0.3538	0.9580	-0.0005	0.0243	0.0006	0.0917	0.9460
		$Un(GG)$	-0.0000	0.0900	0.0081	0.3539	0.9570	-0.0005	0.0243	0.0006	0.0917	0.9460
		$Likelihood$	0.0020	0.0855	0.0073	- - -	- - -	0.0002	0.0225	0.0005	- - -	- - -
		CC	0.0055	0.1161	0.0135	0.4530	0.9510	-0.0019	0.0565	0.0032	0.2205	0.9460
		$QS(Exp)$	0.0021	0.0890	0.0079	0.3537	0.9560	0.0001	0.0241	0.0006	0.0344	0.9330
	0.3	$Un(Exp)$	0.0020	0.0890	0.0079	0.3537	0.9560	0.0001	0.0241	0.0006	0.0348	0.9330
		$QS(Weib)$	0.0016	0.0896	0.0080	0.3558	0.9500	0.0004	0.0248	0.0006	0.0343	0.9320
		$Un(Weib)$	0.0015	0.0896	0.0080	0.3560	0.9530	0.0004	0.0248	0.0006	0.0343	0.9310
		$QS(GG)$	0.0031	0.0898	0.0081	0.3540	0.9560	-0.0007	0.0264	0.0007	0.0335	0.9240
		$Un(GG)$	0.0031	0.0989	0.0081	0.3554	0.9560	-0.0007	0.0264	0.0007	0.0335	0.9260
30%		$Likelihood$	0.0062	0.0903	0.0082	- - -	- - -	-0.0015	0.0257	0.0007	- - -	- - -
		CC	0.0013	0.1140	0.0130	0.4542	0.9590	0.0010	0.0554	0.0031	0.2199	0.9540
		$QS(Exp)$	0.0045	0.0932	0.0087	0.3591	0.9470	-0.0012	0.0263	0.0007	0.0989	0.9410
	0.5	$Un(Exp)$	0.0047	0.0934	0.0087	0.3599	0.9460	-0.0012	0.0265	0.0007	0.0993	0.9400
		$QS(Weib)$	0.0043	0.0938	0.0088	0.3593	0.9430	-0.0012	0.0279	0.0008	0.0989	0.9230
		$Un(Weib)$	0.0043	0.0938	0.0088	0.3593	0.9430	-0.0012	0.0279	0.0008	0.0989	0.9230
		$QS(GG)$	0.0055	0.0974	0.0095	0.3594	0.9340	-0.0021	0.0333	0.0011	0.0990	0.8570
		$Un(GG)$	0.0059	0.0979	0.0096	0.3604	0.9300	-0.0023	0.0340	0.0012	0.0995	0.8500
		$Likelihood$	0.0005	0.0954	0.0091	- - -	- - -	-0.0004	0.0334	0.0011	- - -	- - -
		CC	0.0014	0.1140	0.0130	0.4520	0.9530	-0.0010	0.0560	0.0031	0.2196	0.9480
		$QS(Exp)$	0.0013	0.0971	0.0094	0.3761	0.9480	-0.0007	0.0340	0.0012	0.1243	0.9320
	0.8	$Un(Exp)$	0.0001	0.1009	0.0102	0.3885	0.9460	-0.0004	0.0365	0.0013	0.4282	0.9380
		$QS(Weib)$	-0.0010	0.1004	0.0101	0.3761	0.9420	0.0005	0.0387	0.0015	0.1244	0.8930
		$Un(Weib)$	-0.0024	0.1055	0.0111	0.3880	0.9340	0.0011	0.0429	0.0018	0.1313	0.8750
		$QS(GG)$	0.0018	0.1112	0.0124	0.3765	0.9100	-0.0016	0.0516	0.0027	0.1253	0.7660
		$Un(GG)$	0.0035	0.1215	0.0148	0.3907	0.8460	-0.0032	0.0605	0.0037	0.1330	0.7190
		$Likelihood$	-0.0324	0.0850	0.0083	- - -	- - -	0.0103	0.0218	0.0006	- - -	- - -
		CC	0.0044	0.1868	0.0349	0.7521	0.9510	-0.0113	0.2607	0.0681	1.0403	0.9460
		$QS(Exp)$	-0.0024	0.1203	0.0145	0.4713	0.9430	0.0001	0.0354	0.0013	0.1381	0.9500
	0.2	$Un(Exp)$	-0.0024	0.1204	0.0145	0.4714	0.9410	0.0001	0.0354	0.0013	0.1381	0.9520
		$QS(Weib)$	-0.0029	0.1210	0.0146	0.4731	0.9480	0.0010	0.0378	0.0014	0.1404	0.9350
		$Un(Weib)$	-0.0030	0.1210	0.0147	0.4731	0.9430	0.0010	0.0378	0.0014	0.1404	0.9350
		$QS(GG)$	-0.0038	0.1227	0.0151	0.4763	0.9460	0.0024	0.0503	0.0025	0.1465	0.7980
		$Un(GG)$	-0.0036	0.1224	0.0150	0.4752	0.9470	0.0020	0.0492	0.0024	0.1430	0.7850
		$Likelihood$	0.0012	0.1036	0.0107	- - -	- - -	-0.0000	0.0315	0.0010	- - -	- - -
		CC	0.0048	0.1944	0.0378	0.7544	0.9550	-0.0005	0.2697	0.0727	1.0437	0.9470
		$QS(Exp)$	0.0038	0.1210	0.0147	0.4733	0.9420	-0.0008	0.0378	0.0014	0.1398	0.9300
	0.3	$Un(Exp)$	0.0037	0.1210	0.0147	0.4734	0.9440	-0.0008	0.0378	0.0014	0.1399	0.9310
		$QS(Weib)$	0.0025	0.1224	0.0150	0.4751	0.9360	0.0011	0.0425	0.0018	0.1426	0.8990
		$Un(Weib)$	0.0023	0.1224	0.0150	0.4753	0.9390	0.0011	0.0425	0.0018	0.1426	0.9000
		$QS(GG)$	0.0001	0.1226	0.0159	0.4781	0.9320	0.0038	0.0648	0.0042	0.1479	0.6940
		$Un(GG)$	0.0001	0.1258	0.0158	0.4763	0.9330	0.0037	0.0647	0.0042	0.1440	0.6840
70%		$Likelihood$	0.0083	0.1039	0.0109	- - -	- - -	-0.0021	0.0356	0.0013	- - -	- - -
		CC	0.0003	0.1871	0.0350	0.7513	0.9560	0.0056	0.2599	0.0676	1.0329	0.9430
		$QS(Exp)$	0.0046	0.1187	0.0141	0.4747	0.9590	-0.0011	0.0398	0.0016	0.1451	0.9320
	0.5	$Un(Exp)$	0.0046	0.1187	0.0141	0.4759	0.9590	-0.0011	0.0397	0.0016	0.1453	0.9330
		$QS(Weib)$	0.0030	0.1198	0.0144	0.4762	0.9610	0.0006	0.0507	0.0026	0.1466	0.8480
		$Un(Weib)$	0.0030	0.1197	0.0143	0.4772	0.9610	0.0005	0.0506	0.0026	0.1468	0.8580
		$QS(GG)$	-0.0032	0.1295	0.0168	0.4813	0.9460	0.0110	0.1018	0.0105	0.1588	0.5320
		$Un(GG)$	-0.0023	0.1290	0.0167	0.4775	0.9350	0.0095	0.1014	0.0104	0.1466	0.5042
		$Likelihood$	0.0015	0.1063	0.0113	- - -	- - -	-0.0009	0.0548	0.0030	- - -	- - -
		CC	-0.0020	0.1868	0.0349	0.7517	0.9490	0.0061	0.2562	0.0657	1.0375	0.9580
		$QS(Exp)^{(1)}$	0.0020	0.1246	0.0155	0.4815	0.9430	-0.0016	0.0574	0.0033	0.1723	0.8590
	0.8	$Un(Exp)$	0.0010	0.1277	0.0163	0.4940	0.9480	-0.0015	0.0576	0.0033	0.1747	0.8620
		$QS(Weib)^{(2)}$	-0.0051	0.1330	0.0177	0.4837	0.9260	0.0078	0.0971	0.0095	0.1752	0.6460
		$Un(Weib)$	-0.0061	0.1358	0.0185	0.4961	0.9290	0.0082	0.0977	0.0096	0.1776	0.6560
		$QS(GG)^{(3)}$	-0.0158	0.1762	0.0300	0.4881	0.8530	0.0286	0.2204	0.0494	0.1864	0.2730
		$Un(GG)$	-0.0157	0.1763	0.0313	0.4782	0.8370	0.0272	0.2240	0.0509	0.2176	0.2190

(1) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 95.3% and 93.2% respectively
(2) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 93.0% and 94.3% respectively
(3) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 95.2% and 96.1% respectively

Table 2.2: Simulation when an additional fixed covariate is present. The true values of the parameters β_0, β_1 and β_2 are 3.4, -0.007 and -0.036 respectively. The time variable was generated from an AFT model using the extreme value.

Method	Parameter	Est.	SE	Bias	MSE	Asympt. Cover.	Boots. Cover.
CC	β_0	3.4299	0.5231	0.0299	0.2746	0.9540	-
	β_1	-0.0069	0.0030	4×10^{-5}	9×10^{-6}	0.9370	-
	β_2	-0.0366	0.0095	-0.0007	9×10^{-5}	0.9490	-
QS(Weib)	β_0	3.37927	0.4531	-0.0207	0.2057	0.9090	0.9410
	β_1	-0.0069	0.0012	10^{-5}	10^{-6}	0.7490	0.9510
	β_2	-0.0356	0.0083	0.0003	7×10^{-5}	0.8990	0.9390
QS(GG)	β_0	3.3962	0.4222	-0.0037	0.2136	0.9020	0.9510
	β_1	-0.0074	0.0022	0.0004	5×10^{-6}	0.4210	0.9640
	β_2	-0.0356	0.0084	0.0003	7×10^{-5}	0.8730	0.9590

2.3.3.3 Binary data

The real value of the intercept was set at $\beta_0 = log(9)$ so that $P(Y = 1|X = 0) = 0.9$. If $Y = 1$ denotes a positive marker result, this implies a 90% chance of a positive test at the time of death (!). Three values of β_1 were studied, $\beta_1 = -2, -3, -4$. These β_1 values were chosen in order to avoid total separation and the 'only successes' or 'only failures' scenario. A graphical representation of the scenarios considered is given in Appendix A (Figure A1). The results are presented in Table 2.3. The simulations showed superiority of the proposed method (QS) over the CC in all cases. Moreover, negligible gains in MSE are obtained using the likelihood method. As in the linear case, when the exponential distribution was fitted to the covariate the results were even better, because fewer parameters needed to be estimated. The likelihood method was not implemented here due to its computational cost. We also omitted the Unweighted method since very slow convergence of the Fisher Scoring algorithm was observed. The results are presented in Table 2.3.

2.3.3.4 Count data

For count data we considered the same censoring mechanism as in section 2.3.3.1. The real value of the intercept was $\beta_0 = 1$. For the slope we considered $\beta_1 = -1$ and $\beta_1 = -1/3$ for each scenario. Note that the relative change in mean $(E(Y|X = 0) - E(Y|X = median))/E(Y|X = 0)$ is 0.875 for $\beta_1 = -1$ and 0.5 for $\beta_1 = -1/3$. We set the overdispersion parameter at $\tau = 2$. The proposed method (QS) was superior to the CC in terms of MSE in all cases. The results are presented in Appendix A (Table A5). 14% of the repetitions in QS(GG) were discarded when the censoring was 70% and $\beta_1 = -1/3$, due to violations of the conditions discussed in section 2.2.2.3. Overall, the coverage seems satisfactory. However, we observed that when fitting the GG in the presence of high censoring the proposed method yielded lower coverage when $\beta_1 = -1/3$ with censoring level of 70%. This problem can be circumvented with the use of the percentile bootstrap as in the linear case.

Table 2.3: Simulation results of 1000 repetitions in the case of binary data. Sample size is n=300, with 70% and 30% censoring, half of which is due to the cutpoint (end of study), and the other half due to random censoring. The logit link function was used. The real value of the intercept is $log(9)$ so that $P(Y = 1|X = 0) = 0.9$.

			β_0					β_1				
Cens.	β_1	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
		CC	0.0685	0.3796	0.1488	1.4574	0.9490	-0.0654	0.3081	0.0992	1.1604	0.9530
		$Likelihood$	0.0643	0.3683	0.1405	–	–	-0.0599	0.2927	0.0893	–	–
	-2	$QS(Exp)$	0.0641	0.3693	0.1405	1.4381	0.9390	-0.0598	0.2927	0.0892	1.1226	0.9550
		$QS(Weib)$	0.0644	0.3694	0.1406	1.4382	0.9390	-0.0599	0.2928	0.0893	1.1226	0.9550
		$QS(GG)$	0.0640	0.3695	0.1406	1.4381	0.9390	-0.0595	0.2929	0.0893	1.1223	0.9530
		CC	0.0610	0.4329	0.1912	1.6801	0.9490	-0.0990	0.5124	0.2723	1.9103	0.9360
		$Likelihood$	0.0588	0.4295	0.1879	–	–	-0.0951	0.5000	0.2591	–	–
	-3	$QS(Exp)$	0.0587	0.4295	0.1897	1.1042	0.9520	-0.0951	0.5000	0.2590	1.5126	0.9440
30%		$QS(Weib)$	0.0589	0.4294	0.1879	1.6673	0.9520	-0.0949	0.4998	0.2588	1.8722	0.9440
		$QS(GG)$	0.0589	0.4295	0.1879	1.6672	0.9430	-0.0949	0.5000	0.2590	1.8723	0.9520
		CC	0.0956	0.5025	0.2616	1.9035	0.9610	-0.1716	0.7412	0.5788	2.8175	0.9570
		$Likelihood$	0.0941	0.5001	0.2589	–	–	-0.1712	0.7323	0.5656	–	–
	-4	$QS(Exp)$	0.0941	0.5001	0.2589	1.8928	0.9680	-0.1712	0.7323	0.5656	2.7776	0.9560
		$QS(Weib)$	0.0942	0.5001	0.2590	1.8930	0.9480	-0.1709	0.7321	0.5651	2.7772	0.9560
		$QS(GG)$	0.0942	0.5004	0.2592	1.8930	0.9670	-0.1713	0.7323	0.5656	2.7777	0.9550
		CC	0.0957	0.5662	0.3297	2.1175	0.9520	-0.1039	0.6944	0.4931	2.6001	0.9480
		$Likelihood$	0.0651	0.4370	0.1952	–	–	-0.0620	0.3753	0.1447	–	–
	-2	$QS(Exp)$	0.0623	0.4368	0.1967	1.6518	0.9520	-0.0590	0.3760	0.1449	1.3689	0.9400
		$QS(Weib)$	0.0673	0.4430	0.2008	1.6532	0.9430	-0.0635	0.3937	0.1590	1.3694	0.9270
		$QS(GG)$	0.0689	0.4499	0.2071	1.6524	0.9550	-0.0583	0.4179	0.1781	1.3657	0.9270
		CC	0.0853	0.5381	0.2969	2.0700	0.9550	-0.1490	0.7417	0.5723	2.8563	0.9640
		$Likelihood$	0.0602	0.4759	0.2301	–	–	-0.1002	0.5653	0.3296	–	–
	-3	$QS(Exp)$	0.0587	0.4760	0.2300	1.8079	0.9520	-0.0984	0.5662	0.3303	2.0917	0.9380
70%		$QS(Weib)$	0.0634	0.4780	0.2325	1.8080	0.9540	-0.1054	0.5731	0.3395	2.0905	0.9380
		$QS(GG)$	0.0635	0.4799	0.2343	1.8076	0.9525	-0.1046	0.5764	0.3431	2.0885	0.9360
		CC	0.1163	0.5557	0.3223	2.1594	0.9740	-0.2219	0.9260	0.9083	3.5378	0.9640
		$Likelihood$	0.1049	0.5150	0.2763	–	–	-0.1918	0.7775	0.6413	–	–
	-4	$QS(Exp)$	0.1040	0.5150	0.2761	2.0101	0.9690	-0.1909	0.7777	0.6413	3.0124	0.9660
		$QS(Weib)$	0.1061	0.5156	0.2771	2.0100	0.9710	-0.1938	0.7788	0.6441	3.0112	0.9660
		$QS(GG)$	0.1077	0.5166	0.2785	2.0118	0.9710	-0.1976	0.7811	0.6491	3.0147	0.9650

2.4 Exploring non parametric models for the censored covariate in a GLM

Modeling the covariate by a parametric model has the disadvantage of making strong assumptions that in practice may be not justified. At the other extreme lies the non parametric product limit estimator introduced by Kaplan and Meier (1958) (KM). However, the crude product limit estimator does not allow estimation of survival probabilities beyond the greatest event time. There are settings where the largest time value may correspond to a censored observation. In this case even if the last event is followed by (greater) censored values, the estimation of the survival function is limited to the last event (t_{max}).

Some strategies have been imposed to remedy this drawback. Efron (1967) proposes setting survival probabilities at time points beyond t_{max} equal to 0 while Gill (1980) proposes setting them equal to $\hat{S}(t_{max})$. Efron's and Gill's approaches turn out to be negatively and positively biased respectively. An approach regarding the completion of the tail of the Kaplan Meier estimator by an exponential curve is discussed in Brown et al. (1974) and the use of the more flexible Weibull parametric model was suggested by Klein and Moeschberger (1985). Another approach may be the use of even more flexible parametric models for completing the tails such as the generalized gamma model. In general simulation studies have shown that smooth estimates of the survival function are more efficient than the crude non parametric one (Pan (2000)).

The most celebrated methods to smooth non parametrically the survival function are

based on kernel smoothers and splines. The kernel based methodologies are discussed in detail by Silverman (1986) in the case of no censoring. Standard references that include issues of censoring are Wand and Jones (1995) and Bowman and Azzalini (1997). Kernel smoothers are widely used in survival analysis, but they too suffer from the drawback of the 'last is censored' phenomenon.

In section 2.4.1 we briefly recall the logspline density estimation approach, as well as the standard kernel smoothing method. In section 2.4.1.3 we introduce a new constrained natural spline based approach in estimating the survival function and discuss the case where covariates are present.

2.4.1 Non-Parametric Approaches for Survival Estimation of a Censored Variable

2.4.1.1 Log-spline Models

Logspline models have been studied in Stone and Koo (1986), Stone (1990) and Kooperberg and Stone (1991). In Kooperberg and Stone (1992) logspline density estimation was developed for censored data. Here we briefly recall their methodology.

Consider the data (T_i, D_i) $i = 1, \ldots, n$, where $T_i = min(X_i, C_i)$ is a survival time random variable, C_i is the censoring variable, and D_i is a binary indicator variable taking values 1 for an event and 0 for censoring, i.e. $D_i = I(X_i < C_i)$. Let the integer $K \geq 3$, and the knot sequence $\tau_1, \ldots, \tau_K$ with with $-\infty \leq L < \tau_1 < \ldots < \tau_K \leq U \leq \infty$ where L and U are some numbers.

The logspline density model is stated as

$$f(x; \boldsymbol{\theta}) = exp(\theta_1 B_1(x) + \ldots + \theta_p B_p(x) - \mathcal{C}(\boldsymbol{\theta})), L < x < U, \tag{2.12}$$

where

$$\mathcal{C}(\boldsymbol{\theta}) = log(\int_L^U exp(\theta_1 B_1(x) + \ldots + \theta_p B_p(x))dx)$$

is the normalizing constant and the basis functions $B_1(x), B_2(x), \ldots, B_p(x)$ can be chosen such that B_1 is linear with negative slope on $(L, \tau_1]$, $B_2, \ldots, B_p$ are constant on $(L, \tau_1]$, B_p is linear with positive slope on $[\tau_K, U)$, $B_1, \ldots, B_{p-1}$ are constant on $[\tau_K, U)$, and in each of the intervals $[\tau_1, \tau_2], \ldots, [\tau_{K-1}, \tau_K]$ we have a cubic polynomial. This way a natural cubic spline is formed. Note that for the above model the feasibility condition $\int_L^U exp(\theta_1 B_1(x) + \ldots + \theta_p B_p(x))dx < \infty$ is required. The survival function is then given by

$$S(x; \boldsymbol{\theta}) = 1 - \int_L^x f(z, \boldsymbol{\theta})dz, \quad L < x < U.$$

Note that if we set $U = \infty$, then the density function is exponential on $[\tau_K, \infty)$ and if $L = -\infty$ then the density is exponential on $(-\infty, \tau_1]$. Here, we consider time to event data, thus we expect the density to be positive supported, hence $L = 0$. If X_i is right censored $(T_i < X_i)$ or observed exactly $(T_i = X_i)$, then $A_i = (T_i, U_i)$ or $A_i = X_i$ respectively. Under the assumption that the random sample is independent of the censoring mechanism, a maximum likelihood estimate of the vector $\boldsymbol{\theta}$ is then obtained by maximizing the log-likelihood

$$l(\boldsymbol{\theta}) = \sum_i \varphi(A_i, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta,$$

where $\varphi(A; \boldsymbol{\theta}) = log(\int_A f(x; \boldsymbol{\theta}))$ and $f(x; \boldsymbol{\theta})$ is given by (2.12). The maximization of the likelihood is possible via the Newton Raphson iterative procedure. Note that when there is no censoring the Hessian is globally negative definite and the log likelihood function is strictly concave, and hence the maximum likelihood of $\boldsymbol{\theta}$ is unique. However, in the presence of censoring, this is not always the case (Stone (1990)).

For choosing the number of knots Kooperberg and Stone (1992) apply a stepwise procedure for addition and deletion of knots depending on their statistical significance and model selection is based on AIC or BIC. Since the primary target of their methodology is the estimate of the density, they present a sophisticated initial knot placement based on experience. The knot placement in the density estimation setting is crucial, since many peaks of the density may have to be detected without making the estimate very noisy.

2.4.1.2 Kernel Smoothing

Kernel smoothers have received much attention in the literature. See for example Silverman (1986) and in the case of censoring Wand and Jones (1995) and Bowman and Azzalini (1997). The formula for the kernel density estimator in the case of no censoring is

$$\hat{f}(x; h) = (nh)^{-1} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h}\right\}$$

where $K(\cdot)$ is the kernel function for which $K(x) = 1$ and $h > 0$ is called the bandwidth. The most popular kernels are the Epanechnikov, the Biweight, the Triweight, the Normal, the Triangular and the Uniform, mentioned with order of efficiency. We refer the reader to Wand and Jones (1995). The efficiencies of the Epanechnikov and the Uniform kernel are 1 and 0.930 respectively. It is evident that one loses very little in terms of performance by using a suboptimal kernel. In effect, the choice of the kernel may be relied on computationally simplicity. Note also that in practice the Epanechnikov kernel is sometimes avoided due to its discontinuous first derivative. In contrast to the choice of the kernel function, the bandwidth selection is a crucial issue. Various approaches have been made regarding bandwidth selection, from computationally simple that can be written in closed form, to computationally more cumbersome such as the cross validation based techniques. For a review of common bandwidth selection approaches see Wand and Jones (1995).

In the presence of censoring the kernel density estimator is based on the Kaplan Meier estimator. Let $T_{(i)}, D_{(i)}$, $i = 1, \ldots, n$ be $\{T_i, D_i\}$ ordered with respect to the T_i's. The Kaplan Meier survival estimator is given by:

$$\hat{S}^{KM}(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq T_{(1)} \\ \prod_{i=1}^{j-1} \left(\frac{n-i}{n-i+1}\right)^{D_{(i)}}, & \text{if } T_{(j-1)} \leq x \leq T_{(j)},\ j = 2, \ldots, n. \end{cases}$$

and the estimated kernel density is

$$\hat{f}_X(x) = \sum_{i=1}^{n} s_i K_h(x - T_i),$$

where $K_h(u) = h^{-1}K(u/h)$ and s_i is the size of the jump of the KM estimator at T_i. Survival probabilities can be estimated through the density estimator. Note that $s_i = 0$ if and only if T_i corresponds to a censored observation. This is where the drawback of kernel smoothing in the presence of censoring arises. The kernel estimate is constructed by centering a kernel at each event time. In effect, the density estimator is not extended further than the tail of the kernel placed at the last event, even if the last event is followed by censored data. Similarly, the same problem would rise in the case of estimating the distribution function by integrating the density (see also Azzalini (1980)). In effect, if one wants to use the kernel methodology when censoring is present, then the condition that the largest time is an event time is needed.

2.4.1.3 The HCNS Approach

In this section we consider the use of a positive monotone natural spline to smooth the nonparametric estimate of the cumulative hazard function (HCNS method: Hazard Constrained Natural Spline). The monotone increasing nature of the data obtained from the Kaplan Meier estimator of the cumulative hazard function

$$\hat{H}^{KM}(x) = -log(\hat{S}^{KM}(x))$$

is expected to set the ground for a cubic spline to be adequately flexible to be fitted to the points $(T_i, \hat{H}^{KM}(T_i)|D_i = 1)$, i.e. the jumps of the Kaplan Meier estimator, and provide a smooth estimate of the survival function.

Consider the K knots placed at $\tau_1 < \ldots < \tau_K$ and let the natural spline for the cumulative hazard

$$H(x) = \theta_1 W_1(x) + \theta_2 W_2(x) + \ldots + \theta_{K-2} W_{K-2}(x), \tag{2.13}$$

where for $j = 1, \ldots, K-2$ we have

$$W_j(x) = (x - \tau_j)_+^3 - \frac{(x - \tau_{K-1})_+^3 (\tau_K - \tau_j)}{\tau_K - \tau_{K-1}} + \frac{(x - \tau_K)_+^3 (\tau_{K-1} - \tau_j)}{\tau_K - \tau_{K-1}},$$

where $x_+ = max(0, x)$. It can be shown that $W_j(x)$ is linear in x for $x \geq \tau_K$. Model (2.13) can be written as

$$H(x) = \theta_1 (x - \tau_1)_+^3 + \ldots + \theta_{K-2}(x - \tau_{K-2})_+^3 + \theta_{K-1}(x - \tau_{K-1})_+^3 + \theta_K (x - \tau_K)_+^3 \tag{2.14}$$

where

$$\theta_{K-1} = \frac{\theta_1(\tau_1 - \tau_K) + \theta_2(\tau_2 - \tau_K) + \ldots + \theta_{K-2}(\tau_{K-2} - \tau_K)}{\tau_K - \tau_{K-1}} \tag{2.15}$$

$$\theta_K = \frac{\theta_1(\tau_1 - \tau_{K-1}) + \theta_2(\tau_2 - \tau_{K-1}) + \ldots + \theta_{K-2}(\tau_{K-2} - \tau_{K-1})}{\tau_{K-1} - \tau_K}.$$

Due to need of extrapolation beyond the last knot, Stone and Koo (1985) state advantages of linearly extrapolating the model, which is also the case for our model. The model (2.14) has the following properties:
(i) It is linear beyond the last knot,
(ii) It equals to zero before the first knot,
(iii) Its first and second derivative are continuous,
(iv) Its first derivative is zero at the first knot,
(v) It has $K-2$ parameters to be estimated.

For the model above to be fitted to the cumulative hazard step function monotonicity conditions are required. Necessary and sufficient conditions for the monotonicity of a cubic spline are given in Fritsch and Carlson (1980). Consider a cubic polynomial P in the interval $[\tau_j, \tau_{j+1}]$ and denote its first derivative by P'. Let $a_j = \frac{\tau_{j+1}-\tau_j}{P(\tau_{j+1})-P(\tau_j)} P^{'}(\tau_j)$ and $b_j = \frac{\tau_{j+1}-\tau_j}{P(\tau_{j+1})-P(\tau_j)} P^{'}(\tau_{j+1})$, be the respective ratios of the endpoint derivatives to the slope of the secant line. Consider the region $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, where $\mathcal{M}_1$ is the square defined by $a = 0, 3$ and $b = 0, 3$ and $\mathcal{M}_2$ is the ellipse defined by $\phi(a, b) = (a-1)^2 + (a-1)(b-1) + (b-1)^2 - 3(a+b-2) = 0$, which is tangent to the coordinates (0,3) and (3,0). Monotonicity in the interval $[\tau_j, \tau_{j+1}]$ holds within region $\mathcal{M}$ but outside this region the cubic polynomial is non-monotone. Under the further condition that $0 \leq min(P^{'}(\tau_j), P^{'}(\tau_{j+1}))$ the function is ensured to be non-decreasing. Note that any subregion of $\mathcal{M}$ provides a sufficient condition for monotonicity.

Note that the exact region defines a nonlinear condition for monotonicity and one might be tempted to consider other linear subregions such as $\mathcal{M}_1$ at the cost of excluding other candidate models that may provide a significantly better fit to the data at hand. On the other hand, one may try to approach the problem by using the entire region of monotonicity, $\mathcal{M}$. This would be computationally cumbersome because non linear constraints would be applied for each of the subintervals that are defined by the knots. In effect there would be no guarantee of convergence during an optimization procedure, and very good initial values satisfying the initial constraints would be necessary.

In Figure 2.2 we illustrate a linear approximation, $\mathcal{A}$, of the exact region $\mathcal{M}$ with the use of 16 line segments. These correspond to a linear spline beginning from $(a, b) = (0, 0)$ where the a values are depicted on the horizontal axis. We move counterclockwise starting from $(0, 0)$ for the values of a=(0, 3.0000, 3.4664, 3.8024, 3.9778, 3.9777, 3.8021, 3.4663, 3.0000, 2.4461, 1.8517, 1.2700, 0.7529, 0.3474, 0.0889, 0, 0). These values yield the optimal inscribed decahexagon within region $\mathcal{M}$ in terms of the enclosed area. Note that the captured area from $\mathcal{A}$ is approximately 98.4% of the entire region, $\mathcal{M}$. This linear approximation reduces the problem to a linear programming one.

We derived the values of a that correspond to the optimal decahexagon by numerically maximizing the area with respect to the points in the perimeter of region $\mathcal{M}$. However, there is a suboptimal way with minimum computational cost for calculating a linear approximating inscribed polygon within the area $\mathcal{M}$. In Smith (1970) an easy to implement algorithm for computing a piecewise representation of an ellipse is presented. This algorithm computes the optimal placement of a fixed number of points for an ellipse, in terms of inscribed area. However, only a part of our region $\mathcal{M}$ matches the curve of an ellipse and we need to fix three of our available points to the locations (0,0), (0,3) and (3,0). Thus, one computationally simple way to construct an efficient, but suboptimal, inscribed polygon within region $\mathcal{M}$ would be to fix these three points ((0,0), (0,3) and

(3,0)) and to include any points indicated by the algorithm of Smith (1970) that lie above the line $b = -a+3$. This way one can easily fit very fine linear approximations. Moreover, note that in the case that one chooses to construct a $12k-$gon, $k = 1, 2, \ldots$ to approximate the ellipse $\phi(a, b)$ using Smith's algorithm, then as discussed in the technical details at the end of this chapter, it happens that the points (0,3) and (3,0) are included in the approximation. In this case, we can discard all points below $b = -a + 3$ and use (0,0) instead, to derive the optimal inscribed $(8k + 2)-$gon within region $\mathcal{M}$. Thus by using Smith's algorithm we can easily optimally inscribe any $(8k + 2)-$gon within the region of monotonicity and hence approximate $\mathcal{M}$ to any desired degree of accuracy. See also the Appendix B. In the simulation studies we considered both the 16 line approximation, as well as the 18 line approximation in all scenarios and observed no differences in the results.

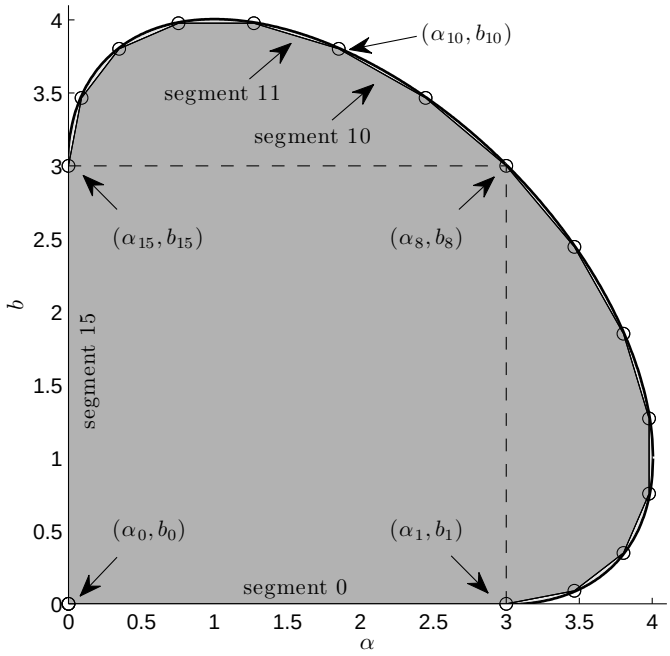

Figure 2.2: Linear approximation with 16 line segments of the entire region of monotonicity ($\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$) which is defined by the square ($a = 0, 3$ and $b = 3, 0$) and the ellipse $\phi(a, b) = (a-1)^2 + (a-1)(b-1) + (b-1)^2 - 3(a+b-2) = 0$ that are overlapping. The linear approximation $\mathcal{A}$ of region $\mathcal{M}$ is given by linearly joining the consecutive points beginning from $(0, 0)$ and moving counterclockwise starting from $(0, 0)$ for the values of a=(0, 3.0000, 3.4664, 3.8024, 3.9778, 3.9777, 3.8021, 3.4663, 3.0000, 2.4461, 1.8517, 1.2700, 0.7529, 0.3474, 0.0889, 0, 0). The area captured by region $\mathcal{A}$ is approximately 98.4% of entire region $\mathcal{M}$.

Under the constraints that force the cumulative hazard to be monotonically increasing, it is easy to show that all moments of the r.v. X exist if we force the derivative of the cumulative hazard to be strictly positive at the last knot ($H'(\tau_K) > 0$). We require

no further monotonicity restrictions beyond τ_K since the linear tail will ensure that the model $H(x)$ will be strictly increasing beyond τ_K.

We can choose the first knot to be placed at $\tau_1 = min(event\ times) = min(T_i|D_i = 1)$, and the last knot at $\tau_K = max(event\ times) = max(T_i|D_i = 1)$. Note that model (2.14) will always be smaller than the Kaplan Meier based estimate of $H(x)$ at $\tau_1 = min(T_i|D_i = 1)$, since we assume to have zero values before the first knot and $H'(\tau_1) = 0$. Thus, we expect to underestimate the cumulative hazard function near the interval $(0, \tau_1 = min(T_i|D_i = 1))$. Alternatively, one may choose to set $\tau_1 = 0$ so as to allow positive values of the cumulative hazard function near zero. The choice of $\tau_1 = min(T_i|D_i = 1)$ may be justifiable if we expect that $min(T_i|D_i = 1) >> 0$.

A usual way of placing the knots is at equally spaced quantiles (Harrell (2001)). Another possible approach is to select equally spaced knots between $min(T_i|D_i = 1)$ and $max(T_i|D_i = 1)$. We expect, again due to the nature of monotonically increasing data, that such knot placement strategies will be robust enough since no modes or valleys are need be detected and a sophisticated procedure such as the one presented in Kooperberg and Stone (1992) may not be needed. Note that if the first knot is placed at $min(T_i|D_i = 1)$, then property (ii) provides an imitation of the behavior of the Kaplan Meier estimator, in that it is zero before the first event.

In this chapter, we consider three knot placement schemes based on the non censored data: 1. Equally spaced knots from $min(T_i|D_i = 1)$ to $max(T_i|D_i = 1)$, 2. Knots at $min(T_i|D_i = 1)$, 5th, 25th, 50th, 75th percentiles and $max(T_i|D_i = 1)$, 3. Knots at 0, 5th, 25th, 50th, 75th percentiles and $max(T_i|D_i = 1)$. In a given application we choose the one that yields the smallest distance to the Kaplan Meier based cumulative Hazard estimator

$$\Psi(\hat{\boldsymbol{\theta}}) = \sum_i (\hat{H}(T_i|D_i = 1) - \hat{H}^{KM}(T_i|D_i = 1))^2,$$

where $\hat{H}$ is the fitted model defined in (2.14) under the appropriate constraints of monotonicity, and $\hat{H}^{KM}$ is the Kaplan Meier based cumulative hazard estimator.

Another approach is to consider the knots as free parameters that have to be estimated. Since we fit the model under study to the Kaplan Meier estimator a criterion regarding the goodness of fit would be to consider again $\Psi = \Psi(\boldsymbol{\theta}, \boldsymbol{\tau})$, where $\boldsymbol{\theta}$ is the vector of the parameters of interest, $\boldsymbol{\tau}$ is the vector of knots $[\tau_1, \tau_2, \ldots, \tau_K]$. After the estimate of the vector parameter $\boldsymbol{\theta}$ is obtained, it can be considered fixed and we can set $\Psi = \Psi(\hat{\boldsymbol{\theta}}, \boldsymbol{\tau})$ where Ψ now depends only on the knot placement. Minimizing Ψ with respect to vector $\boldsymbol{\tau}$, one can derive an improved fit of the model. However, function Ψ depends non linearly on the knots and the convergence of the optimization is not guaranteed. The three previously mentioned knot placement schemes presented no computational problems and seemed to be adequate in the simulation studies.

To derive the linear constraints we regard the points $(0, 0)$, $(0, 3)$ and $(3, 0)$ as included in the set of points that will define a linear approximation of the entire region $\mathcal{M}$. Thus, the set of distinct points we consider to define a linear approximation of $\mathcal{M}$ is $(a_0, b_0) = (a_{Q+2}, b_{Q+2}) = (0, 0)$, $(a_1, b_1) = (3, 0)$, (a_2, b_2), $\ldots$, (a_Q, b_Q), and $(a_{Q+1}, b_{Q+1}) = (0, 3)$. Denote the $i-$th segment that joins (a_i, b_i) with (a_{i+1}, b_{i+1}) by g_i, with its equation given by $g_i(a, b) = b + \xi_1^{(i)} a + \xi_0^{(i)} = 0$, where $i = 1, \ldots, Q$ (see also Figure 2.2 where $Q = 14$).

First, we need to adjust for the direction of the inequality that will allow us to be on

the correct half-plane for each of the Q line segments. The curve on the perimeter of $\mathcal{M}$ is convex for $b < 1$ and concave for $b > 1$.

For the $i-$th segment:

1. If $b_{i+1} \leq 1$ then we require $g_i(a, b) \geq 0$ (and set $v_i = -1$)
2. If $b_i \geq 1$ then we require $g_i(a, b) \leq 0$ (and set $v_i = 1$)
3. If $b_i < 1$ and $b_{i+1} > 1$ then
   - if $a_i < a_{i+1}$ we require $g_i(a, b) \geq 0$ (and set $v_i = -1$)
   - if $a_i \geq a_{i+1}$ we require $g_i(a, b) \leq 0$ (and set $v_i = 1$)

We derive the $Q(K-1)$ constraints defined by (2.16) that deal with capturing the region $\mathcal{A}$. See the technical details section of an example of $K = 6$ knots with an inscribed decahexagon ($Q = 14$). For each of the J, ($J = 1, 2, \ldots, K-1$), knot intervals $[\tau_1, \tau_2], [\tau_2, \tau_3], \ldots, [\tau_{K-1}, \tau_K]$ the restriction required for the $i-$th segment is given by

$$v_i \Bigg\{ \theta_J \left[3(\tau_{J+1} - \tau_J)^2 + \xi_0^{(i)} (\tau_{J+1} - \tau_J)^2 \right] + I_{(J \geq 2)} \sum_{j=1}^{J-1} \theta_j \left[3(\tau_{J+1} - \tau_j)^2 + 3\xi_1^{(i)} (\tau_J - \tau_j)^2 + \xi_0^{(i)} \frac{(\tau_{J+1} - \tau_j)^3 - (\tau_J - \tau_j)^3}{\tau_{J+1} - \tau_J} \right] \Bigg\} \leq 0 \quad (2.16)$$

where $i = 1, 2, \ldots, Q$ and $I_{(J \geq 2)} = 1$ if $J \geq 2$ and 0 otherwise. Further we require that the derivatives before the last knot be non negative. Based on our model (2.14) we have $H'(\tau_1) = 0$. We also require $H'(\tau_j) \geq 0$, for $j = 2, \ldots, K-1$ which yields the following $(K-2)$ constraints

$$-\sum_{j=1}^{J-1} 3\theta_j (\tau_J - \tau_j)^2 \leq 0, \quad J = 2, \ldots, K-1 \quad (2.17)$$

The strict inequality of the derivative at the last knot yields the additional constraint

$$-\sum_{j=1}^{K-1} 3\theta_j (\tau_K - \tau_j)^2 < 0. \quad (2.18)$$

We can add an arbitrary small value $u > 0$ on the left of (2.18) to achieve $u - \sum_{j=1}^{K-1} 3\theta_j (\tau_K - \tau_j)^2 \leq 0$.
Note that the inequalities $a_j \geq 0$ and $b_j \geq 0$, $\forall j = 1, \ldots, K-1$ are trivially satisfied since $\tau_1 < \tau_2 < \ldots < \tau_K$ and due to the restrictions (2.17) and (2.18).

The final equality restriction for θ_{K-1} which is stated in (4). Thus, there is a total of $Q(K-1) + (K-2) + 1 + 1 = Q(K-1) + K$ constraints, consisting of $Q(K-1) + K - 1$ inequalities and one equality. Alternatively we can consider $Q(K-1) + K + 1$ inequality

constraints, since the equality can be written as two inequalities (as in Liew (1976) where the case of an untruncated covariance matrix of the parameters $\boldsymbol{\theta}$ is studied), and so finally the constraints in (2.16), (2.17) and (2.18) and the constraint for θ_{K-1} can be written as

$$\mathbf{A}[\theta_1, \theta_2, \ldots, \theta_{K-1}]' \leq \mathbf{0}.$$

where matrix $\mathbf{A}$ has $n_c = K - 1$ columns and $n_r = Q(K-1) + K + 1$ rows.

We fit model (2.14), (2.15) to the Kaplan Meier based cumulative hazard estimated function $\hat{H}^{KM}$, by minimizing their distance under the constraints (2.16), (2.17), (2.18), i.e. minimizing the function

$$\Psi(\boldsymbol{\theta}) = \sum_i \left(H(T_i|D_i = 1) - \hat{H}^{KM}(T_i|D_i = 1) \right)^2,$$

subject to the inequality constraints of the form $\mathbf{A}[\theta_1, \theta_2, \ldots, \theta_{K-1}]' \leq \mathbf{0}$. The lagrange function is then of the form

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \Psi(\boldsymbol{\theta}) + \boldsymbol{\lambda}\boldsymbol{\alpha},$$

where $\boldsymbol{\lambda}$ is the row vector of the lagrange multipliers and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_r}]$ with α_i be the $i-$th constraint. Note that the problem stated is always convex, and the global minimum can be found through standard algorithms. The case of using the entire region of monotonicity, $\mathcal{M}$, would be computationally cumbersome because non-linear constraints would be applied for each of the subintervals that are defined by the knots. In effect there would be no guarantee of convergence during the minimization of least squares, and very good initial values satisfying the initial constraints would be necessary. After fitting model (2.14), (2.15) under the appropriate constraints we obtain the smooth estimate, $\hat{H}(x)$, of the cumulative hazard function. The proposed estimator of the survival function $S(x)$, is

$$\hat{S}(x) = exp(-\hat{H}(x)),$$

with $\hat{H}(x) = H(x; \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the constrained least squares estimated vector of parameters. For the derivation of confidence intervals of $\hat{S}(x)$ we use the percentile bootstrap technique.

The HCNS approach can be further extended to accommodate other fully observed covariates through the semi-parametric Cox model. Consider the setting where measurements are taken on p additional covariates, $Z_1, Z_2, \ldots, Z_p$, not subject to censoring, and thus the available data are $\{T_i, D_i, Z_{i1}, \ldots, Z_{ip}\} = \{T_i, D_i, \boldsymbol{Z}_i\}$, $i = 1, \ldots, n$. It is well known that under the Cox model which is of the form

$$S(x|\mathbf{Z}) = S_0(x)^{exp(\boldsymbol{\gamma}'\mathbf{Z})},$$

where $\boldsymbol{\gamma}$ is the parameter vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_p]'$ that relates the survival time to the covariates $Z_1, \ldots, Z_p$, one can derive a consistent estimator of survival for given values of the covariates.

We apply model (2.14), (2.15) to the baseline hazard step function derived by the Cox model, to obtain the constrained natural spline estimator of the baseline cumulative hazard $(\hat{H}_0)$. Thus the corresponding estimator of the baseline survival function is

$$\hat{S}_0 = exp(-\hat{H}_0).$$

Given the profile of a subject, the estimated survival function is then given by

$$\hat{S}(x|\mathbf{Z}) = (\hat{S}_0)^{exp(\hat{\gamma}'\mathbf{Z})},$$

where $\hat{\gamma}$ is the estimated parameter vector derived by fitting the Cox model. Conditional logspline estimation has been suggested by Mâsse and Truong (1995). However, their method is not generalized to accommodate censoring. Other approaches are studied in the literature such as the HEFT model (for the unconditional hazard function) and the HARE model (in the case of hazard regression) presented in Kooperberg et al. (1995). An appealing feature of these models is that the proportional hazards model is included as a special case. However, these techniques are computationally cumbersome due to the numerous numerical integrations required for cubic or quadratic splines, and only the case of fitting a linear spline is addressed by the authors.

The use of medians are also used in survival analysis to summarize survival data and a kernel based approach in the presence of a covariate is taken in Beran (1981) as well as in Doksum and Yandell (1982) and Gentleman and Crowley (1991). At the covariate value, weights are computed to construct a weighted Kaplan Meier survival curve, and thus the median. In order to produce a smooth curve Wright and Bowman (1997) propose a nonparametric regression procedure to the data representing the corners of the steps (see also Bowman and Azzalini (1997)). The data used in constructing the Kaplan Meier are extracted from a 'window' of the data, centered at a value of the covariate, from which the local percentile of interest is computed. Repeating the procedure for several values of the covariate, the dependence of the percentile of interest and covariate can be explored. However, as the number of the covariates increases this approach is computationally infeasible.

Our approach is different from the one presented in Herndon and Harrell (1995). They consider a Cox model with a restricted (natural) spline being the baseline hazard. Under their approach all parameters are simultaneously estimated, in contrast to the two stage approach taken in this article. In our approach, once the Cox model is fitted, we do not consider maximizing a likelihood function. Instead, we smooth over the Kaplan Meier step function using a squared distance criterion. In many situations this may not be optimal compared to a likelihood based approach. However, our method, in the first stage, relies on the standard procedure to obtain the regression parameter estimates and during the second stage deals with a restricted least squares problem with linear restrictions. Thus, our method is computationally straightforward whereas likelihood based methods usually involve nonlinear functions and the convergence of any optimization routine is not guaranteed.

Our method can be easily adjusted to accommodate left censoring. However, when dealing with interval censoring in the presence of covariates a simultaneous estimation method of regression parameters and the baseline cumulative hazard function seems inevitable. For this setting we refer the reader to Zhang et al. (2010). They consider a spline-based semiparametric maximum likelihood approach. They approximate the baseline cumulative hazard using monotone $B-$splines and extend the Rosen algorithm to derive maximum likelihood estimates. Their approach is computationally intensive due to the nature of estimation for interval censored data.

In Appendix B we provide a description of a user friendly software written with `MATLAB` that is built to apply the HCNS method. The user can select the number of knots,

where to place them as well as the degree of approximation of the region of monotonicity. There is also an `auto` option for the knot placement where multiple knot placement schemes are tested and the one that minimizes the distance from the corners of the Kaplan Meier is finally chosen. For more details see the Appendix B. In Appendix C we provide some simulation studies to evaluate the HCNS method based on the knot schemes discussed in this chapter. We compare the HCNS approach with the Kaplan Meier estimator, the logspline method, as well as with the restricted cubic spline approach in the case of additional covariates. The comparisons are made with respect to the mean integrated squared error, as well as the obtained coverage.

2.4.2 Using the HCNS model for the censored covariate

Assume again, that we are in the setting of estimating the parameters of a generalized linear regression model when only a single censored covariate is available that might be censored. Our goal is to non parametrically estimate the conditional expectations $E(X_i|T_i,\Delta_i)$ and $E(X_i^2|T_i,\Delta_i)$ that appear in the optimal estimating function discussed in Section 2.2.1.

$$\sum_{i=1}^{n}\left[\frac{y_i-E_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}}{\tau^2 E_{X_i|T_i,\Delta_i}\left\{v(g^{-1}(\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}))\right\}+Var_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_\mathbf{i}'\boldsymbol{\beta})\right\}}E_{X_i|T_i,\Delta_i}\left\{\frac{d(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\right\}\right].$$

Due to the fact that the spline (2.14) is linearly extrapolated beyond the last knot and the derivative at the last knot is forced to be positive, it is obvious that under the above spline modeling approach the second moment $E(X^2)=\int_0^\infty x^2dF(x)$ is always finite, which in turn yields $V(X)<\infty$.

In the case where the optimal estimating function above is employed and the HCNS approach is used for the estimation of the covariate cumulative distribution, then parameter estimation is done with mild assumptions for the distributions of the response and the covariate. Furthermore, the approach is computationally stable, since the optimization problem involved is convex and can be handled by standard software. In effect, convergence is guaranteed and no initial values in fitting the spline model are required.

In the case of count data and when the $log-$link function is used then, again, under this spline formulation it can be shown that $M_X(2\beta_1)<\infty$. As already mentioned, in the case of binary data, the required condition $V(Y)<\infty$ is trivially satisfied.

The conditional expectations that appear in

$$\mathbf{G}^* = \sum_{i=1}^{n}\left[\frac{\left\{y_i-E_{X_i|T_i,\Delta_i,Z_i}(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))\right\}\left\{E_{X_i|T_i,\Delta_i,Z_i}\left(\frac{d(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))}{d\boldsymbol{\beta}}\right)\right\}}{\tau^2 E_{X_i|T_i,\Delta_i,Z_i}\left\{v(g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta}))\right\}+Var_{X_i|T_i,\Delta_i,Z_i}\left\{g^{-1}([\mathbf{x}_i^{'},z_i]\boldsymbol{\beta})\right\}}\right]$$

could be evaluated if a model that would relate the censored time variable with the other observed covariate was available. Assume now that the fully observed covariates, let $Z_1,Z_2,\ldots....,Z_p$, were available. The information they may carry regarding the censored variable must be taken into account and thus the approach is divided in two stages:

- Stage 1: Model the censored variable using $Z_1,Z_2,\ldots....,Z_p$ as covariates.

- Stage 2: Follow the estimating function approach to derive estimations of the parameters $\boldsymbol{\beta}$ of the desired GLM, using the model of the previous stage to calculate all conditional expectations required.

Here, we explore the use of the well known semi parametric approach of the Cox model to regress the censored variable on the fully observed covariates $Z_1, Z_2, \ldots ..., Z_q$. We assume the proportional hazards model of the form:

$$H(x|\mathbf{Z}) = H_0(x)exp(\gamma_1 Z_1 + \gamma_2 Z_2 + \cdots + \gamma_q Z_q), \tag{2.19}$$

where $F_0(x)$ is the baseline cumulative hazard function which is completely unspecified and the parameter vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_q]$ is the parameter vector that relates the possibly censored variable to the covariate matrix $Z = [Z_1, Z_2, \ldots, Z_q]$ in which each column is one fully observed covariate. After fitting model (2.19), we obtain the Cox based estimates, $\hat{\boldsymbol{\gamma}}$, as well as the baseline step cumulative hazard estimate $\hat{S}_0$. Model 2.14 can then be fitted to the the corners, at which events occur, of this step estimator in the same way as discussed in the previous section where only a single covariate was available. Thus, one can derive the CNS estimator for any desirable profile of a subject by

$$\hat{H}^{cns}(x|\mathbf{Z}) = \hat{H}_0^{cns}(x)exp(\hat{\gamma}_1 Z_1 + \hat{\gamma}_2 Z_2 + \cdots + \hat{\gamma}_q Z_q), \tag{2.20}$$

with the corresponding survival estimate being equal to

$$\hat{S}^{cns}(x|\mathbf{Z}) = exp(-\hat{H}^{cns}(x|\mathbf{Z})). \tag{2.21}$$

All conditional expectations required for implementation of the estimating function approach can be based on (2.21).

