# Acknowledgement

Several people have contributed in one way or another to the completion of this tract. I think about my former colleagues in the Netherlands and who I met during my period at the University of California, Berkeley. They have been helpful by stimulating discussions and providing useful suggestions. In particular, I would like to thank Susan Murphy for our long discussions on several relevant topics for this thesis, here in the Netherlands when she was visiting the University of Utrecht for half a year and in Berkeley during the MSRI-semester about semiparametric models. I thank André de Meijer, the computer expert, for answering many questions and being helpful.

In Berkeley I got the opportunity to know Peter Bickel. His enthusiasm and generousity complemented with his broad knowledge and creative mind formed an important ingredient for the successful completion of the project resulting in chapter 4.

Richard Gill has been an excellent supervisor who has made the research period leading to this work into an intriguing and intensive one. He has been helpful with dozens of valuable suggestions and comments. His enthusiasm, broad knowledge and philosophical approach to problems kept the research work exciting and pleasant.

My former office-mate Bart Wijers at the Utrecht University was a good person to talk with and brought humour and atmosphere to our office in the mathematical institute. Our discussions on the line-segment model where helpful for the completion of chapter 5. Another person who gave extra flavour to the days in the office was Richard Starmans from the University of Tilburg. His phone calls lead to discussions about all kinds of sports (in particular chess and Ajax), politics, economics and if there was time left about our work.

Last, but not least, I thank Martine for being of great support and being an excellent organizer and relaxing company.

# Contents

**Introduction.** 1

Overview. . . . . 5

**1 Basic Theory** 7

1.1 Weak convergence in non-separable metric spaces. . . . . 7

1.2 Empirical processes. . . . . 10

1.2.1 Uniform Donsker classes and some basic multivariate techniques. . . . . 12

1.3 The functional delta-method. . . . . 15

1.4 Efficiency theory. . . . . 17

## Part I: Efficiency Theory and Applications for (Non-parametric) Maximum Likelihood Estimators

**2 Efficiency Theory for the (NP)MLE and an Identity for Linear Parameters in Convex Models** 25

2.1 Introduction. . . . . 25

2.2 Efficient score equation for NPMLE. . . . . 27

2.3 Efficiency theorem for NPMLE. . . . . 28

2.4 An Identity for linear parameters in convex models. . . . . 32

2.4.1 Discussion about the (ir)relevance of the identity conditions. . . . . 34

2.4.2 The gain from the identity. . . . . 35

2.5 Application of the identity to convex models which are linear in the parameter. . . . . 36

2.5.1 Invertibility of the information operator. . . . . 38

2.5.2 Example. . . . . 39

2.6 Efficiency theorem for NPMLE of linear parameters in convex models. . . . . 43
2.6.1 One step-estimators. . . . . 44
2.7 Bibliographic remarks. . . . . 45

**3 Efficiency of the sieved-NPMLE for a Class of Missing Data Models with Applications. 49**
3.1 Introduction. . . . . 49
3.2 A class of missing data models. . . . . 53
3.2.1 Verification of assumptions 1 and 2 for the examples. . . 56
3.3 Existence of sieved-NPMLE and EM-equations. . . . . 59
3.3.1 Existence and uniqueness of sieved-NPMLE. . . . . 60
3.3.2 EM-equations for the sieved-NPMLE. . . . . 61
3.3.3 Identifiability of the self-consistency equation. . . . . 63
3.4 Efficiency of the sieved-NPMLE. . . . . 64
3.4.1 Supremum norm invertibility of the information operator. 67
3.4.2 Weak assumption approach. . . . . 70
3.4.3 The P-Donsker class condition. . . . . 71
3.4.4 Verification of assumptions 3, 4, 5.1, 5.2 and 5.3 for the examples. . . . . 73
3.4.5 The P-consistency condition. . . . . 75
3.4.6 Identity condition. . . . . 77
3.5 Final theorems and results for the examples. . . . . 77

**4 Efficient Estimation in the Bivariate Censoring Model and Repairing NPMLE. 79**
4.1 Introduction. . . . . 79
4.2 SOR-MLE for the bivariate censoring model. . . . . 85
4.2.1 Existence and uniqueness of the SOR-MLE and EM-equations. . . . . 88
4.3 Outline of the efficiency proof. . . . . 89
4.4 Proof of efficiency of SOR-MLE. . . . . 93
4.4.1 Uniform consistency of $F_n^h$ for $h_n \to 0$. . . . . 94
4.4.2 Empirical process condition. . . . . 94
4.4.3 Approximation condition. . . . . 99
4.5 Results. . . . . 101
4.6 The bootstrap. . . . . 103
4.7 Technical lemmas. . . . . 105
4.7.1 Proof of lemma 4.2. . . . . 106

4.7.2 Proof of lemma 4.4. . . . . 109
4.7.3 Proof of lemma 4.8. . . . . 110

**5 Efficiency of the NPMLE in the Line-Segment Problem. 115**
5.1 Introduction to the line-segment process problem. . . . . 115
5.2 Existence and uniqueness of the sieved-NPMLE, EM-equations. 123
5.3 Efficiency result and outline of proof. . . . . 125
5.3.1 Construction of confidence intervals. . . . . 128
5.4 Donsker class condition, uniform bound on the variation of the efficient influence function. . . . . 129
5.4.1 Solving for the hardest submodel. . . . . 130
5.4.2 The information after $\tau$. . . . . 134
5.4.3 Estimation of the efficient influence curve in practice. . 134
5.4.4 Bounded variation of the hardest underlying scores. . . 135
5.4.5 Donsker class condition. . . . . 137
5.5 Consistency condition for the efficient influence function. . . . . 138
5.5.1 The identity condition . . . . . 139
5.6 Discussion. . . . . 140
5.6.1 Inhomogeneous Poisson Process. . . . . 141

**Part II: Inefficient Estimation in Semiparametric Models**

**6 Inefficient Estimators of the Bivariate Survival Function in the Bivariate Censoring Model 145**
6.1 Three approaches to estimation. . . . . 145
6.2 Lemmas. . . . . 155
6.3 Differentiability of the Dabrowska, Volterra, and Prentice and Cai representations of $F$. . . . . 162
6.3.1 The Volterra Representation. . . . . 162
6.3.2 The Dabrowska representation. . . . . 167
6.3.3 The Prentice-Cai representation. . . . . 173
6.3.4 Differentiability theorem for the three representations. . 173
6.4 The estimators. . . . . 174
6.5 Asymptotic properties of the estimators. . . . . 175
6.5.1 Final results. . . . . 176
6.6 Influence curves. . . . . 178
6.6.1 Computation of the influence curves. . . . . 180

6.6.2 Optimal influence curve under complete independence. . 182

**7 Modified EM-Estimator of the Bivariate Survival Function 185**

7.1 Introduction. . . . . 185
7.2 A Modified EM-estimator (Pruitt). . . . . 188
7.3 Outline of proof of consistency and weak convergence. . . . . 190
7.4 Invertibility of the derivative of the modified self-consistency equation. . . . . 193
7.5 Functional delta-method for functionals of density estimators. . 197
7.6 Weak convergence of the explicit part. . . . . 198
7.6.1 Uniformly consistent edge-corrected bivariate density estimators. . . . . 200
7.6.2 Bivariate kernel density estimators: simultaneous uniform consistency, consistency of derivative and asymptotic normality of the integrated kernel density estimator. 201
7.6.3 Application of the functional delta-method for density estimators. . . . . 205
7.6.4 Analytical part of the analysis. . . . . 207

**References 212**

**Notation 218**

# Introduction.

Before giving an overview of this tract, we will start with a general discussion about statistics. An experimenter is usually interested to understand a certain observable variable, which we call $X$. If he is in the circumstance that he is able to find deterministic factors, controlled and measured without error, which completely determine $X$, then he has total control over the variable $X$. In this case he does not have to be interested in *statistics*.

Fortunately, life is not that perfect. Usually there will be an essential part of $X$ which is not understood. A natural approach is now to consider $X$ as a *random variable* with a probability distribution which is a member of a certain set in which the knowledge about $X$ is incorporated. Such a set of probability distributions is called a *model*. The experimenter generates now $n$ *independent* and *identically* distributed random variables $X_i$, $i = 1, \ldots, n$, all having the same distribution as $X$, say $P$, by $n$ times repeating the experiment under identical circumstances. He is now concerned with estimating the distribution $P$ of $X$, or a function of $P$, using the observations $X_i$, and knowing that $P$ is an element of the model. The elements which he wants to estimate, $P$ or functionals of $P$, are called *parameters* and an *estimator* of such a parameter is just a function of the observations $X_i$, $i = 1, \ldots, n$.

In order to illustrate the usefulness of considering a variable of interest as a random variable with a probability distribution which lies in a certain model, we give the following example (which is well known from text books):

**Example 0.1 (The quiz-master problem).** I am the quiz-master. There are three doors and I know that behind one of the doors there is placed a car. If you guess the right door, then you get the car. You choose one of the three doors. To keep the game exciting I open one of the remaining two doors, of course one with no car behind it. After having opened such an empty door I give you another opportunity to choose a door. Do you change your mind and choose the other closed door or do you stick to your first decision?

For convenience, assume that your first choice was the first door and that I opened the second door. The fact that I opened the second door tells you that the car is behind the first or third door. However, it tells you more. Assume that the car is behind the first door, then I could also have opened the third door. On the other hand if the car is behind the third door I had only one choice, namely I had to open the second door. Therefore, in order to make what you observed, namely that I opened the second door, most likely you should change your mind and choose the third door.

Another explanation of why you should change your mind is the following. If we play this experiment 1000 times and you always stick to your first decision, then you expect to win 1/3 of the 1000 cars. But if you would have changed your mind each of these thousand times, then you would expect to win the remaining 2/3 of 1000 cars.

Consider now the case where we have 100 doors and I open 98 doors and then ask you if you want to change your mind. Changing your mind this last time would give you the car with probability 99/100.

In this example we are interested in estimating the *parameter* $\theta_0$, which is the number of the door with the car behind it. Then $\theta_0 \in \{1, 2, 3\}$. Assume you chose door 1. Consider the number of the door opened by the quiz-master as a random variable $X$. If we assume that the quiz-master does not open the door with the car and the door which has been chosen by you and that he has no other preferences for opening a door, then the distribution $P$ of $X$ is completely determined by $\theta$ as follows:

for all $\theta$ we have $P_\theta(X = 1) = 0$, and
if $\theta = 1$, then $P_\theta(X = 2) = P_\theta(X = 3) = 1/2$,
if $\theta = 2$, then $P_\theta(X = 2) = 0$, $P_\theta(X = 3) = 1$,
if $\theta = 3$, then $P_\theta(X = 2) = 1$, $P_\theta(X = 3) = 0$.

Assume that the quiz-master opens the second door; in other words we observe $X = 2$. An intuitively appealing *estimator*, a function of the observation $X = x$, $\theta(x)$ of $\theta_0$ based on the observation $x = 2$ is obtained by maximizing $P_\theta(X = 2)$ over $\theta \in \{1, 2, 3\}$; in other words make the observation $X = 2$ as likely as possible. $\theta(x)$ is called the *maximum likelihood* estimator of $\theta_0$ and for $x = 2$ it is given by $\theta(2) = 3$, which agrees with the heuristically derived optimal decision in the example. Notice that $\theta(2) = 3$ would be a MLE for any distribution $P_1$; so the third door would be the right choice even if the quiz-master would have a preference for one of the two empty doors.

The model in this example was given by $\{P_\theta : \theta \in \{1, 2, 3\}\}$. Models for which the distribution is determined up to a finite dimensional vector are called *parametric models.* In this example, the model contains only three elements, but is still very realistic; any reasonable quiz-master will open doors which he is allowed to open with equal probability and will not spoil the game by opening a door with the car behind it. Another nice feature of this three element model is that one observation of $X$ makes one element of the model already essentially more likely than the others. This is a consequence of the fact that the model contains only a few elements. In general we have that the larger the model

is, the more observations one needs in order to do some essential statistical inference about the probability measure of the data.

Heuristically, the reader is hopefully convinced that the maximum likelihood estimator in this example is the only correct estimator to use. Formally, this requires a notion of *efficiency* of estimators. Firstly, one needs a criterium to judge the performance of an estimator. For example, as criterium one might take the supremum over all possible distributions of the data of the expectation (under such a distribution) of the squared difference between the estimator and the parameter. Now one establishes a lower bound on the performance w.r.t. this criterium and one calls an estimator efficient if it attains this bound. The lower bound is only interesting if it is achievable, i.e. if it is the greatest lower bound. For establishing the lower bound for a certain criterium one can decide to restrict oneself to a class of estimators with certain properties. For example, one can restrict oneself to the class of unbiased estimators and as criterium take the variance of the estimator. Then the variance of such an unbiased estimator is always larger than or equal to the well known *Cramér-Rao* bound. A nice feature of the *Cramér-Rao* bound is that the bound can be derived for each number of observations. However, most estimators are not unbiased; unbiased (for each possible $P$) estimators often do not even exist. So far there has not been developed a general *finite sample efficiency theory* for a more interesting class of estimators than just the class of unbiased estimators. Instead one restricts oneself to "asymptotically (number of observations converges to infinity) unbiased" estimators which are known to converge at $\sqrt{n}$-ratein distribution and then establishes the Cramér-Rao bound for the variance of the limit distribution. An estimator is now called *asymptotically efficient* if it attains the Cramér-Rao bound in the limit.

In *parametric models* an asymptotic efficiency theory based on this Cramér-Rao lower bound has been developed. This efficiency theory has been generalized to all models, not only parametric models. For parametric models it has been shown that under some natural assumptions this Cramér-Rao bound is attained by *maximum likelihood* estimators or modifications thereof. For a literature overview and description of a general efficiency theory we refer to Bickel, Klaassen, Ritov and Wellner (1993). The for us relevant theory is summarized in chapter 1.

In many experiments one observes random variables $X$ which are hardly understood. For example, one might be interested in the life-time $X$ of a patient with a completely new and thereby unknown disease. In this case the experimenter does not want to make any essential assumptions on the

distribution $P$ of $X$, because each assumption would be a guess. In this case the model would consist of all probability distributions or a set which lies dense in this set, where dense is in the sense that the models cannot be statistically distinguished. We call such a model a *nonparametric model.* Suppose that the experimenter observes $n$ identically and independent lifetimes $X_1, \ldots, X_n$ of such patients. A natural estimator of $P$ is now the *empirical distribution*

$$P_n(X \in B) \equiv \text{the fraction of } X_i \text{ which fall in } B.$$

If we subtract from $P_n(B)$ its expectation $P(B)$, then we have an average of $n$ i.i.d. random variables $I_B(X_i) - P(B)$ with mean zero, where $I_B(\cdot)$ is the indicator of the set $B$. The *central limit theorem* tells us that such an average multiplied with $\sqrt{n}$ *converges in distribution* to the *normal distribution* with mean zero and variance equal to the variance of $I_B(X)$. The empirical distribution is for this nonparametric model also an *efficient estimator* of $P$ with respect to a *generalized Cramér-Rao* lower bound.

A great deal of literature is concerned with the so called *uniform central limit theorem*, where uniform is concerned with uniform in a collection of sets $B$ and, more generally, uniform in any class of i.i.d. random variables, instead of only indicators $I_B(X_i)$. In this theory one considers $P_n$ as a random element of a space of real valued functions.

This so called *empirical process theory* solves the problem of proving efficiency of estimators of parameters in *nonparametric models* to satisfaction, using the fact that nice parameters are smooth functionals of $P$ and thereby inherit all properties of the empirical distribution. For an overview of the empirical process literature we refer to Wellner (1992). The for us relevant theory will be summarized in chapter 1.

It is not surprising that there is a large remaining area between parametric models and nonparametric models. We call each model which belongs to neither of them a *semiparametric model.* There exists an abundance of interesting and natural applications which are described by semiparametric models. Consider for example the following problem. One observes a random variable $X$ and one knows that $X$ is a known function of a random variable $Y$ which has a completely unknown distribution. Here the distribution of $X$ has a special structure induced by the known function. Therefore the model corresponding with $X$ will often be essentially smaller than the *nonparametric* model consisting of all probability measures. Consequently, one should not be satisfied anymore with the empirical distribution $P_n$ as an estimator of the distribution of $X$. $P_n$ might even not be a member of the model, which already explains

that one can do better.

An important class of models which can be described in this way are the *missing data models*. Here one is interested in a certain random variable $X$ with unknown distribution, but because of a certain irrelevant (random) factor one is only able to observe that $X$ falls in a certain region; so each observation tells you something about the value of $X$ but it does not have to determine $X$ completely. In this case the observation can often be described as a known many to one mapping on $X$ and the irrelevant factor.

## Overview.

We close this section with a short overview of this tract. In the next chapter we will give an introduction to existing relevant theory which forms a basis for this tract: *weak convergence theory*, *empirical process theory*, *efficiency theory* ans some multivariate techniques. The rest of the tract consists of two parts: chapter 2, 3, 4 and 5 form the first part and chapter 6, 7 the second part.

The first part covers general efficiency theory for maximum likelihood estimators (MLE) and applies this theory to a general class of missing data models and two interesting applications. In chapter 2 we present a method for proving efficiency of MLE. The method is applicable to all models. In this theory a lot of significance occurs if the model is *convex*, which means that if one moves along a straight line from one element to another element of the model, then one does not leave the model. Using this convexity we establish a useful *identity* for MLE which in a straightforward manner provides us with consistency, efficiency and validity of the bootstrap under minimal conditions. The theorem can be trivially extended to all kinds of modifications of MLE.

Many *semiparametric models* are convex. In chapter 3 we apply this efficiency theory for convex models to MLE in a general class of *missing data models* and illustrate it with several examples. Moreover, in chapter 4 and 5 we successfully apply this theory to the *bivariate censoring* and *line segment* model, which are models where the standard approaches based on the self-consistency equation require too strong conditions. For the bivariate censoring model we propose and prove efficiency of a MLE which is based on a slight reduction of the data.

For reading chapter 2 one only needs to read the empirical process theory section and efficiency theory section of chapter 1. The general structure of the efficiency proofs in chapter 3, 4 and 5 are applications of the main theorems presented in chapter 2, but except for this the three chapters are self-contained and can be read independently of each other.

In the second part of the tract we study the construction of interesting *inefficient estimators* in the *bivariate censoring model* and analyze them by considering the estimators as functionals of the empirical distribution and establishing the required differentiability of these functionals. These estimators are especially interesting because they are easy to compute, have a good practical performance, and are very robust to changes of the distribution where we sample from so that the bootstrap works well. In chapter 6 we use the generally applicable *functional delta-method* in order to give full analyses of three explicit estimators. This chapter is joint work with Richard Gill and Jon Wellner. In chapter 7 an ad hoc modification of a MLE (a so called $M$-estimator) for the bivariate censoring model, using density estimators, is analyzed by applying the *implicit function theorem* and a refinement of the usual *functional delta-method* as used in chapter 6.

For chapter 6 and 7 one only needs to read the functional delta-method section of chapter 1 and chapter 6 and 7 can be read independently.

We refer to our *notation index* at the end of this book. We have grouped the notation, which is all introduced in the next chapter, in the following categories: general, weak convergence theory, empirical process theory, efficiency theory.

# Chapter 1

# Basic Theory

## 1.1 Weak convergence in non-separable metric spaces.

In our applications we will consider estimators as random elements of a Banach space of functions. Suppose that $(D, \|\cdot\|)$ is such a Banach space, $(\mathcal{X}_n, \mathcal{A}_n, P_n)_{n\geq 0}$ is a sequence of probability spaces, and

$$X_n : \mathcal{X}_n \to D, \text{ for } n = 0, 1, 2, \ldots \text{ are arbitrary maps.}$$

We endow $D$ with the Borel sigma-algebra; the smallest $\sigma$-field containing the open sets and which makes each continuous real valued function measurable.

For many interesting applications it is natural to consider $X_n$ as an element of a non-separable space. In this case the Borel-sigma algebra is often very large and therefore $X_n$ will usually not be measurable. On the other hand, for all known applications the limit random variable $X_0$ lies in a separable (sub)space and thereby will be measurable w.r.t. the Borel sigma-algebra, except for some pathological cases.

Because we are only concerned with the asymptotic behavior of $X_n$, only "asymptotic measurability" should be relevant. Indeed there exists a powerful weak convergence theory for non-separable spaces without giving up the Borel sigma-algebra, but giving up that $X_n$ induces a distribution on the Borel-sigma algebra. In the definition of weak convergence as used in Pollard (1984), which also generalizes the traditional definition of Billingsley (1968), $D$ is endowed with that (often closed-ball) sigma-algebra which makes $X_n$ measurable and thereby makes probabilities for $X_n$ of events in this sigma-algebra well defined. In the modern theory with the Borel sigma-algebra expectations and

probabilities for $X_n$ are replaced by *outer expectations* and *outer probabilities*. This weak convergence theory is due to Hoffmann-Jørgensen (1984) and Dudley (1985) following an evolution from Dudley (1966) and Wichura (1968) and is presented in full details in van der Vaart and Wellner (1995).

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $f : \Omega \to \mathbb{R}$ is an arbitrary function. If $f$ is measurable we define $Pf \equiv \int f dP$. An *outer expectation* is defined as follows:

$$P^*f \equiv \inf\{Ph \equiv \int h dP : \ f \leq h \text{ and } h \text{ is measurable}\}.$$

Similarly, we define *outer probability*:

$$P^*(A) \equiv \inf\{P(B) : B \supset A, \ B \in \mathcal{A}\}. \tag{1.1}$$

It can be verified that $P^*f = Pf^*$ and $P^*(A) = P(A^*)$ for a certain measurable $f^*$ and measurable set $A^*$. In other words, the infimum in the definition of outer expectation and outer probability is attained.

Let $C_b(D)$ be the collection of bounded, continuous functions $h$ from $D$ to $\mathbb{R}$.

**Definition 1.1** *We say that* $X_n$ converges weakly to a Borel measurable random element $X_0$ *in* $(D, \|\cdot\|)$, *and write* $X_n \stackrel{D}{\Longrightarrow} X_0$, *if for every* $h \in C_b(D)$,

$$P_n^* h(X_n) \to P_0 h(X_0) = \int h(X_0)(x) dP_0(x), \quad \text{as } n \to \infty.$$

As can be straightforwardly verified a heuristically appealing equivalent characterization of weak convergence is given by:

$$\lim_{n\to\infty} P^*(X_n \in A) = P(X_0 \in A)$$

for every Borel set $A$ with $P(X_0 \in \partial A) = 0$. We say that $X_0$ is tight if for each $\epsilon > 0$ there exists a compact set $K$ so that $P(X_0 \in K) > 1 - \epsilon$.

As usual define

$$L^2(P) \equiv \{f : (\Omega, \mathcal{A}, P) \to \mathbb{R} : \ f \text{ measurable and } \int f^2 dP < \infty\},$$

which is endowed with the Hilbert space inner-product norm $\|f\|_P \equiv \sqrt{\langle f, f\rangle_P} \equiv \sqrt{\int f^2 dP}$. Let $\mathcal{F} \subset L^2(P)$. Estimators can often be considered as a random element of (i.e. this is our space $D$)

$$l^\infty(\mathcal{F}) \equiv \{H : \mathcal{F} \to \mathbb{R} : \|H\|_{\mathcal{F}} \stackrel{def}{\equiv} \sup_{f\in\mathcal{F}} \mid H(f) \mid < \infty\} \tag{1.2}$$

for some class $\mathcal{F} \subset L^2(P)$. Let $X_n : (\mathcal{X}_n, \mathcal{A}_n) \to \ell^\infty(\mathcal{F})$ for a certain class of functions $\mathcal{F} \subset L^2(P)$. Given $X : (\mathcal{X}, \mathcal{A}) \to \ell^\infty(\mathcal{F})$ we consider the following semi-metric on $\mathcal{F}$

$$\rho_{X,P}(f,g) = \int (X(f) - X(g))^2 \, dP. \tag{1.3}$$

$X_n$ is called *asymptotically uniformly $\rho_{X,P}$-equicontinuous in probability* if for every $\epsilon, \eta > 0$, there exists a $\delta > 0$ such that

$$\limsup_{n \to \infty} P^* \left( \sup_{\rho_{X,P}(f,g) < \delta} | X_n(f) - X_n(g) | > \epsilon \right) < \eta. \tag{1.4}$$

A Borel measurable map $X$ in $l^\infty(\mathcal{F})$ is called *Gaussian* if each of its finite dimensional marginals $(X(f_1), \ldots, X(f_k))$ has a multivariate normal distribution on Euclidean space.

By using a similar proof as the proof of the Ascoli-Arzéla theorem, which gives conditions under which a pointwise convergent sequence converges also uniformly, the following important theorem can be proved:

**Theorem 1.1** *A sequence $X_n$ converges weakly in $\ell^\infty(\mathcal{F})$, $\mathcal{F} \subset L^2(P)$, to a tight Gaussian process $X$ if and only if*

- *the finite dimensional marginals of $X_n$ converge weakly to the corresponding marginals of $X$; and*
- *$X_n$ is asymptotically $\rho_{X,P}$-equicontinuous in probability.*

If $X_n \stackrel{D}{\Longrightarrow} X$, then one might hope that the weak convergence of 'smooth' functionals of $X_n$ is preserved, in order not have to verify the weak convergence for each special application. The following theorem tells us that weak convergence is preserved under continuous mappings, where the mappings are allowed to depend on $n$.

**Theorem 1.2** (Extended continuous mapping theorem). *Let $(D, d)$ and $(E, e)$ be metric spaces. Let $D_n \subset D$ and $g_n : D_n \to E$ satisfy: if $x_n \to x$ with $x_n \in D_n$ for every $n$ and $x \in D_0$, then $g_n(x_n) \to g(x)$, where $D_0 \subset D$ and $g : D_0 \to E$. Suppose that $X \in D_0$ and $g(X)$ are Borel measurable and separable.*

*Then $X_n \stackrel{D}{\Longrightarrow} X$, $X_n \in D_n$, implies that $g_n(X_n) \stackrel{D}{\Longrightarrow} g(X)$.*

When applying this theorem it is effective to choose $D_n$ as small as possible by putting all known properties of $X_n$ in $D_n$. Then for showing $g_n(X_n) \stackrel{D}{\Longrightarrow} g(X)$ we only have to verify the convergence $g_n(x_n) \to g(x)$ for sequences $x_n \in D_n$.

## 1.2 Empirical processes.

Let $X_1, \ldots, X_n$ be a sample of i.i.d. random elements in a measurable space $(\mathcal{X}, \mathcal{A})$ each with law $P$ and let $P_n$ be the empirical measure which puts mass $1/n$ on each $X_i$, $i = 1, \ldots, n$. For a collection $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ consider the map from $\mathcal{F}$ to $\mathbb{R}$ given by

$$f \to P_n f = \int f dP_n.$$

We will consider $P_n = (P_n f : f \in \mathcal{F})$ as a random element of $\ell^\infty(\mathcal{F})$. The normalized version of this map is the $\mathcal{F}$-indexed *empirical process* given by

$$f \to G_n f = \sqrt{n}(P_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf).$$

For a given function of $f$ one has the law of large numbers and the central limit theorem

$$P_n f \xrightarrow{\text{as}} Pf \qquad\qquad G_n f \xRightarrow{D} N\left(0, P(f - Pf)^2\right), \tag{1.5}$$

provided $Pf$ exists and $Pf^2 < \infty$, respectively.

Empirical process theory is concerned with making these statements uniform in $f$ varying over a class $\mathcal{F}$. For a nice literature overview we refer to Wellner (1992) and for a self-contained presentation we refer to van der Vaart and Wellner (1995).

The uniform version of the law of large numbers becomes

$$\|P_n - P\|_{\mathcal{F}} \xrightarrow{\text{a.s.}^*} 0.$$

With a.s.* we mean that the convergence is outer almost surely; there exists a measurable set with $P$-measure zero so that the convergence holds outside this set. A class for which this is true is called a *P-Glivenko Cantelli* class.

For the uniform version of the central limit theorem it is assumed that $\mathcal{F} \subset L^2(P)$ and

$$\sup_{f \in \mathcal{F}} | f(x) - Pf | < \infty, \text{ for every } x.$$

Under this condition the empirical process $(G_n f : f \in \mathcal{F})$ can be viewed as an element of $\ell^\infty(\mathcal{F})$. Consequently, it makes sense to investigate conditions under which

$$G_n \xRightarrow{D} G \text{ in } \ell^\infty(\mathcal{F}),$$

where the limit $G$ is a tight Borel measurable element in $l^\infty(\mathcal{F})$. Then we say that the uniform central limit theorem holds at $P$. Classes $\mathcal{F}$ for which the uniform central limit theorem at $P$ holds are called *$P$-Donsker classes.* Because $\mathcal{F} \subset L^2(P)$ the multivariate central limit theorem characterizes the finite dimensional distributions of $G$ uniquely:

$$(G_n f_1, \ldots, G_n f_k) \stackrel{D}{\Longrightarrow} N(0, \Sigma),$$

where the $k \times k$ matrix $\Sigma$ has $(i,j)$-th element $P(f_i - Pf_i)(f_j - Pf_j)$. A tight Borel measurable random element is completely determined by its finite dimensional distributions. Consequently, if the weak convergence holds, then $G = G_P$ is completely determined. $G_P$ is called the *$P$-Brownian Bridge.*

Notice that $\rho_{G,P}(f,g)$ as defined in (1.3) equals

$$\rho_P(f,g) \equiv \int ((f-g) - P(f-g))^2 \, dP.$$

If $\mathcal{F}$ is a $P$-Donsker class, then by theorem 1.1 we know that $G_n$ is asymptotically $\rho_P$-equicontinuous: for $\delta_n \downarrow 0$

$$\|G_n\|_{\mathcal{F}_{\delta_n}} = \sup_{h \in \mathcal{F}_{\delta_n}} | G_n(h) | \stackrel{P^*}{\to} 0, \tag{1.6}$$

where

$$\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \rho_P(f,g) < \delta\}. \tag{1.7}$$

Notice that we suppressed the dependence on $P$ in the notation $\mathcal{F}_\delta$. In fact theorem 1.1 and the multivariate C.L.T. tells us that condition (1.6) is sufficient for $\mathcal{F} \subset L^2(P)$ to be a $P$-Donsker class.

Weak convergence can be metrized with a metric $d$ so that weak convergence of $G_{n,P} \equiv \sqrt{n}(P_n - P)$ in $\ell^\infty(\mathcal{F})$ to the Brownian Bridge $G_P$ is equivalent with $d(G_{n,P}, G_P) \to 0$. We call $\mathcal{F}$ *Donsker uniformly in* $P \in \mathcal{M}$ for a certain collection of probability measures $\mathcal{M}$ if this convergence is uniform in $P \in \mathcal{M}$.

It is not surprising that if $\mathcal{F}$ is uniform Donsker in $P \in \mathcal{M}$, then we have $\rho_P$-equicontinuity uniformly in $P \in \mathcal{M}$: for $\delta_n \downarrow 0$

$$\sup_{P \in \mathcal{M}} \|G_{n,P}\|_{\mathcal{F}_{\delta_n}} \stackrel{P_*}{\to} 0. \tag{1.8}$$

One might hope that an empirical process indexed by a uniform (in $\mathcal{M}$) Donsker class $\mathcal{F}$ based on sampling from a sequence $P^{(n)} \in \mathcal{M}$ which "converges" to $P$ converges weakly to $G_P$. This statement can be made precise as follows.

For each $n$ let $X_{n1}, \ldots, X_{nn}$ be i.i.d. with probability measure $P^{(n)}$ and let $\widehat{P}_n$ be the corresponding empirical measure. Define

$$G_{n,P^{(n)}} \equiv \sqrt{n}(\widehat{P}_n - P^{(n)}).$$

**Theorem 1.3** (Sampling from approximating measures). *Assume that $\mathcal{F}$ is Donsker uniformly in $\{P^{(n)}\}$. Moreover, suppose that $\rho_{P^{(n)}}$ converges uniformly to $\rho_P$ in the sense that*

$$\sup_{f,g\in\mathcal{F}} |\rho_{P^{(n)}}(f,g) - \rho_P(f,g)| \to 0 \tag{1.9}$$

*and that the marginals of $G_{n,P^{(n)}}$ converge in distribution to the marginals of $G_P$, the P-Brownian Bridge.*

*Then $G_{n,P^{(n)}} \stackrel{D}{\Longrightarrow} G_P$.*

For establishing the convergence of the marginals the following lemma is useful. The lemma is proved in Bickel and Freedman (1981) and uses the Mallows-metric (Mallows, 1972).

**Lemma 1.1** *Suppose $P^{(n)}(f - P^{(n)}f)^2 \to P(f - Pf)^2$ and $f(X^{(n)}) \stackrel{D}{\Longrightarrow} f(X)$, $X^{(n)} \sim P^{(n)}$, $X \sim P$.*

*Then $G_{n,P^{(n)}} f \stackrel{D}{\Longrightarrow} G_P f$.*

Let $F \equiv \sup_{f\in\mathcal{F}} | f |$ be the so called envelope of $\mathcal{F}$. For the case that $P^{(n)}$ is just the empirical measure $P_n$ based on an i.i.d. sample $X_1, \ldots, X_n$ of $P$, we have the following theorem due to Giné and Zinn (1990):

**Theorem 1.4** (Empirical Bootstrap). *$G_{n,P_n} \stackrel{D}{\Longrightarrow} G_P$ given almost all sequences $X_1, X_2, \ldots$ if and only if $\mathcal{F}$ is Donsker and $PF^2 < \infty$.*

### 1.2.1 Uniform Donsker classes and some basic multivariate techniques.

In this tract we will consider several multivariate models in the sense that the i.i.d. observations are multivariate vectors and therefore we will be concerned with estimation of a multivariate distribution function. For these applications we need that the following classes are uniform Donsker.

**Example 1.1 (Indicators).** For $a, b \in \mathbb{R}^2$ we use the partial ordering; $a < b \Leftrightarrow (a_1 < b_1) \wedge (a_2 < b_2)$ and $a \leq b \Leftrightarrow (a_1 \leq b_1) \wedge (a_2 \leq b_2)$. We denote the indicator of the rectangle $(a, b] \equiv \{c : a < c \leq b\} \subset \mathbb{R}^2$ with $I_{(a,b]}$, which is a function from $\mathbb{R}^2 \to \mathbb{R}$. Define

$$\mathcal{F} \equiv \{I_{(a,b]} : a, b \in \mathbb{R}^2, a < b\}.$$

This is a Donsker class uniform in all probability measures on $\mathbb{R}^2$. The same holds for $[a, b)$, $(a, b)$, $[a, b]$ and for the general $\mathbb{R}^k$ case. The class of real valued functions on $\mathbb{R}$ with variation smaller than $M < \infty$ is a generalized

convex hull of indicators and therefore it is also uniformly Donsker (see van der Vaart and Wellner, 1995).

**Example 1.2 (Uniformly bounded uniform sectional variation).** Let $[0,\tau] \subset \mathbb{R}^2$ be a fixed rectangle. Let $f : [0,\tau] \to \mathbb{R}$ be a real valued *bivariate* function on $[0,\tau]$. The *generalized difference of $f$ over $(a,b]$* is defined as

$$f(a,b] \equiv f(b_1,b_2) - f(a_1,b_2) - f(b_1,a_2) + f(a_1,a_2).$$

The *variation norm of $f$*, which will be denoted with $\|f\|_{\mathrm{v}}$, is defined as the supremum over all *lattice* (rectangular) partitions of $[0,\tau]$ of the sum of the absolute values of the generalized differences of $f$ over the elements of the partition; let $\{A_{i,j}\}$ be a collection of disjoint rectangles forming a lattice-partition of $[0,\tau]$, then

$$\|f\|_{\mathrm{v}} \equiv \sup_{\{A_{i,j}\}} \sum_{i,j} \mid f(A_{i,j}) \mid . \tag{1.10}$$

If $\|f\|_{\mathrm{v}} < \infty$, then we say that $f$ is of *bounded variation.* We will say that $f : [0,\tau] \to \mathbb{R}$ is of *bounded uniform sectional variation* if

$$\|f\|_{\mathrm{v}}^* \equiv \max\left(\|f\|_\infty,\ \|f\|_{\mathrm{v}},\ \sup_u \|v \to f(u,v)\|_{\mathrm{v}},\ \sup_v \|u \to f(u,v)\|_{\mathrm{v}}\right) \tag{1.11}$$

is finite. Define the bivariate cadlag function space $D[0,\tau]$ as in Neuhaus (1971):

**Definition 1.2** *$D[0,\tau]$, $[0,\tau] \subset \mathbb{R}^2$, is the vector space of bivariate functions $f : [0,\tau] \to \mathbb{R}$ for which (with $f(s+,t)$ we mean $\lim_{s_n \downarrow s, s_n > s} f(s,t)$)*

$$f(s,t) = f(s+,t+) = f(s+,t) = f(s,t+),$$

*and for which $f(s-,t+)$, $f(s-,t-)$ and $f(s+,t-)$ exist.*

The $k$-variate case is a trivial generalization of this definition (Neuhaus, 1971). Define now for any $0 < M < \infty$:

$$\mathcal{F}_M \equiv \{f \in D[0,\tau] : \|f\|_{\mathrm{v}}^* \leq M\}.$$

$\mathcal{F}_M$ is a Donsker class uniform in all probability measures on $[0,\tau]$.

We prove this by applying the continuous mapping theorem 1.2. Let $P$ be a probability measure on $[0,\tau]$ and let $P_n$ be the empirical distribution. Then, by example 1.1, $G_n(\cdot) \equiv \sqrt{n}(P_n - P)(0,\cdot] \in (D[0,\tau], \|\cdot\|_\infty)$, the empirical process indexed by rectangles $(0,\cdot]$, converges weakly to the $P$-Brownian bridge $G(\cdot)$. Define

$$\Phi : (D[0,\tau], \|\cdot\|_\infty) \to l^\infty(\mathcal{F}_M) :\ G \mapsto \left(f \to \int f dG\right),$$

where $\int f dG$ is defined by integration by parts if $G$ is not of bounded variation (as shown below). For proving that $\mathcal{F}_M$ is $P$-Donsker we need to show that $\Phi(G_n) \stackrel{D}{\Longrightarrow} \Phi(G)$ as elements in $l^\infty(\mathcal{F}_M)$. For this purpose we apply the continuous mapping theorem to $\Phi$. We already have the weak convergence of $G_n \in D[0,\tau]$ to $G$. It remains to prove the required continuity of $\Phi$:

$$\sup_{f \in \mathcal{F}_M} \mid \int f d(G_n - G) \mid \to 0 \tag{1.12}$$

for all sequences $G_n$, with $\|G_n\|_v < \infty$ and which converge in supnorm to $G$. This is proved by applying integration by parts as we will show now. Hildebrandt (1963, p. 108) provides us with:

**Lemma 1.2** *Let $f : [0,\tau] \subset \mathbb{R}^2 \to \mathbb{R}$ be cadlag and of bounded variation, then*

$$f(0,x] = (f_1 - f_2)(0,x],$$

*where $f_1, f_2$ generate positive finite measures on the Borel sigma-algebra on $[0,\tau]$.*

Now, Fubini's theorem provides us with the following *integration by parts formula* for 2 bivariate cadlag functions which are of bounded variation (and thereby by lemma 1.2 generate signed measures):

**Lemma 1.3** (Integration by parts). *Let $f, g \in D[0,\tau]$ and $\|f\|_v^* < \infty$, $\|g\|_v < \infty$.*

$$\begin{aligned}\int_0^s \int_0^t f(u,v) g(du,dv) &= \int_0^s \int_0^t g\left([(u,s) \times (v,t)]\right) f(du,dv) \\ &+ \int_0^s g\left([u,s] \times (0,t]\right) f(du,0) + \int_0^t g\left((0,s] \times [v,t]\right) f(0,dv) \\ &+ f(0,0) g\left((0,s] \times (0,t]\right).\end{aligned}$$

**Proof.** We refer to (Gill, 1992) for the general $\mathbb{R}^k$ case. It works as follows. Substitute

$$f(u,v) = \int_{(0,u] \times (0,v]} f(du',dv') + \int_{(0,u]} f(du',0) + \int_{(0,v]} f(0,dv') + f(0,0)$$

and apply Fubini's theorem. □

Notice that with these formulas we can also define these integrals for $g$ of unbounded variation and that if $F$ is zero at the bottom edges of $[0,\tau]$, then only the first term on the right hand-side is non-zero. With this integration by parts formula we can bound $\int f dg$ by $16\|g\|_\infty \|f\|_v^*$. So we have the following lemma:

**Lemma 1.4** *Let $f$ and $g$ be two bivariate cadlag functions ans suppose that $\|f\|_v^* < \infty$. Then*

$$\int_{[0,\tau]} f dg \leq 16 \|f\|_v^* \|g\|_\infty.$$

Now, (1.12) follows from lemma 1.4 by setting $g = G_n - G$, which proves that $\mathcal{F}_M$ is a $P$-Donsker class. The uniform Donsker class property is proved similarly.

The following lemma is useful:

**Lemma 1.5** *Let $f : \mathbb{R}^2 \to \mathbb{R}$. If $\|f\|_v^* < \infty$ and $f > \delta > 0$, then $\|1/f\|_v^* < \infty$.*

The proof requires some combinatorial arguments following directly from the definition (1.10) of $\|\cdot\|_v$ (it is sketched for general $k$ in Gill, 1993).

With the straightforward extensions of the definitions of $\|f\|_v^*$ and of $D[0,\tau]$, it is proved in the same way that the $k$-variate analogue of $\mathcal{F}_M$ is also uniformly Donsker (see Gill, 1992). We enforced the functions in $\mathcal{F}_M$ to be cadlag in order to guarantee that integrals w.r.t. $f \in \mathcal{F}_M$ are well defined; according to Hildebrandt (1963) this is not necessary if the integrand satisfies some weak continuity conditions.

Finally, we state the following lemma.

**Lemma 1.6** (Telescoping). *Let $a_i$, $i = 1, \ldots, k$, $b_i$, $i = 1, \ldots, k$ be real numbers.*

$$\prod_{i=1}^{k} a_i - \prod_{i=1}^{k} b_i = \sum_{j=1}^{k} \prod_{i=1}^{j-1} a_i (a_j - b_j) \prod_{i=j+1}^{k} b_i.$$

It can be easily verified (compare with the product rule for differentiating a product of functions) and it holds also for matrices (see Gill and Johansen, 1990). It is a very useful lemma for proving that differences of two products converge to zero and that is what we often have to do in differentiability proofs.

## 1.3 The functional delta-method.

Let $P_n$ be the empirical measure based on an i.i.d. sample from $P$. Suppose that $\mathcal{F}$ is a $P$-Donsker class. Then we know that $G_{n,P} = \sqrt{n}(P_n - P) \xrightarrow{D} G_P$ as random elements of $\ell^\infty(\mathcal{F})$. Suppose that we are interested in estimating $\Phi(P)$ for a certain functional $\Phi$. In large semiparametric and nonparametric models a natural and good estimator of $\Phi(P)$ might be $\Phi(P_n)$, assuming it is well defined. A question which naturally arises is if $\sqrt{n}(\Phi(P_n) - \Phi(P))$ also converges weakly,

just as $G_{n,P}$. The following theorem, the so called *functional delta-method*, which answers this question (by setting $Z_n = G_{n,P}$, $Z = G_P$, $F_n = P_n$ and $F = P$) is an immediate consequence of the extended continuous mapping theorem 1.2 by applying it to $g_n(Z_n) \equiv \sqrt{n}\,(\Phi(F + (1/\sqrt{n})Z_n) - \Phi(F))$:

**Theorem 1.5** (Functional delta-method). *Let $\Phi : D_\phi \subset (D, \|\cdot\|) \to (E, \|\cdot\|_1)$, where $(D, \|\cdot\|)$ and $(E, \|\cdot\|_1)$ are normed vector spaces. Endow $D$ and $E$ with the Borel sigma-algebra.*

*Suppose that $D_n, D_0, D_\phi \subset D$ are so that $F_n, F \in D_\phi$; $Z_n \equiv \sqrt{n}(F_n - F) \in D_n$; $Z \in D_0$, $D_0$ separable and*

1. *$Z_n \stackrel{D}{\Longrightarrow} Z$ in $(D, \|\cdot\|)$, where $Z$ is Borel measurable.*
2. *$\Phi$ satisfies the following differentiability property: if $h_n \equiv \sqrt{n}(G_n - G) \to h$, $G_n, G \in D_\phi$, with $h_n \in D_n$ and $h \in D_0$, then*

$$\sqrt{n}\left(\Phi(G + (1/\sqrt{n})h_n)) - \Phi(G)\right) - d\Phi(G)(h) \to 0 \tag{1.13}$$

*for a certain continuous linear mapping $d\Phi(G) : D_0 \subset (D, \|\cdot\|) \to (E, \|\cdot\|_1)$.*

*Then*

$$\sqrt{n}\,(\Phi(F_n) - \Phi(F)) \stackrel{D}{\Longrightarrow} d\Phi(F)(Z) \text{ in } (E, \|\cdot\|_1).$$

The delta-method formulated in this way, by using that the differentiability only has to be verified for sequences $h_n \in D_n$, is essentially more convenient than the usual formulation in terms of Hadamard-differentiability tangentially to a subspace: see Gill, 1989, Reeds, 1976, van der Vaart and Wellner (1995) and Wellner (1993). We will sometimes refer to the differentiability property 1.13 as compact differentiability, meaning (1.13) for an appropriate choice of $D_n$. For a discussion about the utility of the functional delta-method for analyzing estimators and also for obtaining its asymptotic variance we refer to chapter 6.

Let's proceed with our example. Assume that $\Phi$ satisfies the differentiability property (1.13). Then application of theorem 1.5 provides us with:

$$\sqrt{n}\,(\Phi(P_n) - \Phi(P)) \stackrel{D}{\Longrightarrow} d\Phi(P)(G_P).$$

The limit random variable $d\Phi(P)(G_P)$ is often unknown because we do not know $P$. In order to estimate the distribution of $\sqrt{n}(\Phi(P_n) - \Phi(P))$ we can use the bootstrap; let $P^{(n)}$ be an estimator of $P$ and suppose that by using

(e.g.) theorems 1.3 or (if $P^{(n)} = P_n$) theorem 1.4 we establish $G_{n,P^{(n)}} \equiv \sqrt{n}\left(\widehat{P}_n - P^{(n)}\right) \stackrel{D}{\Longrightarrow} G_P$ for almost all sequences $X_1, X_2, \ldots$, then we estimate the distribution of $\sqrt{n}(\Phi(P_n) - \Phi(P))$ with the distribution of $\sqrt{n}(\Phi(\widehat{P}^n) - \Phi(P^{(n)}))$, given $P^{(n)}$.

The following theorem tells us that if we are able to verify (1.13) uniformly in $P^{(n)}$, then the bootstrap works asymptotically for almost all sequences $X_1, X_2, \ldots$. Again, the theorem is an immediate consequence of the extended continuous mapping theorem.

**Theorem 1.6** *Let $\Phi : D_\phi \subset (D, \|\cdot\|) \to (E, \|\cdot\|_1)$, where $(D, \|\cdot\|)$ and $(E, \|\cdot\|_1)$ are normed vector spaces. Endow $D$ and $E$ with the Borel sigma-algebra. Suppose $D_n, D_0, D_\phi \subset D$ are so that $\widehat{F}_n, F_n, F \in D_\phi$; $Z_{n,F_n} \equiv \sqrt{n}(\widehat{F}_n - F_n) \in D_n$; $Z \in D_0$, $D_0$ separable, and*

1. *$Z_{n,F_n} \stackrel{D}{\Longrightarrow} Z$ in $(D, \|\cdot\|)$ for outer almost surely all sequences $F_n$, $Z$ is Borel measurable.*

2. *$\Phi$ satisfies the following differentiability property: if $h_n \equiv \sqrt{n}(\widehat{G}_n - G_n) \to h$, $\widehat{G}_n, G_n \in D_\phi$, with $h_n \in D_n$ and $h \in D_0$, then*

$$\sqrt{n}\left(\Phi(G_n + (1/\sqrt{n})h_n)) - \Phi(G_n)\right) - d\Phi(G)(h) \to 0 \tag{1.14}$$

*for a certain continuous linear mapping $d\Phi(G) : D_0 \subset (D, \|\cdot\|) \to (E, \|\cdot\|_1)$.*

*Then for outer almost surely all sequences $F_n$ we have:*

$$\sqrt{n}\left(\Phi(\widehat{F}_n) - \Phi(F_n)\right) \stackrel{D}{\Longrightarrow} d\Phi(F)(Z) \text{ in } (E, \|\cdot\|_1).$$

## 1.4 Efficiency theory.

Let $\mathcal{M}$ be a model, a set of probability measures, on $(\mathcal{X}, \mathcal{B})$. For two measures $F$ and $G$ on $\mathcal{B}$ we write $F \ll G$ if $F$ is absolute continuous with respect to $G$. Let

$$\mathcal{M}(\mu) \equiv \{P \in \mathcal{M} : P \ll \mu\}. \tag{1.15}$$

For each $P \in \mathcal{M}(\mu)$ we denote its density $dP/d\mu$ with $p$ and the collection of all these densities $p$ corresponding with $\mathcal{M}(\mu)$ will be denoted with $\mathcal{P}(\mu)$.

Let $\vartheta : \mathcal{M} \to \Theta \subset D$ be a parameter and $B$ be a collection of real valued linear mappings $b : D \to \mathbb{R}$. Given an i.i.d. sample $X_1, \ldots, X_n$ from an

unknown $P \in \mathcal{M}$ it is required to estimate the parameter $\theta = \vartheta(P)$, which is done by an estimator $\theta_n = \theta_n(X_1, X_2, \ldots, X_n)$. Here $b\theta_n : (\mathcal{X}^n, \mathcal{B}^n) \to \mathbb{R}$ is a measurable map for all $b \in B$.

We define $L_0^2(P)$ as all elements $f \in L^2(P)$ with $\int f dP = 0$.

**Definition 1.3** *A map $\epsilon \mapsto p_\epsilon$ from $[0,1] \subset \mathbb{R}$ to $\mathcal{P}(\mu)$ is called a* differentiable *(one- dimensional) submodel of $\mathcal{P}(\mu)$ through $p$ if there exist $g \in L_0^2(P)$ with*

$$\int \left( \frac{1}{\epsilon}(\sqrt{p_\epsilon} - \sqrt{p}) - \frac{1}{2} g \sqrt{p} \right)^2 d\mu \to 0 \text{ for } \epsilon \downarrow 0. \tag{1.16}$$

Notice that if the integrand in (1.16) would converge pointwise to zero, then

$$g(x) = \frac{d}{d\epsilon} \log(p_\epsilon(x)) \Big|_{\epsilon=0}.$$

Therefore $g$ can be considered as a $L^2(\mu)$ version of the *score function* of the one dimensional submodel $p_\epsilon$. Submodels $P_\epsilon \subset \mathcal{M}$ with densities $p_\epsilon \subset \mathcal{P}(\mu)$ for certain measure $\mu$ which satisfy (1.16) are called *Hellinger differentiable*. In the rest of this tract, if we write $p_{\epsilon,g} \in \mathcal{P}(\mu)$ or $P_{\epsilon,g} \in \mathcal{M}(\mu)$ we mean a one dimensional differentiable submodel of densities (w.r.t. $\mu$) or measures, respectively, with score $g$ as defined in definition 1.3.

The variance of unbiased (over $P_{\epsilon,g}$) estimators of $b\vartheta(P_{\epsilon,g})$ at $\epsilon = 0$ is bounded from below by the well known Cramér-Rao lower bound, which is given by

$$\frac{1}{n} \left( \frac{\frac{d}{d\epsilon} b\vartheta(P_{\epsilon,g})\big|_{\epsilon=0}}{\|g\|_P} \right)^2, \text{ as shown below,} \tag{1.17}$$

Let $\mathcal{S}(P)$ be class of differentiable submodels. The variance of unbiased (over whole $\mathcal{M}$) estimators of $b\vartheta(P)$ is bounded from below by the supremum over $\mathcal{S}(P)$ of (1.17), assuming that this supremum exists, which leads to the so called *generalized Cramér-Rao lower bound*. Since the Cramér-Rao lower bound (1.17) does only depend through $P_{\epsilon,g}$ on the score $g$ the supremum is in fact a supremum over the collection of scores corresponding with $\mathcal{S}(P)$. In the following definition of this collection of scores, note the difference between $S(P)$ and $\mathcal{S}(P)$.

**Definition 1.4** *A cone $S(P)$ in $L_0^2(P)$ is called a tangent cone at $P \in \mathcal{M}$ of $\mathcal{S}(P)$ if for all $g \in S(P)$ there exists a differentiable one dimensional submodel $P_{\epsilon,g} \in \mathcal{S}(P) \subset \mathcal{M}$ through $P \in \mathcal{M}$ with score $g$.*

Recall that a cone $C$ in a vector space over the reals is a subset which is closed under multiplication by nonnegative scalars: if $g \in C$, then $ag \in C$ for all $a \geq 0$.

As we will see below, the supremum over $S(P)$ in the *generalized Cramér-Rao lower bound* can be replaced by a supremum over the so called *tangent space* $T(P)$:

**Definition 1.5** *For a tangent cone $S(P) \subset L_0^2(P)$ we define the tangent space $T(P) \subset L_0^2(P)$ as the closure of the linear extension of $S(P)$ within $L_0^2(P)$.*

The following differentiability property of $b\vartheta$ guarantees the existence of the supremum over $S(P)$ of (1.17) and that the supremum over $T(P)$ is taken.

**Definition 1.6** *A parameter $b\vartheta : \mathcal{M} \to \mathbb{R}$ is called pathwise differentiable at $P \in \mathcal{M}$ relative to $S(P)$, if there exists a linear mapping $\dot{\vartheta} : T(P) \to (D, \|\cdot\|)$ so that $b\dot{\vartheta} : T(P) \to \mathbb{R}$ is continuous and linear and*

$$\frac{1}{\epsilon}\left(b\vartheta(P_{\epsilon,g}) - b\vartheta(P)\right) - b\dot{\vartheta}(g) \to 0 \text{ for all } g \in S(P).$$

*By the Riesz representation theorem there exists a $\widetilde{I}(P, b\vartheta) \in T(P)$ so that*

$$b\dot{\vartheta}(g) = \int \widetilde{I}(P, b\vartheta)(x) g(x) dP(x). \tag{1.18}$$

We have that the Cramér-Rao lower bound (1.17) equals $1/n \left(b\dot{\vartheta}(g)/\|g\|_P\right)^2$. Consequently, by (1.18) and the Cauchy-Schwarz inequality, this is maximized over $T(P)$ by $g = \widetilde{I}(P, b\vartheta)$ and therefore $P_{\epsilon,g}$, $g = \widetilde{I}(P, b\vartheta)$, can be considered as the so called hardest one dimensional submodel for estimating $b\vartheta(P)$ (if $\widetilde{I}(P, b\vartheta) \notin S(P)$, then we might still think of it as an approximate submodel). Therefore $\widetilde{I}(P, b\vartheta)$ is sometimes called the efficient score. The variance of $\widetilde{I}(P, b\vartheta)$ is the generalized Cramér-Rao lower bound for unbiased estimators, as can be proved as follows (for the technical details see van der Vaart, 1988):

**Generalized Cramér-Rao lower bound.** For a random variable, say $Y$, we denote its expectation with $E(Y)$ and it variance with $\mathrm{Var}(Y)$: we often use an index which describes the distribution of $Y$. To begin with, let $b\theta_1$ be an estimator of $b\theta$ based on one observation $X \sim P$ and suppose that $E_P(b\theta_1(X)) = b\theta$ for all $P \in \mathcal{M}$. Definition 1.6 tells us that we have $b\dot{\vartheta}(g) = \langle g, \widetilde{I}(P, b\vartheta)\rangle_P$. By the definition of pathwise differentiability and the fact that $b\vartheta(P_{\epsilon,g}) = E_{P_{\epsilon,g}}(b\theta_1(X))$ we also have:

$$b\dot{\vartheta}(g) = \lim_{\epsilon \to 0} 1/\epsilon \int b\theta_1(x) d(P_{\epsilon,g} - P)(x)$$

$$= \lim_{\epsilon \to 0} 1/\epsilon \int b\theta_1(x) \left(p_{\epsilon,g}(x) - p(x)\right) \mu(dx)$$

$$= \lim_{\epsilon \to 0} 1/\epsilon \int b\theta_1(x) \left(\frac{p_{\epsilon,g}(x) - p(x)}{p(x)}\right) dP(x). \tag{1.19}$$

By using (1.16) we have $(p_{\epsilon,g}(x) - p(x))/\epsilon p(x) \approx g(x)$, where the approximation can be made rigorous. Using this approximation tells us that (1.19) converges to $E_P(b\theta_1(X)g(X))$. Our calculations do not depend on $g \in S(P)$ and by linearity and continuity of $\dot{\vartheta}(g)$ and $g \to E_P(b\theta_1(X)g(X))$ we have the identity for all $g \in T(P)$. In other words, for all $g \in T(P)$ we have:

$$\langle g, \widetilde{I}(P, b\vartheta)\rangle_P = E_P\left(b\theta_1(X)g(X)\right).$$

By the Cauchy-Schwarz inequality we have that

$$\left(E_P(b\theta_1(X)g(X))\right)^2 \leq \mathrm{Var}\left(b\theta_1(X)\right)\mathrm{Var}(g(X)).$$

This tells us that

$$\mathrm{Var}(b\theta_1(X)) \geq \frac{\langle g, \widetilde{I}(P, b\vartheta)\rangle_P^2}{\langle g, g\rangle_P}.$$

This holds for all $g \in T(P)$ and therefore in particular for $g = \widetilde{I}(P, b\vartheta)$.

This provides us with the Cramér-Rao lower bound:

$$\mathrm{Var}_P\left(b\theta_1(X)\right) \geq \mathrm{Var}_P\left(\widetilde{I}(P, b\vartheta)(X)\right).$$

Suppose now that we have $n$ i.i.d. observations $X_i \sim P$ and let $b\theta_n(X_1, \ldots, X_n)$ be an unbiased estimator of $b\theta$ for all $P \in \mathcal{M}$. Then the same calculations show that:

$$\mathrm{Var}_P\left(b\theta_n(X_1, \ldots, X_n)\right) \geq \frac{\mathrm{Var}_P\left(\widetilde{I}(P, b\vartheta)(X)\right)}{n}.$$

The variance of $\widetilde{I}(P, b\vartheta)$ is also the optimal *asymptotic* variance of $\sqrt{n}(\theta_n - \vartheta(P))$ for so called regular estimators.

**Definition 1.7** *Let $b\theta_n$ be an estimator of $b\theta = b\vartheta(P)$ for which*

$$L_P\left(\sqrt{n}\left(b\theta_n - b\vartheta(P)\right)\right) \to L_b.$$

*$b\theta_n$ is a $S(P)$-regular estimator of $b\theta$ if for all $g \in S(P)$ there exists a $P_{\epsilon,g} \in \mathcal{S}(P)$ so that for $\epsilon_n = 1/\sqrt{n}$*

$$L_{P_{\epsilon_n,g}}\left(\sqrt{n}\left(b\theta_n - b\vartheta(P_{\epsilon_n,g})\right)\right) \to L_b.$$

Notice that the smaller we choose $S(P)$ the larger the class of regular estimators (relative to $S(P)$) and the easier it is to verify pathwise differentiability. On the other hand the lower bound $\sigma^2(\widetilde{I}(P, b\vartheta))$, where $\sigma^2$ denotes variance, represents a supremum over all Cramér-Rao lower bounds for the one dimensional submodels $p_{\epsilon,g}$ and therefore this lower bound can only be attained if $S(P)$ is large enough. Therefore, in order to have *existence* of efficient estimators one has to choose a *rich enough* class $\mathcal{S}(P)$ of one dimensional submodels $p_{\epsilon,g}$.

Most interesting estimators are asymptotically linear:

**Definition 1.8** *An estimator $\theta_n$ of $\theta = \vartheta(P)$ is called $\|\cdot\|_B$-asymptotically linear with influence curve $I(P, b\vartheta) \in L_0^2(P)$, $b \in B$, if*

$$\sqrt{n}(b\theta_n - b\theta) = \sqrt{n}(P_n - P)I(P, b\vartheta) + R_{n,b},$$

*where*

$$\|R_n\|_B \stackrel{def}{=} \sup_{b \in B} \mid R_{n,b} \mid = o_P(1)$$

*and the empirical process $\int I(P, b\vartheta) d\sqrt{n}(P_n - P)$ indexed by $\{I(P, b\vartheta) : b \in B\}$ converges weakly.*

Theorem 2.12 in van der Vaart (1988) tells us that for any regular estimator $b\theta_n$ the limiting distribution $L_b$ has a variance which is larger than $\sigma^2(\widetilde{I}(P, b\vartheta))$ and that equality holds if and only if $b\theta_n$ is asymptotically linear with influence curve equal to $\widetilde{I}(P, b\vartheta)$. One may call this result an asymptotic Cramér-Rao bound. The result also explains the following name:

**Definition 1.9** *$\widetilde{I}(P, b\vartheta) \in T(P)$ is called the efficient influence curve w.r.t. $S(P)$ for estimating $b\theta(P)$ in $\mathcal{M}$.*

The *convolution theorem* tells us that if $S(P)$ is convex, then the limiting distribution $L_b$ of a regular estimator $b\theta_n$ equals the sum of $N(0, \sigma_P^2(\widetilde{I}(P, b\vartheta)))$ and another independent random variable. Moreover, under some extra regularity assumptions the variance of $\widetilde{I}(P, b\vartheta)$ is also a lower bound for minimax estimators (for both statements see e.g. van der Vaart, 1988). This justifies the following definition of efficiency of $\theta_n$:

**Definition 1.10** *Let $\theta, \theta_n \in D$ and $B$ be a collection of real valued linear functions on $D$. Assume that $\theta_n$ is $\|\cdot\|_B$-asymptotically linear with efficient influence curve $\widetilde{I}(P, b\vartheta)$, $b \in B$.*

*Then we say that $\theta_n$ is $\|\cdot\|_B$-efficient.*

The *functional delta-method* theorem 1.5 shows that weak convergence is preserved under differentiability as stated in (1.13). The following theorem tells us that efficiency is also preserved under this kind of differentiability (van der Vaart, 1991):

**Theorem 1.7** *Let $B$ and $B_1$ be a collection of real valued linear functions on vector spaces $D$ and $E$, respectively, so that $(D, \|\cdot\|_B)$ and $(E, \|\cdot\|_{B_1})$ are normed vector spaces. Let $\Phi : D_\phi \subset (D, \|\cdot\|_B) \to (E, \|\cdot\|_{B_1})$ be a functional, where both spaces are endowed with the the Borel sigma-algebra.*

*Suppose that $\theta_n \in D_\phi$ is an $\|\cdot\|_B$-efficient estimator of $\theta \in D_\phi$, and the differentiability condition (1.13) of theorem 1.5 holds for $\Phi(\theta_n)$.*

*Then $\Phi(\theta_n)$ is a $\|\cdot\|_{B_1}$-efficient estimator of $\Phi(\theta)$.*

# Part I
# Efficiency Theory and Applications for (Nonparametric) Maximum Likelihood Estimators

# Chapter 2

# Efficiency Theory for the (NP)MLE and an Identity for Linear Parameters in Convex Models

## 2.1 Introduction.

To begin with we give the general set up of the problem we will investigate. We refer to the efficiency section 1.4 for some notation and definitions. Let $\mathcal{M}$ be a set of probability measures $P$ on a measurable space $(\Omega, \mathcal{A})$. Let $\vartheta : \mathcal{M} \to D$, where $D$ is a vector space, be a $D$-valued parameter. Let $B$ be a collection of real valued linear mappings $b : D \to \mathbb{R}$.

Given an i.i.d. sample $X_1, \ldots, X_n$ from an unknown $P \in \mathcal{M}$ we want to give a $\|\cdot\|_B$-efficiency theory of maximum likelihood estimators of $\theta = \vartheta(P)$ and modifications thereof.

$\mathcal{M}$ is not necessarily dominated by one fixed measure $\mu$. Therefore it does not always make sense to define a maximum likelihood estimator as the maximizer over all $P \in \mathcal{M}$ of the likelihood $\prod_{i=1}^n (dP/d\mu)(X_i)$ for a fixed measure $\mu$. A natural generalization of this last definition to general $\mathcal{M}$, due to Kiefer and Wolfowitz (1956), is now given by: $\mathbb{P}_n \in \mathcal{M}$ is an MLE of $P$ if and only if for each measure $P_1 \in \mathcal{M}$ we have with $\mu \equiv P_1 + \mathbb{P}_n$:

$$\int \log\left(\frac{d\mathbb{P}_n}{d\mu}\right) dP_n \geq \int \log\left(\frac{dP_1}{d\mu}\right) dP_n. \tag{2.1}$$

In other words, $\mathbb{P}_n$ is the winner in each pairwise comparison. $\theta_n = \vartheta(\mathbb{P}_n)$ is

called a *maximum likelihood estimator* (MLE) of $\theta$.

If it exists, then for any measure $\mu_n$ with $\mathbb{P}_n \ll \mu_n$ we have:

$$\mathbb{P}_n = \arg \max_{P \in \mathcal{M}(\mu_n)} \int \log \left( \frac{dP}{d\mu_n} \right) dP_n. \tag{2.2}$$

The log likelihood cares only about the values $p(X_i) \equiv (dP/d\mu_n)(X_i)$, $i = 1, 2, \ldots, n$. Therefore even in semiparametric models it is often possible to identify $\mathbb{P}_n$ with a vector, where its dimension grows with $n$, and define it as a maximum over a compact euclidean set, which makes existence of MLE much easier to prove.

$\mathbb{P}_n$ is often not explicitly known as a function of $X_1, \ldots, X_n$, but once a dominating measure of $\mathbb{P}_n$, given the data, is explicitly known, then it is defined by (2.2) which can usually be computed with certain algorithms. Finding an MLE involves essentially two steps. Firstly, one determines, given the data, explicitly a dominating measure $\mu_n$ (if $\mathbb{P}_n$ is discrete, then it suffices to find its support) and then one maximizes the log likelihood over all possible densities in the model w.r.t. this known $\mu_n$.

One can also decide not to worry about finding an explicitly, given the data, known dominating measure $\mu_n$ of $\mathbb{P}_n$, but choose $\mu_n$ ourselves. $\mu_n$ is usually chosen so that $\mathcal{M}(\mu_n)$ grows with $n$ and more and more closely approximates all of $\mathcal{M}(P)$. In our applications we will do this. Then we can define a so called *"sieved"-maximum likelihood estimator* as a maximizer of the log likelihood over $\mathcal{M}(\mu_n)$, abusing the traditional definition of sieve-MLE. Also in this case we can define this MLE as in (2.2) for some measure $\mu_n$.

In this chapter we present a theory for proving efficiency of a MLE $\theta_n$. Our applications in the next chapter are described by semiparametric models for which MLE are usually called "nonparametric maximum likelihood estimators" (NPMLE). Therefore we will often denote $\theta_n$ with NPMLE, though the efficiency theory is also applicable to parametric models.

The organization of this chapter is as follows. Firstly, in the next section we show that a maximum likelihood estimator often solves an *efficient score equation*. In section 3 we derive natural conditions for efficiency of an NPMLE and show how they trivially generalize to bootstrapped NPMLE and NPMLE based on a transformation, depending on $n$, of the original data (in the latter case everything, also the model, changes with $n$). (The latter kind of NPMLE will be analyzed in chapter 4.) Surprisingly enough we get one of the main conditions for free if the model is *convex* and the parameter $\vartheta : \mathcal{M} \to D$ *linear*. This follows from an identity. In section 4 we prove this identity and explain the gain from this identity for proving efficiency of NPMLE. In section 5 we

apply this identity to convex models which are linear in the parameter, study the required invertibility of the information operator and illustrate the use of the identity in a well known (non trivial) example. In section 6 we formulate the efficiency theorem for the NPMLE in convex models and in section 6.1 we extend this theorem to one-step estimators. Section 6.1 is not important for following the subsequent chapters.

## 2.2 Efficient score equation for NPMLE.

Assume that an MLE $\mathbb{P}_n$ as in (2.2) exists.

Let $\mathcal{S}(\mathbb{P}_n)$ be a class of one dimensional differentiable submodels of $\mathcal{M}$ through $\mathbb{P}_n$ (see definition 1.3). Let $S(\mathbb{P}_n) \subset L_0^2(\mathbb{P}_n)$ be the tangent cone corresponding to *this class* of submodels (see definition 1.4). Recall that a score $g_n \in S(\mathbb{P}_n)$ of a $\mathbb{P}_{n,\epsilon,g_n}$ does by definition depend on $\mathbb{P}_n$. Let $T(\mathbb{P}_n)$ be the tangent space at $\mathbb{P}_n$ (see definition 1.5).

Suppose that $b\vartheta$ is pathwise differentiable relative to $S(\mathbb{P}_n)$ at $\mathbb{P}_n$ with efficient influence curve $\widetilde{I}(\mathbb{P}_n, b\vartheta_n) \in T(\mathbb{P}_n)$ (see definition 1.6). $\mathbb{P}_n$ maximizes in particular the log likelihood of each one dimensional submodel $\mathbb{P}_{n,\epsilon,g_n}$, $g_n \in S(\mathbb{P}_n)$, which is dominated by a certain $\nu_n$. Consequently, if $\mathbb{P}_n$ lies in the interior of each one dimensional submodel of $\mathcal{S}(\mathbb{P}_n)$, then the following equation to hold (let $\nu_n$ be a dominating measure of $\mathbb{P}_{n,\epsilon,g_n}$):

$$0 = \frac{d}{d\epsilon} \int \log \left( \frac{d\mathbb{P}_{n,\epsilon,g_n}}{d\nu_n}(x) \right) dP_n(x) \Big|_{\epsilon=0}. \tag{2.3}$$

Suppose that $g_n$ is also pointwise defined instead of only in $L^2(\mathbb{P}_n)$ sense. Then by exchanging differentiation and integration (2.3) translates into

$$0 = \int g_n(x) dP_n(x). \tag{2.4}$$

This holds for all $g_n \in S(\mathbb{P}_n)$ and by linearity of $g \to \int g dP_n$ also for $\mathrm{Lin}(S(\mathbb{P}_n))$, its linear extension. Consequently, if $\widetilde{I}(\mathbb{P}_n, b\vartheta) \in \mathrm{Lin}(S(\mathbb{P}_n))$, then we have:

$$0 = P_n \widetilde{I}(\mathbb{P}_n, b\vartheta) = \int \widetilde{I}(\mathbb{P}_n, b\vartheta)(x) dP_n(x), \tag{2.5}$$

which we will call the *efficient score equation* or MLE-equation.

If $\widetilde{I}(\mathbb{P}_n, b\vartheta) \in T(\mathbb{P}_n) \backslash S(\mathbb{P}_n)$, then it might still be possible to prove (2.5) by a continuity argument. For our theory we only need that

$$P_n \widetilde{I}(\mathbb{P}_n, b\vartheta) = o_P\left(\frac{1}{\sqrt{n}}\right). \tag{2.6}$$

There exist examples with $\widetilde{I}(\mathbb{P}_n, b\vartheta) \in T(\mathbb{P}_n)\backslash S(\mathbb{P}_n)$ where (2.5) does not hold, while (2.6) is true (see Groeneboom and Wellner, 1992, interval censoring case I, p. 115: if we replace in the efficient score identity $g_n$ by $g$, then we obtain (2.5) and they show that the difference of the two score identities is $o_P(1/\sqrt{n})$).

Since $T(P) \subset L_0^2(P)$ we always have the equations

$$P\widetilde{I}(P, b\vartheta) = 0 \text{ and } \mathbb{P}_n\widetilde{I}(\mathbb{P}_n, b\vartheta) = 0. \tag{2.7}$$

## 2.3 Efficiency theorem for NPMLE.

$\theta_n$ is $\|\cdot\|_B$-efficient (see definition 1.10) if and only if

$$b\theta_n - b\theta = \int \widetilde{I}(P, b\vartheta)d(P_n - P) + R_{n,b}, \tag{2.8}$$

where $\|R_n\|_B = o_P(1/\sqrt{n})$ and $\{\widetilde{I}(P, b\vartheta) : b \in B\}$ is $P$-Donsker. Notice that $\|R_n\|_B = o_P(1/\sqrt{n})$ can be replaced by $\|R_n\|_B = o_P(\|\theta_n - \theta\|_B)$; then (2.8) implies

$$\|\theta_n - \theta\|_B = O_P(1/\sqrt{n}) + o_P(\|\theta_n - \theta\|_B),$$

which implies in its turn trivially that $\|\theta_n - \theta\|_B = O_P(1/\sqrt{n})$.

Assuming (2.6) this tells us that (2.8) holds if (and only if)

$$\sup_{b\in B}\left|b\theta_n - b\theta + \int \widetilde{I}(\mathbb{P}_n, b\vartheta)dP - \int \left(\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta)\right) d(P_n - P)\right| \tag{2.9}$$

is $o_P(\|\theta_n - \theta\|_B)$. Assume that there exists a $P$-Donsker class $\mathcal{F}$ so that $\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta) \in \mathcal{F}$ for all $b \in B$ with probability tending to 1. By the $\rho_P$-uniform continuity of the sample paths of the empirical process indexed by $\mathcal{F}$ (see (1.6)) it follows that if

$$\sup_{b\in B} \rho_P\left(\widetilde{I}(P, b\vartheta), \widetilde{I}(\mathbb{P}_n, b\vartheta)\right) \to 0 \text{ in probability,}$$

then $\sup_{b\in B}\left|\int \left(\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta)\right) d(P_n - P)\right| = o_P(1/\sqrt{n})$. Therefore, once we have this it suffices for proving (2.8) to prove that the sum of the other terms appearing on the left-hand side of (2.9) is $o_P(\|\theta_n - \theta\|_B)$. This proves the following theorem:

**Theorem 2.1** *Let $X \sim P \in \mathcal{M}$ for a model $\mathcal{M}$ and let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of $X$. Let $\theta = \vartheta(P) \in D$, $D$ a vector space, and $B$ be a certain collection of real valued linear mappings on $D$. Suppose that for each $P \in \mathcal{M}$, $b\vartheta$, $b \in B$, is pathwise differentiable at $P$ relative to $S(P)$ with efficient influence function $\widetilde{I}(P, b\vartheta)$.*

*Let $\theta_n \equiv \vartheta(\mathbb{P}_n)$, $\mathbb{P}_n \in \mathcal{M}$ be an estimator of $\theta$ which satisfies the following conditions:*

**Efficient Score Equation.**

$$\sup_{b\in B}\left|\int \widetilde{I}(\mathbb{P}_n, b\vartheta)dP_n\right| = o_P\left(\frac{1}{\sqrt{n}}\right).$$

**Differentiability condition.**

$$\sup_{b\in B}\left|b\vartheta(P) - b\vartheta(\mathbb{P}_n) - \int \widetilde{I}(\mathbb{P}_n, b\vartheta)d(P-\mathbb{P}_n)\right| = o_P(\|\theta_n-\theta\|_B). \quad (2.10)$$

**Empirical process condition.**

$$\sup_{b\in B}\left|\int \left(\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta)\right) d(P_n - P)\right| = o_P(1/\sqrt{n}).$$

*Sufficient conditions for the empirical process condition are:*

**$P$-Donsker class condition.** *There exists a $P$-Donsker class $\mathcal{F}$ so that $\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta) \in \mathcal{F}$ for all $b \in B$ with probability tending to 1.*

**$\rho_P$-consistency.**

$$\sup_{b\in B} \rho_P\left(\widetilde{I}(P, b\vartheta), \widetilde{I}(\mathbb{P}_n, b\vartheta)\right) \to 0 \text{ in probability.}$$

*Then $\theta_n$ is a $\|\cdot\|_B$-asymptotically efficient estimator of $\theta$.*

It is clear that $\widetilde{I}(P, b\vartheta)$ has to be defined in a stronger sense than as an element of $L_0^2(P)$. For example, for verifying the Donsker class condition for a univariate class of functions one might want to bound the variation of $\widetilde{I}(P, b\vartheta)$, which requires that $\widetilde{I}(P, b\vartheta)(x)$ is pointwise well defined for all $x$. This can often be straightforwardly accomplished, as we will see in the next chapters.

If $B = \{b\}$ for a single $b$, then this theorem provides us with efficiency of the real valued parameter $b\vartheta(P) \in \mathbb{R}$. By verifying the conditions uniformly over a larger collection $B$ one obtains $\|\cdot\|_B$-efficiency. It might be clear that for verifying the $\rho_P$-consistency condition $\|\cdot\|_B$-consistency of $\vartheta(\mathbb{P}_n)$ will typically be required.

At first sight it is not clear that $\int \widetilde{I}(\mathbb{P}_n, b\vartheta)d(P-\mathbb{P}_n)$ should be the Gateaux derivative of $\vartheta$ in the direction $P - \mathbb{P}_n$ at $\mathbb{P}_n$; in other words it is not trivial to see that the differentiability condition (2.10) is indeed a differentiability condition. Let $\nu$ be a dominating measure of $\mathbb{P}_n$ and suppose that $P \ll \mathbb{P}_n$. Then we can rewrite the differentiability condition (2.10) as:

$$\sup_{b\in B}\left|b\vartheta(P) - b\vartheta(\mathbb{P}_n) - \int \widetilde{I}(\mathbb{P}_n, b\vartheta)\frac{p-p_n}{p_n}d\mathbb{P}_n\right| = o_P(\|\theta_n - \theta\|_B),$$

where the densities are w.r.t. $\nu$. $\int \widetilde{I}(\mathbb{P}_n, b\vartheta)(p_n - p)/p_n d\mathbb{P}_n$ is the pathwise derivative of $\vartheta$ at $\mathbb{P}_n$ evaluated at score $g = (p_n - p)/p_n$, where $g$ corresponds

with the score of the line $\epsilon P+(1-\epsilon)\mathbb{P}_n$ from $P$ to $\mathbb{P}_n$. Therefore the condition is strongly related to "uniform" (in $B$) pathwise differentiability of $\vartheta$ at a moving sequence $\mathbb{P}_n$. In this argument the condition $P \ll \mathbb{P}_n$ is only due to the fact that in the definition of pathwise differentiability we *linearize in scores* $g = \lim_{\epsilon\to 0}(p_{\epsilon,g} - p)/(\epsilon p)$ of the one dimensional submodels, but this condition would not be necessary if we use a differentiability definition, where we linearize in $\lim_{\epsilon\to 0}(p_\epsilon - p)/\epsilon$. So there is no reason to expect that this condition is essential and examples (see next sections) show that it is indeed not. We discuss this disadvantage of the definition of pathwise differentiability in the next section.

In this tract we are especially concerned with estimation of *linear parameters* in *convex models.* In this case the differentiability condition can be verified to hold with remainder zero, following from the pathwise differentiability of $\vartheta$ with remainder zero, and thereby (2.10) reduces to an identity (there is no $o_P(\|\theta_n - b\theta\|_B)$-term in the differentiability condition). This identity provides us under the $P$-Donsker class condition and efficient score equation with root-$n$-consistency of $\vartheta(\mathbb{P}_n)$ which can in its turn be used to prove the $\rho_P$-consistency condition. We will work this out in detail in the next section.

**Extension to NPMLE based on resampled data or reduced data.** Firstly, we discuss the two applications to which we want to extend theorem 2.1 and then we will show that both can be straightforwardly captured by one extension of theorem 2.1. The extension can also be used to investigate uniformity in $P$ and regularity.

The first application is the *semiparametric bootstrap.* Let $X_i \sim P$, $i = 1,\ldots,n$, be $n$ i.i.d. random variables. Suppose that $\vartheta(\mathbb{P}_n)$ is a $\|\cdot\|_B$-efficient estimator of $\vartheta(P)$. Let $P^{(n)} \in \mathcal{M}$ be an estimator of $P$ and let $X_i^* \sim P^{(n)}$, $i = 1,\ldots,n$, be $n$ i.i.d. random variables. For asymptotic validity of the semiparametric bootstrap we want to show that the limiting distribution of the normalized NPMLE $\sqrt{n}(\vartheta(\mathbb{P}_n^*) - \vartheta(P^{(n)})$, where $\mathbb{P}_n^*$ is a NPMLE based on the resampled data $X_i^*$, equals the limiting distribution of $\sqrt{n}(\vartheta(\mathbb{P}_n) - \vartheta(P))$, which is given by the optimal Gaussian process indexed by the efficient influence functions $\{\widetilde{I}(P, b\vartheta) : b \in B\}$.

In view of an important application (the second application to which we want to extend theorem 2.1) in chapter 4 we do not want to force $P^{(n)}$ to be a member of $\mathcal{M}$. In this application we have $X_i \sim P$ and $X_i^* = \phi_n(X_i) \sim P^{(n)} \equiv P\phi_n^{-1}$ where $\phi_n(X_i)$ is a slight *transformation* (*reduction*) of the original data $X_i$. By using the reduced data $X_i^*$ we arranged to work in an easier model

$\mathcal{M}_{1n} \equiv \{P\phi_n^{-1} : P \in \mathcal{M}\}$ for which all conditions of theorem 2.1 can be verified *uniformly in* $P^{(n)}$. The parameter of interest is given by $\vartheta_n : \mathcal{M}_{1n} \to D$ which is so that $\vartheta_n(P^{(n)}) = \vartheta(P)$; so we are still estimating the same $\theta = \vartheta(P)$. If $\phi_n$ converges to the identity one hopes that asymptotic efficiency is still attained. Here we are concerned with the question if the NPMLE based on this transformed (reduced data) is asymptotically efficient; in other words we want that $\sqrt{n}(\vartheta_n(\mathbb{P}_n^*) - \vartheta_n(P^{(n)}))$ converges to the optimal Gaussian process.

The last problem is captured by the following set up, which also coveres the resampled data by setting $\mathcal{M}_n = \mathcal{M}$, $\vartheta_n = \vartheta$, $S_n(P) = S(P)$ and $\widetilde{I}_n(P_1, b\vartheta_n) = \widetilde{I}(P_1, b\vartheta)$:

Let $P^{(n)} \in \mathcal{M}_n$ be an approximation of $P$ and $\mathcal{M}_n$ be a sequence of models. Let $X_i \sim P$, $X_i^* \sim P^{(n)}$, $i = 1, \ldots, n$, be two collections of $n$ i.i.d. random variables. Suppose that $b\vartheta_n : \mathcal{M}_n \to D$ is pathwise differentiable at each $P_1 \in \mathcal{M}_n$ relative to a tangent cone $S_n(P_1)$ with efficient influence function $\widetilde{I}_n(P_1, b\vartheta_n)$. By verifying the conditions of theorem 2.1 *uniformly in* $P^{(n)}$ (so for the sufficient conditions for the empirical process condition this means that we need to verify a *uniform* $\{P^{(n)}\}$-Donsker class condition and a *uniform* in $P^{(n)}$ $\rho_{P^{(n)}}$-consistency condition), then we have that $\sqrt{n}\left(b\vartheta_n(\mathbb{P}_n^*) - b\vartheta_n(P^{(n)})\right)$ equals (here $P_n^*$ is the empirical distribution of $X_i^*$)

$$\int \widetilde{I}_n(P^{(n)}, b\vartheta_n) d\sqrt{n}(P_n^* - P^{(n)}) + R_{n,b}, \tag{2.11}$$

where $\sup_{b \in B} \mid R_{n,b} \mid = o_{P^{(n)}}(1)$. To see this recall that the uniform Donsker condition provides us with asymptotic $\rho_{P^{(n)}}$-equicontinuity uniformly in $\{P^{(n)}\}$ (see (1.8)). The proof of (2.11) is then a copy of the proof of theorem 2.1 applied to the model $\mathcal{M}_n$.

For proving asymptotic efficiency it remains to prove the following condition:

**Approximation condition.**
The empirical process $\left(\int \widetilde{I}_n(P^{(n)}, b\vartheta_n) d\sqrt{n}(P_n^* - P^{(n)}), b \in B\right)$ converges weakly under $P^{(n)}$ to (the optimal Gaussian process) a mean zero Gaussian process $X_0$ with covariance structure:

$$E(b_1 X_0 b_2 X_0) = E\left(\widetilde{I}(P, b_1\vartheta)(X)\widetilde{I}(P, b_2\vartheta)(X)\right),$$

i.e. the same limit process as when $P^{(n)} = P$. This can be proved by applying theorem 1.3. Theorem 2.1 can also be straightforwardly extended to the non-parametric bootstrap. We show how this works in the bootstrap section 6 of chapter 4.

## 2.4 An Identity for linear parameters in convex models.

Let $\mathcal{M}$ be a convex set of probability measures. Notice that this also implies that $\mathcal{M}(\mu)$ and $\mathcal{P}(\mu)$ are convex.

For each $P_1 \in \mathcal{M}(P)$ the line $\epsilon P_1 + (1-\epsilon)P$, $\epsilon \in [0,1]$, is a submodel of $\mathcal{M}$ through $P$. Let $\mu$ be a dominating measure of $P$. Then the corresponding straight line $\epsilon p_1 + (1-\epsilon)p$ of densities with respect to $\mu$ can be written as:

$$p_{\epsilon,g} = (1+\epsilon g)p, \; g = \frac{p_1 - p}{p}.$$

Consequently, its score is given by $(p_1 - p)/p$. If $g$ has finite supnorm, then $p_{\epsilon,g}$ is clearly Hellinger differentiable; it satisfies the differentiability property (1.16). Therefore for efficiency calculations a natural class of one dimensional submodels through $P$ is given by

$$\mathcal{S}(P) \equiv \left\{ \epsilon P_1 + (1-\epsilon)P, \; \epsilon \in [0,1] : P_1 \in \mathcal{M}(P), \| \frac{dP_1}{dP} \|_\infty < \infty \right\}. \quad (2.12)$$

In terms of densities this class (2.12) is given by:

$$\left\{ p_{\epsilon,g} = (1+\epsilon g)p : \|g = \frac{p_1 - p}{p}\|_\infty < \infty, P_1 \in \mathcal{M}(P) \right\} \subset L_0^2(P).$$

The corresponding tangent cone $S(P)$ and tangent space $T(P)$ are defined as in definitions 1.4 and 1.5.

**Definition 2.1** *A parameter $\vartheta : \mathcal{M} \to \Theta \subset D$ is linear if $\vartheta$ is well defined on $Lin(\mathcal{M})$ and $\vartheta : Lin(\mathcal{M}) \to D$ is a linear mapping.*

**Theorem 2.2** (Identity for linear parameters in convex models). *Suppose that $\mathcal{M}$ is convex and $\vartheta : \mathcal{M} \to D$ is linear. Suppose $P, P_1 \in \mathcal{M}$ and that $b\vartheta$ is pathwise differentiable at $P_1 \in \mathcal{M}$ relative to $S(P_1)$ with efficient influence curve $\widetilde{I}(P_1, b\vartheta)$.*

*Assume that* either *there exists a sequence $P \in \mathcal{M}(P_{1m})$ with $dP/dP_{1m} \in L^2(P_{1m})$ so that for $m \to \infty$*

$$\begin{aligned} \int \widetilde{I}(P_{1m}, b\vartheta) dP &\to \int \widetilde{I}(P_1, b\vartheta) dP. \\ b\vartheta(P_{1m}) &\to b\vartheta(P_1) \end{aligned} \quad (2.13)$$

or *there exists a sequence $P_m \in \mathcal{M}(P_1)$ with $dP_m/dP_1 \in L^2(P_1)$ so that for $m \to \infty$*

$$\begin{aligned} \int \widetilde{I}(P_1, b\vartheta) dP_m &\to \int \widetilde{I}(P_1, b\vartheta) dP. \\ b\vartheta(P_m) &\to b\vartheta(P) \end{aligned} \quad (2.14)$$

*Then*

$$b\vartheta(P) - b\vartheta(P_1) = \int \widetilde{I}(P_1, b\vartheta) d(P - P_1) = \int \widetilde{I}(P_1, b\vartheta) dP. \tag{2.15}$$

Notice that for $P_1 = \mathbb{P}_n$, the NPMLE, (2.15) implies the differentiability condition of theorem 2.1. The proof of theorem 2.2 (below) shows that it is easy to show the identity (2.15) for $P$, $P_1 \in \mathcal{M}$ with $P \ll P_1$. Once we have established this the identity conditions (2.13) (we mean both convergence statements) and (2.14) are just continuous extension conditions which trivially extend the identity to many more $P$, $P_1$. Notice that the sequences $P_m$ and $P_{1m}$ are allowed to depend on $b$ and that the identity conditions are not requiring more than convergence of two real numbers and hence they are very weak conditions. A natural candidate for $P_{1m}$ is $(1 - \epsilon_m)P_1 + \epsilon_m P$ for a sequence $\epsilon_m \to 0$.

**Proof of theorem 2.2.** Suppose that we can prove the identity (2.15) for any $P'$ and $P_1'$ with $P' \ll P_1'$ and $dP'/dP_1' \in L^2(P_1')$. Then the identity (2.15) holds in particular for $P, P_{1m}$ with $dP/dP_{1m} \in L^2(P_{1m})$ and $P_m, P_1$ with $dP_m/dP_1 \in L^2(P_1)$. If now the convergence for $m \to \infty$ holds as stated in (2.13) and (2.14), then it follows that the identity holds also for $P$ and $P_1$. Therefore, it suffices to prove the identity (2.15) for any $P$ and $P_1$ with $P \ll P_1$ and $dP/dP_1 \in L^2(P_1)$:

Let $\mu$ be a dominating measure of $P_1$ and denote the densities of $P$ and $P_1$ with respect to $\mu$ by $p$ and $p_1$, respectively. Define $p_{\epsilon,g} \equiv (1 + \epsilon g)p_1$ for $g = (p - p_1)/p_1$, which corresponds with the straight line $\epsilon P + (1 - \epsilon)P_1$. $p_{\epsilon,g}$ is a differentiable submodel through $P_1$ with score $g \in S(P_1)$ and it is linear in $g$. By linearity of $\vartheta$ we have:

$$\begin{aligned} \frac{1}{\epsilon}(b\vartheta(P_{\epsilon,g}) - b\vartheta(P_1)) &= \frac{1}{\epsilon} b\vartheta(\epsilon(P - P_1)) \\ &= b\vartheta(P) - b\vartheta(P_1). \end{aligned} \tag{2.16}$$

We also have that the left hand side is linear in $g$. By the pathwise differentiability of $\vartheta$ at $P_1$ relative to $S(P_1)$ and the linearity in $g$ of the left hand side, the left hand side equals the pathwise derivative in the direction $g$, which is given by $\int \widetilde{I}(P_1, b\vartheta) g dP_1$. Combining this with (2.16) and recalling that $g = (p - p_1)/p_1$ gives us:

$$b\vartheta(P) - b\vartheta(P_1) = \int \widetilde{I}(P_1, b\vartheta) \frac{p - p_1}{p_1} dP_1 = \int \widetilde{I}(P_1, b\vartheta) dP. \square$$

### 2.4.1 Discussion about the (ir)relevance of the identity conditions.

In semiparametric models the NPMLE $\mathbb{P}_n$ does often not dominate the underlying $P$. If $P \not\ll \mathbb{P}_n$, then the direct proof of theorem 2.2 of the identity does not work because the line $\epsilon P + (1-\epsilon)\mathbb{P}_n$ does not even have a score (it is not Hellinger differentiable) and therefore cannot be linearized in a score; so it does also not make sense to talk about pathwise differentiability along this line. The approach followed by this theorem is to prove the identity for Hellinger differentiable lines and approximate the non-differentiable lines by Hellinger differentiable lines in order to obtain also a prove of the identity for non-differentiable lines.