The approach presented above accommodates the additional covariates under the assumptions of proportional hazards. This assumption, however, may no be justified from the data at hand and a more general approach would be desirable. The Cox model can be generalized to relax the assumption of proportionality by using time functions as interactions with the observed covariates. For an overview of this approach see Klein and Moeschberger (1985). Assume for simplicity that only one observed covariate, let Z, is available, then under this formulation the Cox model takes the form

$$H(x|Z) = H_0(x)exp(\gamma_1 Z + \gamma_2 Z \times \eta(x)), \tag{2.22}$$

where $\eta(x)$ is a time function. Common choices are $\eta(x) = x$, $\eta(x) = \sqrt{x}$, $\eta(x) = log(x)$. Again, after fitting model (2.22), we obtain the step estimator of the baseline cumulative hazard, $\hat{H}_0(x)$, and then the implementation of the presented approach is straightforward following the discussion provided for the proportional hazards model case.

2.4.2.1 Simulation Studies

Here, we consider some simulation studies to evaluate the performance of the estimating function approach when the proposed spline technique is used for the censored covariate. We generated the censored covariate from a Weibull(3,2) distribution. For the censoring mechanism we considered a cutpoint due to which half of the expected censoring level would occur. The censoring variable was taken to be exponentially distributed with a proper value parameter so as to achieve expected total levels of censoring of 30% and 70%. (Half of the censoring are due to the censoring variable, and half due to the cutpoint (end of study)).

Linear Case

The sample sizes were chosen to be $n = 100$ and $n = 300$, while the model that generated the data was a simple linear model of the form $Y = 5 + \beta_1 X + \epsilon$. The slope values were chosen to result a correlation coefficient equal to $\rho(X, Y) = 0.2, 0.3, 0.5, 0.8$ and ϵ was generated from a standard normal distribution. The asymptotic confidence intervals are computed, and is some cases (where the desired coverage is not achieved) we also provide the bootstrap based ones for comparison purposes. The latter, as we observe, seem to have nice coverage properties even in the case of high correlation, high censoring level and small sample sizes, which is not always the case for the asymptotic one.

The simulation results for the linear case with a single censored covariate are presented in Table A4. We observe that for slope values that yield correlation coefficient equal to 0.2, 0.3, and 0.5 the presented approach clearly outperforms the CC approach in terms of MSE, in all cases (i.e. in all sample sizes and censoring levels). The CC approach seem to yield better results when the correlation coefficient is high (=0.8). We also observe that the asymptotic confidence intervals based on the unadjusted variance does not provide nice coverage properties for the QS approach as the slope and the censoring level increases. This can be circumvented with the use of the percentile bootstrap. In this table, results for the computationally simpler Unweighted approach are also presented. We observe minor differences from the QS approach in terms of MSE, Bias and SE and thus one might be tempted to use the Unweighted approach. However, with modern computer technology the computational time of both approaches is not an issue.

We also looked at a case where an additional fully observed covariate, Z, is present. We generated the censored values from a Cox model in which the baseline density was taken to be a Weibull(2,3) and the true value of the coefficient, γ, of the covariate Z, was set to 2. The response variable was then generated by the linear model $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, where the true values of the parameters of interest are $\beta_0 = 5$, $\beta_1 = 0.5$, $\beta_2 = 1$, and $\epsilon \sim N(0, 1)$. Following the two stage approach the parameter γ was estimated by simply using a Cox model. The simulation results are presented in Table (2.5). We observe that the proposed approach outperforms the CC method an all cases yielding narrower asymptotic confidence intervals.

Binary data

In the case of binary data we set the value of the intercept equal to $\beta_0 = log(9)$, which

in turn yields $P(Y = 1|X = 0)$. We looked at the following slope values: $\beta_1 = -2, -3, -4$. The sample size is set to $n = 300$ with expected levels of censoring 30% and 70% using the same censoring mechanism as in the linear case. The presented approach outperformed the CC approach in all cases. The results are presented in Table 2.6. We note here that we do not present results for the Unweighted method since the Fisher Scoring algorithm needed significantly more iterations to converge, due to the absence of a weight matrix, than the QS approach that is based on the optimal estimating functions. Unlike the linear case, in which the Unweighted method involves a simple least squares problem for the imputed data set, in the binary data setting the time the Unweihted method needs to converge may be essentially greater than the corresponding time of the QS approach. Thus, there is no need for attempting to use the suboptimal Unweighted method in such a case.

Count data

For the count data we considered the same censoring mechanism as in the previous settings and the sample size of $n = 300$. The true value of the intercept is set to $\beta_0 = 1$ and the slope values are set equal to $\beta_1 = -1$ and $\beta_1 = -1/3$. The overdispersion parameter τ is set equal to 2 and was estimated by the well known moment estimator (see McCullogh 2001) using the CC approach. In this setting we also observe that the presented approach outperforms the CC analysis in all cases (see Table 2.7).

2.5 Application

We applied our methods to data from a double-blinded randomized placebo controlled clinical trial of the drug D-penicillamine (DPCA) used for treatment of primary biliary cirrhosis (PBC). The trial was conducted at the Mayo Clinic during 1974-1984 (Fleming and Harrington (1990)) and involved 312 subjects. In the analysis of our results, we included an additional 112 subjects that did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of these subjects were lost to follow up, thus the available sample size was 418. The study established that DPCA is not effective for treatment. The data have been used to create a commonly used clinical prediction model which is based on the linear predictor of the Cox model. This predictor includes edema, log(bilirubin), age, log(prothrombin time) and albumin. However, Krzeski et al. (2003) question the applicability of the 'Mayo Model' and argue that selecting a prognostic variable, such as serum bilirubin, as a marker is more appropriate. We looked at a subgroup of the sample in this clinical trial, consisting of women with no edema (n=318). Our model considered log(bilirubin) as a marker (Y), versus survival time (X), adjusted for age (z). The plot of $log(bilirubin)$ versus time is presented in Figure (2.3). Initially, we considered an AFT model of X on z. We fitted both a GG and a Weibull regression AFT model. The likelihood ratio test suggested that the Weibull regression model is adequate (LR=0.3459, p-value=0.5564). However, we performed analyses using both the Weibull and the GG regression models.

First, we considered the simple model

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 X + \alpha_2 z + \epsilon \\ log(X|z) &= \xi_0 + \xi_1 z + \sigma u. \end{aligned} \tag{2.23}$$

Table 2.4: Simulation results for 1000 repetitions for the linear case. Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from $N(0,1)$. The distribution of the censored covariate was estimated using the $HCNS$ method.

			β_0					β_1				
	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
$n = 300$:												
		CC	-0.0039	0.1672	0.0280	0.6722	0.9570	0.0018	0.0729	0.0053	0.2943	0.9600
	0.2	QS	0.0190	0.1301	0.0173	0.5074	0.9510	-0.0018	0.0443	0.0021	0.1674	0.9230
		$Unweighted$	0.0193	0.1300	0.0173	0.5063	0.9520	-0.0110	0.0443	0.0021	0.1662	0.9280
		CC	0.0078	0.1724	0.0298	0.6741	0.9440	-0.0023	0.0758	0.0057	0.2955	0.9460
	0.3	QS	0.0367	0.1341	0.0193	0.5092	0.9270	-0.0186	0.0472	0.0026	0.1686	0.8950
30%		$Unweighted$	0.0373	0.1339	0.0193	0.5047	0.9210	-0.0189	0.0472	0.0026	0.1651	0.8880
Censoring		CC	0.0008	0.1675	0.0281	0.6713	0.9590	0.0012	0.0735	0.0054	0.2942	0.9530
	0.5	QS	0.0617	0.1452	0.0249	0.5173	0.9030	-0.0321	0.0549	0.0040	0.1754	0.8270
		$Unweighted$	0.0664	0.1464	0.0258	0.5051	0.8820	-0.0344	0.0588	0.0043	0.1654	0.7860
		CC	0.0027	0.1694	0.0287	0.6717	0.9570	-0.0012	0.0742	0.0055	0.2942	0.9590
	0.8	QS	0.1006	0.1810	0.0429	0.5547	0.8360	-0.0549	0.0786	0.0092	0.2059	0.7380
		$Unweighted$	0.1413	0.2086	0.0635	0.5058	0.7140	-0.0762	0.0956	0.0149	0.1660	0.5220
		CC	0.0067	0.2829	0.0801	1.1155	0.9510	-0.0071	0.2027	0.0411	0.7853	0.9450
	0.2	QS	0.0476	0.1521	0.0254	0.5853	0.9240	-0.0403	0.0490	0.0040	0.1698	0.7540
		$Unweighted$	0.0476	0.1520	0.0254	0.5867	0.9320	-0.0404	0.0489	0.0040	0.1687	0.7520
		CC	0.0070	0.2815	0.0813	1.1215	0.9490	-0.0033	0.1981	0.0392	0.7887	0.9450
	0.3	QS	0.0772	0.1617	0.0321	0.5924	0.8970	-0.0608	0.0618	0.0075	0.1742	0.6110
70%		$Unweighted$	0.0775	0.1614	0.0321	0.5890	0.8960	-0.0609	0.0617	0.0075	0.1702	0.6020
Censoring		CC	-0.0034	0.2800	0.0784	1.1131	0.9570	0.0048	0.1990	0.0396	0.7832	0.9480
	0.5	QS	0.1386	0.1742	0.0496	0.5989	0.7970	-0.1099	0.0870	0.0196	0.1835	0.4230
		$Unweighted$	0.1411	0.1736	0.0501	0.5894	0.7970	-0.1111	0.0869	0.0199	0.1712	0.3920
		CC	-0.0051	0.2752	0.0758	1.1179	0.9570	0.0051	0.1922	0.0369	0.7858	0.9580
	0.8	QS	0.3070	0.2537	0.1587	0.6280	0.5030	-0.2500	0.1707	0.0917	0.2266	0.2390
		$Unweighted$	0.3229	0.2566	0.1701	0.5891	0.4580	-0.0026	0.1729	0.0975	0.1708	0.1670
$n = 100$:												
		CC	-0.0199	0.2900	0.0845	1.1835	0.9520	0.0090	0.1271	0.0162	0.5207	0.9600
	0.2	QS	0.0254	0.2316	0.0543	0.8618	0.9250	-0.0156	0.0780	0.0063	0.2800	0.8970
		$Unweighted$	0.0258	0.2315	0.0543	0.8606	0.9320	-0.0158	0.0779	0.0063	0.2776	0.8990
		CC	0.0025	0.3021	0.0913	1.1871	0.9500	0.0019	0.1328	0.0176	0.5210	0.9410
	0.3	QS	-0.0220	0.0830	0.0074	0.8725	0.9240	0.0475	0.2374	0.0586	0.2865	0.8850
30%		$Unweighted$	0.0485	0.2372	0.0586	0.8677	0.9220	-0.0225	0.0830	0.0074	0.2809	0.8770
Censoring		CC	0.0046	0.2932	0.0862	1.1875	0.9590	-0.0007	0.1273	0.0162	0.5199	0.9550
	0.5	QS	0.0838	0.2582	0.0737	0.8891	0.8840	-0.0429	0.0987	0.0116	0.2995	0.8110
		$Unweighted$	0.0902	0.2591	0.0752	0.8696	0.8750	-0.0460	0.0999	0.0121	0.2815	0.7850
		CC	0.0068	0.3003	0.0902	1.1858	0.9540	-0.0026	0.1304	0.0170	0.5196	0.9560
	0.8	QS	0.1406	0.3406	0.1358	0.9576	0.8560	-0.0749	0.1516	0.0286	0.3538	0.7950
		$Unweighted$	0.1993	0.3864	0.1890	0.8706	0.7620	-0.1056	0.1783	0.0429	0.2824	0.6140
		CC	-0.0075	0.4977	0.2477	2.0308	0.9510	0.0002	0.3519	0.1238	1.4331	0.9580
	0.2	QS	0.0391	0.2659	0.0722	1.0214	0.9310	-0.0381	0.0895	0.0095	0.2993	0.7960
		$Unweighted$	0.0393	0.2657	0.0722	1.0259	0.9380	-0.0381	0.0895	0.0095	0.2981	0.7980
		CC	0.0062	0.5192	0.2696	2.0331	0.9560	-0.0030	0.3627	0.1316	1.4288	0.9510
	0.3	QS	0.0756	0.2810	0.0847	1.0228	0.9040	-0.0573	0.1086	0.0151	0.3028	0.7000
70%		$Unweighted$	0.0761	0.2801	0.0843	1.0297	0.9150	-0.0575	0.1083	0.0150	0.2994	0.7070
Censoring		CC	0.0289	0.4999	0.2508	2.0411	0.9470	-0.0277	0.3514	0.1240	1.4365	0.9570
	0.5	QS	0.1443	0.3003	0.1110	1.0381	0.8540	-0.1138	0.1435	0.0335	0.3147	0.5640
		$Unweighted$	0.1470	0.2999	0.1116	1.0267	0.8530	-0.1147	0.1438	0.0388	0.2955	0.5300
		CC	-0.0030	0.5184	0.2687	2.0476	0.9500	0.0088	0.3660	0.1341	1.4426	0.9500
	0.8	QS	0.3175	0.4190	0.2764	1.0940	0.6630	-0.2524	0.2806	0.1425	0.3900	0.4120
		$Unweighted$	0.3344	0.4165	0.2853	1.0298	0.6340	-0.2611	0.2817	0.1476	0.2997	0.3360

Table 2.5: Simulation when an additional fixed covariate is present. The true values of the parameters β_0, β_1 and β_2 are 5, 0.5 and 1 respectively. The time variable was generated from a Cox model in which the baseline density (for $Z = 0$) was taken to be a Weibull(2,3).

	Method	Parameter	Bias	SE	MSE	Width	Asympt. Cover.
		β_0	-0.0093	0.2389	0.0572	0.9242	0.9441
	CC	β_1	0.0056	0.1300	0.0169	0.5103	0.9500
30%		β_2	0.0069	0.2467	0.0609	0.9653	0.9500
Censoring		β_0	-0.0051	0.2315	0.0536	0.8872	0.9381
	QS	β_1	0.0032	0.1227	0.0151	0.4763	0.9471
		β_2	-0.0005	0.2263	0.0512	0.8725	0.9491
		β_0	-0.0141	0.3380	0.1144	1.3424	0.9450
	CC	β_1	0.0166	0.2108	0.0447	0.8504	0.9520
70%		β_2	0.0013	0.4322	0.1171	1.3351	0.9540
Censoring		β_0	0.0125	0.3087	0.0954	1.1860	0.9430
	QS	β_1	-0.0035	0.1757	0.0309	0.6647	0.9340
		β_2	-0.0164	0.2550	0.0653	0.9641	0.9390

Table 2.6: Simulation results of 1000 repetitions in the case of binary data. Sample size is $n = 300$, with 70% and 30% censoring, half of which is due to the cutpoint (end of study), and the other half due to random censoring. The logit link function was used. The real value of the intercept is $log(9)$ so that $P(Y = 1|X = 0) = 0.9$.

			β_0					β_1				
	β_1	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
	-2	CC	0.0745	0.5345	0.2913	2.0374	0.9590	-0.0604	0.3427	0.1211	1.2845	0.9520
		QS	0.0630	0.5194	0.2738	1.9818	0.9550	-0.0519	0.3284	0.1105	1.2262	0.9440
30%	-3	CC	0.1512	0.7350	0.5631	2.7753	0.9600	-0.1737	0.6548	0.4589	2.4453	0.9580
Censoring		QS	0.1432	0.7231	0.5434	2.7375	0.9610	-0.1630	0.6360	0.4310	2.3860	0.9550
	-4	CC	0.2496	1.0366	1.1367	3.6842	0.9580	-0.3720	1.2126	1.6087	4.2440	0.9520
		QS	0.2426	1.0319	1.1236	3.6528	0.9590	-0.3651	1.1952	1.5618	4.1730	0.9540
	-2	CC	0.0968	0.7344	0.5487	2.7572	0.9520	-0.0808	0.5517	0.3108	2.0635	0.9560
		QS	-0.0257	0.5900	0.3488	2.2090	0.9330	0.0278	0.3981	0.1593	1.4114	0.9180
70%	-3	CC	0.1697	0.8597	0.7679	3.2288	0.9590	-0.2026	0.8205	0.7143	3.0411	0.9550
Censoring		QS	0.1345	0.7999	0.6579	2.9524	0.9460	-0.1593	0.7211	0.5454	2.6190	0.9530
	-4	CC	0.2746	1.1970	1.5083	4.0398	0.9620	-0.4173	1.4667	2.3254	4.8295	0.9630
		QS	0.2549	1.1500	1.3875	3.8829	0.9590	-0.3937	1.3763	2.0492	4.5006	0.9530

Table 2.7: Simulation results of 1000 repetitions in the case of count data. Sample size is $n = 300$, with 70% and 30% censoring, half of which is due to the cutpoint (end of study), and the other half due to random censoring. The log link function was used. The real value of the intercept is 1 and the values of the slope were set to -1 and -1/3.

			β_0					β_1				
	β_1	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
	-1	CC	-0.0317	0.3852	0.1494	1.4247	0.9340	-0.0115	0.2503	0.0628	0.9287	0.9380
		QS	-0.0345	0.3700	0.1381	1.3456	0.9250	-0.0068	0.2286	0.0523	0.8259	0.9250
30%	-1/3	CC	-0.0130	0.2583	0.0669	1.0016	0.9450	-0.0011	0.1295	0.0168	0.4952	0.9440
Censoring		QS	-0.0222	0.2164	0.0473	0.8523	0.9350	0.0090	0.0919	0.0085	0.3559	0.9250
	-1	CC	-0.0258	0.5031	0.2538	1.9022	0.9280	-0.0247	0.4365	0.1911	1.6478	0.9080
		QS	-0.0839	0.3963	0.1641	1.4788	0.8860	0.0518	0.2632	0.0719	0.9485	0.8550
70%	-1/3	CC	-0.0416	0.3981	0.1602	1.5102	0.9360	0.0136	0.2924	0.0857	1.1263	0.9440
Censoring		QS	-0.0294	0.2347	0.0559	0.9552	0.7190	0.0370	0.1065	0.0127	0.4161	0.6990

The proposed method based on O_F optimality yielded smaller standard errors for all three coefficients compared to the CC analysis. The results for this model are presented in Appendix A (Tables 6 and 7).

Next, we fitted a brokenline type model

$$Y = \beta_0 + \beta_1^{(1)}(X - \tau^*)_- + \beta_1^{(2)}(X - \tau^*)_+ + \beta_2 z + \epsilon \qquad (2.24)$$
$$log(X|z) = \xi_0 + \xi_1 z + \sigma u$$

where $\epsilon \sim (0, \tau^2)$, $x_+ = xI(x \geq 0)$, $x_- = xI(x < 0)$ and τ^* is some time point (changepoint). We obtained an estimate of the changepoint τ^* using the algorithm presented in Kuchenhoff (1997). Although the estimate obtained was $\hat{\tau^*} = 86.6$ we considered it fixed at 84 months (7 years) for illustrative purposes. We performed analyses using the following methods: CC, QS(Weib), Un(Weib), QS(GG) and Un(GG) and the results are presented in Table 2.8. In the same table we also include results that correspond to the optimal QS approach and the Unweighted approach when the HCNS technique is considered for the covariate (QS(HCNS) and Un(HCNS) respectively). Note that the estimate of the dispersion parameter $\hat{\tau} = 0.91$ was provided from the CC analysis. The $p-$value of the covariate of age in the underlying Cox model when using the HCNS approach equals to 0.0014, indicating that the age is a significant predictor for the time to event variable. We computed 95% CI's using both the asymptotic covariance matrix and the bootstrap. For each bootstrap sample we fitted the underlying AFT model for the methods QS(Weib), Un(Weib), QS(GG) and Un(GG). For the methods QS(HCNS) and Un(HCNS) we considered fitting the HCNS spline approach for each bootstrap sample.

To compare the brokenline type versus the linear model we tested $H_0 : \beta_1^{(1)} - \beta_1^{(2)} = 0$ using 1000 bootstrap samples. The 95% CI for the difference $\beta_1^{(1)} - \beta_1^{(2)}$ using QS(Weib) and QS(GG) were (-0.0234, -0.0065) and (-0.0230, -0.0054) respectively, indicating that model (2.24) is more appropriate than the simpler one in (2.23). The spline based QS(HCNS) yielded similar results for this difference as well, i.e. (-0.0233, -0.0042).

The results showed smaller standard errors for the estimates of our methods when compared to the CC analysis. The Unweighted method provided estimates similar to the QS method in both the linear and the brokenline type model. The asymptotic CI's were fairly similar to the ones obtained by the resampling technique. For both the QS and the Unweighted method a small increase in the SE's was observed as we moved from the Weibull to the GG model. When using the HCNS technique for the covariate we observe that we obtain similar results as when using the GG model.

The model in (2.24) reduces to the so called hockey stick model when $\beta_1^{(2)} = 0$. The hockey stick model belongs to the family of models presented in Cai et al. (2006), where they discuss a modeling approach to address the accuracy of a time dependent biomarker. Their approach considers sensitivity and specificity as functions of time between the measurement and the event. A suitably chosen distant time point (changepoint) is used to distinguish the control group from the cases. Subjects that experience the event prior to that time point are considered as cases. Under this modeling approach the marker values for the control group are independent of survival time, given the covariates in the model.

We note that when using the CC method we cannot reject the hypothesis that $\beta_1^{(2)} = 0$, that corresponds to the hockey stick model, whereas the same hypothesis is rejected when using our methods. A major difference here between ignoring the censoring (CC) or

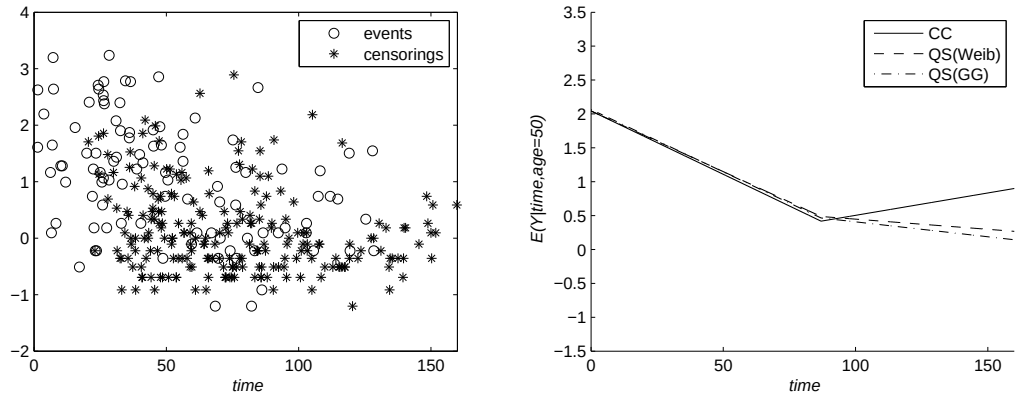

Figure 2.3: Left: Scatter plot for the pbc data of the marker log(bilirubin) vs. time to event for women with no edema (n=318). Right: Fitted brokenline models for the age of 50 based on the CC, the proposed method with a Weibull model for the covariate (QS(Weib)), and the proposed method with a GG model for the covariate (QS(GG)).

taking it into account by our methods concerns the interpretation and use of the resulting models. The CC analysis implies a model that can be used to evaluate the accuracy of time dependent biomarkers as discussed in Cai et al. (2006). On the other hand, our approach implies that the second linear segment (after the changepoint) has a negative slope. A clear framework does not yet exist for assessing the accuracy of the biomarker in this type of setting.

2.6 Discussion

In this chapter we study the problem of estimating the parameters of a generalized linear model when a covariate is censored. We propose a quasi score based method that requires no parametric assumptions for the distribution of the response and is computationally simple. One setting in which the proposed methodology may be applied, is when dealing with modeling a time dependent biomarker. There the marker value plays the role of the response and the time to event, which is subject to censoring, is a covariate (Cai et al. (2006)). The methods presented here differ from the Expected Estimating Equations (EEE) approach in Wang and Pepe (2000) and the regression calibration (RC) technique (see Carroll and Stefanski (1990)), which address measurement error in the covariates. Our approach can be thought of as a compromise between the EEE, where the expectation of the estimating equation given the data is considered, and the RC method, where $E(X_i|\Delta_i)$ is simply plugged in for censored covariate values. Apart from the method based on the optimal estimating function, we presented the computationally simpler Unweighted method.

Simulation studies showed good performance, for various scenarios. In some cases, the use of the percentile bootstrap technique is recommended in order to derive confidence intervals. In the linear case, the simulations showed fairly similar results between the suboptimal Unweighted method and the optimal one. In the case of binary and count data we presented simulation results only for the optimal method since for the

Table 2.8: Estimates of the coefficients of the brokenline type model for the PBC data.

Method	Parameter	Est.	Asympt. SE	Asympt. CI 95%		Bootstrap CI 95%	
CC	β_0	1.4778	0.5571	0.3720	2.5836	–	–
	$\beta_1^{(1)}$	-0.0187	0.0041	-0.0267	-0.0106	–	–
	$\beta_1^{(2)}$	0.0066	0.0104	-0.0141	0.0274	–	–
	β_2	-0.0212	0.0100	-0.0411	-0.0013	–	–
QS(Weib)	β_0	1.9418	0.4368	1.0856	2.7980	1.2003	2.6052
	$\beta_1^{(1)}$	-0.0179	0.0033	-0.0245	-0.0114	-0.0249	-0.0114
	$\beta_1^{(2)}$	-0.0030	0.0011	-0.0052	-0.0008	-0.0055	-0.0008
	β_2	-0.0291	0.0068	-0.0424	-0.0158	-0.0400	-0.0170
QS(GG)	β_0	1.9700	0.4577	1.0729	2.8671	1.2037	2.6320
	$\beta_1^{(1)}$	-0.0184	0.0033	-0.0249	-0.0118	-0.0249	-0.0115
	$\beta_1^{(2)}$	-0.0044	0.0018	-0.0079	-0.0009	-0.0076	-0.0008
	β_2	-0.0301	0.0070	-0.0439	-0.0163	-0.0405	-0.0179
QS(HCNS)	β_0	1.9038	0.4379	1.0455	2.7621	1.2144	2.5911
	$\beta_1^{(1)}$	-0.0181	0.0033	-0.0246	-0.0116	-0.0251	-0.0110
	$\beta_1^{(2)}$	-0.0039	0.0016	-0.0070	-0.0009	-0.0090	-0.0011
	β_2	-0.0286	0.0068	-0.0419	-0.0152	-0.0393	-0.0171
Un(Weib)	β_0	1.9864	0.4380	1.1279	2.8450	1.2150	2.6662
	$\beta_1^{(1)}$	-0.0180	0.0033	-0.0245	-0.0114	-0.0249	-0.0115
	$\beta_1^{(2)}$	-0.0030	0.0011	-0.0052	-0.0007	-0.0056	-0.0009
	β_2	-0.0299	0.0068	-0.0433	-0.0166	-0.0418	-0.0179
Un(GG)	β_0	1.9996	0.4580	1.1018	2.8973	1.2349	2.7065
	$\beta_1^{(1)}$	-0.0184	0.0033	-0.0249	-0.0118	-0.0250	-0.0116
	$\beta_1^{(2)}$	-0.0044	0.0018	-0.0079	-0.0009	-0.0076	-0.0008
	β_2	-0.0307	0.0070	-0.0445	-0.0168	-0.0416	-0.0183
Un(HCNS)	β_0	1.9530	0.4379	1.0948	2.8113	1.1865	2.6081
	$\beta_1^{(1)}$	-0.0182	0.0033	-0.0247	-0.0117	-0.0248	-0.0114
	$\beta_1^{(2)}$	-0.0039	0.0016	-0.0069	-0.0009	-0.0088	-0.0012
	β_2	-0.0295	0.0068	-0.0429	-0.0162	-0.0407	-0.0171

Unweighted method we observed very slow convergence of the Fisher Scoring algorithm.

The extension of the proposed method to accommodate interval censoring is straightforward, when the data include a number of fully observed values of the covariate. When all data are interval censored, as in the setting discussed by Gomez et al. (2003), some aspects need further study, in particular regarding the estimation of the dispersion parameter τ. Due to the presence of censoring a kind of a parametric model for the covariate is needed for the method to be applicable. The presented approach can be straightforwardly generalized in the case of left censoring. In this case the difference would be that the expectations required would be conditioned on $X_i < t_i$ instead of $X_i > t_i$. Left truncation could be also accommodated if the parametric model assumed for the regressor is properly adjusted.

In this chapter, we explored fitting the flexible generalized gamma AFT model to the time to event covariate as well as a new spline based approach. We investigated the fit of a natural non-decreasing spline to smooth the Kaplan Meier based cumulative hazard of the censored covariate. The constraints derived through a linear approximation of a non linear region that defines the necessary and sufficient condition of monotonicity. Thus, the spline fitting problem reduces to a restricted least squares one, with linear restrictions on the parameters. The problem is always convex and the method can be applied in the presence of small sample sizes and/or heavy censoring. In the case of additional covariates our method is easily generalizable, under the proportional hazard assumption, and computationally stable. This is due to the convex linear optimization that the proposed methodology is based on.

Another important generalization which will be discussed in the next chapter is the case when multiple measurements are taken repeatedly on each subject, until the time of event or censoring. This last setting would appear to be the ideal one in which to consider joint modeling for both, marker values and survival, an approach that is potentially more informative. A good source of references on relevant work is given in Lawless (2003). In particular the model with additional covariates presented in Section 2.2.3 is consistent with the modeling approach concerning marker processes discussed in Cox (1999).

2.7 Technical Notes

2.7.1 Proof of Theorem 1

First, note that

$$\boldsymbol{\mu}_i^c = E(Y_i|T_i,\Delta_i) = E_{X_i|T_i,\Delta_i}\left\{E(Y_i|T_i,\Delta_i,X_i)\right\} = E_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}$$

and

$$\begin{aligned} Var(Y_i|T_i,\Delta_i) &= E_{X_i|T_i,\Delta_i}\left\{Var(Y_i|T_i,\Delta_i,X_i)\right\} + Var_{X_i|T_i,\Delta_i}\left\{E(Y_i|T_i,\Delta_i,X_i)\right\} \\ &= \tau^2 E_{X_i|T_i,\Delta_i}\left\{v(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))\right\} + Var_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\right\}. \end{aligned}$$

The estimating functions are unbiased since

$$E\left\{\mathbf{A}(\boldsymbol{\beta})(\mathbf{Y}-\boldsymbol{\mu}^{\boldsymbol{c}})\right\} = E\left[E\left\{\mathbf{A}(\boldsymbol{\beta})(\mathbf{Y}-\boldsymbol{\mu}^{\boldsymbol{c}})|\mathbf{T},\boldsymbol{\Delta}\right\}\right] = E\left\{\mathbf{A}(\boldsymbol{\beta})E(\mathbf{Y}-\boldsymbol{\mu}^{\boldsymbol{c}}|\mathbf{T},\boldsymbol{\Delta})\right\} = \mathbf{0}.$$

We also have

$$\begin{aligned} E(\dot{\mathbf{G}}) &= E\left\{\dot{\mathbf{A}}(\boldsymbol{\beta})(\mathbf{I}_p \otimes (\mathbf{y}-\boldsymbol{\mu}^c)')' - \mathbf{A}(\boldsymbol{\beta})\dot{\boldsymbol{\mu}}^c\right\} \\ &= E\left[E\left\{\dot{\mathbf{A}}(\boldsymbol{\beta})(\mathbf{I}_p \otimes (\mathbf{y}-\boldsymbol{\mu}^c)')' - \mathbf{A}(\boldsymbol{\beta})\dot{\boldsymbol{\mu}}^c|\mathbf{T},\boldsymbol{\Delta}\right\}\right] \\ &= -E\left\{\mathbf{A}(\boldsymbol{\beta})\dot{\boldsymbol{\mu}}^c\right\}. \end{aligned}$$

If we denote the error vector by $\boldsymbol{e} = \mathbf{Y} - \boldsymbol{\mu}^c$ then we derive

$$\begin{aligned} E(\mathbf{G}\mathbf{G}^{*'}) &= E\left[E\left\{\mathbf{A}(\boldsymbol{\beta})\mathbf{e}\mathbf{e}'\mathbf{A}^{*'}(\boldsymbol{\beta})\right\}|\mathbf{T},\boldsymbol{\Delta}\right] \\ &= E\left\{\mathbf{A}(\boldsymbol{\beta})\mathbf{W}^{-1}\mathbf{A}^{*'}(\boldsymbol{\beta})\right\}. \end{aligned}$$

Thus, $(E\dot{\mathbf{G}})^{-1}E(\mathbf{G}\mathbf{G}^{*'})$ is a constant matrix if $\mathbf{A}^{*'}(\boldsymbol{\beta}) = \mathbf{A}^{*'} = \mathbf{W}\dot{\boldsymbol{\mu}}^c$. So the optimal estimating function is $\mathbf{G}^* = \dot{\boldsymbol{\mu}}^{c'}\mathbf{W}(\mathbf{Y}-\boldsymbol{\mu}^c)$ and can be written as

$$\mathbf{G}^* = \sum_{i=1}^{n}\left\{\left(\frac{y_i - E(Y_i|T_i,\Delta_i)}{Var(Y_i|T_i,\Delta_i)}\right)\left(\frac{dE(Y_i|T_i,\Delta_i)}{d\boldsymbol{\beta}}\right)\right\},$$

from which the result in (2.8) is straightforward. This completes the proof.

2.7.2 Optimal points of approximation for the region of monotonicity

We use Smith's algorithm (1970) to derive optimal polygons within the region of monotonicity. Let n be any even number greater or equal to 4. Consider the set $\varphi_{(n)}$ of incremented angles from 0 to 2π by $\frac{2\pi}{n}$. Thus, $\varphi_{(n)}$ is the set

$$\varphi_{(n)} = \left\{0, \frac{2\pi}{n}, \frac{4\pi}{n}, \frac{6\pi}{n}, \frac{8\pi}{n}, , \ldots, 2\pi\right\}.$$

Obviously $\varphi_{(n)} \subset \varphi_{(kn)}$, $\forall k$ integer. Smith's algorithm is based on the calculation of sines and cosines of the incremented quantity, φ, to obtain the optimal approximation (in terms of the inscribed area) of an ellipse, given a fixed number of points. By applying Smith's algorithm to derive the optimal $12-$gon within the ellipse under study, $\phi(a,b)$, we notice that the points (3,0), (0,3) and (3,3) are included in the set of the twelve points. Thus, these points will be included in every $12k-$gon, $k = 1, 2, \ldots$

For the optimal $12-$gon we also notice that three points need to be discarded since they are under the line $b-a+3=0$. Note that for every pair of angles in φ, an extra angle is placed between them if k is increased by 1. In effect, as we move from k to $k+1$, we get additional points, each one lying between two consecutive points obtained in the previous approximation. Thus, for the optimal $12k-$gon we should discard $4k-1$ points due to the fact that they are under the line $b-a+3=0$, and insert the point (0,0) instead. So finally, through Smith's algorithm, we can instantly derive the optimal inscribed polygon within the region of monotonicity, $\mathcal{M}$, defined by the $12k-(4k-1)+1 = 8k+2$ points of approximation.

2.7.3 Constraints for capturing region $\mathcal{A}$

We consider the derivation of the constraints given in (2.16). Suppose that we have 6 knots i.e. $K = 6$. For model (2.14) we can derive the following:

$$
\begin{aligned}
H(\tau_1) &= 0 \\
H(\tau_2) &= \theta_1(\tau_2 - \tau_1)^3 \\
H(\tau_3) &= \theta_1(\tau_3 - \tau_1)^3 + \theta_2(\tau_3 - \tau_2)^3 \\
H(\tau_4) &= \theta_1(\tau_4 - \tau_1)^3 + \theta_2(\tau_4 - \tau_2)^3 + \theta_3(\tau_4 - \tau_3)^3 \\
H(\tau_5) &= \theta_1(\tau_5 - \tau_1)^3 + \beta_2(\tau_5 - \tau_2)^3 + \beta_3(\tau_5 - \tau_3)^3 + \beta_4(\tau_5 - \tau_4)^3 \\
H(\tau_6) &= \theta_1(\tau_6 - \tau_1)^3 + \beta_2(\tau_6 - \tau_2)^3 + \beta_3(\tau_6 - \tau_3)^3 + \beta_4(\tau_6 - \tau_4)^3 + \theta_5(\tau_6 - \tau_5)^3
\end{aligned}
$$

Also for the derivatives of the model at the knots we have:

$$
\begin{aligned}
H'(\tau_1) &= 0 \\
H'(\tau_2) &= 3\theta_1(\tau_2 - \tau_1)^2 \\
H'(\tau_3) &= 3\theta_1(\tau_3 - \tau_1)^2 + 3\theta_2(\tau_3 - \tau_2)^2 \\
H'(\tau_4) &= 3\theta_1(\tau_4 - \tau_1)^2 + 3\theta_2(\tau_4 - \tau_2)^2 + 3\theta_3(\tau_4 - \tau_3)^2 \\
H'(\tau_5) &= 3\theta_1(\tau_5 - \tau_1)^2 + 3\theta_2(\tau_5 - \tau_2)^2 + 3\theta_3(\tau_5 - \tau_3)^2 + 3\theta_4(\tau_5 - \tau_4)^2 \\
H'(\tau_6) &= 3\theta_1(\tau_6 - \tau_1)^2 + 3\theta_2(\tau_6 - \tau_2)^2 + 3\theta_3(\tau_6 - \tau_3)^2 + 3\theta_4(\tau_6 - \tau_4)^2 + 3\theta_5(\tau_6 - \tau_5)^2
\end{aligned}
$$

Consider the region of approximation $\mathcal{A}$ that is defined by a decahexagon inscribed in region $\mathcal{M}$ (thus $Q = 14$). Each of the Q line segments will lead to Q line equations of the form:

$$g_i = g_i(a, b) = b + \xi_1^{(i)} a + \xi_0^{(i)} = 0, \quad i = 1, \ldots, Q$$

For a point (a_j, b_j) and following the notation of section 2.4.1.3 we require

$$v_i g_i(a_j, b_j) \leq 0, \quad i = 1, 2, \ldots, Q \tag{2.25}$$

Note that (2.25) can be written as

$$v_i \left[H'(\tau_{j+1}) + \xi_1^{(i)} H'(\tau_j) + \xi_0^{(i)} \frac{H(\tau_{j+1}) - H(\tau_j)}{\tau_{j+1} - \tau_j} \right] \leq 0 \tag{2.26}$$

Thus, for the interval $[\tau_1, \tau_2]$, using formula (2.26) we derive for the g_i:

$$g_i(a_j, b_j) = \theta_1(3(\tau_2 - \tau_1)^2 + \xi_0^{(i)}(\tau_2 - \tau_1)^2), \; i = 1, \ldots, 14.$$

For the interval $[\tau_2, \tau_3]$:

$$g_i(a_j, b_j) = \theta_1[3(\tau_3 - \tau_1)^2 + 3\xi_1^{(i)}(\tau_2 - \tau_1)^2 + \xi_0^{(i)} \frac{(\tau_3 - \tau_1)^3 - (\tau_2 - \tau_1)^3}{\tau_3 - \tau_2}] +$$

$$\theta_2[3(\tau_3 - \tau_2)^2 + \xi_0^{(i)}(\tau_3 - \tau_2)^2], \quad i = 1, ..., 14$$

For $[\tau_3, \tau_4]$

$$g_i(a_j,b_j) = \theta_1[3(\tau_4-\tau_1)^2 + 3\xi_1^{(i)}(\tau_3-\tau_1)^2 + \xi_0^{(i)}\frac{(\tau_4-\tau_1)^3-(\tau_3-\tau_1)^3}{\tau_4-\tau_3}]+$$
$$\theta_2[3(\tau_4-\tau_2)^2 + 3\xi_1^{(i)}(\tau_3-\tau_2)^2 + \xi_0^{(i)}\frac{(\tau_4-\tau_2)^3-(\tau_3-\tau_2)^3}{\tau_4-\tau_3}]+$$
$$\theta_3[3(\tau_4-\tau_3)^2 + \xi_0^{(i)}(\tau_4-\tau_3)^2] \quad i=1,...,14$$

For $[\tau_4,\tau_5]$

$$g_i(a_j,b_j) = \theta_1[3(\tau_5-\tau_1)^2 + 3\xi_1^{(i)}(\tau_4-\tau_1)^2 + \xi_0^{(i)}\frac{(\tau_5-\tau_1)^3-(\tau_4-\tau_1)^3}{\tau_5-\tau_4}]+$$
$$\theta_2[3(\tau_5-\tau_2)^2 + 3\xi_1^{(i)}(\tau_4-\tau_2)^2 + \xi_0^{(i)}\frac{(\tau_5-\tau_2)^3-(\tau_4-\tau_2)^3}{\tau_5-\tau_4}]+$$
$$\theta_3[3(\tau_5-\tau_3)^2 + 3\xi_1^{(i)}(\tau_4-\tau_3)^2 + \xi_0^{(i)}\frac{(\tau_5-\tau_3)^3-(\tau_4-\tau_3)^3}{\tau_5-\tau_4}]+$$
$$\theta_4[3(\tau_5-\tau_4)^2 + \xi_0^{(i)}(\tau_5-\tau_4)^2] \quad i=1,...,14$$

For $[\tau_5,\tau_6]$

$$\theta_1[3(\tau_6-\tau_1)^2 + 3\xi_1^{(i)}(\tau_5-\tau_1)^2 + \xi_0^{(i)}\frac{(\tau_6-\tau_1)^3-(\tau_5-\tau_1)^3}{\tau_6-\tau_5}]+$$
$$\theta_2[3(\tau_6-\tau_2)^2 + 3\xi_1^{(i)}(\tau_5-\tau_2)^2 + \xi_0^{(i)}\frac{(\tau_6-\tau_2)^3-(\tau_5-\tau_2)^3}{\tau_6-\tau_5}]+$$
$$\theta_3[3(\tau_6-\tau_3)^2 + 3\xi_1^{(i)}(\tau_5-\tau_3)^2 + \xi_0^{(i)}\frac{(\tau_6-\tau_3)^3-(\tau_5-\tau_3)^3}{\tau_6-\tau_5}]+$$
$$\theta_4[3(\tau_6-\tau_4)^2 + 3\xi_1^{(i)}(\tau_5-\tau_4)^2 + \xi_0^{(i)}\frac{(\tau_6-\tau_4)^3-(\tau_5-\tau_4)^3}{\tau_6-\tau_5}]+$$
$$\theta_5[3(\tau_6-\tau_5)^2 + \xi_0^{(i)}(\tau_6-\tau_5)^2] \quad i=1,...,14$$

Thus, it is easy to show that for an interval $[\tau_j,\tau_{j+1}]$ the constraints required to capture region $\mathcal{A}$ are of the form (2.16).

Chapter 3

Generalized Linear Models with a censored covariate for longitudinal data

In many longitudinal clinical studies interest lies in modeling marker values that are repeatedly taken on subjects. These marker values may depend on the time to event variable (among others). For example in prostate cancer studies we are interested in relating the PSA (prostate specific antigen) with the time to relapse. Studies regarding HIV (human immunodeficiency virus) patients may also attempt to investigate the relationship between CD4 cell count and time to death. Another example studied in Schluchter et al. (2002) that focused on cystic fibrosis patients, explores the association between pulmonary function and time to death. They use a standard joint model (also known as shared parameter model) which involves two stages. The first stage specifies the distributions of the subject specific characteristics (or random effects). The second stage involves assumptions regarding the distributions of the longitudinal measurements and the time to event variable given the subject specific effects. For detailed overviews of this modeling approach see Tsiatis and Davidian (2004), Yu et al. (2004). Under the joint model framework, Rizopoulos (2011) discusses estimation of survival probabilities as well as the assessment of the discriminatory capability of a longitudinal marker by appropriately defining sensitivity and specificity followed by ROC analysis.

In this Chapter we follow an estimating function approach considering marginal models. We present an extension of the method introduced in Chapter 2, under the longitudinal framework. When repeated measurements are taken on each subject correlation must be taken into account. In Chapter 2, our approach assumed parametric models for the censored covariate. Here, we relax strict parametric assumptions by employing monotone natural cubic splines for the survival function of the censored covariate. No assumptions are made for the parametric form of the joint distribution of the repeated measurements. We do not specify any parametric form of the conditional distribution of the response given the subject specific effects even though such a modeling approach may be more appropriate in some settings. In this Chapter our primary focus is to make inferences about the population means. We also discuss other computationally simpler approaches and comparisons are made via simulations. We contrast our approach with common linear mixed models and the joint modeling approach.

This Chapter is organized as follows: In Section 3.1 we briefly recall the joint modeling approach and we refer to the linear model. In Sections 3.2 and 3.3 we present the proposed approach and discuss some of the most commonly used link functions as well as the assumptions imposed for the censored covariate. In section 3.4 we discuss the way additional fully observed covariates can be accommodated under semiparametric assumptions. In Section 3.5 we present a real data application and we conclude with a discussion in which we contrast our approach with mixed models.

3.1 Joint Modeling Approach for the Linear Mixed Effect Model

Consider the marker measurements $Y_{ij} = Y_i(s_{ij})$ for subject i taken at the time points s_{ij}, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Denote with X_i the time to event variable and C_i the censoring variable. We only observe a censored version of the time to event variable, let $T_i = min(X_i, C_i)$. A status indicator $\Delta_i = I(X_i < C_i)$ is also available, taking the value 1 for an event and 0 otherwise. A usual choice it to employ a relative risk model to associate the true response $m_i(t)$ with the risk of an event (see also Rizopoulos (2011)).

$$h_i(t|\mathcal{M}_i(t)) = h_0(t)exp(\alpha m_i(t)), \tag{3.1}$$

where $\mathcal{M}_i(t) = \{m_i(u), 0 \leq u < t\}$ is the history of the true longitudinal process until time t and h_0 is the baseline hazard rate that may be completely unspecified. Even though additional covariates can be straightforwardly accommodated, we do not consider them here for simplicity. The linear mixed effect model assumed is of the form:

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= \mathbf{x}_i'(t)\boldsymbol{\beta} + \mathbf{z}_i^{*'}(t)\mathbf{b}_i + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \sigma^2), \end{aligned} \tag{3.2}$$

where $\boldsymbol{\beta}$, $\mathbf{b}_i$ are the vectors of fixed and random effects respectively, and $\mathbf{x}_i(t)$, $\mathbf{z}_i^*(t)$ are the row vectors of the corresponding design matrices for the fixed and random effects. The measurement error ϵ is assumed to be normally distributed. The random effects $\mathbf{b}_i$ are assumed to be independent of $\epsilon_i(t)$ and to follow a multivariate normal distribution with some covariance matrix and zero mean. Maximum likelihood estimators can then be derived by assuming a joint distribution for the available data T_i, δ_i, y_i. Under the assumption that the repeated measurements of the response are conditionally independent from the time to event variable, given the subject specific effects we have:

$$f(T_i, \delta_i, y_i|\mathbf{b}_i; \boldsymbol{\theta}) = f(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta}) f(y_i|\mathbf{b}_i; \boldsymbol{\theta}) \tag{3.3}$$

$$f(y_i|\mathbf{b}_i; \boldsymbol{\theta}) = \prod_j f(y_i(s_{ij})|\mathbf{b}_i; \boldsymbol{\theta}) \tag{3.4}$$

where $\boldsymbol{\theta}$ is the parameter vector. The contribution of the $i-$th subject to the log-likelihood is given by

$$logf(T_i, \delta_i, y_i; \boldsymbol{\theta}) = log \int f(T_i, \delta_i|\mathbf{b}_i; \boldsymbol{\theta}) \prod_j f(y_i(s_{ij})|\mathbf{b}_i; \boldsymbol{\theta}) f(\mathbf{b}_i; \theta) d\mathbf{b}_i \tag{3.5}$$

where for the survival function we have

$$S_i(t|\mathcal{M}_i(t)) = P(X > t|\mathcal{M}_i(t)) = exp\left(-\int_0^t h_i(s|\mathcal{M}_i(s);\boldsymbol{\theta})ds\right). \tag{3.6}$$

Computational issues may arise during maximization of the above log-likelihood function since the integrals in (3.6) and (3.5) do not generally have a closed form solution. For illustration purposes, consider the special case where $\mathbf{b}_i = [b_{0i}, b_{1i}]$, all data are fully observed, and $h_0(t)$ is constant, $h_0(t) = \lambda$, that is the baseline hazard rate correspond to an exponential distribution, then the survival function in (3.6) is tractable we derive

$$\begin{aligned} S_i(t|b_{0i}, b_{1i}) &= exp\left(-\int_0^t h_0(s)e^{am_i(s)}ds\right) \\ &= exp\left(-\int_0^t \lambda e^{a[(\beta_0+b_{0i})+(\beta_1+b_{1i})s]}ds\right) \\ &= exp\left(-\lambda e^{a(\beta_0+b_{0i})}\frac{e^{at(\beta_1+b_{1i})}-1}{a(\beta_1+b_{1i})}\right) \end{aligned} \tag{3.7}$$

and

$$f(t|b_{0i}, b_{1i}) = \left(\lambda e^{a[(\beta_0+b_{0i})+(\beta_1+b_{1i})t]}\right) exp\left(-\lambda e^{a(\beta_0+b_{0i})}\frac{e^{at(\beta_1+b_{1i})}-1}{a(\beta_1+b_{1i})}\right). \tag{3.8}$$

If for the $i-$th subject $\theta^* = \lambda e^{a(\beta_0+b_0)} > 0$ and $a^* = a(b_1+\beta_1) > 0$ then (3.7) and (3.8) correspond to the Gompertz distribution with density function

$$f(x) = \theta^* e^{a^* x} exp\left(\frac{\theta^*}{\alpha^*}(1-e^{a^* x})\right), \quad x \geq 0.$$

Numerical procedures have to be employed for the maximization of the log-likelihood and the evaluation of its integrals. Even then, the approach is not computationally stable and convergence is not guaranteed.

3.2 Marginal Models in the Presence of a Censored Covariate

We propose a population based modeling approach. We are interested in estimating the parameters of a generalized linear model (GLM) when longitudinal data are available and one covariate suffers from censoring. We assume that the other, fully observed covariates, that may be available refer only to baseline measurements and are not longitudinal in nature. We denote the time lag exists between the marker measurement and the occurence of the event by $X_{ij}^* = X_i - s_{ij} > 0$. It is expected that marker measurements taken closer to the time of the occurence of the event would be higher (if higher marker

values are more indicative of the disease). For a detailed discussion about such a setting see Cai et al. (2006). Consider the GLM of the form

$$\begin{aligned} E(Y_{ij}|X^*_{ij}=x^*_{ij},\mathbf{z}_i) &= \mu_{ij} \\ g(\mu_{ij}) &= [\mathbf{x}^{*'}_{ij},\mathbf{z}_i^{'}]'[\beta_0,\beta_1,\boldsymbol{\beta}_2^{'}]^{'} \\ Var(Y_{ij}|X^*_{ij}=x^*_{ij},\mathbf{z}_i) &= \tau^2 v(\mu_{ij});\ i=1,\ldots,N,\ \ j=1,\ldots,n_i \end{aligned}$$

where $g(\cdot)$ is the link function, β_0 is the intercept, β_1 is the coefficient of the censored covariate, $\boldsymbol{\beta}_2$ is the vector of coefficients corresponding to the fully observed covariates, the vector $\mathbf{x}^*_{ij}=[1,x^*_{ij}]'$ refers to the covariate values that may be censored, the vector $\mathbf{z}_i^{'}=[z_{i1},\ldots,z_{i,p-1}]$ refers to the fully observed covariate values, $v(\mu_i)$ is the variance function and τ^2 is the dispersion parameter (McCullogh and Searle (2001)). The within subject association is denotes as ρ_{jk}. Assuming that there are K subjects in total, and each one exhibits n_i measurements, $i=1,\ldots,K$, then model (3.9) in vector form can be written as

$$\mathrm{E}(\mathbf{Y}|\mathbf{X})=\boldsymbol{\mu},\ \ \mathbf{g}(\boldsymbol{\mu})=\mathbf{X}\boldsymbol{\beta},\ \ \mathrm{Var}(\mathbf{Y}|\mathbf{X})=\tau^2\mathbf{V}(\boldsymbol{\mu}), \tag{3.9}$$

where $\mathbf{Y}=[y_{11},\ldots,y_{1n_1},y_{21},\ldots,y_{2n_2},\ldots,y_{K1},\ldots,y_{Kn_K}]'$, $\mathbf{X}=[(\mathbf{x}^{*'}_1,\mathbf{z}^{'}_1),\ldots,(\mathbf{x}^{*'}_n,\mathbf{z}^{'}_n)]'$, $\boldsymbol{\beta}=[\beta_0,\beta_1,\boldsymbol{\beta}_2^{'}]^{'}$, $\boldsymbol{\mu}=[\mu_{11},\ldots,\mu_{Kn_K}]'$, $\mathbf{g}(\boldsymbol{\mu})=[g(\mu_{11}),\ldots,g(\mu_{Kn_K})]'$ and for the variance we have that $\mathbf{V}=\mathbf{V}(\boldsymbol{\mu})$ which would be a symmetric positive-definite block diagonal matrix of known functions with number of rows and columns equal to $N=\sum_i^K n_i$.