In order to carry out a direct proof we should have a notion of differentiability which also applies to lines from $P$ to $\mathbb{P}_n$, i.e. which also applies to non-Hellinger-differentiable submodels. In example 2.2 in the next section we show that if we give up linearizing in $(p-p_1)/p_1$ as in the proof of the theorem, i.e. we give up using the definition of pathwise differentiability, but instead only concentrate on linearizing in $p-p_1$ (densities w.r.t. e.g. $\mu = P+P_1$), then there is a direct proof, still using Hilbertspace structures, of an identity of similar nature. The use of this approach is not yet clear for us, but it established a kind of differentiability of $\vartheta$ along a line from any $P_1$ to $P$ and it can be easily generalized. We did not use this approach for our applications in the next chapters, because here the identity conditions (2.14) could be straightforwardly proved. But this alternative approach of proving a similar identity clarifies the irrelevance of the condition $P \ll P_1$ and hence that with a different set up one should be able to formulate a similar theorem, without requiring the identity conditions, but requiring a different kind of differentiability.

In the following example the identity can be explicitly written down and it also shows how weak the identity conditions are.

**Example 2.1 (Nonparametric Model).** Let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of $X \sim P$, where $P$ is a completely unknown distribution. We want to estimate $b\vartheta(P) = P(B)$ for a measurable set $B$. Let $\mathbb{P}_n$ be the NPMLE. In this nonparametric model we have $\widetilde{I}(\mathbb{P}_n, B)(X) = I(X \in B) - \mathbb{P}_n(B)$. Therefore the efficient score equation for $\mathbb{P}_n$ tells us that $\mathbb{P}_n$ equals the empirical distribution $P_n$. The identity of theorem 2.2 is trivially true:

$$(\mathbb{P}_n - P)(B) = -P\widetilde{I}(\mathbb{P}_n, B).$$

The condition (2.14) for $P_1 = \mathbb{P}_n$ is satisfied if there exists a $\widetilde{P}_n^m$ with $\widetilde{P}_n^m \ll P_n$

so that $\widetilde{P}_n^m(B) \to P_n(B)$, which holds of course trivially. The latter condition provides us by application of theorem 2.2 with an indirect proof of the identity.

In spite of the fact that the identity conditions (2.13) and (2.14) are very weak, implicitness of the efficient influence curve makes them not always easy to verify. In our applications in chapter 3,4 and 5 we verify the identity condition by proving the following much too strong condition, but which is still rather straightforwardly verifiable; there exists a sequence $P_{1m}$ with $dP/dP_{1m}$ finite, $b\vartheta(P_{1m}) \to b\vartheta(P_1)$, $\|P_{1m} - P_1\|_P \to 0$ (or we take another $L^2$ norm) and (using this) show that

$$\int \left|\widetilde{I}(P_{1m}, b\vartheta) - \widetilde{I}(P_1, b\vartheta)\right| dP \to 0.$$

From now on if we write condition (2.14) we mean just one of the two conditions (2.13) and (2.14).

### 2.4.2 The gain from the identity.

Let $\mathbb{P}_n \in \mathcal{M}$ be a NPMLE of $P \in \mathcal{M}$ and suppose that (2.14) with $P_1 = \mathbb{P}_n$ and $P$ holds. Then (2.15) applied to $P_1 = \mathbb{P}_n$ provides us with:

$$b\vartheta(\mathbb{P}_n) - b\vartheta(P) = -\int \widetilde{I}(\mathbb{P}_n, b\vartheta) dP, \tag{2.17}$$

which is the differentiability condition of theorem 2.1, but with remainder zero.

Combining this equation with the efficient score equation (2.5)

$$P_n \widetilde{I}(\mathbb{P}_n, b\vartheta) = 0,$$

provides us with

$$b\vartheta(\mathbb{P}_n) - b\vartheta(P) = \int \widetilde{I}(\mathbb{P}_n, b\vartheta) d(P_n - P). \tag{2.18}$$

This is a very powerful identity because the right-hand side can be considered as an empirical process indexed by $\widetilde{I}(\mathbb{P}_n, b\vartheta)$. Therefore if the Donker class condition of theorem 2.1 holds, then it provides us already with root-$n$ consistency of $b\vartheta(\mathbb{P}_n)$. Now, the $\rho_P$-consistency condition remains to be verified, where we can use this consistency result. In other words, the convexity of the model and linearity of the parameter often gives us the differentiability condition of theorem 2.1 and consistency for free and we can concentrate our attention on the Donsker class condition.

In our examples we apply theorem 2.2 to models of a special type, discussed in the next section.

## 2.5 Application of the identity to convex models which are linear in the parameter.

Consider a model $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset D$ is a convex set and $\theta \rightarrow P_\theta$ is linear. Then $\mathcal{M}$ is convex. For clarity and because it holds for all our applications we will assume that $\Theta$ is a convex set of *probability measures* on a certain fixed measurable space (in our case $\Theta$ is a convex set of distribution functions of probability measures on $\mathbb{R}^k$), but this is not essential.

Consider a parameter $\vartheta(P_\theta) = \Psi(\theta)$, where $\Psi$ is linear (see definition 2.1). Define $\vartheta(P_\theta) = \Psi(\theta)$ for all $\theta \in \mathrm{Lin}(\Theta)$, which we can do because $\Psi$ is linear on $\mathrm{Lin}(\Theta)$. Then

$$\begin{aligned}\vartheta(\alpha P_{\theta_1} + \beta P_\theta) &= \vartheta(P_{\alpha\theta_1+\beta\theta}) \\ &= \Psi(\alpha\theta_1 + \beta\theta) \\ &= \alpha\Psi(\theta_1) + \beta\Psi(\theta) \\ &= \alpha\vartheta(P_{\theta_1}) + \beta\vartheta(P_\theta),\end{aligned}$$

where we used linearity of $\theta \rightarrow P_\theta$, extended definition of $\vartheta$, linearity of $\Psi$ and the definition of $\vartheta$, respectively. This proves that $\vartheta$ is linear. Therefore if we can establish pathwise differentiability of $\vartheta$, then application of theorem 2.2 provides us with the identity (2.23) below.

Let $\theta_{\epsilon,g}$ be a line from $\theta_1$ to $\theta$, $\theta_1 \ll \theta$ with score $g = (d\theta_1 - d\theta)/d\theta \in L_0^2(\theta)$. By linearity of $\theta \rightarrow P_\theta$ this line implies a submodel $P_{\theta_{\epsilon,g}}$ with score $(dP_{\theta_1} - dP_\theta)/dP_\theta \in L_0^2(P_\theta)$, assuming that $dP_{\theta_1}/dP_\theta$ exists and that it square integrable. $(dP_{\theta_1} - dP_\theta)/dP_\theta$ is linear in the underlying score $g$:

$$\frac{dP_{\theta_1} - dP_\theta}{dP_\theta} = \frac{dP_{\int g d\theta}}{dP_\theta} \equiv A_\theta(g). \tag{2.19}$$

Define now $\mathcal{S}(\theta)$ as all lines $\theta_{\epsilon,g} = \epsilon\theta_1 + (1-\epsilon)\theta$ for which $(dP_{\theta_1} - dP_\theta)/dP_\theta \in L_0^2(P_\theta)$. In all our applications $\mathcal{S}(\theta)$ is just the same set of lines as defined in (2.12), but now in the space $\Theta$ instead of in $\mathcal{M}$. Let $S(\theta) \subset L_0^2(\theta)$ be the corresponding tangent cone and $T(\theta) \subset L_0^2(\theta)$ be the tangent space.

Now, $A_\theta : S(\theta) \rightarrow S(P_\theta) \subset L_0^2(P_\theta)$, where $S(P_\theta)$ is the tangent cone corresponding with $P_{\theta_{\epsilon,g}}$, $g \in S(\theta)$. Suppose that $A_\theta$ can be continuously extended to $T(\theta) \subset L_0^2(\theta)$. Then the so called *score operator* $A_\theta$ can be defined as a linear Hilbertspace operator:

$$A_\theta : T(\theta) \subset L_0^2(\theta) \rightarrow L_0^2(P_\theta). \tag{2.20}$$

Now, the tangent space $T(P_\theta)$ is given by

$$T(P_\theta) = \overline{A_\theta(T(\theta))} \subset L_0^2(P_\theta).$$

We will now give conditions for pathwise differentiability of $\vartheta$ relative to $S(P_\theta)$.

Suppose that $b\Psi$ is pathwise differentiable at $\theta$ relative to $S(\theta)$ with efficient influence function $\widetilde{\kappa}(\theta, b\Psi)$. Then we have with the line $\theta_{\epsilon,g} = \epsilon\theta_1 + (1-\epsilon)\theta \in \mathcal{S}(\theta)$ with score $g = (d\theta_1 - d\theta)/d\theta$:

$$\begin{aligned} \frac{1}{\epsilon}\left(b\vartheta(P_{\theta_{\epsilon,g}}) - b\vartheta(P_\theta)\right) &= \frac{1}{\epsilon}\left(b\Psi(\theta_{\epsilon,g}) - b\Psi(\theta)\right) \\ &= \int \widetilde{\kappa}(\theta, b\Psi) g d\theta. \end{aligned} \tag{2.21}$$

This is a linear mapping in $g$. For pathwise differentiability of $\vartheta(P_\theta)$ relative to $S(P_\theta)$ we need to rewrite this as a continuous linear mapping in the scores $A_\theta(g)$ of $P_{\theta_{\epsilon,g}}$.

For this purpose let $A_\theta^\top : L_0^2(P_\theta) \to T(\theta)$ be the *adjoint* of $A_\theta$: for all $g \in T(\theta)$ and $v \in L_0^2(P_\theta)$ we have

$$\langle A_\theta(g), v\rangle_{P_\theta} = \langle g, A_\theta^\top(v)\rangle_\theta. \tag{2.22}$$

We have the following important result:

**Lemma 2.1** *Assume that $b\Psi$ is pathwise differentiable at $\theta$ relative to $S(\theta)$. $b\vartheta$ is pathwise differentiable at $P_\theta$ relative to $S(P_\theta)$ if and only if $\widetilde{\kappa}(\theta, b\Psi)$ lies in the range of $A_\theta^\top : L_0^2(P_\theta) \to T(\theta)$.*

This lemma is due to van der Vaart (1991). The proof is straightforward: Suppose that we have pathwise differentiability with efficient influence function $\widetilde{I}(P_\theta, b\Psi)$. Then by (2.21) and the definition of pathwise differentiability we have for all $g \in S(\theta)$:

$$\int \widetilde{\kappa}(\theta, b\Psi) g d\theta = \int \widetilde{I}(P_\theta, b\Psi) A_\theta(g) dP_\theta.$$

Using the definition of adjoint we see that

$$\int \left(A_\theta^\top(\widetilde{I}(P_\theta, b\Psi)) - \widetilde{\kappa}(\theta, b\Psi)\right) g d\theta = 0 \text{ for all } g \in T(\theta),$$

which proves one direction of the lemma.

For the other direction, we have to express (2.21) as a linear mapping in $A_\theta(g) \in S(P_\theta)$. Suppose that there exists a $l \in L_0^2(P_\theta)$ so that $A_\theta^\top(l) = \widetilde{\kappa}(\theta, b\Psi)$. Then by definition of the adjoint

$$\int \widetilde{\kappa}(\theta, b\Psi) g d\theta = \int l A_\theta(g) dP_\theta,$$

which, by the Cauchy-Schwarz inequality, is a continuous linear operator on $S(P)$. This proves the pathwise differentiability with efficient influence function $\widetilde{I}(P_\theta, b\vartheta) = \Pi(l \mid T(P_\theta)$ ($\Pi$ denotes $L^2$-projection) and hence the following theorem:

**Theorem 2.3** (Identity in convex linear models). *Consider a model* $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, *where* $\Theta \subset D$ *is a convex set of probability measures on a measurable space and* $\theta \to P_\theta$ *is linear. Let* $\vartheta(P_\theta) = \Psi(\theta)$, *where* $\Psi$ *is linear. If*

- $b\Psi$ *is pathwise differentiable at* $\theta_1$ *relative to* $S(\theta_1)$ *with efficient influence function* $\widetilde{\kappa}(\theta_1, b\Psi)$.
- *There exists an* $l \in L_0^2(P_\theta)$ *so that* $A_{\theta_1}^\top(l) = \widetilde{\kappa}(\theta_1, b\Psi)$,

*then* $b\vartheta(P_{\theta_1})$ *is pathwise differentiable with efficient influence function* $\widetilde{I}(P_{\theta_1}, b\Psi) = \Pi(l \mid T(P_\theta))$ *and for all* $P_\theta$ *and* $P_{\theta_1}$ *which satisfy (2.14) we have the identity:*

$$b\Psi(\theta) - b\Psi(\theta_1) = \int \widetilde{I}(P_{\theta_1}, b\Psi) d(P_\theta - P_{\theta_1}) = \int \widetilde{I}(P_{\theta_1}, b\Psi) dP_\theta. \tag{2.23}$$

A simple corollary of lemma 2.1, which provides us with a formula for $\widetilde{I}(P_\theta, b\Psi)$, is given by the following corollary (here $I_\theta^-(f)$ is just any element $g$ for which $I_\theta(g) = f$):

**Corollary 2.1** *If* $\widetilde{\kappa}(\theta, b\Psi)$ *lies in the range of the so called* **information operator** $I_\theta \equiv A_\theta^\top A_\theta : T(\theta) \to T(\theta)$, *then* $\vartheta(P_\theta)$ *is pathwise differentiable at* $P_\theta$ *relative to* $S(P_\theta)$ *with efficient influence function*

$$\widetilde{I}(P_\theta, b\Psi) = A_\theta I_\theta^-(\widetilde{\kappa}(\theta, b\Psi)). \tag{2.24}$$

### 2.5.1 Invertibility of the information operator.

If the information operator is invertible and onto we can apply corollary 2.1 and theorem 2.3. Therefore the following invertibility lemma is useful:

**Lemma 2.2** *Let* $I_\theta = A_\theta^\top A_\theta : T(\theta) \subset (L_0^2(\theta), \|\cdot\|_\theta) \to T(\theta)$ *be the information operator as defined above. Assume that for all* $h \in T(\theta)$ *with* $\|h\|_\theta > 0$ *we have* $\|A_\theta(h)\|_{P_\theta} > 0$. *Then* $I_\theta$ *is 1-1.*

*Assume that there exists a* $\delta > 0$ *so that for all* $h \in T(\theta) \subset H(\theta)$ *we have* $\|\|h\|_{theta} \geq \|A_\theta(h)\|_{P_\theta} \geq \delta \|h\|_\theta$ *for some* $1 > \delta > 0$. *Then* $I_\theta$ *is onto and has bounded inverse with operator norm smaller than or equal to* $1/\delta^2$. *Its inverse is given by:*

$$I_\theta^{-1} = \sum_{i=0}^{\infty} (I - I_\theta)^i.$$

*Moreover, the range of* $(A_\theta(T(\theta))$ *is closed and therefore equals* $T(P_\theta)$.

**Proof.** Let $||h||_\theta = 1$, then we have by the Cauchy-Schwarz inequality:

$$\begin{aligned} ||A_\theta^\top A_\theta(h)||_\theta &= ||A_\theta^\top A_\theta(h)||_\theta ||h||_\theta \\ &\geq \langle A_\theta^\top A_\theta(h), h\rangle_\theta \\ &= \langle A_\theta(h), A_\theta(h)\rangle_{P_\theta}. \end{aligned}$$

If $||A_\theta(h)||_{P_\theta} > 0$ for all $||h||_\theta = 1$, then $I_\theta$ is 1–1 and hence invertible. We have that $I_\theta = I - (I - I_\theta)$. If $||A_\theta(h)||_{P_\theta} > \delta$, then we have that $||I_\theta(h)||_\theta > \delta^2$. Because $I - I_\theta$ is self-adjoint its norm is given by:

$$\sup_{||h||_\theta = 1} \langle h, (I - I_\theta)(h)\rangle_\theta.$$

Because $\langle h, I_\theta(h)\rangle = \langle A_\theta(h), A_\theta(h)\rangle \geq \delta^2$ it follows that this norm is smaller than $1 - \delta^2$. Consequently, the inverse of $I_\theta$ is given by the Neumann series of $I - I_\theta$ which converges for all $h \in T(\theta)$. This proves that $I_\theta$ is onto and has bounded inverse with operator norm bounded by $1/\delta^2$. The final statement is also straightforward to check by using Cauchy sequences and the completeness of a Hilbert space.□

By using this lemma it is often easy to find natural conditions for (bounded) invertibility of the information operator $I_\theta : (T(\theta), ||\cdot||_\theta) \to (T(\theta), ||\cdot||_\theta)$. In particular this is true for missing data models as the next example and the models we cover in chapter 3, 4 and 5.

### 2.5.2 Example.

We already gave a trivial example of the completely nonparametric model where the efficient influence function is known and thereby the identity (2.18) for the NPMLE could *also* be explicitly verified. A non-trivial (well known) example, where this identity had not yet been discovered and can be explicitly written down, is the following. Here we will also show efficiency of the Kaplan-Meier estimator using identity (2.18) and we will give a method of proving a kind of differentiability of a parameter along any line, instead of only along lines with a score.

**Example 2.2 (Univariate Censoring Model).** Let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of a real valued $X$ with distribution function $F$, where $F$ is completely unknown. Let $C_1, \ldots, C_n$ be $n$ i.i.d. copies of a real valued $C$ with distribution

function $G$, where $G$ is completely unknown. $X$ and $C$ are independent. We observe

$$Y_i = (Z_i, D_i) = \Phi(X_i, C_i) \equiv (X_i \wedge C_i, I(X_i \leq C_i)) \sim P_{F,G}.$$

We are interested in estimating the *survival function* $\vartheta(P_{F,G}) = S(t) \equiv 1-F(t)$. If $G$ was known, then the model corresponding with $Y$ would be a convex model which is linear in the parameter $F$. Therefore under the conditions of theorem 2.3 identity (2.23) holds for the efficient influence function for the model where $G$ is known. However, in this model it can be shown (as done in chapter 4 for the bivariate censoring model) that the efficient influence function for estimating $F(t)$ in the model where $G$ is unknown equals the efficient influence function for estimating $F(t)$ in the model where $G$ is known. This follows straightforwardly from the fact that the conditional density of $X$ given what we observe about $C$ (so $C = z$ for the censored ($D = 0$) observations and $C > z$ for the uncensored ($D = 1$) observations) equals the unconditional density of $X$. So the identity for $G$ is known equals the identity for the model where $G$ is unknown (i.e. the univariate censoring model). The identity conditions of theorem 2.3 can be verified by writing down the score operator, information operator, applying lemma 2.2, and using formula (2.24) for the efficient influence function, as we will do in chapter 3,4 and 5, which would give a proof of the identity without explicitly knowing the efficient influence function.

Below, we will explicitly verify it: Define

$$\begin{aligned} N_n(t) &\equiv \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq t, D_i = 1) \\ Y_n(t) &\equiv \frac{1}{n}\sum_{i=1}^{n} I(Z_i \geq t) \\ \Lambda(t) &\equiv \int_0^t \frac{dF(s)}{1-F(s-)}. \end{aligned}$$

It is well known that the NPMLE of $S(t)$ is given by the Kaplan-Meier estimator $S_n(t) = \prod_{(0,t]}(1 - dN_n/Y_n)$, where $\prod_{(0,t]}$ is a *product integral* and stands for a limit of approximating finite products over partitions of $(0, t]$ as the partitions become finer.

This estimator has been extensively analyzed. For an overview of work done in this field we refer to Andersen, Borgan, Gill and Keiding (1993). Let $H = 1 - G$, $N = E_{P_F}(N_n)$ and $Y = E_{P_F}(Y_n)$. It is well known (e.g. Wellner, 1982, Gill, 1993) that if $H(t) > 0$, then the efficient influence function for

estimating $S(t)$ is given by:

$$\widetilde{I}(S,t)(z,d) = -S(t)\int_0^t \frac{I(z \in dv, d=1) - I(z \geq v)d\Lambda(v)}{S(v)H(v-)}.$$

This formula equals $A_F I_F^{-1}(I_{(t,\infty)} - S(t))$, where $A_F$ is the score operator and $I_F$ the information operator (see corollary 2.1 and Wellner, 1982). Consequently, the efficient score-equation for the NPMLE $S_n$ is given by:

$$P_n\widetilde{I}(S_n,t) = S_n(t)\int_0^t \frac{dN_n(v) - Y_n d\Lambda_n(v)}{S_n(v)H(v-)}.$$

By using that $S_n(t) = \prod_{(0,t]}(1 - dN_n/Y_n)$, it follows that $d\Lambda_n = dN_n/Y_n$ and consequently $P_n\widetilde{I}(S_n,t) = 0$. This verifies the efficient score-equation. It remains to verify the identity (2.23), i.e. $S_n(t) - S(t) = -P_F\widetilde{I}(S_n,t)$, which is here given by:

$$S_n(t) - S(t) = S_n(t)\int_0^t \frac{dN(v) - Yd\Lambda_n(v)}{S_n(v)H(v-)}. \tag{2.25}$$

We know that $dN = Yd\Lambda$, $Y = S_-H_-$. So $dN - Yd\Lambda_n = S_-H_-(d\Lambda - d\Lambda_n)$, where $H_-$ cancels with the denominator. Therefore (2.25) is equivalent to:

$$\begin{aligned} S_n(t) - S(t) &= \int_0^t S(v-)d(\Lambda_n - \Lambda)(v)\frac{S_n(t)}{S_n(v)} \\ &= \int_0^t \prod_{(0,v)}(1 - d\Lambda(v))\,(\Lambda_n - \Lambda)\,(dv)\prod_{(v,t]}(1 - d\Lambda_n(v))\,, \end{aligned}$$

where we used that $S_n(t)/S_n(v) = \prod_{(v,t]}(1 - dN_n/Y_n)$. This is the well known *Duhamel equation* for the univariate product integral (Gill and Johansen, 1990). This proves the identity for the NPMLE in the univariate censoring model:

$$S_n(t) - S(t) = (P_n - P_F)\widetilde{I}(S_n,t). \tag{2.26}$$

Because we have already verified the identity we might as well finish the efficiency proof by verifying the $P$-Donsker class and $\rho$-consistency conditions of theorem 2.1.

Notice that $\widetilde{I}(S_n,t)(z,d)$ is a sum of two monotone functions and both parts are bounded by $c/H(t)$ for a constant $c$. The class of bounded monotone functions is a uniform-Donsker class (see example 1.1). Application of this to the right-hand side of (2.26) provides us with the following whole line result:

$$\sup_{t\in[0,\infty]} H(t)\,|S_n(t) - S(t)| = O_P(1/\sqrt{n}).$$

Fix $\tau$ so that $H(\tau) > 0$. It follows trivially that $\sup_{t\in[0,\tau]} \|\widetilde{I}(S_n, t, \cdot) - \widetilde{I}(S, t, \cdot)\|_{P_F} \to 0$ in probability. Application of theorem 2.1 provides us now with supremum norm efficiency of $S_n$ on $[0, \tau]$. Gill (1993) is able to obtain a few refined results for the Kaplan-Meier estimator by using identity (2.26).

**Alternative method of proving a similar identity as (2.23).** We end this example with an alternative method for proving an identity without using the explicit form of the efficient influence function and without any need to verify the identity condition (2.14). Let $F_{1\epsilon} = \epsilon F + (1-\epsilon)F_1$ be a line from $F$ to $F_1$. This line is dominated by $\mu \equiv F + F_1$. Let $f, f_1$ be the densities of $F, F_1$ w.r.t. $\mu$. We characterize $F_{1\epsilon}$ by the direction (an equivalent of the score)

$$h \equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon}(F_{1\epsilon} - F_1) = f_1 - f \in L_0^2(\mu),$$

assuming that $\int (f_1 - f)^2 d\mu < \infty$. Let $\mu_1 \equiv (\mu \times G)\Phi^{-1}$ be the measure induced by $\mu \times G$ on the sample space. Then $d\mu_1(z, 1) = H(z)d\mu(z)$ and $d\mu_1(z, 0) = \mu(z, \infty)dG(z)$. $P_{F_{1\epsilon,h}}$ is dominated by $\mu_1$ and we denote its line of densities with $p_{F_{1\epsilon,h}}$. $F_{1\epsilon,h}$ induces a direction for $P_{F_{1\epsilon,h}}$ which we denote with $B_{F_1}(h)$:

$$B_{F_1}(h) \equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon}(p_{F_{1\epsilon,h}} - p_F) = p_{F_1} - p_F \in L_0^2(\mu_1).$$

Consider $B_{F_1} : L_0^2(\mu) \to L_0^2(\mu_1)$ as an operator (an equivalent of the score operator). This operator is given by:

$$B_{F_1}(h)(z, d) = H(z)h(z)I(d = 1) + g(z)\int_z^\infty h(x)d\mu(x)I(d = 0).$$

Let $B_{F_1}^\top : L_0^2(\mu_1) \to L_0^2(\mu)$ be the adjoint of $B_{F_1}$. Then we can define an equivalent of the information operator by

$$J_{F_1} \equiv B_{F_1}^\top B_{F_1} : L_0^2(\mu) \to L_0^2(\mu).$$

Lemma 2.2 tells us that $J_{F_1}$ is onto and invertible if $\|B_{F_1}(h)\|_{\mu_1} \geq \delta\|h\|_\mu$ for certain $\delta > 0$. We have by only integrating over the complete observations and assuming that $H > \delta > 0$ (this can be arranged by artificially censoring all observations at $\tau$, where $H(\tau) > 0$, which does not influence the Kaplan-Meier estimator at $t < \tau$):

$$\begin{aligned}\int (B_{F_1}(h))^2 d\mu_1 &\geq \int (H(z)h(z))^2 H(z)d\mu(z) \\ &\geq \delta^3 \int h^2(z)d\mu(z).\end{aligned}$$

This proves that $J_{F_1}$ has bounded inverse and is onto. Let $c(t) \equiv \int_t^\infty d\mu$ and $\kappa_t(\cdot) \equiv I_{(t,\infty)}(\cdot) - c(t) \in L_0^2(\mu)$. Now, we are ready to prove "the identity":

$$\begin{aligned} F(t) - F_1(t) &= \int I_{(t,\infty)}(x)(f - f_1)(x)d\mu(x) \\ &= \langle \kappa_t, (f - f_1)\rangle_\mu \\ &= \langle J_{F_1} J_{F_1}^{-1}(\kappa_t), (f - f_1)\rangle_\mu \\ &= \langle B_{F_1} J_{F_1}^{-1}(\kappa_t), B_{F_1}(f - f_1)\rangle_{\mu_1} \\ &= \langle B_{F_1} J_{F_1}^{-1}(\kappa_t), p_F - p_{F_1}\rangle_{\mu_1} \\ &= \int B_{F_1} J_{F_1}^{-1}(\kappa_t) d(P_F - P_{F_1}). \end{aligned}$$

## 2.6 Efficiency theorem for NPMLE of linear parameters in convex models.

In the subsection "The gain from the identity" we proved that combining theorem 2.1 and theorem 2.2 provides us with a general efficiency result for the NPMLE of linear parameters in convex models. We summarize it here:

**Theorem 2.4** (Efficiency theorem for NPMLE of linear parameters in convex models). *Let $X \sim P \in \mathcal{M}$ for a convex model $\mathcal{M}$ and let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of $X$. Let $\theta = \vartheta(P) \in D$ be a linear parameter and $B$ be a certain collection of real valued linear mappings on $D$. Suppose that for each $P \in \mathcal{M}$, $b\vartheta$, $b \in B$, is pathwise differentiable at $P$ relative to $S(P)$ with efficient influence function $\widetilde{I}(P, b\vartheta)$.*

*Let $\theta_n = \vartheta(\mathbb{P}_n)$, $\mathbb{P}_n \in \mathcal{M}$, be an estimator of $\theta = \vartheta(P)$ for which (2.14) holds with $P$ and $P_1 = \mathbb{P}_n$ and for which the following conditions hold:*

**Efficient Score Equation.**

$$\sup_{b \in B} \left| \int \widetilde{I}(\mathbb{P}_n, b\vartheta) dP_n \right| = o_P(1/\sqrt{n}).$$

***P*-Donsker class condition.** *There exists a $P$-Donsker class $\mathcal{F}$ so that $\widetilde{I}(P, b\vartheta) - \widetilde{I}(\mathbb{P}_n, b\vartheta) \in \mathcal{F}$ for all $b \in B$ with probability tending to 1.*

*Then $\|\theta_n - \theta\|_B = O_P(1/\sqrt{n})$.*

**$\rho_P$-consistency condition.**

$$\sup_{b \in B} \rho_P \left( \widetilde{I}(P, b\vartheta), \widetilde{I}(\mathbb{P}_n, b\vartheta) \right) \to 0 \text{ in probability.}$$

*Then $\theta_n$ is a $\|\cdot\|_B$-asymptotically efficient estimator of $\theta$.*

### 2.6.1 One step-estimators.

Theorem 2.4 can immediately be extended to one-step estimators.

**Corollary 2.2** (Efficiency of one-step estimators of linear parameters in convex models). *Let $X \sim P \in \mathcal{M}$ for a convex model $\mathcal{M}$ and let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of $X$. Let $\theta = \vartheta(P) \in D$ be a linear parameter and $B$ be a certain collection of real valued linear mappings on $D$. Suppose that for each $P \in \mathcal{M}$, $b\vartheta$, $b \in B$, is pathwise differentiable at $P$ relative to $S(P)$ with efficient influence function $\widetilde{I}(P, b\vartheta)$.*

*Let $\theta_n = \vartheta(\widetilde{P}_n)$, $\widetilde{P}_n \in \mathcal{M}$, be any estimator of $\theta = \vartheta(P)$ for which the condition (2.14), the $P$-Donsker condition and the $\rho_P$-consistency of theorem 2.4 hold.*

*Then*

$$\theta_n^1 \equiv \theta_n + \int \widetilde{I}(\widetilde{P}_n, b\vartheta) dP_n$$

*is an $\|\cdot\|_B$-asymptotically efficient estimator of $\theta$.*

By application of the identity (2.18) we also obtain an efficiency result for the one-step estimator with sample splitting as will be defined below. Let $n_1 + n_2 = n$, where $n_i \to \infty$ if $n \to \infty$, $i = 1, 2$, and split the sample in $X_1, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$. Define $P_{n_1}^{(1)} \equiv 1/n \sum_{i=1}^{n_1} \delta_{X_i}$ and $P_{n_2}^{(2)} \equiv 1/n \sum_{i=n_1+1}^{n} \delta_{X_i}$; the empirical distributions of the first and second sample.

Let $\theta_{n_i} = \vartheta(\widetilde{P}_{n_i}^{(i)})$ be an estimator of $\theta$ based on $P_{n_i}^{(i)}$, $i = 1, 2$. Suppose that $\widetilde{P}_{n_i}^{(i)}$ satisfies condition (2.14). Then the identity holds for $\theta_{n_i}$, $i = 1, 2$:

$$b\theta_{n_i} - b\theta = -\int \widetilde{I}(\widetilde{P}_{n_i}^{(i)}, b\vartheta) dP. \tag{2.27}$$

The *one-step estimator with sample splitting* is defined as follows:

$$b\theta_n^{\mathrm{sp}} \equiv \frac{1}{2}(b\theta_{n_1} + b\theta_{n_2}) + \frac{1}{2}\left(P_{n_1}^{(1)} \widetilde{I}(\widetilde{P}_{n_2}^{(2)}, b\vartheta) + P_{n_2}^{(2)} \widetilde{I}(\widetilde{P}_{n_1}^{(1)}, b\vartheta)\right). \tag{2.28}$$

By (2.27) we have the following identity for $b\theta_n^{\mathrm{sp}}$:

$$b\theta_n^{\mathrm{sp}} - b\theta = \frac{1}{2}\left((P_{n_1}^{(1)} - P)\widetilde{I}(\widetilde{P}_{n_2}^{(2)}, b\vartheta) + (P_{n_2}^{(2)} - P)\widetilde{I}(\widetilde{P}_{n_1}^{(1)}, b\vartheta)\right).$$

The crucial difference between this identity and the identity for the one step estimator lies in the fact that here the right-hand side consists (conditionally) of sums of i.i.d. variables and thereby weak convergence is obtained under essentially weaker conditions. In order to show this we will consider

$$V_n \equiv (P_{n_1}^{(1)} - P)\widetilde{I}(\widetilde{P}_{n_2}^{(2)}, b\vartheta). \tag{2.29}$$

Conditioned on $X_{n_1+1}, \ldots, X_n$ this is a sum of i.i.d. mean zero random variables. In the following arguments all statements are conditioned on this sample, or equivalently on $P_{n_2}^{(2)}$.

By lemma 1.1 it suffices for weak convergence of $V_n$ to have that $\mathrm{Var}(\widetilde{I}(\widetilde{P}_{n_2}^{(2)}, b\vartheta)) \to \mathrm{Var}(\widetilde{I}(P, b\vartheta))$ and $\widetilde{I}(\widetilde{P}_{n_2}^{(2)}, b\vartheta)(Y) \xRightarrow{D} \widetilde{I}(P, b\vartheta)(Y)$, $Y \sim P$. In other words, if these two conditions hold almost surely for all sequences $P_{n_2}^{(2)}$, then (2.29) converges in distribution to a normal distribution $V = N(0, \widetilde{I}(\theta, b))$. By definition of weak convergence this tells us that for all bounded and continuous functions $h : \mathbb{R} \to \mathbb{R}$ we have that $E(h(V_n) \mid P_{n_2}^{(2)}) \to E(h(V))$ a.s. For such a $h$ we have that $E(h(V_n) \mid P_{n_2}^{(2)}) \leq \|h\|_\infty < \infty$. Therefore the dominated convergence theorem provides us with $E(E(h(V_n) \mid P_{n_2}^{(2)})) \to E(h(V))$ which proves the unconditional weak convergence of $V_n$ to $N(0, \widetilde{I}(\theta, b))$. This proves the following corollary for the one-step estimator with sample splitting:

**Corollary 2.3** (Efficiency of one-step estimator with sample splitting). *Let $X \sim P \in \mathcal{M}$ for a convex model $\mathcal{M}$ and let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of $X$. Let $\theta = \vartheta(P) \in D$ be a linear parameter and let $b$ be a real valued linear mapping on $D$. Suppose that for each $P \in \mathcal{M}$ $b\vartheta$ is pathwise differentiable at $P$ relative to $S(P)$ with efficient influence function $\widetilde{I}(P, b\vartheta)$.*

*Let $\theta_{n_i} = \vartheta(\widetilde{P}_{n_i}^{(i)})$, $\widetilde{P}_{n_i}^{(i)} \in \mathcal{M}$, be an estimator of $\theta$ based on $P_{n_i}^{(i)}$, $i = 1, 2$. Suppose that $\widetilde{P}_{n_i}^{(i)}$ satisfies condition (2.14), $Var(\widetilde{I}(\widetilde{P}_{n_i}^{(i)}, b\vartheta)) \to Var(\widetilde{I}(P, b\vartheta))$ and $\widetilde{I}(\widetilde{P}_{n_i}^{(i)}, b\vartheta)(Y) \xRightarrow{D} \widetilde{I}(P, b\vartheta)(Y)$, $Y \sim P$, for almost all sequences $\widetilde{P}_{n_i}^{(i)}$, $i = 1, 2$.*

*Then $b\theta_n^{\mathrm{sp}}$ is an asymptotically efficient estimator of $b\theta$.*

This theorem provides us with minimal conditions for constructing efficient estimators of linear parameters in convex models; it tells us that if we can estimate the efficient influence function consistently, then there exists an efficient estimator.

## 2.7 Bibliographic remarks.

Klaassen (1987) considers the problem of finding necessary and sufficient conditions for constructing efficient one-step estimators, using sample splitting, in models $P_{\theta,g}$, $\theta \in \mathbb{R}^d$ and $g \in \mathcal{G}$, where $\mathcal{G}$ is a class of functions. One of his crucial conditions is essentially the same as our differentiability condition in theorem 2.1. Our corollary 2.3 has the same nature as his result, except that we restricted ourselves to convex models and did not prove that the conditions are also necessary.

Greenwood and Wefelmeier (1991) consider NPMLE as a solution of the efficient score equation and also have as main condition the differentiability condition of theorem 2.1. They are concerned with efficiency of solutions of the efficient score equation in nonparametric filtered models and they point out the power of their approach for proving efficiency of the NPMLE. They formulate a similar efficiency theorem as theorem 2.1, but assume root-$n$-consistency and do not make the connection with empirical process theory. The discovery of the identity for convex models (theorem 2.2) by linking the differentiability condition to pathwise differentiability along lines is not found in the preceding literature. Linking to pathwise differentiability seems also to be the right approach for verifying the differentiability condition for non-convex models.

Gill and van der Vaart (1993) and van der Vaart (1992b), following the approach of Gill (1989), follow a different approach and concentrate on explaining efficiency of the NPMLE in semiparametric models. They base their analysis of NPMLE on a set of score equations, but where the directions do not depend on the NPMLE itself and thereby do not allow the efficient score equations which we use. As shown by Greenwood and Wefelmeier (1991) several models can be constructed where such a characterization of the NPMLE does not exist. The approach has the disadvantage that it does not separate conditions for pointwise efficiency from conditions for supremum norm efficiency. They have to use implicit function theorem requirements like invertibility of the derivative and smoothness of the generalized score equation as a whole. These assumptions are in many interesting applications not satisfied (for example, the bivariate censoring model and the line-segment model of chapter 4 and 5). Gill and van der Vaart (1993) assume that the NPMLE converges weakly as a process to a Gaussian process and then show under sharp regularity conditions, which are unfortunately hard to verify, that the covariance structure of this Gaussian process must be the optimal one. Van der Vaart (1992b) bases his analysis on a general theorem for proving weak convergence of M-estimators, assuming consistency and strong smoothness assumptions on the set of score equations. This approach does not exploit the specific structure of the efficient score equation and hence does not give optimal results for the NPMLE.

In the case that the model is convex and the parameter linear, then theorem 2.4 provides us even with consistency. In the general case (theorem 2.1) the differentiability condition can often be verified by just assuming consistency.

With Greenwood and Wefelmeier's, Klaassen's and our approach the conditions for efficiency depend on the efficiency result: the smaller $B$ the more easily the conditions can be verified. For example, very weak conditions suf-

fice for obtaining pointwise efficiency of the one step estimator with sample splitting.

The main results of this chapter can also be found in van der Laan (1993a).

# Chapter 3

# Efficiency of the sieved-NPMLE for a Class of Missing Data Models with Applications.

## 3.1 Introduction.

Assume one is concerned with *nonparametric* estimation of a distribution function $F$ based on i.i.d. observations $X_i \sim F$, $i = 1, \ldots, n$. However, there is another random variable $C$ which causes that one only observes a many to one mapping $Y = \Phi(X, C)$ of these $X_i$'s and thereby one gets only partial information about $X_i$ in the sense that one knows that $X_i$ lies in a certain region. Such models are called (nonparametric) *missing data models* or incomplete data models. We will restrict our attention to models where $X$ is completely observed with positive probability.

Several of such nonparametric missing data models are covered in the literature. Well known examples are: *univariate censoring* model (see 2.2), *double censoring* model (see Chang and Yang, 1987, Chang, 1990, Gu and Zhang, 1993), *multivariate censoring* model (see chapter 4 and its bibliography), the class of *Ibragimov-Has'minskii* (IH) models (see Ibragimov and Has'minskii, 1983, Bickel and Ritov, 1992, van der Vaart, 1992a), in particular the *Vardi-Zhang* model (Vardi and Zhang, 1992).

In order to have identifiability of $F$ one assumes that the conditional distribution $G(\cdot \mid x)$ of $C$, given $X = x$, is known or that it implies a *coarsening*

*at random* on $X$ (Heitjan and Rubins, 1991, Jacobsen, Keiding, 1994, Gill, van der Laan, Robins, 1995). One says that $X$ is coarsened at random if the observation $Y$ does not carry more information on $X$ than that $X$ lies in the with $Y$ associated region (this region is called the coarsening); i.e. the observed information on $C$ is noninformative about the location of $X$ in the associated region. A useful heuristic way to think about CAR is that $C$ should only depend on $X$ through $Y$ (i.e. what we observe). In Gill, van der Laan, Robins (1995) a general definition of CAR is given and it is shown that in nonparametric missing data models $G$ satisfies CAR if $G(\cdot \mid x)$ is dominated by a $\mu_0(\cdot \mid x)$ which is CAR itself and for which the density of $G(\cdot \mid x)$ w.r.t. $\mu_0(\cdot \mid x)$ is only a function of $(c, \Phi(c, x))$. Since one can always represent $Y = \Phi(C, X)$ with $C = Y$ and $\Phi(C, X) = C$, CAR holds if (in fact, and only if) the same density statement holds for the conditional distribution of $Y$, given $X = x$. Moreover it is shown that if this is the only assumption on $G$, then the model for $Y$ is nonparametric.

An important consequence of coarsening at random (and the only relevant consequence for us) is that the density of the distribution of $Y$ w.r.t. any by $\Phi$ induced measure factorizes in a $F$-part and a $G$-part: For all $F$ with $dF(x)G(dc \mid x) \ll \mu$ we have

$$p_{F,G}(y) = p_F(y)p_G(y), \tag{3.1}$$

where $p_{F,G}$ is the density of $Y$ w.r.t. $\mu\Phi^{-1}$, $p_F$ does not depend on $G$ and $p_G$ does not depend on $F$. Under CAR this factorization holds for all measures $\mu$. The *factorization of the likelihood* implies that a NPMLE of $F$ is computed by maximizing the $F$-part of the likelihood and hence does not depend on the knowledge on $G$. It also implies that the generalized Cramér-Rao lower bound for estimation of functionals of $F$ is independent of the knowledge on $G$. Hence CAR-missing data models are from an analytical point of view not really different from then missing data models for which we assume that $G$ is known. In this paper we are concerned with the behavior of the NPMLE in CAR-missing data models and in missing data models with $G$ known.

In a model where one only assumes CAR (as we will do in our general class of missing data models) the model is completely nonparametric which implies that all asymptotically linear estimators are asymptotically equivalent and efficient; in other words, one needs to use the NPMLE-principle to come up with sensible estimators. Hence ad hoc estimators can only be constructed by assuming stronger assumptions on $G$ than just CAR, for example, by assuming independence between $X$ and $C$. Therefore an NPMLE is less sensitive to

dependent censoring than ad hoc estimators which is an essential advantage of a NPMLE relative to an ad hoc estimator; in the bivariate censoring model as covered in chapter 4, 6, 7 and 8 many ad hoc estimators have been proposed which are all inconsistent if one allows for dependent censoring, but CAR.

An NPMLE solves each score equation corresponding with a one-dimensional submodel through the NPMLE itself. Gill (1989) shows that for missing data models a natural set of score equations for the NPMLE corresponds with the well known self-consistency equation (Efron, 1967). A solution of the self-consistency equation can be found with the *EM-algorithm* (Dempster, Laird and Rubin, 1977, Turnbull, 1976) which does in fact nothing else than iterating the self-consistency equation. In section 4 we show that any solution of the self-consistency equation which is equivalent (= absolutely continuous w.r.t. each other) with an MLE is an MLE; so if we iterate the self-consistency equation with an estimator with a certain support, then iterating the self-consistency equation provides us with the MLE $F_n$ which maximizes the likelihood over all $F$ with the same support. So in order to compute an MLE one will first need to agree on its support. The support points should include at least one point in each associated region of the $X_i$'s. A natural set of support points are the observed $X_i$ and a point in each region associated with a censored $X_i$ which does not contain uncensored $X_i$. The corresponding MLE will be refered to as Sieved-NPMLE.

Because the EM-algorithm is easy to understand it teaches us a lot about MLE's. By studying the EM-algorithm we learn that there is one crucial condition which makes the EM-algorithm work in the sense that the NPMLE will be asymptotically efficient: With probability tending to 1 each region for $X_i$ implied by the incomplete observations contains several (in the limit even infinitely many) completely observed $X_j$. This condition can be used as a rule of thumb for deciding if the NPMLE in a semiparametric CAR-missing data model, allowing complete observations, will be asymptotically efficient or not.

In the chapter 4 we cover the bivariate right-censoring model which has associated regions which are half-lines in the plane which causes the inconsistency of the NPMLE. We show how to slightly reduce the data so that the NPMLE based on the reduced data is efficient; the method is generally applicable to any missing data model where the associated regions do not have full dimension. This example also functionates as a motivation for this chapter, i.e. for making rigorous that once one has regions of full dimension, then the NPMLE is efficient.

In section 2 we define the general class of models satisfying the conditions

that $X$ is observed with positive probability and that each region contains with positive probability several observed $X_i$. By enforcing the two conditions to hold in the more stringent sense that the words *positive* and *several* in the two conditions are replaced by *larger than* $\delta > 0$ and *a fraction*, we can carry through an efficiency proof for the general class of models without getting into a very refined, technical and specific analysis. We show that the assumption that $X$ is observed with positive probability implies that each one dimensional submodel has information bounded away from zero.

The total number of completely observed $X_i$ and the number of completely observed $X_j$ in an incomplete region should be considered as measures of stability of the estimators (this appears also from simulation results in chapter 8). We show in the applications that the conditions can often be arranged by reducing the data to a compact subset of the whole support of the data. This makes the NPMLE more stable at cost of a small loss of information; it is essentially comparable with a truncated mean. This appears to be a right thing to do in practice, most of the times, since otherwise the NPMLE is less reliable for finite samples because of its large sensitivity to outliers. The assumption that given $X = x$ it is observed with probability larger than $\delta > 0$ forces one already to reduce data to a compact support. Therefore, the additional assumption that each region has $F_0$-probability larger than $\delta > 0$ is normally not an extra assumption so that it is not worthwhile to weaken that assumption while not weakening the other. In some applications these more stringent conditions cannot be artificially arranged. Hence in this chapter we will formulate two sets of assumptions which guarantee efficiency of the NPMLE, one based on the stringent conditions and one set based on weak conditions, but less worked out. In this way our set of assumptions provide a framework for verifying efficiency of NPMLE in any missing data model.

In section 3 we prove existence and uniqueness of a "sieved"-NPMLE, discuss the EM-algorithm and prove identifiability of the self-consistency equation. These missing data models are essentially (or one is allowed to work as if the censoring distribution is known or it is known) *convex* and *linear* in the parameter $F$. As shown in the preceding chapter this leads to a useful *identity* for the NPMLE; efficiency can be proved by applying theorem 2.3 and verifying the boldfaced conditions of theorem 2.4. The general efficiency proof is given in section 4. In order to make the proof work we need invertibility of the so called information operator, a Donsker class condition and continuity condition for the efficient influence curve. The invertibility of the information operator is established in section 4.1. In the remaining subsections of section

4 the Donsker class and continuity condition for the efficient influence function are covered. Section 5 contains the final theorems. Our efficiency result is successfully applied to the mentioned examples.

## 3.2 A class of missing data models.

Firstly, we will describe the class of missing data models which we want to analyze and we introduce the necessary notation. Let $\mathcal{X}$ and $\mathcal{C}$ be two vector spaces (in our applications we have $\mathcal{X} = \mathcal{C} = \mathbb{R}^k$). $(X, C)$ is a $\mathcal{X} \times \mathcal{C}$-valued random element of a probability space $(\mathcal{X} \times \mathcal{C}, \mathcal{B}, P_{X,C})$, endowed with sigma-algebra $\mathcal{B}$ and with distribution $P_{X,C}$ of $(X, C)$ is determined as follows: $X \sim F_0$, where $F_0$ is completely unknown, $C \mid X = x$ has distribution $Q(\cdot \mid X = x)$ with density $q(c \mid x)$ w.r.t. a fixed $\mu_2$ (same for each $X = x$). So $C \sim Q$ where $Q$ has density with respect to $\mu_2$ given by:

$$q(c) = \int q(c \mid x) dF_0(x).$$

$(X_i, C_i)$, $i = 1, \ldots, n$, are $n$ i.i.d. copies of $(X, C)$. We are interested in estimating the distribution $F_0$. For this purpose we try to observe $(X_i, C_i)$, but we can only get partial information about $(X_i, C_i)$ in the sense that we observe $Y_i = \Phi(X_i, C_i)$, $i = 1, \ldots, n$, where $\Phi$ is a known many to one mapping from $\mathcal{X} \times \mathcal{C}$ to a certain vector space $\mathcal{Y}$. It is assumed that $q$ is known or that $Q$ satisfies CAR:

$$q(c \mid x) = h(c, \Phi(c, x)) \text{ for some function } h. \tag{3.2}$$

In the CAR-models the likelihood factorizes in a part which only depends on $F_0$ and a part which only depends on $q$, so that one can compute the NPMLE of $F_0$ by just maximizing the first part. Moreover, for the purpose of information calculations one can also do as if $q$ is known.

We endow $\mathcal{Y}$ with the image $\sigma$-algebra $\mathcal{A} \equiv \Phi(\mathcal{B})$ and the image probability measure $P_{X,C}\Phi^{-1}$ will be denoted with $P_{F_0}$. Then for any $A \in \mathcal{A}$ we have

$$P_{F_0}(A) \stackrel{def}{\equiv} P(Y \in A) = P\left((X, C) \in \Phi^{-1}(A)\right) = P_{X,C}(\Phi^{-1}(A)). \tag{3.3}$$

We assume that $\mathcal{X}$ is a normed vector space. Let $\mathcal{B}_X$ be a sigma-algebra on $\mathcal{X}$, $\mathcal{B}_C$ be a sigma algebra on $\mathcal{C}$ and suppose that $\mathcal{B}$ is the product sigma-algebra of $\mathcal{B}_X$ and $\mathcal{B}_C$. For obtaining efficiency of the NPMLE of $F_0$ we need some assumptions on $\Phi$, $F_0$ and $Q$. These assumptions are displayed and numbered in the sequel.

If we define for a region $B \in \mathcal{B}$

$$\begin{aligned} B_x &\equiv \{c : (x,c) \in B\} \in \mathcal{B}_C \\ B_c &\equiv \{x : (x,c) \in B\} \in \mathcal{B}_X, \end{aligned}$$

assuming that these *sections* are elements of $\mathcal{B}_C$ and $\mathcal{B}_X$ (which is well known for the Borel product sigma-algebra in the Euclidean case), then for any set $B \in \mathcal{B}$ we have $P_{X,C}(B) = P((X,C) \in B) = \int_{B_1} Q(B_x \mid x) dF_0(x)$. Let $\mu_1$ be a dominating measure of $F_0$ and denote its density with $f_0$. Then $(X,C)$ has density $q(c \mid x) f_0(x)$ w.r.t. $\mu_1 \times \mu_2$. Moreover, we have that $P_{F_0} \ll (\mu_1 \times \mu_2)\Phi^{-1}$ and we denote the corresponding density with $p_{F_0}$ or $p_0$.

Let's write down the *model* $\mathcal{M}$ of probability measures on $(\mathcal{Y}, \mathcal{A})$ as described above. Let $\mathcal{F}$ be the nonparametric model consisting of all probability measures on $(\mathcal{X}, \mathcal{B}_X)$. Then

$$\mathcal{M} = \{P_F : F \in \mathcal{F}\}, \text{ where } P_F \text{ is defined as in (3.3).}$$

The parameter $F$ which we want to estimate is now formally defined by $\vartheta : \mathcal{M} \to D$, where $\vartheta(P_F) = F$ and $D$ is any vector space which contains the measures $\mathcal{F}$. For example, for $D$ we can take the space of all signed measures. If $\mathcal{X} = \mathbb{R}^k$, then we will identify $F$ with its distribution function so that the space of multivariate cadlag functions space is a natural candidate for $D$.

Each observation $Y_i$ tells us that $(X_i, C_i) \in B(Y_i) \equiv \Phi^{-1}(\{Y_i\})$. In order to define the notion of a *complete observation* we define for a set $B \in \mathcal{B}$:

$$\begin{aligned} B_1 &\equiv \{x : (x,c) \in B \text{ for some } c\} \in \mathcal{B}_X \\ B_2 &\equiv \{c : (x,c) \in B \text{ for some } x\} \in \mathcal{B}_C, \end{aligned}$$

assuming (which holds for the Euclidean case) that the *projections* $B_1, B_2$ are elements of $\mathcal{B}_X$, $\mathcal{B}_C$, respectively.

Then $B(y)_1 = \Phi^{-1}(\{y\})_1 = \{x : \Phi(x,c) = y \text{ for some } c \in C\}$ which is the projection of the region $\Phi^{-1}(\{y\})$ on the $\mathcal{X}$-space; in other words the observation $Y$ tells us that $X \in B(Y)_1$. Similarly we have $B(y)_2 = \{c : \Phi(x,c) = y \text{ for some } x \in \mathcal{X}\}$. Decompose $\mathcal{Y} \cap \text{supp}(P_0) = A_1 \cup A_2$, where

$$A_1 \equiv \{y \in \mathcal{Y} : B(y) = B(y)_1 \times B(y)_2,\ B(y_1) = \{x\}\}$$

and $A_2$ is the complement of $A_1$ within $\mathcal{Y} \cap \text{supp}(P_0)$. Notice that $Y_i \in A_1$ tells us that $B(Y_i)_1 = \{X_i\}$. Therefore an observation $Y_i$ is called *complete* (for $X$) if $Y_i \in A_1$.

For each $x \in \text{supp}(F_0)$ we denote with $y(x)$ the collection of $y \in A_1$ with $B(y)_1 = x$. For all our applications we have that $y(x)$ consists of one point. For convenience, in the sequel we will assume this, but it is not relevant for the results.

We will now state the two main assumptions on $F_0$, $Q$ and $\Phi$. The first crucial assumption we make is that, given $X = x$, the probability that this $X = x$ will be completely observed (i.e. $Y \in A_1$) is larger than $\delta > 0$. (In the sequel we will use the symbol $\delta > 0$ as a symbol for some small number larger than zero. So the same $\delta$'s are used for different numbers.) Formally, this is written down as follows:

**Assumption 1**

$$P(Y \in A_1 \mid X = x) = Q\left(B(Y(x))_2 \mid x\right) > \delta > 0, \text{ for all } x \in \text{supp}(F_0).$$

Notice that assumption 1 requires completely observed $X_i$ on the *whole* support of $F_0$. We will refer to assumption 1* as the weakened version of this assumption obtained by setting $\delta = 0$.

The next assumption involves an assumption related to the density of the distribution of $X$ given $Y$, $Y \in A_2$, w.r.t. $F$, assuming that this density exists. The following guarantees the existence of such a density and can also be shown to be necessary: Suppose that the conditional distribution of $Y$ on $A_2$ given $X$, i.e. $Q(\Phi(x,\cdot)^{-1} \mid x)$, has a density $q_1(y \mid x)$ w.r.t. a fixed measure $\mu_3$ (this does not hold on $A_1$, but we only need this assumption on $A_2$). Then the joint distribution of $X, Y$, $Y \in A_2$, is $q_1(y \mid x)F(dx)\mu_3(dy)$. However, assuming $F_0(B(Y)_1) > 0$ we have

$$q_1(y \mid x)F(dx)\mu_3(dy) = \int_{B(y)_1} q_1(y \mid x)F(dx)\mu_3(dy)\frac{q_1(y \mid x)F(dx)}{\int_{B(y)_1} q_1(y \mid x)F(dx)},$$

where $\int_{B(y)_1} q_1(y \mid x)F(dx)\mu_3(dy)$ is the marginal distribution of $Y$ on $A_2$. Consequently, the conditional distribution of $X$ given $Y$, $Y \in A_2$, is given by $q_1(y \mid x)F(dx)/\int_{B(y)_1} q_1(y \mid x)F(dx)$ which proves that the distribution of $X$ given $Y$ has a density proportional to $q_1(y \mid x)$ w.r.t. $F$. We can decompose $q_1(y \mid x) = \alpha(y)q_1'(y \mid x)$, where $\alpha(y)$ is the factor in $q_1(y \mid x)$ which does not depend on $x$. Then

$$\frac{q_1(y \mid x)F(dx)}{\int_{B(y)_1} q_1(y \mid x)F(dx)} = \frac{q_1'(y \mid x)F(dx)}{\int_{B(y)_1} q_1'(y \mid x)F(dx)}.$$

We denote $p'_F(y) \equiv \int_{B(y)_1} q_1'(y \mid x)F(dx)$, which is the normalizing constant for the distribution of $X$ given $Y = y$. If $Q$ satisfies CAR, then $q_1(y \mid x)$ is only a function of $y$ and hence then we have $p'_F(y) = F(B(y)_1)$.

**Assumption 2** Suppose that the conditional distribution of $Y$ on $A_2$ given $X = x$, i.e. $Q(\Phi(x,\cdot)^{-1} \mid x)$, has a density $q_1(y \mid x)$ w.r.t. a fixed measure $\mu_3$. Moreover, assume that

$$p_0'(y) > \delta > 0 \text{ and } P(X \in B(y)_1) = F_0(B(y)_1) > \delta > 0 \text{ for all } y \in A_2.$$

and $\{I_{B(Y)_1} : Y \in A_2\}$ is a $F_0$-Glivenko-Cantelli class. *If $Q$ satisfies CAR, then $p_0'(y) = F_0(B(y)_1)$.*

We will refer to assumption 2* as the same assumption, but with $\delta$ replaced by 0. The Glivenko-Cantelli assumption tells us that

$$\|\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(Y)_1) - F_0(B(Y)_1)\|_{y \in A_2^j} \to 0 \text{ a.s.}$$

Because $F_0(B(Y)_1) > 0$ this implies that all the incomplete regions $B(y)_1$ will contain (for $n$ large enough) a fraction of the underlying $X_j$ and by assumption 1 each underlying $X_j$ has probability larger than 0 to be completely observed. It is now straightforward to verify that this implies that with probability tending to 1 each $B(Y_j)_1$ will contain completely observed $X_i$ and the minimal number of completely observed $X_i$ in $B(Y_j)_1$, $Y_j \in A_2$, converges to infinity.

We conclude that for $n$ large enough assumption 1* and 2* provide us with probability tending to 1 with the following picture of the data: a fraction of the $X_i$ will be completely observed and if $X_i$ is not completely observed, then $X_i \in B(Y_i)_1$, where $B(Y_i)_1$ is a region which contains many completely observed $X_j$.

### 3.2.1 Verification of assumptions 1 and 2 for the examples.

**Example 3.1 (Univariate censoring).**
**Model.** We have $n$ i.i.d. copies $X_i \in \mathbb{R}_{\geq 0}$ of $X \sim F_0$, where $F_0$ is completely unknown. We have $n$ i.i.d. copies $C_i \in \mathbb{R}_{\geq 0}$ of $C \sim G_0$, where $G_0$ is completely unknown. $X$ and $C$ are independent. Denote the survival functions of $F_0$ and $G_0$ with $S_0$ and $H_0$, respectively. Let $F_0 \ll \mu_1$, $G_0 \ll \mu_2$ with densities $f_0$, $g_0$, respectively. We observe:

$$Y_i = (Z_i, D_i) = \Phi(X_i, C_i) \equiv (\min(X_i, C_i), I(X_i \leq C_i)), \ i = 1, \ldots, n.$$

We are interested in estimating $S_0$. If $Y = (z,1) \in A_1$ (uncensored), then $B(z,1) = \{z\} \times (z, \infty)$. If $Y = (z,0) \in A_2$ (censored), then $B(z,0) = (z,\infty) \times \{z\}$. We have $p_0(z,d) = f_0(z)H_0(z)I(d=1) + S_0(z)g_0(z)I(d=0)$, where $p_0$ is

the density induced by $F_0 \times G_0$ with respect to $(\mu_1 \times \mu_2)\Phi^{-1}$. The sample space is $\mathcal{Y} = \mathbb{R}_{\geq 0} \times \{0,1\}$. $\mathcal{Y} = A_1 \cup A_2$, where $A_1 = \mathbb{R}_{\geq 0} \times \{1\}$ and $A_2 = \mathbb{R}_{\geq 0} \times \{0\}$.
**Assumption 1.** $P(Y \in A_1 \mid X = x) = H_0(x)$. So assumption 1 requires that $H_0(x) > \delta > 0$ $F_0$ a.e.
**Assumption 2.** We have $dP_{G_0}((z,0) \mid x) = I(x > z)dG_0(z)$. So $q_1'((z,0) \mid x) = dP_0((z,0) \mid x)/dG_0(z) = I(x > z)$ and therefore $p_0'(z,0) = S_0(z)$. So assumption 2 requires that $S_0(z) > \delta > 0$ $P_0(\cdot,0)$ a.e., or in other words $S_0(z) > 0$ for all possible censored $z$.
**How to arrange assumption 1 and 2?** Fix $\tau < \infty$ so that $S_0(\tau) > \delta > 0$ and $H_0(\tau) > \delta > 0$. Make each observation $z > \tau$ uncensored at $\tau$. This does not influence the NPMLE on $[0,\tau)$; by the EM-algorithm (as explained in the next section) we know that all uncensored and right-censored observations after $\tau$ put only mass on $(\tau,\infty)$. Then these truncated observations are coming from $F_0^\delta$ which equals $F_0$ on $[0,\tau)$, but which has an atom at $\tau$ so that $F_0^\delta(\tau) = 1$. Now, $S_0(z) > \delta$ for all censored $(z,0)$ and $H_0(\tau) > \delta > 0$ and thereby assumption 1 and 2 are satisfied.

**Example 3.2 (Double censoring).**
**Model.** We have $n$ i.i.d. copies $X_i$ of $X \sim F_X$, $F_X$ unknown. We have $n$ i.i.d. copies of $(Z_i, Y_{1i})$ of $(Z, Y_1) \sim G_{Z,Y_1}$ unknown, except that $P(Y_1 > Z) = 1$. Let $Y_1 \sim F_{Y_1}$ and $Z \sim F_Z$. $X$ and $(Z, Y_1)$ are independent. Let $W \equiv \min(\max(Z,X), Y_1)$ and $D = 1$ if $W = X$, $D = 2$ if $W = Y_1$ and $D = 3$ if $W = Z$. We observe:

$$Y = (W,D) = \Phi(X,Z,Y_1) \equiv (\min(\max(Z,X), Y_1), D).$$

So if $Z < X < Y_1$, then $X$ is completely observed and if $X \geq Y_1$, then $X$ is right censored at $Y_1$ and if $X \leq Z$, then $X$ is left censored at $Z$. We have $A_1 = \{(w,1) : w \geq 0\}$, $A_2 = \{(w,2),(w,3) : w \geq 0\}$ and the sample space is given by $\mathcal{Y} = A_1 \cup A_2$. The underlying probability space corresponding with $(X,(Z,Y_1))$ is $\mathbb{R}^3_{\geq 0}$ endowed with the Borel sigma algebra. We are interested in estimating $F_X$.

We have

$$\begin{aligned} B(w,1) &= \{w\} \times [0,w) \times (w,\infty) \\ B(w,2) &= (w,\infty) \times [0,w) \times \{w\} \\ B(w,3) &= [0,w) \times \{w\} \times (w,\infty). \end{aligned}$$

$p(w,2) = S_X(w) f_{Y_1}(w)$ and $p(w,3) = F_X(w) f_Z(w)$.
**Assumption 1.** $P(Y \in A_1 \mid X = x) = G_{Z,Y_1}([0,x) \times (x,\infty)) > \delta > 0$.

**Assumption 2.** $dP_{G_0}((w,2) \mid x) = I(w \leq x)G_0([0,w] \times \{w\}) = I(w \leq x)F_Z[0,w]dF_Y(w)$. So $dP_{G_0}((w,2) \mid x)/dF_Y(w) = I(w \leq x)F_Z(0,w)$. So $q_1'((w,2) \mid x) = I(w \leq x)$ and $p'(w,2) = S_X(w)$. $dP_{G_0}((w,3) \mid x) = I(w > x)G_0(\{w\} \times [w,\infty)) = I(w > x)F_y([w,\infty))dF_Z(w)$. So $dP_{G_0}((w,3) \mid x)/dF_Z(w) = I(w > x)F_y([w,\infty))$ and hence $q_1'((w,3) \mid x) = I(w > x)$. Consequently, we have $p'(w,2) = S_X(w)$ and $p'(w,3) = F_X(w)$. Assumption 2 requires $F_X(z) > \delta > 0$ for all observed $z$ and $S_X(y_1) > \delta > 0$ for all observed $y_1$.

**How to "arrange"?** Fix a $\tau < \infty$. Assume $F_Z(\tau) = 1$ and $S_{Y_1}(\tau) > 0$. This means that after $\tau$ we only have uncensored ($d = 1$) and right-censored ($d = 2$) observations. Then as in the univariate censoring model we can make all observations after $\tau$ uncensored and by the same reason this does not influence the NPMLE on $[0,\tau)$. Then these truncated observations are coming from $F_X^\delta$ which equals $F_X$ on $[0,\tau)$, but which has an atom at $\tau$ so that $F_X^\delta(\tau) = 1$.

Now, we assume that there exists a $\delta_1$ so that $F_Z[0,\delta_1] = F_Z(\{0\}) > 0$, i.e. it has an atom at 0 and no mass immediately after this atom. To summarize: by assuming that $F_Z(\tau) = 1$ and $S_{Y_1}(\tau) > 0$ we could arrange by artificial censoring that:

(i): $F_X(\tau-) < 1$, $F_X(\tau) = 1$.

And we have also to make the following assumption:

(ii): $F_Z[0,\delta_1] = F_Z(\{0\}) > 0$ and $F_X(\delta_1) > 0$ for certain $\delta_1 > 0$.

Under these assumptions (which are the same as the assumptions as used in Chang-Yang (1990) for proving asymptotic normality), assumption 1 and 2 can be proved as follows.

$$G_{Z,Y_1}(Z \in [0,w), Y_1 \in (w,\infty)) \geq G_{Z,Y_1}(Z = \{0\}, Y_1 > Z) = F_Z(\{0\}) > \delta$$

by assumption (ii). This proves assumption 1. Furthermore, $P((0,\delta_1],3) = 0$ (i.e. there are no observed $z \in (0,\delta_1)$) and therefore for $F_X(z) > \delta > 0$ we need $F_X(\delta_1) > 0$ and that holds by assumption (ii). We have $P([\tau,\infty),2) = 0$ by assumption (i) (i.e. there are no observed $y_1 \in (\tau,\infty)$) and therefore for $S_X(y_1) > \delta > 0$ it suffices to have $S_X(\tau-) > 0$ and that holds by assumption (i). This proves assumption 2. The conditions (i) and (ii) are not very easy to arrange. Gu and Zhang (1993) succeeded, by a specific analysis, to weaken these conditions. Of course, our assumptions 1* and 2* are easily satisfied.

**Example 3.3 (Ibragimov-Has'minskii (IH) Model).**

**Model.** $X_1,\ldots,X_n$ are i.i.d. copies of a $\mathcal{X}$-valued random variable $X$ which is distributed according to an unknown distribution $F_0$. $C_1\ldots,C_n$ are i.i.d. copies of $C$ of a $\mathcal{C}$-valued random variable with conditional distribution of $C$

given $X = x$ given by a known kernel $Q(\cdot, x)$ which has a density $q(\cdot \mid x)$ w.r.t. $\mu_2$. Let $\Delta$ be a third variable independent of $X$ and $C$ so that $\Delta \in \{0, 1\}$ and $P(\Delta = 1) = \lambda \in (0, 1]$. We observe the following many to one mapping of $(X, C, \Delta) \in \mathcal{X} \times \mathcal{C} \times \{0, 1\}$:

$$Y = (Z, \Delta) = \Phi(X, C, \Delta) \equiv (\Delta X + (1 - \Delta)C, \Delta).$$

$A_1 = \{(z, 1) : z \in \mathcal{X}\}$ and $A_2 = \{(z, 0) : z \in \mathcal{C}\}$. $B(Y) \subset \mathcal{X} \times \mathcal{C}$ and

$$\begin{aligned} B(z, 1) &= \{z\} \times \mathcal{C} \times \{1\} \\ B(z, 0) &= \mathcal{X} \times \{z\} \times \{0\}. \end{aligned}$$

**Assumption 1.** $P(Y \in A_1 \mid X = x) = \lambda > 0$. So assumption 1 is satisfied.
**Assumption 2.** $dP_Q(z, 0 \mid x) = (1 - \lambda)q(z \mid x)d\mu_2(z)$ and hence $dP_Q(z, 0 \mid x)/d\mu_2(z) = (1 - \lambda)q(z \mid x)$. So assumption 2 requires that $p_0'(z, 0) = \int q(z \mid x)dF_0(x) > \delta > 0$ for $P_0(\cdot, 0)$ almost each $z$. By artificially censoring the $C_i$ it will often not be hard to arrange this assumption.

For example, let's consider the Vardi and Zhang (1992) special IH model: $q(c \mid x) = \frac{1}{x} I(c \leq x)$ and $x$ and $c$ are real valued. Then assumption 2 requires $p_0'(c, 0) = \int_c^\infty \frac{dF_0(x)}{x} > \delta > 0$ for almost each observed $c$.
**How to arrange in the Vardi-Zhang model?** Let $\tau$ be so that $F_0(\tau) < 1$. By artificially censoring the observed $C_i$ at $\tau$ we obtain the following model: if $x < \tau$, then $x \sim F_0$ and $C \mid x$ is uniform $[0, x]$ and if $x \geq \tau$, then $C \mid x$ has density $1/x$ on $[0, \tau]$ and it puts mass $1 - \tau/x$ at $\tau$. Then it is clear that assumption 2 holds. It is also satisfied if $F_0$ has compact support on $[0, \tau]$ and an atom at $\tau$.

## 3.3 Existence of sieved-NPMLE and EM-equations.

We avoid the search for a dominating measure of the NPMLE (as defined in Kiefer and Wolfowitz, 1956) by analyzing the so called *sieved*-NPMLE of $F_0$ which is purely discrete and only puts mass on the completely observed $X_i$ in each region $B(Y_j)_1$ and if $B(Y_j)_1$ does not contain completely observed $X_i$, then it puts mass on one point in $B(Y_j)_1$, chosen by us.

Let $P_n$ be the empirical distribution function of the data $Y_1, \ldots, Y_n$. Let $\{x_1, \ldots, x_{m(n)}\}$ be the set consisting of completely observed $X_i$ and the chosen points in the regions $B(Y_j)_1$ which contain no completely observed $X_i$. Let $\mu_n$ be the counting measure on $\{x_1, \ldots, x_{m(n)}\}$ and define $\mathcal{F}(\mu_n)$ as the set of all distributions $F$ with $F \ll \mu_n$.

Now, we define the sieved-NPMLE by:

$$F_n \equiv \arg \max_{F \in \mathcal{F}(\mu_n)} \int \log(p_F) dP_n. \tag{3.4}$$

Sometimes $F_n$ coincides with a NPMLE, and is certainly as natural because $F_n$ eventually puts all its mass on the uncensored observations (what it should do, see discussion of EM-algorithm, below). In this section we prove existence and uniqueness of the sieved-NPMLE under the weak assumption 1* and 2*. Here there is no need for the stronger assumptions since the statements are statements for fixed $n$.

### 3.3.1 Existence and uniqueness of sieved-NPMLE.

Denote $\mathcal{F}(\mu_n)$ by $\mathcal{F}_n$ and let $\mathcal{P}_n$, be the class of densities $p_F$, $F \in \mathcal{F}_n$, w.r.t. $(\mu_n \times \mu_2)\Phi^{-1}$. For each $F \in \mathcal{F}_n$ we denote $f \equiv dF/d\mu_n$. Notice that each region $B(Y_i)_1$ contains one or more elements of $\{x_1, \ldots, x_{m(n)}\}$. We will show that $F_n$ exists and is unique.

Consider the set $\mathcal{F}_n(\delta_1) \equiv \{F \in \mathcal{F}_n : F\{X_i\} \geq \delta_1, i = 1, \ldots, m\}$. Notice that $\mathcal{F}_n(\delta_1)$ is isomorphic with a *compact convex* subset in $\mathbb{R}^m$. Firstly, we show that $p_F \in \mathcal{P}_n$ is uniformly bounded away from zero on this set. By assumption 1* we have that if $Y_i \in A_1$ and $F \in \mathcal{F}_n(\delta_1)$, then $p_F(Y_i) > \delta\delta_1$. If $Y_i \in A_2$, then

$$p_F(Y_i) \geq \delta_1 \sum_{j=1}^{m} I(x_j \in B(Y_i)_1) Q(B(Y_i)_{x_j} \mid x_j), \tag{3.5}$$

which is larger than $\delta_1 \max_{x_j \in B(Y_i)_1} Q(B_{x_j}(Y_i) \mid x_j) > \delta_2\delta_1$, for certain $\delta_2 > 0$. Consequently, $\min_i p_F(Y_i)$ is uniformly in $F \in \mathcal{F}_n(\delta_1)$ bounded away from zero.

Therefore $F \to \int \log(p_F) dP_n$ is a *continuous* function on $\mathcal{F}_n(\delta_1)$ for each $\delta_1 > 0$. Moreover, $\mathcal{F}_n(\delta_1)$ is compact. Consequently the MLE, say $F_n(\delta_1)$, over $\mathcal{F}_n(\delta_1)$ exists. Now, we want to show that there exists a $\delta > 0$ so that $F_n(\delta)$ lies in the interior of $\mathcal{F}_n(\delta)$. Assume that there does not exist such a $\delta > 0$. That means that $\lim_{\delta \downarrow 0} \min_{i \in \{1,\ldots,m\}} F_n(\delta)(X_i) = 0$ and consequently $\int \log(p_{F_n(\delta)}) dP_n \to -\infty$ for $\delta \to 0$, which contradicts that $\int \log(p_{F_n(\delta_1)}) dP_n \geq \int \log(p_{F_n(\delta_2)}) dP_n$ for $\delta_1 < \delta_2$. This proves that $F_n$ exists, namely it equals the interior maximum $F_n(\delta)$ for $\delta$ small enough. (The EM-equations below tell us that $\delta < 1/n$ suffices.)

We will now show uniqueness of $F_n$. If $F_1(\{X_i\}) \neq F_2(\{X_i\})$ for one of the completely observed $X_i$, then by assumption 1 $p_{F_1} \neq p_{F_2}$ w.r.t. the counting measure on the complete observations $Y_1, \ldots, Y_m$. This follows from $p_{F_1}(Y_i) -$

$p_{F_2}(Y_i) = (f_1 - f_2)(X_i)Q(B(Y_i)_2 \mid X_i)$, $i = 1, \ldots, m$, and that by assumption 1* $Q(B(Y_i)_2 \mid X_i) > 0$. Suppose now that $F_1(\{x_i\}) \neq F_2(\{x_i\})$ for a point $x_i$ in an empty region $B(Y_j)_1$. Then $F_1(B(Y_j)_1) = F_1(\{x_i\}) \neq F_2(\{x_i\}) = F_2(B(Y_j)_1)$ and therefore it follows that $p_{F_1}(Y_j) \neq p_{F_2}(Y_j)$ (see (3.5)). This shows that if $F_1 \neq F_2$ w.r.t. $\mu_n$, then $P_{F_1} \neq P_{F_2}$ w.r.t. $P_n$.

By linearity of $F \to p_F$ and the strict concavity of "log" this implies *strict concavity* of the log likelihood $F \to \int \log(p_F) dP_n$ on $\mathcal{F}(\mu_n)$. Now, the uniqueness follows from the fact that a strictly concave function on a convex set has a unique maximum.

### 3.3.2 EM-equations for the sieved-NPMLE.

Suppose that assumptions 1* and 2* hold. Let $\mathcal{S}(F_n)$ be the class of lines $\epsilon F_1 + (1-\epsilon)F_n$, $F_1 \in \mathcal{F}$, through $F_n$ with score $h = d(F_1 - F_n)/dF_n \in L_0^2(F_n)$. By convexity of $\mathcal{F}_n$ we have that these lines are submodels of $\mathcal{F}_n$. Let $S(F_n)$ be the corresponding tangent cone (=collection of scores) and notice that it includes all $h \in L_0^2(F_n)$ with finite supremum norm. Then it is trivial to verify that the tangent space $T(F_n)$ (= closure of linear extension of $S(F_n) \subset L_0^2(F_n)$) equals $L_0^2(F_n)$. The lines $F_{n,\epsilon,h} \in \mathcal{S}(F_n)$ with score $h$ generate one-dimensional submodels $P_{F_{n,\epsilon,h}}$ through $P_{F_n}$ with score $A_{F_n}(h) \in L_0^2(P_{F_n})$, where $A_{F_n}$ is the so called score operator.

The score operator is given by:

$$A_{F_n} : L_0^2(F_n) \to L_0^2(P_{F_n}) : A_{F_n}(h)(Y) = E_{F_n}(h(X) \mid Y).$$

This is a result which holds for any missing data model (see van der Vaart, 1988, Gill, 1989, Bickel et al., 1993, section 6.6).

$F_n$ maximizes the log likelihood over $\mathcal{F}_n$. By differentiating the log likelihood along $P_{F_{n,\epsilon,h}}$ we obtain:

$$E_{P_n}(A_{F_n}(h)) = 0 \text{ for all } h \in S(F_n) \text{ with } \|h\|_\infty < \infty. \tag{3.6}$$

In particular, this holds for $h = I_E - F_n(E)$ for a collection of sets $E \in \mathcal{E} \subset \mathcal{B}_X$. Let $\mathcal{E}$ be so that each $F \in \mathcal{F}_n$ is uniquely determined by $F(E)$, $E \in \mathcal{E}$. Then this equation reduces to the well known self-consistency equation:

$$F_n(E) = \int P_{F_n}(X \in E \mid Y)\, dP_n(Y) \text{ for all } E \in \mathcal{E} \tag{3.7}$$

or equivalently, with $f_n \equiv dF_n/d\mu_n$,

$$f_n(X_i) = \int P_{F_n}(X = X_i \mid Y) dP_n(Y) \text{ for all } X_i,\ i = 1, \ldots, m.$$

By copying the proof of theorem 3.1 below one shows that equation (4.6) has a unique solution in the class $\{F : F \equiv \mu_n\}$, where $F \equiv \mu_n$ means that the two measures are equivalent. A solution of (4.6) is computed with the EM-algorithm. Dempster et al. (1977) and Turnbull (1976) (see also Meilijson, 1989)) show that the strictly concave likelihood $\int \log(p_{F_n^k}) dP_n$ (in $F$) increases at each step and converges to its unique maximum at $F_n$.

The EM-algorithm does the following. Start with a $F_n^0 \equiv \mu_n$. Now, for $k = 0, 1, \ldots$ and $i = 1, \ldots, m$ we can compute

$$f_n^{k+1}(X_i) = \int P_{f_n^k}(X = X_i \mid Y)\, dP_n(Y) = \frac{1}{n} \sum_{j=1}^{n} P_{f_n^k}(X = X_i \mid Y_j). \quad (3.8)$$

This means that each observation $Y_j$ has mass $1/n$ which it redistributes over $B(Y_j)_1$ as follows: a point $X_i \in B(Y_j)_1$ gets mass $1/n \times P_{f_n^k}(X = X_i \mid Y_j)$. This step is natural because our ultimate goal should be to give the $X_j$ mass $1/n$, but because of the random censoring we only know that $X_j \in B(Y_j)_1$; so we redistribute the $1/n$ over $B(Y_j)_1$ according to a good estimate of the conditional density over $B(Y_j)_1$, namely $P_{f_n^k}(X = \cdot \mid Y_j)$.

Define $P_c(A) \equiv P(X \in A,\ Y \in A_1) = P(B(Y)_1 \in A, Y \in A_1)$, which is the distribution of the *completely observed* $X_i$. Let $P_{cn}(A) = 1/n \sum_{i=1}^{n} (I(X_i \in A, Y_i \in A_1)$ be the empirical distribution of $P_c$. The completely observed $X_i$ get the full mass $1/n$ from $Y_i$ and therefore $f_n^k(X_i) \geq \#\{j : X_j = X_i\}/n$, $i = 1, \ldots, m$. This tells us that for each $A \in \mathcal{B}_X$ we have $F_n(A) \geq P_{cn}(A)$. If assumption 1* holds, then this implies that $\limsup F_n(A) \geq \delta F_0(A)$ for all $A \in \mathcal{B}_X$.

We summarize the obtained results in the following lemma (notation: $P_n f = \inf f dP_n$):

**Lemma 3.1** *Let assumptions 1* and 2* hold. With probability tending to 1 we have that each $B(Y_i)_1$ contains completely observed $X_j$.*

1. *The sieved-NPMLE $F_n \in \mathcal{F}(\mu_n)$ over $\mathcal{F}(\mu_n)$ exists and is unique.*
2. *For each set $A$ we have $F_n(A) \geq P_{cn}(A)$ and if assumption 1 holds, then* $\limsup F_n(A) \geq \delta F_0(A)$.
3. *The score operator at $P_F$ is given by:*

$$A_F : L_0^2(F) \to L_0^2(P_F) : h \mapsto E_F(h(T) \mid Y).$$

*$F_n$ solves*

$$P_n\left(A_{F_n}(h - F_n(h)) = 0 \text{ for all } h \in L^2(F_n) \text{ with } \|h\|_\infty < \infty. \quad (3.9)\right.$$

4. *$F_n$ is found by iterating the EM-algorithm above from some $F_n^0 \equiv \mu_n$. $F_n$ is the unique solution of (3.9) in $\{F \in \mathcal{F}_n : F \equiv \mu_n\}$.*

*Heuristics of assumption 1 and 2.* We end this subsection with the heuristic explanation of why $F_n$ will have a good performance under assumptions 1 and 2. By lemma 3.1 the sieved-NPMLE exists and is unique. Our assumptions 1* and 2* will take care that with probability tending to 1 each of the $B(Y_j)_1$ corresponding with $Y_j \in A_2$ will contain one or more completely observed $X_i$, $i = 1, \ldots, m$. Then $F_n$ puts only mass on the completely observed $X_i$.

The EM-algorithm starts with an initial estimator $F_n^0$ which puts mass only on each of the completely observed $X_i$. Each completely observed $X_i$ gets mass $1/n$ from itself and $1/nP_{F_n^k}(X = X_i \mid Y_j)$ from all other observations $Y_j$. Because there is only mass on complete observations, it follows that the estimate $P_{F_n^k}(\cdot \mid Y_j)$ is determined by mass given to the completely observed $X_i \in B(Y_j)_1$ and hence can only improve if $B(Y_j)_1$ contains completely observed $X_i$. Therefore these incomplete $Y_j$ can only redistribute consistently if $B(Y_j)_1$ contain enough completely observed $X_i$, and this is exactly guaranteed by assumption 1 and 2 with probability tending to 1.

### 3.3.3 Identifiability of the self-consistency equation.

The following theorem is useful for proving consistency of solutions of the self-consistency equation (see e.g. Gu and Zhang, 1993, identifiability of the self-consistency equation is a crucial ingredient of any consistency proof based on the self-consistency equations). Moreover, it tells us in what sense the self-consistency equation determines the NPMLE. Let $\mathcal{F}_0$ be the set of all distributions on $\mathcal{X}$ which are equivalent with $F_0$ and for which $\|dF_0/dF\|_\infty < \infty$. Denote the densities $dF/d\mu_1$ by $f$.

**Theorem 3.1** *Suppose that $F \in \mathcal{F}_0$, $F \neq F_0$, implies $p_F \neq p_{F_0}$ $P_{F_0}$ a.e. (e.g. assumption 1 holds). For $F \in \mathcal{F}_0$ we denote*

$$S_0(F) \equiv \left\{ h = \frac{f_1 - f}{f} : F_1 \in \mathcal{F}_0 \right\}.$$

*Then*

$$P_{F_0}(A_F(h)) = 0, \forall h \in S_0(F),\ F \in \mathcal{F}_0 \implies F = F_0.$$

*In other words: the self-consistency equation in $P_{F_0}$ has a unique solution in $\mathcal{F}_0$, namely $F_0$.*

**Proof.** Assume $F \neq F_0$, $F \equiv F_0$, $\|dF_0/dF\|_\infty < M$ and $U_h(F, P_{F_0}) \equiv P_{F_0}(A_F(h)) = 0$ for all $h \in S_0(F)$. We want to get a contradiction (then we have to conclude that $F = F_0$). Set $h = (f_0 - f)/f$ and notice that $h \in S_0(F)$ and $\|h\|_\infty < \infty$. Define for this $h$ a function $\Phi_h : I \subset \mathbb{R} \to \mathbb{R}$ on a closed interval $I$ around zero given by:

$$\Phi_h(\epsilon) \equiv \int \log(p_{F_{\epsilon,h}}) dP_{F_0},$$

where $F_{\epsilon,h}$ is a line through $F$ with score $h$; in terms of densities it is given by $f_{\epsilon,h} = (1 + \epsilon h)f$. $U_h(F, P_{F_0}) = 0$ tells us that $\frac{d}{d\epsilon}\Phi_h(\epsilon) \mid_{\epsilon=0} = 0$. However, by linearity of $f \to p_f$ and the strict concavity of the "log" (and our identifiability assumption) we have:

$$\begin{aligned}
\Phi_h(\epsilon) &= \int \log\left(p_{(1+\epsilon(f_0-f)/f)f}\right) dP_0 \\
&= \int \log\left((1-\epsilon)p_f + \epsilon p_{f_0}\right) dP_0 \\
&> (1-\epsilon)\int \log(p_f) dP_0 + \epsilon \int \log(p_0) dP_0 \\
&= (1-\epsilon)\Phi_h(0) + \epsilon \int \log(p_0) dP_0
\end{aligned}$$

and hence

$$\begin{aligned}
\Phi_h(\epsilon) - \Phi_h(0) &> \epsilon\left(\int \log(p_0) dP_0 - \int \log(p_f) dP_0\right) \\
&> \epsilon \cdot \delta,
\end{aligned}$$

where by the Jensen inequality $\delta > 0$, using that $p_f \neq p_0$ $P_0$ a.e. So

$$\frac{1}{\epsilon}\left(\Phi_h(\epsilon) - \Phi_h(0)\right) > \delta > 0,$$

which contradicts that $\frac{d}{d\epsilon}\Phi_h(\epsilon) \mid_{\epsilon=0} = 0$. □

The crucial ingredients in this proof were the convexity of $\mathcal{F}$, the linearity of $F \to P_F$ and the identifiability of $F$ from $P_F$ as stated in the theorem.

## 3.4 Efficiency of the sieved-NPMLE.

For each $F \in \mathcal{F}$ define $\mathcal{S}(F)$ as all lines $\epsilon F_1 + (1-\epsilon)F$, $F_1 \in \mathcal{F}$, $F_1 \ll F$, with scores $h = d(F_1 - F)/dF \in L_0^2(F)$. Let $S(F)$ be the tangent cone of $\mathcal{S}(F)$ and recall that the tangent space $T(F)$ of $\mathcal{S}(F)$ equals $L_0^2(F)$.

**Lemma 3.2** *The score operator at $F$ is given by:*

$$A_F : L^2(F) \to L^2(P_F) : A_F(h)(Y) = E_F(h(X) \mid Y).$$

*The adjoint of $A_F$ is given by:*

$$A_F^\top : L^2(P_F) \to L^2(F) : A_F^\top(v)(X) = E_F(v(Y) \mid X).$$

*The information operator is defined by:*

$$I_F = A_F^\top A_F : L^2(F) \to L^2(F) : I_F(h)(X) = E_F\left(E_F(h(X) \mid Y) \mid X\right).$$

**Proof.** For the derivation of the score operator see Gill (1989), Bickel et al. (1993). We have:

$$\begin{aligned}
\langle A_F(h), v\rangle_{P_F} &= E_F\left(E_F\left(h(X) \mid Y\right) v(Y)\right) \\
&= E_F\left(E_F\left(h(X)v(Y) \mid Y\right)\right) \\
&= E_F\left(h(X)v(Y)\right) \\
&= E_F\left(E_F\left(h(X)v(Y) \mid X\right)\right) \\
&= E_F\left(h(X)E_F\left(v(Y) \mid X\right)\right) \\
&= \langle h, E_F\left(v(Y) \mid X\right)\rangle_F.
\end{aligned}$$

So $A_F^\top(v) = E_F(v(Y) \mid X)$. □

If $\inf_{\|h\|_F=1} \|I_F(h)\|_F = 0$, then $\inf_{\|h\|_F=1} \langle A_F(h), A_F(h)\rangle_{P_F} = 0$ and hence there exist one dimensional submodels $P_{F_{\epsilon,h}}$ with arbitrarily low information. In this case it might be hard or impossible to estimate $F$ as a whole at root-$n$ rate. So the following lemma indicates the strength of assumption 1.

**Lemma 3.3** *If assumption 1 holds (with $\delta > 0$), then $I_F : L^2(F) \to L^2(F)$ is onto and has a bounded inverse. Moreover, the bound does not depend on $F$:*

$$\|I_F^{-1}(h)\|_F \leq \frac{1}{\delta}\|h\|_F.$$

*If assumption 1* holds, then $I_F$ is 1-1, but not necessarily onto.*

**Proof.** By lemma 3.3, for the onto and bounded invertibility of $I_F$ it suffices to show that $\|A_F(h)\|_{P_F} \geq \sqrt{\delta}\|h\|_F$. By only integrating over the complete observations, application of the substitution rule $\int_B h(T^{-1}(x))dF(T^{-1})(X) =$

$\int_{T^{-1}(B)} h(y)dF(y)$ in the third line below, using that $y \to B(y)_1$ is 1-1 from $A_1$ to Supp$(F_0)$, and assumption 1 at the fourth line provides us with

$$\begin{aligned}
\|A_F(h)\|^2_{P_F} &\geq \|h(B(y)_1)I(y \in A_1)\|^2_{P_F} \\
&= \int_{A_1} h^2(B(y)_1)dP_F(y) \\
&= \int_{A_1} h^2(B(y)_1)Q\left(B(y)_2 \mid B(y)_1\right)dF(B(y)_1) \\
&= \int_{\text{supp}(F)} h^2(x)Q\left(B(y(x))_2 \mid x\right)dF(x) \\
&\geq \delta \int h^2(x)dF(x) \\
&= \delta\|h\|^2_F.\square
\end{aligned}$$

Let $S(P_F)$ be the tangent cone at $P_F$ corresponding with the submodels $P_{F_{\epsilon,h}}$, $h \in S(F)$; in other words $S(P_F)$ equals the range of $S(F)$ under $A_F$. Application of corrolary 2.1 to the parameter $\vartheta(P_F) = F$ provides us now with the following result.