The available data are of the form

$$\{\mathbf{Y},\mathbf{T},\boldsymbol{\Delta},\mathbf{Z}\}=\left(\begin{array}{ccccccc} y_{11} & t_1 & s_{11} & \delta_1 & z_{1,1} & \ldots & z_{p-1,1} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \\ y_{1n_1} & t_1 & s_{1n_1} & \delta_1 & z_{1,1} & \ldots & z_{p-1,1} \\ \hline y_{21} & t_1 & s_{21} & \delta_2 & z_{1,2} & \ldots & z_{p-1,2} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \\ y_{2n_2} & t_2 & s_{2n_2} & \delta_2 & z_{1,2} & \ldots & z_{p-1,2} \\ \hline \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \\ \hline y_{K1} & t_K & s_{Kn1} & \delta_K & z_{1,K} & \ldots & z_{p-1,K} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \\ y_{Kn_K} & t_K & s_{Kn_K} & \delta_K & z_{1,K} & \ldots & z_{p-1,K} \end{array}\right), \tag{3.10}$$

where $t_i=min(x_i,c_i)$, c_i is censoring time for the $i-$th subject and $\delta_i=I_{(x_i<c_i)}$, where $I_{(A)}$ denotes the indicator function of the event A. We denote with t^*_{ij} is the censored time lag, namely, $t^*_{ij}=t_i-s_{ij}$.

3.3 Estimating Function Approach

Here, we extend the approach developed in Chapter 2 when repeated measurements are available. We consider employing the estimating equations

$$\mathbf{G}^*=\dot{\boldsymbol{\mu}}^{c'}\mathbf{W}(\mathbf{Y}-\boldsymbol{\mu}^c) \tag{3.11}$$

where the $i,j-th$ row of $\boldsymbol{\mu}^c$ is $\mu_{ij}^c = E_{X_i|T_i,\Delta_i}\left\{g^{-1}(\mathbf{x}_{ij}^{*\prime}\boldsymbol{\beta})\right\}$. We assume that the unconditional variance of the response is finite and $E\left\{\dot{\boldsymbol{\mu}}^{c\prime}\mathbf{W}\dot{\boldsymbol{\mu}}^c\right\}$ is nonsingular, where $\mathbf{W}^{-1}=\mathbf{V}$.

In the case where there is no censoring these estimating function (3.11) reduces to the usual estimating function employed in generalized linear models dealing with longitudinal models under the marginal approach presented in Fitzmaurice et al. (2004). That is, the well known approach of Generalized Estimating Equations (GEE). A computationally easier alternative is to consider the Unweighted estimating function $\mathbf{G}^{Un}=\dot{\boldsymbol{\mu}}^{\mathbf{c}\prime}(\mathbf{Y}-\boldsymbol{\mu}^{\mathbf{c}})$. In practice, for solving $\mathbf{G}^*=\mathbf{0}$, estimates for the dispersion parameter and the correlation parameter(s) are required. The CC approach can provide consistent estimates for both the correlation and the dispersion parameters. Given these estimates one can proceed by using the Fisher Scoring algorithm to solve the proposed quasi score (QS) estimating equations:

$$\hat{\boldsymbol{\beta}}_{(m+1)}=\hat{\boldsymbol{\beta}}_{(m)}+\left(\dot{\boldsymbol{\mu}}^{c\prime}_{(m)}\mathbf{W}_{(m)}\dot{\boldsymbol{\mu}}^{c}{}_{(m)}\right)^{-1}\mathbf{A}^*_{(m)}(\mathbf{Y}-\boldsymbol{\mu}^c_{(m)}).$$

One can use $\boldsymbol{\beta}^{(CC)}$, that is the obtained estimate of the CC approach, as initial estimate values of parameter vector $\boldsymbol{\beta}$ for the above iterative procedure.

3.3.1 Examples of commonly used link functions

Here we present some examples of three commonly used link functions, that is the identity, the logit and the log. We show the form of the QS estimating functions and discuss what assumptions are trivially satisfied due to the nature of the data.

Identity link function for continuous data:

In this case we consider a simple regression setting relating the response to the time lag. We have

$$\mu_{ij}^c(\boldsymbol{\beta})=E(Y_{ij}|T_i,\Delta_i)=\begin{cases}\beta_0+\beta_1(t_i-s_{ij}), & \text{if } \Delta_i=1\\ \beta_0+\beta_1(E(X_i|X_i>t_i)-s_{ij}), & \text{if } \Delta_i=0.\end{cases}$$

For the variance and covariance we derive:

$$Var(Y_{ij}|T_i,\Delta_i)=\begin{cases}\tau^2 & \text{if } \Delta_i=1\\ \beta_1^2Var(X_i|X_i>t_i)+\tau^2, & \text{if } \Delta_i=0\end{cases}$$

and

$$Cov(Y_{ij},Y_{ik}|T_i,\Delta_i)=\begin{cases}\rho\tau^2 & \text{if } \Delta_i=1\\ \beta_1^2Var(X_i|X_i>t_i)+\rho\tau^2, & \text{if } \Delta_i=0.\end{cases}$$

Thus in this case, the matrix $\dot{\boldsymbol{\mu}}^c$ is an $(N=\sum_{i=1}^K n_i)\times 2$ matrix with its first column equal to $\mathbf{1}_N$. The elements of the second column equal to $t_{ij}^*=t_i-s_{ij}$ when the covariate is observed and $E(X_i^*|X_i>t_i)=E(X_i|X_i>t_i)-s_{ij}$ when censoring occurs. Let $\mathbf{A}$ be equal to the $(N\times N)$ diagonal matrix $diag\{\tau^2\}_{i=1}^N$. Let also $\mathbf{R}$ be the block diagonal matrix $\mathbf{R}=diag\{\mathbf{R}_1,\mathbf{R}_2,\ldots,\mathbf{R}_K\}$ where $\mathbf{R}_i$ is the correlation matrix of the measurements of the $i-$th subject. Hence, matrix $\mathbf{V}$ in this case turns out to be

$$\mathbf{V} = \mathbf{A}^{\frac{1}{2}}\mathbf{R}\mathbf{A}^{\frac{1}{2}} + \mathbf{C}$$

where $\mathbf{C} = diag\{\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C}_K\}$ with $\mathbf{C}_i = \beta_1^2 Var(X_i|T_i, \Delta_i) \times \mathbf{J}_{n_i}$. Note that if we consider that all correlations are equal to ρ then $\mathbf{R}_i = \rho\mathbf{J}_{n_i}$. We note that in the identity link case, our approach simply reduces to mean imputation of the censored covariate. Then the appropriate weights, that is matrix $\mathbf{W} = \mathbf{V}^{-1}$, are calculated. Finally, to solve the corresponding estimating equations an estimate of τ^2 and ρ are needed, and in practice can be obtained by the CC approach. The assumption that the unconditional variance of the response is finite is satisfied in this case if $E(X_i^2) < \infty$. Note that matrix $\mathbf{V}$ is similar to the corresponding matrix used in the common setting of no censoring (see Fitzmaurice et al. (2004)), only now the correction $\mathbf{C}$ is added.

Binary data with the logit link function:

In the case of binary data and when the logit link is employed we derive

$$\mu_{ij}^c(\boldsymbol{\beta}) = E(Y_{ij}|T_i, \Delta_i) = \begin{cases} \frac{exp(\beta_0+\beta_1(t_i-s_{ij}))}{1+exp(\beta_0+\beta_1(t_i-s_{ij}))}, & \text{if } \Delta_i = 1 \\ E\left(\frac{exp(\beta_0+\beta_1(X_i-s_{ij}))}{1+exp(\beta_0+\beta_1(X_i-s_{ij}))}|X_i > t_i\right), & \text{if } \Delta_i = 0 \end{cases}$$

where $E(Y_{ij}|T_i, \Delta_i) = P(Y_{ij} = 1|T_i, \Delta_i)$ and it can be shown that

$$Var(Y_{ij}|T_i, \Delta_i) = \mu_{ij}^c(\boldsymbol{\beta})(1 - \mu_{ij}^c(\boldsymbol{\beta})).$$

For the covariance we have

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}|T_i, \Delta_i) &= E(Y_{ij}, Y_{ik}|T_i, \Delta_i) - E(Y_{ij}|T_i, \Delta_i)E(Y_{ik}|T_i, \Delta_i) \\ &= P(Y_{ij} = 1, Y_{ik} = 1|T_i, \Delta_i) - P(Y_{ij} = 1|T_i, \Delta_i)P(Y_{ik} = 1|T_i, \Delta_i) \end{aligned}$$

The within subject association is defined through the log odds ratio. Under the unstructured pairwise log odds ratio pattern we have

$$\gamma_{ijk} = log(OR(Y_{ij}, Y_{ik}|T_i, \Delta_i)) = log\left(\frac{P(Y_{ij} = 1, Y_{ik} = 1|T_i, \Delta_i)P(Y_j = 0, Y_k = 0|T_i, \Delta_i)}{P(Y_j = 1, Y_k = 0|T_i, \Delta_i)P(Y_j = 0, Y_k = 1|T_i, \Delta_i)}\right).$$

Following the ideas of Diggle (1992) and Carey et al. (1993) it is easy to show that

$$logit(P(Y_{ij} = 1|Y_{ik} = y_{ik}, T_i, \Delta_i)) = \gamma_{ijk}y_{ik} + w_{ijk} \tag{3.12}$$

where

$$w_{ijk} = log\left(\frac{P(Y_{ij} = 1|T_i, \Delta_i) - P(Y_{ij} = 1, Y_{ik} = 1|T_i, \Delta_i)}{1 - P(Y_{ij} = 1|T_i, \Delta_i) - P(Y_{ik} = 1|T_i, \Delta_i) + P(Y_{ij} = 1, Y_{ik} = 1|T_i, \Delta_i}\right).$$

In the simplest case we may assume that $\gamma_{ijk} = \alpha$. Thus, α is considered simply as a regression coefficient in a logistic regression model where Y_{ik} plays the role of a covariate and w_{ijk} is considered as an offset. Note that given the probabilities $P(Y_{ij} = y_{ij}|T_i, \Delta_i)$, $y_{ij} = 0, 1$ and the value of the $OR = \exp(\gamma_{ijk})$, then one can obtain the probabilities of the form $P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}|T_i, \Delta_i)$, $y_{ij} = 0, 1$, $y_{ijk} = 0, 1$. As in Carey et al. (1993) the estimation procedure iterates between the following two steps:

- Step 1: In the $r-$th iteration, given the estimates of $\hat{\alpha}^{(r)}$ and $\hat{\boldsymbol{\beta}}^{(r)}$, evaluate the offset w_{ijk} and obtain $\hat{\alpha}^{(r+1)}$ through (3.12).
- Step 2: Given $\hat{\alpha}^{(r+1)}$ and $\hat{\boldsymbol{\beta}}^{(r)}$ one can obtain the probabilities $P(Y_{ij}=1,Y_{ik}=1|T_i,\Delta_i)$, $P(Y_{ij}=1|T_i,\Delta_i)$, and $P(Y_{ik}=1|T_i,\Delta_i)$ and hence $Cov(Y_{ij},Y_{ik}|T_i,\Delta_i)$. Thus, one can solve the proposed estimating equations obtained by (3.11) to derive the updated estimate $\hat{\boldsymbol{\beta}}^{(r+1)}$.

In the more general setting one may assume that $\gamma_{ijk}=\boldsymbol{\alpha}\mathbf{r}_{ijk}^{'}$ where $r_{ijk}^{'}$ is a known vector of pair specific covariates which specify the form of the association of Y_{ij} and Y_{ik}. For example, r_{ijk} might involve indicators that inform us about the relation of the subjects within clusters, such as husband-wife, parent-child etc (see Carey et al. (1993)).

We note that the design matrix $\dot{\boldsymbol{\mu}}^c$ is an ($\sum n_i \times 2$) matrix with the first element of the $i,j-$th row equal to $\frac{exp(\beta_0+\beta_1(t_i-s_{ij}))}{\{1+exp(\beta_0+\beta_1(t_i-s_{ij}))\}^2}$ when the covariate is observed exactly and $E\left(\frac{exp(\beta_0+\beta_1(X_i-s_{ij}))}{\{1+exp(\beta_0+\beta_1(X_i-s_{ij}))\}^2}|X_i>t_i\right)$ when the covariate is right censored. The second element of $i,j-$th row equals to $\frac{(t_i-s_{ij})exp(\beta_0+\beta_1(t_i-s_{ij}))}{\{1+exp(\beta_0+\beta_1(t_i-s_{ij}))\}^2}$ when the covariate is observed exactly and $E\left((X_i-s_{ij})\frac{exp(\beta_0+\beta_1(X_i-s_{ij}))}{\{1+exp(\beta_0+\beta_1(X_i-s_{ij}))\}^2}|X_i>t_i\right)$ when right censoring occurs. Due to the binary nature of the data, the unconditional variance of the response is always finite, that is $Var(Y_{ij})<\infty$.

Count data with the log link function:

In the case of count data and when the log link is employed we have

$$\mu_{ij}^c=E(Y_{ij}|T_i,\Delta_i)=\begin{cases} exp(\beta_0+\beta_1(t_i-s_{ij})), & \text{if } \Delta_i=1 \\ E\left\{exp(\beta_0+\beta_1(X_i-s_{ij}))|X_i>t_i\right\}, & \text{if } \Delta_i=0. \end{cases}$$

Based on the usual mean to variance relationship the conditional variance given the time to event or censoring can be written as

$$Var(Y_{ij}|T_i,\Delta_i)=\begin{cases} \tau^2\mu_{ij}(\boldsymbol{\beta}), & \text{if } \Delta_i=1 \\ \tau^2E\left\{\mu_{ij}(\boldsymbol{\beta})|X_i>t_i\right\}+Var\left\{\mu_{ij}(\boldsymbol{\beta})|X_i>t_i\right\}, & \text{if } \Delta_i=0. \end{cases}$$

If we denote $\mu_{ij}=E(Y_{ij}|X_i=x_i)$ and assume that we have only one correlation parameter to estimate as in the linear case (let ρ), then for the conditional covariance that corresponds to a censored covariate value we derive

$$\begin{aligned} Cov(Y_{ij},Y_{ik}|X_i>t_i) &= E(Y_{ij}Y_{ik}|X_i>t_i)-E(Y_{ij}|X_i>t_i)E(Y_{ik}|X_i>t_i) \\ &= E(\rho\tau^2\sqrt{\mu_{ij}}\sqrt{\mu_{ik}}|X_i>t_i)+Cov(\mu_{ij},\mu_{ik}|X_i>t_i) \end{aligned}$$

since

$$\begin{aligned} E(Y_{ij}Y_{ik}|X_i>t_i) &= E(E(Y_{ij}Y_{ik}|X_i=x_i)|X_i>t_i) \\ &= E(Cov(Y_{ij}Y_{ik}|X_i=x_i)+E(Y_{ij}|X_i=x_i)E(Y_{ik}|X_i=x_i)|X_i>t_i) \\ &= E(\rho\tau^2\sqrt{\mu_{ij}}\sqrt{\mu_{ik}}+\mu_{ij}\mu_{ik}|X_i>t_i) \end{aligned}$$

and $E(Y_{ij}|X_i > t_i) = \mu_{ij}^c = E(E(Y_{ij}|X_i = x_i)|X_i > t_i) = E(\mu_{ij}|X_i > t_i)$. Thus, for the covariance we have

$$Cov(Y_{ij}, Y_{ik}|T_i, \Delta_i) = \begin{cases} \rho\tau^2\sqrt{\mu_{ij}}\sqrt{\mu_{ik}}, & \text{if } \Delta_i = 1 \\ E(\rho\tau^2\sqrt{\mu_{ij}}\sqrt{\mu_{ik}}|X_i > t_i) + Cov(\mu_{ij}, \mu_{ik}|X_i > t_i), & \text{if } \Delta_i = 0. \end{cases}$$

In this case the first element of the $i, j-$th row of matrix $\dot{\boldsymbol{\mu}}^c$ equals to $exp(\beta_0 + \beta_1(t_i - s_{ij}))$ when the covariate value is observed exactly and $E\left\{exp(\beta_0 + \beta_1(X_i - s_{ij}))|X_i > t_i\right\}$ when the covariate value is right censored. The second element of $i, j-$th row of matrix $\dot{\boldsymbol{\mu}}^c$ equals to $(t_i - s_{ij})exp(\beta_0 + \beta_1(t_i - s_{ij}))$ and $E\left\{(X_i - s_{ij})exp(\beta_0 + \beta_1(X_i - s_{ij})|X_i > t_i\right\}$ when right censoring occurs. For the proposed method to be applicable in this case, we require $Var\left\{exp(\beta_0 + \beta_1(X_i - s_{ij}))\right\} < \infty$ or equivalently $M_X(2\beta_1) < \infty$, where $M_X(.)$ is the moment generating function (m.g.f) of the distribution of the time to event variable.

3.4 HCNS approach for modeling the censored covariate

In the case where only the censored covariate is available our approach is a two stage one. Due to the need of calculation of the conditional expectations presented above some assumptions regarding the distribution of X are required. One approach would be to assume a parametric approach for the covariate distribution. Since the censored covariate refers to a time to event random variable one may choose among the commonly used parametric models used in survival analysis such as Exponential, Weibull, Gamma etc. Note that for each model the applicability of our approach should be checked based on the finite variance condition as also mentioned in the previous chapter. This is discussed separately in the above examples that refer to the three commonly used link functions.

Here we explore the spline approach that was utilized to model the censored covariate in Chapter 2, namely model (2.14) under the linear restrictions (2.15) that are required to impose monotonicity. After fitting model (2.14) under the necessary constraints of monotonicity we derive $\hat{\boldsymbol{\theta}}$, and an estimator of the smoothed version of the cumulative hazard based on this constrained natural spline (HCNS) approach is obtained. Based on the HCNS technique all conditions for the application of our method are satisfied. In this case our approach is summarized in the following four stages:

- *Stage 1:* Obtain the usual Kaplan Meier based cumulative hazard step estimate based only on the information provided by the censored covariate.
- *Stage 2:* Fit the spline model in (3.4) under the necessary constraints of monotonicity
- *Stage 3:* Perform CC analysis to obtain initial estimates for the parameters of interest, as well as consistent estimates for the dispersion and within subject association parameters.
- *Stage 4:* Based on the previous three steps, solve $\mathbf{G}^* = \mathbf{0}$ using the Fisher Scoring algorithm.

Consider the case that another fully observed covariate, let Z, is available and need to be accommodated in our modeling approach. We assume that this covariate, is not longitudinal in nature and refers only to baseline or time invariant measurements of the subjects. In this case the corresponding QS estimating functions can be shown to be $\mathbf{G}^* = \dot{\boldsymbol{\mu}}^{c'}\mathbf{W}(\mathbf{Y} - \boldsymbol{\mu}^c)$ where the $i,j-$th row of $\boldsymbol{\mu}^c$ is $\boldsymbol{\mu}^c_{ij} = E_{X_i|T_i,\Delta_i,Z_i}\left\{g^{-1}([\mathbf{x}_i', z_i]\boldsymbol{\beta})\right\}$. The accommodation of other fully observed covariates, let $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_p]$, can be done straightforwardly in a similar fashion. The information that the fully observed covariates may carry about the censored one should be taken into account. Here, we explore spline models that relate the censored covariate with the fully observed ones based on semiparametric assumptions in order to model $X|Z_1, Z_2, \ldots, Z_p$. We utilize the well known Cox model and smooth the step function of the baseline cumulative hazard as described in the previous section. That is, our approach is summarized in the following stages:

- *Stage 1:* Perform usual Cox analysis and derive a consistent estimate of parameter vector $\hat{\boldsymbol{\gamma}}$, along with the cox estimate of the baseline cumulative hazard function.
- *Stage 2:* Fit the spline model in (2.14) to the corners of the estimated step cumulative hazard function of the previous stage.
- *Stage 3:* Perform CC analysis to obtain initial estimates for the parameters of interest along with consistent estimates of the dispersion and the within subject association parameters.
- *Stage 4:* Based on the estimated vector coefficients of the cox model used in stage 1, the spline modeling employed in stage 2, and the obtained estimates of stage 3, solve $\mathbf{G}^* = \mathbf{0}$.

3.5 Application

We apply our methods to a data set that involves HIV patients. A clinical trial that involves 467 patients with advanced HIV infection who had previously failed or were intolerant to treatment with zidovudine (AZT) is presented in Abrams et al. (1994). The aim of this study is to compare two antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC) regarding their effect on time to death. Subjects were randomized to the two therapy groups and were monitored by measuring their CD4 cell counts at 2, 6, 12 and 18 months. The CD4 cell count is widely used as a biomarker for AIDS progression and one might be interested in how the CD4 cell count is affected by the time lag or time event among other covariates. The level of censoring in this study is about 60%. For more details of this study we refer to Abrams et.at. (1994). The data set is included in JM R package written by Rizopoulos (2010) where an analysis based on the joint modeling approach is provided. These data are also analyzed in Guo and Carlin (2004) where parametric models are considered to link the time to survival to the other covariates. The available variables of this data set are: the id of the patients, the treatment group (ddC=0, ddI=1), previous opportunistic infection (prevOI=1 for AIDS, prevOI=-1 for No Aids), gender (male=1, female=-1), and AZT (intolerance=-1 and failure=1). This is the coding also used in SAS in Littell et al. (2006). We consider the square root of $CD4$

as the biomarker response. It is common to work with $\sqrt{CD4}$ instead of $CD4$ since its distribution may be skewed.

First, we provide the graphs that show the actual $\sqrt{CD4}$ measurements over time for each profile and each patient (Figures 3.1, 3.2, 3.3 and 3.4). Note that there are no patients with the following two profiles: (i) $Drug = ddC$, $Gender = Female$, $PrevOI = AIDS$, $AZT = failure$, and (ii) $Drug = ddI$, $Gender = Male$, $PrevOI = NoAIDS$, $AZT = failure$. Next, we consider exploring the survival curves that will account for all the time invariant covariates and correspond to all possible profiles of a patient (see Figure 3.5). Apart from the fact that this analysis is required to proceed to our modeling approaches, it is also of clinical interest to explore the effect of the fixed covariates on time to death. We observe that the Cox based survival curves that correspond to subjects with $PrevOI = NoAIDS$ yield greater survival probabilities in all cases (i.e. in all subprofiles). Along with the Cox survival curves the spline curves are also plotted in Figure 3.5.

Here, we consider relating the square root of $CD4$ cells to the other covariate by:

$$\begin{aligned} Y_i = \sqrt{CD4}_i &= \beta_0 + \beta_1^{(C)} X_i + \beta_1^{(L)}(X_i - s_{ij}) \\ &+ \beta_2 Drug_i + \beta_3 Drug_i X_i + \beta_4 Drug_i (X_i - s_{ij}) \\ &+ \beta_5 Gender_i + \beta_6 PrevOI + \beta_7 AZT + \epsilon_{ij} \end{aligned} \tag{3.13}$$

Following our approaches presented above, we evaluate the underlying conditional expectation $E(X_i|T_i, \Delta_i, Drug_i, Gender_i, AZT_i)$. This is done by fitting the $HCNS$ model to the Cox based cumulative baseline hazard function. The results of the usual Cox analysis are: $\hat{\gamma}_{Drug} = 0.2168, (SE = 0.1388, p - value = 0.1464)$, $\hat{\gamma}_{Gender} = -0.1710 (SE = 0.1227, p - value = 0.1636)$, $\hat{\gamma}_{prevOI} = 0.6459 (SE = 0.1135, p - value = 0.0000)$, $\hat{\gamma}_{AZT} = 0.0768 (SE = 0.0817, p - value = 0.3359)$.

Before applying the QS method we first discuss the Unweighted method which easier to implement. After imputing the censored observations by replacing the time to censoring with $\hat{E}(X_i|T_i, \Delta_i, Drug_i, Gender_i, AZT_i)$ which is evaluated based on the spline survival estimate $\hat{S}(X_i|T_i, \Delta_i, Drug_i, Gender_i, AZT_i)$, we can straightforwardly perform the Unweighted method. This is done by using the CC within subjects correlation coefficient estimate ($\hat{\rho}^{(CC)} = 0.688$) as well as the CC dispersion parameter estimate ($\hat{\sigma}^{(CC)} = 2.9283$). These CC estimates can be obtained by standard software since routines are readily available in SAS (`proc genmod`) and R (`gee`). Once the imputation is employed then all times are regarded as fully observed and during the repetitive procedure of solving the estimating equations $\hat{\rho}^{(CC)}$ and $\hat{\sigma}^{(CC)}$ are held fixed and considered known. The results of the Unweighted method along with the ones of the CC are presented in Table 3.1. To obtain 95% confidence intervals for the Unweighted method we considered the percentile bootstrap with 1000 bootstrap samples. For the CC we present the usual asymptotic confidence intervals.

We observe that the Unweighted Method yields for almost all parameters essentially narrower confidence intervals compared to ones obtained by the CC. An interesting fact is that the cross sectional effect associated with the time to event variable turns out to be statistically significant according to the Unweighted method while the CC yields the asymptotic confidence interval $(-0.1795, 0.0437)$ for this parameter. One might have expected that the time to event variable not to be significant since at the time of the event

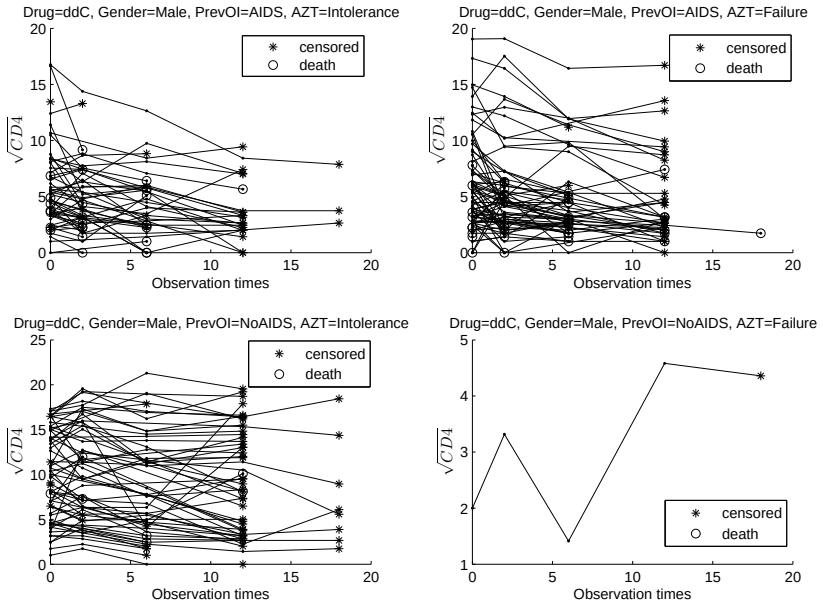

Figure 3.1: The actual measurements of $\sqrt{CD4}$ of each patient over time for the profiles mentioned separately in each title

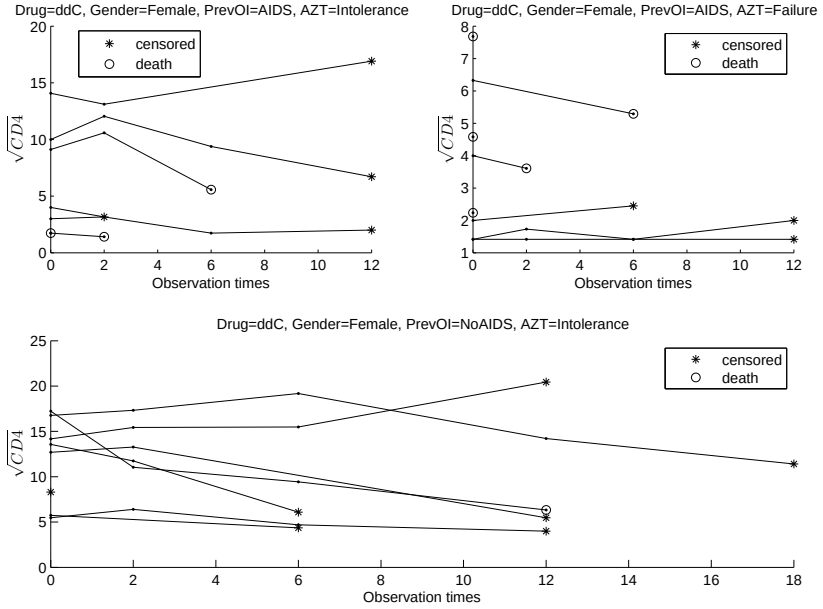

Figure 3.2: The actual measurements of $\sqrt{CD4}$ of each patient over time for the profiles mentioned separately in each title

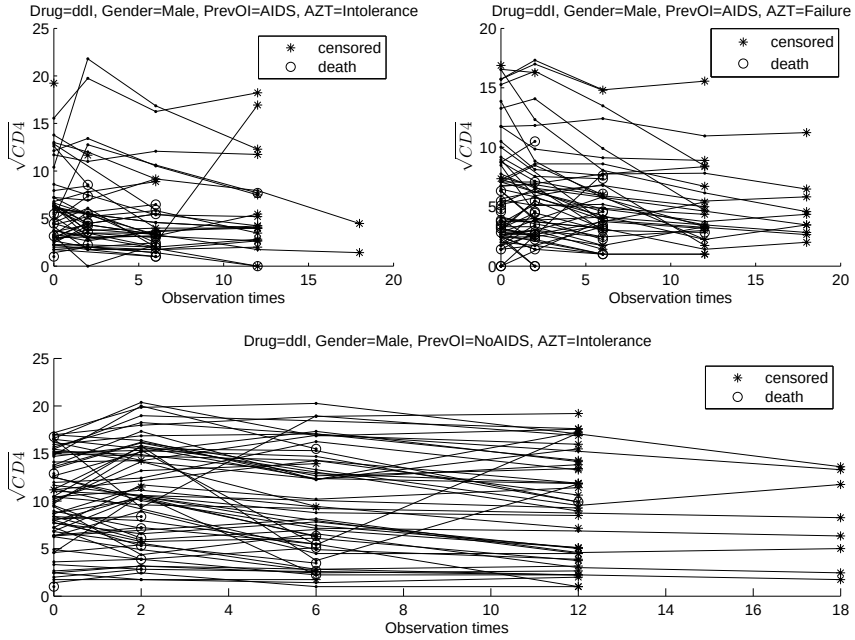

Figure 3.3: The actual measurements of $\sqrt{CD4}$ of each patient over time for the profiles mentioned separately in each title

Table 3.1: Parameter estimates and corresponding 95% confidence intervals as obtained by the CC, the Unweighted method and the QS Method. The asterisk indicates statistical significance for the corresponding estimates of the QS method at $\alpha = 0.05$. ($\hat{\rho}^{(CC)} = 0.688$ and $\hat{\sigma}^{(CC)} = 2.9283$)

	CC			Unweighted method			QS method		
Parameters	Estimates	95% CIs		Estimates	95% CIs		Estimates	95% CIs	
$intercept$	5.1079	3.6779	6.5379	5.6980	4.4871	6.8030	5.5429	4.4410	6.7044*
X_i	-0.0679	-0.1795	0.0437	-0.1093	-0.1537	-0.0570	-0.1093	-0.1527	-0.0631*
$timelag$	0.1883	0.0804	0.2962	0.1573	0.1152	0.1985	0.1597	0.1202	0.1991*
$drug \times timelag$	-0.0356	-0.1069	0.1781	-0.0053	-0.0629	0.0534	-0.0078	-0.0615	0.0519
$drug$	0.0004	-1.8307	1.8316	0.1828	-0.7826	1.1976	-0.0555	-1.0348	0.9649
$drug \times X_i$	-0.0310	-0.2322	0.1703	0.0264	-0.0379	0.0912	0.0395	-0.0263	0.1007
$gender$	-0.1143	-0.9512	0.7226	-0.4784	-1.1508	0.1043	-0.2376	-0.9312	0.4080
$prevOI$	-1.7649	-2.6152	-0.9146	-0.9591	-1.6467	-0.2827	-1.0815	-1.6787	-0.5357*
AZT	-0.2267	-0.6069	0.1535	-0.0486	-0.4122	0.3388	-0.0449	-0.3955	0.3434

the $\sqrt{CD4}$ cell counts should not significantly differ from subject to subject. However, in such studies where new drugs are under study and comparison, it might be the case that individuals do not die form the actual disease causes but due to drug side effects (see Palella et al. (2006)). In their study they conclude that the proportion of deaths attributable to non-AIDS diseases may include hepatic, cardiovascular, and pulmonary disorders, as well as non-AIDS malignancies.

We continue by applying the QS estimating function approach assuming the same model as before. To assess this method we need to obtain an estimate for the conditional

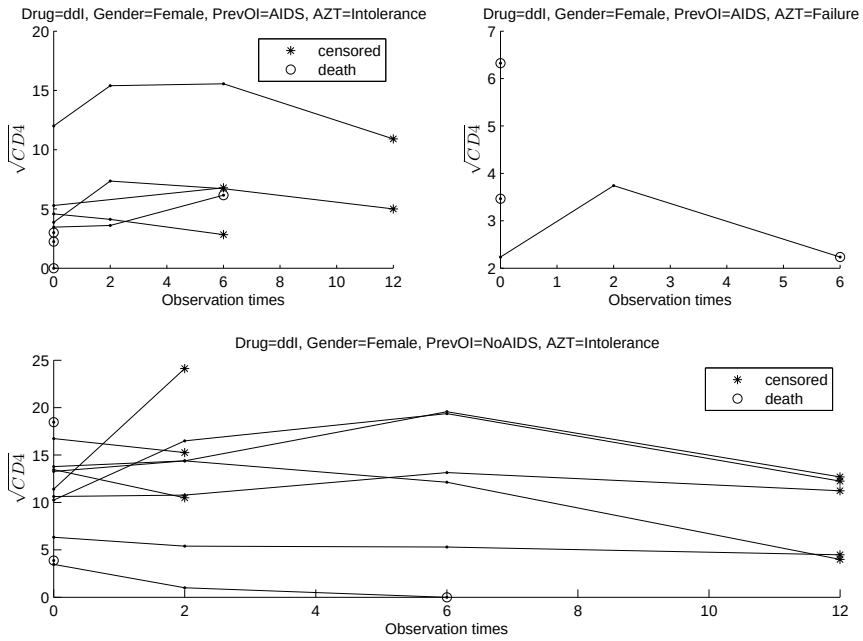

Figure 3.4: The actual measurements of $\sqrt{CD4}$ of each patient over time for the profiles mentioned separately in each title

variance $V(Y_{ij}|X_i, \Delta_i, Drug_i, Gender_i, PrevOI_i, AZT_i) = V(Y_{ij}|X_i, \Delta_i, \mathbf{Z}_i)$, where with $\mathbf{Z}_i$ we denote the $i-$th row of matrix $\mathbf{Z}$ which has as columns all fixed covariates, namely $Drug_i, Gender_i, PrevOI_i, AZT_i$, $i = 1, ...467$. Thus, for the conditional variance of the model under study given an event we have $Var(Y_{ij}|X_i, \Delta_i = 1, \mathbf{Z}_i) = \tau^2$ while for a censored time we derive:

$$Var(Y_{ij}|X_i, \Delta_i = 0, \mathbf{Z}_i) =$$
$$\tau^2 + Var(X_i|X_i > t_i, \mathbf{Z}_i) \times \left\{ [(\beta_1^{(C)} + \beta_1^{(L)}) + Drug_i(\beta_3 + \beta_4)]^2 \right\} \quad (3.14)$$

The conditional variance $Var(X_i|X_i > t_i, \mathbf{Z}_i)$ can be evaluated through the spline based survival estimate for any given patient's profile. As in the computational simpler Unweighted approach, we use the CC estimates for the within subjects correlation as well as for the dispersion parameter, which are considered fixed and known at their estimated values. The results of the parameter estimates based on this approach are also shown in Table 3.1. We observe that results are consistent with the ones provided by the Unweighted method, yielding also essentially narrower confidence intervals compared to the ones yielded by the CC.

3.6 Discussion

In this chapter we explore a generalization of the method presented in Chapter 2 when marker measurements are taken repeatedly over time. The proposed method is a popu-

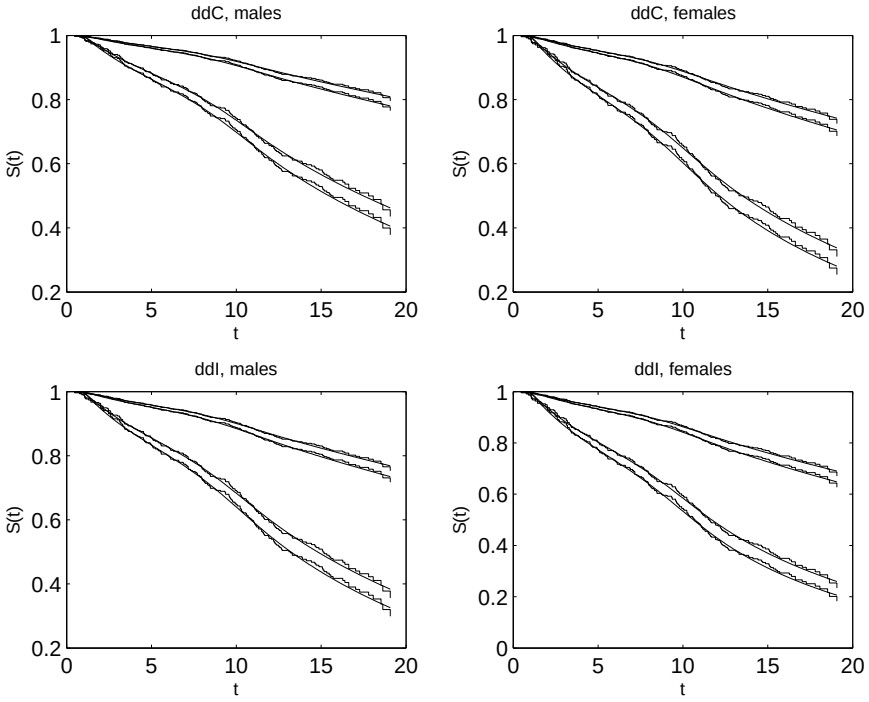

Figure 3.5: Cox and Spline based estimates of the survival curves. For each of the four graphs the curves refer (moving upwards) to the following profiles (i) PrevOI=AIDS, AZT=Failure (ii) PrevOI=AIDS AZT=Intolerance, (iii) PrevOI=NoAIDS AZT=Failure (iv) PrevOI=Noaids AZT=Intolerance.

lation oriented one, and is based on an estimating equation approach. As such, it can be considered as a generalization of the well known Generalized Estimating Equation approach (see Fitzmaurice for a detailed overview) since it accommodates a censored covariate. Our approach differs from the one taken in Rizopoulos where joint models are employed. Joint models are subject specific oriented models for which inference is made based on maximum likelihood techniques. Hence, parametric models need to be assumed for the involved random effects as well as for the response given the covariates. Our approach does not require such parametric assumptions. Even though a parametric model might be chosen for the distribution of the time to event variable given the other fully observed covariates, a flexible spline model may be considered instead, as a non-parametric alternative. We consider using the spline approach developed in Chapter 2 which has the merit of being numerically stable since convex optimization is involved and convergence is guaranteed. Additional, fully observed, covariates are incorporated using the same ideas developed Chapter 2 using semi parametric models for the survival function. We note that the linear case can be easily viewed as a natural generalization of the GEE approach while for the case of binary data our approach is related to the one presented in Carey et al. (1993).

A drawback of our approach is that it cannot address subject specific trajectories over

time, as in the case of joint models which can be considered as the extension of linear mixed models in the case of censoring (see Rizopoulos (2011)). In longitudinal studies one should proceed with caution sice there might be two potential sources of information, that is the cross sectional and the longitudinal. Longitudinal studies primarily focus on characterizing the within subject change of the response over time. However, cross sectional information must be also taken into account. These two sources of information may conflict, hence it is essential to build models that accommodate both longitudinal and cross sectional effects. By considering different parameters for these two sources of information, and performing simultaneous estimation, one may also proceed to further comparisons for these effects or even provide an estimate of a combined effect. For an overview about separating cross sectional and longitudinal effects we refer to Fitzmaurice et al. (2004).

Consider our setting, where the response marker measurements are taken repeatedly over time on each subject. The baseline marker measurement would refer to cross sectional information and the time-lag to longitudinal information. Assuming that its association to the response is linear we consider

$$E(Y_{ij}) = \beta_0 + \beta_1^{(C)} X_i + \beta_1^{(L)} (X_i - s_{ij}) \tag{3.15}$$

where $\beta_1^{(C)}$ refers to the cross sectional effect of X_i and $\beta_1^{(L)}$ to the longitudinal, with $\beta_1^{(C)} \neq \beta_1^{(L)}$. The proposed approach can be implemented in a straightforward fashion to account both longitudinal and cross sectional information.

Note that if the data are generated from a mixed effect model with a random intercept, b_{0i}, that is

$$Y_{ij} = \beta_0 + \beta_1^{(C)} X_i + \beta_1^{(L)} (X_i - s_{ij}) + b_{0i} + e_{ij}. \tag{3.16}$$

then this model could be estimated by assuming that the within subject association is common and equal to ρ. However, it would be interesting to investigate the robustness of our estimating function approach by using this simple correlation structure to broader models that may contain random effects for the time dependent slope. In the classical setting, when no censoring is present, the obtained estimator of the parameters of interest is robust to misspecification which makes the GEE approach attractive. An appealing property of our approach is that, unlike likelihood based approaches such as joint modeling, we make no distributional assumptions regarding the parametric form of the response distribution. The covariate distribution is also non-parametrically estimated through splines.

However, the above advantages of our approach come at cost. Consider for example what happens when the mechanism that generated the data is of the form

$$Y_{ij} = \beta_0 + \beta_1^{(C)} X_i + \beta_1^{(L)} (X_i - s_{ij}) + b_{0i} + b_{1i} s_{ij} + e_{ij}. \tag{3.17}$$

For the marginal expected value of the change of the response of the $i-$th subject from the $j-$th time of measurement until the time of death we get

$$\begin{aligned} E(Y_{ij} - Y_{iX_i}) &= \beta_1^{(L)}(X_i - s_{ij}) + E(b_{1i}(X_i - s_{ij})) \\ &= \beta_1^{(L)}(X_i - s_{ij}) + E(b_{1i}X_i) \end{aligned} \tag{3.18}$$

For the expectation of the right-hand side to be zero, one should assume independence of b_{1i} and X_i which may be an unjustified assumption. The joint modeling approach has the advantage of relating the survival function based on the subject specific trajectory of the subjects through a Cox model. When the association of b_{1i} and X_i is weak, our approach could be employed as an alternative that relaxes all parametric forms of the underlying distributions of the parameters imposed in the joint modeling approach. Furthermore our approach is computationally simpler compared to the joint model based approaches. Specifically, the Unweighted method can be straightforwardly be applied based on mean imputation given the covariate model. In the following chapter we will explore using the estimating function based approaches previously developed to construct time dependent ROC and evaluate the time dependent accuracy of the corresponding markers utilized in the PBC and HIV data sets.

Chapter 4

Assessing the accuracy of a marker based on GLMs: Two examples

In this chapter we explore the construction of time dependent ROC curves based on the approaches in the previous two chapters, that deal with a single baseline measurement (one measurement per subject) as well as with longitudinal marker measurements (multiple measurements per subject). We consider estimating the sensitivity and specificity with the use of generalized linear models that incorporate the censored random variable of the time to event as a covariate. Similar models have been explored in Cai et al. (2006) where an example regarding the Framingham risk score (FR-score) is discussed. They consider the FR-score as a marker for the future risk of cardiovascular events which may occur after the score is ascertained. The time-lag is taken into account since the biomarker might be more indicative of the disease when the measurement is taken closer to the time of the event. In this chapter we re-visit the PBC and HIV data to construct the corresponding time dependent ROC curves. We recall the incident based and cumulative incident based definitions of the sensitivity and FPR (see also Cai et al. (2006), Heagerty et al. (2000) and Pepe (2003)) for the PBC data set where a changepoint might be used to separate the diseased form the controls. We also explore the cumulative incident based time dependent ROC and the area under it for the HIV data set discussed in the previous chapter.

4.1 Time dependent ROC for continuous marker measurements

Recall that (see Introduction) the true positive and false positive rates are defined respectively as

$$TPR(c) = P(Y \geq c|D = 1), \;\; FPR(c) = P(Y \geq c|D = 0) \tag{4.1}$$

where c is the threshold value used to define a positive biomarker result. In the case where Y is continuous the ROC curve is plotted as the pairs of $TPR(c)$ and $FPR(c)$ for all possible threshold values of c.

The sensitivity and specificity can be extended in a situation where the true disease status is based on a time to event variable, X. The marker measurement, Y, may be

more predictive if marker measurements are taken at a timepoint closer to the event. We denote with s the time at which a measurement is taken. That is, measurements taken at time s are denoted with $Y(s)$. A distant time point τ^* is used to separate the diseased from the control group. That is, subjects that experience the event prior to τ^* are considered as cases, while subjects that manage to survive beyond τ^* are defined as the control group. Under this setting the sensitivity and FPR are defined as

$$TPR_{s,x}(c) = P(Y(s) \geq c | X - s = x), \;\; x \leq \tau^* \tag{4.2}$$

where the timelag between the time to event and the time of the measurement is $X - s = x$. This means that given that a subject experienced the event at X, the $TPR_{s,x}(c)$ is the probability of a positive marker measurement at x time units prior to the event. The false positive rate function is defined as

$$FPR_{s,\tau^*}(c) = P(Y(s) \geq c | X - s > \tau^*) \tag{4.3}$$

For these "incidence" based definitions and some additional references see Cai et al. (2006). In the simple case where $s = 0$ is considered and a linear model is used for modeling the marker measurements based on the time to event covariate X that is subject to censoring as well on other fully observed covariates $Z_1, Z_2, \ldots, Z_p$ we employ the parametric model:

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X - \tau^*)_- + \beta_2 Z_1 + \ldots + \beta_{p-1} Z_p + \epsilon \\ log(X|z) &= \xi_0 + \xi_1 Z_1 + \xi_2 Z_2 + \cdots + \xi_p Z_p + \sigma_{AFT} u \end{aligned} \tag{4.4}$$

where an AFT model is used as a submodel to account the information that the fully observed covariates carry for the censored one. The error term is assumed to be normally distributed. Under a non-parametric method we may assume the HCNS approach to model the censored covariate based on the other fully observed ones. Observe that:

$$\begin{aligned} TPR &= P(Y > c | D = 1) \\ &= P(\beta_0 + \beta_1 (X - \tau^*)_- + \beta_2 Z_1 + \ldots + \beta_{p-1} Z_p + \epsilon > c) \\ &= 1 - P(\epsilon < (c - \beta_0 + \beta_1 (X - \tau^*)_- + \beta_2 Z_1 + \ldots + \beta_{p-1} Z_p)) \end{aligned} \tag{4.5}$$

and similarly for the false positive rate we obtain

$$\begin{aligned} FPR &= P(Y > c | D = 0) \\ &= P(\beta_0 + \beta_2 Z_1 + \ldots + \beta_{p-1} Z_p + \epsilon > c) \\ &= 1 - P(\epsilon < c - \beta_0 + \beta_2 Z_1 + \ldots + \beta_{p-1} Z_p) \end{aligned} \tag{4.6}$$

For the error term one could assume normality if this is justified by the data at hand. Alternatively one may use a kernel based approach to non-parametrically estimate the density of the error term. However, this would involve imposing restrictions that would force a zero mean as we will see in the first example to follow. Note that the TPR depends on the time to event variable whereas the FPR does not. This is a logical

assumption since for the controls the marker measurement is not expected to vary over time. Heagerty et al. (2000) introduce the cumulative incidence based TPR and FPR functions, namely:

$$TPR_{s,x}^{(CI)} = P(Y(s) \geq c | X - s \leq x) \tag{4.7}$$

$$FPR_{s,x}^{(CI)} = P(Y(s) \geq c | X - s > x). \tag{4.8}$$

Note that the cumulative incidence based true and false positive rates can be evaluated by the corresponding incidence based TPR and FPR functions, and this is a reason that makes the former more attractive.

For the cumulative based definition one can derive that

$$\begin{aligned} TPR_{s,x}^{(CI)} = P(Y(s) \geq c | X - s \leq x) &= \frac{P(Y(s) \geq c, X \leq x + s)}{P(X \leq x + s)} \\ &= \frac{\int_0^{s+x} \int_c^{\infty} f_{Y|X}(u_1|u_2) f_X(u_2) du_1 du_2}{F_X(s+x)}, \end{aligned} \tag{4.9}$$

and the expression for the FPR is similarly obtained. Note again, that if common parametric models cannot be justified then we may use the HCNS approach or some other non-parametric approach to estimate the distribution of $Y|X$ and X. The corresponding ROC curves as obtained by the above two definitions, at a given time point x, are respectively:

$$ROC_{s,x} = \{FPR_{s,\tau^*}(c), TPR_{s,x}(c), c \in (-\infty, \infty)\} \tag{4.10}$$

$$ROC_{s,x}^{(CI)} = \{FPR_{s,x}^{(CI)}(c), TPR_{s,x}^{(CI)}(c), c \in (-\infty, \infty)\} \tag{4.11}$$

The generalization of the above definitions for the case where additional covariates are observed is straightforward and the ROC can be defined at a specific time point as well as at a specific covariate profile.

Under a longitudinal framework where multiple measurements per subject are available, the data for analysis are of the form $\{Y_{ij}, s_{ij}, T_i, \Delta_i, \mathbf{Z}_i\}$, where Y_{ij} is the $j-$th measurement of the $i-$th subject, s_{ij} is the time of the $j-$th measurements on the $i-$th subject, T_i is the time to event or censoring, Δ_i the corresponding event indicator ($\Delta_i = 1$ for an event) and $\mathbf{Z}_i$ is a vector of covariates that define the profile of the $i-$th subject that may affect the diagnostic accuracy of the biomarker under study.

In this case the ROC curve at a given covariate profile and a given time point is given based on the usual definition:

$$ROC_{s,x} = \{FPR_{\tau^*,s,\mathbf{Z}_i}(c), TPR_{x,s,\mathbf{Z}_i}(c), c \in (-\infty, \infty)\}$$

where the $TPR_{x,s,\mathbf{Z}_i}$ and the $FPR_{\tau^*,\mathbf{Z}_i}$ functions can either be defined using the "incident" based definitions if a distant time-point is available to separate the diseased from the controls, or the "cumulative incident" based definition in a similar fashion as previously described.

4.2 Time dependent ROC for the PBC data

Recall the PBC data set analyzed in Chapter 2 that involved subjects that participated in a randomized placebo controlled clinical trial where $log(bilirubin)$ is considered as a marker for primary biliary cirrhosis. To use Cai's definitions of TPR and FPR we now consider a "hockey stick" model for the marker measurement given the censored time to event variable, taking also into account the covariate of age (Z). As in Chapter 2, we distinguish cases and controls based on a distant time point (changepoint) that equals to 84 months and is considered fixed and known. We assume as discussed in the previous sections that the FPR function will not depend on the time to event variable, hence for this application we will explore fitting a "hockey stick" stick model. That is, after the changepoint τ^* the slope that refers to the effect of time will be set equal to zero.

Parametric modeling:

The assumed parametric model of the time to event variable given the age, $X|Z$, is an AFT generalized gamma regression model, upon which the mean imputation is based. The "hockey stick" model and the corresponding survival submodel will be of the form:

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X - \tau^*)_- + \beta_2 z + \epsilon \\ log(X|z) &= \xi_0 + \xi_1 z + \sigma_{AFT} u \end{aligned} \tag{4.12}$$

and the corresponding estimated parameters as obtained by using the optimal estimating function approach discussed in Chapter 2 are presented in Table 4.1. We remind that results for the fitted AFT model are presented in Table A7 of Appendix A. The hockey stick models are plotted for the age of 50 in Figures 4.1.

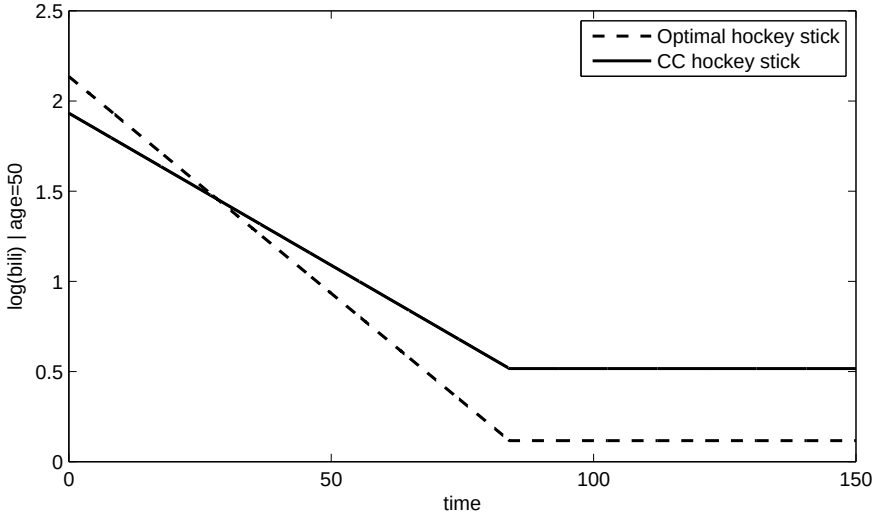

Figure 4.1: Hockey stick models for the Complete Case (CC) and the Optimal Estimating function approach for the PBC data set given that the age is equal to 50 years.

Table 4.1: Estimates of the coefficients of the hockey stick type model for the PBC data.(The estimated standard deviation based on the CC analysis is $\sigma_{(CC)} = 0.9125$)

Method	Parameter	Est.	Asympt. SE	Asympt. CI 95%	
CC	β_0	1.5354	0.5478	0.4480	2.6229
	β_1	-0.0169	0.0035	-0.0238	-0.0100
	β_2	-0.0206	0.0100	-0.0405	-0.0007
QS(GG)	β_0	1.0640	0.2649	0.5448	1.5833
	β_1	-0.0240	0.0024	-0.0287	-0.0194
	β_2	-0.0192	0.0053	-0.0296	-0.0087
QS(HCNS)	β_0	1.0624	0.2646	0.5830	1.5418
	β_1	-0.0240	0.0024	-0.0287	-0.0194
	β_2	-0.0190	0.0053	-0.0294	-0.0086

For this application we only have one baseline measurement for each subject, namely $s_{ij} = 0$, and the $log(bilirubin)$ is considered as the response. Assuming normality we obtain:

$$\hat{TPR}_{x,age}(c) = 1 - \Phi\left(\frac{c - (1.0640 + 0.0240x + 0.0192 \times age)}{0.9125}\right) \tag{4.13}$$

$$\hat{FPR}_{\tau^*,age}(c) = 1 - \Phi\left(\frac{c - (1.0640 + 0.0192 \times age)}{0.9125}\right) \tag{4.14}$$

We observe (see Figure 4.2) that the TPR function is a decreasing function of time, as expected. This means that when the time of the baseline measurement is close to event, that is the time of death, then the sensitivity of the marker is high. The sensitivity reduces essentially for larger time-lags. For example, given a positivity threshold of 1.5 the TPR at early event times is above 0.9 while near 6 years (72 months) the sensitivity reduces to approximately 0.5. Similarly the AUC is a decreasing function of time as expected (see Figure 4.3). For early times the log(bilirubin) turns out to be a very accurate marker yielding AUC above 0.95 while at near 70 months the AUC is decreased to approximately 0.6.

Non-Parametric modeling:

We also consider the HCNS approach to model the censored covariate instead of using the AFT submodel presented in 4.12. Even though strict parametric assumptions are relaxed the values of the estimated coefficients are similar when using the spline model and the GG AFT model (see Table 4.1). Even though the spline model is employed and the coefficients of the GLM are estimated with no assumptions regarding the parametric form of the distribution of the response, we need a density estimate for the response to proceed with the estimation of the time dependent ROC. Hence, to further robustify the modeling approach in this example we relax the assumption of the normal distribution assumed for the residuals. We instead employ a kernel density approach to estimate the density of the estimated residuals, $\hat{\epsilon}_i$ with $i = 1, 2, \ldots, 98$ obtained after performing the

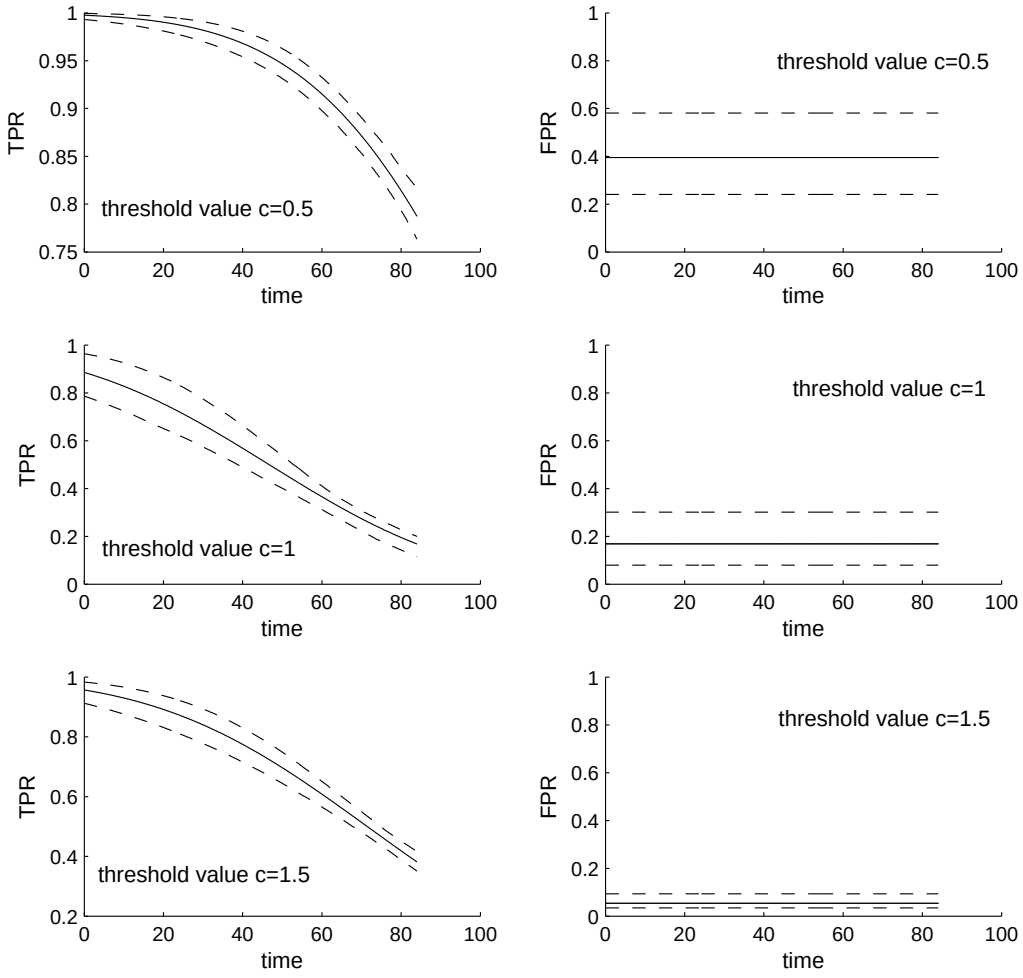

Figure 4.2: The TPR and FPR functions versus time for the PBC data given the age of 50. The positivity threshold is taken to be equal to 0.5, 1 or 1.5. The dashed lines refer to 95% confidence intervals obtained by the percentile bootstrap using 1000 bootstrap samples.

CC analysis (there are 98 fully observed data). We employ the normal kernel with the following bandwidth that is optimal for normal densities (see also Bowman and Azzalini (1997)):

$$h = \left(\frac{4}{3n}\right)^{1/5} \tilde{\sigma},$$

where $\tilde{\sigma} = median(|\hat{\epsilon}_i - median(\hat{\epsilon}_i)|)/0.6745$. Even though, based on these residuals, normality cannot be rejected (Kolmogorov Smirnov $p-$value=0.9892), we provide the ker-

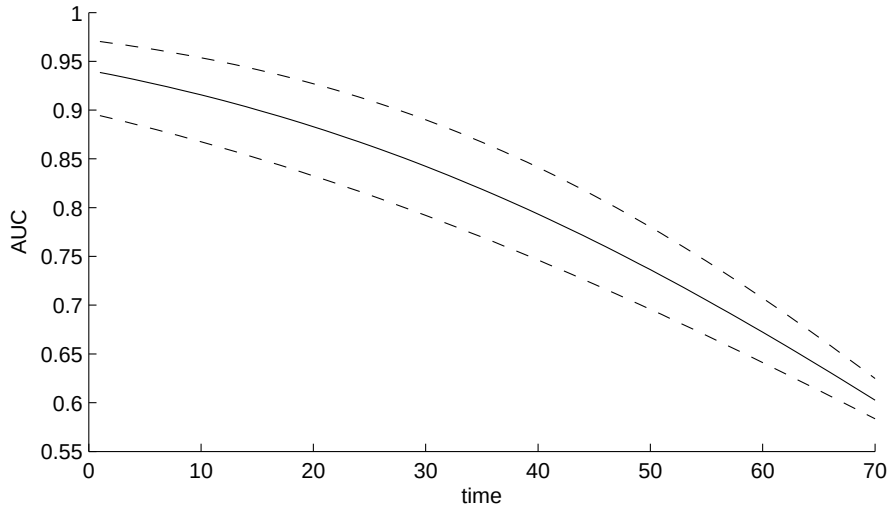

Figure 4.3: The AUC of the time dependent ROC curve for the PBC data at the age of 50. The dashed lines refer to 95% confidence intervals obtained by the percentile bootstrap. Normality is assumed for the response

nel based approach for illustration purposes. The qq-plot along with the kernel density estimate for the CC based residuals are given in Figure 4.4. We observe that the obtained kernel density estimate is very close to the standard normal density. The absolute value of the estimated kernel based mean for the residual distribution is $< 10^{-4}$. The corresponding standard deviation is estimated as 0.9776. Luckily, there is no need to apply any restrictions that would force the kernel based estimated density to have zero mean since this is intrinsically satisfied in this application. We proceeded by estimating the TPR and FPR functions over time using this estimated kernel density for the error term. The results are given in Figure 4.5. Based on this approach we do not make any strict parametric approaches at no stage of analysis since the HCNS approach is employed to model the censored covariate based on the fully observed covariate of age, and the kernel approach is used for the response. For the derivation of the confidence intervals we considered the bootstrap technique using 1000 bootstrap samples. However, we did not consider re-estimating the error distribution with the kernel approach previously discussed. A more preferable approach would involve re-estimating the error distribution for every bootstrap sample by also forcing a restriction for a zero mean, and taking censoring into account. This is a very interesting point for future research since there is no such approach available currently in the literature (to our knowledge). An exception is the paper of Hall and Presnell (1999) in which they present a weighted bootstrap based approach for density estimation with moment constraints. However, as noted by the authors, there might be cases that will result in negative weights. Furthermore, their approach cannot account for censoring. Moreover, in many cases where such an approach is to be employed for every bootstrap sample, it might be computationally very intense since this would involve a "bootstrap within a bootstrap" technique. Here, we limit our analysis by considering the estimated density based on the complete data

which we keep fixed when we apply the bootstrap. The time dependent AUC is presented in Figure 4.6. We observe that the trajectory of the AUC vs time is similar for the cases when normality is assumed, and when normality is relaxed by using the kernel approach for the response.

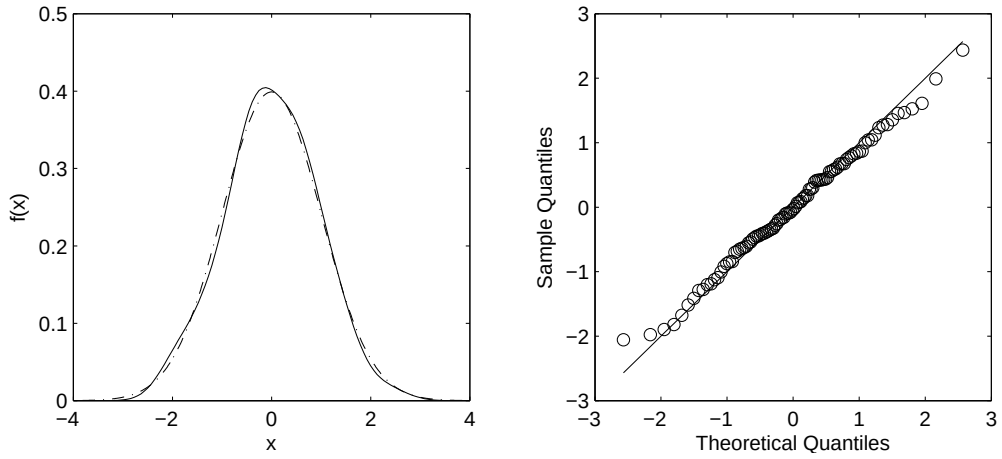

Figure 4.4: Left: Kernel density estimate of the raw residuals based on the CC method for the PBC data (solid line) vs standard normal density (dashed line). Right: The corresponding qq-plot for normality checking

4.3 Time dependent ROC for the HIV data

Recall the HIV data set discussed in the previous chapter, where marker measurements were taken repeatedly over time for each of the 467 patient. Let's focus on the covariate profile of $\mathbf{Z}$: $Drug = ddC$, $Gender = Male$, $PrevOI = AIDS$, $AZT = Intolerance$ for which there are 57 patients. See also Figure 3.5 for the subjects' trajectory over time. We desire to construct the underlying time dependent ROC curve that refers to this specific covariate profile and explore the discriminatory capability of the $CD4$ marker. Here, we use the "cumulative incident" definitions for the FPR and TPR, and consider assessing the accuracy when $s = 0$, namely we have:

$$\begin{aligned} TPR_x^{(CI)} = P(Y < c | X \leq x, \mathbf{Z}) &= \frac{P(Y < c, X \leq x | \mathbf{Z})}{P(X \leq x | \mathbf{Z})} \\ &= \frac{\int_0^x \int_0^c f_{Y|X,\mathbf{Z}}(u_1|u_2) f_{X|\mathbf{Z}}(u_2) du_1 du_2}{F_{X|\mathbf{Z}}(x)}, \end{aligned}$$

where $Y = \sqrt{CD4}$. For the estimation of the density $f_{X|\mathbf{Z}}$ we employ the HCNS approach, and assume normality for $Y|X, \mathbf{Z}$, that is we assume (based on model (3.13) and results of Table 3.1 for the covariate profile previously mentioned) that $Y|X, \mathbf{Z} \sim N((\beta_0 - \beta_7) + \beta_1^{(C)} x + \beta_1^{(L)}(x - s), \sigma^2_{(CC)})$. For the double integral that appears in the

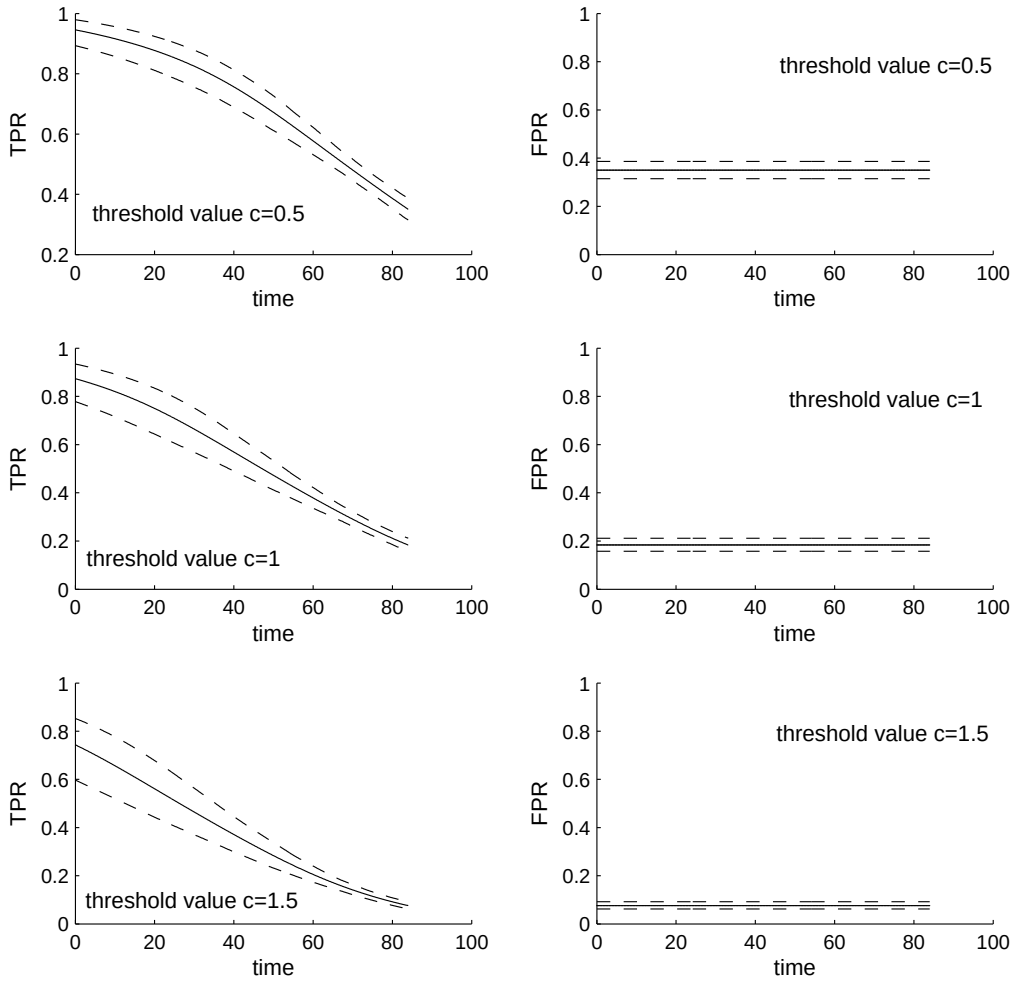

Figure 4.5: The TPR and FPR functions versus time for the PBC data given the age of 50 when the kernel density estimate is used for the error term. The positivity threshold is taken to be equal to 0.5, 1 or 1.5. The dashed lines refer to 95% confidence intervals obtained by the percentile bootstrap using 1000 bootstrap samples.

denominator of the TPR numerical integration is required and this makes the "cumulative incident" definition computationally more intense compared to the "incident" based definition employed in the previous example. Here, we do not consider a "distant" time point that defines the true health status of the subjects. Furthermore, note that all subjects are HIV patients, and one could argue that they are all cases. The ROC obtained in this example refers to the marker's capability of distinguishing at a given time point x, and a given covariate profile $\mathbf{Z}$ subjects that are able to survive beyond x and subjects who cannot. Obtaining the TPR trajectories versus time would be computationally

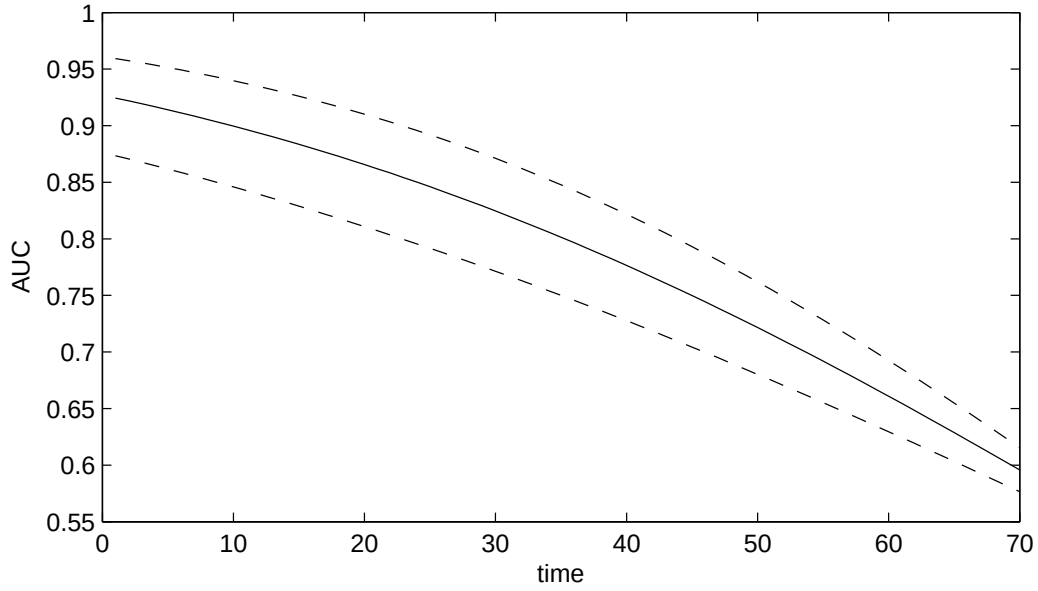

Figure 4.6: The AUC of the time dependent ROC curve for the PBC data at the age of 50. The dashed lines refer to 95% confidence intervals obtained by the percentile bootstrap. The kernel approach is employed for estimating the responce distribution.

very expensive and for this example we only consider providing the estimated ROCs at time $x = 2.5, 5, 7.5, 10$ and 12.5 months (see Figure 4.7). We observe that these ROC curves almost coincide. Furthermore the resulting AUC at the timepoints mentioned along with their bootstrap based confidence intervals are the respectively the following: 0.5919 (0.5563, 0.6911), 0.5942 (0.5511, 0.6946), 0.5962 (0.5578, 0.7028), 0.5994 (0.5667, 0.7062), 0.6049 (0.5685, 0.7132). Unfortunately, we do not observe that the discriminatory capability is better for early times. We observe that it is the same over time yielding an AUC around 0.6. This result may be related to the fact the in HIV studies it is not uncommon that subjects die not from the actual disease but due to drug side effects as mentioned in Chapter 3. However, as will be discussed in the further research issues of Chapter 6, the development of an analogous to the so called "sandwich" estimator (see Fitzmaurice et al. (2004)) for the error variance might improve estimation. Under the examined approach a working correlation matrix is assumed, the robustness of which needs to be further explored.

At this stage, one could also employ more robust approaches that would relax the normality assumption $Y|X, \mathbf{Z}$ or the underlying estimated residuals. As in the previous example, here too, this would involve developing a non-parametric technique that would accommodate the observed covariates, could account for censoring, and provided the option to impose first moment restrictions. To our knowledge, such a technique does not exist and would be of interest for further research as also discusses in Chapter 6.

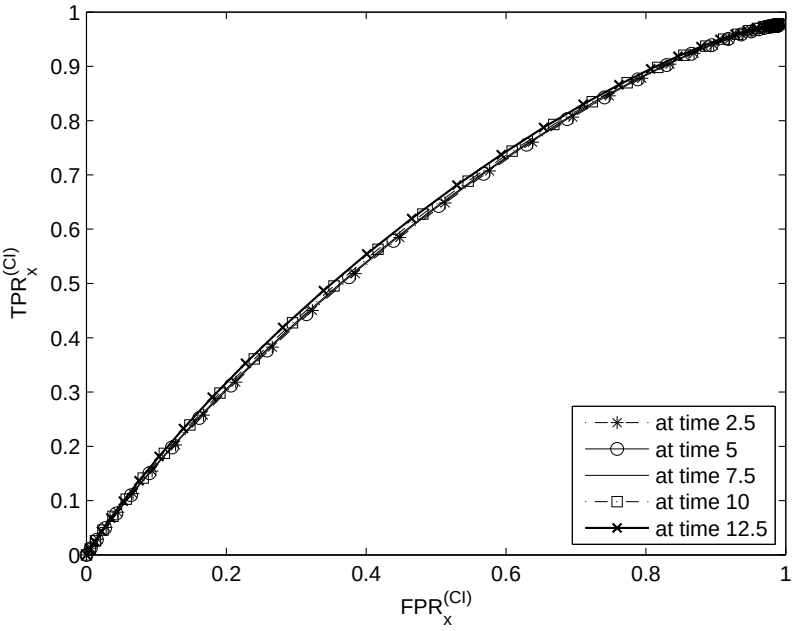

Figure 4.7: The cumulative incident based ROC curves for the HIV data at times 2.5, 5, 7.5, 10 and 12.5. The curves almost coincide yielding AUCs equal to 0.5919, 0.5942, 0.5962, 0.5994, 0.6049 respectively

Chapter 5

ROC curves and surfaces for biomarkers with a limit of detection

Until now we considered settings of time dependent biomarkers where the time to event covariate was subject to censoring. Censoring can also occur on the biomarker itself due to practical reasons regarding the mechanism/nature of the marker. Here, we consider biomarkers that yield continuous values, Y with high marker values being indicative of the presence of the disease. As mentioned in the introduction, the ROC curve may be represented through the survival functions of the diseased and the healthy populations, denoted with S_1 and S_0 respectively, as

$$\text{ROC}(t) = S_1(S_0^{-1}(t)) \tag{5.1}$$

In some studies marker measurements cannot be obtained above or below some value (d_U and d_L respectively) due to practical limitations (for an example of a biomarker with an upper limit of detection see Jafarzadeh et al. (2010)). Then, only a possibly censored version, $Y^{(c)}$, of the true marker measurement, Y, is observed. Some ad-hoc methods have been used to deal with biomarker measurements in the presence of a lower limit of detection (LOD). One approach is to set measurements that are left censored at d_L equal to the LOD value and then perform the usual ROC analysis (see Hughes (2000) for an overview of approaches regarding inference in the presence of a lower LOD). Another approach is to impute the undetectable lower values with $d_L/2$, and is based on the technique presented by Nehls and Akland (1973). This is equivalent to assuming that all values from 0 to d_L are equally likely, i.e. assuming a uniform distribution in the interval $[0, d_L]$ for the undetectable values. Another approach for data subject to a lower LOD is to impute the unobserved values by $d_L/\sqrt{2}$ (see Hornung and Reed (1990)). These simple imputation techniques attempt to decrease the bias induced by simply ignoring the censored nature of data subject to an LOD. Obviously such a technique provides a computationally simple way to deal with the problem but it suffers from some drawbacks. First, it cannot accommodate right censored data (upper LOD). Second, it assumes that the marker values are positive and this is not always the case. Even in a situation where the marker measurements are intrinsically positive, it is often the case that we work with transformations that may project the measurements to the real line. Moreover, Perkins et al. (2007) showed that any replacement value for the censored data as described above

will induce bias in the AUC index.

Some parametric approaches are also available. Mumford (2006) focused only on normally distributed biomarkers. Perkins et al. (2007), as well as Vexler et.al.(2008), focus on the use of common parametric models for the distribution of the biomarker such as the binormal and the bigamma models. These authors also consider the case of multiple biomarkers subject to left LODs. A standard parametric approach is to assume a distribution for each of the two populations and perform classical ROC analysis.

We consider data of the form $\{Y_i^{(c)}, D_i, \Delta_i\}$, $i = 1, \ldots, n$, where $Y_i^{(c)}$ are the possibly censored marker values, D_i indicates the group of the $i-$th subject, and Δ_i is the censoring indicator taking the value 0 for a censored observation and 1 otherwise. For the moment, assume that the marker values are only right censored due to an upper LOD ($d_L = -\infty$). When the parametric form of the distributions for the healthy, Y_0, and the diseased group, Y_1, is known, one can simply maximize the corresponding likelihood for each group. Let the sample be ordered with respect to the health status. For simplicity assume that the first r subjects are the healthy individuals and the remaining $n - r$ individuals belong to the diseased group. If we denote with $y_i^{(c)}$, δ_i, the realizations of the random variables $Y_i^{(c)}, \Delta_i$ then for the likelihood we have:

$$\prod_{i=1}^{r} f_0(y_i^{(c)}; \boldsymbol{\theta}_0)^{\delta_i} S_0(y_i^{(c)}; \boldsymbol{\theta}_0)^{1-\delta_i} \prod_{i=r+1}^{n} f_1(y_i^{(c)}; \boldsymbol{\theta}_1)^{\delta_i} S_1(y_i^{(c)}; \boldsymbol{\theta}_1)^{1-\delta_i}$$

where f_0 and f_1 are the densities corresponding to S_0 and S_1 with $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ being their parameter vectors respectively. Once these parameters are estimated by maximizing the above likelihood then one can construct the smooth parametric form of an ROC curve that corresponds to the two populations based on the ROC representation in (5.1). The likelihood in the case of a lower LOD is obtained in a similar fashion.

However, if a parametric assumption cannot be justified by the data at hand then a likelihood based approach may not be appropriate. Note also that if more complex parametric models are entertained, such as mixture distributions, then parameter estimation is not trivial and computational problems may arise. Under a linear regression framework where the covariate is subject to a limit of detection, Schisterman et al. (2006) use least squares estimation to determine the suitable replacement value for the non detectable values. In this chapter we employ the spline based HCNS approach discussed in the second chapter. Hence, unlike other maximum likelihood based methods that involve splines or mixtures of distributions, the proposed constrained natural spline (CNS) ROC estimate is computationally stable since it involves convex optimization.

This chapter is organized as follows. In Section 5.1 we present the CNS based ROC curve estimate and discuss how it can be generalized in the presence of covariates. In Section 3 we discuss the generalization of the proposed technique to ROC surfaces. In Section 4 we present simulation studies comparing our approach to the simple imputation techniques and the likelihood approach. We conclude with a discussion and point out some issues for future research.

5.1 CNS ROC curve/surface estimation

The CNS (constrained natural spline) estimator of the survival function $S(y)$, with K knots which is defined as

$$\hat{S}^{cns}(y) = exp(-\hat{H}^{cns}(y)),$$

with $\hat{H}^{cns}(y) = H(y;\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \ldots, \hat{\beta}_K]$ is the constrained least squares estimated vector of parameters as discussed in the second chapter. One might consider various knot placement schemes and select the one that yields the smallest distance from the corners of the nonparametric maximum likelihood step estimator, that is the one that minimizes the quantity $\sum(\hat{S}^{cns}(Y_i^{(c)}|\Delta_i = 1) - S^{KM}(Y_i^{(c)}|\Delta_i = 1))^2$, where S^{KM} is the Kaplan Meier survival estimation. Since this is a simple optimization problem, one might want to consider multiple (even hundreds) knot placement schemes. With modern computer technology this strategy is not time consuming, and the required minimizations can be easily carried out. In effect, it is feasible to base inference on resampling techniques such as the bootstrap as we will see later.

The proposed CNS estimator of an ROC curve is of the form

$$\hat{ROC}^{(cns)}(t) = \hat{S}_1^{(cns)}(\hat{S}_0^{(cns)^{-1}}(t)). \tag{5.2}$$

The AUC estimate is simply given by

$$\hat{AUC}^{(cns)} = \int_0^1 \hat{ROC}^{(cns)}(t)dt$$

The proposed estimator is based on the smooth survival estimator $\hat{S}^{cns}(y)$ for each of the two populations. This estimator expands the survival estimation through an exponential curve beyond the last censored observation, which in the setting of a censored biomarker is usually the limit of detection. This holds also for a lower LOD, which is more often the case, if the ordering of the measurements is reversed (e.g. by simply multiplying all values by -1). In this case the proposed estimator is of the form $\hat{ROC}^{(cns)}(t) = \hat{F}_1^{(cns)}(\hat{F}_0^{(cns)^{-1}}(t))$. A simulated example of the proposed estimator of an ROC curve compared to the Naive method of ignoring the censored nature of the data is given in Figure 1.

Our approach can be straightforwardly applied to the case where three populations are under study. In the classical setting the aim is to evaluate a biomarker that distinguishes between the three populations with corresponding marker distributions F_1, F_2, and F_3. In the three class case two thresholds c_1 and c_2 are used for which $c_1 < c_2$. As discussed in the introduction the three true positive rates are obtained by: $TPR_1 = P(Y < c_1|D = 1), TPR_2 = P(c_1 < Y < c_2|D = 2)$, and $TPR_3 = P(Y > c_2|D = 3)$, where D denotes the population. A three dimensional plot can then be used to visualize the ROC surface constructed in the unit cube. The parametric form of the ROC surface is given by

$$ROC(TPR_1, TPR_3) = F_2(F_3^{-1}(1 - TPR_3)) - F_2(F_1^{-1}(TPR_1)).$$

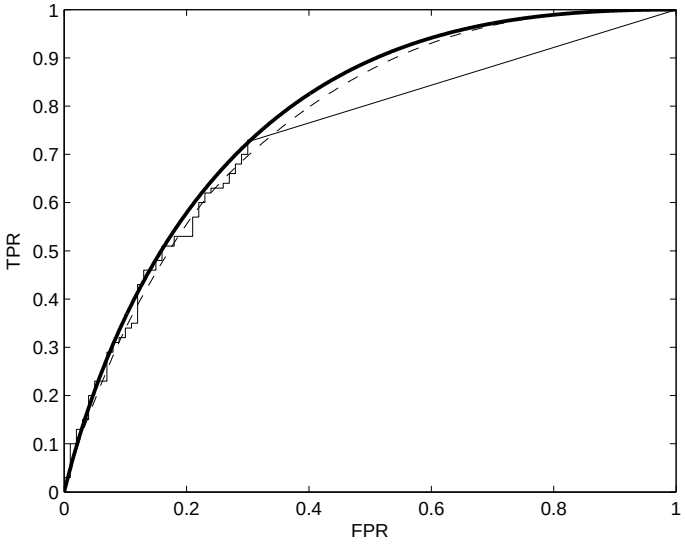

Figure 5.1: An ROC curve from a simulated data set with a lower limit of detection that yields 50% censoring where $Y_0 \sim N(3,1)$ and $Y_1 \sim N(4, 0.8^2)$. The solid bold line is the true ROC curve. The thin step solid line is the empirical ROC obtained by ignoring the censored nature of the data (Naive method). The dashed line is the ROC curve obtained by the proposed method.

In the two class case Perkins et al. (2007) showed that any replacement value below the limit of detection d_L induces bias to the AUC index. The following proposition generalizes this result also for the three class case where an ROC surface is to be constructed.

Proposition 1. *Let $Y_1 < Y_2 < Y_3$ be marker measurements from distributions F_1, F_2 and F_3 respectively and d_L be a lower limit of detection. Let a, be any replacement value that is less than the limit of detection ($a < d_L$). Denote the imputed marker scores that are constructed by imputing the left censored measurements with a, as M_1, M_2, M_3 respectively. For every $a < d_L$, the volume under the ROC surface based on the imputed marker scores (M_1, M_2, M_3) is biased.*

Proof. It can be shown that

$$\begin{aligned} VUS_M &= P(M_3 > M_2 > M_1) \\ &+ \frac{1}{2}P(M_3 = M_2 > M_1) + \frac{1}{2}P(M_3 > M_2 = M_1) + \frac{1}{6}P(M_3 = M_2 = M_1) \\ &= \int_d^\infty F_2(x) f_3(x) dx - S_3(d) \left(\int_d^\infty F_2(z) f_1(z) dz - F_1(d) F_2(d) \right) \\ &+ \frac{1}{2} S_1(d) F_2(d) F_1(d) + \frac{1}{6} F_3(d) F_2(d) F_1(d), \end{aligned}$$

which is independent of the replacement value a. The details are given in the technical notes at the end of this chapter. □

When the measurements of a trichotomous marker are available we estimate a smooth version of the ROC surface by using the CNS method to obtain a smooth estimate of the cumulative distributions, F_1^{cns}, F_2^{cns} and F_3^{cns}, of each one of the three populations, Y_1, Y_2 and Y_3, respectively. Then the proposed estimator is of the form

$$\hat{ROC}^{cns}(TPR_1, TPR_3) = \hat{F}_2^{cns}(\hat{F}_3^{cns-1}(1 - TPR_3)) - \hat{F}_2^{cns}(\hat{F}_1^{cns-1}(TPR_1)). \quad (5.3)$$

The discriminatory capability of such a trichotomous marker is summarized by the corresponding volume under the surface, and provides an estimator of the probability of a correct classification, $P(Y_1 < Y_2 < Y_3)$:

$$\hat{VUS}^{cns} = \int_0^1 \int_0^1 \hat{ROC}^{cns}(TPR_1, TPR_3) dTPR_1 dTPR_3. \quad (5.4)$$

The variance of the estimated volume under the surface following the proposed methodology, $\hat{VUS}^{cns}$, is obtained using the percentile bootstrap technique. Some graphical examples of the proposed method estimation in the case of a trichotomous marker appear in Figure 5.2 where the true setting is trinormal.

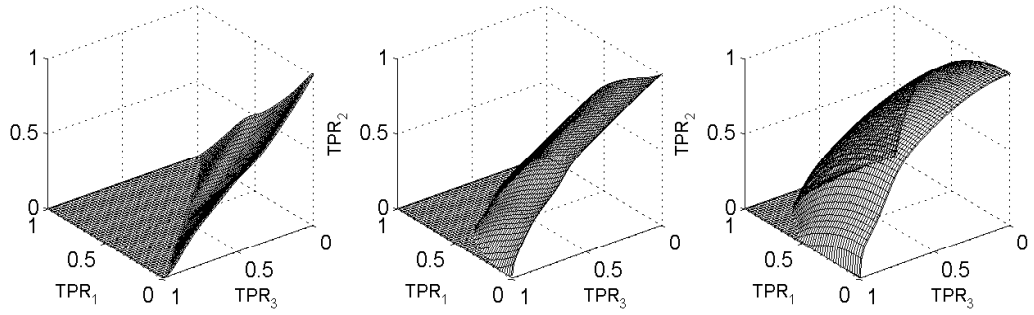

Figure 5.2: ROC^{cns} surfaces from a simulated example. An upper LOD was used so as to achieve expected level of censoring of 30%. Left: The obtained ROC^{cns} referring to $Y_1 \sim N(0,1), Y_2 \sim N(0,1), Y_3 \sim N(0,1)$. Middle: The obtained ROC^{cns} referring to $Y_1 \sim N(0,1), Y_2 \sim N(0.5,1), Y_3 \sim N(1,1)$. Right: The obtained ROC^{cns} referring to $Y_1 \sim N(0,1), Y_2 \sim N(1,1), Y_3 \sim N(2,1)$.

5.2 Adjusting for Covariates

In many settings it is natural to assume that covariates may affect the marker value or even the diagnostic ability of a marker as measured by the area under the ROC curve or the volume under the ROC surface. It is known that if one ignores the covariates then bias may be induced regarding the inference about the test accuracy (Pepe 2003). Note that these covariates may or may not be the same for each population.

Consider the two population setting. Without loss of generality assume that the covariates for the healhty and the diseased group are the same. Denote the covariates

with $Z_1, Z_2, \ldots, Z_p$. Using matrix notation the covariates can be denoted more compactly as the matrix $\mathbf{Z}$ with each column representing one covariate. The ROC curve is then defined as

$$\text{ROC}(t) = S_{1:\mathbf{Z}}(S^{-1}_{0:\mathbf{Z}}(t)), \tag{5.5}$$

where $S_{i:\mathbf{Z}}(t) = S(t|D = i, \mathbf{Z}) = P(Y \geq t|D = i, \mathbf{Z})$. To construct an ROC curve in the presence of covariates we investigate a two stage approach. At the first stage, in order to assess the influence of the covariates on each one of the groups we explore the use of a proportional hazards model of the form

$$S_{j:\mathbf{Z}}(t) = S_{j:\mathbf{Z}=\mathbf{0}}(t)^{exp(\mathbf{Z}\boldsymbol{\gamma}_j)}, j = 0, 1. \tag{5.6}$$

where $\boldsymbol{\gamma}_0$ is the parameter vector $[\alpha_1, \alpha_2, \ldots, \alpha_p]'$ that relates the covariates with the marker measurements for the healthy group, and $\boldsymbol{\gamma}_1$ is the parameter vector $[\gamma_1, \gamma_2, \ldots, \gamma_k]'$ that relates the covariates with the marker measurements for the diseased group.

The proportional hazards model (also known as the Cox model) is primarily used in survival analysis, and the target is to model survival time, that is subject to censoring, in the presence of other covariates.

In our setting, we aim to derive a consistent estimator of $S_{j:\mathbf{Z}}(t)$. Suppose for simplicity (and without loss of generality) that we only have one covariate that takes two values (0 or 1) and its corresponding coefficient is γ_1. Model (5.6) implies that, given that a subject belongs to group i, if $exp(\gamma_1) < 1$ then the probability of getting a measurement greater than a threshold value c is higher for a subject for which $Z = 1$, that is $P(Y > c|Z = 1, D = j) > P(Y > c|Z = 0, D = j)$. Similarly, if $exp(\gamma_1) > 1$ then $P(Y > c|Z = 1, D = j) < P(Y > c|Z = 0, D = j)$. The baseline functions $S_{j:\mathbf{Z}}(t), j = 0, 1$ are completely unspecified. In the case of a three way analysis then three Cox models are required in order to derive an ROC surface (in this case $j = 1, 2, 3$).

The second stage of our approach deals with smoothing the derived survival functions of the first stage. We fit the constrained natural spline to the baseline step function of the cumulative hazard, $H(y|\mathbf{Z} = 0, D = j) = -log(S(y|\mathbf{Z} = 0, D = j))$, obtained via fitting a Cox model for each j. Thus we aim to minimize the following functions

$$\Psi(\boldsymbol{\beta}^{(j)}) = \sum_i (H(Y_i^{(c)}|\mathbf{Z} = \mathbf{0}, D_i = j, \Delta_i = 1) - H^{KM}(Y_i^{(c)}|\mathbf{Z} = \mathbf{0}, D_i = j, \Delta_i = 1)^2, \tag{5.7}$$

where $j = 0, 1$ (for a three way analysis $j = 1, 2, 3$), and $\boldsymbol{\beta}^{(j)} = [\beta_1^{(j)}, \ldots, \beta_K^{(j)}]'$ are the spline coefficients referring to group j, under the constraints

$$\mathbf{A}[\beta_1^{(j)}, \beta_2^{(j)}, \ldots, \beta_{K-1}^{(j)}]' \leq \mathbf{0}.$$

Thus, for the construction of an ROC curve we deal with minimizing two functions under linear constraints. Note again that the minimizations refer to convex optimization that can be applied with standard software. Under this approach we can naturally derive the survival function for any profile of covariate values based on the Cox model's formulation:

$$\hat{S}^{cns}_{j:\mathbf{Z}}(t) = (\hat{S}^{cns}_{j:\mathbf{Z}=\mathbf{0}})^{exp(\mathbf{Z}\hat{\boldsymbol{\gamma}}_j)}, j = 0, 1 \tag{5.8}$$

and construct the corresponding ROC curve

$$\hat{ROC}^{cns}_{\mathbf{Z}}(t) = \hat{S}^{cns}_{1:\mathbf{Z}}(\hat{S}^{cns-1}_{0:\mathbf{Z}}(t)). \tag{5.9}$$

The corresponding estimate for the area under the curve is

$$\hat{AUC}^{cns}_{\mathbf{Z}} = \int_0^1 \hat{ROC}^{cns}_{\mathbf{Z}}(t)dt \tag{5.10}$$

and is a measure of the discriminatory capability of the marker for the diseased and healthy group given the profiles of the two populations based on their covariate values.

Similarly, the proposed two stage approach for constructing an ROC surface involves the fit of three Cox models in the first stage, and at the second stage the derivation of the ROC surface:

$$\begin{aligned} \hat{ROC}^{cns}_{\mathbf{z}}(TPR_1, TPR_3) &= \hat{F}^{cns}_{2:\mathbf{Z}}(\hat{F}^{cns-1}_{3:\mathbf{Z}}(1 - TPR_3)) \\ &- \hat{F}^{cns}_{2:\mathbf{Z}}(\hat{F}^{cns-1}_{1:\mathbf{Z}}(TPR_1)), \end{aligned} \tag{5.11}$$

where $\hat{F}^{cns}_{j:\mathbf{Z}}(t) = 1 - \hat{S}^{cns}_{j:\mathbf{Z}}(t), j = 1, 2, 3$. And the corresponding estimation for the volume under the surface is

$$\hat{VUS}^{cns} = \int_0^1 \int_0^1 \hat{ROC}^{cns}_{\mathbf{Z}}(TPR_1, TPR_3)dTPR_1dTPR_3. \tag{5.12}$$

This is a measure of the discriminatory capability of the marker for group 1, 2, and 3 given the profiles of interest.

It may be the case that the proportional hazards assumption that allows the use of the Cox model in (5.8) may not be justified for marker data. For example if $log(Y)$ follows a simple linear regression model with normal errors (in survival settings this is a lognormal AFT model) then it is known that the proportional hazards assumption does not hold for the distribution of the marker value. In our approach the proportional hazards assumption can be relaxed by allowing an interaction with some function $g(y)$. This is also the approach taken when modeling time dependent covariates in a survival setting (see also Klein and Moeschberger (2003)). Thus the cox model can be written as

$$S_{j:\mathbf{Z}}(t) = exp\left(-\int_{-\infty}^{y} h_0(u)exp(\boldsymbol{\gamma}'\mathbf{Z}g(u))du\right) \tag{5.13}$$

where h_0 is completely unspecified and $g(y)$ is a known function. When the marker measurements are positive supported then one can consider $g(y) = log(y)$ or $g(y) = \sqrt{y}$. When the marker measurements can also take negative values then one can consider

$g(y) = y$. These are popular choices used in survival settings. A more flexible approach would be to use an interaction with polynomials or splines. After fitting model (5.13) the estimate $\hat{\boldsymbol{\gamma}}$, as well as the baseline cumulative hazard function $\hat{H}_0$ can be obtained. At the second stage model (2.14) is fitted on the cumulative step baseline hazard function $\hat{H}_0$ and the $\hat{H}_0^{cns}$ is obtained. Given a profile of a subject the covariate adjusted survival function can be estimated as

$$\hat{S}_{j:\mathbf{Z}}(t) = exp\left(-\int_{\tau_1}^{y} \hat{h}_0^{cns}(u) exp(\hat{\boldsymbol{\gamma}}'\mathbf{Z}g(u))du\right) \tag{5.14}$$

where $\hat{h}_0^{cns}(y)$ is the first derivative based on model (2.14) with respect to y. Note that we integrate from τ_1 because model (2.14) equals to zero before the first knot. The construction of the ROC curve or surface, for any given profile of a subject, is straightforward.

5.3 Simulation Studies

We conducted some simulation studies to evaluate the proposed method and compare it with the $d_L/\sqrt{2}$, $d_L/2$ and d_L imputation schemes as well as with the likelihood approach. The naive approach of proceeding with the empirical ROC curve/surface by ignoring the censored nature of the data is also considered. Note that the imputation based approaches were considered only in the left censoring scenarios. In all simulations presented in this chapter we examined sample sizes of $n_i = 100$, and $n_i = 200$ where $i = 1, 2$ for an ROC curve and $i = 1, 2, 3$ for an ROC surface. Simulation results tables that refer to the ROC curve with equal sample sizes considered for the two populations are Table 5.1, and Tables D1 and D2 of Appendix D regarding scenarios based on the normal, the gamma and the non-central t distribution respectively as we will see right next. For the corresponding simulation results in the three class case see Table 5.2 and Tables D3 and D4 of Appendix D. We also considered scenarios with unequal sample sizes where $n_1 = 100$, $n_2 = 300$, the results of which are presented in the Appendix D (see D5, D6, D7 for unequal sample size scenarios involving normals, gammas and non-central t distributions respectively). The limit of detection was selected to achieve $10\%, 30\%$ and 50% expected levels of censoring (these percentages refer to the total sample size (both populations)). The tables that refer to our simulation results also show in parentheses the two (or three) population specific censoring levels obtained in each simulation. As expected the likelihood method is superior when the underlying distributional assumptions hold.

We explored the use of six knots for the proposed spline based technique and consider the following scheme regarding their location: We consider 10 equally spaced points between the first and last fully observed marker measurements. (The first and last fully observed marker measurements are included in this set of 10 points). These points are candidates for placing the knots. Thus, there are $10!/(6!4!) = 210$ combinations of knot placement schemes to choose from. The scheme that is finally selected, is the one that yields the smallest distance from the corners of the nonparametric maximum likelihood step estimator, that is the one that minimizes the quantity $\sum(\hat{S}^{cns}(Y_i^{(c)}|\Delta_i = 1) - \hat{S}^{KM}(Y_i^{(c)}|\Delta_i = 1))^2$, where $\hat{S}^{KM}$ is the Kaplan Meier survival estimation.

Initially we performed some simulations regarding the ROC curve using the AUC as a comparison criterion since this index is usually of most interest. First, we considered a binormal scenario where $Y_0 \sim N(3,1)$, $Y_1 \sim N(4,0.8^2)$. We looked at left as well as right censoring. In the presence of left censoring and for moderate censoring levels the proposed approach compared to the naive and the simple imputation approaches yielded minor differences in terms of mean squared error (MSE) of the estimated AUC. However, in the cases where the expected level of censoring is 50% we observe essential differences in terms of MSE due to the essentially lower bias that is provided by the proposed approach. The likelihood approach yielded somewhat better results, as expected, since it assumes the correct model for each of the two populations. The results were similar for the cases when the measurements where generated from two gamma distributions ($Y_1 \sim Gamma(25,0.2)$, $Y_2 \sim Gamma(35,0.2)$). The results of the binormal are presented in Table 5.1. The results that correspond to the bi-gamma scenario are presented in the Appendix D (see D1). The likelihood approach in these two cases assumed the correct models for the two distributions. We also performed a simulation where the two populations were generated from two non-central t distributions ($Y_1 \sim t(4,7)$ and $Y_2 \sim t(5,10)$) where the likelihood approach falsely assumed normality for the two groups. These distributions are heavy tailed (see Appendix D, Figure D1). However, in the presence of an upper limit of detection the tails are in the undetectable region and hence one could falsely assume normality in such a scenario. The likelihood approach seems to be fairly robust in this setting (see Appendix D). However, in a setting where the distributions of the two groups are bimodal, or a common parametric model cannot be assumed, computational problems may occur during the maximization of the likelihood (for example identifiability problems in the case of mixture distributions). For the derivation of confidence intervals we considered the percentile bootstrap technique in all methods (i.e. the proposed, the naive and the simple imputation methods). In all cases of heavier censoring, the proposed method yielded markedly better coverage compared to its competitors (apart from the likelihood approach when it assumes the correct models for the two groups).

One may argue that 50% censoring is unlikely to occur. The merit of our approach is however evident when one considers the partial AUC, even with low censoring rates. It is often the case, that clinicians focus their interest on a specific range of FPR values. We also explored the estimation of the partial area under the curve for $0.8 \leq FPR \leq 1$, in the case of a lower LOD for the binormal scenario mentioned above with unequal sample sizes ($n_1 = 100$, $n_2 = 300$). The expected total censoring was set at 10%. The coverage, as obtained by the percentile bootstrap, using the imputation schemes $d_L/\sqrt{2}$, $d_L/2$ and d_L is respectively 0.9080, 0.8690 and 0.0240. The proposed (CNS) approach and the likelihood approach yielded coverages of 0.9490 and 0.9370 respectively.

We note that in the case of an uninformative or almost uninformative biomarker the naive method might be preferred since as it is easier to compute. In that case the linear extension of the ROC curve from the FPR point that corresponds to the lower limit of detection, to the point (1,1) may provide an adequate approximation to the true ROC. This might even be the case when the true ROC beyond (or prior in the case of an upper LOD) the FPR point that corresponds to the LOD is approximately linear. However, the true ROC is not known in practice and use of the naive method is not to be preferred. The other simple imputation approaches still have the merit of being computationally easier

than our approach but may not perform as well (particularly in terms of coverage).