**Lemma 3.4** *Let assumption 1 hold. For each $E \in \mathcal{B}_X$ we define $b_E : D \to \mathbb{R}$ by $b_E F = F(E)$. For each $b_E$ we have that $b_E\vartheta : \mathcal{M} \to \mathbb{R}$ is pathwise differentiable at $P_F$ relative to $S(P_F)$ with efficient influence function given by:*

$$\widetilde{I}(F,E) = A_F I_F^{-1}\left(I_E - F(E)\right) \in L_0^2(P_F). \tag{3.10}$$

*In other words,*

$$\frac{1}{\epsilon}b_E\vartheta(P_{F_{\epsilon,h}}) - b_E\vartheta(P_F) = \langle\widetilde{I}(F,E), A_F(h)\rangle_{P_F} \tag{3.11}$$

*If only assumption 1* holds and $I_E - F(E)$ lies in the range of $I_F$, then the same statements hold.*

For proving efficiency we apply theorem 2.4.

The model $\mathcal{M}$ is convex and $F \to P_F$ is linear. Theorem 2.3 says now that we have the following identity;

$$F_1(E) - F_0(E) = -\int \widetilde{I}(F_1,E)dP_{F_0},$$

for all $F_1$ with $F_0 \ll F_1$ and $dF_0/dF_1 \in L_0^2(F_1)$. We want to apply this identity to $F_1 = F_n$. Usually $F_n$ does not dominate $F$ so that this identity cannot be directly applied. However, notice that the identity holds in particular for $F_1 = F_n(\alpha) \equiv (1-\alpha)F_n + \alpha F$ for any $\alpha \in (0,1]$. Hence if $\widetilde{I}(F_n(\alpha),E)$ converges

to $\widetilde{I}(F_n, E)$ in $L^1(P_{F,G})$ for $\alpha \to 0$, then the identity also holds for $F_n$. Since $F_n(\alpha)$ converges to $F_n$ w.r.t. each norm this is a weak continuity condition in $F$ on the efficient influence function, which in particular follows from our verification of a stronger $\rho_{P_0}$-consistency property below.

If $\widetilde{I}(F_n, E)$ is a score of $P_{F_{n,\epsilon,h}}$ for a certain one dimensional line $F_{n,\epsilon,h}$ through $F_n$ with score $h$ with finite supremum norm, then by lemma 3.1 we have the so called *efficient score equation*:

$$P_n \widetilde{I}(F_n, E) = 0.$$

By (3.10), for this it suffices to show that $I_{F_n}^{-1}(I_E)$ has finite supremum norm on $\mathcal{X}$, which is proved by lemma 3.5 in the next subsection.

Then combining the last two identities provides us with the identity: for each $E \in \mathcal{B}_X$

$$(F_n - F_0)(E) = \int \widetilde{I}(F_n, E) d(P_n - P_{F_0}).$$

Let $B \equiv \{b_E : E \in \mathcal{E}\}$. Suppose now that with probability tending to 1 $\widetilde{I}(F_n, E)$ lies in a $P$-Donsker class for all $E \in \mathcal{E}$, This $P$-Donsker class condition will be studied in section 5.2 and sufficient conditions are given. This provides us with $\|F_n - F_0\|_{\mathcal{E}} = O_P(1/\sqrt{n})$. The $\rho_{P_0}$-consistency condition requires to verify: if

$$\sup_{E \in \mathcal{E}} \|\widetilde{I}(F_n, E) - \widetilde{I}(F_0, E)\|_{P_{F_0}} \to 0 \text{ in probability,}$$

then $\sup_{E \in \mathcal{E}} (P_n - P_0)(\widetilde{I}(F_n, E) - \widetilde{I}(F_0, E)) = o_P(1/\sqrt{n})$. The latter provides us with supnorm-efficiency of $F_n$.

In the next subsections we prove, or rather give verifiable conditions for, the $P_0$-Donsker condition, the supremum norm invertibility of the information operator and the $\rho_{P_0}$-consistency condition. The conditions will be verified for the examples. Finally, we summarize our conclusions in the final efficiency theorems in section 5.

### 3.4.1 Supremum norm invertibility of the information operator.

We will now write down the score operator. Recall that by assumption 2 the distribution of $X$ given $Y = y$, $y \in A_2$, has a density $q_1'(y \mid x)/p_F'(y)$ w.r.t. $F$ which only lives on $B(y_1)$. So we have:

$$A_F(h)(Y) = E_F(h(X) \mid Y) I(Y \in A_1) + E_F(h(X) \mid Y) I(Y \in A_2)$$

$$= h(B(Y)_1)I(Y \in A_1) + \frac{\int_{B(Y)_1} h(x)q_1'(y \mid x)F(dx)}{p_F'(Y)} I(Y \in A_2).$$

Notice that $E(h(B(Y)_1)I_{A_1}(Y) \mid X = x) = Q(B(y(x))_2 \mid x)\, h(x)$ and by assumption 2* $E(V(Y)I_{A_2}(Y) \mid X = x) = \int_{A_2} V(y)dP_Q(y \mid x) = \int_{A_2} V(y)q_1(y \mid x)d\mu_3(y)$. So the information operator is given by:

$$I_F(h)(x) = Q(B(y(x))_2 \mid x)h(x) + \int_{A_2} \frac{\int h(u)q_1'(y \mid u)F(du)}{\int q_1'(y \mid u)F(du)} q_1(y \mid x)d\mu_3(y).$$

Consider the equation $I_F(h)(x) = f(x)$ for some pointwise well defined $f$ with finite supremum norm. Define

$$v_1(x) \equiv Q(B(y(x)_2 \mid x) \tag{3.12}$$

and notice that assumption 1* tells us that $v_1(x) > 0$ for $F$-all $x$. Then the equation $I_F(h)(x) = f(x)$ is equivalent with the following equation:

$$h(x) = \frac{1}{v_1(x)} \left\{ f(x) - \int_{A_2} \frac{\int h(u)q_1'(y \mid u)F(du)}{p_F'(y)} dP_Q(y \mid x) \right\}. \tag{3.13}$$

For the moment denote the right-hand side by $C_F(h, f)(x)$: i.e. we consider the equation $h(x) = C_F(h, f)(x)$. If we assume that $f$ lies in the range of $I_F$ (so in particular if assumption 1* holds), then we know by lemma 3.3 that there exists a $h' \in L^2(F)$, which is unique in $L^2(F)$, with $\|I_F(h') - f\|_F = 0$: i.e. $\|h' - C_F(h', f)\|_F = 0$. Notice that if $\|h - g\|_F = 0$, then for each $x$ $C_F(h - g, f)(x) = 0$. So even if $h'$ is only uniquely determined in $L^2(F)$, then $C_F(h', f)(x)$ is uniquely determined for each $x$. Now, we can define $h(x) \equiv C_F(h', f)(x)$. Then $\|h - h'\|_F = \|C_F(h', f) - h'\|_F = 0$. So in this way we have found a solution $h$ of (3.13) which holds for each $x$ instead of only in $L^2(F)$ sense.

By assumption 2 we have that $p_F' > \delta > 0$. So if $\sup_{Y \in A_2} \|u \to q_1'(y \mid u)\|_F < \infty$, then it follows trivially (as shown below) by the Cauchy-Schwarz inequality that $\|h\|_\infty < \infty$. Moreover, $I_F(h) = 0$ implies $\|h\|_F = 0$ and that implies (see (3.13)) that $h = 0$ in supnorm. So we have now shown that $I_F$ is 1-1 and onto in supnorm sense. From now on, if we talk about $I_F^{-1}(f)$, we mean this pointwise well defined solution.

Let $(B(K), \|\cdot\|_\infty)$ be the Banach space of functions on $K$, where $K$ is the support of $F$, with finite supremum norm $\|\cdot\|_\infty$. We have shown:

**Lemma 3.5** *Let assumption 1 and 2 hold and suppose that* $\sup_{Y \in A_2} \|u \to q_1'(y \mid u)\|_F < \infty$. *Then* $I_F : (B(K), \|\cdot\|_\infty) \to (B(K), \|\cdot\|_\infty)$ *is 1-1 and onto.*

For the Donsker class condition we have to consider the solution $h_E^n$ of $I_{F_n}(h_E^n) = I_E$ for $E \in \mathcal{E}$. We will show that $\|h_E^n\|_\infty < M\|I_E\|_\infty$ for some $M < \infty$. So we want a uniform (in $n$) bound on the norm of the mapping $I_{F_n}^{-1}$ w.r.t. the supremum norm.

For this purpose we consider $I_{F_n}(h_n) = f$. The approach to be followed is to bound the right-hand side of (3.13) with $F = F_n$ in the supremum norm of $f$ and $\|h^n\|_{F_n}$, where we can use that the latter is uniformly bounded by the $L^2(F_n)$ norm of $f$, by lemma 3.3.

Firstly, we need to bound $p'_{F_n}$ away from zero. The following assumption will take care of this.

**Assumption 3** Let $P_c(B) = P(X \in B, Y \in A_1)$ be the distribution of the completely observed $X_i$ and $P_{nc}$ be the empirical distribution function of $P_c$. Assume that

$$\sup_{y \in A_2} \left| \int q_1'(y \mid x) d(P_{nc} - P_c)(x) \right| \to 0 \text{ in probability.}$$

This is equivalent to saying that $\{x \to q_1'(y \mid x) : y \in A_2\}$ is a $P_c$-Glivenko-Cantelli class.

**Lemma 3.6** *If assumptions 1,2 and 3 hold, then uniformly in* $y \in A_2$ $p'_{F_n}(y) > \delta_1 > 0$ *for some* $\delta_1 > 0$ *with probability tending to 1.*

**Proof.** We have by lemma 3.1

$$p'_{F_n}(y) = \int q_1'(y \mid x) dF_n(x) \geq \int q_1'(y \mid x) dP_{nc}(x).$$

Assumption 3 tells us that this converges uniformly in $y \in A_2$ to $\int q_1'(y \mid x) dP_c(x)$ in probability. We have by assumption 1: $dP_c(x) = F(dx)P(Y \in A_1 \mid X = x) \geq \delta F(dx)$. By assumption 2 we have: $\int q_1'(y \mid x) dF(x) = p'_F(y) > \delta > 0$. □

By lemma 3.6 we have that the denominators $p'_{F_n}$ in (3.13) are uniformly bounded away from zero for $n$ large enough.

Furthermore, we want to bound $\mid \int h^n(u) q_1'(Y \mid u) F_n(du) \mid$ by $M\|h^n\|_{F_n}$ for an $M < \infty$ which is independently of $Y \in A_2$ and $n$.

By the Cauchy-Schwarz inequality, a sufficient condition for this is given by:

**Assumption 4**

$$\int_{Y \in A_2} \|q_1'(Y \mid u)\|_{F_n} dP_Q(Y \mid x) < M < \infty.$$

Then by bounding $1/v_1$ and the denominator by $1/\delta$ we have:

$$h^n(x) = C_n(h^n, f)(x) \leq \frac{1}{\delta^2}\left(\|f\|_\infty + \|h^n\|_{F_n} M\right).$$

By lemma 3.3 (the uniform in $n$ bounded invertibility of $I_{F_n}$ w.r.t. $L^2(F_n)$) we have

$$\|h^n\|_{F_n} \leq M_1\|f\|_{F_n} \leq M\|f\|_\infty \text{ for some } M < \infty.$$

Consequently, we have $\|h^n\|_\infty = \|I_{F_n}^{-1}(f)\|_\infty \leq M\|f\|_\infty$ for some $M < \infty$. This proves the following lemma.

**Lemma 3.7** *If assumptions 1, 2, 3 and 4 hold, then* $I_{F_n} : (B(K), \|\cdot\|_\infty) \to (B(K), \|\cdot\|_\infty)$ *is onto and has bounded inverse. Moreover we have that* $\|I_{F_n}^{-1}(h)\|_\infty \leq M\|h\|_\infty$ *for certain* $M < \infty$ *which does not depend on* $F_n$ *and* $h$.

### 3.4.2 Weak assumption approach.

In the case that assumption 1 and 2 needs to be weakened one can follow the same approach without using that $v_1$ and $p_F$ are bounded away from zero. Realize that our only goal is to prove that $\|h^n\|_\infty < M$ for some $M < \infty$ independent of $n$, with probability tending to 1. One can replace assumption 4 by the following weak version of it:

**Assumption 4*:** Assume that $f^*(x) \equiv f(x)/v_1(x)$ has finite supnorm over the support of $F$. Assume that $\|I_{F_n}^{-1}(f)\|_{F_n} \leq M$ with probability tending to 1, where $M < \infty$ does not depend on $n$. Finally, assume

$$\frac{1}{v_1(x)}\int_{A_2} \frac{\sqrt{\int q_1'(y \mid u)^2 dF_n(u)}}{p_{F_n}(y)} dP_Q(y \mid x) < M < \infty,$$

with probability tending to 1.

The first two assumptions provide us by Cauchy-Schwarz, as above, with the following bound

$$\|h^n\|_\infty \leq \|f^*\|_\infty + \frac{M}{v_1(x)}\int_{A_2} \frac{\sqrt{\int q_1'(y \mid u)^2 dF_n(u)}}{p_{F_n}(y)} dP_Q(y \mid x).$$

The third assumption just says that this is bounded uniformly in $n$ with probability tending to 1. One can use that $dF_n(x) \geq dP_{nc}(x)$ (and hence $p_{F_n}(y) \geq \int q_1'(y \mid x) dP_{nc}(x)$) and apply refined empirical process results in a specific analysis.

### 3.4.3 The P-Donsker class condition.

Lemma 3.7 tells us that for each $E \in \mathcal{E}$ $h_E^n \equiv I_{F_n}^{-1}(I_E)$ exists pointwise, $\sup_{E,n} \|h_E^n\|_\infty < \infty$ and solves

$$h_E^n(x) = C_n(h_E^n, I_E)(x), \tag{3.14}$$

where

$$C_n(h_E^n, I_E)(x) = \frac{1}{v_1(x)} \left( I_E(x) - \int_{A_2} A_{F_n}(h_E^n)(y) dP(y \mid X = x) \right).$$

Here $dP(y \mid x)$ can be replaced by $q_1(y \mid x) d\mu_3(y)$. Define

$$\mathcal{F}_{1n} \equiv \left\{ \frac{1}{v_1} \left( I_E(\cdot) - \int_{A_2} A_{F_n}(g)(y) dP(y \mid X = \cdot) \right) : \|g\|_\infty < 1, E \in \mathcal{E} \right\}.$$

Now, it follows that for proving the $P_0$-Donsker-class condition it suffices to have:

**Assumption 5** Assume that there exists a $P_0$-Donsker class $\mathcal{F} \subset L^2(P_0)$ so that

$A_{F_n}(\mathcal{F}_{1n})) \subset \mathcal{F}$ with probability tending to 1.

**Lemma 3.8** *If assumption 1-5 or assumption 1*,2*,3,4*,5 hold, then* $\{\widetilde{I}(F_n, E) : E \in \mathcal{E}\} \subset \mathcal{F}$ *with probability tending to* 1.

As mentioned in the general efficiency proof (begin of section 5) the assumptions of lemma 3.8 provide us also with $\|F_n - F_0\|_{\mathcal{E}} = O_P(1/\sqrt{n})$. For efficiency it remains to verify the $\rho_{P_0}$-consistency condition.

Firstly, we work out assumption 5 for the general practical relevant case that $\mathcal{X} = \mathbb{R}^k$.

**Conditions for assumption 5 in the case of multivariate observations.** Let $\mathcal{X} = \mathbb{R}^k$ for certain $k \in \mathbb{N}$ and let $\mathcal{E} \equiv \{(0,t] : t \in \text{supp}(F_0)\}$. We refer to the following results stated in chapter 1. A real valued function on $[0,\tau] \subset \mathbb{R}^k$ is called to be of bounded uniform sectional variation if the variations of all sections ($s \to f(s,t)$ is a section of the bivariate function f, etc.) and of the function itself is uniformly (in all sections) bounded. The corresponding norm is denoted with $\|\cdot\|_v^*$. In chapter 1 we proved that the class of functions with uniform sectional variation smaller than $M < \infty$ is a Donsker class. We also stated that if $f > \delta > 0$, then $\|1/f\|_v^* \leq M\|f\|_v^*$ for some $M < \infty$ which does not depend on $f$ (Gill, 1993). We will now give properties of $\Phi$ guaranteeing

that $h_t^n \equiv I_{F_n}^{-1}(I_{(0,t]})$ is of bounded uniform sectional variation uniformly in $n$ and thereby for any reasonable score operator the efficient influence function $\widetilde{I}(F_n, t) = A_{F_n}(h_t^n)$ will lie in a fixed Donsker class.

The next conditions will be denoted with 5.1, 5.2 and 5.3, where 5 stands for assumption 5. Recall from assumption 2 that for $y \in A_2$ we defined $q_1(y \mid x) \equiv dP(\cdot \mid X = x)/d\mu_3$.

**Condition 5.1** Assume

$$\int_{A_2} \|x \mapsto q_1(y \mid x)\|_v^* d\mu_3(y) < \infty.$$

**Condition 5.2**

$$\|v_1\|_v^* < \infty.$$

Let $f : \mathbb{R}^k \to \mathbb{R}$ be any function with $\|f\|_v^* < \infty$. Condition 5.2 tells us that $\|f/v_1\|_v^* < \infty$. Using this, the uniform bound on $A_{F_n}(h^n)$ and condition 5.1 it follows that $\|C_n(h^n, f)(\cdot)\|_v^* < M\|f\|_v^*$, uniformly in $n$. Consequently, (3.14) and $\|I_{(0,t]}\|_v^* = 2^k$ tells us that $\|h_t^n\|_v^* < M$, as required.

Define $(B(K), \|\cdot\|_v^*)$ as the Banach space of functions on $K$, the support of $F_0$, which are of bounded uniform sectional variation endowed with the uniform sectional variation norm. We showed that:

**Lemma 3.9** *Consider the case where $\mathcal{X} = \mathbb{R}^k$. If assumptions 1–4, 5.1 and 5.2 hold, then $I_F : (B(K), \|\cdot\|_\infty) \to (B(K), \|\cdot\|_\infty)$ and $I_F : (B(K), \|\cdot\|_v^*) \to (B(K), \|\cdot\|_v^*)$ are onto and have a bounded inverse with an operator norm which is bounded uniformly in $F \in \mathcal{F}$.*

Finally, assume that:

**Condition 5.3** Let $\mathcal{G}_1(M) \equiv \{f : \|f\|_v^* < M\}$. Assume that there exists a $P_0$-Donsker class $\mathcal{G} \subset L^2(P_0)$ and $M < \infty$ such that

$$A_{F_n}(\mathcal{G}_1(M))) \subset \mathcal{G} \text{ with probability tending to 1.}$$

Because $\mathcal{G}_1(M)$ is a Donsker class, condition 5.3. is a rather weak assumption, which will hold in most practical examples. We can now state the following lemma:

**Lemma 3.10** *Consider the case where $\mathcal{X} = \mathbb{R}^k$. If assumptions 1–4, 5.1, 5.2 and 5.3 hold, then there exists a $P_0$ Donsker class $\mathcal{G} \subset L^2(P_0)$ so that*

$$\widetilde{I}(F_n, t) \in \mathcal{G} \text{ with probability tending to } 1.$$

*In other words, then assumption 5 holds.*

**The weak assumption approach.** Instead of assuming condition 5.1,5.2, we can just assume $\int_{A_2} \|x \to q_1(y \mid x)/v_1(x)\|_v^* d\mu_3(y) < \infty$, which does not necessarily require that $v_1(x) > \delta > 0$.

### 3.4.4 Verification of assumptions 3, 4, 5.1, 5.2 and 5.3 for the examples.

**Example 3.4 (Univariate Censoring).**
**Assumption 3 and 4.** By the independence of $X$ and $C$ we had $q_1'((z,0) \mid x) = dP_0((z,0) \mid x)/dG_0 = I(x > z)$ and hence assumption 3 and 4 are trivially satisfied.
**Condition 5.1.** We have that $q_1((z,0) \mid x) = q_1'((z,0) \mid x)$ and clearly $z \to I(x > z)$ is of bounded variation uniformly in $x$. This proves 5.1.
**Condition 5.2.** $v_1(x) = H(x)$. So 5.2 holds.
**Condition 5.3.**

$$A_F(h)(z,\delta) = \delta h(z) + (1-\delta)\frac{\int_z^\infty h dF}{S(z)}.$$

By lemma 3.6 we know that $S_n(z) > \delta S(z) > \delta_1 > 0$ with probability tending to 1 and therefore the denominator is uniformly bounded away from zero with probability tending to 1. Because the variation norm of a distribution function, as $S_n$, is bounded, it follows that the variation of $A_{F_n}(h)(z,0)$ and $A_{F_n}(I(0,t])(z,1)$ are uniformly (in $n$ and $t$) bounded with probability tending to 1. This proves 5.3.

**Example 3.5 (Double Censoring.)**
**Assumption 3 and 4.** Similar as for univariate censoring.
**Condition 5.1.** Recall $q_1(w,2) \mid x) = (dP_{G_0}((w,2) \mid x)/dF_Y(w) = I(w \leq x)F_Z(0,w)$ and $q_1((w,3) \mid x) = dP_{G_0}((w,3) \mid x)/dF_Z(w) = I(w > x)F_Y([w,\infty))$ and consequently 5.1 holds.
**Condition 5.2.** $v_1(x) = F_{z,y}([0,x] \times [x,\infty])$. So condition 5.2 holds.
**Condition 5.3.**

$$A_F(h)(w,d) = h(w)I(d=1) + I(d=2)\frac{\int_w^\infty h dF_X}{\bar{F}_x(w)} + I(d=3)\frac{\int_0^w h dF_X}{F_X(w)}.$$

The same proof as for the univariate censoring model holds.

**Example 3.6 (IH-Models, Vardi-Zhang.)**
For a general kernel $q(\cdot \mid \cdot)$ the best thing to require is that assumption 3, 4 and 5 have to hold. We will work out what this means for the Vardi-Zhang

model.

**Assumption 3.** For this we need that

$$\sup_{w\in(0,\tau]}\left|\int_w^\infty \frac{d(P_{nc}-P_c)(x)}{x}\right| \to 0 \text{ a.s.}$$

We know that the convergence holds for each $w \in (0,\tau]$. Recall now that a sequence of distribution functions which converges pointwise to a continuous limit is uniformly convergent. Therefore it suffices to show that

$$\int_0^\infty \frac{d(P_{nc}-P_c)(x)}{x} \to 0 \text{ a.s.}$$

By the Glivenko-Cantelli theorem this holds if $P_c(1/x) < \infty$.

**Assumption 4.** Assumption 4 requires that $\int_0^\tau dF_n(x)/x^2 < M$ with probability tending to 1. Assume $F_n(\{X_i\}) \leq MP_{nc}(\{X_i\})$ for $X_i \in [0,\delta_1]$ for certain $\delta_1 > 0$ with probability tending to 1 (this is easy to verify by using the EM-steps, because $X_i \in [0,\delta_1]$ gets mass $1/n$ from itself and mass $O(1/n^2)$ from other observations; we will not get into the details here). Then it suffices to show that

$$\int_0^\tau \frac{P_{nc}(dw)}{w^2} \leq M, \text{ with probability tending to 1.}$$

Again, by the Glivenko-Cantelli theorem it suffices now to assume $P_c(1/w^2) < \infty$.

**Assumption 5.** Recall that $dP_Q(z,0 \mid x)/d\mu_2(z) = (1-\lambda)q(z \mid x)$ and hence Condition 5.1 requires that

$$\int \|x \mapsto q(z \mid x)\|_v^* d\mu_2(z) < \infty.$$

Consider the Vardi-Zhang kernel: $x \to I(z \leq x)/x$. Because of the singularity at zero 5.1 does not hold in this case. Therefore, we will verify assumption 5 directly. Then it follows that it suffices to show that the following statement holds with probability tending to 1:

$$\int \frac{\int g(u)q(c \mid u)dF_n(u)}{\int q(c \mid u)dF_n(u)} q(c \mid x)dc$$

is of bounded variation uniformly in $\|g\|_\infty < 1$. Substitute the Vardi-Zhang kernel. Then we have

$$\frac{1}{x}\int_0^x \frac{\int_c^\infty g(u)\frac{dF_n(u)}{u}}{\int_c^\infty \frac{dF_n(u)}{u}} dc.$$

Define $f_n(c) \equiv \left(\int_c^\infty g(u)dF_n(u)/u\right) / \left(\int_c^\infty dF_n(u)/u\right)$. Then we have to consider the derivative of $f_{1n}(x) \equiv \frac{1}{x}\int_0^x f_n(c)dc$. We have

$$f'_{1n}(x) = \frac{1}{x}\int_0^x \frac{f_n(x) - f_n(c)}{x}dc.$$

However,

$$\begin{aligned}\frac{1}{x}\int_0^x \frac{f_n(x) - f_n(c)}{x}dc &= \frac{1}{x}\int_0^x \frac{\int_c^x g dF_n(u)/u}{x}dc \\ &\leq \|g\|_\infty \frac{1}{x}\int_0^x \frac{\int_c^x dF_n(u)/u}{x}dc.\end{aligned}$$

Therefore, it suffices to show that $1/x\int_0^x dF_n(u)/u$ is bounded with probability tending to 1. Now, $1/x\int_0^x dF_n(u)/u \leq \int_0^x dF_n(u)/u^2$. As already shown above this is bounded with probability tending to 1 if $P_c(1/x^2) < \infty$.

In the cases where we have to deal with singularities, as in the Vardi-Zhang model, one should verify assumption 5 directly, instead of verifying the sufficient conditions 5.1, 5.2, 5.3.

### 3.4.5 The P-consistency condition.

We will formulate this condition as an assumption, because it is something which can be straightforwardly verified for any application.

**Assumption 6** Suppose that

$$\sup_{E\in\mathcal{E}} \|\widetilde{I}(F_n, E) - \widetilde{I}(F, E)\|_{P_F} \to 0 \text{ in probability,}$$

where we can use that $\|F_n - F\|_{\mathcal{E}} \to 0$ in probability.

Denote $h_{nE} = I_{F_n}^{-1}(I_E)$ and $h_E = I_F^{-1}(I_E)$, as defined in the invertibility section 3.5.2. We have

$$\begin{aligned}A_{F_n}I_{F_n}^{-1}(I_E) - A_F I_F^{-1}(I_E) &= (A_{F_n} - A_F)I_F^{-1}(I_E) + A_{F_n}I_{F_n}^{-1}(I_{F_n} - I_F)I_F^{-1}(I_E) \\ &= (A_{F_n} - A_F)(h_E) + A_{F_n}I_{F_n}^{-1}(I_{F_n} - I_F)(h_E). \quad (3.15)\end{aligned}$$

For assumption 6 we need to show that the $P_F$-norm of these terms converges to zero in probability, uniformly in $E \in \mathcal{E}$. Recall the score operator $A_F$ and that the denominator $p'_{F_n}$ is uniformly bounded away from zero by lemma 3.6. By telescoping the first term, it is written as a sum of two differences. The first difference is given by:

$$\int \left(\frac{I_{A_2}(Y)}{p'_F(Y)}\int_{B(Y)_1} h_E(x)q'_1(Y \mid x)d(F_n - F)(x)\right)^2 dP_F(Y) \qquad (3.16)$$

and the second difference is similar and shown to converge to zero in the same way. Standard techniques, like the Cauchy-Schwarz inequality, dominated convergence theorem, integration by parts and invertibility of the information operator w.r.t. supremum norm and $L^2$ norm (lemma 3.9) lead under mild conditions to a straightforward proof of the convergence of (3.16) to zero. Also the other terms are dealt in this way.

Consider the case where $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{E} = \{(0,t] : t \in [0,\tau]\}$. For proving assumption 6 we need to prove convergence of terms

$\int h_t q_1' d(F_n - F)$, which can be bounded by integration by parts (lemma 1.3) by $C\|F_n - F\|_\infty \|h_t q_1'\|_v^*$. By lemma 3.9 we know that $h_t$ is of bounded uniform sectional variation. Consequently, the following condition is sufficient, but certainly not necessary, for proving assumption 6.

**Condition 6.1.** There exists an $M < \infty$ so that for $P_F$ almost each $Y \in A_2$ we have that $\|x \to q_1'(Y \mid x)\|_v^* < M < \infty$.

We can state this as a lemma:

**Lemma 3.11** *For the case that $\mathcal{X} = \mathbb{R}^k$ we have that under assumptions 1–4, 5.1, 5.2, 5.3 and assumption 6.1 the following holds:*

$$\sup_{t\in[0,\tau]} \|\widetilde{I}(F_n,t) - \widetilde{I}(F_0,t)\|_{P_{F_0}} \to 0 \ a.s.$$

Because in the univariate censoring model and double censoring model $q_1'$ are indicators, condition 6.1 holds trivially in these models. Condition 6.1 is in general not true for the Vardi-Zhang kernel. However, a direct proof of assumption 6 following the general proof above works as follows: Consider the term (3.16). It is given by:

$$\int \left( \int_c^\infty \frac{h_t(x)}{x} d(F_n - F_0)(x) \right)^2 q(c)dc. \tag{3.17}$$

We want to show that this term converges to zero uniformly in $t$. By lemma 3.9 we know that $h_t(x)$ is of bounded variation and bounded in supremum norm uniformly in $t$. Therefore by integration by parts we can bound $f_n^t(c) - f^t(c) \equiv \int_c^\tau \frac{h_t(x)}{x} d(F_n - F_0)(x)$ by $\|F_n - F_0\|_\infty$ times a constant which involves the variation of $h_t$ and $1/x$ on $(c,\tau)$. This converges to zero for each $c > 0$, uniformly in $t$. For application of the dominated convergence theorem it remains to verify that $f_n^t + f^t$ is bounded by a constant, uniformly in $t$. $h_t$ is bounded uniformly in $t$ and we already showed in the preceding example that $\int_0^\infty dF_n(x)/x < M$ with probability tending to 1, using the assumption

that $P_c(1/x) < \infty$. This proves the convergence of (3.17). The other terms appearing in the general proof are dealt similarly.

### 3.4.6 Identity condition.

Let $F_n(\alpha) = (1-\alpha)F_n + \alpha F$. Rewrite $A_{F_n(\alpha)}I^{-1}_{F_n(\alpha)}(I_E)$ as in (3.15). Since $F_n(\alpha) - F_n = \alpha(F_n - F)$ it is now trivially verified that both terms converge to zero if $\alpha \to 0$, assuming assumption 1* and 2*. For example, term (3.16) above equals $\alpha^2$ times a term which is bounded (here $n$ is just fixed, so we do not need terms to be bounded away from zero uniformly in $n$).

## 3.5 Final theorems and results for the examples.

We proved the following theorem:

**Theorem 3.2** *Let $\mathcal{E}$ be any collection of sets for which assumptions 1-6 or 1*,2*,3,4*,5,6 hold. Then $F_n$ is asymptotically supremum norm (over $\mathcal{E}$) efficient.*

Now, we formulate the less general theorem for the case that $\mathcal{X} = \mathbb{R}^k_{\geq 0}$ and $\mathcal{E} = \{[0,t] : t \in K \subset \text{supp}(F_0)\}$.

**Theorem 3.3** *Under assumptions 1-4, 5.1, 5.2, 5.3, 6.1 we have: $F_n$ is asymptotically supremum norm efficient.*

We already verified the assumptions for our examples. Application of these theorems provides us under the stated assumptions with efficiency of the sieved-NPMLE for the univariate censoring, double censoring, and the IH-models. We will summarize these results below. The results are not new, but they are obtained by straightforward verification of the conditions of the general theorem 3.2 and 3.3 which can be applied to any CAR-missing data model or a missing data model with $G$ known.

**Result 3.1** (Univariate Censoring).
*Let $[0,\tau] \subset \mathbb{R}_{\geq 0}$ be an interval such that $H_0(\tau) > 0$ and $S_0(\tau) > 0$. Then $F_n$ is supremum norm asymptotically efficient on $[0,\tau]$.*

**Result 3.2** (Double Censoring).
*Let $[0,\tau] \subset \mathbb{R}_{\geq 0}$ be an interval such that $F_Z(\tau) = 1$, $S_{Y_1}(\tau) > 0$. Furthermore, assume that there exists a $\delta > 0$ so that $F_Z(\delta) = F_Z(\{0\}) > 0$ and $F_X(\delta_1) > 0$. Then $F_n$ is supremum norm asymptotically efficient on $[0,\tau]$.*

**Result 3.3** (Vardi-Zhang model). *Let* $[0, \tau] \subset \mathbb{R}_{\geq 0}$ *be an interval such that* $\int_\tau^\infty \frac{dF_0(x)}{x} > \delta > 0$. *Moreover assume that* $\int 1/x^2 dF_0(x) < \infty$. *By artificially censoring the observed* $C_i$ *at* $\tau$ *we have a model: if* $x < \tau$, *then* $x \sim F_0$ *and* $C \mid x$ *is uniform* $[0, x]$ *and if* $x \geq \tau$, *then* $C \mid x$ *has a density* $1/x$ *is* $[0, \tau]$ *and it puts mass* $1 - (\tau/x)$ *at* $\tau$. *Now,* $F_n$ *is supremum norm asymptotically efficient on* $[0, \tau]$.

In the general class of IH-models the assumptions 1–7 give clear conditions on the kernel $q(\cdot \mid \cdot)$, which can be easily analyzed for each kernel. We could also state a result involving sufficient conditions on the kernel, but verification of assumption 1–7, just as we did for the Vardi-Zhang model, provides us with the sharpest results.

# Chapter 4

# Efficient Estimation in the Bivariate Censoring Model and Repairing NPMLE.

## 4.1 Introduction.

In this chapter we are concerned with estimation of the bivariate survival function of two dependent survival times. For example, one might be interested in estimation of the bivariate survival function of twins with a certain disease. Suppose that for each twin one observes two calendar times $(U_1, U_2)$ at which the disease started for twin1 and twin2 and that one keeps track of the bivariate survival time $(T_1, T_2)$ of the twin measured from $(U_1, U_2)$ till a given calendar point $t_0$. At $t_0$ one wants to use the available data to estimate the bivariate survival function of $(T_1, T_2)$. In this setting, $T_1$ will be potentially (i.e. if $T_1 > C_1$) right randomly censored at the observed censoring time $C_1 = t_0 - U_1$ and similarly $T_2$ will be potentially right randomly censored at the observed censoring time $C_2 = t_0 - U_2$.

In this chapter we propose an estimator for the *bivariate survival function* of $T = (T_1, T_2)$ based on *bivariate right randomly censored data*, assuming that the censoring times $C = (C_1, C_2)$ are always observed, as in the example above, or assuming that the censoring times are discrete. We prove asymptotic efficiency of this estimator. In the case that the censoring times are not observed for the failures and the censoring times are not discrete, then we propose a simulation of the unobserved censoring variables and conjecture (no proof, but heuristic argument) that our estimator based on these simulated censoring

variables will also be asymptotically efficient.

We found it useful not to use a special notation for vectors in $\mathbb{R}^2$; if we do not mean a vector this will be clear from the context. So if we write $T$ we usually mean $T = (T_1, T_2) \in \mathbb{R}^2_{\geq 0}$ and if we write $\leq$, $\geq$, $<$, $>$ then this should hold componentwise: for example if $x, y \in \mathbb{R}^2$ then $x \leq y \Leftrightarrow x_1 \leq y_1,\ x_2 \leq y_2$. We will write $T_i$, $i = 1, \ldots, n$, as notation for $n$ i.i.d. bivariate survival times with the same distribution as $T$, while we write $T_1$ and $T_2$ for the components of $T$.

Bivariate right randomly censored data can be modelled as follows: $T$ is a positive bivariate lifetime vector with bivariate distribution $F_0$ and survival function $S_0$; $F_0(t) \equiv \Pr(T \leq t)$ and $S_0(t) \equiv \Pr(T > t)$. Let $C$ be a positive bivariate censoring vector with bivariate distribution $G_0$ and survivor function $H_0$; $G_0(t) \equiv \Pr(C \leq t)$ and $H_0(t) \equiv \Pr(C > t)$. Assume that $T$ and $C$ are independent; $(T, C) \in \mathbb{R}^4$ has distribution $F_0 \times G_0$. Let $(T_i, C_i)$, $i = 1, \ldots, n$ be $n$ independent copies of $(T, C)$. We observe the following many to one mapping $\Phi$ of $(T_i, C_i)$:

$$Y_i \equiv \Phi(T_i, C_i) \equiv (T_i \wedge C_i, I(T_i \leq C_i)) \equiv (\tilde{T}_i, D_i),$$

with components given by:

$$\tilde{T}_{ij} = \min\{T_{ij}, C_{ij}\},\ D_{ij} = I(T_{ij} \leq C_{ij}),\ j = 1, 2.$$

In other words, the minimum and indicator are taken componentwise, so that $\tilde{T}_i \in [0, \infty)^2$ and $D_i \in \{0, 1\}^2$ are bivariate vectors. The observations $Y_i$ are elements of $[0, \infty)^2 \times \{0, 1\}^2$ and $Y_i \sim P_{F_0, G_0} = (F_0 \times G_0)\Phi^{-1}$. We are concerned with estimation of $S_0$.

Each observation $Y_i$ tells us that $(T_i, C_i) \in B(Y_i) \equiv \Phi^{-1}(Y_i) \subset \mathbb{R}^2 \times \mathbb{R}^2$, where $B(Y_i) = B(Y_i)_1 \times B(Y_i)_2$ for the projections $B(Y_i)_1 \subset \mathbb{R}^2$ and $B(Y_i)_2 \subset \mathbb{R}^2$ of $B(Y)$ on the $T$ and $C$ space, respectively. The kind of region $B(Y_i)_1$ for $T_i$ (point, vertical half-line, horizontal half-line, quadrant) generates a classification of the observations $Y_i = (\tilde{T}_i, D_i)$ in 4 groups:

*Uncensored.* If $D_i = (1, 1)$, then the observation $Y_i$ is called uncensored, and it tells us that $T_i \in B(Y_i)_1 = \{\tilde{T}_i\}$. So $T_i = \tilde{T}_i$.

*Singly censored.* If $D_i = (0, 1)$ or $D_i = (1, 0)$, then the observation $Y_i$ is called singly censored. If $D_i = (0, 1)$, then it tells us that $T_i \in B(Y_i)_1 = \{(\tilde{T}_{i1}, \infty) \times \{\tilde{T}_{i2}\}\}$ (horizontal half-line), and if $D_i = (1, 0)$ that $T_i \in B(Y_i)_1 = \{\{\tilde{T}_{i1}\} \times (\tilde{T}_{i2}, \infty)\}$ (vertical half-line).

*Doubly censored.* If $D_i = (0,0)$, then the observation $Y_i$ is called doubly censored, and it tells us that $T_i \in B(Y_i)_1 = \{(\widetilde{T}_{i1}, \infty) \times (\widetilde{T}_{i2}, \infty)\}$ (upper quadrant).

The uncensored observations are the *complete* observations and the singly-censored and doubly censored are incomplete observations. An NPMLE solves the self-consistency equation (Efron, 1967, Gill, 1989) and a solution of the self-consistency can be found with the *EM-algorithm* (Dempster, Laird and Rubin, 1977, Turnbull, 1976), which does in fact nothing else than iterating the self-consistency equation. In the EM-algorithm each observation $Y_i$ gets mass $1/n$ which it need to redistribute over $B(Y_i)_1$ in a self-consistent way. The incomplete observations $Y_i$ need to get information from the observed $T_i$ about how to redistribute their mass $1/n$ over $B(Y_i)_1$, and for this purpose they need complete observations in $B(Y_i)_1$; the EM-algorithm listens only to the observations with a region $B(Y_j)_1$ which has an intersection with $B(Y_i)_1$. It is only possible to have uncensored observations in $B(Y_i)_1$ if $F_0\left(B(Y_i)_1\right) > 0$, which is typically not true for the singly-censored observations; if $F_0$ is continuous, then the probability that $T$ falls on a line is zero. Indeed it is well known that the NPMLE for continuous data is not consistent (Tsai, Leurgans and Crowley, 1986).

Many proposals for estimation of the bivariate survival function in the presence of bivariate censored data have been made. Because the usual NPML and self-consistency principle do not lead to a consistent estimator for continuous data, most proposals are explicit estimators based on representations of the bivariate survival function in terms of distribution functions of the data: among them Tsai, Leurgans and Crowley (1986), Dabrowska (1988, 1989), Burke (1988), the so called Volterra estimator of P.J. Bickel (see Dabrowska, 1988), Prentice and Cai (1992a, 1992b).

Prentice and Cai (1992a) proposed a nice estimator which is closely related to Dabrowska's estimator except that this one also uses the Volterra structure of Bickel's suggestion. Dabrowska's multivariate product-limit estimator, based on a very clever representation of a multivariate survival function in terms of its conditional multivariate hazard measure, and the Prentice-Cai estimator have a better practical performance in comparison w.r.t. the Volterra, pathwise estimator and the estimator proposed in Tsai, Leurgans and Crowley (1986) (see Bakker, 1990, Prentice and Cai, 1992b, Pruitt, 1992, and chapter 8 of van der Laan, 1993d). It is expected that Dabrowska's and Prentice-Cai's estimators are certainly better than the other proposed explicit estimators. Besides, these two estimators are smooth functionals of the empirical distri-

butions of the data so that such results as consistency, asymptotic normality, correctness of the bootstrap, consistent estimation of the variance of the influence curve, LIL, all hold by application of the functional delta method: see Gill (1992) and Gill, van der Laan and Wellner (1993) and van der Laan (1990). In Gill, van der Laan and Wellner (1993), here chapter 6, Dabrowska's results about her estimator are reproved and new ones are added by application of the functional delta method and similar results are proved for the Prentice-Cai estimator. Moreover, it is proved that the Dabrowska and Prentice-Cai estimator are efficient in the case that $T_1, T_2, C_1, C_2$ are all independent.

All the estimators proposed above are ad hoc estimators which are not asymptotically efficient (except at some special points $(F, G)$). This is also reflected by the fact that most of these estimators put a non negligible proportion of negative mass to points in the plane (Pruitt, 1991a, Bakker, 1990).

Pruitt (1991b) proposed an interesting implicitly defined estimator which is the solution of an ad hoc modification of the self-consistency equation. Pruitt points out why the original self-consistency equation has a wide class of solutions and his estimator tackles this non-uniqueness problem in a very direct way by estimating conditional densities over the half-lines implied by the singly-censored observations. Uniform consistency, $\sqrt{n}$-weak convergence, and the bootstrap for his normalized estimator is proved in chapter 7 under some smoothness assumptions which are due to the fact that his estimator uses kernel density estimators. However this estimator is not asymptotically efficient (except at some special points) and its practical performance is (somewhat surprisingly) worse, except at the tail where one hardly finds uncensored observations, (as shown in chapter 8 of van der Laan (1993d)) than Dabrowska's and Prentice and Cai's estimators. In the case that the sampling distribution is smooth, Pruitt's estimator appeared (as expected) to improve by using large bandwidths.

As noticed by Pruitt (1991) the inconsistency of the NPMLE is due to the fact that the singly-censored observations imply half-lines for $T$ which do not contain any uncensored observations. Based on this understanding we propose in section 2 to (slightly) interval censor the singly censored observations in the sense that we replace the uncensored component (say) $T_{1i}$ of the singly censored observations by the observation that $T_{1i}$ lies in a small predetermined interval around $T_{1i}$. These intervals are determined by a grid partition $\pi_h$ with a width $h = h_n$. Now, for these interval censored singly censored observations $Y_i^h$ the regions $B(Y_i^h)_1$ are strips which contain with positive probability uncensored observations

The interval censoring of the singly censored observations causes one problem. The joint likelihood for $F$ and $G$ does not factorize anymore in a $F$-term and $G$-term, which is due to the fact that the region for $(T, C)$ implied by the interval censored singly censored observations is not rectangular anymore. This tells us that for computing the NPMLE of $F$ we also need to estimate $G$ by maximizing over $G$. Because of similar reasons as for the NPMLE of $F$ the NPMLE of $G$ will only be good if we do a symmetric reduction (lines should be strips for C as well as for T). In other words, an extra reduction of the data will be necessary. Because the involvement of $G$ in computing the NPMLE $F_n^h$ certainly complicates the analysis and it makes the estimator more computer intensive we decided to choose a reduction of the data which recovers the orthogonality (i.e. factorization of the likelihood), while at the same time, as will appear, not losing asymptotic efficiency. The further reduction is based on the insight that if $G_0$ is purely discrete on $\pi^h$, then $p^h_{F_0,G_0}(\cdot, d)$ factorizes, as shown in section 2. Hence if the actual $G$ is discrete, then by choosing $\pi_h$ (which can be done with probability tending to 1 if the number of observations converges to infinity) so that censoring variables lie on the grid $\pi^h$ we still have factorization of the likelihood. If the actual $G$ is not discrete, but we observe $C_1, \ldots, C_n$, then we can discretize (to the left) these $C_i$'s to $C_i^h$ on $\pi_h$, 2) replace the original $Y_i$'s by $\Phi(T_i, C_i^h)$, and 3) replace the singly-censored observations of $\Phi(T_i, C_i^h)$ by interval singly-censored observations $Y_i^h$. In this way, we constructed new observations $Y_i^h$ for which the density factorizes in a $F$ and $G$ part.