We also conducted simulations for the three class case for the same sample sizes for each population and the same expected levels of censoring. Again, we considered three scenarios. In the first scenario the measurements were generated from three normal distributions ($Y_1 \sim N(5,1)$, $Y_2 \sim N(6,1)$, $Y_3 \sim N(7,1)$). In the second scenario we consider a tri-gamma setting where $Y_1 \sim Gamma(25, 0.2)$, $Y_2 \sim Gamma(35, 0.2)$, and $Y_3 \sim Gamma(45, 0.2)$. In both of these cases the likelihood approach assumes the correct models for estimation of the distribution parameters. In the third scenario marker measurements were generated from three non central t distributions, $Y_1 \sim t(4,7)$, $Y_2 \sim t(5,10)$ and $Y_3 \sim t(6,12)$ (the figure of the densities of all all distribution is given in Appendix D (see D2)). In this setting the likelihood approach falsely assumed normality for each of the distributions. The conclusions from these simulations are the same as in the two class case (Table 5.2 and Tables D3, D4 of Appendix D). Note again, that for higher censoring, the coverage of the bootstrap based confidence intervals is essentially better when using the proposed approach compared to the corresponding obtained by the simple imputation techniques or the naive method.

Other knot placement schemes may also be employed. In most applications involving splines, knots are placed at equally spaced percentiles (see Harrell (2001) for example). However, in our case restrictions of monotonicity are imposed. Moreover the CNS cumulative hazard function is forced to be zero before the first knot. Hence, it may be necessary that more knots need to be placed near the minimum measurement value of each group so that the proposed splines achieve enough flexibility. This is allowed by providing 10 equally spaced points at which the six knots can be placed. A criterion that could account for the sample size and the number of knots would be of great interest and could be considered as future research. The simulations for the two class case with unequal sample sizes and 6 knots for each spline are given in Appendix D. Another issue might be the way the initial points (that the candidate knots are to be placed) are spread or how many initial points should be considered. In Appendix D we also provide the results of simulations that were carried out, using 5 or 7 knots for the two population and a lower LOD scenarios (see Appendix D , Tables D8, D9 and D10).

For the estimation of the ROC and the AUC or VUS computational time is not an issue. For the construction of the ROC curve, when the sample size of each population equals to 200, 6 knots are used and the level of censoring is 30% the required CPU time is approximately 2 seconds with a 2.8GHz processor. The corresponding time for an ROC surface is about 5 seconds. For the calculation of the AUC or VUS numerical integration is required. However, the bootstrap procedure is a more intensive task and the computational time depends on how many bootstrap samples are to be used for inference. The computational time needed can be found with multiplication of the mentioned required times for estimation, with the number of bootstrap samples.

5.4 Application

To demonstrate our method we use a liver cancer data set generated by the surface-enhanced laser desorption/ionization (SELDI) time of flight mass spectrometer. The serum samples of the liver cancer data were taken at Shanghai Chang-zheng Hospital, China. There were three groups to discriminate: hepatoma (H) patients, chronic liver

Table 5.1: Simulation results for 1000 repetitions for the bi-normal scenario. The likelihood approach assumes the correct model for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True AUC equals to 0.7826)

					AUC		
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_0 : 19.1%, Y_1 : 0.9%)	$Likelihood$	-0.0005	0.0332	0.0011	0.9330
			$Naive$	-0.0015	0.0340	0.0012	0.9530
			$d_L/\sqrt{2}$	-0.0001	0.0337	0.0011	0.9340
			$d_L/2$	-0.0020	0.0325	0.0011	0.9360
			d_L	-0.0026	0.0355	0.0013	0.9330
			CNS	0.0006	0.0351	0.0012	0.9500
		30% (Y_0 : 50%, Y_1 : 10%)	$Likelihood$	-0.0006	0.0342	0.0012	0.9330
			$Naive$	-0.0123	0.0346	0.0014	0.9440
			$d_L/\sqrt{2}$	-0.0066	0.0369	0.0014	0.9340
			$d_L/2$	-0.0048	0.0367	0.0014	0.9360
			d_L	-0.0289	0.0374	0.0022	0.8720
			CNS	-0.0015	0.0352	0.0012	0.9550
		50% (Y_0 : 71%, Y_1 : 29%)	$Likelihood$	-0.0025	0.0387	0.0015	0.9440
			$Naive$	-0.0502	0.0346	0.0037	0.6750
			$d_L/\sqrt{2}$	-0.0351	0.0385	0.0027	0.8290
			$d_L/2$	-0.0320	0.0390	0.0025	0.8540
			d_L	-0.0696	0.0369	0.0062	0.4800
			CNS	-0.0025	0.0440	0.0019	0.9630
	$n = 200$	10% (Y_0 : 19.1%, Y_1 : 0.9%)	$Likelihood$	-0.0008	0.0227	0.0005	0.9450
			$Naive$	-0.0015	0.0229	0.0005	0.9490
			$d_L/\sqrt{2}$	0.0001	0.0230	0.0005	0.9440
			$d_L/2$	-0.0017	0.0222	0.0005	0.9410
			d_L	-0.0025	0.0243	0.0006	0.9460
			CNS	-0.0001	0.0233	0.0005	0.9510
		30% (Y_0 : 50%, Y_1 : 10%)	$Likelihood$	-0.0010	0.0235	0.0006	0.9440
			$Naive$	-0.0125	0.0236	0.0007	0.9470
			$d_L/\sqrt{2}$	-0.0068	0.0254	0.0007	0.9320
			$d_L/2$	-0.0050	0.0254	0.0007	0.9420
			d_L	-0.0291	0.0255	0.0015	0.7960
			CNS	-0.0012	0.0241	0.0006	0.9490
		50% (Y_0 : 71%, Y_1 : 29%)	$Likelihood$	-0.0021	0.0265	0.0007	0.9560
			$Naive$	-0.0503	0.0235	0.0031	0.4560
			$d_L/\sqrt{2}$	-0.0352	0.0261	0.0019	0.7440
			$d_L/2$	-0.0321	0.0264	0.0017	0.7820
			d_L	-0.0700	0.0255	0.0055	0.1770
			CNS	0.0006	0.0292	0.0009	0.9460
Right Censoring	$n = 100$	10% (Y_0 : 4%, Y_1 : 16%)	$Likelihood$	-0.0005	0.0336	0.0011	0.9340
			CNS	0.0025	0.0345	0.0012	0.9320
		30% (Y_0 : 14%, Y_1 : 46%)	$Likelihood$	-0.0010	0.0352	0.0012	0.9420
			CNS	0.0020	0.0357	0.0013	0.9420
		50% (Y_0 : 29%, Y_1 : 71%)	$Likelihood$	-0.0027	0.0403	0.0016	0.9450
			CNS	0.0085	0.0453	0.0021	0.9500
	$n = 200$	10% (Y_0 : 4%, Y_1 : 16%)	$Likelihood$	-0.0007	0.0228	0.0005	0.9520
			CNS	0.0012	0.0231	0.0005	0.9460
		30% (Y_0 : 14%, Y_1 : 46%)	$Likelihood$	-0.0008	0.0234	0.0005	0.9540
			CNS	0.0017	0.0241	0.0006	0.9490
		50% (Y_0 : 29%, Y_1 : 71%)	$Likelihood$	-0.0019	0.0271	0.0007	0.9520
			CNS	0.0091	0.0300	0.0010	0.9460

Table 5.2: Simulation results for 1000 repetitions for the tri-normal scenario. The likelihood approach assumes the correct model for the three populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True VUS equals to 0.5362)

				VUS			
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_1 : 29.4%, Y_2 : 4.7%, Y_3 : 0.4%)	$Likelihood$	0.0008	0.0344	0.0012	0.9290
			$Naive$	-0.0015	0.0351	0.0012	0.9490
			$d_L/\sqrt{2}$	-0.0070	0.0351	0.0013	0.9280
			$d_L/2$	-0.0183	0.0361	0.0016	0.9160
			d_L	-0.0041	0.0342	0.0012	0.9360
			CNS	0.0061	0.0367	0.0014	0.9610
		30% (Y_1 : 61.6%, Y_2 : 24%, Y_3 : 4.4%)	$Likelihood$	0.0001	0.0376	0.0014	0.9350
			$Naive$	-0.0226	0.0339	0.0017	0.8920
			$d_L/\sqrt{2}$	-0.0229	0.0345	0.0017	0.8840
			$d_L/2$	-0.0374	0.0342	0.0026	0.8020
			d_L	-0.0306	0.0326	0.0020	0.8460
			CNS	0.0014	0.0436	0.0019	0.9760
	$n = 200$	10% (Y_1 : 29.4%, Y_2 : 4.7%, Y_3 : 0.4%)	$Likelihood$	-0.0007	0.0234	0.0005	0.9470
			$Naive$	-0.0027	0.0238	0.0006	0.9530
			$d_L/\sqrt{2}$	-0.0080	0.0236	0.0006	0.9430
			$d_L/2$	-0.0195	0.0242	0.0010	0.8840
			d_L	-0.0049	0.0234	0.0006	0.9450
			CNS	0.0018	0.0245	0.0006	0.9520
		30% (Y_1 : 61.6%, Y_2 : 24%, Y_3 : 4.4%)	$Likelihood$	-0.0008	0.0262	0.0007	0.9440
			$Naive$	-0.0234	0.0230	0.0011	0.8330
			$d_L/\sqrt{2}$	-0.0231	0.0233	0.0011	0.8550
			$d_L/2$	-0.0375	0.0231	0.0019	0.6400
			d_L	-0.0309	0.0227	0.0015	0.7330
			CNS	0.0031	0.0285	0.0008	0.9570
Right Censoring	$n = 100$	10% (Y_1 : 0.37%, Y_2 : 4.7%, Y_3 : 24.93%)	$Likelihood$	0.0009	0.0340	0.0012	0.9350
			CNS	0.0054	0.0369	0.0014	0.9500
		30% (Y_1 : 4.4%, Y_2 : 24%, Y_3 : 61.6%)	$Likelihood$	-0.0003	0.0382	0.0015	0.9340
			CNS	0.0002	0.0428	0.0018	0.9720
	$n = 200$	10% (Y_1 : 0.37%, Y_2 : 4.7%, Y_3 : 24.93%)	$Likelihood$	-0.0007	0.0236	0.0006	0.9410
			CNS	0.0020	0.0245	0.0006	0.9490
		30% (Y_1 : 4.4%, Y_2 : 24%, Y_3 : 61.6%)	$Likelihood$	-0.0017	0.0257	0.0007	0.9540
			CNS	0.0024	0.0279	0.0008	0.9600

disease (LD) patients, and normal individuals (No) with sample sizes 54, 39 and 52 respectively. Wang and Chang (2011) studied a procedure of marker selection for these data using pairwise analysis. They propose a wrapper-type algorithm for selecting the best linear combination of markers that has high TPR rate within a specificity range. We will explore the discriminatory capability of marker 4271.37 which was ranked in the top five of markers in the study of Wang and Chang (2011).

Since there are three groups for classification, an ROC surface approach would be preferable for the evaluation of such a marker. We explored both an ROC surface analysis as well as an ROC curve analysis. The ROC curve analysis was conducted after merging the hepatoma patients with chronic liver disease patients to a single 'diseased' group. This is in line with the results of the Kolmogorov-Smirnov test for the equality of the H and LD marker distributions ($p-$value=0.239), as well as with the results of the Mann-Whitney test ($p-$value=0.131). For discriminating the 'diseased' from the non-diseased the empirical ROC curve yields an AUC equal to 0.8563 (see also Figure D3 of Appendix D). Our CNS estimate yielded an AUC equal to 0.8898 (see Table 5.3). The two ROC curve estimates in Appendix D. The empirical ROC surface as well as the CNS estimated surface are shown in Figure 5.3. The empirical VUS equals to 0.4563 and the CNS based volume is 0.4796. The empirical survival estimates along with the CNS survival estimates are shown in the same figure. In Appendix D we provide the projections of the ROC surface needed for pairwise analysis (see Figure D4 of Appendix D). These projections are equivalent to ROC curves even though it is not the sensitivity against the false positive rate that is plotted. However, the interpretation is similar with higher AUC values indicating better pairwise discriminatory capability of the marker. Thus, using the ROC surface approach provides the merit of evaluating the marker simultaneously for all three groups without foregoing a pairwise analysis. A discussion is also provided in Yannoutsos et.al. (2008). All results regarding the three way analysis along with the corresponding confidence intervals are shown in Table 5.3.

To investigate the performance of our method in this particular application in the presence of an upper (lower) LOD we censored the marker values. For the ROC curve analysis we used LOD values to achieve 30% and 50% censoring. For the ROC surface analysis we used LOD values to achieve 10% and 30% censoring. Higher censoring was not investigated for ROC surfaces since it is very likely that marker measurements for a specific group may all end up censored. The results and the bootstrap based confidence intervals are presented in Table 5.3. The corresponding ROC curves are presented in Figures 5.4, and 5.5 for the cases of no censoring as well as the cases where a lower detection limit was used. In the case of right censoring we compared our approach with a d_U replacement technique to see how estimates change as censoring increases. We observe that the 'naive' approach yielded an essentially reduced estimate in the case of 50% censoring whereas our estimate does not seem to be very sensitive to the level of censoring. In the presence of left censoring the naive approach yielded similar results to the proposed one. This is probably due to the almost linear part of the ROC curve at high TPR values. Figure 5.6 shows the obtained CNS ROC surfaces for 10% and 30% right censoring along the obtained survival curves which are extrapolated.

In Table 5.4 we present results derived from the projections of the corresponding ROC surfaces in the unit cube. Bootstrap based confidence intervals are also reported for each pair of the health status. Even though the pair LD-H yields an AUC estimate above 0.5,

Table 5.3: AUC and VUS estimates for the liver data. The 95% confidence intervals are derived using the percentile bootstrap using 500 bootstrap samples (LCL and UCL are lower and upper confidence limits respectively).

AUC:					
Direction	Censoring	Method	Estimation	95% LCL	95% UCL
No censoring	0%	Empirical	0.8563	0.7866	0.9205
		CNS	0.8898	0.7992	0.9492
	30%	$Naive$	0.8507	0.7879	0.9129
$Right$		CNS	0.8594	0.7892	0.9265
	50%	$Naive$	0.7964	0.7302	0.8565
		CNS	0.8503	0.7426	0.9531
	30%	$Naive$	0.8478	0.7798	0.9161
$Left$		CNS	0.8513	0.7485	0.9190
	50%	$Naive$	0.8511	0.7763	0.9160
		CNS	0.8606	0.7830	0.9228
VUS:					
No censoring	0%	Empirical	0.4563	0.3497	0.5683
		CNS	0.4796	0.3733	0.5913
	10%	$Naive$	0.4522	0.3498	0.5635
$Right$		CNS	0.4477	0.3488	0.5758
	30%	$Naive$	0.4315	0.3295	0.5333
		CNS	0.4447	0.3226	0.5594
	10%	$Naive$	0.4534	0.3564	0.5509
$Left$		CNS	0.4514	0.3548	0.5674
	30%	$Naive$	0.4474	0.3509	0.5486
		CNS	0.4658	0.3366	0.5755

the confidence intervals indicate a non-informative biomarker for this pair. This is also the case when one uses the naive approach. This is consistent with the results of the non-parametric tests for the equality of these distributions mentioned above. Overall, the marker seems to discriminate well between the healthy and each one of the remaining groups, yielding somewhat better performance for the $H - No$ pair, as expected from the survival curves shown in Figure 5.3. This is also evident by looking at the projections of the ROC surfaces that allow visualization of a pairwise analysis (see Figure D5 of Appendix D).

5.5 Discussion

In this chapter we considered a method of constructing an ROC curve or surface in the case of a biomarker with a limit of detection. The proposed approach is a spline based one and allows exponential extrapolation of the survival function when multiple measurements that are left (or right) censored pile up at the limit of detection. Usual ROC analysis can then be performed using the proposed CNS estimates for the survival functions of each population.

Table 5.4: AUC estimates for the liver data based on the projections of the ROC surface on the sides of the unit cube. The 95% confidence intervals are derived using the percentile bootstrap using 500 bootstrap samples (LCL and UCL are lower and upper confidence limits respectively).

Direction	Censoring	Pair	AUC Estimation	95% LCL	95% UCL
			CNS		
		LD-No	0.8713	0.7322	0.9385
	No censoring	LD-H	0.5954	0.4705	0.7177
		H-No	0.9035	0.8204	0.9598
		LD-No	0.8314	0.7234	0.9273
$Right$	10% censoring	LD-H	0.5852	0.4849	0.7193
		H-No	0.8736	0.7957	0.9420
		LD-No	0.8222	0.6899	0.9067
	30% censoring	LD-H	0.5938	0.4809	0.7093
		H-No	0.8840	0.8097	0.9424
		LD-No	0.8372	0.7517	0.9337
	10% censoring	LD-H	0.5845	0.4478	0.6959
$Left$		H-No	0.8744	0.8046	0.9400
		LD-No	0.8279	0.7394	0.9244
	30% censoring	LD-H	0.6035	0.4338	0.7057
		H-No	0.8726	0.7818	0.9349
			IMPUTATION		
		LD-No	0.8343	0.7465	0.9103
	No censoring	LD-H	0.5921	0.4810	0.7123
		H-No	0.8722	0.7999	0.9409
		LD-No	0.8314	0.7414	0.9162
$Right$	10% censoring	LD-H	0.5926	0.4684	0.6982
		H-No	0.8738	0.8006	0.9370
		LD-No	0.8247	0.7340	0.9122
	30% censoring	LD-H	0.5933	0.4651	0.6911
		H-No	0.8741	0.7954	0.9348
		LD-No	0.8348	0.7490	0.9152
	10% censoring	LD-H	0.5878	0.4696	0.7004
$Left$		H-No	0.8723	0.7883	0.9370
		LD-No	0.8346	0.7401	0.9083
	30% censoring	LD-H	0.5668	0.4608	0.6797
		H-No	0.8691	0.7869	0.9252

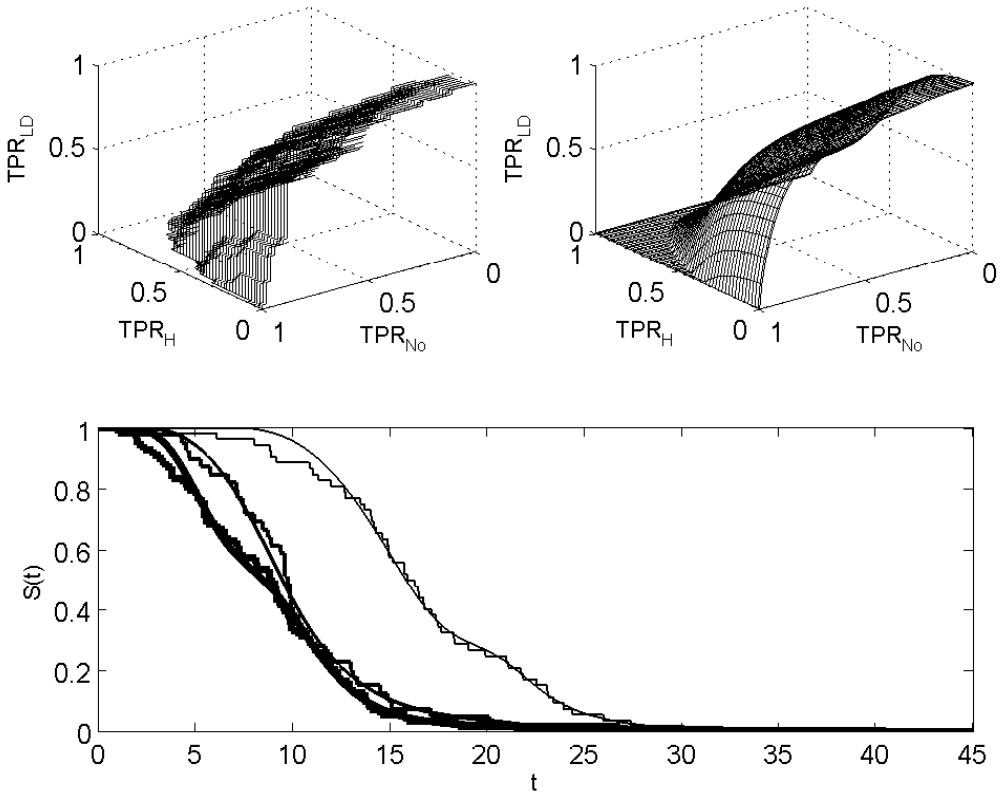

Figure 5.3: Empirical and CNS ROC surface estimates for the liver data. Down: Kaplan Meier and CNS estimates for the three survival functions (one for each group H, LD and No, from left to right respectively).

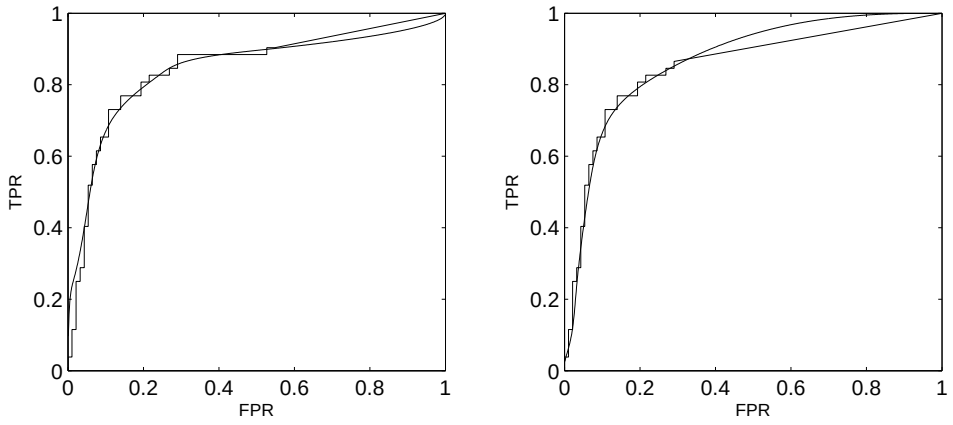

Figure 5.4: ROC curves for the liver data when an upper limit of detection is chosen to censor the measurements. Left: 30% censoring. Right: 50% censoring.

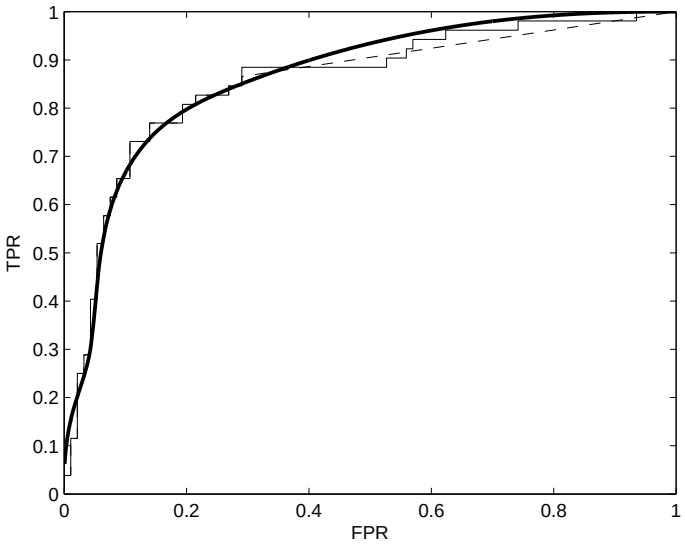

Figure 5.5: ROC curves when the H and LD group are combined to single 'diseased' group plotted together for 0% and 50% left censoring. Thin line: Empirical ROC with using all available data. Dashed line: The empirical ROC when 50% of the data are left censored (Naive approach). Thick line: CNS estimate when 50% of the data are left censored.

Unlike traditional methods that consider imputation of a single value for observation below the lower limit of detection the proposed approach provides a smooth estimate of the ROC curve or surface. The proposed approach provides a flexible way of estimating the underlying survival distributions, yet with no computational problems since it involves least squares problems with linear restrictions. The function to be minimized is always convex and the procedure can be applied with standard software. The corresponding algorithms have already been developed for such optimization tasks by some packages (we used `lsqlin` of MATLAB). Furthermore, the inverse of the survival function based on the proposed approach can be derived in a closed form, since it is the real root of a cubic polynomial. The inversion of the survival function might not be such a simple task when a common parametric model is not justified by the available data. Our method can be generalized to take into account information carried by covariates. This is made under the formulation of a Cox model for biomarker measurements.

Simulations that were performed under various scenarios for AUC (or VUS) estimation showed satisfactory performance of the proposed method yielding in some cases differences in terms of bias (and MSE) from its naive and imputation based competitors. The confidence intervals for the AUC (or VUS) that was obtained by the percentile bootstrap technique yielded values close to the nominal level. This was not always the case for the naive and the simple imputation techniques. Our method was also compared to the maximum likelihood approach assuming the correct models for the populations and we observed small disfferences in terms of MSE. In some cases, the proposed approach turned out to be more efficient when the parametric assumption was misspecified.

When left censoring is present (left censoring is more common than right censoring) the proposed approach can be promising in settings where the ROC region of high FPR

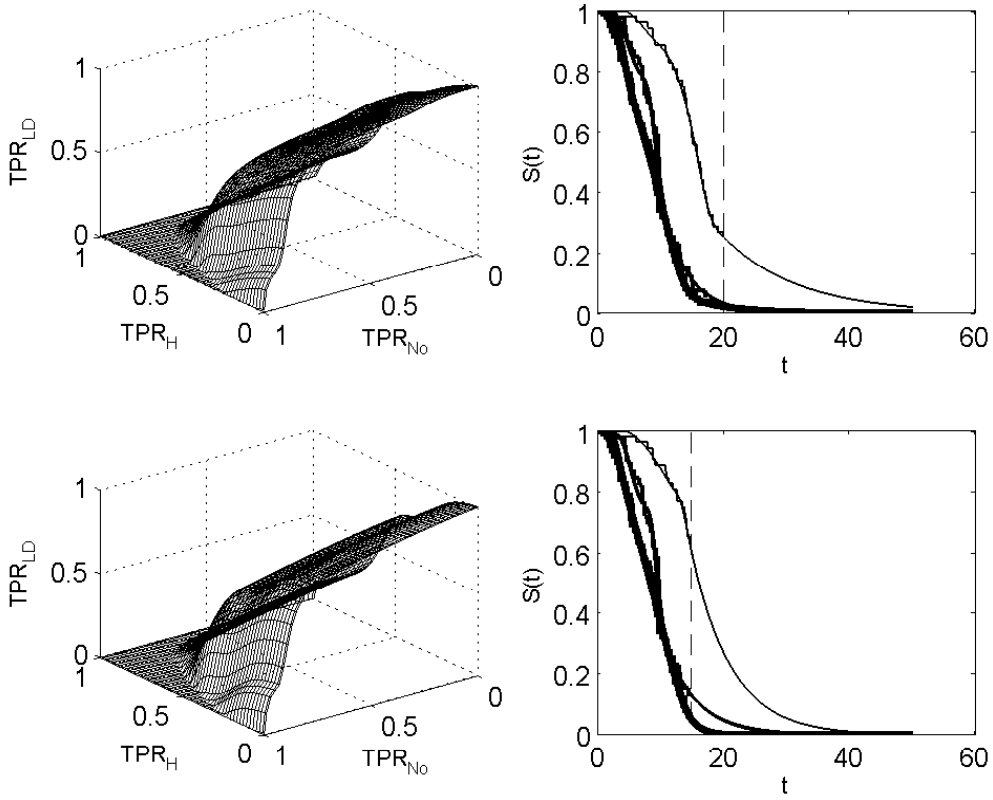

Figure 5.6: CNS ROC surface estimates for the liver data when an upper limit of detection is used. Up: CNS ROC surface estimate with a limit of detection that causes a level of 10% censoring along with the three survival estimates (KM and CNS). It is shown how the cns survival estimates are extended beyond the limit of detection (vertical dashed line). Down: ROC surface estimate with a limit of detection that causes a level of 30% censoring.

rates is of interest. For example, the assessment of the discriminatory capability of PSA-related biomarkers could involve investigation of the ROC curve for higher FPR's (see Miyakubo et al. (2009)). Exploring the partial area of an ROC curve in its high sensitivity part may be more reasonable in some cases, since it may lead to avoiding unnecessary procedures. Another example is given in Jiang et al. (1996) who study a partial AUC index in the case of highly sensitive diagnostic tests. They use a mammography example to indicate the need of focusing on the high sensitivity region of an ROC curve. Screening mammography demands high sensitivity which can contribute to reducing mortality of women with breast cancer. In this case, we desire high sensitivity because women with false negative tests cannot take advantage of early detection and treatment (see Kopans (1985)). As seen in our simulation studies the proposed approach is expected to be more efficient than the naive and the simple imputation techniques.

We note here, that these replacement value imputation methods should not be con-

fused with the multiple imputation based techniques, mostly used in the concept of missing data. The latter should be further explored in the setting of a biomarker with an LOD. However, some strict parametric assumptions should also be imposed, as in the maximum likelihood approach. A recent strategy based on multiple imputation under a regression framework is proposed by Arunajadai and Rauh (2012). They use a 7 stage multiple imputation based procedure, following the ideas of Rubin (2004) for a setting where the LOD refers to the covariate of a regression model. Their approach might be adapted to the ROC setting.

Another interesting case may regard biomarkers applied on patients in different locations (centers). If the LOD of the biomarker differs from center to center then the censored data will not be piled up at one limit of detection d, but spread out at the center-specific limits of detection. In such a case one still has the knowledge (a-priori) which measurement is censored and which is exactly observed for the same reason described above (the LOD is known a-priori for each center). Our approach can accommodate such data with no modifications.

5.6 Technical Notes

5.6.1 Bias of the VUS using a replacement value $a < d$

We drop the subscript of d_L for convenience ($d_L = d$). Consider that we use a value a, $a < d$, to impute the values that are left censored. The marker values are obtained by:

$$M_{3_i} = \begin{cases} Y_{3_i}, & \text{if } Y_{3_i} \geq d \\ a, & \text{if } Y_{3_i} < d \end{cases}$$

$$M_{2_i} = \begin{cases} Y_{2_i}, & \text{if } Y_{2_i} \geq d \\ a, & \text{if } Y_{2_i} < d \end{cases}$$

$$M_{1_i} = \begin{cases} Y_{1_i}, & \text{if } Y_{1_i} \geq d \\ a, & \text{if } Y_{1_i} < d. \end{cases}$$

The volume under the surface based on the imputed values is:

$$\begin{aligned} VUS_M &= P(M_3 > M_2 > M_1) \\ &+ \frac{1}{2}P(M_3 = M_2 > M_1) + \frac{1}{2}P(M_3 > M_2 = M_1) \\ &+ \frac{1}{6}P(M_3 = M_2 = M_1). \end{aligned}$$

Let

$$P(M_3 > M_2 > M_1) = A,$$

$$P(M_3 > M_2 = M_1) = B,$$

$$P(M_3 = M_2 > M_1) = C,$$

$$P(M_3 = M_2 = M_1) = D.$$

We derive:

$$\begin{aligned} P(M_3 > M_2 > M_1) &= P(Y_3 > Y_2 > Y_1, Y_3 \geq d, Y_2 \geq d, Y_1 \geq d) \\ &+ P(Y_3 > Y_2 > a, Y_3 \geq d, Y_2 \geq d, Y_1 < d). \end{aligned}$$

We also derive:

$$\begin{aligned} A &= P(Y_3 > Y_2 > Y_1, Y_3 \geq d, Y_2 \geq d, Y_1 \geq d) + P(Y_3 > Y_2 > a, Y_3 \geq d, Y_2 \geq d, Y_1 < d) \\ &= P(Y_3 > Y_2 > Y_1, Y_3 \geq d, Y_2 \geq d, Y_1 \geq d) + P(Y_3 > Y_2 > d, Y_1 < d) \\ &= \int_d^\infty \int_d^\infty P(y_1 < Y_2 < y_3) f(y_3, y_1) dy_3 dy_1 + \int_d^\infty P(d < Y_2 < y_3) f(y_3) dy_3 F_1(d) \\ &= \int_d^\infty \int_d^\infty (F_2(y_3) - F_2(y_1)) f_3(y_3) f_1(y_1) dy_3 dy_1 + \int_d^\infty (F_2(y_3) - F_2(d)) f_3(y_3) dy_3 F_1(d) \\ &= \int_d^\infty F_2(y_3) f_3(y_3) S_1(d) dy_3 - \int_d^\infty F_2(y_1) f_1(y_1) S_3(d) dy_1 \\ &+ \left(\int_d^\infty F_2(y_3) f_3(y_3) dy_3 - \int_d^\infty F_2(d) f_3(y_3) dy_3 \right) F_1(d) \\ &= \int_d^\infty F_2(y_3) f_3(y_3) dy_3 - \int_d^\infty F_2(y_1) f_1(y_1) S_3(d) dy_1 - F_1(d) \int_d^\infty F_2(d) f_3(y_3) dy_3 \\ &= \int_d^\infty F_2(y_3) f_3(y_3) dy_3 - S_3(d) \left(\int_d^\infty F_2(y_1) f_1(y_1) dy_1 - F_1(d) F_2(d) \right). \end{aligned}$$

Note that this probability is independent of a. It can be shown that $B = 0$ and for C and D we have respectively:

$$\begin{aligned} C &= P(M_3 > M_2 = M_1) \\ &= P(Y_3 > Y_2 = Y_1, Y_3 \geq d, Y_2 < d, Y_1 < d) \\ &= P(Y_3 > a = a, Y_3 \geq d, Y_2 < d, Y_1 < d) \\ &= P(Y_3 > a, Y_3 \geq d, Y_2 < d, Y_1 < d) \\ &= P(Y_3 \geq d, Y_2 < d, Y_1 < d) \\ &= S_3(d) F_2(d) F_1(d) \end{aligned}$$

$$D = P(M_3 = M_2 = M_1) = P(Y_3 < d, Y_2 < d, Y_1 < d) = F_3(d) F_2(d) F_1(d).$$

Thus,

$$\begin{aligned} VUS_M &= A + \frac{1}{2}B + \frac{1}{2}C + \frac{1}{6}D \\ &= \int_d^\infty F_2(y_3) f_3(y_3) dy_3 - S_3(d) \left(\int_d^\infty F_2(y_1) f_1(y_1) dy_1 - F_1(d) F_2(d) \right) \\ &+ \frac{1}{2} S_3(d) F_2(d) F_1(d) + \frac{1}{6} F_3(d) F_2(d) F_1(d). \end{aligned}$$

Note that if we set $d = -\infty$ we derive that $VUS_M = VUS$ (as expected):

$$\begin{aligned}
VUS_M &= \int_{-\infty}^{\infty} F_2(y_3) f_3(y_3) dy_3 - \int_{-\infty}^{\infty} F_2(y_1) f_1(y_1) dy_1 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_2(y_3) f_3(y_3) f_1(y_1) dy_3 dy_1 - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_2(y_1) f_3(y_3) f_1(y_1) dy_3 dy_1 \\
&= \int_{-\infty}^{\infty} \int_{d}^{\infty} (F_2(y_3) - F_2(y_1)) f_3(y_3) f_1(y_1) dy_3 dy_1 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(y_1 < Y_2 < y_3) f(y_3, y_1) dy_3 dy_1 \\
&= P(Y_1 < Y_2 < Y_3) \\
&= VUS.
\end{aligned}$$

Thus, since the probabilities A, B, C, and D are independent of the value a, we conclude that the volume under the ROC surface VUS_M will be biased for any replacement value $a < d$.

Chapter 6

Summary and further research

Diagnostic marker evaluation is a developing field. Strategies that deal with modeling the marker measurements based on covariates have been proposed by many authors (see Pepe (2003) and the references provided therein). A typical setting would involve the use of a generalized linear model that could relate the predictors with the biomarker as a response. Based on such a model one could then proceed to constructing the corresponding ROC for a specific covariate profile. Settings that involve the evaluation or modeling of a biomarker in the presence of censoring have been recently developed (see Cai et al. (2006), Heagerty et al. (2000)). Censoring is typically present when a time to death (or more generally a time to event) variable is involved. When fatal diseases are under study, it is desired to model the biomarker values based on the time to death variable. It is expected that marker values taken closer to the event are more indicative.

Maximum likelihood and semi-parametric approaches have been proposed to model such time-dependent biomarkers (see Austin and Hoch (2004), Cai et al. (2006), Heagerty et al. (2000)). In this thesis Estimating Function approaches are proposed that do not assume any parametric form for the distribution of the biomarker values and can accommodate a censored covariate. Settings that involve biomarker values that are taken repeatedly over time are also included and discussed. The repeated measurements setting needs to be explored through simulation studies and compared with the joint modeling approach when parametric assumptions are satisfied or violated for the latter. Particularly when binary data are involved and ALR type models are employed the computational cost must be explored and evaluated. A parametric model might be used to model the censored covariate, however a spline based approach that relaxes strict parametric assumptions can also be used for this purpose. We propose a new spline based approach for survival estimation that has the appealing property of always converging since convex optimization is involved, unlike the splines fitted by maximum likelihood.

Even though in most cases there is only one censored variable available along with other fully observed variables for analysis, there might be cases that two or more censored variables are available. Hence it would be interesting to see how our methods can be extended to such a setting when building a generalized linear regression model. In these cases the covariance between the two (or more) censored variables must be addressed. This might involve obtaining a smooth version of the bivariate distribution of the two censored covariates by extending our spline approach to account for bivariate survival

functions (or even higher dimensions). This might involve smoothing the bivariate Kaplan Meier survival function introduced by Dabrowska (1988). As a further extension, the same setting could be studied when measurements are taken repeatedly over time. In that case, the covariance of the two censored variables as well as the longitudinal nature of the data must be taken into account. Another interesting point would be to consider cases where apart from the censored covariate, the response values are also censored. This could be observed in a setting where modeling a biomarker that is subject to a lower (or an upper) limit of detection and also depends on a time to event variable. Further exploration is also required in different types of censoring such as the more general interval censoring. In our estimating function approach interval censoring is a case that can only be accommodated only if at least some data are exactly observed, since the dispersion parameter, as well as the correlation parameters, are estimated based solely on the fully observed data. The modification of our method that would accommodate this kind of censoring when all data are interval censored is challenging since it might involve the extension of the estimating equations to also account for the dispersion parameter. Furthermore, the exploration/construction of an analogous to the so called "sandwich" variance estimator (see Fitzmaurice et al. (2004)) would be of great interest. This would involve defining the appropriate residuals and studying their properties.

As seen in Chapter 4, the derivation of the time dependent sensitivity might imply the use of a parametric model for the underlying residuals. It would be of great interest (particularly under a longitudinal framework) to develop a non-parametric density estimation technique that could accommodate the censored nature of the data and could also provide the option to impose equality restrictions for the mean (and maybe the variance). The extension of our HCNS approach under this notion might be a computationally challenging task since these equality constraints are not linear with respect to the spline parameters and need to be employed simultaneously with the monotonicity constraints.

Apart from cases where time to event variables are present, the censoring phenomenon may appear in the biomarker itself. The most common setting is the case of a lower limit of detection (LOD) where due to technical limitations measurements cannot be taken below some limit. Some simple imputation approaches have been employed to deal with such settings however our simulation studies have shown may be inefficient. Maximum likelihood approaches have also been proposed (see Perkins et al. (2007), Vexler et al. (2008)) but strict parametric assumptions must be imposed. We prove that in the three class case bias is invoked in estimating the VUS when simple replacement values are used and we explore our spline approach to construct the ROC curve (or surface in the three class case). Our approach does not require any strict parametric assumption and is shown to be more efficient than simple replacement value based approaches via simulations. The proposed approach can also accommodate right censored data as well as marker values that lie on the real line, unlike the simple replacement value methods which can only deal with let censored biomarker values and positive valued biomarkers.

Regarding the HCNS approach and its use for ROC curve/surface estimation further research could focus on more sophisticated knot placement strategies and methods of selecting the number of knots. Another point of interest for further study might be the use of the disease status variable, D, as a covariate in the Cox model used in section 5.2. Building such a model could allow testing whether the populations are stochastically ordered. This approach is taken in Gonen and Heller (2010) in which the Lehmann family

or ROC curves is studied (see also Lehmann (1953)). When the available data consist of a possibly censored marker and the disease status, they consider fitting a cox model where the disease status plays the role of a covariate. If another covariate is also present they consider also an interaction term with the disease status. Under this formulation the proposed method would involve fitting the constrained spline model to the baseline survival and thus force a stochastic ordering of the population distributions.

It is worth mentioning that the approaches introduced in this thesis are not limited to settings in the field of biostatistics. In many cases econometricians are asked to model variables where income plays the role of a covariate, and in most studies income is right censored above some specific value. Environmetrics is another field that the approaches introduced in this thesis may apply. For example wind speed might be subject to a lower limit of detection and might play the role of a covariate in forecasting wildfire hazard (see also Chang (2007)). Another example might refer to the field of entomology where wind speed might affect fly activity (see Steelman et al. (1993)).

Bibliography

Armstrong, B. (1985). Measurement error in generalized linear models *Communications in Statistics-Simulation and Computation* **14**, 529-544.

Arunajadai, S.G., Rauh, V.A. (2012). Handling covariates subject to limits of detection *Environ Ecol Stat* **19**, 369-391.

Austin, P. C. and Hoch, J. S. (2004). Estimating linear regression models in the presence of a censored independent variable. *Statistics in Medicine* **23**, 411-429.

Azzalini A. (1980). A note on the estimation of a distribution function and quantiles by a kernel method *Biometrika* **68**, 326-328.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the Receiver Operating Characteristic Graph *J. Math Psychol* **12**, 387-415.

Bantis, L.E., Tsimikas, J.V. and Georgiou, S.D. (2011). Survival estimation through the cumulative hazard with constrained natural cubic splines. *Lifetime Data Analysis* **18(3)**, 364-396.

Bantis, L.E., Tsimikas, J.V. and Georgiou, S.D. (2011). Smooth ROC curves and surfaces for markers subject to a limit of detection using monotone natural cubic splines. *Biometrical Journal (to appear)*.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report, University of California, Berkeley*.

Biomarker Definitions Working Group (2002). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* **69,** 89-95.

Bowman A. W. and Azzalini, A. (1997). Smoothing techniques for data analysis *Oxford University Press*.

Brown, J. B. W., Hollander, M. and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. *In Reliability and Biometry: Statistical Analysis of Lifelength* **1**, 327-354.

Cai, T., Pepe, M. S., Lumley, T., Zheng, Y., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7,** 187-197.

Carey, V., Zeger, S.L., Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions *Biometrika* **80(3),** 517-526.

Carroll, R. J., and Stefanski L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* **85,** 652-663.

Chang, C-H. (2007). *Separability testing for point processes with covariates and an application to wildfire hazard assessment* Ph.D. Thesis, University of California, Los Angeles.

Colburn, W.A. (2003). Biomarkers in drug discovery and development. From target identification through drug marketing. *Journal of Clinical Pharmacology* **43,** 329-341.

Cox, C., Haitao, C., Schneider, M., F., and Mūnoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine* **26,** 4352-4374.

Cox D. R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life . *Statistics in Medicine* **5,** 307-314.

Dabrowska, M.D. (1988). Kaplan-Meier Estimate on the plane. *Annals of Statistics* **16(4)**, 1475-1489.

Doksum, K. and Yandell, B. S. (1982). Properties of regression estimates based on censored survival data. *In: A Festschrift for Erich Lehmann edited by P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr. Wadsworth International Group, Belmont, CA.* **4**, 831-853.

Dreiseitl, S., Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis *Med Decis Making* **20(3)**, 323-331.

Efron, B. (1967). The two sample problem with censored data. *In Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability, New York: Prentice-Hall* **4**, 831-853.

Fitzmaurice G. M., Laird, N. M, Ware, J. H. (2004). Applied longitudinal analysis. *Wiley Series in Propability and Statistics.*

Fleming, T. R., and Harrington, D. P. (1990). Counting processes and survival analysis. *Wiley Series in Propability and Statistics.*

Fritsch F. N., Carlson, R. E. (1980). Monotone piecewise cubic interpolation *SIAM Journal of Numerical Analysis* **2**, 238-246.

Gentleman, R. and Crowley, J. (1991). Graphical methods for censored data *Journal of the American Statistical Association* **86**, 678-683.

Gill, R. D. (1980). Censoring and stochastic integrals. *Mathematical Centre Tracts.* **124**

Gomes O., Combes C., Dussauchoy A. (2008). Parameter estimation of the generalized gamma distribution. *Mathematics and Computers in Simulation* **79,** 955-963.

Grundy, S., Balady, G., Criqui, M., Fletcher, G., Greenland, P., Hiratzka, L., Houston-Miller, N., Kris-Etherton, P., Krumholz, H., LaRosa, J., et al. (1998). Primary intervention of coronary heart disease: guidance from Framingham-a statement for healthcare professionals from the AHA Task Force on risk reduction. *Circulation* **97**, 1876-1887.

Gõmez, G., Espinal, A., and Lagakos, S. W.(2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine* **22,** 409-425.

Gonen, M., Heller, G. (2010). Lehmann family of ROC curves. *Medical Decision Making* **30**, 509-517.

Green, D.M., Swets, J.A. (1966). *Signal detection theory and psychophysics* Wiley, New York.

Hall, P., and Presnell, B. (1999). Density estimation under constraints, *Journal of Computational and Graphical Statistics* **8**, 259-277.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve *Radiology* **143(1)**, 29-36.

Hardin, J.W., Schmiediche, and Carroll R.J., (2003). The regression calibration method for fitting generalized linear models with additive measurement error *The Stata Journal* **4**, 361-372.

Harrell, F. E. (2001). Regression modeling strategies (with applications to linear models, logistic regression, and Survival Analysis). *Springer Series in Statistics*

Heagerty, P.J., Lumley, T. and Pepe, M.S. (2000). Time-Dependent ROC Curves for censored survival data and a diagnostic Marker. *Biometrics* **4**, 337-344.

Herndon, E. J. II, Harrell F. E. Jr. (1995). The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in Medicine* **14**, 2119-2129.

Heyde, C. C. (1997). Quasi-Likelihood and its application: A general approach to optimal parameter estimation. Springer-Verlag New York.

Hornung, W. R., Reed, D.L. (1990). Estimation of average concentration in the presence of nondetectable values *Applied Occupational and Environmental Hygiene* **5**, 46-51.

Hsiao, C. (1983), Regression Analysis with a categorized explanatory variable. *Studies in Econometrics, Time Series and Multivariate Statistics* edited by S. Karlin, T Anemiya, and L. Goodman. New York: Academic Press

Hughes, M.D.(2000). Analysis and design issues for studies using censored biomarker neasurements with an example of viral load measurements in HIV clinical trials. *Statistics in Medicine* **19** 3171-3191.

Jafarzadeha, S.R., Johnson, W.O., Utts, J.M. and Gardnera, I.A. (2010). Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. *Statistics in Medicine* **29**, 2090-2106.

Jiang, Y., Metz, E.C., Nishikawa, R.M. (1996). A Receiver Operating Characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201**, 745-750.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Kardaun, O. (1983). Statistical analysis of male larynx cancer patients. A Case Study. *Statistical Nederlandica* **37**, 103-126.

Klein, J. P. (1991). Small sample moments of some estimators of the variance of the Kaplan Meier and Nelson Aalen Estimators. *Scandinavian Journal of Statistics* **18**, 333-340.

Klein, J. P., and Moeschberger, M. L. (2003). Survival Analysis, Techniques for censored and truncated data. *Springer Verlag*.

Kooperberg, C., Stone, J. C., Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78-94.

Kooperberg, C., and Stone, J. C. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**, 301-328.

Kooperberg, C., Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis* **12**, 327-347.

Kopans, D.P. (1993). Mammography screening for breast cancer. *Cancer* **72**, 1809-1812.

Krzeski, P., Zych, W., Kraszewska, E., Milewski, B., Butruk, E., and Habior, A. (2003). Is serum bilirubin concentration the only valid prognostic marker in primary biliary cirrhosis?. *Hepatology* **30,** 865-869.

Kuchenhoff, H. (1997). An exact algorithm for estimating breakpoints in segmented generalized linear models. *Computational Statistics* **12,** 235-247.

Lawless F. J. (2003). Statistical models and methods for lifetime Data. *Wiley Series in Probability and Statistics*.

Lee T., and Wang J. (2003). Statistical methods for survival data analysis (third edition). *John Wiley & Sons*.

Leisenring, W., Pepe, M.S., and Longton, G.L. (1997). A marginal regression modeling framework for evaluating medical diagnostic tests. *Statistics in Medicine* **16**, 1263-1281.

Lehmann, E.L. (1953). The power of rank tests. *Ann. Math. Stat.* **24**, 23-43.

Liew, K. C. (1976). Inequality constrained least squares estimation. *Journal of the American Statistical Association* **71**, 746-751.

Link, C. L. (1984). Confidence intervals for the survival function using Cox's proportional hazards model with covariates. *Biometrics* **40**, 601-610.

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science* **171**, 1217-1219.

Mâsse, R. B., Truong, K. Y. (1999). Conditional logspline density estimation. *The Canadian Journal of Statistics* **27**, 819-832.

McCullagh, P., and Nelder, J. A. (1983). Generalized Linear Models (second edition). Chapman & Hall.

McCulloch, C. E., and Searle, S. R. (2001). Generalized, Linear, and Mixed Models. *John Wiley & Sons*.

Miyakubo, M., et al. (2009). Proprostate-specific antigen: Its usefulness in the era of multiple-core prostate biopsy. *International Journal of Urology* **16(6)**, 561-565.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making* **19**, 78-89.

Mumford, S. L. (2006). Pooling biospecimens and limits of detection: Effects on ROC curve analysis. *Biostatistics* **7**, 585-598.

Nakas, C.T., Yannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* **23**, 3437-3449.

Naylor, S. (2005). Overview of biomarkers in disease, drug discovery and development *Drug Discovery World Spring* 21-30.

Nehls, G.J., Akland, G.G. (1973). Procedures for handling aerometric data. *Journal of Air Pollution Control Association* **23**, 180-184.

Palella, F.J. Jr, Baker, R.K., Moorman, A.C., Chmiel, J.S., Wood, K.C., Brooks, J.T., Holmberg, S.D. (2006). Mortality in the highly active antiretroviral therapy era: changing causes of death and disease in the HIV outpatient study. *J Acquir Immune Defic Syndr* **43(1)**, 27-34.

Pan, W. (2000). Smooth Estimation of the Survival Function for interval censored data. *Statistics in Medicine* **19**, 2611-2624.

Pawitan, Y., and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* **88,** 719-726.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction* Oxford University Press.

Perkins, N.J., Schisterman, E.F., and Vexler, A. (2006). A Receiver Operating Characteristic Curve inference from a sample with a limit of detection *Americal Journal of Epidemiology* **165(3)** 325-333.

Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in failure time regression models. *Biometrika* **69** 331-342.

Rigobon, R., and Stoker, T. (2007). Estimation with censored regressors (Basic Issues). *International Economic Review* **40,** 1441-1467.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-Event data. *Biometrics* **67(3),** 819-829.

Rubin, D (2004). Multiple imputation for nonresponse in surveys. *John Wiley and Sons Inc, Hoboken*.

Schluchter, M. D., Konstan, M. W., and Davis, P. B. (2002). Jointly modeling the relationship between survival and pulmonary function in cystic fibrosis patients. *Statistics in Medicine* **21**, 1271-1287.

Silverman, B. W. (1986). Density estimation for statistics and data Analysis. *Monographs on Statistics and Applied Probability, London: Chapman and Hall*

Smith, L. B. (1970). Drawing ellipses, hyperbolas or parabolas with a fixed number of points and maximum inscribed area. *The Computer Journal* **14**, 81-86.

Stablein, D. M. and Koutrouvelis, I. A. A (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored Data. *Biometrics* **41**, 643-652.

Stacy, E.W. (1962). A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics* **33(3)**, 1187-1192.

Steelman, C.D., Gbur, E.E., Tolley, G., and Brown, A.H.Jr (1993). Variation in population density of the face fly, Musca autumnalis De Geer, Among Selected Breeds of Beef Cattle. *J. Agric. Entomol.* **10(2)**, 97-106.

Stone, C. J. (1990). Large sample inference for log-Spline models. *The Annals of Statistics* **18**, 717-741.

Stone, C. J. and Koo, C.-Y. (1986). Logspline density estimation. *AMS Contemporary Mathematics Series* **29**, 1-15.

Stone, C. J., and Koo, C.-Y. (1985). Additive splines in Statistics. *In Proceedings of the Statistical Computing Section ASA*, 45-48.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated Data. *Journal of the Royal Statistical Society (Series B)* **38**, 290-295.

Tsiatis, A., Davidian, M.(2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809-834.

Tsimikas, J. V., Bantis, L. E., Georgiou S. D. (2012). Inference in Generalized Linear Regression Models with a censored covariate. *Computational Statistics and Data Analysis* **56**, 1854-1868.

Vexler, A., Liu, A., Eliseev, E. and Enrique, F., Schisterman (2008). Maximum likelihood ratio tests for comparing the disciminatory ability of biomarkers subject to limit of detection *Biometrics* **64(3)**, 895-903.

Wang C. Y., Hsu, L., Feng Z. D., and Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics* **53,** 131-145.

Wang, C. Y., and Pepe, M. S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B* **62,** 509-524.

Wand, M. P., and Jones, M. C. (1995). Kernel smoothing. *Monographs on Statistics and Applied Probability London: Chapman and Hall*

Wang, Z., Chang, Y. C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12(2)**, 369-385.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton Method. *Biometrika* **61,**, 439-447.

Wright, E. M., Bowman, A. W. (1974). Exploration of survival data using non parametric quantile estimators. *Tecnhcal Report, University of Glasgow.*

Yiannoutsos, C.T., Nakas, C.T., Naviac, B.A. (2008). Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR Spectroscopy (MRS) in HIV-related neurological injury. *Neuroimage* **40(1)**, 248-255.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14**, 835-862.

Yuan, K. H., and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis* **65**, 245-260.

Zhang Y., Hua L. and Huang J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338-354.

Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). *Statistical methods in diagnostic medicine* Wiley, New York.

Appendices

1 Appendix A

1.1 Additional Simulations for continuous, count and binary data

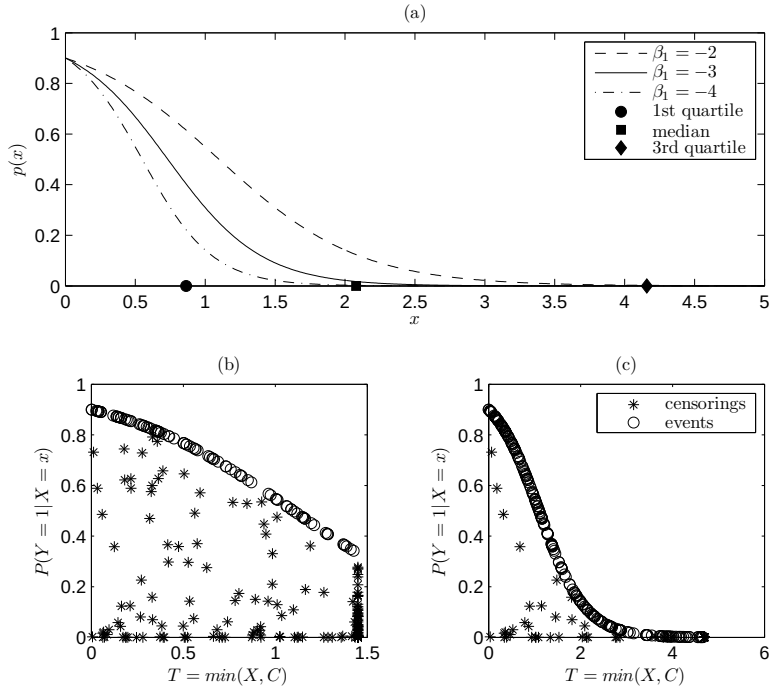

Figure A1: (a): The three sigmoidals corresponding to the values of $\beta_1 = -2, -3, -4$ at which the simulations were conducted using the logit link function. On the $x-$axis the three quartiles for the true distibution of the covariate (Exponential with mean 3) are plotted. (b): Example of simulated data set where the probability of a positive result given the true value of X, versus the censored covariate ($T = min(X, C)$) is plotted. The slope value is -2 and the total censoring 70%, half of which is due to the cutpoint (end of study). The value of β_0 is $log(9)$. (c): Example of simulated data set where the probability of a positive result given the true value of X versus the censored covariate ($T = min(X, C)$) is plotted. The slope value is -2 and the total censoring 30%, half of which is due to the cutpoint (end of study). The value of β_0 is $log(9)$.

Table A1: Simulation results of 1000 repetitions for the linear case ($n = 100$). Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from $N(0, 1)$.

			β_0					β_1				
Cens.	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
30%	0.2	$Likelihood$	-0.0380	0.1280	0.0178	- - -	- - -	0.0119	0.0277	0.0009	- - -	- - -
		CC	-0.0135	0.1954	0.0384	0.7963	0.9580	0.0065	0.0961	0.0093	0.3910	0.9540
		$QS(Exp)$	-0.0063	0.1586	0.0252	0.6131	0.9500	0.0010	0.0413	0.0017	0.1607	0.9450
		$Un(Exp)$	-0.0063	0.1587	0.0252	0.6133	0.9510	0.0010	0.0413	0.0017	0.1608	0.9460
		$QS(Weib)$	-0.0069	0.1596	0.0255	0.6151	0.9480	0.0012	0.0424	0.0018	0.1515	0.9420
		$Un(Weib)$	-0.0070	0.1597	0.0256	0.6151	0.9470	0.0012	0.0424	0.0018	0.1615	0.9450
	0.3	$Likelihood$	-0.0006	0.1483	0.0220	- - -	- - -	0.0028	0.0374	0.0014	- - -	- - -
		CC	0.0029	0.2037	0.0415	0.7957	0.9440	0.0025	0.0986	0.0097	0.3895	0.9500
		$QS(Exp)$	0.0047	0.1596	0.0255	0.6146	0.9380	0.0008	0.0430	0.0018	0.1626	0.9410
		$Un(Exp)$	0.0046	0.1595	0.0255	0.6149	0.9380	0.0008	0.0430	0.0018	0.1627	0.9410
		$QS(Weib)$	0.0037	0.1605	0.0258	0.6167	0.9440	0.0012	0.0443	0.0020	0.1637	0.9390
		$Un(Weib)$	0.0036	0.1603	0.0257	0.6170	0.9430	0.0012	0.0443	0.0020	0.1639	0.9400
	0.5	$Likelihood$	0.0101	0.1585	0.0252	- - -	- - -	-0.0026	0.0461	0.0021	- - -	- - -
		CC	0.0030	0.1999	0.0400	0.7962	0.9470	0.0002	0.0970	0.0094	0.3896	0.9550
		$QS(Exp)$	0.0070	0.1651	0.0273	0.6215	0.9420	-0.0020	0.0477	0.0023	0.1719	0.9270
		$Un(Exp)$	0.0073	0.1647	0.0272	0.6232	0.9400	-0.0021	0.0476	0.0023	0.1727	0.9280
		$QS(Weib)$	0.0057	0.1667	0.0278	0.6230	0.9390	-0.0015	0.0509	0.0026	0.1726	0.9050
		$Un(Weib)$	0.0058	0.1663	0.0277	0.6248	0.9380	-0.0016	0.0510	0.0026	0.1735	0.9000
	0.8	$Likelihood$	0.0080	0.1660	0.0276	- - -	- - -	-0.0032	0.0577	0.0033	- - -	- - -
		CC	0.0029	0.2018	0.0407	0.7969	0.9470	-0.0011	0.0957	0.0092	0.3891	0.9560
		$QS(Exp)$	0.0087	0.1709	0.0293	0.6552	0.9420	-0.0043	0.0590	0.0035	0.2154	0.9250
		$Un(Exp)$	0.0076	0.1766	0.0313	0.6743	0.9380	-0.0040	0.0637	0.0041	0.2278	0.9230
		$QS(Weib)$	0.0046	0.1772	0.0314	0.6530	0.9300	-0.0025	0.0673	0.0045	0.2158	0.8780
		$Un(Weib)$	0.0034	0.1850	0.0342	0.6742	0.9320	-0.0023	0.0748	0.0056	0.2274	0.8640
70%	0.2	$Likelihood$	-0.0565	0.1475	0.0250	- - -	- - -	0.0183	0.0377	0.0018	- - -	- - -
		CC	-0.0108	0.3337	0.1115	1.3594	0.9510	0.0051	0.4687	0.2198	1.8982	0.9500
		$QS(Exp)$	-0.0103	0.2087	0.0436	0.8172	0.9420	0.0027	0.0614	0.0038	0.2418	0.9520
		$Un(Exp)$	-0.0103	0.2086	0.0436	0.8175	0.9420	0.0027	0.0614	0.0038	0.2419	0.9520
		$QS(Weib)$	-0.0127	0.2117	0.0450	0.8250	0.9390	0.0053	0.0696	0.0049	0.2506	0.9220
		$Un(Weib)$	-0.0127	0.2116	0.0449	0.8253	0.9410	0.0053	0.0696	0.0049	0.2506	0.9230
	0.3	$Likelihood$	-0.0095	0.1649	0.0273	- - -	- - -	0.0060	0.0489	0.0024	- - -	- - -
		CC	0.0028	0.3498	0.1223	1.3631	0.9490	-0.0002	0.4801	0.2304	1.8938	0.9530
		$QS(Exp)$	0.0044	0.2073	0.0430	0.8163	0.9390	0.0009	0.0647	0.0042	0.2424	0.9410
		$Un(Exp)$	0.0044	0.2069	0.0428	0.8166	0.9390	0.0009	0.0646	0.0042	0.2425	0.9390
		$QS(Weib)$	-0.0005	0.2107	0.0444	0.8260	0.9340	0.0070	0.0768	0.0060	0.2564	0.8990
		$Un(Weib)$	-0.0005	0.2102	0.0442	0.8265	0.9360	0.0070	0.0768	0.0059	0.2565	0.8990
	0.5	$Likelihood$	0.0174	0.1813	0.0332	- - -	- - -	-0.0045	0.0624	0.0039	- - -	- - -
		CC	0.0162	0.3251	0.1060	1.3755	0.9600	-0.0257	0.4515	0.2045	1.9145	0.9620
		$QS(Exp)$	0.0039	0.2142	0.0459	0.8258	0.9290	-0.0008	0.0708	0.0050	0.2520	0.9130
		$Un(Exp)$	0.0039	0.2145	0.0460	0.8282	0.9300	-0.0007	0.0710	0.0050	0.2524	0.9140
		$QS(Weib)$	-0.0028	0.2178	0.0475	0.8347	0.9340	0.0075	0.0920	0.0085	0.2635	0.8250
		$Un(Weib)$	-0.0028	0.2180	0.0475	0.8371	0.9370	0.0076	0.0922	0.0086	0.2641	0.8270
	0.8	$Likelihood$	0.0108	0.1792	0.0322	- - -	- - -	-0.0021	0.0959	0.0092	- - -	- - -
		CC	-0.0001	0.3433	0.1178	1.3635	0.9480	0.0068	0.4770	0.2276	1.8964	0.9400
		$QS(Exp)$	0.0047	0.2163	0.0468	0.8348	0.9380	-0.0017	0.1024	0.0105	0.2996	0.8490
		$Un(Exp)$	0.0036	0.2198	0.0483	0.8576	0.9440	-0.0015	0.1020	0.0104	0.3040	0.8520
		$QS(Weib)$	-0.0123	0.2327	0.0543	0.8448	0.9260	0.0194	0.1676	0.0285	0.3115	0.6480
		$Un(Weib)$	-0.0135	0.2343	0.0551	0.8675	0.9280	0.0200	0.1678	0.0286	0.3159	0.6530

Table A2: Simulation results of 1000 repetitions for the linear case ($n = 300$). Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from Student t with 4 d.f.

			β_0					β_1				
Cens.	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
		$Likelihood$	-0.1491	0.2669	0.0935	- - -	- - -	0.0457	0.0691	0.0069	- - -	- - -
		CC	0.0009	0.1652	0.0273	0.6315	0.9510	-0.0010	0.0805	0.0065	0.3062	0.9440
		$QS(Exp)$	-0.0006	0.1275	0.0163	0.4944	0.9530	0.0002	0.0329	0.0011	0.1285	0.9490
	0.2	$Un(Exp)$	-0.0006	0.1275	0.0162	0.4983	0.9530	0.0002	0.0329	0.0011	0.1284	0.9500
		$QS(Weib)$	-0.0013	0.1279	0.0163	0.4954	0.9520	0.0005	0.0335	0.0011	0.1291	0.9470
		$Un(Weib)$	-0.0013	0.1278	0.0163	0.4954	0.9520	0.0005	0.0335	0.0011	0.1292	0.9480
		$QS(GG)$	-0.0003	0.1287	0.0166	0.4948	0.9490	-0.0003	0.0347	0.0012	0.1285	0.9280
		$Un(GG)$	-0.0003	0.1286	0.0165	0.4949	0.9500	-0.0003	0.0346	0.0012	0.1285	0.9310
		$Likelihood$	-0.0906	0.2460	0.0687	- - -	- - -	0.0274	0.0680	0.0054	- - -	- - -
		CC	0.0032	0.1611	0.0259	0.6358	0.9560	-0.0009	0.0775	0.0060	0.3091	0.9490
		$QS(Exp)$	0.0032	0.1290	0.0166	0.4990	0.9500	-0.0008	0.0350	0.0012	0.1312	0.9340
	0.3	$Un(Exp)$	0.0032	0.1295	0.0167	0.4997	0.9360	-0.0008	0.0350	0.0012	0.1309	0.9450
		$QS(Weib)$	-0.0071	0.1356	0.0185	0.5002	0.9350	0.0017	0.0354	0.0013	0.1312	0.9370
		$Un(Weib)$	-0.0070	0.1355	0.0184	0.5003	0.9360	0.0017	0.0354	0.0013	0.1313	0.9360
		$QS(GG)$	-0.0056	0.1367	0.0187	0.4998	0.9310	0.0006	0.0381	0.0015	0.1307	0.9050
		$Un(GG)$	-0.0055	0.1366	0.0187	0.5000	0.9310	0.0006	0.0382	0.0015	0.1308	0.9020
30%		$Likelihood$	-0.0405	0.2366	0.0576	- - -	- - -	0.0102	0.0651	0.0043	- - -	- - -
		CC	-0.0059	0.1609	0.0259	0.6335	0.9490	0.0002	0.0818	0.0067	0.3081	0.9390
		$QS(Exp)$	-0.0044	0.1268	0.0161	0.5038	0.9550	-0.0002	0.0364	0.0013	0.1395	0.9520
	0.5	$Un(Exp)$	-0.0042	0.1272	0.0162	0.5053	0.9560	-0.0003	0.0365	0.0013	0.1402	0.9520
		$QS(Weib)$	-0.0043	0.1276	0.0163	0.5043	0.9530	-0.0003	0.0385	0.0015	0.1397	0.9410
		$Un(Weib)$	-0.0042	0.1283	0.0165	0.5056	0.9540	-0.0004	0.0387	0.0015	0.1403	0.9350
		$QS(GG)$	-0.0022	0.1322	0.0175	0.5042	0.9430	-0.0018	0.0463	0.0021	0.1397	0.8540
		$Un(GG)$	-0.0018	0.1330	0.0177	0.5057	0.9440	-0.0022	0.0470	0.0022	0.1404	0.8520
		$Likelihood$	-0.0140	0.1610	0.0261	- - -	- - -	0.0034	0.0507	0.0026	- - -	- - -
		CC	-0.0078	0.1584	0.0251	0.6320	0.9610	0.0021	0.0787	0.0062	0.3078	0.9590
		$QS(Exp)$	-0.0057	0.1311	0.0172	0.5268	0.9560	0.0005	0.0483	0.0023	0.1746	0.9290
	0.8	$Un(Exp)$	-0.0070	0.1339	0.0180	0.5456	0.9560	0.0008	0.0520	0.0027	0.1851	0.9290
		$QS(Weib)$	-0.0075	0.1344	0.0181	0.5271	0.9460	0.0013	0.0546	0.0030	0.1748	0.8780
		$Un(Weib)$	-0.0089	0.1385	0.0193	0.5501	0.9500	0.0016	0.0611	0.0037	0.1851	0.8570
		$QS(GG)$	-0.0058	0.1520	0.0231	0.5274	0.9190	0.0001	0.0736	0.0054	0.1757	0.7620
		$Un(GG)$	-0.0037	0.1634	0.0267	0.5482	0.9130	-0.0022	0.0865	0.0075	0.1870	0.7160
		$Likelihood$	-0.1982	0.2679	0.1111	- - -	- - -	0.0631	0.0798	0.0104	- - -	- - -
		CC	-0.0012	0.2684	0.0721	1.0406	0.9520	-0.0043	0.3634	0.1321	1.4399	0.9540
		$QS(Exp)$	-0.0030	0.1731	0.0300	0.6524	0.9390	0.0007	0.0510	0.0026	0.1925	0.9420
	0.2	$Un(Exp)$	-0.0030	0.1730	0.0300	0.6525	0.9400	0.0006	0.0510	0.0026	0.1924	0.9440
		$QS(Weib)$	-0.0046	0.1738	0.0302	0.6553	0.9390	0.0026	0.0545	0.0030	0.1965	0.9310
		$Un(Weib)$	-0.0045	0.1737	0.0302	0.6554	0.9390	0.0026	0.0545	0.0030	0.1965	0.9320
		$QS(GG)$	-0.0059	0.1768	0.0313	0.6591	0.9400	0.0047	0.0749	0.0056	0.2028	0.7890
		$Un(GG)$	-0.0054	0.1761	0.0310	0.6670	0.9430	0.0041	0.0738	0.0055	0.2163	0.7760
		$Likelihood$	-0.1454	0.2764	0.0975	- - -	- - -	0.0453	0.0865	0.0095	- - -	- - -
		CC	0.0019	0.2619	0.0686	1.0479	0.9580	-0.0111	0.3647	0.1331	1.4506	0.9470
		$QS(Exp)$	-0.0037	0.1661	0.0276	0.6591	0.9590	-0.0004	0.0516	0.0026	0.1963	0.9320
	0.3	$Un(Exp)$	-0.0038	0.1663	0.0277	0.6694	0.9460	-0.0004	0.0517	0.0027	0.1974	0.9540
		$QS(Weib)$	-0.0053	0.1666	0.0278	0.6614	0.9550	0.0008	0.0572	0.0033	0.1986	0.9050
		$Un(Weib)$	-0.0055	0.1669	0.0279	0.6596	0.9560	0.0008	0.0573	0.0033	0.1987	0.9060
		$QS(GG)$	-0.0104	0.1844	0.0341	0.6746	0.9380	0.0056	0.0925	0.0086	0.2069	0.6800
		$Un(GG)$	-0.0102	0.1844	0.0341	0.6717	0.9360	0.0053	0.0924	0.0086	0.2006	0.6700
70%		$Likelihood$	-0.0521	0.2850	0.0839	- - -	- - -	0.0126	0.0729	0.0055	- - -	- - -
		CC	0.0020	0.2618	0.0686	1.0482	0.9580	-0.0114	0.3646	0.1330	1.4510	0.9470
		$QS(Exp)$	-0.0038	0.1672	0.0280	0.6647	0.9580	-0.0002	0.0572	0.0033	0.2038	0.9150
	0.5	$Un(Exp)$	-0.0046	0.1681	0.0283	0.7740	0.9560	-0.0000	0.0573	0.0033	0.2041	0.9150
		$QS(Weib)$	-0.0067	0.1685	0.0284	0.6645	0.9480	0.0024	0.0727	0.0053	0.2060	0.8410
		$Un(Weib)$	-0.0072	0.1694	0.0288	0.6662	0.9480	0.0025	0.0729	0.0053	0.2064	0.8390
		$QS(GG)$	-0.0121	0.1821	0.0333	0.6704	0.9400	0.0129	0.1398	0.0197	0.2181	0.5280
		$Un(GG)$	-0.0124	0.1827	0.0335	0.6647	0.9358	0.0125	0.1405	0.0199	0.2034	0.5041
		$Likelihood$	0.0130	0.1694	0.0289	- - -	- - -	-0.0034	0.0798	0.0064	- - -	- - -
		CC	-0.0085	0.2622	0.0688	1.0511	0.9640	0.0125	0.3652	0.1336	1.4519	0.9620
		$QS(Exp)$	-0.0017	0.1741	0.0303	0.6740	0.9520	-0.0003	0.0828	0.0069	0.2426	0.8610
	0.8	$Un(Exp)$	-0.0027	0.1784	0.0318	0.6926	0.9500	-0.0000	0.0834	0.0070	0.2462	0.8580
		$QS(Weib)$	-0.0083	0.1834	0.0337	0.6762	0.9390	0.0078	0.1314	0.0173	0.2453	0.6450
		$Un(Weib)$	-0.0093	0.1862	0.0348	0.6950	0.9380	0.0082	0.1321	0.0175	0.2489	0.6640
		$QS(GG)^{(1)}$	-0.0189	0.2413	0.0586	0.6804	0.8370	0.0295	0.3018	0.0920	0.2588	0.2700
		$Un(GG)$	-0.0190	0.2453	0.0606	0.6660	0.8200	0.0277	0.3074	0.0953	0.2051	0.2180

(1) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 95.8% and 96.7% respectively

Table A3: Simulation results of 1000 repetitions for the linear case $(n = 100)$. Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from Student t with 4 d.f.

			β_0					β_1				
Cens.	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
30%	0.2	$Likelihood$	-0.0970	0.2388	0.0664	- - -	- - -	0.0304	0.0572	0.0042	- - -	- - -
		CC	-0.0127	0.2792	0.0781	1.1067	0.9630	0.0060	0.1396	0.0195	0.5411	0.9560
		$QS(Exp)$	-0.0063	0.2188	0.0479	0.8526	0.9540	0.0016	0.0575	0.0033	0.2225	0.9490
		$Un(Exp)$	-0.0063	0.2187	0.0479	0.8529	0.9530	0.0016	0.0575	0.0033	0.2225	0.9500
		$QS(Weib)$	-0.0075	0.2202	0.0485	0.8553	0.9510	0.0020	0.0588	0.0035	0.2236	0.9450
		$Un(Weib)$	-0.0075	0.2202	0.0485	0.8555	0.9510	0.0020	0.0588	0.0035	0.2237	0.9470
	0.3	$Likelihood$	-0.0421	0.2401	0.0594	- - -	- - -	0.0147	0.0623	0.0041	- - -	- - -
		CC	0.0015	0.2815	0.0792	1.1095	0.9530	0.0001	0.1377	0.0190	0.5436	0.9540
		$QS(Exp)$	0.0007	0.2216	0.0491	0.8567	0.9470	0.0017	0.0602	0.0036	0.2280	0.9360
		$Un(Exp)$	0.0006	0.2220	0.0493	0.8572	0.9480	0.0018	0.0603	0.0036	0.2282	0.9370
		$QS(Weib)$	-0.0011	0.2229	0.0497	0.8601	0.9490	0.0026	0.0616	0.0038	0.2301	0.9360
		$Un(Weib)$	-0.0013	0.2232	0.0498	0.8606	0.9490	0.0027	0.0618	0.0038	0.2303	0.9370
	0.5	$Likelihood$	-0.0010	0.2480	0.0615	- - -	- - -	0.0004	0.0689	0.0047	- - -	- - -
		CC	0.0131	0.2790	0.0780	1.1112	0.9580	-0.0049	0.1359	0.0185	0.5419	0.9470
		$QS(Exp)$	0.0082	0.2230	0.0498	0.8699	0.9500	-0.0022	0.0632	0.0040	0.2419	0.9340
		$Un(Exp)$	0.0072	0.2238	0.0501	0.8727	0.9520	-0.0019	0.0637	0.0041	0.2433	0.9310
		$QS(Weib)$	0.0051	0.2260	0.0511	0.8725	0.9510	-0.0007	0.0675	0.0046	0.2435	0.9070
		$Un(Weib)$	0.0041	0.2271	0.0516	0.8753	0.9520	-0.0003	0.0684	0.0047	0.2448	0.9110
	0.8	$Likelihood$	-0.0010	0.2447	0.0599	- - -	- - -	0.0027	0.0829	0.0069	- - -	- - -
		CC	-0.0122	0.2835	0.0805	1.1041	0.9430	0.0074	0.1375	0.0190	0.5390	0.9470
		$QS(Exp)$	-0.0027	0.2432	0.0592	0.9053	0.9430	0.0017	0.0844	0.0071	0.3023	0.9210
		$Un(Exp)$	-0.0038	0.2526	0.0638	0.9403	0.9500	0.0020	0.0924	0.0085	0.3218	0.9150
		$QS(Weib)$	-0.0099	0.2500	0.0626	0.9068	0.9370	0.0053	0.0961	0.0093	0.3029	0.8790
		$Un(Weib)$	-0.0116	0.2619	0.0687	0.9404	0.9320	0.0057	0.1082	0.0117	0.3213	0.8580
70%	0.2	$Likelihood$	-0.1574	0.2559	0.0894	- - -	- - -	0.0492	0.0721	0.0076	- - -	- - -
		CC	-0.0176	0.4539	0.2063	1.8495	0.9600	0.0186	0.6242	0.3899	2.5771	0.9620
		$QS(Exp)$	-0.0111	0.2874	0.0827	1.1042	0.9470	0.0033	0.0869	0.0076	0.3281	0.9320
		$Un(Exp)$	-0.0112	0.2877	0.0829	1.1046	0.9500	0.0033	0.0870	0.0076	0.3282	0.9320
		$QS(Weib)$	-0.0148	0.2922	0.0856	1.1180	0.9490	0.0086	0.0995	0.0100	0.3464	0.9160
		$Un(Weib)$	-0.0149	0.2926	0.0858	1.1184	0.9490	0.0086	0.0997	0.0100	0.3466	0.9140
	0.3	$Likelihood$	-0.0861	0.2660	0.0782	- - -	- - -	0.0298	0.0827	0.0077	- - -	- - -
		CC	0.0081	0.4754	0.2261	1.8802	0.9550	0.0032	0.6747	0.4542	2.6143	0.9370
		$QS(Exp)$	0.0100	0.2969	0.0833	0.8249	0.9320	-0.0014	0.0908	0.0082	0.3351	0.9270
		$Un(Exp)$	0.0090	0.2973	0.0882	0.8274	0.9370	-0.0012	0.0907	0.0082	0.3360	0.9260
		$QS(Weib)$	0.0030	0.3020	0.0912	1.1408	0.9370	0.0071	0.1081	0.0117	0.3565	0.8900
		$Un(Weib)$	0.0028	0.3023	0.0914	1.1419	0.9380	0.0072	0.1082	0.0117	0.3566	0.8900
	0.5	$Likelihood$	-0.0089	0.2644	0.0700	- - -	- - -	0.0050	0.0929	0.0087	- - -	- - -
		CC	0.0191	0.4537	0.2062	1.8637	0.9590	-0.0163	0.6399	0.4097	2.5836	0.9570
		$QS(Exp)$	0.0092	0.2950	0.0871	1.1265	0.9430	-0.0005	0.0990	0.0098	0.3447	0.9080
		$Un(Exp)$	0.0091	0.2950	0.0871	1.1303	0.9440	-0.0005	0.0986	0.0097	0.3499	0.9110
		$QS(Weib)$	-0.0021	0.3025	0.0915	1.1407	0.9350	0.0147	0.1327	0.0178	0.3691	0.8130
		$Un(Weib)$	-0.0023	0.3023	0.0914	1.1445	0.9390	0.0148	0.1326	0.0178	0.3698	0.8180
	0.8	$Likelihood$	0.0134	0.2694	0.0726	- - -	- - -	-0.0013	0.1376	0.0189	- - -	- - -
		CC	-0.0262	0.4686	0.2203	1.8754	0.9430	0.0140	0.6555	0.4299	2.6080	0.9610
		$QS(Exp)$	-0.0172	0.3019	0.0915	1.1451	0.9360	0.0044	0.1418	0.0201	0.4174	0.8590
		$Un(Exp)$	-0.0181	0.3094	0.0961	1.1819	0.9400	0.0044	0.1424	0.0203	0.4273	0.8680
		$QS(Weib)$	-0.0447	0.3285	0.1099	1.1605	0.9170	0.0389	0.2437	0.0609	0.4360	0.6390
		$Un(Weib)$	-0.0458	0.3342	0.1138	1.1974	0.9240	0.0396	0.2454	0.0618	0.4428	0.6550

Table A4: Simulation results of 1000 repetitions for the linear case ($n = 1000$). Half of the censoring is due to the cutpoint (end of study). The noise ϵ is from $N(0, 1)$.

			β_0					β_1				
Cens.	ρ	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
		$Likelihood$	-0.0117	0.0409	0.0018	- - -	- - -	0.0042	0.0086	0.0001	- - -	- - -
		CC	0.0014	0.0639	0.0041	0.2464	0.9490	-0.0003	0.0303	0.0009	0.1197	0.9420
		$QS(Exp)$	0.0010	0.0498	0.0025	0.1937	0.9520	-0.0000	0.0126	0.0002	0.0501	0.9560
	0.2	$Un(Exp)$	0.0010	0.0497	0.0025	0.1937	0.9520	-0.0000	0.0126	0.0002	0.0501	0.9560
		$QS(Weib)$	0.0009	0.0499	0.0025	0.1930	0.9490	0.0000	0.0127	0.0002	0.0501	0.9570
		$Un(Weib)$	0.0009	0.0499	0.0025	0.1939	0.9490	0.0000	0.0127	0.0002	0.0501	0.9560
		$QS(GG)$	0.0013	0.0500	0.0025	0.1937	0.9480	-0.0002	0.0131	0.0002	0.0500	0.9470
		$Un(GG)$	0.0013	0.0500	0.0025	0.1937	0.9490	-0.0002	0.0131	0.0002	0.0500	0.9470
		$Likelihood$	-0.0022	0.0487	0.0024	- - -	- - -	0.0008	0.0129	0.0002	- - -	- - -
		CC	-0.0046	0.0630	0.0040	0.2466	0.9410	0.0018	0.0307	0.0009	0.1196	0.9470
		$QS(Exp)$	-0.0028	0.0493	0.0024	0.1945	0.9380	0.0010	0.0131	0.0002	0.0509	0.9350
	0.3	$Un(Exp)$	-0.0027	0.0493	0.0024	0.1945	0.9380	0.0010	0.0131	0.0002	0.0509	0.9360
		$QS(Weib)$	-0.0030	0.0495	0.0025	0.1946	0.9360	0.0011	0.0134	0.0002	0.0510	0.9370
		$Un(Weib)$	-0.0029	0.0495	0.0025	0.1945	0.9360	0.0011	0.0134	0.0002	0.0510	0.9360
		$QS(GG)$	-0.0027	0.0501	0.0025	0.1946	0.9400	0.0009	0.0144	0.0002	0.0510	0.9220
		$Un(GG)$	-0.0026	0.0501	0.0025	0.1946	0.9390	0.0009	0.0144	0.0002	0.0509	0.9230
30%		$Likelihood$	0.0014	0.0509	0.0026	- - -	- - -	-0.0003	0.0141	0.0002	- - -	- - -
		CC	0.0030	0.0684	0.0047	0.2465	0.9220	-0.0016	0.0325	0.0011	0.1196	0.9390
		$QS(Exp)$	0.0011	0.0525	0.0028	0.1966	0.9380	-0.0002	0.0145	0.0002	0.0541	0.9330
	0.5	$Un(Exp)$	0.0010	0.0524	0.0027	0.1971	0.9380	-0.0002	0.0144	0.0002	0.0543	0.9330
		$QS(Weib)$	0.0009	0.0530	0.0028	0.1967	0.9350	-0.0001	0.0154	0.0002	0.0542	0.9220
		$Un(Weib)$	0.0008	0.0530	0.0028	0.1971	0.9370	-0.0001	0.0155	0.0002	0.0543	0.9220
		$QS(GG)$	0.0016	0.0554	0.0031	0.1967	0.9230	-0.0007	0.0184	0.0003	0.0541	0.8620
		$Un(GG)$	0.0016	0.0555	0.0031	0.1970	0.9210	-0.0007	0.0186	0.0003	0.0543	0.8530
		$Likelihood$	-0.0006	0.0521	0.0027	- - -	- - -	-0.0004	0.0177	0.0003	- - -	- - -
		CC	-0.0011	0.0629	0.0040	0.2468	0.9210	-0.0004	0.0298	0.0009	0.1198	0.9650
		$QS(Exp)$	-0.0009	0.0542	0.0029	0.2063	0.9430	-0.0004	0.0184	0.0003	0.0680	0.9420
		$Un(Exp)$	-0.0010	0.0561	0.0031	0.2130	0.9390	-0.0003	0.0199	0.0004	0.0718	0.9230
	0.8	$QS(Weib)$	-0.0018	0.0559	0.0031	0.2063	0.9290	0.0002	0.0208	0.0004	0.0681	0.8960
		$Un(Weib)$	-0.0020	0.0582	0.0034	0.2131	0.9260	0.0003	0.0232	0.0005	0.0718	0.8750
		$QS(GG)$	-0.0017	0.0620	0.0039	0.2100	0.9090	0.0000	0.0278	0.0008	0.0681	0.7710
		$Un(GG)$	-0.0013	0.0669	0.0045	0.2133	0.8810	-0.0002	0.0323	0.0010	0.0720	0.7360
		$Likelihood$	-0.0164	0.0486	0.0026	- - -	- - -	0.0057	0.0122	0.0002	- - -	- - -
		CC	0.0035	0.1064	0.0113	0.4066	0.9390	-0.0027	0.1453	0.0211	0.5616	0.9500
		$QS(Exp)$	0.0023	0.0675	0.0046	0.2588	0.9500	-0.0005	0.0195	0.0004	0.0760	0.9510
	0.2	$Un(Exp)$	0.0021	0.0675	0.0046	0.2588	0.9500	-0.0005	0.0195	0.0004	0.0760	0.9510
		$QS(Weib)$	0.0021	0.0678	0.0046	0.2592	0.9460	-0.0001	0.0206	0.0004	0.0765	0.9410
		$Un(Weib)$	0.0021	0.0677	0.0046	0.2592	0.9460	-0.0001	0.0206	0.0004	0.0765	0.9420
		$QS(GG)$	0.0009	0.0695	0.0048	0.2606	0.9360	0.0022	0.0311	0.0010	0.0791	0.8030
		$Un(GG)$	0.0009	0.0695	0.0048	0.2606	0.9380	0.0022	0.0311	0.0010	0.0791	0.8020
		$Likelihood$	-0.0014	0.0602	0.0036	- - -	- - -	0.0004	0.0183	0.0003	- - -	- - -
		CC	-0.0082	0.1043	0.0110	0.4079	0.9440	0.0075	0.1416	0.0201	0.5633	0.9570
		$QS(Exp)$	-0.0039	0.0672	0.0045	0.2598	0.9420	0.0012	0.0209	0.0004	0.0768	0.9270
	0.3	$Un(Exp)$	-0.0038	0.0671	0.0045	0.2598	0.9420	0.0012	0.0209	0.0004	0.0768	0.9230
		$QS(Weib)$	-0.0043	0.0674	0.0046	0.2601	0.9420	0.0017	0.0229	0.0005	0.0773	0.9120
		$Un(Weib)$	-0.0042	0.0673	0.0046	0.2601	0.9400	0.0017	0.0229	0.0005	0.0773	0.9120
		$QS(GG)$	-0.0060	0.0702	0.0050	0.2615	0.9329	0.0048	0.0404	0.0017	0.0773	0.6764
		$Un(GG)$	-0.0059	0.0701	0.0050	0.2615	0.9329	0.0048	0.0404	0.0017	0.0773	0.6743
70%		$Likelihood$	0.0037	0.0592	0.0035	- - -	- - -	-0.0010	0.0195	0.0004	- - -	- - -
		CC	0.0086	0.1038	0.0108	0.4071	0.9490	-0.0091	0.1382	0.0192	0.5614	0.9570
		$QS(Exp)$	0.0038	0.0703	0.0050	0.2604	0.9350	-0.0011	0.0228	0.0005	0.0796	0.9130
	0.5	$Un(Exp)$	0.0036	0.0703	0.0050	0.2604	0.9280	-0.0011	0.0228	0.0005	0.0796	0.9010
		$QS(Weib)$	0.0029	0.0706	0.0050	0.2608	0.9320	0.0002	0.0281	0.0008	0.0802	0.8450
		$Un(Weib)$	0.0029	0.0706	0.0050	0.2614	0.9360	0.0002	0.0281	0.0008	0.0803	0.8450
		$QS(GG)$	-0.0018	0.0800	0.0064	0.2629	0.9020	0.0091	0.0686	0.0048	0.0838	0.4680
		$Un(GG)$	-0.0020	0.0800	0.0064	0.2636	0.9040	0.0092	0.0687	0.0048	0.0839	0.4640
		$Likelihood$	0.0009	0.0556	0.0031	- - -	- - -	-0.0017	0.0293	0.0009	- - -	- - -
		CC	0.0004	0.1038	0.0108	0.4078	0.9510	-0.0024	0.1400	0.0196	0.5621	0.9470
		$QS(Exp)^{(1)}$	0.0002	0.0684	0.0047	0.2649	0.9400	-0.0015	0.0315	0.0010	0.0944	0.8500
	0.8	$Un(Exp)$	0.0005	0.0699	0.0049	0.2663	0.9400	-0.0015	0.0315	0.0010	0.0946	0.8510
		$QS(Weib)^{(2)}$	-0.0018	0.0715	0.0051	0.2653	0.9430	0.0012	0.0510	0.0026	0.0949	0.6340
		$Un(Weib)$	-0.0015	0.0726	0.0053	0.2720	0.9430	0.0012	0.0511	0.0026	0.0976	0.6510
		$QS(GG)^{(3)}$	-0.0127	0.1069	0.0116	0.2671	0.8020	0.0220	0.1525	0.0237	0.0980	0.2550
		$Un(GG)$	-0.0128	0.1086	0.0120	0.2749	0.8020	0.0222	0.1539	0.0242	0.0993	0.2600

(1) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 94.2% and 93.6% respectively
(2) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 94.5% and 93.6% respectively
(3) Coverage of CI for β_0 and β_1 based on 100 bootstrapped samples per iteration is 94.5% and 96.1% respectively

Table A5: Simulation results of 1000 repetitions in the case of count data. Sample size is n=300. Half of the censoring is due to the cutpoint (end of study). The real value of the β_0 is 1.

			β_0					β_1				
Cens.	β_1	Method	Bias	SE	MSE	Width	Cover	Bias	SE	MSE	Width	Cover
		CC	-0.0059	0.2032	0.0413	0.7760	0.9470	-0.0210	0.2038	0.0420	0.7912	0.9470
		$QS(Exp)$	-0.0036	0.1985	0.0394	0.7563	0.9430	-0.0221	0.1908	0.0369	0.7288	0.9440
	-1	$QS(Weib)$	-0.0034	0.1986	0.0394	0.7562	0.9430	-0.0222	0.1908	0.0369	0.7287	0.9440
		$QS(GG)$	-0.0044	0.1994	0.0398	0.7566	0.9410	-0.0214	0.1913	0.0370	0.7290	0.9430
30%		CC	-0.0061	0.1602	0.0257	0.6267	0.9540	-0.0064	0.1012	0.0103	0.3954	0.9500
		$QS(Exp)$	-0.0035	0.1433	0.0206	0.5631	0.9480	-0.0055	0.0710	0.0051	0.2775	0.9390
	-1/3	$QS(Weib)$	-0.0033	0.1433	0.0205	0.5632	0.9380	-0.0055	0.0709	0.0051	0.2776	0.9390
		$QS(GG)$	-0.0042	0.1437	0.0207	0.5632	0.9460	-0.0047	0.0715	0.0051	0.2776	0.9350
		CC	-0.0169	0.2762	0.0766	1.0265	0.9360	-0.0288	0.4766	0.2280	1.7851	0.9370
		$QS(Exp)^{(1)}$	-0.0070	0.2189	0.0479	0.8366	0.9400	-0.0229	0.2336	0.0551	0.8751	0.9330
	-1	$QS(Weib)^{(1)}$	-0.0040	0.2195	0.0482	0.8374	0.9260	-0.0272	0.2378	0.0573	0.8767	0.9440
		$QS(GG)^{(2)}$	-0.0014	0.2199	0.0483	0.8376	0.9380	-0.0389	0.2471	0.0626	0.8734	0.9140
70%		CC	-0.0174	0.2360	0.0560	0.9385	0.9470	-0.0106	0.3454	0.1194	1.4013	0.9560
		$QS(Exp)^{(3)}$	-0.0116	0.1656	0.0275	0.6647	0.9500	-0.0200	0.0968	0.0098	0.3786	0.9480
	-1/3	$QS(Weib)^{(4)}$	-0.0126	0.1666	0.0279	0.6671	0.9540	-0.0217	0.1006	0.0106	0.3810	0.9460
		$QS(GG)^{(5)}$	0.0173	0.1691	0.0289	0.6708	0.9530	-0.0379	0.1234	0.0167	0.3898	0.8991

(1) Approximately 1.3% of the repetitions were discarded due to violations of the required restrictions.
(2) Approximately 6.0% of the repetitions were discarded due to violations of the required restrictions.
(3) Approximately 5.7% of the repetitions were discarded due to violations of the required restrictions.
(4) Approximately 14% of the repetitions were discarded due to violations of the required restrictions.
(5) Approximately 23% of the repetitions were discarded due to violations of the required restrictions.

Table A6: Estimates of the coefficients of the linear model for the PBC data.

Method	Parameter	Est.	Asympt. SE	Asympt. CI 95%		Bootstrap CI 95%	
CC	α_0	2.8111	0.5563	1.7068	3.9154	–	–
	α_1	-0.0129	0.0028	-0.0184	-0.0074	–	–
	α_2	-0.0203	0.0102	-0.0405	-0.0002	–	–
QS(Weib)	α_0	3.2785	0.4084	2.4780	4.0790	2.5280	4.0026
	α_1	-0.0065	0.0007	-0.0079	-0.0051	-0.0088	-0.0043
	α_2	-0.0354	0.0070	-0.0491	-0.0217	-0.0485	-0.0220
QS(GG)	α_0	3.5526	0.3964	2.7755	4.3296	2.9565	4.2475
	α_1	-0.0090	0.0009	-0.0108	-0.0072	-0.0108	-0.0075
	α_2	-0.0384	0.0065	-0.0512	-0.0256	-0.0490	-0.0285
Un(Weib)	α_0	3.5001	0.4146	2.6875	4.3128	2.6599	4.3321
	α_1	-0.0063	0.0007	-0.0077	-0.0050	-0.0088	-0.0041
	α_2	-0.0400	0.0071	-0.0538	-0.0261	-0.0556	-0.0246
Un(GG)	α_0	3.6853	0.4028	2.8958	4.4748	2.5527	4.1720
	α_1	-0.0089	0.0009	-0.0107	-0.0071	-0.0106	-0.0033
	α_2	-0.0411	0.0066	-0.0429	-0.0393	-0.0498	-0.0226

Table A7: Estimates of the coefficients of the AFT models for the PBC data.

Model	Parameter	Est.	SE	CI 95%	
Weibull AFT	ξ_0	6.5038	0.4741	5.5745	7.4330
	ξ_1	-0.0263	0.0086	-0.0431	-0.0096
	σ	0.7828	0.0684	0.6596	0.9291
GG AFT	ξ_0	6.5359	0.5955	5.3686	7.7031
	ξ_1	-0.0242	0.0078	-0.0394	-0.0090
	σ	0.3269	0.6799	0.0055	19.2673
	δ	2.7674	5.8436	-8.6857	14.2206

2 Appendix B

2.1 A MATLAB package for survival estimation with constrained splines

In section 2.2 we recall the modeling technique along with the required constraints of monotonicity. In section 2.3 we give a description of the routine HCNS (Hazard Constraint Natural Spline), along with some examples that can be straightforwardly re-produced by the reader using MATLAB. In section 2.4 we present a small simulation study and in section 2.5 an application, showing the necessary code for analysis.

2.2 Background methodology

Consider again data of the form T_i, Di where $T_i = min(X_i, C_i)$ with C_i being the censoring variable, X_i the time to event variable and Δ_i is the event indicator taking the value 1 for an event and zero otherwise (Note that in MATLAB all routines that accommodate censoring have the opposite coding for the events/censorings, however we decide to use the common coding in theory to avoid confusion. In the code examples we repeat the coding used when necessary). Recall the spline based approach introduced in Chapter 2 (section 2.5.2.). Survival estimation based on the available data T_i, Δ_i is obtained in two stages. In the first stage the Kaplan Meier based cumulative step hazard function is obtained. In the second stage, model (2.14) is fitted to the corners of the cumulative step hazard function. The steps of the KM estimator occur only at event times, that is $T_i|\Delta_i = 1$. However the cumulative hazard function monotone, and thus monotonicity restrictions must be imposed.

As discussed in chapter 2, the monotonicity region, $\mathcal{M}$, is a non linear region and if one attempts to fit model 2.14 under the implied conditions of this then computational problems may occur. If one considers a linearly defined subregion, $\mathcal{A}$, of $\mathcal{M}$, then monotonicity would be achieved but other candidate models would be excluded. On the other hand, a linearly defined region has the merit of reducing the problem to be a restricted least squares one, with linear restrictions on the parameters. Thus, convergence is guaranteed since the function to minimize is always convex. We explore a linear approximation $\mathcal{A}$, of the entire region of approximation $\mathcal{M}$ by using optimal (in terms of inscribed area) polygons within $\mathcal{M}$. Smith (1970) presents an algorithm for deriving the optimal inscribed polygon within an ellipse. Using Smith's algorithm we showed that any optimal inscribed $(8k + 2)$-gon, $k = 1, 2, \ldots$ within $\mathcal{M}$ can be exactly calculated. The MATLAB code given in the appendixend of this appendix provides the optimal inscribed polygon along with the corresponding plot for region $\mathcal{M}$. The user is only expected to provide the value of k in the first line.

We already have explored the use of an optimal 18-gon to approximate region $\mathcal{M}$ ($k = 2$). In Figure B1, we show the approximation obtained from the optimal 10-gon $(k = 1)$, 18-gon $(k = 2)$, 26-gon $(k = 3)$ and 34-gon $(k = 4)$, and these figures can be reproduced by the code given in the end of this appendix.

We recommend the use of at least six knots due to the restrictions that force the model to be zero before the first knot. For the same reason one may choose placing more knots near the first events (i.e. for smaller T_i's) where additional flexibility is required to avoid underestimation of the cumulative hazard function in that region. Regarding the knot placement, in this chapter, we consider the following strategy: We derive

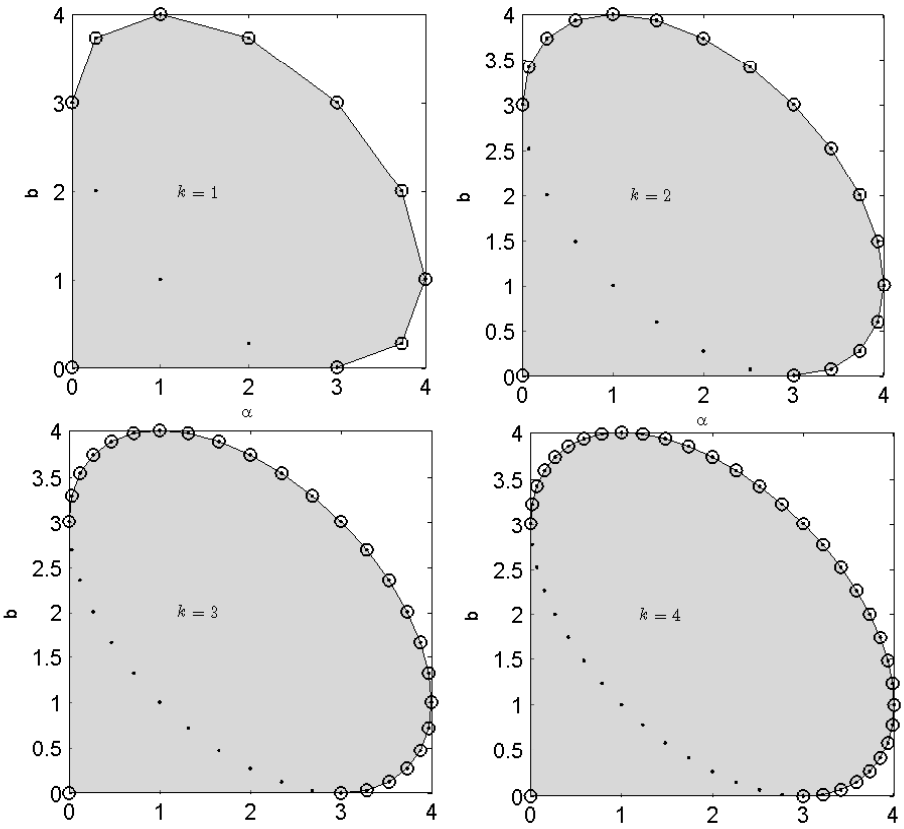

Figure B1: The shaded region is the linear approximation of the monotonicity region for $k = 1, 2, 3, 4$. The circles refer to the approximation of the region $\mathcal{M}$. The dots refer to the approximation of the ellipse $\phi(a, b)$.

10 equally spaced points expanding from $min(event\ times) = min(T_i|\Delta_i = 1)$ up to $max(event\ times) = max(T_i|\Delta_i = 1)$, and each of these points is a candidate for placing a knot. Using 6 knots, there are $10!/(6!4!) = 210$ possible combinations, and thus 210 possible knot schemes. Next, we consider 10 points at the following percentiles that are calculated only by the fully observed data: $0, 2.5th, 5th, 10th, 20th, 40th, 50th, 60th, 80th$, and $100th$. Exploring again all possible combinations, there are 210 additional combinations (knot schemes) to be explored. In a given application, and if asked by the user as we will see in the next section, all 420 knot schemes are tested by fitting model (2.14) to the Kaplan Meier based cumulative hazard function. Finally, the knot scheme that results to the smallest distance for the corners of the step function is the one chosen. That is, the knot scheme selection is based on the criterion

$$\Psi(\hat{\boldsymbol{\theta}}) = \sum_i (\hat{H}(T_i|\Delta_i = 1) - \hat{H}^{KM}(T_i|\Delta_i = 1))^2, \tag{.1}$$

where $\hat{H}$ is the fitted model defined in (2.14) under the appropriate constraints of monotonicity, and $\hat{H}^{KM}$ is the Kaplan Meier based cumulative hazard estimator.

This may seem a difficult task from a computing time point of view. However, with current computer technology, this procedure is only a matter of seconds, as can be seen from the simulation studies later on. In effect, resampling methods are feasible for the inference of a given data set. The use of percentile bootstrap showed satisfactory coverage of the corresponding confidence intervals (see also Appendix C).

Recall that the function to minimize is

$$\Psi(\hat{\boldsymbol{\theta}}) = \sum_i (\hat{H}(T_i|\Delta_i = 1) - \hat{H}^{KM}(T_i|\Delta_i = 1))^2,$$

where $\hat{H}$ is the fitted model defined in (2.14) under the appropriate constraints of monotonicity, and $\hat{H}^{KM}$ is the Kaplan Meier based cumulative hazard estimator.

Denote with Q the linear segments that form the approximation of the entire region of monotonicity $\mathcal{M}$ without including the ones that lie on the horizontal and vertical axis (for example, $Q = 16$ for the the optimal 18-gon). There are $Q(K-1)+K$ constraints, consisting of $Q(K-1)+K-1$ inequalities and one equality given in Chapter 2. Alternatively, one can consider $Q(K-1)+K+1$ inequality constraints, since the equality can be written as two inequalities, and so finally all constraints can be written as

$$\mathbf{A}[\theta_1, \theta_2, \ldots, \theta_{K-1}]' \leq \mathbf{0}.$$

where matrix $\mathbf{A}$ has $n_c = K-1$ columns and $n_r = Q(K-1)+K+1$ rows.

Thus, the problem stated is restricted least squares one with linear restrictions. The function to minimize is always convex and convergence is guaranteed. MATLAB's built in function for these kind of optimization problems is `lsqlin`.

The generalization of the approach to accommodate covariates is done under the assumption of proportional hazards. The usual Cox model can be fitted to the data in order to obtain the baseline cumulative step hazard estimate. Then model 2.14 is fitted to this crude estimate under the same constraints discussed above. Thus the two stage analysis required in a setting with p covariates $Z_1, Z_2, \ldots, Z_p$ is

- Stage 1: Fit the Cox model which of the form $H(x) = H_0(x)exp(\gamma_1 Z_1 + \gamma_2 Z_2 + \ldots + \gamma_p Z_p)$ and derive $\hat{H}_0^{cox}(x)$

- Stage 2: Fit model (2.14) to the corners of the $\hat{H}_0^{cox}(x)$ (that is where events occur) under the constraints $\mathbf{A}[\theta_1, \theta_2, \ldots, \theta_{K-1}]' \leq \mathbf{0}$ and derive the corresponding estimate $\hat{H}_0$.