This further reduction leads also to a good practical estimator as appears in the simulations in chapter 8 of van der Laan (1993d); its performance for a small value of $h$ is better than Dabrowska's, Prentice and Cai's and Pruitt's estimator, except at the tail, and under complete independence of $T_1, T_2, C_1, C_2$. We show that if $h_n \to 0$ at a rate slower than $n^{-1/18}$, then the estimator is asymptotically efficient and if $h$ is fixed, then one still has an asymptotically normal estimator with an asymptotic variance arbitrarily close (small $h$) to the asymptotic optimal variance. Our derived lower bound is purely of theoretical value since it shows the existence of rates $h = h_n$ for which the estimator is efficient, but quicker rates will also provide efficient estimators. Obtaining theoretical insight about the precise rate at which $h_n$ should converge to zero if $n \to \infty$ is very hard and not very useful because constants are not available. Simulations show that if $n = 200$, and the range of the observations is transformed back to $[0, 1] \times [0, 1]$, then choosing the width of the strips equal to $h = 0.02$ gives a very good estimator; so a few observations in each strip

is already effective. The estimator gets essentially worse if we increase $h$ independent of the smoothness of $(F, G)$. This bandwidth-behavior is explained as follows. A large $h$ means a large reduction of the data and hence an increase in asymptotic variance. On the other hand, we needed a $h > 0$ so that the EM-algorithm is able to use the uncensored observations in the strips around the singly-censored half-lines for obtaining a redistribution of mass $1/n$ over the half-lines. However, our primary interest is not the distribution over the half-line, but the survival function itself (which integrates over the distributions over the half-lines), which explains that a smaller bandwidth than the one advised by density estimation literature will suffice. In practice, a sensible method for programming a sensible grid $\pi^h$ would be to set the width for the horizontal axis equal to a fixed proportion of the cross-validated bandwidth $h_1^*$ using the observed $T_{1i}$'s and similarly compute the vertical width.

If we do not observe $C_i$, then we can draw a $C_i'$ from a conditional distribution of $C$, given $C \in B(Y_i)_2$, and consider these simulated $C_i'$ as the observed $C_i$'s above. For example, if we observe that $C_{1i} \in (T_{1i}, \infty)$ we set $C_{1i}' = T_{1i} + U_i$, where $U_i$ is a realization from a known distribution on $(0, \tau]$. Then $Y_i' = \Phi(T_i, C_i') = Y_i$, but we now observe $C_i'$. $C_i'$, $i = 1, \ldots, n$, are still i.i.d, but $C_i'$ depends on $T_i$ only through $Y_i$. However, if the density of $C$, given $T = t$, depends only on $T$ through $Y = \Phi(C, T)$, then the censoring mechanism satisfies *coarsened at random* (see Heitjan, Rubin, 1991) which implies that the density of $Y$ still factorizes, where the $F$ part of the density of $Y'$ is still the same as the $F$ part of the density of $Y$, i.e. where $C$ and $T$ are independent. Consequently, we have that the efficient influence function for estimating $F$ based on $Y_i'$ equals the efficient influence function for estimating $F$ based on $Y_i$. Hence, if we construct an estimator of $F$ based on $(C_i', Y_i')$ which is efficient, then it is also efficient for the original data $Y_i$. In other words, without any loss we arranged that we have available a set of observed $C_i'$'s. However, because of the dependence between $C'$ and $T$ the likelihood does not factorize anymore for the data $\Phi(T, C_h')$ based on the discretized $C_h'$ so that our proposed estimator is not a NPMLE for the interval censored $\Phi(T, C_h')$ and hence has a bias. On the other hand, we let $h$ converge to zero when the numbers of observations converge to infinity so that this bias converges to zero. Therefore, we conjecture (no proof) that our estimator based on these simulated $C'$ is asymptotically efficient if $h = h_n$ converges to zero at an appropriate rate (not too slow and not too quick). In the sequel it will be assumed that the $C_i$'s are observed or that $G_0$ is discrete.

We will call the MLE based on a reduction, or call it a slight transformation,

of the data a "Sequence of Reductions"-MLE and will abbreviate it with SOR-MLE. It is a general way to repair the real NPMLE in problems where the real NPMLE does not work. If one understands why the usual NPMLE does not work, then one can hope to find a natural choice for the transformation of the data. Moreover, if we do not lose the identifiability, we have for a *fixed* transformation consistency, asymptotic normality and efficiency of the NPMLE among estimators based on the transformed data; while we obtain efficiency by letting amount of reduction of the data converge to zero slowly enough if $n$ converges to infinity.

In the next section we will define, in detail, the SOR-MLE for the bivariate censoring model. In section 3 we will give an outline of the efficiency proof, which is based on an *identity* for the SOR-MLE which holds in general for convex models which are linear in the parameter (van der Laan, 1993a). This identity lies a direct link between efficiency of the SOR-MLE and properties of the efficient influence function corresponding with the data $Y_h$. In section 4 we prove the ingredients of this general proof. The crucial lemmas of this section are proved in section 6. We summarize the results in section 5.

## 4.2 SOR-MLE for the bivariate censoring model.

Our original data is given by:

$$(\tilde{T}_i, D_i) = \Phi(T_i, C_i) \sim P_{F_0,G_0}(\cdot,\cdot),\ i = 1,\ldots,n.$$

Let $P_{11}(\cdot) = P_{F_0,G_0}(T \leq \cdot, D = (1,1))$ be the subdistribution of the (doubly) uncensored observations and similarly let $P_{01}$, $P_{10}$ and $P_{00}$ be the subdistributions corresponding with $D = (0,1)$, $D = (1,0)$ and $D = (0,0)$, respectively. Then

$$\begin{aligned} P_{F_0,G_0}(\cdot, D = d) &= P_{11}(\cdot)I(d = (1,1)) + P_{01}(\cdot)I(d = (0,1)) \\ &\quad + P_{10}(\cdot)I(d = (1,0)) + P_{00}(\cdot)I(d = (0,0)), \end{aligned} \tag{4.1}$$

Let $f_0 \equiv dF_0/d\mu$ for some finite measure $\mu$ which dominates $F_0$. Similarly, let $G_0 \ll \nu$ with density $g_0$. $S_0(x_1,\cdot)$ generates a measure on $\mathbb{R}_{\geq 0}$. This measure is absolutely continuous w.r.t. $\mu((x_1,\infty),\cdot)$; the marginal of the measure $\mu$ restricted to $(x_1,\infty) \times \mathbb{R}_{\geq 0}$. Now, we define $S_{02}(x_1,x_2) \equiv -S_0(x_1,dx_2)/\mu((x_1,\infty),dx_2)$ as the Radon-Nykodim derivative and similarly we define $S_{01}(x_1,x_2) \equiv -S_0(dx_1,x_2)/\mu(dx_1,(x_2,\infty))$, $H_{01}(x_1,x_2) \equiv -H_0(dx_1,x_2)/\nu(dx_1,(x_2,\infty))$

and $H_{02}(x_1, x_2) \equiv -H_0(x_1, dx_2)/\nu((x_1, \infty), dx_2)$. Then the density $p_{F_0,G_0}$ of $P_{F_0,G_0}$ w.r.t. $(\mu \times \nu)\Phi^{-1}$ is given by

$$\begin{aligned} p_{F_0,G_0}(x,d) &= f_0(x)H_0(x)I(d=(1,1)) + S_{01}(x)H_{02}(x)I(d=(1,0)) \\ &\quad + S_{02}(x)H_{01}(x)I(d=(0,1)) + S_0(x)g_0(x)I(d=(0,0)) \\ &\equiv p_{11}(x)I(d=(1,1)) + p_{10}(x)I(d=(1,0)) \\ &\quad p_{01}(x)I(d=(0,1)) + p_{00}(x)I(d=(0,0)) \\ &= \sum_{\delta \in \{1,0\}^2} p_\delta(x)I(d=\delta). \end{aligned} \tag{4.2}$$

Suppose that we observe $C_i$ and $(\widetilde{T}_i, D_i)$, $i = 1, \ldots, n$. We will transform $(\widetilde{T}_i, D_i)$ and base our NPMLE on the transformed data. The transformation depends on a *grid*. For this purpose let $\pi^h = (u_k, v_l)^h$ be a nested grid in $h = h_n$ of $[0, \tau]$ which depends on a scalar $h = h_n$ in the following way: $\epsilon h_n < u_{k+1} - u_k < M h_n$, where $\epsilon$ and $M$ are independent of $n, k$, and similarly for $v_{l+1} - v_l$. With nested we mean that the grid points of $\pi_{h_n}$ are a subset of the grid-points of $\pi_{h_{n+m}}$ (we use this in order to make martingale arguments work for conditional expectations, given increasing sigma-fields) In other words, the grid must have a width between $\epsilon h_n$ and $M h_n$. This tells us that the grid $\pi^h$ has (in order of magnitude) $1/h_n^2$ points $(u_k, v_l)$. Let $R_{k,l} \equiv (u_k, u_{k+1}] \times (v_l, v_{l+1}]$.

Move each $C_i$ to the left lower corner $(u_k, v_l)$ of the rectangle $R_{k,l}$ of $\pi^h$ which contains $C_i$. Denote these discretized $C_i$ with $C_i^h$. Then $C_i^h \sim G_h$ where $G_h$ is the step function with jumps on $\pi^h$ corresponding with $G_0$:

$$P(C^h = (u_k, v_l)) = \int_{R_{k,l}} dG_0(c).$$

Consider now the $n$ i.i.d. observations

$$Y_i(T_i, C_i^h) = \Phi(T_i, C_i^h) \sim P_{F_0,G_h}.$$

Notice that we are able to observe these $Y_i(T_i, C_i^h)$ because for this we only need to know $Y_i(T_i, C_i)$. If $h = h_n$ converges to zero, then one the distribution of $\Phi(T, C^h)$ converges to the distribution of $\Phi(T, C)$.

For convenience we will denote $\Phi(T_i, C_i^h)$ with $Y_i = (\widetilde{T}_i, D_i)$, again, and still use the notation $p_{11}$, $p_{10}$, $p_{01}$ and $p_{00}$, suppressing the dependence on $h$, but we have to realize that all censored $\widetilde{T}_{1i}$ equal $u_k$ for some $k$ and $\widetilde{T}_{2i}$ equal $v_l$ for some $l$. Now, we can define the *reduced data* $(\widetilde{T}_i, D_i)^h$ which we will use for our estimator:

$$Y_i^h = (\widetilde{T}_i, D_i)^h = \Phi^h(T_i, C_i^h) \equiv \mathrm{Id}^h((\widetilde{T}_i, D_i)) = \mathrm{Id}^h(\Phi(T_i, C_i^h)),$$

where $\mathrm{Id}^h$ is a many to one mapping on the data $(\widetilde{T}_i, D_i)$ which is defined as follows.

$$\begin{aligned}
&\mathrm{Id}^h(\widetilde{T}, D) = (\widetilde{T}, D) \text{ if } D = (1,1)\\
&\mathrm{Id}^h(\widetilde{T}, D) = ((u_i, \widetilde{T}_2), D) \text{ for } u_i \text{ s.t. } \widetilde{T}_1 \in (u_i, u_{i+1}], \text{ if } D = (1,0)\\
&\mathrm{Id}^h(\widetilde{T}, D) = ((\widetilde{T}_1, v_j), D) \text{ for } v_j \text{ s.t. } \widetilde{T}_2 \in (v_j, v_{j+1}], \text{ if } D = (0,1)\\
&\mathrm{Id}^h(\widetilde{T}, D) = (\widetilde{T}, D) \text{ if } D = (0,0).
\end{aligned}$$

Notice that $\mathrm{Id}^h$ equals the identity for the uncensored and doubly censored observations and it groups all singly-censored observations $(T_1, C_2, I(T_1 \leq C_1) = 1, I(T_2 \leq C_2) = 0)$ with $T_1 \in (u_k, u_{k+1}]$ to one observation and similarly with the singly-censored observations with $D = (0,1)$. We used the notation $\mathrm{Id}^h$ (Id from Identity) because for $h \to 0$ (in other words, if the partition gets finer) this transformation converges to the identity mapping. We will still call the $Y^h$ with $D = (1,0)$ and $D = (0,1)$ singly censored observations, in spite of the fact that they are really censored singly censored observations. $Y_i^h$ are i.i.d. observations with a distribution which is indexed by the (same as for $Y_i$) parameters $F_0$ and $G_h$.

To be more precise, we have

$$Y^h \sim P^h_{F_0,G_h}(\cdot,\cdot),$$

where

$$\begin{aligned}
P^h_{F_0,G_h}(x, D = d) &= P_{11}(\cdot)I(d = (1,1)) + P^h_{01}(\cdot)I(d = (0,1))\\
&\quad + P^h_{10}(\cdot)I(d = (1,0)) + P_{00}(\cdot)I(d = (0,0)), \qquad (4.3)
\end{aligned}$$

where the density $p^h_{F_0}$ of $P^h_{F_0,G_h}$ w.r.t. $(\mu \times \nu_h)\Phi_h^{-1}$, $\nu_h$ being the counting measure on $\pi_h$, is given by:

$$\begin{aligned}
p_{11}(y_1, y_2) &= f_0(y_1, y_2)H_h(y_1, y_2)\\
p_{00}(v_k, v_l) &= S_0(v_k, v_l)g_h(v_k, v_l),
\end{aligned}$$

and

$$\begin{aligned}
p^h_{01}(v_k, v_l) &= \int_{(v_l, v_{l+1}]} p_{01}(v_k, y_2)\mu((v_k, \infty), dy_2)\\
&= \int_{(v_l, v_{l+1}]} S_{02}(v_k, y_2)H_{01}(v_k, v_l)\mu((v_k, \infty), dy_2)\\
&= F_0((v_k, \infty), (v_l, v_{l+1}])H_{01}(v_k, v_l).
\end{aligned}$$

Similarly, $p^h_{10}(u_k, y_2) = S_{01}((v_k, v_{k+1}], v_l)H_{02}(v_k, v_l)$. Notice that $p^h_0(\cdot, d)$, $d \neq (1,1)$, is discrete on $\pi_h$. The independence between $C_h$ and $T$ and the fact

that $C_h$ is discrete on $\pi_h$ implied that the density $p^h_{F_0}(\cdot, d)$ also factorized for $d = (1,0)$ and $d = (0,1)$.

Let $P^h_n$ be the empirical distribution function based on $n$ i.i.d. $Y^h_i(T_i, C^h_i) \sim P^h_{F_0,G_h}$, which is the distribution of the data corresponding with $T \sim F_0$, $C \sim G_h$, where $G_h$ is discrete on the grid $\pi^h$, and the singly censored observations are interval censored by $Id^h$ (i.e. halflines are grouped to strips). Let $\{x_1, \ldots, x_{m(n)}\}$ consist of the uncensored $T_i$ and one point of each $B(Y_j)_1$ which does not contain uncensored $T_i$. Let $\mu_n$ be the counting measure on $\{x_1, \ldots, x_{m(n)}\}$. Now, we let $\mathcal{F}(\mu_n)$ be the set of all distributions which are absolutely continuous w.r.t. $\mu_n$.

We define our SOR-MLE $F^h_n$ of $F_0$ which we will analyze;

$$F^h_n = \arg \max_{F \in \mathcal{F}(\mu_n)} \int \log(p^h_{F,G_h}) dP^h_n, \tag{4.4}$$

where the maximum can be determined without knowing $G_h$ by maximizing the term which only depends on $F$. We define $S^h_n$ as the survival function corresponding with $F^h_n$.

### 4.2.1 Existence and uniqueness of the SOR-MLE and EM-equations.

In lemma 3.1 for a general class of missing data models it is proved that the MLE over all $F$ with support $\{x_1, \ldots, x_{m(n)}\}$ exists and is unique, if the following two assumptions hold: $H_0 > \delta > 0$ $F_0$ a.e. and $F_0(B(Y^h_i)_1) > 0$ for all censored $Y^h_i$ ($D = (1,0)$, $D = (0,1)$, $D = (0,0)$). This holds if all data lives on a rectangle $[0, \tau] \subset \mathbb{R}_{\geq 0}$, where $\tau$ is such that $H_0(\tau) > 0$, $S_0(\tau-) > 0$, $F_0(\tau) = 1$, $F_0(T_1 \in [u_i, u_{i+1}], T_2 > \tau_2) > 0$ and $F_0(T_1 > \tau_1, T_2 \in [v_j, v_{j+1}]) > 0$ for all grid points $(u_i, v_j)$. By making all observations $\widetilde{T}_i \in [0, \tau]^c$ uncensored at the projection point on the edge of $[0, \tau]$ we obtain truncated observations with distribution $P^h_{F^\tau_0,G_h}$, where $F^\tau_0$ equals $F_0$ on $[0, \tau)$, but puts all ($= 1$) its mass on $[0, \tau]$. This means that our efficiency result proves efficiency for data reduced to $[0, \tau]$. For obtaining full efficiency we can let $\tau = \tau_n$ converge slowly enough to infinity for $n \to \infty$. In our analysis this will mean an extra singularity of magnitude $1/H(\tau_n)$ and therefore our analysis can be straightforwardly extended to this case.

Let $g \in L^2(F^h_n)$ have finite supnorm. We will use the notation $F(g) = \int g dF$. We have that $dF^h_{n,\epsilon} = (1 + \epsilon(g - F^h_n(g)) dF^h_n$, $\epsilon \in (-\delta, \delta)$, $\delta > 0$ small enough, is a one-dimensional submodel through the MLE $dF^h_n$ and hence by

definition of $F_n^h$

$$\epsilon \to \int \log(p^h_{F^h_{n,\epsilon},G_h}) dP_n^h$$

is maximized at $\epsilon = 0$. Consequently, the derivative of this real valued function on $(-\delta, \delta)$ at $\epsilon = 0$ equals zero so that exchanging integration and differentiation provides us with:

$$P_n^h(A^h_{F_n}(g - F_n^h(g)) = 0 \text{ for all } g \in L^2(F_n^h) \text{ with } \|g\|_\infty < \infty, \tag{4.5}$$

where the so called score operator $A_F^h$ for a distribution function $F$ is given by:

$$A_F^h : L^2(F) \to L^2(P^h_{F,G_h}) : g \mapsto E_F(g(T) \mid Y^h).$$

The form of the score operator follows from the general fact that the score operator in missing data models equals the conditional expectation operator (see Gill, 1989, Bickel, Ritov, Klaassen, Wellner, (1993), section 6.6). In particular, by setting $g(T) = I_{(0,t]}(T)$ in (4.5) one obtains the well known self-consistency equation (Efron, 1967):

$$F_n^h(t) = \frac{1}{n} \sum_{i=1}^{n} P_{F_n^h}(T \leq t \mid Y_i^h), \; t \in [0, \tau], \tag{4.6}$$

where $P_F(T \leq t \mid Y^h) = P_F(T \leq t \mid T \in B(Y^h)_1)$, where $B(Y^h)_1$ is a point, horizontal strip, vertical strip, or an upper quadrant, where the strips and quadrants start at the grid points. The SOR-MLE $F_n^h$ is computed by iterating this equation with an initial estimator of $F$ which puts mass on each point of the support of $F_n^h$. The self-consistency equation tells us that $F_n^h$ puts at least mass $1/n$ on each uncensored observation, which provides us with the following useful bound: for each set $A$:

$$F_n^h(A) \geq P_{11}^n(A). \tag{4.7}$$

## 4.3 Outline of the efficiency proof.

Firstly, we define the models corresponding with the data $Y^h$ and $Y$. Let $\mathcal{F}$ be the set of all bivariate distributions on $[0, \infty)$ and $\mathcal{F}_h$ be the set of all possible bivariate distributions $G_h$ which live on $\pi^h$. Then the model corresponding with $Y^h$ (see (4.3)) is given by

$$\mathcal{M}_h \equiv \left\{ P^h_{F,G_h} : F \in \mathcal{F}, G_h \in \mathcal{F}_h \right\}$$

and the model corresponding with $Y$ (see (4.1)) by

$$\mathcal{M} \equiv \left\{ P_{F,G} : F, G \in \mathcal{F} \right\}.$$

Let $D[0,\tau]$ be the space of bivariate cadlag functions on $[0,\tau]$ as defined in Neuhaus (1971). We are interested in estimating the parameter

$$\vartheta_h : \mathcal{M}_h \to D[0,\tau] : \vartheta_h(P^h_{F,G_h}) = S.$$

Similarly, we define

$$\vartheta : \mathcal{M} \to D[0,\tau] : \ \vartheta(P_{F,G}) = S.$$

To begin with we will prove pathwise differentiability of these parameters (see e.g. BKRW, 1993, chapter 3, van der Vaart, 1988).

Let $\mathcal{S}(F)$ the class of lines $\epsilon F_1 + (1-\epsilon)F$, $F_1 \in \mathcal{F}$, with score $h = d(F_1 - F)/dF \in L^2_0(F)$, through $F$. By convexity of $\mathcal{F}$ this is a class of submodels. Let $S(F) \subset L^2_0(F)$ be the corresponding tangent cone (i.e. set of scores). It is easily verified that the tangent space $T(F)$ (the closure of the linear extension of $S(F)$) equals $L^2_0(F)$. Each submodel of $\mathcal{S}(F)$ with score $g$ will be denoted with $F_{\epsilon,g}$. The score of the one dimensional submodels $P^h_{F_{\epsilon,g},G_h} \subset \mathcal{M}_h$, $g \in S(F)$, is given by $A^h_F(g)$ where $A^h_F$ is called the score operator:

$$A^h_F : L^2(F) \to L^2(P^h_{F,G_h}) : A^h_F(g)(Y^h) = E_F(g(T) \mid Y^h),$$

which is a well known result which holds in general for missing data models (van der Vaart, 1988, Gill, 1989, BKRW, 1993, section 6.6). The score operator $A_F$ for the one dimensional submodels $P_{F_{\epsilon,g},G} \subset \mathcal{M}$, $g \in S(F)$, is given by:

$$A_F : L^2(F) \to L^2(P_{F,G}) : A_F(g)(Y) = E_F(g(T) \mid Y).$$

Let $G_{h,\epsilon,g_1} \subset \mathcal{M}_h$ be a line through $G_h$ with score $g_1$. Because of factorization of $p^h_{F,G_h}(y)$ and $p_{F,G}(y)$ the scores $B^h_G(g_1)$ of $P^h_{F,G_{h,\epsilon,g_1}}$ and the scores $B_G(g_1)$ of $p_{F,G_{\epsilon,g_1}}$ are orthogonal to the range of $A_F$ and $A^h_F$, respectively.

Lemma 3.2 says that the adjoint of $A_F$ is given by

$$A^\top_F : L^2(P_{F,G}) \to L^2(F) : A^\top_F(v)(T) = E_{F,G}(v(Y) \mid T)$$

and similarly that the adjoint of $A^h_F$ is given by:

$$A^{h\top}_F : L^2(P^h_{F,G^h}) \to L^2(F) : A^{h\top}_F(v)(T) = E_{F,G^h}(v(Y^h) \mid T).$$

Hence the corresponding information operator $I^h_F = A^{h\top}_F A^h_F : L^2(F) \to L^2(F)$ is defined by:

$$I^h_F(g)(X) = E_{F,G_h}\left(E_{F,G_h}(g(X) \mid Y^h) \mid X\right).$$

If $H > \delta > 0$, then it is trivially verified that $||A_F(h)||_{P_F} > \sqrt{\delta}||h||_F$. Now, application of lemma 3.2 tells us that this implies that $I^h_F : L^2(F) \to L^2(F)$ has a bounded inverse, uniformly in $F \in \mathcal{F}$. And the same result holds for $I_F : L^2(F) \to L^2(F)$. This proves:

**Lemma 4.1** *Let $I_{F,G} = A_F^{\top} A_F : L^2(F) \to L^2(F)$ be the information operator for $\mathcal{M}$. We have: If $H > \delta > 0$ $F$-a.e., for certain $\delta > 0$ then $I_{F,G}$ has bounded inverse $I_{F,G}^{-1}$ with norm smaller than $1/\delta$ and is onto. The same holds for the information operator $I_{F,G_h}^h : L^2(F) \to L^2(F)$ for $\mathcal{M}_h$ with inverse $I_{h,F,G_h}^{-1}$, where the bound is uniform in $h$.*

Let $b_t : D[0,\tau] :\to \mathbb{R}$ be defined by $b_t F = F(t)$. Define $\kappa_t \equiv I_{(t,\infty)} - S(t)$. For each one dimensional submodel $P^h_{F_{\epsilon,g},G_{h,\epsilon,g_1}}$, we have

$$\begin{aligned}
\frac{1}{\epsilon}\left(b_t\vartheta_h(P^h_{F_{\epsilon,g},G_{h,\epsilon,g_1}}) - b_t\vartheta_h(P^h_{F,G_h})\right) &= \int_{(t,\infty)} g dF \\
&= \langle I_{(t,\infty)} - S(t), g\rangle_F \\
&= \langle \kappa_t, g\rangle_F \\
&= \langle I_F^h I_{h,F}^{-1}(\kappa_t), g\rangle_F \\
&= \langle A_F^h I_{h,F}^{-1}(\kappa_t), A_F^h(g)\rangle_{P^h_{F,G_h}} \\
&= \langle A_F^h I_{h,F}^{-1}(\kappa_t), A_F^h(g) + B_G^h(g_1)\rangle_{P^h_{F,G_h}},
\end{aligned}$$

where we used the orthogonality of the scores at the last step. The same holds for $\vartheta$ and $P_{F,G}$ without $h$. This proves by definition (see e.g. BKRW, 1993) that for each $t \in [0,\tau]$, $b_t\vartheta_h$ is pathwise differentiable at $P^h_{F,G_h}$ for each one dimensional submodel $P_{F_{\epsilon,g},G_{h,\epsilon,g_1}}$ at $P^h_{F,G_h}$ with efficient influence function (suppressing the $G$ in the notation) given by:

$$\widetilde{I}^h(F,t)(\cdot) = A_F^h I_{h,F}^{-1}(\kappa_t)(\cdot). \tag{4.8}$$

And similarly for $\vartheta$ at $P_{F,G}$ with

$$\widetilde{I}(F,t)(\cdot) = A_F I_F^{-1}(\kappa_t)(\cdot). \tag{4.9}$$

Notice that these are the same efficient influence curves as we would have found in the models where $G = G_0$ would have been known. In the sequel $G_0$ does not vary and therefore we can skip the $G$ in the notation; $P_F^h \equiv P^h_{F,G_h}$ and $P_F \equiv P_{F,G_0}$, $I_F = I_{F,G}$ etc.

Our goal is to prove efficiency of $S_n^h$ as an estimator of $\vartheta(P_{F_0}) = S_0$. It should be remarked that for fixed $h$ application of theorem 3.3 provides us under the assumptions as stated in section 2.1, by simple verification, with efficiency of $S_n^h$, among estimators based on the data $Y_i^h$, $i = 1,\ldots,n$, as an estimator of $\vartheta_h(P^h_{F_0}) = S_0$. However, we want more than efficiency for a fixed reduction. For this purpose we will follow the same analysis as followed for the general class of missing data models, except that we look carefully what happens if $h_n \to 0$ when the number of observation converges to infinity.

It works as follows: The model $\mathcal{M}_h$ is *convex* and the $F \to P_F^h$ is *linear*. Theorem 2.2 says now that we have the following identity; for each $t \in [0, \tau]$ we have

$$S_1(t) - S_0(t) = -\int \tilde{I}^h(S_1, t) dP_{F_0}^h,$$

for all $F_1$ with $F_0 \ll F_1$ and $dF_0/dF_1 \in L_0^2(F_1)$. So in particular this identity holds for

$$S_n^h(\alpha) = \alpha S_0 + (1-\alpha) S_n^h, \ \alpha \in (0, 1],$$

which provides us with the identity:

$$S_n^h(\alpha)(t) - S_0(t) = -\int \tilde{I}^h(S_n^h(\alpha), t) dP_{F_0}^h, \ \alpha \in (0, 1]. \tag{4.10}$$

Notice now that $S_n^h(\alpha) - S_n^h = \alpha(S_n^h - S_0)$. If $\alpha \to 0$ the left-hand side of (4.10) converges to $S_n^h(t) - S_0(t)$ and it has been verified for the general class of missing data models that the right-hand side converges to $-\int \tilde{I}^h(S_n^h, t) dP_{F_0}^h$; in fact in our proof we show that $\int (I^h(S_n^h, t) - I^h(S_0, t))^2 dP_{F_0}^h \to 0$ which basically proves this much weaker result (notice that $S_n^h(\alpha)$ converges to $S_n^h$ w.r.t. each norm). It follows that we have the following identity:

$$S_n^h(t) - S_0(t) = -\int \tilde{I}^h(S_n^h, t) dP_{F_0}^h. \tag{4.11}$$

It remains to verify:

**Efficient score equation.** For all $t \in [0, \tau]$

$$\int \tilde{I}^h(F_n^h, t) dP_n^h = 0.$$

The score equations (4.5) tell us that it suffices to prove that $I_{F_n^h}^{-1}(I_{(t,\infty)})$ has finite supnorm. This is proved by lemma 4.12 in section 6 of this chapter.

The efficient score equation and the identity (4.11) provide us with the crucial identity

$$S_n^h(t) - S_0(t) = \int \tilde{I}^h(F_n^h, t) d(P_n^h - P_{F_0}^h). \tag{4.12}$$

**Empirical process condition.** Now, we will show for an appropriate rate $h_n \to 0$ that

$$\sup_{t \in [0,\tau]} \left| \int \left( \tilde{I}^h(F_n^h, t) - \tilde{I}^h(F_0, t) \right) d(P_n^h - P_{F_0}^h) \right| = o_{P_{F_0}^h}\left(1/\sqrt{n}\right).$$

This condition requires a lot of hard work (done in section 4 and 7). The reason for this is that we are not able to prove that $\widetilde{I}(F_0, t)$ has any nice properties, except that it exists as an element in $L_0^2(P_{F_0})$, due to the very complicated form of the information operator $I_{F_0}$. Therefore $\widetilde{I}^h(F_n^h, t)$ cannot be shown to be an element of a fixed Donsker-class when $h_n \to 0$. In other words the $P$-Donsker class and $\rho_P$-consistency condition as used in the proof for the general class of missing data models in chapter 3 do not help us here. More sophisticated conditions are needed. The technique will be to determine how quickly $\widetilde{I}^h(F_n^h, t)$ looses its Donsker class properties for $h_n \to 0$ and then to use (4.12) in order to obtain a rate for $||S_n^h - S_0||_\infty$ so that terms can be shown to converge to zero if $h_n \to 0$ slowly enough.

The empirical process condition provides us with (see e.g. Pollard, 1990)

$$S_n^h(t) - S_0(t) = \int \widetilde{I}^h(F_0, t) d(P_n^h - P_{F_0}^h) + o_{P_{F_0}^h}\left(1/\sqrt{n}\right),$$

where the remainder holds uniformly in $t$.

**Approximation condition.** Finally, we need to show

$$\int \widetilde{I}^h(F_0, t) d\sqrt{n}(P_n^h - P_{F_0}^h) \xrightarrow{D} N\left(0, \sigma^2(\widetilde{I}(F_0, t))\right).$$

Notice that the left-hand side is a sum of i.i.d. random variables given by: $1/\sqrt{n}\sum_{i=1}^n X_i^h(t)$ where $X_i^h(t) \equiv \widetilde{I}^h(F_0, t)(Y_i^h)$. By Bickel and Freedman (1981) we have that if for $h = h_n \to 0$ $X_i^h(t) \xrightarrow{D} X_i(t)$ and $\mathrm{Var}(X_i^h(t)) \to \mathrm{Var}(X_i(t))$, then this sum converges weakly to a normal distribution with mean zero and variance equal to $\mathrm{Var}(X_i(t))$. These two conditions are proved by lemma 4.8.

We also show the approximation condition for the case that we consider the left and right-hand side as a random element of a $L^2$-space of functions in $t$, which provides us with pointwise and $L^2$-efficiency.

## 4.4 Proof of efficiency of SOR-MLE.

Recall the assumptions made in section 2.1: in particular $F_0(\tau) = 1$ and hence $P_{F_0}^h(\cdot, d)$ lives on $[0, \tau]$. In all statements the width (of grid) $h$ converges to zero for $n \to \infty$; the problem is to find a lower bound for the rate at which $h$ should converge to zero.

### 4.4.1 Uniform consistency of $F_n^h$ for $h_n \to 0$.

The starting point of the analysis is (4.12). The indicators are a uniform Donsker class. This tells us that $\sup_h \|P_n^h - P_{F_0}^h\|_\infty = O_P(1/\sqrt{n})$.

A real valued function on $[0,\tau] \subset \mathbb{R}^2$ is called to be of bounded *uniform sectional variation* if the variations of all sections ($s \to f(s,t)$ is a section of the bivariate function f) and of the function itself is uniformly (in all sections) bounded. The corresponding norm is denoted with $\|\cdot\|_v^*$. Recall from chapter 1 that the class of functions with uniform sectional variation smaller than $M < \infty$ is a uniform Donsker class and that if $f > \delta > 0$, then $\|1/f\|_v^* \leq M\|f\|_v^*$ for some $M < \infty$ which does not depend on $f$ (Gill, 1993). We have:

**Lemma 4.2** (Uniform sectional variation of efficient influence curve). *Let $E_{k,l}^h(1,0) \equiv (u_k, u_{k+1}] \times [v_l, \infty)$ be the the vertical strips of $\pi^h$ and $E_{k,l}^h(0,1)$ be the horizontal strips. Suppose that the grid $\pi^h$ is so that $F_0(E_{k,l}^{h_n}) > \delta h_n$ for certain $\delta > 0$. Let $r_1(h_n) = 1/h_n^{3/2}$.*

*For all $d \in \{0,1\}^2$ we have that for some $M < \infty$ $\widetilde{I}^h(F_n^h, t)(\cdot, d) \in D[0,\tau]$ and*

$$\sup_{t\in[0,\tau]} \|\widetilde{I}^h(F_n^h, t)(\cdot, d)\|_v^* \leq M r_1(h) \text{ with probability tending to 1.}$$

**Proof.** See section 6.

Consider an integral $\int F_1 dH_1$ where $F_1 \in D[0,\tau]$ and $H_1 \in D[0,\tau]$ are bivariate real valued cadlag functions which are of bounded uniform sectional variation. By integration by parts lemma 1.3 we can bound it by $C\|H_1\|_\infty \|F_1\|_v^*$. Because $\widetilde{I}^h(F_n^h, t)(\cdot, d)$ generates a signed measure we can apply this to (4.12) with $F_1 = \widetilde{I}^h(F_n^h, t)(\cdot, d)$ and $H_1 = (P_n^h - P_{F_0}^h)(\cdot, d)$ and apply lemma 4.2 to $F_1$. This proves the following lemma:

**Lemma 4.3** (Uniform consistency). *Under the assumption of lemma 4.2 we have:*

$$\|F_n^{h_n} - F_0\|_\infty = O_P\left(\frac{r_1(h_n)}{\sqrt{n}}\right) = O_P\left(\frac{1}{\sqrt{nh_n^3}}\right).$$

So if $h \to 0$ slower than $n^{-1/3}$, then $F_n^h$ is uniformly consistent (also for $h$ is fixed).

### 4.4.2 Empirical process condition.

Define $Z_n^h \equiv \sqrt{n}(P_n^h - P_{F_0}^h)$ and $f_{nt}^h \equiv \widetilde{I}^h(F_n^h, t) - \widetilde{I}^h(F_0, t)$. We will show that $\int f_{nt}^h dZ_n^h$ converges to zero uniformly in $t$ with probability tending to 1. By

using that $\|F_n^h - F_0\|_\infty = O_P(r_1(h_n)/\sqrt{n})$ (lemma 4.3) we are able to show that:

**Lemma 4.4** (Supnorm convergence of efficient influence curve). *Under the assumption of lemma 4.2 we have for all* $d \in \{1,0\}^2$, *with* $r_2(h_n) = 1/h_n^3$:

$$\|f_{nt}^h(\cdot, d)\|_\infty = O_P\left(r_1(h_n)r_2(h_n)/\sqrt{n}\right) = O_P\left(1/\sqrt{nh_n^9}\right).$$

**Proof.** See appendix.

**Analysis of the uncensored term.** Let's first analyze $\int f_{nt}^h I(d=(1,1))dZ_n^h$. Recall that $Z_n^h I(d=(1,1)) = Z_n I(d=(1,1)) = \sqrt{n}(P_{11}^n - P_{11})$, where $p_{11} \Rightarrow f_0 H_h$. We will assume that $F_0 = F_0^d + F_0^c$, where $F_0^c$ is absolute continuous w.r.t. the Lebesque measure with continuous density which is bounded away from zero and $F_0^d$ is purely discrete with finite support. Then we can decompose $P_{11} = P_{11}^d + P_{11}^c$, where $p_{11}^d = f_0^d H_h$ is purely discrete on the finite number of support points of $F_0^d$ and $P_{11}^c$ is absolutely continuous w.r.t. Lebesque measure with density bounded away from zero.

For $P_{11}^n$ we have a corresponding decomposition $P_{11}^n = P_{11}^{nd} + P_{11}^{nc}$, where $P_{11}^{nd}$ only counts the number of observations coming from $P_{11}^d$. Firstly consider the integral w.r.t. $\sqrt{n}(P_{11}^{nd} - P_{11}^d)$. Let $p_{11}^d$ be the density of $P_{11}^d$ w.r.t. the counting measure, say $\mu_k$, which lives on the support of $P_{11}^d$. We have that $\int \mid p_{11}^{nd} - p_{11}^d \mid d\mu_k = O_P(1/\sqrt{n})$. Therefore, with $Z_{nd} \equiv \sqrt{n}(P_{11}^{nd} - P_{11}^d)$ we have

$$\begin{aligned}
\int f_{nt}^h I(d=(1,1))dZ_{nd} &= \sqrt{n}\int f_{nt}^h I(d=(1,1))(p_{11}^{nd} - p_{11}^d)d\mu_k \\
&\leq \sqrt{n}\|f_{nt}^h I(d=(1,1))\|_\infty \int \mid (p_{11}^{nd} - p_{11}^d) \mid d\mu_k \\
&= \sqrt{n} O_P\left(\frac{1}{\sqrt{nh_n^9}}\right) O_P\left(\frac{1}{\sqrt{n}}\right) \\
&= O_p\left(\frac{1}{\sqrt{nh_n^9}}\right),
\end{aligned}$$

where the bound does not depend on $t$. Consequently, if $nh_n^9 \to \infty$, then $\int f_{nt}^h I(d=(1,1))dZ_{nd} = o_P(1)$.

Consider now $\int f_{nt}^h I(d=(1,1))dZ_n^c$, where $Z_n^c I(d=1,1) = \sqrt{n}(P_{11}^{nc} - P_{11}^c)$. For convenience, we denote $Z_n^c$ with $Z_n$, again. We construct a lattice-grid $\pi^{a_n} = (t_i, t_j)$, with maximal mesh $a_n < h_n$, on $[0,\tau] = [0,\tau_1] \times [0,\tau_2]$, which we force to be so that $\pi^{h_n} \subset \pi^{a_n}$. Now

$$[0,\tau] = \bigcup_{i,j} A_{i,j}(a_n), \text{ where } A_{i,j}(a_n) \equiv ((t_i, t_{i+1}] \times (t_j, t_{j+1}]) \cap [0,\tau]$$

and the union is over all partition elements $A_{i,j}(a_n)$, $i = 1, \ldots, n_1(a_n)$, $j = 1, \ldots, n_2(a_n)$. The number of partition elements will be denoted by $n(a_n)$ and it is clear that $n(a_n) = O(1/a_n^2)$. Now, we define an approximation of $Z_n$ as follows:

$$Z_n^{a_n}(t) \equiv Z_n(t_i, t_j) \text{ if } t \in A_{i,j}(a_n).$$

So $Z_n^{a_n}$ is constant on each $A_{i,j}(a_n)$ with value $Z_n(t_i, t_j)$.

By using integration by parts it is clear that we have for $d = (1,1)$ (the integral is over $y \in [0, \tau]$, fixed $d$):

$$\begin{aligned}
\int f_{nt}^h(y,d)dZ_n(y,d) &= \int f_{nt}^h(y,d)d(Z_n - Z_n^{a_n})(y,d) \\
&\quad + \int f_{nt}^h(y,d)dZ_n^{a_n}(y,d) \\
&\leq C\|f_{nt}^h(\cdot,d)\|_v^* \|(Z_n - Z_n^{a_n})(\cdot,d)\|_\infty \\
&\quad + \|f_{nt}^h(\cdot,d)\|_\infty \|Z_n^{a_n}(\cdot,d)\|_v^* \\
&\leq O_P\left(r_1(h_n)\right) \|(Z_n - Z_n^{a_n})(\cdot,d)\|_\infty \\
&\quad + O_P\left(\frac{r_1(h_n)r_2(h_n)}{\sqrt{n}}\right) \|Z_n^{a_n}(\cdot,d)\|_v^*.
\end{aligned}$$

In order to show that $\int f_{nt}^h(y,d)dZ_n(y,d) = o_P(1)$ for a rate $h_n \to 0$, it suffices to show that there exists a rate $a_n$ for which the last two terms converge to zero in probability.

For convenience we will neglect the $d$ in our notation. Define:

$$W_{i,j}^n(a_n) \equiv \sup_{s,t \in A_{i,j}(a_n)} | Z_n(s) - Z_n(t) |,$$

and

$$W_n(a_n) \equiv \max_{i,j} W_{i,j}^n(a_n).$$

In other words, $W_n(a_n)$ is a *modulus of continuity* of a bivariate empirical process. Firstly, we will bound the two terms in $W^n(a_n)$.

We have $\|Z_n^{a_n} - Z_n\|_\infty \leq \max_{i,j} W_{i,j}^n(a_n)$. Therefore

$$P\left(\|Z_n^{a_n} - Z_n\|_\infty > \epsilon\right) \leq P\left((W_n(a_n) > \epsilon\right). \tag{4.13}$$

Furthermore we have

$$\|Z_n^{a_n}\|_v^* \leq \sum_{i,j} W_{i,j}^n(a_n) \leq \frac{c}{a_n^2} W_n(a_n). \tag{4.14}$$

**Analysis of the modulus of continuity.** For a rectangle $R$ we define $Z_n(R)$ as the measure of $R$ assigned by the bivariate signed measure $Z_n$. Define $W_{n,R}(a_n) \equiv \sup_{R:|R|\leq a_n} \mid Z_n(R) \mid$. Einmahl's (1987) inequality 6.4, for $W_{n,R}(a_n)$ holds for an empirical process from a sample of a continuous density which is bounded away from zero and infinity on $[0, \tau]$ and is given by:

$$P\left(W_{n,R}(a_n) > \lambda\right) \leq \frac{C}{a_n} \exp\left(\frac{-c_1\lambda^2}{a_n}\Psi(\frac{\lambda}{\sqrt{na_n}})\right) \text{ for any } \lambda > 0, \quad (4.15)$$

where $\Psi(x) \geq 1/(1 + 1/3x)$. Notice that $W_n(a_n)$ is a bound on the measure assigned by $Z_n$ to strips instead of rectangles. However, the strips are a union of at most $c/a_n$ rectangles $A_{i,j}(a_n)$ and on each rectangle $A_{i,j}(a_n)$ of these strips $p_{11}^c$ is bounded away from zero and infinity and is continuous (here we use the nesting of $\pi^{h_n}$ in $\pi^{a_n}$) and hence for the modulus of continuity on the sets $A_{i,j}(a_n)$ the discontinuities on $\pi_h$ play no role. Consequently, (4.15) can be applied to each rectangle $A_{i,j}(a_n)$ in the strips. So the bound (4.15) implies the following bound for $W_n(a_n)$:

$$\begin{aligned} P\left(W_n(a_n) > \lambda\right) &\leq \frac{c}{a_n} P\left(W_{n,R}(a_n) > \lambda\right) \\ &\leq \frac{C}{a_n^2} \exp\left(\frac{-c_1\lambda^2}{a_n}\Psi(\frac{\lambda}{\sqrt{na_n}})\right) \text{ for any } \lambda > 0, \end{aligned}$$

where $\Psi(x) \geq 1/(1+1/3x)$ and where the $C$ is now different from the preceding one.

By using this inequality with $\lambda = a_n^{0.5-\epsilon}$ it is trivial to see that if $na_n \to \infty$ at an arbitrarily small polynominal rate ($n^\epsilon$), then for each $\epsilon > 0$ there exists a sequence $\delta_n \to 0$ and an $\epsilon' > 0$ so that

$$P\left(\frac{W_n(a_n)}{a_n^{0.5-\epsilon}} > \delta_n\right) \leq \frac{C}{a_n^2} \exp\left(-C_1/a_n^{\epsilon'}\right). \quad (4.16)$$

So $W_n(a_n)/a_n^{0.5-\epsilon}$ converges to zero in probability exponentially fast.

Assume $na_n \to \infty$ at a polynomial rate. Applying (4.16) to (4.13) provides us with:

$$P\left(\frac{\|Z_n^{a_n} - Z_n\|_\infty}{a_n^{0.5-\epsilon}} > \epsilon\right) \leq \frac{C}{a_n^2} \exp\left(-C_1/a_n^{\epsilon'}\right) = o(1).$$

So $\|Z_n^{a_n} - Z_n\|_\infty = o_P\left(a_n^{0.5-\epsilon}\right)$. This proves that $r_1(h_n)\|(Z_n - Z_n^{a_n})(\cdot, d)\|_\infty = o_P\left(r_1(h_n)a_n^{0.5-\epsilon}\right)$ for any $\epsilon > 0$.