Once the model 2.14 is fitted, then one can easily derive any survival estimation for any profile of a subject based on $\hat{H}(x) = \hat{H}_0(x)exp(\hat{\gamma_1} Z_1 + \hat{\gamma_2} Z_2 + \ldots + \hat{\gamma_p} Z_p)$, where $\hat{\gamma_i}, i = 1, 2, \ldots, p$ are simply the estimates provided by the usual Cox model fit.

2.3 Software Description

The routine has been developed using MATLAB R2011a and is available for download from the authors' website (or upon request). The algorithm of the program is summarized by the flowchart given in Figure B3.

The functions included in the file are `HCNS`, `HCNSboots`, `HCNSsup`, `HCNScox`, `HCNScoxsup`, `conlsqlin`, `cnsk`, and `approxM`. Apart from functions `HCNS` and `HCNSboots`, all others are interior functions that are called depending on the choices of the user. Right next we provide a description of these two functions which are the only ones that are of interest to the user. The inputs of the `HCNS` function are the following (mentioned with the order that are required from the user):

INPUT

- `time`: an array that may contain event times and/or right censored times
- `status`: a boolean array taking values 0 or 1 if the corresponding element of `time` is an event time or a censoring time respectively. (Note that MATLAB uses this coding in all its "survival" related functions, which is the opposite of the common coding used in a survival settings. We developed the code using MATLAB's coding)
- `Z`: a covariate matrix (each column corresponds to one covariate). If there are no covariates available then use "`[]`" instead.
- `knots`: The knots provided by the user. There is also an option of setting this field to "auto" and 6 knots will be used after checking all 210 combinations of knot schemes described in section 2.2.
- `k`: A positive integer greater or equal to 1. Based on the value of `k`, the optimal in terms of inscribed area $8k + 2$-gon will be used for approximating the region of monotonicity.

OPTIONAL INPUT

- `plots`: Can be set as `cumulative hazard`, `survivor`, `cdf` or `none` to plot the corresponding functions along with the corresponding empirical function. In the case where covariates (Z) are available the baseline corresponding functions are plotted (i.e. at Z=0, for all covariates). The 'none' option does not create any graph and allows you to proceed to the next optional input arguments.
- `profil`: It may contain a specific profile of covariate values and the estimates produced will refer to that specific profile. It is an array with length equal to the number of columns of the covariate matrix Z. (The 'plots' input argument must be given if the 'profil' argument is to be used).
- `profilplots`: Can be set as 'cumulative hazard', 'survivor', 'cdf' to plot the corresponding functions along with the empirical corresponding function for the specific profile given in 'profil'.

OUTPUT

- `bhat`: the estimated spline coefficients
- `Hx`: a "function handle" that can yield the value of the estimate of the cumulative hazard estimator based on the presented method, for any value(s) of x.

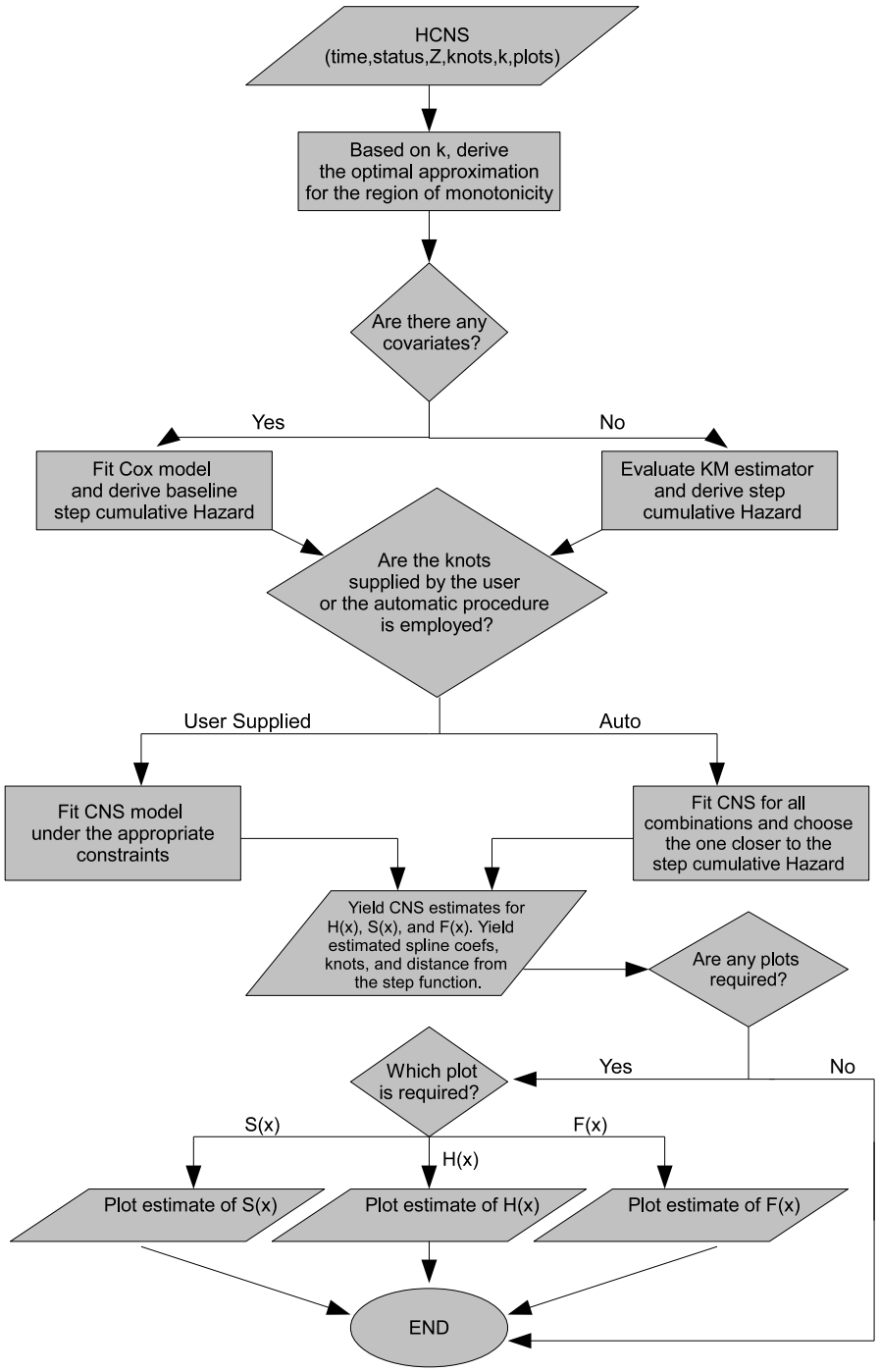

Figure B3: Flowchart of the underlying algorithm of the HCNS approach

- `Fx`: a "function handle" that can yield the value of the estimate of the cumulative distribution estimator based on the presented method, for any value(s) of x.

- `Sx`: a "function handle" that can yield the value of the estimate of the survival estimator based on the presented method, for any value(s) of x.
- `knots`: The knots used. If the knots were provided by the user then these knots are simply returned. If the knots were set to `auto` then the selected knot scheme is returned.
- `KMdist`: The sum of squares of the spline model from the corners of the step cumulative hazard function (that is the quantity presented in (.1)).
- `ghat`: the estimated cox coefficients if covariates are available. If there are no covariates then ghat is returned to be `NaN` (i.e. "not a number")

Next, we provide some examples that can be straightforwardly reproduced by the reader in MATLAB so as to clarify the input/output arguments:

Example 1. Generate some data ($n = 300$) from the Weibull distribution with parameters 2 and 3 and then apply the presented method:

```
n=300
x=wblrnd(2,3,n,1); %Generate some data from Weib(2,3)
c=wblrnd(2,3,n,1); %Generate the censoring variable (Weib(2,3))
xcen=min(x,c); %Create the censored data (expected censoring=50%)
status=(x>c); %Derive the censoring indicator

% The data for analysis are the variables xcen and status.
% The presented approach is carried out from the following line:

[bhat Hx Fx Sx KMdist knots gcoxhat]=HCNS(xcen, status, [],  ...
'auto', 2, 'survivor')
```

An optimal 18-gon is used to approximate the monotonicity region (since `k` is set to 2). The array `bhat` contains the spline coefficients estimates, and the `gcoxhat` is returned to be `NaN` since no covariates are available. The knot placement procedure is set to `'auto'`, thus 6 knots are used and 210 knot placement schemes are tested. The knots returned for this specific data set generated are `knots=[ 0.2591 0.5761 0.8931 1.2101 2.4781 2.7951]` (as can also be seen by B4) and the sum of the squared distance from the corners of the KM estimator is `KMdist=0.1114`. Alternatively the user could manually provide the knots desired. The plot of the survival estimate is asked by the user and the plot that is generated by the above code is given in Figure B4. The survival estimate is plotted up to the last event time, however the user can also plot the presented estimates beyond the last event time. The 'function handles' `Hx, Fx`,and `Sx` can be used to evaluate the proposed estimate at any time value. Of course, caution is needed when extrapolating the curves. For example, to evaluate the estimate at 0.5 1 1.5 and 2 we request:

```
Hx([0.5 1 1.5 2])
```

which yields as a result `0.0087 0.1276 0.3855 1.1723`. Similarly we derive:

```
Fx([0.5 1 1.5 2])
```
which yields `0.0087 0.1198 0.3199 0.6903` and
```
Sx([0.5 1 1.5 2])
```
which yields `0.9913 0.8802 0.6801 0.3097`

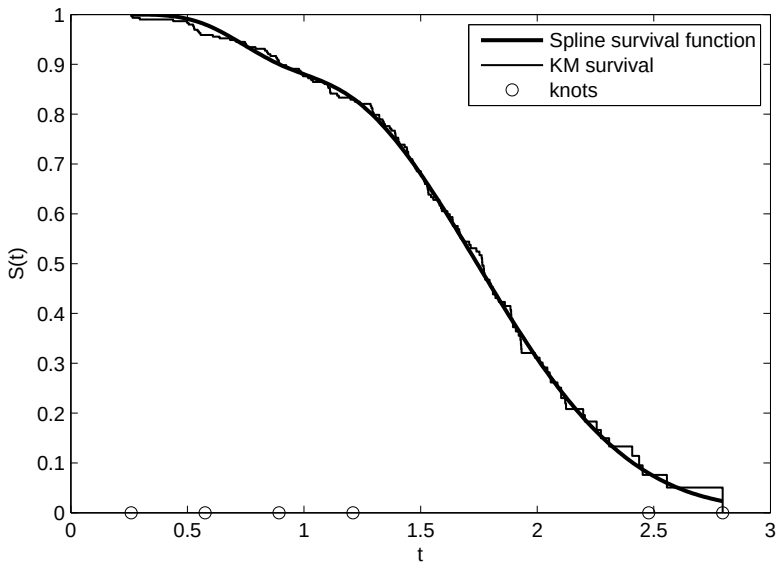

Figure B4: Survival estimate of the presented method as generated by the code in Example 1.

Example 2. In this example we will generate some time values from a Cox model and then use the presented routine for estimation:

```
n=300;
u=rand(n,1); % Generate n number from the Uniform(0,1)
a=2;b=3;g=2; % True parameters of the Weibull baseline survival,
%and true value of cox coefficient g=2.
z=exprnd(0.3,n,1); % Generate exponentially distributed
% covariate with mean 0.3
x=(-log(u)./(a.^(-b).*exp(g.*z))).^(1./b); %Generate values from
%the Cox model
c=exprnd(4,n,1); % Generate the censoring variable
status=(x>c); % Derive the censoring indicator
xcen=min(x,c); % Derive censored time values

% The data for analysis now are the variables xcen, status,
% and z (the covariate)
```

```
% We apply the presented approach by using only the
%following line:
[bhat Hx Fx Sx KMdist knots gcoxhat]=HCNS(xcen, status, z, ...
'auto', 2, 'survivor')
```

The interpretation of the output is similar to the one provided in Example 1. Here we derive a value of `ghat=2.041` which is the estimated coefficient of the covariate based on the usual Cox model. The plot requested now, is the survival estimate for the baseline survival, that is for $Z = 0$, and is presented in Figure B5.

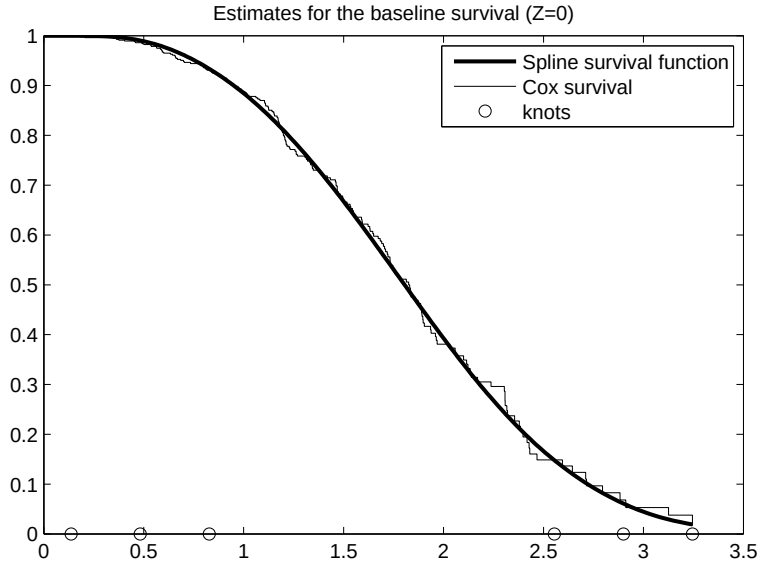

Figure B5: Survival estimate of the baseline survival (Z=0) as generated by the code in Example 2.

We base inference on the percentile bootstrap technique. `HCNSboots` can be used to obtain 95% confidence intervals for the cumulative hazard, survival, or cumulative distribution functions. The input/output arguments of the `HCNSboots` are the following:

INPUT:

- `time`: As defined in the `HCNS` routine.
- `status`: As defined in the `HCNS` routine.
- `Z`: As defined in the `HCNS` routine.
- `knots`: As defined in the `HCNS` routine. The `auto` option is still available. If chosen, then all knot combinations will be explores for each bootstrap sample.
- `kgon`: As defined in the `HCNS` routine.

- `CIat`: Time values at which the 95% confidence intervals are to be obtained.
- `boots`: The number of the bootstrap samples.

OPTIONAL INPUT:

- `profil`: As defined in the `HCNS` routine. If given, then the confidence intervals will be derived for the selected covariate profile. Obviously, input argument `Z` must be also given if this optional input argument is to be used.

OUTPUT:

- `CIH`: A two column matrix that contains the derived 95% bootstrap based confidence intervals for the cumulative hazard. Its left column are the lower confidence limits and its left refer to the upper ones. These confidence intervals refer to the specific profile `profil`, if provided.
- `CIS`: Confidence intervals for the survival function.
- `CIF`: Confidence intervals for the cumulative distribution function.

A usage example for the `HCNSboots` is given in the Application section.

2.4 Simulation Study

We present a small simulation study to evaluate the approach when the automatic knot placement is selected and compare the results with the KM estimator using the Mean Integrated Squared Error criterion $(= E(\int(\hat{S} - S)^2))$. Since the KM estimator cannot provide estimation beyond the last event time we consider that the KM estimate of the survival beyond the last event time equal to $\hat{S}^{KM}(t_{max})$, where $t_{max} = max(event\ time)$. This is proposed by Efron (1967). For this reason we integrated the squared error from 10th to 90th, 20th to 80th and 30th to 70th percentiles of the true survival functions. That is, we obtained $ISE_{(10-90)} = \int_{10th}^{90th}(\hat{S} - S)^2$, $ISE_{(20-80)} = \int_{20th}^{80th}(\hat{S} - S)^2$, and $ISE_{(30-70)} = \int_{30th}^{70th}(\hat{S} - S)^2$ for 1000 repetitions and then we evaluated the average to obtain the corresponding MISE's: $MISE_{(10-90)}, MISE_{(20-80)}$ and $MISE_{(30-70)}$. We used the knot placement set in `auto` and $k = 2$. This means that an optimal 18-gon is used to approximate the region of monotonicity for each iteration. We also report the CPU time needed on average for each iteration. All simulations were carried out on a laptop with a 2.8GHz processor.

We observe that in nearly all cases the presented approach outperforms the KM estimator. Most differences however seem to appear for heavier censoring and/or lower sample sizes. From Table 1 we observe that the computational time in using the HCNS routine is low (about 2 seconds in most cases). We also checked the time needed to perform analysis for simulated data sets where the sample size was 10,000 with 50% expected censoring. That was about 5.5 seconds.

2.5 Examples

We also use the routine to apply the HCNS approach to a real data set with an available covariate. The data set is presented by Kardaun (1983) and refer to 90 males with cancer

Table B1: Simulation results of 1000 repetitions for the KM and HCNS approach. The $MISE$ is calculated for the 10th to 90th, 20th to 80th, and 30th to 70th percentile of the real underlying densities. The mean CPU time of the HCNS approach is also reported.

			Kaplan Meier			HCNS			CPU time
Distribution	n	Censoring	10-90	20-80	30-70	10-90	20-80	30-70	sec./iteration
		30%	0.0030	0.0023	0.0016	0.0035	0.0024	0.0015	2.0051
	300	50%	0.0040	0.0030	0.0020	0.0043	0.0030	0.0019	1.7762
		70%	0.0069	0.0049	0.0031	0.0065	0.0044	0.0028	1.7452
		30%	0.0089	0.0068	0.0045	0.0095	0.0068	0.0044	1.6796
Wei(2,3)	100	50%	0.0121	0.0090	0.0059	0.0121	0.0086	0.0055	1.6824
		70%	0.0210	0.0144	0.0092	0.0189	0.0134	0.0083	1.8009
		30%	0.0179	0.0135	0.0091	0.0181	0.0131	0.0084	1.7070
	50	50%	0.0245	0.0180	0.0119	0.0237	0.0171	0.0108	1.7350
		70%	0.0437	0.0302	0.0192	0.0374	0.0284	0.0184	1.6439
		30%	0.0021	0.0016	0.0011	0.0020	0.0015	0.0010	2.1706
	300	50%	0.0029	0.0022	0.0015	0.0026	0.0020	0.0013	1.9289
		70%	0.0046	0.0035	0.0024	0.0043	0.0032	0.0022	1.9150
		30%	0.0063	0.0049	0.0033	0.0060	0.0045	0.0030	1.8411
Wei(7,8)	100	50%	0.0086	0.0066	0.0045	0.0083	0.0062	0.0041	1.8484
		70%	0.0138	0.0107	0.0072	0.0135	0.0100	0.0065	1.8685
		30%	0.0129	0.0099	0.0067	0.0131	0.0094	0.0062	1.8511
	50	50%	0.0176	0.0135	0.0091	0.0177	0.0129	0.0084	1.8230
		70%	0.0287	0.0218	0.0146	0.0286	0.0209	0.0134	1.7030
		30%	0.0028	0.0023	0.0018	0.0028	0.0023	0.0018	2.3636
	300	50%	0.0038	0.0032	0.0024	0.0037	0.0031	0.0024	2.0029
		70%	0.0059	0.0050	0.0038	0.0057	0.0046	0.0036	1.9281
		30%	0.0080	0.0069	0.0053	0.0078	0.0065	0.0050	1.9943
Mixture	100	50%	0.0113	0.0096	0.0074	0.0107	0.0089	0.0068	1.9582
Weibulls		70%	0.0185	0.0156	0.0119	0.0172	0.0145	0.0109	1.9862
		30%	0.0158	0.0134	0.0104	0.0152	0.0125	0.0095	2.1252
	50	50%	0.0227	0.0192	0.0147	0.0220	0.0183	0.0137	2.1016
		70%	0.0389	0.0325	0.0247	0.0355	0.0303	0.0230	1.7753

of the larynx at a Dutch hospital during the period 1970-1978. The data contain the time to event (death) or censoring, the age of the patient as well as the stage of the disease. There are four stages which are ordered from the least serious to the most serious (stage 1 is the least serious). The data are publicly available and a usual Cox analysis is also presented in Klein and Moeschberger (2003). This data is analyzed with the proposed method in Appendix C with a much simpler knot placement rule than the one provided by the `auto` option of the software described here. Here, we present the following code to derive survival estimates for each stage at the mean of age (64.6111) and the reproduction of Figure B6.

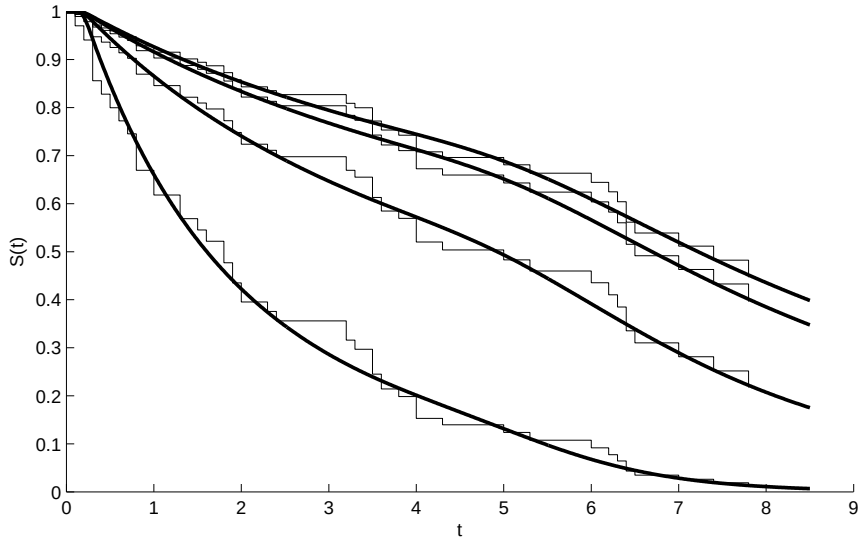

Figure B6: Survival estimates for the four stages of larynx cancer at the mean age (64.6111). The solid lines refer to the HCNS approach, and the step functions refer to the corresponding Cox estimates. For the HCNS approach the `auto` procedure was used for the knot placement.

```
data=[...]  % A matrix that contains the data.  Each column is
%one variable.
% 1st column contains the cancer stage variable,
% 2nd:time variable, 3rd:age, 4th:status
stage=data(:,1);time=data(:,2); %derive variables from the
%data matrix.
age=data(:,3);status=data(:,4);
% In the original data set code 0 is for censoring and
%1 for death.  MATLAB needs the opposite coding:
status=status.*(-1)+1;
Z=[(stage==2) (stage==3) (stage==4) age]; % build the
%covariate matrix

% Now the data are ready for analysis
```

```
% Fit cox model to derive baseline S0 (for Z=0):
[bcox logL H stats] = coxphfit(Z,time,'censoring',status, ...
'baseline',0);
S0cox=exp(-H(:,2)); % this is the cox baseline survival.

% Apply the HCNS routine to estimate baseline functions:
[bhat H0 F0 S0 KMdist knots gcoxhat]=HCNS(time, status, Z, ...
'auto', 2);
g2=gcoxhat(1);g3=gcoxhat(2);g4=gcoxhat(3);gage=gcoxhat(4);
% These are the cox coefs.

gr=0:0.01:8.5; % construct a grid of points over which the
%survival estimates will be plotted.
figure
hold on
% Plot the HCNS survival estimates for each stage:
plot(gr,S0(gr).^(exp(gage.*mean(age))),'k','LineWidth',2)
plot(gr,S0(gr).^(exp(gage.*mean(age)+g2)),'k','LineWidth',2)
plot(gr,S0(gr).^(exp(gage.*mean(age)+g3)),'k','LineWidth',2)
plot(gr,S0(gr).^(exp(gage.*mean(age)+g4)),'k','LineWidth',2)

% Plot the Cox survival estimates for each stage:
stairs(H(:,1),S0cox.^(exp(gage.*mean(age))),'k')
stairs(H(:,1),S0cox.^(exp(gage.*mean(age)+g2)),'k')
stairs(H(:,1),S0cox.^(exp(gage.*mean(age)+g3)),'k')
stairs(H(:,1),S0cox.^(exp(gage.*mean(age)+g4)),'k')
xlabel('t');ylabel('S(t)')
```

From the generated Figure B6 we observe that, at the mean of age (64.6111), the survival curve of stage 1 provides higher survival probabilities. As we move to stage 4 we observe that the survival curves yield lower survival probabilities which is what we expect.

In the previous example we estimated the baseline survival and used the Cox model formulation to obtain estimates for the desired profile. If, for example, we were interested in estimating directly the survival of an individual with age equal to 64.6111 and the most serious cancer stage, i.e. $S(t|age = 64.6111, stage = 4)$ for $t = 1, 2, \ldots, 6$ along with the corresponding 95% confidence intervals for the survival estimate, based on 300 bootstrap samples, we could simply type:

```
% Derive the desired estimates for the specific profile
%[0 0 1 mean(age)]:
[bhat Hx Fx Sx]=HCNS(time, status, Z, 'auto', 2,...
'none', [0 0 1 mean(age)] );
```

```
Shat=Sx(1:6) % obtained survival estimate values for t=1,2,...,6
Shat=
0.9266 0.8535 0.7945 0.7446 0.6885 0.6094
%Now derive the corresponding 95% CI's for these estimates:
[CIH CIS CIF]=HCNSboots(x, status, Z, 'auto', 2, [1:6], 300, [0 0
1 mean(age)])
CIS=
0.8861 0.9898
0.7912 0.9304
0.7224 0.8835
0.6409 0.8508
0.5572 0.8324
0.4739 0.7702
```

The left column of `CIS` contains lower confidence limits and the right the upper ones. In tables `CIH` and `CIF` the corresponding confidence intervals for the cumulative hazard and cumulative distribution respectively are also provided. Note that in this example we used the `auto` procedure for the knot placement. Hence, in each bootstrap sample 420 minimizations are employed. In effect, even though the procedure is feasible, it can be time consuming. It took about 20 minutes to derive the CI's mentioned above. In the case where the knots are supplied by the user then this time is significantly reduced. We used `HCNSboots` with knots set equal to the ones derived by the `HCNS` estimate analysis using the `auto` option. The time needed for the `HCNSboots` to perform all calculations (using again 300 bootstrap samples) was about 8 seconds, yielding the following CI's:

```
CIS =
0.8835 0.9695
0.7757 0.9232
0.6938 0.8882
0.6309 0.8706
0.5559 0.8397
0.4660 0.7776
```

2.6 Code for approximating the region of monotonicity

The following code provides, for a given $k = 1, 2, \ldots$, the optimal in terms of inscribed area $8k + 2$-gon for approximating the region of monotonicity $\mathcal{M}$. (If k is not set to be a positive integer then an error will appear.)

```
k=3;% This can be set to any positive integer
%for finer approximations
xc=2;yc=2; % center of the ellipse
theta=-pi/4; % tilt of the major axis
a=2.44949; % major semi-axis of the ellipse
b=sqrt(2); % minor semi-axis of the ellipse
N=12*k; % Points to approximate the ellipse
```

```
N=N+1;

% ------Apply Smith's algorithm to approximate the ellipse:
df=2*pi/(N-1);
CT=cos(theta);ST=sin(theta);
x=zeros(1,N);y=x; %preallocation
for n=1:N
    xp=a*CNDP;
    yp=b*SNDP;
    x(n)=xc+xp*CT-yp*ST;
    y(n)=yc+xp*ST+yp*CT;
    TEMP=CNDP*CDP-SNDP*SDP;
    SNDP=SNDP*CDP+CNDP*SDP;
    CNDP=TEMP;
end

% Plot the approximation of the ellipse:
plot(x,y,'.k');hold on
axis square;axis([0 4 0 4]);

%Re arrange data so as the first point to be (x,y)=(3,0):
n=N-1;kk=n/12;
M=[x' y'];M(max(size(M)),:)=[];M=circshift(M,kk);
x=M(:,1);y=M(:,2);

%Plot the approximation of the region of monotonicity:
plot(x(1:(8*k+1)),y(1:(8*k+1)),'o-k')
plot([3 0 3 0],[3 3 0 0],'ok')
xlabel('a');ylabel('b')

%Shade the approximated region:
fill([x(1:(8*k+1));0],[y(1:(8*k+1));0],[0.7,0.7,0.7]);alpha(0.5)
hold off
```

3 Appendix C

3.1 Simulation Studies for the evaluation of the HCNS method

In the simulation studies we considered the proposed (HCNS, Hazard Constrained Natural Spline) method, as well as the logspline and the Kaplan Meier methods. For the logspline approach we use the `oldlogspline` function provided by Kooperberg and Stone (1992) in the `polspline` library of `R`, with stepwise deletion of knots (in contrast to the `logspline` function provided also by Kooperberg, the `oldlogspline` function can handle censored data). For the HCNS method we used the `lsqlin` function of `MATLAB` to minimize the constrained least squares. The `lsqlin` function allows the user to directly provide both equality and inequality constraints, thus it is not necessary to write our single equality constraint as two inequalities. The `MATLAB` code is available

upon request. We used the MISE criterion (= $E(\int(\hat{S}-S)^2)$) for comparison of the methodologies based on 1000 repetitions for each scenario. For all three approaches we considered their performances in the 10th to 90th, 20th to 80th and 30th to 70th percentiles of the true survival functions (that is, we calculated the $ISE_{(10-90)} = \int_{10th}^{90th}(\hat{S}-S)^2$, $ISE_{(20-80)} = \int_{20th}^{80th}(\hat{S}-S)^2$, and $ISE_{(30-70)} = \int_{30th}^{70th}(\hat{S}-S)^2$ for 1000 repetitions and we used the average to obtain the associated MISE's). Due to the fact that the Kaplan Meier estimator is limited up to the last event (t_{max}), we considered expanding the estimation of the survival function to be $\hat{S}^{KM}(t_{max})$ beyond t_{max} when necessary (i.e. in the repetitions where the last event occurred earlier that X_{90}). A small sample study presented in Klein (1991) concludes that this approach, which is taken in Gill (1980), is more preferable for small sample sizes than the one one taken in Efron (1967) where $S(t) = 0, \forall\, t > t_{max}$.

For the HCNS approach we considered 6 knots. Harrell (2001) suggests that the principal decision is between 3, 4, or 5 knots for a natural cubic spline. However, the proposed approach deals with monotonicity constraints. Moreover, our spline model for the cumulative hazard is zero before the first knot and linear beyond the last knot. Thus, the flexibility of the spline would be seriously compromised when using just 3 knots. As seen by the simulation results the choice of 6 knots performed very well overall. We also conducted simulations studies for 4 and 5 knots but the results were unsatisfactory and are omitted for brevity.

We considered three main scenarios for the true density of X shown in Figure B7. In all scenarios we considered the proposed, the Kaplan Meier and the logspline method. The censoring variable C was taken to be exponentially distributed with an appropriate parameter so as to achieve expected levels of censoring of 30%, 50% and 70%. The sample sizes considered were 60, 100 and 300 (we did not consider smaller sample sizes due to convergence problems of the logspline approach that increased for small samples and heavy censoring). The results are presented in Table C1.

At the first scenario we generated data from a positive skewed distribution, $Weib(1.5, 4)$. We observed that the logspline technique yielded somewhat better results in all cases (i.e. all sample sizes and censoring levels) in terms of MISE while the proposed method yielded somewhat better results than the Kaplan Meier in nearly all cases. Note that as the level of censoring increases the logspline technique had a few problems of convergence where the proposed technique did not. In these cases we discarded the repetitions that the logspline technique yielded a convergence warning.

At the second scenario we generated data from an approximately bell shaped distribution, the $Weib(3, 3)$. We observed that the logspline technique, again, yielded better results in terms of MISE while the proposed method yielded somewhat better results from the Kaplan Meier. However, the logspline approach suffered from multiple convergence problems in cases of heavy censoring (70%). In cases where the convergence problems are about 10% or more, the results are not presented.

At the third scenario we considered the bimodal mixture of distributions $0.5Weib(5, 4)+$ $0.5Weib(4, 2)$. We observe that the HCNS method yields smaller MISE with minor differences from the logspline approach and the Kaplan Meier in cases of smaller sample sizes. Note again that in cases of 70% censoring the logspline approach has approximately 10% convergence problems and simulation results are not presented.

Generally, the simulations did not reveal major differences regarding the performance of the compared methods. The differences seem to be in favor of the logspline density

Table C1: Simulation results of 1000 repetitions for the Kaplan Meier, HCNS and logspline methodologies. Sample size is $n = 300$, $n = 100$ and $n = 60$ with 70%, 50% and 30% censoring. The mean integrated squared error is calculated for the 10th to 90th, 20th to 80th, and 30th to 70th percentile of the real corresponding distribution functions.

			Kaplan Meier			HCNS			Logspline		
Dist.	n	Cens.	10-90	20-80	30-70	10-90	20-80	30-70	10-90	20-80	30-70
		30%	0.0050	0.0038	0.0025	0.0055	0.0041	0.0026	0.0044	0.0034	0.0024
	300	50%	0.0073	0.0052	0.0034	0.0075	0.0053	0.0034	0.0059	0.0044	0.0029
		70%	0.0166	0.0099	0.0059	0.0139	0.0088	0.0052	$0.0105^{(1)}$	$0.0071^{(1)}$	$0.0043^{(1)}$
		30%	0.0151	0.0112	0.0074	0.0166	0.0121	0.0076	0.0119	0.0090	0.0061
Wei(1.5,4)	100	50%	0.0221	0.0156	0.0099	0.0215	0.0149	0.0093	0.0161	0.0120	0.0079
		70%	0.0529	0.0318	0.0185	0.0388	0.0275	0.0168	$0.0293^{(3)}$	$0.0208^{(3)}$	$0.0130^{(3)}$
		30%	0.0257	0.0194	0.0129	0.0263	0.0194	0.0124	0.0199	0.0154	0.0105
	60	50%	0.0388	0.0272	0.0174	0.0345	0.0249	0.0157	0.0268	0.0202	0.0134
		70%	0.0879	0.0549	0.0323	0.0606	0.0455	0.0291	$0.0487^{(2)}$	$0.0347^{(2)}$	$0.0222^{(2)}$
		30%	0.0021	0.0016	0.0011	0.0020	0.0015	0.0009	0.0019	0.0015	0.0010
	300	50%	0.0030	0.0023	0.0015	0.0027	0.0020	0.0014	0.0025	0.0020	0.0014
		70%	0.0047	0.0035	0.0024	0.0042	0.0032	0.0021	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$
		30%	0.0063	0.0049	0.0033	0.0059	0.0044	0.0030	0.0049	0.0040	0.0027
Wei(3,3)	100	50%	0.0086	0.0066	0.0044	0.0079	0.0060	0.0040	$0.0063^{(1)}$	$0.0051^{(1)}$	$0.0035^{(1)}$
		70%	0.0140	0.0105	0.0070	0.0121	0.0095	0.0063	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$
		30%	0.0109	0.0085	0.0058	0.0100	0.0077	0.0051	0.0084	0.0067	0.0046
	60	50%	0.0151	0.0117	0.0080	0.0132	0.0104	0.0069	0.0107	0.0087	0.0060
		70%	0.0249	0.0186	0.0122	0.0206	0.0164	0.0110	$0.0155^{(4)}$	$0.0125^{(4)}$	$0.0089^{(4)}$
		30%	0.0027	0.0023	0.0017	0.0030	0.0023	0.0017	0.0026	0.0021	0.0016
	300	50%	0.0038	0.0031	0.0023	0.0040	0.0032	0.0023	0.0035	0.0029	0.0022
		70%	0.0068	0.0053	0.0037	0.0064	0.0051	0.0036	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$
		30%	0.0079	0.0066	0.0049	0.0075	0.0062	0.0045	0.0081	0.0069	0.0050
Mixt.	100	50%	0.0110	0.0090	0.0066	0.0102	0.0085	0.0061	$0.0119^{(3)}$	$0.0100^{(3)}$	$0.0074^{(3)}$
		70%	0.0207	0.0161	0.0113	0.0178	0.0150	0.0108	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$
		30%	0.0144	0.0120	0.0088	0.0133	0.0110	0.0080	0.0149	0.0128	0.0094
	60	50%	0.0193	0.0157	0.0115	0.0170	0.0142	0.0103	$0.0191^{(1)}$	$0.0167^{(1)}$	$0.0123^{(1)}$
		70%	0.0350	0.0275	0.0193	0.0284	0.0243	0.0181	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$

(1) 0.5% of the iterations were discarded due to convergence problems
(2) 1.5% of the iterations were discarded due to convergence problems
(3) 2.0% of the iterations were discarded due to convergence problems
(3) 4.0% of the iterations were discarded due to convergence problems
(5) Simulation was not conducted due to $\geq 10\%$ convergence problems

Table C2: Simulation results of 1000 repetitions for the coverage of the bootstrap (300 samples for each repetition) for the 10-th to 90-th percentile of the true underlying density. Sample size is $n = 300$, $n = 100$ and $n = 60$ with 70%, 50% and 30% censoring.

			Coverage of the HCNS method for the corresponding percentiles								
Dist.	Sample	Cens.	X_{10}	X_{20}	X_{30}	X_{40}	X_{50}	X_{60}	X_{70}	X_{80}	X_{90}
		30%	0.772	0.948	0.972	0.961	0.946	0.957	0.957	0.952	0.945
	300	50%	0.902	0.964	0.968	0.945	0.953	0.958	0.957	0.935	0.900
		70%	0.962	0.978	0.957	0.956	0.947	0.953	0.940	0.912	0.886
		30%	0.870	0.958	0.976	0.966	0.959	0.957	0.956	0.945	0.917
Wei(1.5,4)	100	50%	0.929	0.965	0.951	0.934	0.946	0.961	0.956	0.916	0.878
		70%	0.958	0.947	0.942	0.947	0.942	0.913	0.892	0.871	0.896
		30%	0.890	0.951	0.959	0.952	0.939	0.946	0.941	0.923	0.874
	60	50%	0.924	0.968	0.953	0.949	0.950	0.946	0.944	0.912	0.863
		70%	0.973	0.945	0.933	0.928	0.895	0.864	0.845	0.838	0.905
		30%	0.958	0.955	0.956	0.953	0.947	0.945	0.943	0.946	0.932
	300	50%	0.964	0.958	0.957	0.958	0.950	0.950	0.949	0.937	0.911
		70%	0.970	0.956	0.951	0.951	0.946	0.942	0.934	0.932	0.901
		30%	0.964	0.954	0.948	0.949	0.950	0.952	0.938	0.922	0.900
Wei(3,3)	100	50%	0.953	0.941	0.944	0.950	0.955	0.944	0.932	0.922	0.884
		70%	0.966	0.944	0.941	0.945	0.940	0.935	0.924	0.880	0.861
		30%	0.958	0.943	0.942	0.948	0.950	0.942	0.935	0.919	0.894
	60	50%	0.968	0.937	0.938	0.943	0.937	0.928	0.916	0.883	0.834
		70%	0.939	0.925	0.925	0.930	0.920	0.892	0.854	0.820	0.824
		30%	0.929	0.954	0.956	0.954	0.959	0.941	0.948	0.947	0.937
	300	50%	0.941	0.946	0.956	0.966	0.956	0.947	0.938	0.939	0.917
		70%	0.944	0.942	0.953	0.962	0.954	0.944	0.931	0.917	0.917
		30%	0.964	0.945	0.936	0.949	0.954	0.949	0.946	0.931	0.899
Mixt.	100	50%	0.964	0.935	0.922	0.940	0.939	0.944	0.933	0.913	0.913
		70%	0.952	0.950	0.949	0.951	0.953	0.950	0.918	0.903	0.930
		30%	0.972	0.950	0.936	0.941	0.938	0.941	0.925	0.918	0.912
	60	50%	0.957	0.956	0.950	0.950	0.964	0.941	0.925	0.887	0.900
		70%	0.925	0.916	0.915	0.926	0.905	0.874	0.843	0.819	0.892

Table C3: Simulation results of 1000 repetitions for the coverage of the Greenwood formula for the 10-th to 90$-th$ percentile of the true underlying density. Sample size is $n = 300$, $n = 100$ and $n = 60$ with 70%, 50% and 30% censoring.

			Coverage of the Greenwood formula for the corresponding percentiles								
Dist.	Sample	Cens.	X_{10}	X_{20}	X_{30}	X_{40}	X_{50}	X_{60}	X_{70}	X_{80}	X_{90}
		30%	0.951	0.946	0.950	0.948	0.946	0.952	0.947	0.937	0.942
	300	50%	0.937	0.940	0.944	0.940	0.935	0.945	0.954	0.928	0.927
		70%	0.948	0.936	0.951	0.966	0.955	0.938	0.945	0.928	0.833
		30%	0.911	0.923	0.942	0.937	0.934	0.951	0.930	0.935	0.916
Wei(1.5,4)	100	50%	0.927	0.926	0.947	0.960	0.954	0.952	0.935	0.926	0.869
		70%	0.924	0.937	0.945	0.938	0.934	0.929	0.913	0.855	0.588
		30%	0.916	0.933	0.943	0.951	0.956	0.943	0.936	0.906	0.885
	60	50%	0.916	0.933	0.939	0.942	0.933	0.929	0.931	0.921	0.793
		70%	0.910	0.919	0.918	0.927	0.915	0.907	0.884	0.773	0.533
		30%	0.942	0.939	0.942	0.937	0.945	0.952	0.962	0.945	0.939
	300	50%	0.944	0.952	0.939	0.954	0.949	0.949	0.945	0.956	0.945
		70%	0.949	0.936	0.932	0.942	0.946	0.940	0.937	0.917	0.897
		30%	0.927	0.939	0.954	0.947	0.940	0.943	0.926	0.933	0.925
Wei(3,3)	100	50%	0.929	0.923	0.942	0.930	0.935	0.938	0.934	0.932	0.904
		70%	0.913	0.938	0.936	0.934	0.936	0.936	0.936	0.922	0.812
		30%	0.895	0.926	0.941	0.939	0.939	0.950	0.942	0.932	0.909
	60	50%	0.900	0.923	0.931	0.929	0.931	0.929	0.925	0.914	0.867
		70%	0.885	0.919	0.921	0.923	0.924	0.918	0.898	0.863	0.678
		30%	0.937	0.943	0.952	0.945	0.946	0.947	0.942	0.933	0.924
	300	50%	0.929	0.940	0.948	0.951	0.953	0.953	0.953	0.948	0.918
		70%	0.937	0.943	0.944	0.945	0.952	0.948	0.932	0.922	0.896
		30%	0.948	0.943	0.941	0.951	0.949	0.946	0.943	0.918	0.926
Mixt.	100	50%	0.922	0.933	0.936	0.949	0.945	0.948	0.941	0.916	0.883
		70%	0.913	0.934	0.932	0.944	0.931	0.933	0.923	0.899	0.745
		30%	0.920	0.944	0.937	0.949	0.951	0.959	0.941	0.941	0.883
	60	50%	0.906	0.909	0.921	0.933	0.931	0.939	0.931	0.906	0.854
		70%	0.879	0.931	0.941	0.932	0.941	0.933	0.893	0.816	0.581

Table C4: Simulation results of 1000 repetitions for the coverage of the Greenwood formula for the 10-th to 90$-th$ percentile of the true underlying density. Sample size is $n = 300$, $n = 100$ and $n = 60$ with 70%, 50% and 30% censoring.

			Coverage of the Greenwood formula for the corresponding percentiles								
Dist.	Sample	Cens.	X_{10}	X_{20}	X_{30}	X_{40}	X_{50}	X_{60}	X_{70}	X_{80}	X_{90}
		30%	0.951	0.946	0.950	0.948	0.946	0.952	0.947	0.937	0.942
	300	50%	0.937	0.940	0.944	0.940	0.935	0.945	0.954	0.928	0.927
		70%	0.948	0.936	0.951	0.966	0.955	0.938	0.945	0.928	0.833
		30%	0.911	0.923	0.942	0.937	0.934	0.951	0.930	0.935	0.916
Wei(1.5,4)	100	50%	0.927	0.926	0.947	0.960	0.954	0.952	0.935	0.926	0.869
		70%	0.924	0.937	0.945	0.938	0.934	0.929	0.913	0.855	0.588
		30%	0.916	0.933	0.943	0.951	0.956	0.943	0.936	0.906	0.885
	60	50%	0.916	0.933	0.939	0.942	0.933	0.929	0.931	0.921	0.793
		70%	0.910	0.919	0.918	0.927	0.915	0.907	0.884	0.773	0.533
		30%	0.942	0.939	0.942	0.937	0.945	0.952	0.962	0.945	0.939
	300	50%	0.944	0.952	0.939	0.954	0.949	0.949	0.945	0.956	0.945
		70%	0.949	0.936	0.932	0.942	0.946	0.940	0.937	0.917	0.897
		30%	0.927	0.939	0.954	0.947	0.940	0.943	0.926	0.933	0.925
Wei(3,3)	100	50%	0.929	0.923	0.942	0.930	0.935	0.938	0.934	0.932	0.904
		70%	0.913	0.938	0.936	0.934	0.936	0.936	0.936	0.922	0.812
		30%	0.895	0.926	0.941	0.939	0.939	0.950	0.942	0.932	0.909
	60	50%	0.900	0.923	0.931	0.929	0.931	0.929	0.925	0.914	0.867
		70%	0.885	0.919	0.921	0.923	0.924	0.918	0.898	0.863	0.678
		30%	0.937	0.943	0.952	0.945	0.946	0.947	0.942	0.933	0.924
	300	50%	0.929	0.940	0.948	0.951	0.953	0.953	0.953	0.948	0.918
		70%	0.937	0.943	0.944	0.945	0.952	0.948	0.932	0.922	0.896
		30%	0.948	0.943	0.941	0.951	0.949	0.946	0.943	0.918	0.926
Mixt.	100	50%	0.922	0.933	0.936	0.949	0.945	0.948	0.941	0.916	0.883
		70%	0.913	0.934	0.932	0.944	0.931	0.933	0.923	0.899	0.745
		30%	0.920	0.944	0.937	0.949	0.951	0.959	0.941	0.941	0.883
	60	50%	0.906	0.909	0.921	0.933	0.931	0.939	0.931	0.906	0.854
		70%	0.879	0.931	0.941	0.932	0.941	0.933	0.893	0.816	0.581

when we are dealing with a unimodal density, and in favor of the proposed technique in the case of the bimodal density for smaller sample sizes. The main advantage of the proposed approach is that it converges where the logspline technique might not (i.e in cases of small sample sizes or heavy censoring (or both)). The Kaplan Meier yielded consistently greater values of MISE compared two the other two approaches.

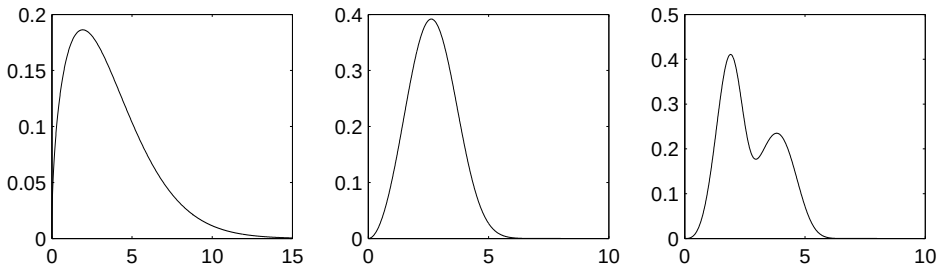

Figure B7: Real densities used in simulations, Left: Weibull(1.5,4), Middle: Weibull(3,3), Right: $0.5Weib(5,4) + 0.5Weib(4,2)$.

For the derivation of confidence intervals of the survival function we explored the percentile bootstrap technique. We consider resampling pairs of the form $\{T_i, D_i\}$ from the available data and considered 300 bootstrap samples in each repetition. We estimated the coverage of confidence intervals for the $10-th$, $20-th$, ..., and $90-th$ percentile of the true underlying distribution. The results are presented in Table C2. These coverage estimates were compared to those based on the well known Greenwood formula used to derive confidence intervals for the Kaplan Meier survival funcion presented in Table C2) of the Greenwood formula are presented in Table C4. We observe that the coverage provided by the percenile bootstrap for the HCNS approach is satisfactory in the range of the $20-th$ to the $80-th$ percentile where enough data are available. Moreover, the coverage of the confidence intervals based on our proposed methodology is satisfactory. Furthermore, the coverage achieved by the HCNS confidence intervals for the $90-th$ percentile is markedly better in all cases compared to the Greenwood's one. For example, in the scenario of the Weibull mixture when $n = 100$ with 70% censoring, we achieved 93% coverage while the corresponding coverage for the Kaplan Meier estimator yields 74.5%.

We conducted an additional simulation study to evaluate and compare our method to the semiparametric estimator of the survival function in the Cox model. We also consider the method taken in Herndon and Harrell (1995) (restricted cubic spline (or RCS method) in that the spline is restricted to have linear tails). They restate the Cox model using a natural cubic spline as a baseline hazard and the estimation is done simultaneously for all parameters. In their approach, no restrictions are imposed to any of the parameters of interest and initial values are required. We used the fitted Cox model to derive an initial value for the coefficient of the covariate. In their approach the spline model also includes an intercept, the initial value of which was derived by assumming an exponential distribution for the baseline hazard rate. For all other parameters we set zeros as initial values. We used 4 knots for their approach placed at equally spaced percentiles that are calculated by the uncensored data.

One covariate, Z, was considered, taken to be exponentially distributed with mean equal to 0.3. The sample sizes and expected levels of censoring were set identical to the ones in the previous simulations. The real value of the coefficient γ of the covariate of the Cox model was set equal to 2. The baseline distribution (at $Z = 0$) was taken to be a Weibull(2,3). We used the estimated coefficient $\hat{\gamma}$ obtained by the fitted Cox model to estimate $S(x|Z = 0.3)$. The ISE for the percentiles was used as in the previous simulations (for the semiparametric estimator of the survival function beyond t_{max} we set $\hat{S}(t_{max}|Z = z)$ when the larger time was a censoring time). The results for 1000 repetitions are shown in Table C5. For larger sample sizes we observe minor differences between the two approaches. However, the proposed approach yields in almost all cases smaller MISE, particularly for smaller sample sizes and/or heavier censoring. We note that in the scenario where $n = 60$ with 70% censoring, $min(T_i|D_i = 1)$ was very often identical to the $5-$th percentile, which led to convergence problems concerning the knot placement scheme that uses $min(T_i|D_i = 1)$ as the first knot. When this occurred only the other two schemes were considered. We note that the RCS method yielded in general smaller MISE but suffered from some convergence problems.

For the derivation of confidence intervals we used the percentile bootstrap technique considering the covariate values fixed and resampling pairs of the form (T_i, D_i). The simulation presented in Table C6 shows satisfactory coverage properties of this resampling technique. In the same table the corresponding coverages of the survival confidence intervals for the Cox model are also presented (see Link (1984)). We observe that at X_{10} and X_{90} the coverage provided by the proposed method is significantly better as the censoring gets heavier. For example, at X_{90} the coverage based on the traditional approach for the Cox model is 79%, for $n = 100$ and 70% censoring. This drops to 73% for $n = 60$ and 70% censoring. In contrast, the coverages provided by the proposed technique are 95.1% and 96.7% respectively. Similar results were obtained when comparing the coverage of the two methods at X_{10} for small samples and heavy censoring.

Table C5: Simulation results of 1000 repetitions for the Kaplan Meier, HCNS and RCS methods in the case of a covariate Z. The mean integrated squared error is calculated for the 10th to 90th, 20th to 80th, and 30th to 70th percentile of the real corresponding distribution functions. The covariate is exponentially distributed with mean 0.3. The HCNS and Cox model estimate $S(x|Z = 0.3)$ (baseline is a Weibull(3,2) and the real value of the parameter of the Cox model is $\gamma = 2$.)