Furthermore, applying (4.16) to (4.14) provides us with:

$$\|Z_n^{a_n}\|_v^* = O\left(1/a_n^2\right) o_P\left(a_n^{0.5-\epsilon}\right) = o_P\left(a_n^{-(1.5+\epsilon)}\right).$$

Consequently, this tells us that for each $\epsilon > 0$ we have: If $na_n \to \infty$ (at least at a polynomial rate), then

$$\int f_{nt}^h(y,1,1)dZ_n(y) = o_P\left(r_1(h_n)a_n^{0.5-\epsilon}\right) + o_P\left(\frac{r_1(h_n)r_2(h_n)}{\sqrt{n}a_n^{1.5+\epsilon}}\right). \quad (4.17)$$

For the first term it suffices that $a_n$ converges quicker to zero than $h_n^3$. Substituting this in the second term tells us that we it suffices to let $h_n$ converge to zero slower than $n^{-1/18}$. This proves the following lemma:

**Lemma 4.5** *Suppose that $F_0 = F_0^d + F_0^c$, where $F_0^c$ is absolutely continuous w.r.t. Lebesque measure with continuous density which is bounded away from zero on $[0,\tau]$ and $F_0^d$ is purely discrete with finite support on $[0,\tau]$.*

*If $h_n$ converges to zero slower than $n^{-1/18}$, then $\int f_{nt}^h I(D=(1,1))dZ_n^h = o_P(1)$.*

**Analysis of the censored terms.** We will now analyze the terms $\int f_{nt}^h I(D \neq (1,1))dZ_n^h$. Recall that $P_{F_0}^h I(D \neq (1,1))$ is purely discrete on the grid $\pi^h$, which contains $O(1/h_n^2)$ points. Let $p_{F_0}^h$ and $p_n^h$ be the densities of $P_{F_0}^h$ and $P_n^h$ w.r.t. $\nu_h$, respectively. So $p_{00}^{h,n}(v_i,v_j) \equiv p_n^h(v_i,v_j,0,0)$ is the fraction of doubly censored observations which falls on $(v_i,v_j)$ and similarly for $D=(1,0)$ and $D=(0,1)$. It is clear that for fixed $h_n$ we have $\|p_n^h - p_{F_0}^h\|_\infty = O_P(1/\sqrt{n})$. In the following result for $h_n \to 0$ we do not make any assumptions. Under weak assumptions the rate would be $O_p(1/\sqrt{h_n^2 n})$, but this improvement is not interesting because of the slow rate in lemma 4.5.

**Lemma 4.6** *We have that*

$$\|p_{01}^{hn} - p_{01}^h\|_{L_1(\nu_h)} = O_p\left(\frac{1}{\sqrt{h^4 n}}\right),$$

*and we have the same rate result for $p_{10}^{hn}$ and $p_{00}^{hn}$.*

**Proof.** We give the proof for the first term, the others are dealt with similarly. Because we are just dealing with a multinomial distribution on the grid $\pi^h$ we have that $E(p_{01}^{nh}(u_k,v_l)) = p_{01}^h(u_k,v_l)$ and $\mathrm{Var}(p_{01}^{nh}(u_k,v_l)) = \frac{1}{n}p_{01}^h(u_k,v_l)(1 - p_{01}^h(u_k,v_l))$. $\pi^h$ has $O(h_n^2)$ grid points $(u_k,v_l)$ by definition of $\pi^h$. Now, we have

$$\begin{aligned} E\left(\sum_{k,l} \mid (p_{01}^{hn} - p_{01}^h)(u_k,v_l)) \mid\right) &= \sum_{k,l} E\left(\mid (p_{01}^{hn} - p_{01}^h)(u_k,v_l) \mid\right) \\ &\leq \frac{1}{\sqrt{n}} \sum_{k,l} \sqrt{p_{01}^h(u_k,v_l)(1 - p_{01}^h(u_k,v_l))} \\ &\leq \frac{1}{\sqrt{n}} \frac{1}{h^2}. \square \end{aligned}$$

Again, we will neglect the $d$ in our notation, but the reader should remember that we only integrate over the singly censored and doubly censored observations. Now, we have:

$$\begin{aligned}\int f_{nt}^h dZ_n^h &= \sqrt{n}\int f_{nt}^h(p_n^h - p_{F_0}^h)d\nu_h \\ &\leq \sqrt{n}\|f_{nt}^h\|_\infty \|p_n^h - p_{F_0}^h\|_{L_1(\nu_h)} \\ &= \sqrt{n}O_P\left(\frac{1}{\sqrt{h_n^9 n}}\right)O_P\left(\frac{1}{\sqrt{nh_n^4}}\right) \\ &= O_P\left(\frac{1}{\sqrt{h_n^{13}n}}\right).\end{aligned}$$

This proves the following lemma:

**Lemma 4.7** *If $h_n$ converges to zero slower than $n^{-1/13}$, then $\int f_{nt}^h I(D = d)dZ_n^h = o_P(1)$ for $d \in \{(1,0),(0,1),(0,0)\}$.*

Lemma 4.5 and lemma 4.7 prove the empirical process condition for a rate of $h_n$ slower than $n^{-1/18}$. Recall that all the derived lower bounds are derived without any knowledge about $\widetilde{I}(F_0,t)$, except that it has a finite variance, and therefore they only have a theoretical value.

### 4.4.3 Approximation condition.

**Pointwise convergence.**

Let $t \in [0,\tau]$ be fixed. Define $V_n^h(t) \equiv \int \widetilde{I}^h(F_0,t)(y)dZ_n^h(y)$. $V_n^h(t)$ is a sum of i.i.d. mean zero random variables given by: $1/\sqrt{n}\sum_{i=1}^n X_i^h(t)$ where $X_i^h(t) \equiv \widetilde{I}^h(F_0,t)(Y_i^h)$. By Bickel and Freedman (1981) we have that if for $h = h_n \to 0$ $X_i^h(t) \xRightarrow{D} X_i(t)$ and $\mathrm{Var}(X_i^h(t)) \to \mathrm{Var}(X_i(t))$, then this sum converges weakly to a normal distribution with mean zero and variance equal to $\mathrm{Var}(X_i(t))$. We will prove these two conditions:

**Lemma 4.8** *Define the following real valued random variables $X^h(t) \equiv \widetilde{I}^h(F_0,t)(Y^h)$, $Y^h \sim P_{F_0}^h$ and $X(t) \equiv \widetilde{I}(F_0,t)(Y)$, $Y \sim P_{F_0}$. We have for each $t \in [0,\tau]$ that for $h_n \to 0$*

$$E((X^{h_n}(t) - X(t))^2) \to 0$$

*and*

$$E(X^{h_n}(t)X^{h_n}(s)) \to E(X(t)X(s)) \text{ uniformly in } s,t \in [0,\tau].$$

**Proof.** See section 6. Lemma 4.8 has the following corollary

**Corollary 4.1** *The empirical process* $\int \widetilde{I}^{h_n}(F_0,t)(y)dZ_n^{h_n}(y)$ *converges in distribution to a normal distribution with mean zero and variance equal to* $Var_{P_{F_0}}(\widetilde{I}^0(F_0,t))$.

**Hilbert space convergence.**

For showing that $V_n^h$ converges weakly as a process in $(D[0,\tau], \|\cdot\|_\infty)$ we need to show at least that $\{\widetilde{I}(F_0,t) : t \in [0,\tau]\}$ is a $P_{F_0}$-Donsker class. We have not been able to do this. Therefore we concentrate on proving weak convergence as a process in a Hilbert space. We use the following result which can be found in Parthasarathy (1967, p. 153).

**Lemma 4.9** *Let* $Z_n, Z_0$ *be random processes in a Hilbert space* $\mathcal{H}$ *endowed with the Borel sigma algebra* $\mathcal{B}$. *Let* $e_1, e_2, \ldots$ *be an orthonormal basis of* $\mathcal{H}$. *If* $\langle e_j, Z_n\rangle \xrightarrow{D} \langle e_j, Z_0\rangle$ *for all* $j$ *and* $\lim_{N\to\infty} \sup_n E(\sum_{j=N+1}^{\infty}\langle e_j, Z_n\rangle^2) = 0$, *then* $Z_n \xrightarrow{D} Z_0$ *in* $\mathcal{H}$.

Let $V_n(t) = 1/\sqrt{n}\sum_{i=1}^n X_i(t)$. Firstly, we will prove the first condition of lemma 4.9 with $Z_n = V_n^h$ and $Z_0 = V_0$, the optimal Gaussian process. We have

$$\langle e_j, V_n^h\rangle = \langle e_j, V_n^h - V_n\rangle + \langle e_j, V_n\rangle.$$

Firstly, we will show that $\langle e_j, V_n^h - V_n\rangle = o_P(1)$. The fact that $V_n^h$ and $V_n$ are sums of i.i.d. random variables $X_i^h$ and $X_i$, respectively, and the Cauchy-Schwarz inequality tell us:

$$\begin{aligned}\mathrm{Var}\left(\langle e_j, V_n^h - V_n\rangle\right) &= \mathrm{Var}\left(\langle e_j, X^h - X\rangle\right)\\ &\leq \mathrm{E}\left(\langle e_j, X^h - X\rangle^2\right)\\ &\leq \langle e_j, e_j\rangle E\langle X^h - X, X^h - X\rangle.\end{aligned}$$

Assume now that $\mathcal{H} = L^2(\lambda)$ for a certain finite measure $\lambda$. By lemma 4.8 we have $\mathrm{Var}(X^{h_n}(t))$ converges to $\mathrm{Var}(X(t))$ and $E((X^{h_n}(t) - X(t))^2) \to 0$, both uniformly in $t$. Therefore,

$$E\langle X^h - X, X^h - X\rangle \leq \sup_{s\in[0,\tau]} |\, E((X^h - X)(s)^2)\,| \int d\lambda(s) \to 0,$$

which proves the convergence of $\langle e_j, V_n^h - V_n\rangle$ to zero in probability. Furthermore, we have

$$\langle e_j, V_n\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int e_j(s)X_i(s)d\lambda(s),$$

which is just a sum of i.i.d. mean zero random variables. By the CLT, for showing that this converges in distribution to $\langle e_j, V_0 \rangle$ it suffices to have that $\mathrm{Var}(\int e_j(s) X_i(s) d\lambda(s)) < \infty$. This follows immediately from the fact that $\|E(X^2(s))\|_\infty < \infty$. This proves the weak convergence of $\langle e_j, V_n^h \rangle$ to $\langle e_j, V_0 \rangle$.

We will now verify the tightness condition. We have:

$$\begin{aligned}
\mathrm{E}\left( \sum_{i=N+1}^{\infty} \langle e_i, V_n^h \rangle^2 \right) &= \sum_{i=N+1}^{\infty} \mathrm{E}\left( \langle e_i, V_n^h \rangle^2 \right) \\
&= \sum_{i=N+1}^{\infty} \mathrm{E}\left( \int\int e_i(s) e_i(t) V_n^h(s) V_n^h(t) d\lambda(s) d\lambda(t) \right) \\
&= \sum_{i=N+1}^{\infty} \int\int e_i(s) e_i(t) \mathrm{E}\left( V_n^h(s) V_n^h(t) \right) d\lambda(s) d\lambda(t) \\
&= \sum_{i=N+1}^{\infty} \int\int e_i(s) e_i(t) \left( \mathrm{E}\left( V_0(s) V_0(t) \right) + o(1) \right) d\lambda(s) d\lambda(t) \\
&= o(1) \left( \sum_{i=N+1}^{\infty} \left( \langle e_i, 1 \rangle \right)^2 \right) + \sum_{i=N+1}^{\infty} \langle e_i, V_0 \rangle^2 .
\end{aligned}$$

At the first, second, third equality we used Fubini's theorem, then we use the uniform convergence of $\mathrm{E}(V_n^h(s) V_n^h(t))$ to $\mathrm{E}(V_0(s) V_0(t))$, by lemma 4.8, and finally we again apply Fubini's theorem but now in the reversed order. The last bound does not depend on $n$ anymore. Because $\|V_0\|^2 = \sum_{i=1}^{\infty} \langle V_0, e_i \rangle^2$ and similarly for the function 1 it follows that if we take the limit for $N \to \infty$, then both (tail) series converge to zero.

Application of lemma 4.9 provides us now with:

**Lemma 4.10** *Suppose the same assumption as in lemma 4.8. If $\lambda$ is a finite measure and $h_n \to 0$, then $V_n^{h_n} \stackrel{D}{\Longrightarrow} V_0$ as random elements in $L^2(\lambda)$.*

## 4.5 Results.

We will summarize the necessary notation for the theorem. Recall the reduced i.i.d. data $Y_i^h \sim P_{F_0,G_h}^h$, obtained by generating $n$ i.i.d. $C_i \sim G_h$ and the $\pi^h$-interval-censoring of the singly censored observations. We defined $E_{k,l}^h(1,0)$ and $E_{k,l}^h(0,1)$ as the vertical and horizontal strips of $\pi^h$ starting at $(u_k, v_l)$. We defined $Z_n^h \equiv \sqrt{n}(P_n^h - P_{F_0,G_h}^h)$ as the empirical process corresponding with the reduced data, $\widetilde{I}^h(F_0, t)$ as the efficient influence function for estimating $F_0(t)$ using the reduced data and $\widetilde{I}(F_0, t)$ as the efficient influence function for estimating $F_0(t)$ using the original data.

We have proved all ingredients of the general efficiency proof of section 3 in section 4. Recalling lemma 4.3 (uniform consistency) and that for fixed $h$ we have efficiency (among all estimators based on the reduced data) under the assumptions as stated in subsection 2.1 provides us with the following theorem:

**Theorem 4.1** *Let $[0,\tau] \subset \mathbb{R}_{\geq 0}$ be a rectangle so that $H(\tau) > 0$, $S_0(\tau-) > 0$, $F_0(\tau) = 1$ (data reduced to $[0,\tau]$).*

**Fixed grid efficiency.** *Suppose that we do not change the grid $\pi^h$ for $n \to \infty$ and that for each grid point $(u_k, v_l)$ $F_0\left(E^h_{k,l}(1,0)\right) > 0$ and $F_0\left(E^h_{k,l}(0,1)\right) > 0$.*

*Then $S^h_n$ is a supnorm-efficient estimator of $S_0$ for the data $Y^h_i$, $i = 1,2,\ldots,n$:*

$$\sqrt{n}(F^h_n - F_0)(t) = \int \widetilde{I}^h(F_0,t)dZ^h_n + R^h_n(t),$$

*where $\|R^h_n\|_\infty = o_P(1)$ and $\int \widetilde{I}^h(F_0,t)dZ^h_n$ converges weakly in $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$ to a Gaussian process $N_h$ with mean zero finite dimensional distributions and covariance structure given by:*

$$E(N_h(s)N_h(t)) = E_{P^h_{F_0}}(\widetilde{I}^h(F_0,s)\widetilde{I}^h(F_0,t)).$$

**Uniform consistency.** *Suppose that the grid $\pi^h$ is such that $F_0\left(E^{h_n}_{k,l}(1,0)\right) > \delta h_n$ and $F_0\left(E^{h_n}_{k,l}(0,1)\right) > \delta h_n$ for some $\delta > 0$.*

*Then for any rate $h_n \to 0$*

$$\|S^{h_n}_n - S_0\|_\infty = O_P\left(1/\sqrt{nh^3_n}\right).$$

**Efficiency.** *Suppose $F_0 = F^d_0 + F^c_0$, where $F^d_0$ is purely discrete with finite support and $F^c_0$ is absolutely continuous w.r.t. Lebesque measure with continuous density uniformly bounded away from zero on $[0,\tau]$.*

*We have that for $h_n \to 0$*

$$E_{P^h_{F_0}}\left(\widetilde{I}^h(F_0,s)(Y^h)\widetilde{I}^h(F_0,t)(Y^h)\right) \to E_{P_{F_0}}\left(\widetilde{I}(F_0,s)(Y)\widetilde{I}(F_0,t)(Y)\right)$$

*uniformly in $s,t \in [0,\tau]$.*

*If $h_n$ converges to zero, but slower than $n^{-1/18}$, then we have that $\|R^h_n\|_\infty = o_P(1)$ and for each $t \in [0,\tau]$ $V^h_n(t) \equiv \int \widetilde{I}^h(F_0,t)dZ^h_n$ converges in distribution to the normal distribution $N_0(t)$ with mean zero and variance:*

$$Var(N_0(t)) = Var\left(\widetilde{I}(F_0,t)\right).$$

*Moreover, for any finite measure $\lambda$ $V^h_n$ converges weakly as a process in $L^2(\lambda)$ to $N_0$.*

*This implies that $F_n^{h_n}(t)$ is an efficient estimator of $F_0(t)$, pointwise and as an element in $L^2(\lambda)$.*

We see that if $nh_n^3 \to \infty$, then $F_n^{h_n}$ converges uniformly to $F_0$. Therefore, we think that $n^{-1/3}$ can also be used as a lower bound for asymptotic efficiency, though we did not prove this.

## 4.6 The bootstrap.

By using the identity (4.12), verification of the bootstrap is immediate. We follow the line of the generalized version of theorem 2.4.

**Semiparametric bootstrap.** Let $F_n$ and $G_n$ be estimators of $F_0$ and $G_0$, respectively. Draw a sample of $n$ i.i.d. observations $Y_i^* \sim P^h_{F_n,G_n^h}$. Let $P_n^{h*}$ be the empirical distribution of $Y_i^*$, $i = 1, \ldots, n$. Let $F_n^{h*}$ be the SOR-NPMLE as defined in (4.4) based on this bootstrap sample. We still have the identity:

$$\sqrt{n}(F_n^{h*} - F_n^h)(t) = \int \widetilde{I}^h(F_n^{h*}, t) d\sqrt{n}(P_n^{h*} - P^h_{F_n^h}).$$

Let $h$ be fixed. Assume that $\|F_n - F_0\|_\infty \to 0$ a.s. and $\|G_n - G_0\|_\infty \to 0$ a.s. Then it is easily verified by applying theorem 1.3 that the bootstrap works for the empirical distribution $P_n^{h*}$; $\sqrt{n}(P_n^{h*} - P^h_{F_n^h}) \stackrel{D}{\Longrightarrow} Z_h$ for $n \to \infty$, where $Z_h$ is the limit distribution of $\sqrt{n}(P_n^h - P_0^h)$. By lemma 4.2 we know that $\|\widetilde{I}(F_n^{h*}, t)\|_v^* < M < \infty$ with probability tending to 1. Integration by parts tells us now that $\|F_n^{h*} - F_n^h\|_\infty = O_P(1/\sqrt{n})$. This implies as in the proof of lemma 4.4 that $\|\widetilde{I}^h(F_n^{h*}, t) - \widetilde{I}^h(F_n^h, t)\|_\infty \to 0$. The class of functions of bounded uniform sectional variation form a uniform Donsker class (example 1.2). Therefore by the uniform continuity of the sample paths of the empirical process indexed by the functions of bounded uniform sectional variation (see (1.8)), we have now

$$\sqrt{n}(F_n^{h*} - F_n^h)(t) = \sqrt{n}(P_n^{h*} - P^h_{F_n^h}) \left(\widetilde{I}^h(F_n^h, t)\right) + R_n^{h*}(t),$$

where $\|R_n^{h*}\|_\infty = o_P(1)$. $\sqrt{n}(P_n^{h*} - P^h_{F_n^h})(\widetilde{I}^h(F_n^h, t))$ is a sum of i.i.d. mean zero random variables. Therefore, for convergence to the optimal normal distribution it suffices (lemma 1.1) again to show that these random variables converge in distribution to $\widetilde{I}^h(F_0, t)(Y)$, $Y \sim P_0^h$ and that their variance converges to the variance of $\widetilde{I}^h(F_0, t)(Y)$, $Y \sim P_0^h$. For this we just copy the proof of lemma 4.8. Because the class of functions of bounded variation form a uniform Donsker

class the process $\sqrt{n}(P_n^{h*} - P_{F_n^h}^h)(\widetilde{I}^h(F_n^h, t))$ is also tight and thereby it converges weakly to the Gaussian process $N_h$ given in our theorem. This proves the semiparametric bootstrap for fixed grid $\pi^h$.

Assume now that $h_n \to 0$. Then we know that $N_{h_n} \stackrel{D}{\Longrightarrow} N_0$ by application of lemma 4.8. This tells us that if $h_n \to 0$ slowly enough, then

$$\sqrt{n}(F_n^{h*} - F_n^h)(t) = N_0(t) + R_n^h(t),$$

where $R_n^h(t)$ converges to zero in probability.

In order to obtain a lower bound for the rate at which $h_n$ should converge to zero we need to copy our general proof. For this it was substantial that $F_0 = F_{0d} + F_{0c}$, where $F_{0c}$ has a continuous density which is uniformly bounded away from zero on $[0, \tau]$. Therefore we also need to assume that $F_n = F_{nc} + F_{nd}$, where $F_{nc}$ has a continuous density which is uniformly (also in $n$) bounded away from zero on $[0, \tau]$ and $F_{nd}$ is purely discrete on a finite support which does not depend on $n$. Then $F_n$ and $G_n$ suffice all assumptions (uniformly in $n$) which we needed for the general proof and hence it is easy to copy the general proof for $F_n \to F$ and $G_n \to G$.

**Nonparametric bootstrap.** Let $h_n$ be fixed. Assume now that we sample from $P_n^h$ and that $F_n^h$ is the SOR-NPMLE of the original sample. The identity tells us again:

$$\begin{aligned}
\sqrt{n}(F_n^{h*} - F_n^h)(t) &= \sqrt{n}(P_n^{h*} - P_{F_n^h}^h)(\widetilde{I}^h(F_n^{h*}, t)) \qquad (4.18)\\
&= \sqrt{n}(P_n^{h*} - P_n^h)(\widetilde{I}^h(F_n^{h*}, t))\\
&\quad + \sqrt{n}(P_n^h - P_{F_n^h}^h)(\widetilde{I}^h(F_n^{h*}, t))\\
&= \sqrt{n}(P_n^{h*} - P_n^h)(\widetilde{I}^h(F_n^{h*}, t))\\
&\quad + \sqrt{n}(P_n^h - P_{F_n^h}^h)\left(\widetilde{I}^h(F_n^{h*}, t) - \widetilde{I}^h(F_n^h, t)\right).
\end{aligned}$$

Firstly, we use the identity (4.18) to obtain consistency of $F_n^{h^*}$. The first term after the last equality is dealt in exactly the same way as the term $\sqrt{n}(P_n^{h*} - P_{F_n^h}^h)(\widetilde{I}^h(F_n^h, t))$ which we had to cover in the semiparametric bootstrap analysis. By theorem 1.4 the bootstrap works for $\sqrt{n}(P_n^{h*} - P_n^h)$, so we do not need assumptions for this. Consider now the second term for which we need to show that it converges to zero in probability. We have

$$\sqrt{n}(P_n^h - P_{F_n^h}^h) = \sqrt{n}(P_n^h - P_0^h) - \sqrt{n}(P_{F_n^h}^h - P_0^h).$$

The integral w.r.t. the first term is a standard empirical process and hence convergence can be shown by the Donsker class condition (lemma 4.2) and the

$\rho$-consistency condition (lemma 4.4). For integral w.r.t. the second term we assume that $\sqrt{n}(F_n^h - F_0)$ and $\sqrt{n}(G_n^h - G_h)$ converge weakly as elements of $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$. Then the second term can been shown to converge weakly by considering $P_{F,G}^h$ as a functional in $(F,G)$ and applying the functional delta method theorem. Then it is easy to show by using integration by parts and $\|\widetilde{I}^h(F_n^{h*},t) - \widetilde{I}^h(F_n^h,t)\|_\infty \to 0$ (lemma 4.4) that the second term converges to zero in probability. This proves that the nonparametric bootstrap works for $h_n$ fixed.

Our general proof for determining a lower bound for the rate at which $h_n$ should converge to 0 cannot be copied because $P_n^h$ is purely discrete. However, just as with the semiparametric bootstrap we still have that the nonparametric bootstrap works for $h_n \to 0$ slowly enough.

## 4.7 Technical lemmas.

In formulas the score operator $A_{F_0}^h$ evaluated at observation $Y^h = (\widetilde{T}, D)^h$ is given by (recall that $\widetilde{T}$ for $D \neq (1,1)$ lives on the grid $\pi^h$):

$$\begin{aligned}
A_{F_0}^h(g)(\widetilde{T},D)^h &= g(\widetilde{T})I(D=(1,1)) \\
&+ \int_{(u_k,u_{k+1}]}\int_{(v_l,\infty)} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0((u_k,u_{k+1}],[v_l,\infty))}I(D=(1,0)) \\
&+ \int_{(u_k,\infty)}\int_{(v_l,v_{l+1}]} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0([u_k,\infty),(v_l,v_{l+1}])}I(D=(0,1)) \\
&+ \int_{(u_k,\infty)}\int_{(v_l,\infty)} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0([u_k,\infty),(v_l,\infty))}I(D=(0,0)).
\end{aligned}$$

Recall that $(u_k, v_l)$ is a function of $\widetilde{T}$ and therefore it is natural to consider $v_l$ as a function in $\widetilde{T}_2$: $v_l(\widetilde{T}_2) = v_l$ if $\widetilde{T}_2 \in (v_l, v_{l+1}]$ and similarly for $u_k$. In this way all four terms can be considered as functions on $[0,\tau]$, where the last three are stepfunctions on $\pi^h$.

In formulas $I_0^h$ is given by:

$$\begin{aligned}
I_{F_0,G_h}^h(g)(T) &= g(T)H_h(T) \\
&+ \int_0^{T_2}\left(\int_{(u_k,u_{k+1}]}\int_{(v_l,\infty)} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0((u_k,u_{k+1}],[v_l,\infty))}\right)G_h((u_k,\infty),\{v_l\}) \\
&+ \int_0^{T_1}\left(\int_{(u_k,\infty)}\int_{(v_l,v_{l+1}]} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0([u_k,\infty),(v_l,v_{l+1}])}\right)G_h(\{u_k\},(v_l,\infty)) \\
&+ \int_{(0,T]}\left(\int_{(u_k,\infty)}\int_{(v_l,\infty)} g(s_1,s_2)\frac{F_0(ds_1,ds_2)}{F_0([u_k,\infty),(v_l,\infty)}\right)G_h(\{u_k\},\{v_l\}).
\end{aligned}$$

We will write down the singly censored term (2nd above) of $I_{F_0,G_0} : L^2(F_0) \to L^2(F_0)$:

$$\int_0^{T_2} \left( \int_{(v_2,\infty)} h(T_1, s_2) \frac{F_{01}(T_1, ds_2)}{F_{01}(T_1, [v_2, \infty))} \right) H_0(T_1, dv_2).$$

### 4.7.1 Proof of lemma 4.2.

**Lemma 4.11** *Let $E^h_{k,l}(1,0) \equiv (u_k, u_{k+1}] \times [v_l, \infty)$ be the the vertical strips of $\pi^h$ and $E^h_{k,l}(0,1)$ be the horizontal strips. Suppose that $H_0(\tau) > 0$ and $F_0(E^{h_n}_{k,l}) > \delta h_n$ for certain $\delta > 0$.*

*Then there exists an $\epsilon > 0$ so that for any sequence $h_n$ which converges to zero slower than $1/\sqrt{n}$ we have*

$$\min_{k,l} F_n^{h_n}(E^{h_n}_{k,l}(1,0)) \geq \epsilon h_n, \text{ with probability tending to 1.}$$

*Similarly, for $E^{h_n}_{k,l}(0,1)$.*

**Proof.** We use the notation $E^h_{k,l}$ for both strips. Firstly, by the EM-equations (see (4.7)) we have

$$F_n^h(E^h_{k,l}) \geq P_{11}^n(E^h_{k,l}), \tag{4.19}$$

where $P_{11}^n$ is the empirical distribution of the uncensored observations of $Y_i^h \sim P^h_{F_0,G_h}$. We have

$$P_{11}(E^{h_n}_{k,l}) \geq H_0(\tau) F_0(E^{h_n}_{k,l}) > \delta_1 h_n \text{ for some } \delta_1 > 0. \tag{4.20}$$

Furthermore, $\{I_{E^h_{k,l}} : h \in (0,1], k, l\}$, the collection of indicators of $E^h_{k,l}$ over all $(u_k, v_l) \in \pi^h$ and for all $h \in (0,1]$, is a uniform Donsker class. Consequently, we have for any $\epsilon > 0$ and rate $r(n)$ slower than $\sqrt{n}$ that

$$P\left( \sup_{k,l} | (P_{11}^n - P_{11})(E^{h_n}_{k,l}) | > \frac{\epsilon}{r(n)} \right) \to 0. \tag{4.21}$$

Assume that there exists an $\epsilon < \delta_1$ so that

$$\limsup_{n\to\infty} P\left( \min_{k,l} P_{11}^n(E^{h_n}_{k,l}) \leq \epsilon h_n \right) > \delta > 0 \text{ for some } \delta > 0. \tag{4.22}$$

We will prove that this leads to a contradiction if $h_n$ converges to zero slower than $1/\sqrt{n}$. The contradiction proves that for each $\epsilon < \delta_1$ and $h_n$ slower than $\sqrt{n}$

$$\lim_{n\to\infty} P\left( \min_{k,l} P_{11}^n(E^{h_n}_{k,l}) \geq \epsilon h_n \right) = 1,$$

which combined with (4.19) proves the lemma. So it remains to prove the contradiction. We have by (4.20) and (4.22), respectively,

$$\limsup_{n\to\infty} P\left(\sup_{k,l}\left|(P_{11}^n - P_{11})(E_{k,l}^{h_n})\right| > \delta_1 h_n - \epsilon h_n\right) \geq P\left(\min_{k,l} P_{11}^n(E_{k,l}^{h_n}) \leq \epsilon h_n\right) > \delta > 0.$$

However, we also have (4.21). These two contradict if $h_n$ converges to zero slower than $1/\sqrt{n}$. □

For obtaining a bound for the uniform sectional variation norm of the efficient influence function consider the equation: $I_F^h(g)(x) = f(x)$ for certain $f \in L^2(F)$. We can write $I_F^h(g) = H_h g + K_F^h(g)$, where $K_F^h(g)$ is the sum of the three terms corresponding with the censored observations. Then this equation is equivalent with the following equation:

$$g(x) = \frac{1}{H_h(x)}\left\{f(x) - K_F^h(g)(x)\right\}. \tag{4.23}$$

For the moment denote the right-hand side with $C_F^h(g,f)(x)$: i.e. we consider the equation $g(x) = C_F^h(g,f)(x)$.

We know by lemma 4.1 that for each $f$ there exists a $g' \in L^2(F)$, which is unique in $L^2(F)$, with $\|I_F^h(g') - f\|_F = 0$: i.e. $\|g' - C_F^h(g',f)\|_F = 0$. Notice that if $\|g_1 - g\|_F = 0$, then for each $x$ $C_F^h(g_1 - g, f)(x) = 0$. So even if $g'$ is only uniquely determined in $L^2(F)$, then $C_F^h(g',f)(x)$ is uniquely determined for each $x$. Now, we can define $g(x) \equiv C_F^h(g',f)(x)$. Then $\|g - g'\|_F = \|C(g',f) - g'\|_F = 0$. So in this way we have found a solution $g$ of (4.23) which holds for each $x$ instead of only in $L^2(F)$ sense.

To summarize, we have $g_h = I_{h,F}^{-1}(f)$ is given by $g_h(x) = C_F^h(g_h', f)(x)$, where $g_h' = I_{h,F}^{-1}(f)$ in $L^2(F)$ sense. Moreover, by the bounded invertibility of $I_F^h$ w.r.t. the $L^2(F)$-norm we have that $\|g_h'\|_F \leq C\|f\|_F$, where $C \leq 1/\delta$ does not depend on the width $h$.

Assume that $\|f\|_v^* < 1$. Now, we can conclude that $\|g_h\|_\infty \leq M\|K_F^h(g_h)\|_\infty$ and $\|g_h\|_v^* \leq M\|K_F^h(g_h)\|_v^*$, for certain $M < \infty$.

Therefore it remains to bound the *supnorm* and *uniform sectional variation norm* of $K_F^h(g)$ and find out how this bound depends on the width $h_n$. It suffices to do this for one of the singly censored terms of $K_F^h(g_h)$. We take the $D = (1,0)$ term which is given by:

$$W(T) \equiv \int_0^{T_2}\left(\int_{(u_k,u_{k+1}]}\int_{(v_l,\infty)} g_h(s_1,s_2)\frac{F(ds_1,ds_2)}{F((u_k,u_{k+1}],[v_l,\infty))}\right) G_h(u_k,\{v_l\}).$$

For convenience, we will often denote $E_{k,l}(1,0)$ by $E_{k,l}$.

**Supnorm.** Recall that $\|f\|_\infty \leq 1$. By the Cauchy-Schwarz inequality and $\|g_h\|_F \leq C\|f\|_F$ we have:

$$\begin{aligned}\int_{u_k}^{u_{k+1}} \int_{v_l}^{\infty} g_h(s) \frac{F(ds)}{F((v_k, u_{k+1}], [v_l, \infty))} &= \int I_{E_{k,l}}(s) g_h(s) \frac{F(ds)}{F(E_{k,l})} \\ &\leq \frac{1}{\sqrt{F(E_{k,l})}} \|g_h\|_F \\ &\leq \frac{C}{\sqrt{F(E_{k,l})}}.\end{aligned}$$

By lemma 4.11 we can assume that $F_n^{h_n}(E_{k,l}) > \epsilon h_n$ for certain $\epsilon > 0$. This proves, by replacing $F$ (above) by $F_n^h$:

**Lemma 4.12** *There exists a $C < \infty$ so that:*

$$\sup_{\|f\|_\infty = 1} \|I_{h,F_n^h}^{-1}(f)\|_\infty \leq \frac{C}{\sqrt{h_n}} \text{ with probability tending to 1.}$$

**Uniform sectional variation norm over $[0, \tau]$.** Notice that $W$ is purely discrete with jumps at the grid points $(u_k, v_l)$. Therefore the uniform sectional variation norm of $W$ equals the sum of the absolute values of all jumps. We have

$$W(T_1, \{v_l\}) = \int_{(u_k, u_{k+1}]} \int_{(v_l, \infty)} g_h(s_1, s_2) \frac{F(ds_1, ds_2)}{F((v_k, u_{k+1}], [v_l, \infty))} G_h(u_k, \{v_l\}).$$

So

$$\begin{aligned}\Delta W(u_k, v_l) = \Delta H_h(u_k, v_l) &\int_{E_{k,l}} g_h(s_1, s_2) \frac{F(ds_1, ds_2)}{F(E_{k,l})} \\ &+ \frac{-\int_{E_{k,l}} g_h(s_1, s_2) F(ds_1, ds_2)}{F(E_{k,l})^2} \left(F(E_{k+1,l}) - F(E_{k,l})\right) H_h(u_k, \{v_l\}) \\ &+ \frac{\left(\int_{E_{k+1,l}} g_h(s_1, s_2) F(ds_1, ds_2) - \int_{E_{k,l}} g_h(s_1, s_2) F(ds_1, ds_2)\right)}{F(E_{k,l})} H_h(u_k, \{v_l\}).\end{aligned}$$

Now, doing nothing more sophisticated than (we use lemma 4.12, at the first inequaltiy, and lemma 4.11 at the second)

$$\frac{\int_{E_{k,l}} g_h dF}{F(E_{k,l})} \leq \|g_h\|_\infty \leq M/\sqrt{h_n} \text{ and } F(E_{k,l}) > \epsilon h_n \tag{4.24}$$

we obtain the following bound:

$$\begin{aligned}|\Delta W(u_k, v_l)| \leq\ & |\Delta H_h(u_k, v_l)| \frac{M}{\sqrt{h_n}} \\ &+ \frac{C}{h_n^{3/2}} \left(F_n^h(E_{k,l}) + F_n^h(E_{k+1,l})\right) |H_h(u_k, \Delta v_l)|.\end{aligned}$$

Consequently, we have for the variation of $W$ with $F$ replaced by $F_n^h$:

$$\begin{aligned}\sum_{k,l}|\Delta W(u_k,v_l)| &\leq \frac{1}{\sqrt{h_n}}\sum_{k,l}|\Delta H_h(u_k,v_l)| + \frac{C}{h_n^{3/2}}\sum_{k,l}F_n^h(E_{k,l})\,|H_h(u_k,\Delta v_l)| \\ &\leq \frac{1}{\sqrt{h_n}} + \frac{C}{h_n^{3/2}} = O\left(\frac{1}{h_n^{3/2}}\right),\end{aligned}$$

where the bounds hold with probability tending to 1. So we proved the following:

**Lemma 4.13** *There exists a $C<\infty$ so that*

$$\sup_{\|f\|_v^*=1}\|I^{-1}_{h,F_n^h}(f)\|_v^* \leq \frac{C}{h_n^{3/2}} \text{ with probability tending to 1.} \tag{4.25}$$

Let $g = I^{-1}_{h,F_n^h}(f)$. The uniform sectional variation of the uncensored term of $A_{F_n^h}(g)$ is bounded by a constant times the uniform sectional variation of $g$ and the uniform sectional variation of the censored terms can be bounded as above using (4.24) by $C/h_n^{3/2}$. Therefore the uniform sectional variation of the efficient influence curve is also bounded by the rate given in (4.25). This completes the proof of lemma 4.2 (the cadlag property follows also trivially).

### 4.7.2 Proof of lemma 4.4.

We will suppress the $d$ in our notation. We have:

$$\begin{aligned}\|f_{nt}^h\|_\infty &= \|\widetilde{I}^h(F_n^h,t) - \widetilde{I}^h(F_0,t)\|_\infty \\ &\leq |\,(S_n^h - S_0)(t)\,| + \|A_n^h I_{h,n}^{-1}(\kappa_t) - A_0^h I_{h,0}^{-1}(\kappa_t)\|_\infty.\end{aligned}$$

We know that $\|F_n^h - F_0\|_\infty = O_P\left(1/(\sqrt{nh_n^3})\right)$. The rate will be determined by the second term. Let $g_{0t}^h \equiv I_{h,0}^{-1}(\kappa_t)$. We rewrite the second term as a sum of two differences:

$$\begin{aligned}A_n^h I_{h,n}^{-1}(\kappa_t) - A_0^h I_{h,0}^{-1}(\kappa_t) &= (A_n^h - A_0^h)I_{h,0}^{-1}(\kappa_t) + A_n^h I_{h,n}^{-1}(I_n^h - I_0^h)I_{h,0}^{-1}(\kappa_t) \\ &= (A_n^h - A_0^h)(g_{0t}^h) + A_n^h I_{h,n}^{-1}(I_n^h - I_0^h)(g_{0t}^h) \quad (4.26)\end{aligned}$$

Firstly, we will consider the first term. It suffices to do the analysis for one of the singly censored terms; we consider the $d=(1,0)$ term. We have by telescoping:

$$\begin{aligned}(A_n^h - A_0^h)(g_{0t}^h)(u_k,v_l,d) &= \frac{\int_{E(k,l)} g_{0t}^h dF_n^h}{F_n^h(E_{k,l})} - \frac{\int_{E(k,l)} g_{0t}^h dF_0}{F_0(E_{k,l})} \\ &= \frac{\int_{E(k,l)} g_{0t}^h d(F_n^h - F_0)}{F_0(E_{k,l})} + \frac{(F_n^h - F_0)(E_{k,l})\int_{E(k,l)} g_{0t}^h dF_n^h}{F_n^h(E_{k,l})F_0(E_{k,l})}.\end{aligned}$$

At the first term, we can apply integration by parts. So the first term is bounded by:

$$C\|F_n^h - F_0\|_\infty \frac{\|g_{0t}^h\|_v^*}{F_0(E_{k,l})}.$$

By lemma 4.13 we have $\|g_{0t}^h\|_v^* = O(1/\sqrt{h_n^3})$ and we have $F_0(E_{k,l}) > \delta h$. Therefore the first term is bounded by

$$O_P\left(\frac{1}{\sqrt{nh_n^3}}\right) O_P\left(\frac{1}{\sqrt{h_n^3}}\right) O\left(\frac{1}{h_n}\right) = O_P\left(\frac{1}{\sqrt{nh_n^8}}\right).$$

The second term is bounded by:

$$C\|F_n^h - F_0\|_\infty \|g_{0t}^h\|_\infty \frac{1}{F_0(E_{k,l})} = O_P\left(\frac{1}{\sqrt{nh_n^6}}\right).$$

This proves that

$$\|(A_n^h - A_0^h)(g_{0t}^h)\|_\infty = O_P\left(\frac{1}{\sqrt{nh_n^8}}\right).$$

Consider now the second term of (4.26). Because $A_0^\top$ does only depend on $G$, we have for the term $(I_n^h - I_0^h)(g_{0t}^h)$:

$$(I_n^h - I_0^h)(g_{0t}^h) = A_0^{h\top}(A_n^h - A_0^h)(g_{0t}^h).$$

Because $A_0^{h\top}$ is just a conditional expectation we have that $\|A_0^{h\top}(g)\|_\infty \leq \|g\|_\infty$. Therefore, we also have that $\|(I_n^h - I_0^h)(g_{0t}^h)\|_\infty = O(1/\sqrt{nh_n^8})$. Now, we apply lemma 4.12 which tells us that $\|I_{h,n}^{-1}(g)\|_\infty \leq 1/\sqrt{h_n}\|g\|_\infty$. This tells us that

$$\|A_n^h I_{h,n}^{-1}(I_n^h - I_0^h)(g_{0t}^h)\|_\infty = O\left(\frac{1}{\sqrt{nh_n^9}}\right).$$

This completes the proof of lemma 4.4.

### 4.7.3 Proof of lemma 4.8.

Lemma 4.8 will be proved as a corollary of the next lemma.

**Lemma 4.14** *Let $C \subset L^2(F_0)$ be any compact set in $L^2(F_0)$. Then we have:*

$$\sup_{g\in C} \|(I_0^h - I_0)(g)\|_{F_0} \to 0, \tag{4.27}$$

*and*

$$\sup_{g\in C} E\left(A_0^h(g) - A_0(g)\right)^2 \to 0 \text{ for } h = h_n \to 0.$$

**Proof.** By the compactness of $C$ and the continuity of $I_0^h : L^2(F_0) \to L^2(F_0)$ the supremum in (4.27) is attained by some $g_0 \in C$. Let $g_k$ be a sequence so that $\|g_k - g_0\|_{F_0} \to 0$ and $\|g_k\|_\infty < \infty$ for $k = 1, 2, \ldots$. We have:

$$\|(I_0^h - I_0)(g_0)\|_{F_0} \leq \|(I_0^h - I_0)(g_0 - g_k)\|_{F_0} + \|(I_0^h - I_0)(g_k)\|_{F_0}.$$

$\|(I_0^h - I_0)(g_0 - g_k)\|_{F_0} \leq 2\|g_0 - g_k\|_{F_0}$ which converges to zero for $k \to \infty$. Therefore it suffices now to show that $\|(I_0^{h_n} - I_0)(g_k)\|_{F_0} \to 0$ for each fixed $k$. Now, we have:

$$(I_0^h - I_0)(g_k) = A_0^{h\top}(A_0^h - A_0)(g_k) + (A_0^{h\top} - A_0^\top)(A_0(g_k)).$$

The difference in the first term are comparable because all can be considered as functions of $(C, T)$ and thereby are defined on the same probability space. Firstly, we will consider the second term. It suffices to deal with one of the singly censored terms. Let $d = (1, 0)$ and $f_k \equiv A_0(g_k)I(D = d)$. We have:

$$(A_0^{h\top} - A_0^\top)(f_k)(T_1, T_2) = \int_0^{T_2} f_k(T_1, v)(G_h - G_0)((T_1, \infty), dv).$$

Let $T = (T_1, T_2)$ be fixed and let $T_2$ be a point where $H_0(T_1, \Delta T_2) = 0$. By definition of weak convergence of $H_h(T_1, dv)$ to $H_0(T_1, dv)$ we have now that if $v \to f_k(T_1, v)$ is bounded and continuous $H_0(T_1, \cdot)$ a.e., then $(A_0^{h\top} - A_0^\top)f_k(T_1, T_2) \to 0$ for this $T$. The boundedness follows from: $\|f_k\|_\infty \leq \|g_k\|_\infty < \infty$. We have that $v \to f_k(T_1, v)$ is given by:

$$v \to \frac{\int_v^\infty g_k(T_1, v_2)F_{01}(T_1, dv_2)}{F_{01}(T_1, (v, \infty))}.$$

This function is continuous at $v$ if $v \to F_{01}(T_1, v)$ is continuous at $v$. Consequently, we need that $F_{01}(T_1, dv)$ puts no mass at a point where $H_0(T_1, dv)$ puts mass. By our convention that if $T = C$, then the observation is uncensored, this is satisfied. This proves the pointwise convergence of $f_h \equiv (A_0^{h\top} - A_0^\top)(f_k)$ to zero $F$-a.e. We need to show that $\int f_h^2 dF_0 \to 0$. However, we also have $\|f_h\|_\infty \leq 2\|g_k\|_\infty$ and therefore the dominated convergence theorem provides us with $\int f_h^2 dF_0 \to 0$.

Let's now consider the first term $A_0^{h\top}(A_0^h - A_0)(g_k)$. Because $A_0^h$ is a conditional expectation its second moment is bounded by the second moment of $(A_0^h - A_0)(g_k)$. Therefore it suffices to show that $E_{X,C}\left((A_0^h - A_0)(g_k))^2\right) \to 0$ for $h \to 0$, where we consider $A_0^h$ and $A_0$ as functions in $(T, C)$ via $Y^h$ and $Y$, respectively.

Recall how we constructed the data $(\widetilde{T}, D)^h$: 1) we have a nested sequence of partitions $\pi^h$ and we observed i.i.d. $C_1, \ldots, C_n \sim G$, 2) Now, we discretize

$C_i$ such that $C_i^h \sim G_h$ where $G_h$ lives on $\pi^h$. This provides us with data $(\widetilde{T}, D)_h \sim P_{F_0,G_h}$. 3) Finally we discretized $(\widetilde{T}, D)_h$ in order to obtain $Y^h = (\widetilde{T}, D)^h \sim P^h_{F_0,G_h}$. Denote the sigma-field generated by $Y^h$ with $\mathcal{A}^h$. Because $\pi^h$ is nested and the sigma field generated by $\pi^h$ converges to the Borel sigma-field on $[0, \tau]$ we have that $\mathcal{A}^h \uparrow \mathcal{A}^\infty$ for $h \to 0$, where $\mathcal{A}^\infty$ is the sigma field generated by $Y = (\widetilde{T}, D))$, $Y \sim P_{F_0,G_0}$.