		Cox model			HCNS			RCS		
Sample	Cens.	10-90	20-80	30-70	10-90	20-80	30-70	10-90	20-80	30-70
300	30%	0.0013	0.0010	0.0007	0.0013	0.0010	0.0006	$0.0012^{(1)}$	$0.0009^{(1)}$	$0.0007^{(1)}$
	50%	0.0017	0.0013	0.0009	0.0017	0.0013	0.0008	$0.0015^{(2)}$	$0.0012^{(2)}$	$0.0008^{(2)}$
	70%	0.0028	0.0021	0.0015	0.0027	0.0020	0.0013	$0.0023^{(4)}$	$0.0018^{(4)}$	$0.0012^{(4)}$
100	30%	0.0039	0.0030	0.0021	0.0039	0.0029	0.0019	$0.0031^{(3)}$	$0.0024^{(3)}$	$0.0016^{(3)}$
	50%	0.0051	0.0039	0.0026	0.0051	0.0037	0.0024	$0.0040^{(3)}$	$0.0032^{(3)}$	$0.0021^{(3)}$
	70%	0.0087	0.0064	0.0042	0.0081	0.0061	0.0039	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$
60	30%	0.0062	0.0048	0.0033	0.0062	0.0047	0.0031	$0.0054^{(2)}$	$0.0042^{(2)}$	$0.0028^{(2)}$
	50%	0.0086	0.0066	0.0045	0.0084	0.0064	0.0043	$0.0072^{(2)}$	$0.0056^{(2)}$	$0.0038^{(2)}$
	70%	0.0150	0.0110	0.0074	0.0135	0.0106	0.0070	$-^{(5)}$	$-^{(5)}$	$-^{(5)}$

(1) Approximately 5% of the iterations were discarded due to convergence problems
(2) Approximately 6% of the iterations were discarded due to convergence problems
(3) Approximately 7% of the iterations were discarded due to convergence problems
(4) Approximately 9% of the iterations were discarded due to convergence problems
(5) Simulation was not conducted due to $>$ 10% convergence problems

Table C6: Simulation results of 1000 repetitions for the coverage of the bootstrap (300 samples for each repetition) for the 10-th to 90-th percentile of the true underlying distribution. The covariate Z is exponentially distributed with mean 0.3 and the following coverages refer to $S(x|Z = 0.3)$ where the baseline was considered to be for $Z = 0$

			Coverage of the HCNS method for the corresponding percentiles								
Dist.	Sample	Cens.	X_{10}	X_{20}	X_{30}	X_{40}	X_{50}	X_{60}	X_{70}	X_{80}	X_{90}
HCNS	300	30%	0.936	0.956	0.944	0.940	0.944	0.948	0.944	0.948	0.933
		50%	0.935	0.946	0.952	0.960	0.951	0.942	0.940	0.920	0.918
		70%	0.945	0.943	0.939	0.949	0.951	0.947	0.926	0.923	0.953
	100	30%	0.930	0.950	0.952	0.943	0.938	0.943	0.936	0.916	0.911
		50%	0.931	0.952	0.953	0.954	0.947	0.944	0.925	0.903	0.919
		70%	0.947	0.963	0.957	0.942	0.935	0.913	0.887	0.914	0.951
	60	30%	0.928	0.941	0.961	0.948	0.942	0.937	0.918	0.907	0.917
		50%	0.926	0.951	0.958	0.948	0.943	0.932	0.926	0.924	0.946
		70%	0.944	0.955	0.957	0.938	0.919	0.908	0.932	0.941	0.967
Cox	300	30%	0.939	0.935	0.940	0.949	0.942	0.958	0.959	0.963	0.951
		50%	0.932	0.948	0.960	0.946	0.944	0.944	0.938	0.942	0.932
		70%	0.943	0.943	0.940	0.942	0.944	0.944	0.949	0.959	0.951
	100	30%	0.940	0.944	0.942	0.929	0.950	0.936	0.938	0.934	0.925
		50%	0.929	0.929	0.942	0.939	0.946	0.942	0.943	0.930	0.900
		70%	0.916	0.936	0.949	0.939	0.936	0.920	0.924	0.887	0.795
	60	30%	0.934	0.940	0.937	0.942	0.961	0.950	0.947	0.959	0.930
		50%	0.911	0.936	0.932	0.945	0.951	0.941	0.936	0.925	0.867
		70%	0.884	0.902	0.925	0.935	0.915	0.905	0.876	0.832	0.732

4 Appendix D

4.1 Additional simulations for the ROC curve and surfaces subject to an LOD

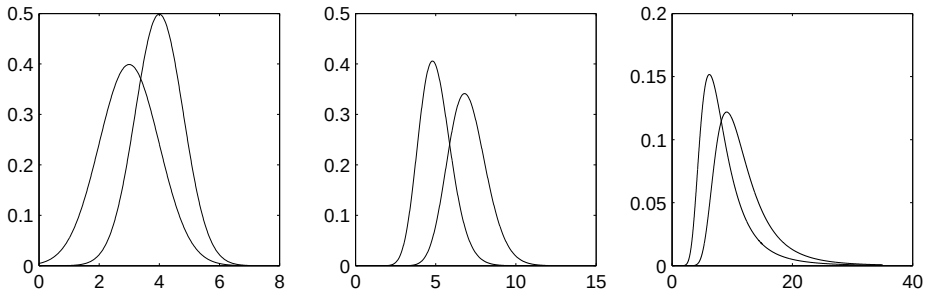

Figure D1: Densities used in the simulation studies in the two class case. Left: $Y_1 \sim N(3,1)$, $Y_2 \sim N(4, 0.8^2)$. Middle: $Y_1 \sim Gamma(25, 0.2)$, $Y_2 \sim Gamma(35, 0.2)$. Right: Noncentral t distributions $Y_1 \sim t(4,7)$ and $Y_2 \sim t(5,10)$

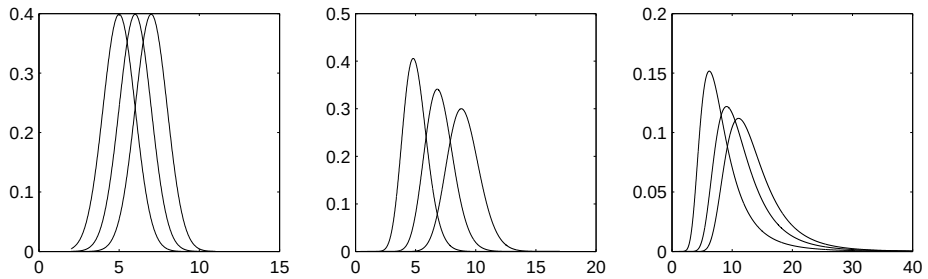

Figure D2: Densities used in the simulation studies in the three class case. Left: $Y_1 \sim N(5,1)$, $Y_2 \sim N(6,1)$, $Y_3 \sim N(7,1)$. Middle: $Y_1 \sim Gamma(25, 0.2)$, $Y_2 \sim Gamma(35, 0.2)$, $Y_3 \sim Gamma(45, 0.2)$. Right: Noncentral t distributions $Y_1 \sim t(4,7)$, $Y_2 \sim t(5,10)$ and $Y_3 \sim t(6,12)$

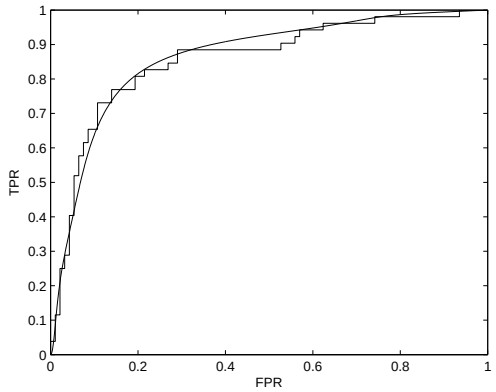

Figure D3: Empirical ROC curve and CNS ROC curve for the liver cancer data when the H and LD group are combined to single 'diseased' group.

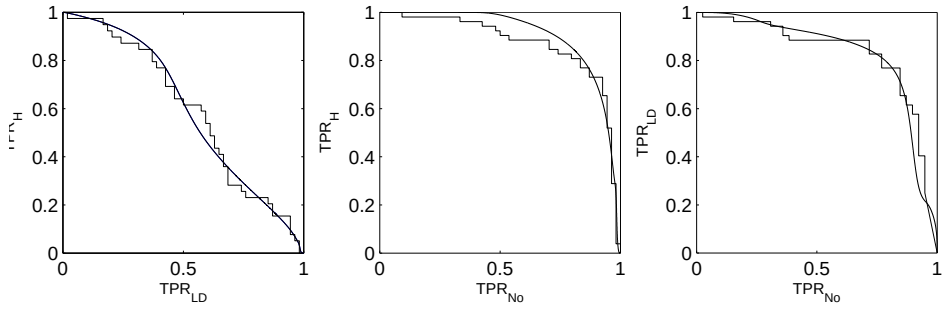

Figure D4: The projections of the ROC surface on the sides of the unit cube are equivalent of a pairwise ROC analysis. The corresponding ROC curves for each pair of disease status for the liver data are shown.

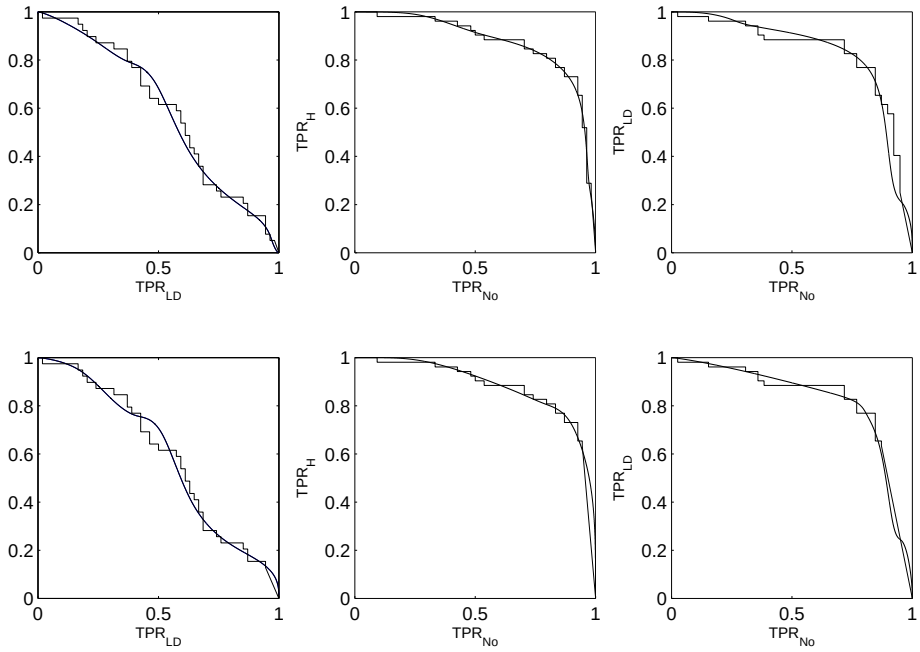

Figure D5: Projections of the empirical and CNS ROC surfaces on the sides of the unit cube for the liver data. Up: Corresponding ROC curves in the case of an upper LOD that causes approximately 10% censoring. Down: Corresponding ROC curves in the case of an upper LOD that causes approximately 30% censoring.

Table D1: Simulation results for 1000 repetitions for the bi-gamma scenario. The likelihood approach assumes the correct model for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True AUC equals to 0.9037)

				AUC			
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_0 : 19.5%, Y_1 : 0.25%)	$Likelihood$	-0.0004	0.0202	0.0004	0.9440
			$Naive$	-0.0007	0.0208	0.0004	0.9570
			$d_L/\sqrt{2}$	-0.0119	0.0192	0.0005	0.9210
			$d_L/2$	-0.0360	0.0181	0.0016	0.4540
			d_L	0.0052	0.0209	0.0005	0.9270
			CNS	0.0032	0.0212	0.0005	0.9410
		30% (Y_0 : 56%, Y_1 : 4%)	$Likelihood$	-0.0004	0.0206	0.0004	0.9480
			$Naive$	-0.0063	0.0219	0.0005	0.9390
			$d_L/\sqrt{2}$	0.0027	0.0230	0.0005	0.9480
			$d_L/2$	-0.0090	0.0246	0.0007	0.9320
			d_L	-0.0075	0.0231	0.0006	0.9400
			CNS	0.0000	0.0212	0.0004	0.9520
		50% (Y_0 : 82.2%, Y_1 : 17.8%)	$Likelihood$	-0.0018	0.0241	0.0006	0.9560
			$Naive$	-0.0432	0.0254	0.0025	0.5850
			$d_L/\sqrt{2}$	-0.0086	0.0281	0.0009	0.9350
			$d_L/2$	-0.0087	0.0302	0.0010	0.9370
			d_L	-0.0576	0.0238	0.0039	0.2930
			CNS	0.0011	0.0270	0.0007	0.9610
	$n = 200$	10% (Y_0 : 19.5%, Y_1 : 0.25%)	$Likelihood$	-0.0005	0.0146	0.0002	0.9430
			$Naive$	-0.0007	0.0148	0.0002	0.9470
			$d_L/\sqrt{2}$	-0.0119	0.0140	0.0003	0.8550
			$d_L/2$	-0.0361	0.0132	0.0015	0.1500
			d_L	0.0051	0.0152	0.0003	0.9060
			CNS	0.0019	0.0149	0.0002	0.9400
		30% (Y_0 : 56%, Y_1 : 4%)	$Likelihood$	-0.0005	0.0149	0.0002	0.9380
			$Naive$	-0.0062	0.0155	0.0003	0.9530
			$d_L/\sqrt{2}$	0.0024	0.0169	0.0003	0.9280
			$d_L/2$	-0.0094	0.0182	0.0004	0.9140
			d_L	-0.0078	0.0165	0.0003	0.9160
			CNS	0.0001	0.0152	0.0002	0.9500
		50% (Y_0 : 82.2%, Y_1 : 17.8%)	$Likelihood$	-0.0009	0.0173	0.0003	0.9370
			$Naive$	-0.0436	0.0177	0.0022	0.2450
			$d_L/\sqrt{2}$	-0.0087	0.0198	0.0005	0.9160
			$d_L/2$	-0.0087	0.0213	0.0005	0.9190
			d_L	-0.0581	0.0172	0.0037	0.0530
			CNS	0.0041	0.0183	0.0004	0.9640
Right Censoring	$n = 100$	10% (Y_0 : 0.45%, Y_1 : 19.55%)	$Likelihood$	-0.0005	0.0201	0.0004	0.9460
			CNS	0.0041	0.0210	0.0005	0.9470
		30% (Y_0 : 4.7%, Y_1 : 55.3%)	$Likelihood$	-0.0008	0.0207	0.0004	0.9480
			CNS	0.0012	0.0216	0.0005	0.9600
		50% (Y_0 : 17.8%, Y_1 : 82.2%)	$Likelihood$	-0.0021	0.0250	0.0006	0.9560
			CNS	0.0043	0.0257	0.0008	0.9700
	$n = 200$	10% (Y_0 : 0.45%, Y_1 : 19.55%)	$Likelihood$	-0.0005	0.0146	0.0002	0.9470
			CNS	0.0029	0.0151	0.0002	0.9300
		30% (Y_0 : 4.7%, Y_1 : 55.3%)	$Likelihood$	-0.0005	0.0149	0.0002	0.9440
			CNS	0.0010	0.0155	0.0002	0.9430
		50% (Y_0 : 17.8%, Y_1 : 82.2%)	$Likelihood$	-0.0012	0.0174	0.0003	0.9330
			CNS	0.0058	0.0194	0.0004	0.9320

Table D2: Simulation results for 1000 repetitions for the scenario of two populations that follow a non-central t distribution. The likelihood approach falsely assumes the normal model for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True AUC equals to 0.7355)

				AUC			
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_0 : 18.9%, Y_1 : 1.1%)	$Likelihood$	-0.0452	0.0385	0.0035	0.8080
			$Naive$	0.0015	0.0355	0.0013	0.9540
			$d_L/\sqrt{2}$	-0.0528	0.0421	0.0046	0.7830
			$d_L/2$	-0.0470	0.0417	0.0040	0.8240
			d_L	-0.0628	0.0423	0.0057	0.6960
			CNS	-0.0047	0.0383	0.0015	0.9530
		30% (Y_0 : 47.5%, Y_1 : 12.5%)	$Likelihood$	-0.0295	0.0349	0.0021	0.8820
			$Naive$	-0.0118	0.0361	0.0014	0.9330
			$d_L/\sqrt{2}$	-0.0585	0.0421	0.0052	0.7370
			$d_L/2$	-0.0454	0.0417	0.0038	0.8340
			d_L	-0.0852	0.0418	0.0090	0.4330
			CNS	-0.0080	0.0297	0.0009	0.9480
		50% (Y_0 : 67.3%, Y_1 : 32.7%)	$Likelihood$	-0.0221	0.0440	0.0024	0.9240
			$Naive$	-0.0507	0.0356	0.0038	0.7080
			$d_L/\sqrt{2}$	-0.0812	0.0415	0.0083	0.4630
			$d_L/2$	-0.0677	0.0412	0.0063	0.6380
			d_L	-0.1133	0.0409	0.0145	0.1350
			CNS	-0.0078	0.0388	0.0016	0.9620
	$n = 200$	10% (Y_0 : 18.9%, Y_1 : 1.1%)	$Likelihood$	-0.0496	0.0280	0.0032	0.5830
			$Naive$	-0.0003	0.0249	0.0006	0.9530
			$d_L/\sqrt{2}$	-0.0575	0.0310	0.0043	0.5360
			$d_L/2$	-0.0515	0.0308	0.0036	0.6010
			d_L	-0.0676	0.0310	0.0055	0.3800
			CNS	-0.0025	0.0346	0.0012	0.9420
		30% (Y_0 : 47.5%, Y_1 : 12.5%)	$Likelihood$	-0.0322	0.0248	0.0017	0.7350
			$Naive$	-0.0134	0.0251	0.0008	0.9040
			$d_L/\sqrt{2}$	-0.0631	0.0308	0.0049	0.4380
			$d_L/2$	-0.0496	0.0304	0.0034	0.6310
			d_L	-0.0900	0.0306	0.0090	0.1240
			CNS	-0.0049	0.0270	0.0008	0.9420
		50% (Y_0 : 67.3%, Y_1 : 32.7%)	$Likelihood$	-0.0221	0.0323	0.0015	0.8710
			$Naive$	-0.0513	0.0249	0.0033	0.4550
			$d_L/\sqrt{2}$	-0.0855	0.0301	0.0082	0.1620
			$d_L/2$	-0.0716	0.0297	0.0060	0.3170
			d_L	-0.1178	0.0299	0.0148	0.0070
			CNS	-0.0058	0.0367	0.0014	0.9420
Right Censoring	$n = 100$	10% (Y_0 : 5.8%, Y_1 : 14.1%)	$Likelihood$	-0.0181	0.0379	0.0018	0.9390
			CNS	0.0098	0.0368	0.0015	0.9300
		30% (Y_0 : 17.8%, Y_1 : 42.2%)	$Likelihood$	0.0064	0.0392	0.0016	0.9520
			CNS	0.0072	0.0388	0.0016	0.9450
		50% (Y_0 : 32.7%, Y_1 : 67.3%)	$Likelihood$	0.0218	0.0431	0.0023	0.9050
			CNS	0.0133	0.0534	0.0030	0.9600
	$n = 200$	10% (Y_0 : 5.8%, Y_1 : 14.1%)	$Likelihood$	-0.0196	0.0267	0.0011	0.8770
			CNS	0.0066	0.0259	0.0007	0.9500
		30% (Y_0 : 17.8%, Y_1 : 42.2%)	$Likelihood$	0.0048	0.0272	0.0008	0.9330
			CNS	0.0050	0.0275	0.0008	0.9510
		50% (Y_0 : 32.7%, Y_1 : 67.3%)	$Likelihood$	0.0198	0.0299	0.0013	0.8840
			CNS	0.0134	0.0375	0.0016	0.9620

Table D3: Simulation results for 1000 repetitions for the tri-gamma scenario. The likelihood approach assumes the correct model for the three populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True VUS equals to 0.7747)

				VUS			
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_1 : 29.4%, Y_2 : 0.6%, Y_3 : 0%)	$Likelihood$	0.0010	0.0269	0.0007	0.9360
			$Naive$	-0.0001	0.0273	0.0007	0.9380
			$d_L/\sqrt{2}$	-0.0086	0.0273	0.0008	0.9420
			$d_L/2$	-0.0320	0.0290	0.0019	0.8190
			d_L	0.0061	0.0278	0.0008	0.9250
			CNS	0.0096	0.0280	0.0009	0.9300
		30% (Y_1 : 76.65%, Y_2 : 13.1%, Y_3 : 0.25%)	$Likelihood$	0.0001	0.0284	0.0008	0.9340
			$Naive$	-0.0281	0.0284	0.0017	0.8460
			$d_L/\sqrt{2}$	-0.0154	0.0294	0.0011	0.9290
			$d_L/2$	-0.0432	0.0301	0.0028	0.7460
			d_L	-0.0367	0.0284	0.0022	0.7740
			CNS	0.0096	0.0280	0.0009	0.9300
	$n = 200$	10% (Y_1 : 29.4%, Y_2 : 0.6%, Y_3 : 0%)	$Likelihood$	0.0000	0.0197	0.0004	0.9540
			$Naive$	-0.0006	0.0200	0.0004	0.9390
			$d_L/\sqrt{2}$	-0.0099	0.0201	0.0005	0.9160
			$d_L/2$	-0.0339	0.0216	0.0016	0.6270
			d_L	0.0051	0.0203	0.0004	0.9350
			CNS	0.0052	0.0207	0.0005	0.9501
		30% (Y_1 : 76.65%, Y_2 : 13.1%, Y_3 : 0.25%)	$Likelihood$	-0.0001	0.0207	0.0004	0.9550
			$Naive$	-0.0285	0.0209	0.0012	0.7578
			$d_L/\sqrt{2}$	-0.0162	0.0219	0.0007	0.8830
			$d_L/2$	-0.0441	0.0226	0.0025	0.4970
			d_L	-0.0382	0.0204	0.0019	0.5400
			CNS	0.0052	0.0207	0.0005	0.9501
Right Censoring	$n = 100$	10% (Y_1 : 0%, Y_2 : 1.75%, Y_3 : 28.25%)	$Likelihood$	0.0007	0.0271	0.0007	0.9360
			CNS	0.0130	0.0288	0.0010	0.9340
		30% (Y_1 : 0.34%, Y_2 : 16.6%, Y_3 : 73.06%)	$Likelihood$	-0.0003	0.0292	0.0009	0.9570
			CNS	0.0025	0.0343	0.0012	0.9767
	$n = 200$	10% (Y_1 : 0%, Y_2 : 1.75%, Y_3 : 28.25%)	$Likelihood$	0.0000	0.0197	0.0004	0.9490
			CNS	0.0080	0.0208	0.0005	0.9359
		30% (Y_1 : 0.34%, Y_2 : 16.6%, Y_3 : 73.06%)	$Likelihood$	-0.0006	0.0217	0.0005	0.9580
			CNS	0.0064	0.0228	0.0006	0.9780

Table D4: Simulation results for 1000 repetitions for the scenario of three non central t distributions. The likelihood approach falsely assumes the normal model for the three populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True VUS equals to 0.4185)

				VUS			
Direction	Sample	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	$n = 100$	10% (Y_1 : 27.08%, Y_2 : 2.74%, Y_3 : 0.18%)	$Likelihood$	-0.0609	0.0418	0.0055	0.6580
			$Naive$	-0.0001	0.0344	0.0012	0.9620
			$d_L/\sqrt{2}$	-0.0748	0.0430	0.0074	0.5730
			$d_L/2$	-0.0663	0.0431	0.0062	0.6380
			d_L	-0.0890	0.0422	0.0097	0.4070
			CNS	-0.0013	0.0411	0.0017	0.9600
		30% (Y_1 : 59.6%, Y_2 : 23.2%, Y_3 : 7.2%)	$Likelihood$	-0.0401	0.0445	0.0036	0.8520
			$Naive$	-0.0314	0.0324	0.0020	0.8350
			$d_L/\sqrt{2}$	-0.0885	0.0404	0.0095	0.3830
			$d_L/2$	-0.0730	0.0392	0.0069	0.5270
			d_L	-0.1203	0.0404	0.0161	0.1250
			CNS	-0.0013	0.0411	0.0017	0.9600
	$n = 200$	10% (Y_1 : 27.08%, Y_2 : 2.74%, Y_3 : 0.18%)	$Likelihood$	-0.0639	0.0326	0.0051	0.4210
			$Naive$	-0.0008	0.0243	0.0006	0.9460
			$d_L/\sqrt{2}$	-0.0779	0.0334	0.0072	0.2620
			$d_L/2$	-0.0693	0.0336	0.0059	0.3960
			d_L	-0.0922	0.0328	0.0096	0.1250
			CNS	-0.0055	0.0391	0.0016	0.9614
		30% (Y_1 : 59.6%, Y_2 : 23.2%, Y_3 : 7.2%)	$Likelihood$	-0.0403	0.0353	0.0029	0.7490
			$Naive$	-0.0321	0.0230	0.0016	0.6960
			$d_L/\sqrt{2}$	-0.0914	0.0313	0.0093	0.1030
			$d_L/2$	-0.0756	0.0303	0.0066	0.2320
			d_L	-0.1235	0.0313	0.0162	0.0080
			CNS	-0.0078	0.0376	0.0015	0.9495
Right Censoring	$n = 100$	10% (Y_1 : 4.2%, Y_2 : 9.85%, Y_3 : 15.95%)	$Likelihood$	-0.0325	0.0352	0.0023	0.8290
			CNS	0.0162	0.0377	0.0017	0.9240
		30% (Y_1 : 12.4%, Y_2 : 30.1%, Y_3 : 47.5%)	$Likelihood$	-0.0029	0.0392	0.0015	0.9420
			CNS	0.0080	0.0441	0.0020	0.9760
	$n = 200$	10% (Y_1 : 4.2%, Y_2 : 9.85%, Y_3 : 15.95%)	$Likelihood$	-0.0344	0.0251	0.0018	0.7110
			CNS	0.0110	0.0269	0.0008	0.9330
		30% (Y_1 : 12.4%, Y_2 : 30.1%, Y_3 : 47.5%)	$Likelihood$	-0.0031	0.0273	0.0008	0.9440
			CNS	0.0079	0.0313	0.0010	0.9670

Table D5: Simulation results for 1000 repetitions for the bi-normal scenario with unequal sample sizes for the two populations (100 and 300 for Y_0 and Y_1 respectively). The likelihood approach assumes the correct model for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True AUC equals to 0.7826)

			AUC			
Direction	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	10% (Y_0 : 30.94%, Y_1 : 3.02%)	$Likelihood$	-0.0008	0.0281	0.0008	0.9440
		$Naive$	-0.0036	0.0284	0.0008	0.9420
		$d_L/\sqrt{2}$	-0.0006	0.0289	0.0008	0.9460
		$d_L/2$	-0.0013	0.0279	0.0008	0.9470
		d_L	-0.0097	0.0304	0.0010	0.9390
		CNS	-0.0013	0.0292	0.0009	0.9430
	30% (Y_0 : 62.05%, Y_1 : 19.32%)	$Likelihood$	-0.0011	0.0295	0.0009	0.9470
		$Naive$	-0.0291	0.0278	0.0016	0.8300
		$d_L/\sqrt{2}$	-0.0195	0.0310	0.0013	0.9020
		$d_L/2$	-0.0168	0.0310	0.0012	0.9110
		d_L	-0.0496	0.0307	0.0034	0.6220
		CNS	-0.0012	0.0314	0.0010	0.9530
	50% (Y_0 : 78.39%, Y_1 : 40.35%)	$Likelihood$	-0.0048	0.0385	0.0015	0.9390
		$Naive$	-0.0798	0.0253	0.0070	0.0240
		$d_L/\sqrt{2}$	-0.0564	0.0308	0.0041	0.5350
		$d_L/2$	-0.0536	0.0311	0.0038	0.5750
		d_L	-0.0925	0.0298	0.0094	0.0450
		CNS	-0.0011	0.0461	0.0021	0.9860
Right Censoring	10% (Y_0 : 2.71%, Y_1 : 12.43%)	$Likelihood$	-0.0011	0.0281	0.0008	0.9400
		CNS	0.0003	0.0288	0.0008	0.9450
	30% (Y_0 : 10.17%, Y_1 : 36.61%)	$Likelihood$	-0.0016	0.0291	0.0008	0.9470
		CNS	0.0003	0.0292	0.0009	0.9550
	50% (Y_0 : 21.06%, Y_1 : 59.65%)	$Likelihood$	-0.0023	0.0312	0.0010	0.9330
		CNS	0.0014	0.0345	0.0012	0.9470

Table D6: Simulation results for 1000 repetitions for the bi-gamma scenario with unequal sample sizes for the two populations (100 and 300 for Y_0 and Y_1 respectively). The likelihood approach assumes the correct model for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. The case of 50% is not presented for left censoring since the level of censoring for Y_0 is over 90%. (True AUC equals to 0.9037)

			AUC			
Direction	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	10% (Y_0 : 36.53%, Y_1 : 1.16%)	$Likelihood$	-0.0002	0.0167	0.0003	0.9380
		$Naive$	-0.0013	0.0169	0.0003	0.9380
		$d_L/\sqrt{2}$	-0.0059	0.0168	0.0003	0.9540
		$d_L/2$	-0.0264	0.0172	0.0010	0.6720
		d_L	0.0034	0.0174	0.0003	0.9290
		CNS	0.0017	0.0172	0.0003	0.9300
	30% (Y_0 : 78%, Y_1 : 14%)	$Likelihood$	-0.0006	0.0183	0.0003	0.9490
		$Naive$	-0.0313	0.0175	0.0013	0.5852
		$d_L/\sqrt{2}$	-0.0020	0.0206	0.0004	0.9460
		$d_L/2$	-0.0037	0.0226	0.0005	0.9420
		d_L	-0.0449	0.0176	0.0023	0.1780
		CNS	0.0014	0.0198	0.0004	0.9623
Right Censoring	10% (Y_0 : 0.22%, Y_1 : 13.26%)	$Likelihood$	-0.0003	0.0165	0.0003	0.9400
		CNS	0.0029	0.0172	0.0003	0.9350
	30% (Y_0 : 1.99%, Y_1 : 39.34%)	$Likelihood$	-0.0003	0.0164	0.0003	0.9400
		CNS	0.0017	0.0171	0.0003	0.9460
	50% (Y_0 : 7.41%, Y_1 : 64.19%)	$Likelihood$	-0.0007	0.0175	0.0003	0.9390
		CNS	0.0008	0.0189	0.0004	0.9600

Table D7: Simulation results for 1000 repetitions for the scenario where the two populations follow two non-central t distributions. The sample sizes for the two populations (100 and 300 for Y_0 and Y_1 respectively). The likelihood approach falsely assumes normality for both populations. The coverage is derived by using the percentile bootstrap with 200 samples for each repetition. (True AUC equals to 0.7355)

			AUC			
Direction	Censoring	Method	Bias	SE	MSE	Coverage
Left Censoring	10% (Y_0 : 29.66%, Y_1 : 3.45%)	$Likelihood$	0.1607	0.0333	0.0269	0.7750
		$Naive$	0.1988	0.0342	0.0407	0.9500
		$d_L/\sqrt{2}$	0.1472	0.0392	0.0232	0.7020
		$d_L/2$	0.1565	0.0389	0.0260	0.7810
		d_L	0.1305	0.0389	0.0185	0.5200
		CNS	0.2002	0.0318	0.0411	0.9460
	30% (Y_0 : 57.39%, Y_1 : 20.87%)	$Likelihood$	0.1732	0.0327	0.0311	0.8440
		$Naive$	0.1746	0.0307	0.0314	0.8400
		$d_L/\sqrt{2}$	0.1326	0.0384	0.0191	0.5270
		$d_L/2$	0.1465	0.0380	0.0229	0.6680
		d_L	0.1026	0.0375	0.0119	0.1350
		CNS	0.1887	0.0375	0.0370	0.9440
	50% (Y_0 : 73.59%, Y_1 : 42.14%)	$Likelihood$	0.1760	0.0462	0.0331	0.8880
		$Naive$	0.1292	0.0277	0.0175	0.1850
		$d_L/\sqrt{2}$	0.1067	0.0366	0.0127	0.1620
		$d_L/2$	0.1192	0.0361	0.0155	0.3060
		d_L	0.0754	0.0358	0.0070	0.0030
		CNS	0.1289	0.0276	0.0174	0.9560
Right Censoring	10% (Y_0 : 4.86%, Y_1 : 11.72%)	$Likelihood$	-0.0232	0.0332	0.0016	0.8940
		CNS	0.0065	0.0327	0.0011	0.9420
	30% (Y_0 : 14.55%, Y_1 : 35.15%)	$Likelihood$	-0.0011	0.0338	0.0011	0.9390
		CNS	0.0025	0.0335	0.0011	0.9460
	50% (Y_0 : 26.47%, Y_1 : 57.85%)	$Likelihood$	0.0134	0.0358	0.0015	0.9000
		CNS	0.0044	0.0407	0.0017	0.9460

Table D8: Simulation results for 1000 repetitions in the case of a lower LOD when 5 or 6 or 7 knots are used with the proposed method ($CNS_{(5)}$, $CNS_{(6)}$, and $CNS_{(7)}$ respectively). The sample size for each of the two populations equals to 100. (Results that correspond to $CNS_{(6)}$ are restated here for convenience)

			AUC		
True Models	Censoring	Method	Bias	SE	MSE
Bi-Normal	10% (Y_0 : 19.1%, Y_1 : 0.9%)	$CNS_{(5)}$	0.0005	0.0351	0.0012
		$CNS_{(6)}$	0.0006	0.0351	0.0012
		$CNS_{(7)}$	-0.0016	0.0344	0.0012
	30% (Y_0 : 50%, Y_1 : 10%)	$CNS_{(5)}$	-0.0016	0.0352	0.0012
		$CNS_{(6)}$	-0.0015	0.0352	0.0012
		$CNS_{(7)}$	-0.0025	0.0350	0.0012
	50% (Y_0 : 71%, Y_1 : 29%)	$CNS_{(5)}$	-0.0026	0.0424	0.0018
		$CNS_{(6)}$	-0.0025	0.0440	0.0019
		$CNS_{(7)}$	-0.0028	0.0436	0.0019
Bi-Gamma	10% (Y_0 : 19.5%, Y_1 : 0.25%)	$CNS_{(5)}$	0.0034	0.0212	0.0005
		$CNS_{(6)}$	0.0032	0.0212	0.0005
		$CNS_{(7)}$	0.0004	0.0210	0.0004
	30% (Y_0 : 56%, Y_1 : 4%)	$CNS_{(5)}$	0.0002	0.0213	0.0005
		$CNS_{(6)}$	0.0000	0.0212	0.0004
		$CNS_{(7)}$	-0.0012	0.0213	0.0005
	50% (Y_0 : 82.2%, Y_1 : 17.8%)	$CNS_{(5)}$	0.0012	0.0265	0.0007
		$CNS_{(6)}$	0.0011	0.0270	0.0007
		$CNS_{(7)}$	0.0009	0.0277	0.0008
Bi-Non central t	10% (Y_0 : 18.9%, Y_1 : 1.1%)	$CNS_{(5)}$	-0.0003	0.0415	0.0017
		$CNS_{(6)}$	-0.0047	0.0383	0.0015
		$CNS_{(7)}$	-0.0018	0.0402	0.0016
	30% (Y_0 : 47.5%, Y_1 : 12.5%)	$CNS_{(5)}$	-0.0031	0.0368	0.0014
		$CNS_{(6)}$	-0.0080	0.0297	0.0009
		$CNS_{(7)}$	-0.0033	0.0366	0.0014
	50% (Y_0 : 67.3%, Y_1 : 32.7%)	$CNS_{(5)}$	-0.0046	0.0451	0.0021
		$CNS_{(6)}$	-0.0078	0.0388	0.0016
		$CNS_{(7)}$	-0.0047	0.0455	0.0021

Table D9: Simulation results for 1000 repetitions in the case of a lower LOD when 5 or 6 or 7 knots are used with the proposed method ($CNS_{(5)}$, $CNS_{(6)}$, and $CNS_{(7)}$ respectively). The sample size for each of the two populations equals to 200. (Results that correspond to $CNS_{(6)}$ are restated here for convenience)

			AUC		
True Models	Censoring	Method	Bias	SE	MSE
Bi-Normal	10% (Y_0 : 19.1%, Y_1 : 0.9%)	$CNS_{(5)}$	-0.0001	0.0234	0.0005
		$CNS_{(6)}$	-0.0001	0.0233	0.0005
		$CNS_{(7)}$	-0.0014	0.0230	0.0005
	30% (Y_0 : 50%, Y_1 : 10%)	$CNS_{(5)}$	-0.0012	0.0239	0.0006
		$CNS_{(6)}$	-0.0012	0.0241	0.0006
		$CNS_{(7)}$	-0.0018	0.0240	0.0006
	50% (Y_0 : 71%, Y_1 : 29%)	$CNS_{(5)}$	-0.0005	0.0281	0.0008
		$CNS_{(6)}$	0.0006	0.0292	0.0009
		$CNS_{(7)}$	-0.0004	0.0276	0.0008
Bi-Gamma	10% (Y_0 : 19.5%, Y_1 : 0.25%)	$CNS_{(5)}$	0.0026	0.0152	0.0002
		$CNS_{(6)}$	0.0019	0.0149	0.0002
		$CNS_{(7)}$	-0.0009	0.0152	0.0002
	30% (Y_0 : 56%, Y_1 : 4%)	$CNS_{(5)}$	0.0001	0.0152	0.0002
		$CNS_{(6)}$	0.0001	0.0152	0.0002
		$CNS_{(7)}$	0.0033	0.0185	0.0004
	50% (Y_0 : 82.2%, Y_1 : 17.8%)	$CNS_{(5)}$	0.0034	0.0180	0.0003
		$CNS_{(6)}$	0.0041	0.0183	0.0004
		$CNS_{(7)}$	0.0001	0.0149	0.0002
Bi-Non central t	10% (Y_0 : 18.9%, Y_1 : 1.1%)	$CNS_{(5)}$	-0.0050	0.0381	0.0015
		$CNS_{(6)}$	-0.0025	0.0346	0.0012
		$CNS_{(7)}$	-0.0061	0.0371	0.0014
	30% (Y_0 : 47.5%, Y_1 : 12.5%)	$CNS_{(5)}$	-0.0050	0.0381	0.0015
		$CNS_{(6)}$	-0.0049	0.0270	0.0008
		$CNS_{(7)}$	-0.0052	0.0269	0.0008
	50% (Y_0 : 67.3%, Y_1 : 32.7%)	$CNS_{(5)}$	-0.0048	0.0367	0.0014
		$CNS_{(6)}$	-0.0058	0.0367	0.0014
		$CNS_{(7)}$	-0.0049	0.0370	0.0014

Table D10: Simulation results for 1000 repetitions in the case of a lower LOD when 5 or 6 or 7 knots are used with the proposed method ($CNS_{(5)}$, $CNS_{(6)}$, and $CNS_{(7)}$ respectively). The sample sizes are 100 and 300 for Y_0 and Y_1 respectively. (Results that correspond to $CNS_{(6)}$ are restated here for convenience)

			AUC		
True Models	Censoring	Method	Bias	SE	MSE
Bi-Normal	10% (Y_0 : 19.1%, Y_1 : 0.9%)	$CNS_{(5)}$	-0.0014	0.0292	0.0009
		$CNS_{(6)}$	-0.0013	0.0292	0.0009
		$CNS_{(7)}$	-0.0030	0.0288	0.0008
	30% (Y_0 : 50%, Y_1 : 10%)	$CNS_{(5)}$	-0.0024	0.0314	0.0010
		$CNS_{(6)}$	-0.0012	0.0314	0.0010
		$CNS_{(7)}$	-0.0026	0.0314	0.0010
	50% (Y_0 : 71%, Y_1 : 29%)	$CNS_{(5)}$	-0.0029	0.0447	0.0020
		$CNS_{(6)}$	-0.0011	0.0461	0.0021
		$CNS_{(7)}$	-0.0027	0.0468	0.0022
Bi-Gamma	10% (Y_0 : 19.5%, Y_1 : 0.25%)	$CNS_{(5)}$	0.0018	0.0171	0.0003
		$CNS_{(6)}$	0.0017	0.0172	0.0003
		$CNS_{(7)}$	-0.0003	0.0173	0.0003
	30% (Y_0 : 56%, Y_1 : 4%)	$CNS_{(5)}$	0.0009	0.0193	0.0004
		$CNS_{(6)}$	0.0014	0.0198	0.0004
		$CNS_{(7)}$	0.0006	0.0198	0.0004
Bi-Non central t	10% (Y_0 : 18.9%, Y_1 : 1.1%)	$CNS_{(5)}$	-0.0042	0.0346	0.0012
		$CNS_{(6)}$	-0.0013	0.0292	0.0009
		$CNS_{(7)}$	-0.0054	0.0343	0.0012
	30% (Y_0 : 47.5%, Y_1 : 12.5%)	$CNS_{(5)}$	-0.0138	0.0372	0.0016
		$CNS_{(6)}$	-0.0012	0.0314	0.0010
		$CNS_{(7)}$	-0.0145	0.0374	0.0016
	50% (Y_0 : 67.3%, Y_1 : 32.7%)	$CNS_{(5)}$	-0.0178	0.0513	0.0029
		$CNS_{(6)}$	-0.0011	0.0461	0.0021
		$CNS_{(7)}$	-0.0184	0.0521	0.0031

ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΗΣΗΣ ΔΙΑΓΝΩΣΤΙΚΩΝ ΕΛΕΓΧΩΝ ΠΑΡΟΥΣΙΑ ΛΟΓΟΚΡΙΣΙΑΣ

Περίληψη :

Η χρήση διαγνωστικών ελέγχων (ή βιοδεικτών) για την ανίχνευση κάποιας ασθένειας είναι σημαντική γιατί συμβάλλει τόσο στην κατανόηση του μηχανισμού της ασθένειας αλλά και στην βελτίωση της ποιότητας της ζωής των ασθενών ή και στην πλήρη αντιμετώπισή της. Όταν έχουμε να κάνουμε με καταληκτικές ασθένειες τότε η έγκαιρη διάγνωση μπορεί να συμβάλλει στην επιμήκυνση του υπολοιπόμενου χρόνου ζωής. Ωστόσο η διάγνωση είναι μια ατελής διαδικασία. Έτσι, η αξιολόγηση των διαγνωστικών ελέγχων είναι κρίσιμης σημασίας.

Στην περίπτωση που έχουμε δύο ομάδες προς διαχωρισμό (π.χ. ασθενείς και υγιείς) και ο διαγνωστικός έλεγχος δίνει συνεχείς ή μετρήσεις διατεταγμένης κλίμακας τότε η πιο διαδεδομένη τεχνική για την αξιολόγησή του, είναι η καμπύλη ROC (receiver operating characteristic) (βλέπε Pepe (2003) για μια λεπτομερή ανασκόπηση των καμπυλών ROC). Ένας δείκτης που συνοψίζει την διαχωριστική ικανότητα ενός τέτοιου διαγνωστικού ελέγχου είναι το εμβαδό κάτω από την καμπύλη ROC (area under the curve ή AUC). Σε προβλήματα διαχωρισμού τριών πληθυσμών χρησιμοποιείται η επιφάνεια ROC (βλέπε Mossman 1999). Ο δείκτης ο οποίος συνοψίζει την διαχωριστική ικανότητα ενός τέτοιου διαγνωστικού ελέγχου είναι ο όγκος κάτω από την επιφάνεια ROC (volume under the surface ή VUS).

Πολλοί διαγνωστικοί έλεγχοι εξαρτώνται από το χρόνο. Για παράδειγμα το Framingham risk score (FR-score) είναι ένας διαγνωστικός έλεγχος που θεωρείται δεικτικός του εμφράγματος του μυοκαρδίου και του εγκεφαλικού (βλέπε και Wilson et al. (1998) και Grundy et al. (1998)). Το FR-score (διαφορετικό για γυναίκες και άντρες) βασίζεται σε παράγοντες όπως η τιμή της χοληστερίνης, του σακχαρώδη διαβήτη, της ηλικίας και της πίεσης. Για τέτοιους διαγνωστικούς ελέγχους αναμένεται ότι μετρήσεις που λαμβάνονται πιο κοντά στο «γεγονός» (που συχνά είναι ο θάνατος) είναι μεγαλύτερες. Πολλές φορές, όταν μελετώνται καταληκτικές ασθένειες παρατηρείται το φαινόμενο τη λογοκρισίας. Αυτό συμβαίνει επειδή κάποιοι ασθενείς αποφασίζουν για διάφορους λόγους να παραιτηθούν από την έρευνα και έτσι η μόνη πληροφορία που έχουμε για τα εν λόγω άτομα είναι ότι κατάφεραν να επιζήσουν πέραν του χρονικού σημείου που παραιτήθηκαν από την έρευνα. Έτσι οι χρόνοι μέχρι τον θάνατο αυτών των ατόμων θεωρούνται λογοκριμένοι μιας και δεν παρατηρούνται πλήρως. Αυτό είναι ένα παράδειγμα δεξιάς λογοκρισίας. Αριστερή λογοκρισία έχουμε όταν ένα άτομο έχει υποστεί το υπό μελέτη γεγονός πριν την εισαγωγή του στην έρευνα. Ένα άλλο είδος λογοκρισίας είναι η λογοκρισία σε διάστημα, όπου το

μόνο που μας είναι γνωστό για κάποιον λογοκριμένο ασθενή είναι ότι υπέστη το γεγονός σε κάποιο κλειστό χρονικό διάστημα.

Σε αυτήν την διδακτορική διατριβή αναπτύσσουμε νέες στατιστικές μεθόδους που συνεισφέρουν στην αξιολόγηση διαγνωστικών ελέγχων παρουσία λογοκρισίας. Η μοντελοποίηση ενός χρονοεξαρτώμενου διαγνωστικού ελέγχου μπορεί να συνεισφέρει στην κατασκευή χρονοεξαρτώμενων καμπυλών ROC και κατ'επέκταση στην μελέτη της διαγνωστικής ακρίβειας του ελέγχου στον χρόνο. Σε αυτήν την εργασία επικεντρωνόμαστε στη χρήση των γενικευμένων γραμμικών μοντέλων για τη μοντελοποίηση τέτοιων διαγνωστικών ελέγχων. Ωστόσο κατά την προσαρμογή των γενικευμένων γραμμικών μοντέλων πρέπει να ληφθεί υπόψη η λογοκριμένη συμμεταβλητή του χρόνου μέχρι το γεγονός. Για αυτό το σκοπό αναπτύσσουμε μία νέα μεθοδολογία που βασίζεται σε βέλτιστες εκτιμητικές εξισώσεις. Με τη μεθοδολογία αυτή δε χρειάζεται καμία υπόθεση για την παραμετρική μορφή της κατανομής της απόκρισης. Για την κατανομή της λογοκριμένης συμμεταβλητής του χρόνου, μπορεί κανείς να υποθέσει είτε παραμετρικά μοντέλα είτε άλλες μη παραμετρικές τεχνικές. Εμείς, για την μοντελοποίηση της λογοκριμένης συμμεταβλητής αναπτύσσουμε μια νέα spline τεχνική. Προσαρμόζουμε μία φυσική κυβική spline στην αθροιστική συνάρτηση κινδύνου κάτω από περιορισμούς μονοτονίας. Με αυτήν την μέθοδο επιτυγχάνουμε εκτίμηση της κατανομής της λογοκριμένης συμμεταβλητής και πέραν της τελευταίας (χρονικά) παρατήρησης ακόμα και αν αυτή είναι λογοκριμένη. Κάτι τέτοιο δεν είναι δυνατό με τη γνωστή μη παραμετρική εκτίμηση μέγιστης πιθανοφάνειας των Kaplan και Meier (1959). Επίσης, κάτι τέτοιο δεν είναι δυνατό ούτε με γνωστές τεχνικές λείανσης όπως αυτή των πυρήνων (βλέπε Wand and Jones (1995) για μια λεπτομερή ανασκόπηση αυτών των τεχνικών). Κάτω από την spline τεχνική που αναπτύσσουμε, εξασφαλίζουμε εγγυημένη σύγκλιση μιας και το πρόβλημα ελαχιστοποίησης αφορά σε άθροισμα τετραγώνων με γραμμικούς περιοριορισμούς ως προς τις παραμέτρους. Κάτι τέτοιο δεν εξασφαλίζεται από spline τεχνικές που αφορούν σε μέγιστη πιθανοφάνεια όπως η logspline τεχνική (βλέπε Kooperberg (1991)) ή λεγόμενη restricted cubic spline (RCS) προσέγγιση (βλέπε Harrell (2001)). Επιπλέον, επεκτείνουμε τη μεθοδολογία των εκτιμητικών εξισώσεων και σε καταστάσεις όπου έχουμε επαναλαμβανόμενες μετρήσεις. Δηλαδή σε περιπτώσεις που ασθενείς παρακολουθούνται στο χρόνο και από τους οποίους λαμβάνονται μετρήσεις ανά τακτά χρονικά διαστήματα μέχρι το γεγονός ή τη λογοκρία. Επίσης, παρουσιάζουμε κάποιες εφαρμογές με πραγματικά δεδομένα που αφορούν είτε σε περιπτώσεις όπου λαμβάνεται μία μέτρηση για κάθε ασθενή είτε σε περιπτώσεις που λαμβάνονται περισσότερες από μια μετρήσεις και οι ασθενείς παρακολουθούνται στο χρόνο.

Εκτός από το φαινόμενο της λογοκρισίας στην μεταβλητή του χρόνου από την οποία μπορεί να εξαρτάται ένας διαγνωστικός έλεγχος, λογοκρισία μπορεί να εμφανιστεί και σε αυτές καθεαυτές τις τιμές του διαγνωστικού ελέγχου. Κάτι τέτοιο μπορεί να συμβεί όταν εξαιτίας τεχνικών δυσκολιών που αφορούν στην τεχνολογία ή την φύση του διαγνωστικού ελέγχου, δε μπορούν να ληφθούν τιμές κάτω ή πάνω από κάποιο όριο (limit of detection ή LOD). Κάποιες τεχνικές που έχουν προταθεί για την αξιολόγηση τέτοιων διαγνωστικών ελέγχων βασίζονται είτε σε απλές τιμές αντικατάστασης των λογοκριμένων παρατηρήσεων είτε σε μεθόδους μέγιστης πιθανοφάνειας ύστερα από αυστηρές παραμετρικές υποθέσεις για την κατανομή των μετρήσεων. Αποδεικνύουμε ότι οι τεχνικές απλών τιμών αντικατάστασης προκαλούν μεροληψία κατά την εκτίμηση του VUS γενικεύοντας το αποτέλεσμα του Perkins et al. (2007) που αφορά στην περίπτωση διαχωρισμού δύο πληθυσμών. Για την αξιολόγηση τέτοιων διαγνωστικών ελέγχων μελετάμε τη χρήση της spline μεθόδου που αναφέραμε στην προηγούμενη παράγραφο για να λάβουμε μια λεία εκτίμηση της αθροιστικής συνάρτησης

κινδύνου των μετρήσεων. Έτσι είναι δυνατή η κατασκευή της αντίστοιχης ROC καμπύλης ή επιφάνειας και η εκτίμηση του AUC ή VUS αντίστοιχα. Μέσω προσομοιώσεων δείχνουμε ότι η τεχνική μας δίνει καλύτερες εκτιμήσεις ως προς αυτούς τους δύο δείκτες σε σχέση με άλλες γνωστές τεχνικές. Επίσης η προτεινόμενη spline μεθοδολογία μπορεί να φανεί ιδιαίτερα χρήσιμη όταν το ενδιαφέρον μας επικεντρώνεται σε υψηλές τιμές εσφαλμένα θετικών μετρήσεων του διαγνωστικού ελέγχου όπου σαν δείκτης αξιολόγησης χρησιμοποιείται η μερική επιφάνεια κάτω από την καμπύλη ROC (partial AUC).

Οι τεχνικές που παρουσιάζονται σε αυτήν την διδακτορική διατριβή μπορούν να εφαρμοστούν και σε άλλα επιστημονικά πεδία πέραν της βιοστατιστικής. Για παράδειγμα στην Οικονομετρία πολλές φορές το εισόδημα παίζει το ρόλο της συμμεταβλητής και υπόκειται συχνά σε δεξιά ή αριστερή λογοκρισία μιας και δεν είναι γνωστή η ακριβής τιμή του αν αυτή βρίσκεται πάνω/κάτω από κάποιο συγκεκριμένο όριο. Παραδείγματα επίσης αφορούν στην πρόβλεψη πυρκαγιών όπου σαν συμμεταβλητή χρησιμοποιείται η ένταση του ανέμου η μέτρηση της οποίας μπορεί να λογοκρίνεται εξαιτίας κάποιου ορίου ανίχνευσης (βλέπε Chang (2007)). Άλλα παραδείγματα μπορεί να αφορούν στην εντομολογία όπου η ένταση του ανέμου επιδρά στην δραστηριότητα εντόμων (βλέπε Steelman et al. (1993)).