Consequently $M_{h_n} \equiv E_{X,C}(g_k(T) \mid \mathcal{A}^{h_n})$ is a martingale in $n$ and it is well known that if $\sup_h E(M_h^2) < \infty$, then $E((M_h - M_0)^2) \to 0$. We have $\sup_h E(E(g_k(T) \mid \mathcal{A}^h)^2) \leq \|g_k\|_\infty < \infty$ and consequently we have $\|(A_0^h - A_0)(g_k)\|_{F_0 \times G_0} \to 0$. This also proves the second statement in lemma 4.14. □

**Corollary 4.2** *We make the same assumptions as in lemma 4.14. For each set $C \subset L^2(F_0)$ which is compact w.r.t. $\|\cdot\|_{F_0}$ we have for $h \to 0$:*

$$\sup_{g \in C} \|(I_0^{-1} - I_{h,0}^{-1})(g)\|_{F_0} \to 0. \tag{4.28}$$

*This implies*

$$\sup_{g, g_1 \in C} \left| \langle A_0^h I_{h,0}^{-1}(g), A_0^h I_{h,0}^{-1}(g_1) \rangle_{P_0^h} - \langle A_0 I_0^{-1}(g), A_0 I_0^{-1}(g_1) \rangle_{P_{F_0}} \right| \to 0.$$

*Moreover, we have*

$$\sup_{g \in C} E \left( A_0^h I_{h,0}^{-1}(g) - A_0 I_0^{-1}(g) \right)^2 \to 0.$$

**Proof.** We have:

$$\begin{aligned} (I_{h,0}^{-1} - I_0^{-1})(g) &= I_{h,0}^{-1} \left( I_0 I_0^{-1} - I_0^h I_0^{-1} \right)(g) \\ &= -I_{h,0}^{-1}(I_0^h - I_0) I_0^{-1}(g). \end{aligned}$$

Firstly, notice that by the bounded $L^2$-invertibility of $I_0$ (lemma 4.1) $I_0^{-1}(C)$ is compact in $L^2(F_0)$. Now, by the preceding lemma we have that $\sup_{g \in C} \|(I_0^h - I_0) I_0^{-1}(g)\|_{F_0} \to 0$. Finally, we know by lemma 4.1 that $\sup_h \|I_{h,0}^{-1}\|_{F_0} < \infty$. This proves the first statement. For the second statement notice that:

$$\begin{aligned} \langle A_0^h I_{h,0}^{-1}(g), A_0^h I_{h,0}^{-1}(g_1) \rangle_{P^h_{F_0}} &= \langle I_{h,0}^{-1}(g), g_1 \rangle_{F_0} \\ &= \langle I_{h,0}^{-1}(g) - I_0^{-1}(g), g_1 \rangle_{F_0} + \langle I_0^{-1}(g), g_1 \rangle_{F_0}. \end{aligned}$$

The first term converges to zero by the Cauchy-Schwarz inequality and (4.28). The second term equals $\langle A_0 I_0^{-1}(g), A_0 I_0^{-1}(g_1) \rangle_{P_{F_0}}$.

It remains to prove the last statement. By the compactness of $C$ and continuity of $A_0 I_0^{-1}$ and $A_0^h I_{h,0}^{-1}$ it suffices to show the statement for a fixed $g \in L_0^2(F_0)$. We have

$$A_0^h I_{h,0}^{-1}(g) - A_0 I_0^{-1}(g) = (A_0^h - A_0) I_0^{-1}(g) + A_0^h (I_{h,0}^{-1} - I_0^{-1})(g).$$

The first term converges to zero by the second statement of lemma 4.14.

For the second term we have:

$$\|A_0^h(I_{h,0}^{-1} - I_0^{-1})(g)\|_{P_0^h} \leq \|(I_{h,0}^{-1} - I_0^{-1})(g)\|_{F_0} \to 0 \text{ by (4.28).}\square$$

Notice that $C \equiv \{I(0,t] : t \in [0,\tau]\} \subset L^2(F_0)$ is a compact set. Application of the corollary to this set $C$ provides us with lemma 4.8.

# Chapter 5

# Efficiency of the NPMLE in the Line-Segment Problem.

## 5.1 Introduction to the line-segment process problem.

The spatial line-segment process problem, observing line-segments in a two dimensional window, was introduced by Laslett (1982). Laslett derives the log likelihood for this spatial problem and shows how a version of the EM-algorithm can be used to find the NPMLE, but the behavior of the NPMLE has not been studied. Wijers (1991) considers the one-dimensional line-segment problem, so now we observe line-segments on the real line through an interval, and shows how it can be formulated as a nonparametric missing data model and thereby that the NPMLE can be characterized by the self-consistency equation as introduced by Efron (1967) (see Gill, 1989). By using an elegant technique based on the log likelihood he proves uniform consistency of the NPMLE. In this chapter we prove efficiency (and bootstrap results also follow easily from the analysis) of a "sieved"-NPMLE for the one-dimensional line-segment problem, where sieved means that we maximize the loglikelihood over all distributions which put mass on the uncensored observations.

In Gill, van der Laan, Wijers (1995) a self-contained overview of Laslett (1982), Wijers (1991) and this chapter is given and it is made clear how and how far the analysis followed in this chapter can be generalized to the spatial line-segment problem. In the spatial line-segment problem there is an additional

unknown parameter, namely the distribution of the orientation of the line-segments. For a known orientation distribution and convex window the results for the NPMLE can be proved as essentially carried out in Wijers (1994) (he did it for a circular window). Suggestions for solving the unknown orientation distribution case are given using an extended identity as proved in van der Laan (1994). Also the NPMLE for non-convex windows is a completely open problem. A version of Gill, van der Laan, Wijers (1995) is also found in Gill (1993).

The one-dimensional line-segment problem has the following statistical motivation. Suppose one is interested in the time a specific patient spends in the hospital. For this purpose one observes incoming and outgoing patients over a period $[0, \tau]$ of time-length $\tau$. The variable of interest is the time $X$ between the arrival time and the departure time of the patient. One can identify with each patient a line-segment with length $X$ and start-point $T$, where $T$ is the arrival time of the patient. If the arrival time is smaller than 0, then the line-segment is left-censored and if the departure time is larger than $\tau$, then the line-segment is right-censored. Hence one will observe four kind of observations on the line-segment: singly left-censored, singly right-censored, doubly censored and uncensored. The goal is to estimate the distribution of the lenght $X$ of the line-segments from these observations. The one-dimensional line-segment problem has also several economic applications. For example, if one is interested in estimation of the disgtribution of the time of unemployment and one has only information available over a period $[0, \tau]$, then the same model applies.

One clear feature of this model is that $X$ is censored. In particular, if follows that one is not able to estimate the distribution (of length) after $\tau$. Another less obvious feature of this problem is that long line-segments are more likely to be observed (i.e. to hit the window $[0, \tau]$) than short line-segments and therefore the observed line-segments cannot be considered as an i.i.d. sample from the random process which generates the line-segments. So there is a so called *length bias* problem; the empirical distribution function of all complete lengths of all (partially) observed line-segments does not converge to the distribution of the length of the line-segments.

We will assume that the starting points, say $T$, of the line-segments follow a homogeneous Poisson point process on $\mathbb{R}$ with rate $\lambda$. Furthermore, assume that the length $X > 0$ corresponding with the line-segment starting at $T$ is independent of the Poisson process and has the common distribution $F$. Preferably, one would like to have an estimator available which does not

utilize knowledge on the starting point distribution. In other words, just as in the univariate right-censoring model and many other missing data models, one prefers to model the distribution of $(T, X)$ such that the likelihood of the data factorizes in a part which only depends on $F$ and a part which only depends on the distribution of $T$ so that knowledge on the distribution of $T$ is irrelevant for computation of the NPMLE and for the information bound. This can only be achieved in missing data models where the observation $Y$ is equivalent with observing that $X \in D_1(Y)$ and $T \in D_2(Y)$ for some regions $D_1(Y)$ and $D_2(Y)$. However, in the line-segment model the censoring by the interval $[0, \tau]$ causes a dependence between $T$ and $X$ regionwise, just as it does for mixture models where one observes a convolution of two variables.

In our last section we discuss how our results can be extended to the case where $T$ follows an inhomogeneous poisson process with a known or estimated rate $\lambda(t)$.

It can now be shown (Karlin, 1981, Stoyan, 1987) that $(T, X)$ follows a Poisson point process on $\mathbb{R} \times \mathbb{R}_{\geq 0}$ with intensity measure

$$\lambda dt dF(x).$$

Let $B$ be the set of all $(X, T)$ for which the corresponding line-segment hits $[0, \tau]$. If $\mu$ is the mean of $F$, then $\int_B dF(x)dt$ equals $\tau + \mu$, as can be trivially verified.

A well known fact about Poisson processes tells us that if we condition on the number of $(X, T) \in B$, then these $(X, T) \in B$ can be represented as i.i.d. observations from a distribution given, for $A \subset B$, by:

$$P((X, T) \in A) = \frac{\lambda \int_A dF(x)dt}{\lambda \int_B dF(x)dt} = \frac{\int_A dF(x)dt}{\tau + \mu}.$$

However, the latter we can rewrite as:

$$\frac{\int_A dF(x)dt}{\tau + \mu} = \int_A \frac{(\tau + x)dF(x)}{\tau + \mu} \frac{dt}{\tau + x},$$

where $(\tau + x)dF(x)/(\tau + \mu)$ corresponds with a probability distribution because it integrates to 1 and $dt/(\tau + x)$ is the density of the uniform distribution over $(-x, \tau)$. Consequently, we can represent the distribution of the observed line-segments as follows: the lengths $X$ of the line-segments which hit $[0, \tau]$ (so which are at least partially observed) have distribution $V$ with

$$dV(x) \equiv (\tau + x)dF(x)/(\tau + \mu), \tag{5.1}$$

and the starting point $T$, given $X = x$, is uniformly distributed over $(-x, \tau)$. This does not describe, yet, the distribution of the data because many of these $(X, T)$ are censored.

$V$ is called the *length biased version* of $F$ because it takes into account that long line-segments are more likely to hit the window $[0, \tau]$ than the short line-segments.

Heuristically one expects to be able to estimate $V$ on $[0, \tau)$ by using the uncensored and singly censored line-segments. Because of the $\mu$ in the relation (5.1) an estimator of $V$ on $[0, \tau)$ does not uniquely determine an estimator of the parameter of interest $F$ on $[0, \tau)$. However, we also observe the fraction of doubly censored observations. Let

$$h_V \equiv P(T < 0, X + T \geq \tau) = \int_{[\tau,\infty)} \frac{x - \tau}{\tau + x} dV(x))$$

be the probability that the line-segment will be doubly censored. If it is not confusing, then we will skip the $V$ in the notation $h_V$. As shown in Wijers (1991) and easy to verify we have the following relation:

$$2\tau/(\tau + \mu) = 1 - h_V + \int_{(0,\tau)} (\tau - x)/(\tau + x) dV(x). \tag{5.2}$$

This tells us that estimators of $V$ on $[0, \tau)$ and $h$ (for $h$ the NPMLE is simply the fraction of doubly censored line-segments) determine estimators of the parameter of interest $F$ on $[0, \tau)$ and its mean $\mu$.

Moreover, Wijers (1991) shows that this relation between $F, \mu$ and $V, h$ is 1-1 and onto: let $V$ and $F$ be distributions on $[0, \infty)$, we have that each $V$ on $[0, \tau)$ and $h$ determine uniquely a $F$ on $[0, \tau)$ and its mean $\mu < \infty$ and each $F$ on $[0, \tau)$ and its mean $\mu < \infty$ determine uniquely a $V$ on $[0, \tau)$ and its $h$. Hence the collection of possible $V$'s on $[0, \tau)$ as defined by (5.1) obtained by varying $F$ nonparametrically (i.e. over all distributions) consists of all possible subdistributions on $[0, \tau)$. So instead of parametrizing the length biased distribution $V$ as $dV(x) = (\tau + x)dF(x)/(\tau + \mu)$, $F$ completely unknown with finite mean $\mu$, we can replace this model by all distributions $V$ and concentrate on estimating $V$ on $[0, \tau)$ and $h \in [0, 1]$. This leads to the following formal model of the line-segment problem due to Wijers (1991):

**Formal description of line-segment model.** We can model the line-segment problem as follows: $X_1, \ldots, X_n$ are $n$ i.i.d. real valued random variables with distribution function $V_0$, which is completely unknown. $T_1, \ldots, T_n$ are $n$ i.i.d. real valued random variables, and given $X_i$ they are uniformly

distributed over $(-X_i, \tau)$. We are concerned with estimation of $(V_0, h_0)$, using observations $Y_i = \Phi(X_i, T_i)$, where $\Phi(X_i, T_i)$ is a many to one mapping of these random variables $(X_i, T_i)$. In order to describe this mapping $\Phi$ we need to define the following sets which form a partition of the probability space for $(X, T)$ (here we follow the notation of Wijers, 1991):

$$\begin{aligned} A_1 &\equiv \{(X,T):\ T<0,\ 0<X+T<\tau\} \\ A_2 &\equiv \{(X,T):\ T<0,\ \tau\le X+T\} \\ A_3 &\equiv \{(X,T):\ 0<T<\tau,\ \tau\le X+T\} \\ A_4 &\equiv \{(X,T):\ 0<T<\tau,\ 0<X+T<\tau\}. \end{aligned}$$

$T$ is the starting point of the line-segment and $X + T$ is the right end point of the line-segment. Consequently if $T > 0$ and $X + T < \tau$ (i.e. $(X, T) \in A_4$), then the line segment is completely observed. If $T < 0$ and $0 < X + T < \tau$ then it is left censored, but not right censored $(A_1)$; if $T < 0$ and $X + T \geq \tau$, then it is doubly censored $(A_2)$; if $0 < T < \tau$ and $X + T \geq \tau$, then it is right censored, but not left censored $(A_2)$.

Let $Y = (\widetilde{X}, D) = \Phi(X, C) \in (0, \tau] \times \{0, 1a, 1b, 2\}$ be the many to one mapping $\Phi$ from $(X, C)$ to $(0, \tau] \times \{0, 1a, 1b, 2\}$ described as follows:

$$Y = (\widetilde{X}, D) = \begin{cases} (T+X, 1a) & \text{if } (X,T) \in A_1 \\ (\tau, 2) & \text{if } (X,T) \in A_2 \\ (\tau - T, 1b) & \text{if } (X,T) \in A_3 \\ (X, 0) & \text{if } (X,T) \in A_4 \end{cases}$$

(The definition of $D$ here is temporary and will be modified in a moment.) Observing $Y$ means now that we know if the line-segment is left, right, doubly or uncensored ($D$ tells us this) and we observe a number $\widetilde{X} \in (0, \tau]$, the length of the intersection of the line-segment with the window, of which we know how it depends on $X$ and $T$. We observe $Y_i = (\widetilde{X}_i, D_i)$, $i = 1, 2, \ldots, n$. This is a *missing data model*: each observation $Y_i$ tells us that $(X_i, T_i)$ has fallen in the region $\Phi^{-1}(Y_i)$. These regions are given by:

$$\Phi^{-1}(\widetilde{X}, D) = \begin{cases} \{(X,T):\ X = \widetilde{X},\ 0<T<\tau-\widetilde{X}\} & \text{if } D = 0 \\ \{(X,T):\ X+T = \widetilde{X},\ T\le 0\} & \text{if } D = 1a \\ \{(X,T):\ T = \tau-\widetilde{X},\ X\ge\widetilde{X}\} & \text{if } D = 1b \\ A_2 & \text{if } D = 2 \end{cases}$$

Notice that $h_0 = P_{V_0}(Y \in A_2)$ and therefore a trivial and in fact the only sensible estimate of $h_0$ is the fraction of doubly censored line segments, which is also the NPMLE as we will see in the sequel.

**Grouping together the singly censored observations.** $D = 0$ corresponds with the uncensored observed line-segments in $(0, \tau)$; $D = 1a, 1b$ with the singly left-censored and singly right-censored observed line-segments; $D = 2$ with the doubly (at left and right) censored observed line-segments. As expected, the distribution of $(\widetilde{X}, 1a)$ equals the distribution of $(\widetilde{X}, 1b)$. Therefore, it makes sense to group together these two kinds of observations to one kind of observation. $D = 0$ for uncensored (as above), $D = 1$ for singly censored ($D = 1a, 1b$ above) and $D = 2$ for doubly censored (as above). The distribution of $(\widetilde{X}, 1)$ equals now two times the distribution of the preceding $(\widetilde{X}, 1a)$, or preceding $(\widetilde{X}, 1b)$. Denote the probability distribution of the data with:

$$P_{V_0}(y, d) \equiv P_{V_0}(\widetilde{X} \leq y, D = d), \ d \in \{0, 1\}$$

and $h_0 = P_{V_0}(D = 2)$. In formulas we have:

$$\begin{aligned} P_{V_0}(d\widetilde{x}, 0) &= \frac{\tau - \widetilde{x}}{\tau + \widetilde{x}} dV_0(\widetilde{x}) && (5.3) \\ P_{V_0}(d\widetilde{x}, 1) &= 2 \int_{[\widetilde{x}, \infty)} \frac{dV_0(x)}{\tau + x} d\widetilde{x} && (5.4) \\ &\equiv 2g_0(\widetilde{x}) d\widetilde{x} && (5.5) \\ h_0 &= P_{V_0}(D = 2). \end{aligned}$$

In order to determine the exact integration area one should realize that for computing probabilities for censored line-segments one should also integrate over the edges; a line-segment which ends exactly at $\tau$ is observed as right-censored at $\tau$. The density (w.r.t. the Lebesgue measure) $g_0$ will appear in our analysis in denominators and therefore plays a crucial role. Define $\overline{V}_0(t) \equiv 1 - V_0(t-)$. The distribution of the data is uniquely determined by $V_0$ on $[0, \tau)$ and $h_0$. This follows from the following identity

$$\overline{V}_0(\tau) = 2\tau g_0(\tau) + h_0, \tag{5.6}$$

which is found by a simple rewriting of $h_0 + P_{V_0}([0, \tau), 0) + P_{V_0}([0, \tau), 1) = 1$. Therefore we can parametrize the distribution of the data as $P_{V_0, h_0}$, where $V_0$ is restricted to $[0, \tau)$, $P_{V_0, h_0}(D = 2) = h_0$ and $P_{V_0, h_0}(y, d) = P_{V_0}(y, d)$, $d \neq 2$, as defined by (5.3) and (5.4). Moreover, (5.6) tells us also that if $\overline{V}_0(\tau)$ is fixed, then $h_0$ can still have any value between 0 and $\overline{V}_0(\tau)$, which means that $V$ as subdistribution on $(0, \tau]$ and $h_V$ can vary quite freely in the parametrization $P_{V,h}$.

Let's now formally write down the *model.* Let $\mathcal{F}_\tau$ be the collection of all subdistributions on $[0,\tau) \subset \mathbb{R}_{\geq 0}$. Then the model is given by:

$$\mathcal{M} \equiv \left\{ P_{V,h} : V \in \mathcal{F}_\tau, h \in [0, \overline{V}(\tau)] \right\}.$$

We only have uncensored ($D = 0$) observations $X_i = \widetilde{X}_i$ on $(0,\tau)$. Therefore, there is only hope to estimate $V_0$ well on $[0,\tau)$ (of course, the model only allows $V_0$ which give probability zero to line-segments with length 0). It can be easily shown by applying theorem 3.1 of van der Vaart (1991) that the parameter $\vartheta(P_{V,h}) = V(\tau-)$ is *not* pathwise differentiable and thereby that there exist no regular estimators of $V_0(\tau-)$ (see Gill, 1993a,b) and it does not help to assume that $V_0$ is continuous at $\tau$. It follows from (5.4) that estimating $g(\tau)$ and hence estimating $V(\tau-)$ is equivalent with estimating the density of the singly-censored line-segments at $\tau$. This explains why $g(\tau)$ and $V(\tau-)$ are not estimable at root-$n$ rate.

In order to suppress the influence of the irregularity of $V_0(\tau-)$ we consider estimation of the following parameter:

$$\vartheta(P_{V_0,h}) = (W_0, h_0) \equiv \left( \int_0^{(\cdot)} (\tau - x) dV_0(x), h_0 \right) \in D[0,\tau] \times [0,1],$$

where $D[0,\tau]$ is defined as the space of cadlag (right-continuous with left-hand limits) on $[0,\tau]$. Because $P_{V_0}(t,0) = \int_{(0,t]} \frac{\tau - x}{\tau + x} dV_0(x)$ (the distribution of the uncensored observations) there exists a regular estimator of $W_0$, namely $\int_0^t (\tau + x) dP_n(x,0)$, where $P_n(\cdot,0)$ is the empirical subdistribution of the uncensored ($D = 0$) observations.

This parameter is also exactly the parameter we need for estimating $F_0$ on $[0,\tau-\epsilon]$ for all $\epsilon > 0$ and $\mu_0$. This is seen as follows: We will prove that the NPMLE $(W_n, h_n)$ is an efficient estimator of $(W_0, h_0) \in D[0,\tau] \times \mathbb{R}$. In order to show that the corresponding $\mu_n$ and $F_n$ on $[0,\tau-\epsilon]$ are efficient for all $\epsilon > 0$, it suffices to show that $(\mu, F)$ can be written as a compactly differentiable functional of $(V,h)$ (van der Vaart, 1991). Because of the relation $dV_0(x) = (\tau + x) dF_0(x)/(\tau + \mu)$ it suffices to show that $\mu$ is a compactly differentiable functional of $(W,h)$. We have the relation $P_{V_0}(A_3 \cup A_4) = P_{V_0}(A_1 \cup A_4) = \tau/(\tau + \mu)$. The latter tells us that $2\tau/(\tau + \mu) = 1 - h_0 + P_{V_0}(A_4)$. However, $P_{V_0}(A_4) = \int_{(0,\tau)} (\tau - x)/(\tau + x) dV_0(x) = \int_{(0,\tau)} dW_0(x)/(\tau + x)$. It follows that $\mu$ is a compactly differentiable functional of $(W_0, h_0) \in D[0,\tau] \times \mathbb{R}$ and therefore efficiency of $(W_n, h_n)$ provides us with efficiency of $\mu_n$ and $F_n$ on $[0,\tau-\epsilon]$ for all $\epsilon > 0$.

The problem of establishing weak convergence at root-$n$ rate in this model has appeared to be difficult and seems not possible by just using the self-consistency equations. However, the *identity* approach as followed in the preceding chapters , followed here is successful. Let's first try to understand why this estimation problem is rather difficult. In chapter 3 we studied a general class of nonparametric missing data models and explained that there are essentially 2 crucial assumptions, 1 and 2, under which the NPMLE will be efficient. The second assumption says that the censored regions for $X$ implied by an observation $Y$ should have positive probability w.r.t. $V$; this is clearly satisfied. Assumption 1 required that $P(Y \in A_4 \mid X = x) > \delta > 0$, which says that, given $X = x$, the probability that the line-segment $(X, T)$ will be uncensored is larger than $\delta > 0$ $V_0$ a.e. Assumption 1 guarantees that the information operator has a bounded $L^2$-inverse (e.g. see van der Laan, 1993b). For $X > \tau$ this probability is zero. This means that the information operator is not invertible and therefore we firstly need to recover the essential part of the information operator for estimation on $[0, \tau)$. Indeed, as will appear, this is one important ingredient of the analysis. So let's now verify assumption 1 for $x \in [0, \tau)$. $P(Y \in A_4 \mid X = x)$ equals $(\tau - x)/(\tau + x)$ and therefore converges to zero for $x \to \tau$. Heuristically this means that there will be hardly any uncensored observations close to $\tau$ and therefore it is very hard to estimate $V_0$ close to $\tau$. But if, for example, $V_0(\tau - \epsilon, \tau] = 0$ for certain $\epsilon > 0$, then assumption 1 is satisfied. However, because $\tau$ certainly does not represent the tail of $V_0$ this is an unacceptable assumption, and the assumption can also not be easily arranged by artificial censoring as in the examples in chapter 3. Consequently, in the analysis we will have to deal with an inverse of an information operator with singularity $1/(\tau - x)$. By exploiting a well understood *Volterra* structure which appears in the information operator we are able to show that this singularity is not disturbing for any parameter which does not use the value $V(\tau)$ and hence in particular for $W$ on $[0, \tau]$.

Our results are based on the assumption that $g_n(\tau)$ is consistent. It appeared that there was a mistake in the proof of this result in Wijers (1991) so that this result has not been established; the consistency on $[0, \tau - \epsilon)$ of $V_n$ still holds. However, Wijers (1993) proposed a slight modification of the data by censoring the line-segments by $[0, \tau - epsilon]$ for some $\epsilon > 0$ in such a way that $g_n(\tau - \epsilon)$ is consistent and hence that our result applies to the sieved-NPMLE of this transformed data. We decided not to change all our results to this setting, but discuss the transformation in section 6.

We will now give an overview of this chapter. In the next section we define

the "sieved"-NPMLE as the maximizer of the log likelihood over the class of discrete distributions which put only mass on the observed line segments and we discuss existence, uniqueness and the EM-algorithm for computation of the estimator in practice. In section 3 we formulate the efficiency theorem and give the general efficiency proof and we specify the three ingredients *identity condition*, *Donsker class condition* and the *$\rho$-consistency condition* which need to be proved in the subsequent sections. We also discuss the bootstrap and estimation of the limit distribution for construction of confidence intervals, where it is shown in subsection 4.3 that the latter can be carried out very quickly. In section 4 we prove the Donsker class condition by proving that the so called efficient influence function is of bounded variation uniformly in $V$ and $t$. The proof consists of two parts. Firstly, we have to split the inversion of the information operator for functions on $[0, \tau)$ and for a remaining non-unique easy part. Secondly, we prove that the first part (the hardest) can be nicely inverted w.r.t. the supnorm and variation norm. In section 5 we prove the $\rho$-consistency condition and identity condition.

## 5.2 Existence and uniqueness of the sieved-NPMLE, EM-equations.

Let $P_n$ be the empirical distribution of $Y_i$, $i = 1, \ldots, n$ and let $m = m(n)$ be the number of uncensored line-segments $X_i$ (i.e. $D_i = 0$). Denote the counting measure on these $X_i$ with $\mu_m$. Let $\mathcal{F}_\tau(\mu_m)$ be those $F \in \mathcal{F}_\tau$ with $F \ll \mu_m$. For a $V \in \mathcal{F}_\tau(\mu_m)$ and $h$ we define $p_{V,h}(y,d)$, $d \neq 2$, as the density of $P_{V,h}$ w.r.t. $(\mu_m \times dt)\Phi^{-1}$, where $dt$ stands for the Lebesgue measure. Define

$$B_m \equiv \left\{(V,h) : V \in \mathcal{F}_\tau(\mu_m), h \in [0, \overline{V}(\tau)]\right\}.$$

Now, we define a sieved-NPMLE as follows

$$(V_n, h_n) = \arg\max_{B_m} \int \log(p_{V,h}) dP_n.$$

By substitution of (5.3) and (5.4) it follows that the empirical loglikelihood is given by:

$$\frac{1}{n}\sum_{i=1}^{n}\left(I(D_i = 0)\frac{\tau - \widetilde{X}_i}{\tau + \widetilde{X}_i}V(\{\widetilde{X}_i\}) + I(D_i = 1)2g(\widetilde{X}_i)\right) + hP_n(D = 2), \quad (5.7)$$

where $g(x) = g(\tau) + \int_x^\tau dV(u)/\tau + u$ and $g(\tau)$ can be expressed in $V, h$ by (5.6). Assuming $g_0(\tau) > 0$ Wijers (1991) shows existence and uniqueness of

the NPMLE $(V_n, h_n)$. His proof is also applicable to this sieved-NPMLE. We will now derive the score-equations and in particular the EM-equations. Recall that $V_n \in \mathcal{F}_\tau(\mu_m)$ is a subdistribution on $[0, \tau)$. Consider the class of lines $\epsilon V_1 + (1-\epsilon)V_n$ through $V_n \in \mathcal{F}_\tau(\mu_m)$ with scores $g$ in $L_0^2(V_n)$ with finite supnorm; so for all the lines $V_{n\epsilon,g}$ we have that $\overline{V}_{n\epsilon,g}(\tau) = \overline{V}_n(\tau)$. Therefore $h$ can still vary freely. Furthermore, consider the following one dimensional submodel through $h_n$ with score $(I(D=2) - h_n)$:

$$h_{n\epsilon} = (1 + \epsilon\,(I(D=2) - h_n))\, h_n \in [0, \overline{V_n}(\tau)).$$

Differentiating of $\int \log(p_{V_n,h_{n\epsilon}})dP_n$ w.r.t. $\epsilon$ provides us with:

$$h_n = P_n(D=2), \text{ the fraction of doubly censored observations.}$$

Differentiating of $\int \log(p_{V_{n\epsilon,g},h_n})dP_n$ provides us with the familiar score operator (see Gill, 1989, Bickel et al., 1993, section 6.6) $A_{V_n}(g) : L_0^2(V_n) \to L_0^2(P_{V_n,h_n})$ for missing data models:

$$E_{V_n,h_n}(g(X) \mid Y) = 0 \text{ for } \|g\|_\infty < \infty. \tag{5.8}$$

For a score $g(x) = I_{(0,t]}(x) - V_n(t)$, $t \in [0, \tau)$, this equation reduces to the self-consistency equation (Gill, 1989) as written out below. Finally, we have the relation (5.6) for determining $g_n(\tau)$.

*Computation of sieved-NPMLE.* Let $v_n \equiv dV_n/d\mu_m$ be the point masses of $V_n$ on the observed $X_i$ and let $P_{n0}(x) = 1/n \sum_{i=1}^n I(\widetilde{X}_i \le x, D_i = 0)$, $P_{n1}(x) = 1/n \sum_{i=1}^n I(\widetilde{X}_i \le x, D_i = 1)$ be the empirical distributions of the uncensored and singly-censored line-segments. We conclude that we have the following set of EM-equations:

$$\begin{aligned}
v_n(x_i) &= \sum_{j=1}^n \frac{1}{n} P_{V_n}(X = x_i \mid Y_j) \\
&= P_{n0}(\{x_i\}) + \frac{v_n(x_i)}{\tau + x_i} \left( \frac{1}{n} \frac{\sum_{j=1}^n I(D_j = 1, \widetilde{X}_j \le x_i)}{\int_{x_j}^\tau 1/(\tau+u) dV_n(u) + g_n(\tau)} \right) \\
h_n &= P_n(D=2), \text{ the fraction of doubly censored} \\
\overline{V}_n(\tau) &= 2\tau g_n(\tau) + h_n.
\end{aligned}$$

The solution of the equations can be found with the EM-algorithm (Dempster et al., 1977, Turnbull, 1978, Meilijson, 1989): initiate the right-hand side of the first equation with $v_n^0$ and $g_n^0(\tau)$, this provides us with a $v_n^1$, now obtain a $g_n^1(\tau)$ using the third equation, and repeat these steps till convergence is established. For a much faster algorithm to solve these equations see Gill (1993a,b).

In practice, one might find that $V_n$ becomes quite noissy close to $\tau$, remember that $V(\tau-)$ is not root-n estimable, which will certainly also be reflected in the constructed confidence bands for which methods will be given. Therefore, interpolation of $V_n$ on $[0, \tau - \epsilon)$ to $V_n(\tau)$ is a sensible method for obtaining a more reliable estimate close to $\tau$.

**Remark: Parametric model.** Suppose now that one wants to assume a parametric model $F_\theta$ for $F$. One obtains the parametric likelihood by replacing $dV(u) = \tau + u/\tau + \mu dF_\theta(x)$ in the nonparametric likelihood (5.7) and in particular in the expression of $g(x) = \int_x^\infty 1/\tau + x dV(x)$ and in the expression of $h$. Then the likelihood depends only on $\theta$. One can now compute the $k$ score equations corresponding with differentiation of the loglikelihood with respect to the $k$ coordinates of $\theta$. This provides us with a $k$-equations with $k$ unknowns which can be solved with Newton-Raphson or other numerical procedures. In this case it is not necessary to split up the distribution of the data in $h$ and $V$ though it seems natural to set $h$ equal to $h_n = P_n(D = 2)$ and maximize over $V_\theta$ restricted to $[0, \tau]$.

## 5.3 Efficiency result and outline of proof.

From now on $V_0$ will denote the underlying distribution on $\mathbb{R}$ and we will parametrize the distribution of the data with $P_{V_0}$, $V_0 \in \mathcal{F}$ again, instead of $P_{V_0,h_0}$, $V_0 \in \mathcal{F}_\tau$. Define $\mathcal{S}(V_0)$ as the class of all lines $\epsilon V + (1-\epsilon)V_0$ through $V_0$, $V \ll V_0$, which are submodels by convexity of $\mathcal{F}$, with score $g \in L_0^2(V_0)$. Let $S(V_0) \subset L_0^2(V_0)$ be the corresponding tangent cone (=set of corresponding scores) and $T(V_0)$ be the tangent space (=closure of linear extension of $S(V_0)$). It is easy to see that $T(V_0) = L_0^2(V_0)$. As discussed in the preceding section each one dimensional submodel $P_{V_0,\epsilon,g}$ has a score given by $A_{V_0}(g)$, where $A_{V_0}$ is the so called score operator:

$$A_{V_0} : L_0^2(V_0) \to L_0^2(P_{V_0}) : A_{V_0}(g)(Y) = E_{V_0}(g(X) \mid Y).$$

The Cramér-Rao lower bound for estimation of $\Psi(\epsilon) \in \mathbb{R}$ for some $\Psi$ at $\epsilon = 0$ along the one-dimensional model $P_{V_0,\epsilon,g}$ equals:

$$\frac{(d/d\epsilon \Psi(\epsilon) \mid_{\epsilon=0})^2}{\|A_{V_0}(g)\|^2}. \tag{5.9}$$

Let

$$A_{V_0}^\top : L_0^2(P_{V_0}) \to L_0^2(V_0) : \; A_V^\top(\eta)(X) = E_{V_0}(\eta(\widetilde{X}, D)|X)$$

be the adjoint of $A_{V_0}$. Now, the information operator is defined by:

$$I_{V_0} = A_{V_0}^{\top} A_{V_0} : L_0^2(V_0) \to L_0^2(V_0).$$

Let's first prove efficiency of $h_n = P_n(D = 2)$ as an estimator of $h_0 = P_{V_0}(D = 2)$. We have that $h_n - h_0 = 1/n \sum_{i=1}^{n} (I(D_i = 2) - h_0)$. Consequently $h_n$ is asymptotically linear with influence curve equal to $I(D = 2) - h_0$. However, $I(D = 2) - h_0$ is a score of a one-dimensional submodel (as defined in the preceding section) and hence it must be the efficient influence curve (see Bickel et al., 1993).

Define

$$\kappa_t(x) \equiv I_{(0,t]}(x)(\tau - x),$$

and notice that

$$\frac{1}{\epsilon}\left(W_{0\epsilon,g}(t) - W_0(t)\right) = \int g(x)\kappa_t(x)dV_0(x) = E_{V_0}\left(g(X)\kappa_t(X)\right). \quad (5.10)$$

For any $V \in \mathcal{F}$ and $t \in (0, \tau]$ we prove that if $g(\tau) > 0$, then there exists a $h_t \in D[0, \infty)$ with finite supnorm which solves $I_V(h_t) = \kappa_t$ (see lemma 5.2 in the next section). For convenience we will denote any such $h_t$ with $I_V^-(\kappa_t)$. This shows that $\vartheta_1(P_V)(t) = W(t)$ is pathwise differentiable along each path $P_{V_{\epsilon,g}}$, $g \in S(V)$, with efficient influence function given by:

$$\widetilde{I}(W,t)(\widetilde{X}, D) = A_V I_V^- \left(\kappa_t - W(t)\right)(\widetilde{X}, D) = A_V(h_t)(\widetilde{X}, D) - W(t). \quad (5.11)$$

Let $\mathcal{G} = \{\widetilde{I}(W_0, t) : t \in [0, \tau]\}$.

Notice that $\mathcal{M}$ is *convex* and that $V \to P_V$ is linear. Theorem 2.2 says now that

$$W_1(t) - W_0(t) = \int \widetilde{I}(W_1, t)dP_{V_0} \text{ for } t \in [0, \tau]. \quad (5.12)$$

for all $V_1$ with $V_0 \ll V_1$ and $dV_0/dV_1 \in L_0^2(V_1)$. In particular, it holds for $V_1 = V_n(\alpha) \equiv \alpha V + (1 - \alpha)V_n$. By letting $\alpha \to 0$ we prove (see lemma 5.6) that this identity also holds at $V_1$ equal to the NPMLE $V_n$. So we have:

$$W_n(t) - W_0(t) = -\int \widetilde{I}(W_n, t)dP_{V_0} \text{ for } t \in [0, \tau]. \quad (5.13)$$

Because $h_{V_n t}$ has a finite supnorm it is a score in $S(V_n)$ of a one dimensional line through $V_n$ and thereby (5.8) provides us with the *efficient score equation*:

$$E_{P_n}(\widetilde{I}(W_n, t)(Y)) = 0 \text{ for all } t \in [0, \tau].$$

Combining the last two identities provides us with the crucial identity:

$$W_n(t) - W_0(t) = \int \tilde{I}(W_n, t) d(P_n - P_{V_0}) \text{ for } t \in [0, \tau]. \tag{5.14}$$

Suppose now that (the *$P_0$-Donsker class condition*) $\tilde{I}(W_n, t)$ lies with probability tending to 1 in a $P_0$-Donsker class $\mathcal{G}$. We prove this for $\mathcal{G}$ equal to the functions of variation less than or equal to $M < \infty$ for some (sufficiently large) $M < \infty$ (corollary 5.1 in the next section) using that $g_n(\tau) > \delta > 0$ with probability tending to 1. Wijers (1991) showed consistency of $g_n(\tau)$ under the assumption that $g(\tau) > 0$, which proves the Donsker class condition. So we have now by the definition of Donsker class:

$$\sup_{t \in [0,\tau]} | (W_n - W_0)(x) | = O_P\left(\frac{1}{\sqrt{n}}\right).$$

This implies that for each $\epsilon > 0$, $\sup_{[0,\tau-\epsilon]} | (V_n - V_0)(x) | \to 0$, and in particular $V_n(x) \to V(x)$ for all $x \in [0, \tau)$, in probability. Suppose that $V_0$ is continuous on $[\tau - \epsilon_1, \tau]$ for some $\epsilon_1 > 0$. The consistency of $g_n(\tau)$ implies by (5.6) consistency of $V_n(\tau)$. If a sequence of monotone functions converges pointwise to a continuous function, then it converges uniformly. Applying this result to $V_n$ on $[\tau - \epsilon, \tau]$ provides us with uniform consistency of $V_n$ on $[0, \tau]$.

The $\rho_{P_0}$-consistency condition is proved by lemma 5.5, using the consistency of $g_n(\tau)$. This completes the supnorm-efficiency proof for $W_n$.

This proves the following theorem, under the assumption that $g_n(\tau)$ is consistent (see section 6).

**Theorem 5.1** *$h_n$ is an efficient estimator of $h_0$. Assume $g_0(\tau) > 0$. Then*

$$\sup_{x \in [0,\tau)} | (W_n - W_0)(x) | = O_P(1/\sqrt{n}).$$

*If $V_0$ is continuous on $[\tau - \epsilon, \tau]$ for an $\epsilon > 0$, then $V_n$ is uniformly consistent on $[0, \tau]$ and $W_n$ is a supnorm-asymptotically efficient estimator of $W_0$:*

$$\sqrt{n}(W_n(t) - W_0(t)) = \int \tilde{I}(W_0, t) d\left(\sqrt{n}(P_n - P_{V_0})\right) + R_{n,t},$$

*where $P(\|R_{n,\cdot}\|_\infty > \epsilon) \to 0$ for all $\epsilon > 0$ and $Z_n(\cdot) \equiv \sqrt{n}(P_n - P_{V_0})(\tilde{I}(W_0, \cdot))$ converges weakly in $(D[0, \tau], \mathcal{B}, \|\cdot\|_\infty)$ to a (pointwise) mean zero measurable Gaussian process $Z$ with covariance structure given by:*

$$E(Z(s)Z(t)) = E_0\left(\tilde{I}(W_0, s)(Y)\tilde{I}(W_0, t)(Y)\right).$$

If $g_0(\tau) = 0$, i.e. $V_0(\tau) = 1$, then one can just shrink down the window $[0, \tau]$ to $[0, \tau']$ where $\tau'$ is chosen so that a fraction of the singly censored or uncensored observations become doubly censored. Then the theorem tells us that the NPMLE based on these reduced observations is efficient for these reduced observations and one can obtain efficiency by letting this fraction converge to zero slowly enough for $n \to \infty$.

### 5.3.1 Construction of confidence intervals.

We will discuss two methods for construction of a pointwise confidence interval for $W(t)$ and estimation of the variance of the estimator $W_n(t)$. The first method is the nonparametric bootstrap. This means that we estimate the distribution of $W_n(t)$ by computing a large number of estimators $W_n(t)^\#$ based on samples $Y_1^\#, \ldots, Y_n^\#$ drawn from the empirical distribution function $P_n$. The asymptotic validity of the nonparametric bootstrap is proved straightforwardly by using identity (5.14)

$$W_n^\#(t) - W_n(t) = (P_n^\# - P_{V_n,h_n})(\widetilde{I}(W_n^\#, t)),$$

and repeating our efficiency proof, in exactly the same way (see section 4.6). Hereby we use that all our results (lemmas) are established uniformly in $V$. Similarly, asymptotic validity of the semiparametric bootstrap holds. In this case we would resample from $P_{V_n,h_n}$, where $V_n, h_n$ is the sieved-NPMLE.

In the second method we estimate the limiting distribution of $W_n(t)$ by estimation of the variance of $\widetilde{I}(W,t)(Y)$. We can estimate $\widetilde{I}(W,t) = A_V I_V^-(\kappa_t)$, which depends on $(V,h)$, by substitution of $(V_n, h_n)$, which provides us with an estimate $\widetilde{I}_n(W,t)$. Now, we can estimate the variance with:

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{I}_n(W,t)(Y_i)^2. \tag{5.15}$$

The computation of $\widetilde{I}_n(W,t)(Y_i)$ involves inverting the information operator. However, as we will see this comes down to inverting a Volterra integral operator which is an infinite dimensional equivalent of a lower-triangular matrix and as shown in subsection 4.3 inversion at a discrete $V_n$ with $k$ support points is inverting a lower-triangular $k$ by $k$ matrix.

The remaining program is as follows: In section 4 we will prove the (hardest) $P_0$-Donsker class condition and thereby in the course of establishing this that $h_{V_n,t}$ is bounded in supnorm. In section 5 the $\rho_P$ consistency and identity conditions will be verified.

## 5.4 Donsker class condition, uniform bound on the variation of the efficient influence function.

In the sequel integrals $\int_0^x$, from zero to any point $x$, are integrals over $[0,x)$ and integrals $\int_y^z$, from $y$ to any $z$, are integrals over $[y,z)$. The score operator $A_{V_0}$ is given by:

$$\begin{aligned} E_V(h(X)|\widetilde{X}=\widetilde{x},\ D=d)) &= h(\widetilde{x})I(d=0)+\frac{\int_{\widetilde{x}}^{\infty}\frac{h(x)}{\tau+x}dV(x)}{\int_{\widetilde{x}}^{\infty}\frac{1}{\tau+x}dV(x)}I(d=1) \\ &\quad +\frac{\int_{\tau}^{\infty}\frac{(x-\tau)h(x)}{\tau+x}dV(x)}{\int_{\tau}^{\infty}\frac{x-\tau}{\tau+x}dV(x)}I(d=2). \end{aligned}$$

The adjoint $A_V^{\top}: L_0^2(P_V)\to L_0^2(V)$ of $A_V$ is given by:

$$\begin{aligned} A_V^{\top}(\eta)(x) &= E(\eta(\widetilde{X},D)|X=x) \\ &= \frac{\tau-x}{\tau+x}\eta(x,0)I(x\le\tau)+\frac{2}{\tau+x}\int_0^{x\wedge\tau}\eta(\widetilde{x},1)d\widetilde{x} \\ &\quad +\frac{x-\tau}{x+\tau}\eta(\tau,2)I(x\ge\tau). \end{aligned}$$

Consequently, the information operator $I_V=A_V^{\top}A_V: L_0^2(V)\to L_0^2(V)$ is given by:

$$\begin{aligned} I_V(h)(x) &= \frac{\tau-x}{\tau+x}h(x)I(x<\tau)+\frac{2}{\tau+x}\int_0^{x\wedge\tau}\frac{\int_{\widetilde{x}}^{\infty}\frac{h(u)}{\tau+u}dV(u)}{\int_{\widetilde{x}}^{\infty}\frac{1}{\tau+u}dV(u)}d\widetilde{x} \\ &\quad +\frac{x-\tau}{x+\tau}\frac{\int_{\tau}^{\infty}\frac{(u-\tau)h(u)}{\tau+u}dV(u)}{\int_{\tau}^{\infty}\frac{u-\tau}{\tau+u}dV(u)}I(x\ge\tau). \end{aligned} \tag{5.16}$$

Consider the equation (in $h_t$)

$$I_V(h_t)(x)=\kappa_t(x)=I_{(0,t]}(x)(\tau-x),$$

which should be considered as solving for the *hardest submodel* in $\mathcal{F}$ for estimating $W(t)$ using data from $P_V$. Notice that $h_t$ depends on $V$. Instead of $h_{tV_n}, h_{tV}$ we use the notation $h_{tn}, h_t$, respectively. In the next subsection we will show that this hardest underlying score $h_t - W(t) = I_V^-(\kappa_t) - W(t)$ exists and that it is given by a Neumann series.

By analyzing this Neumann series we will show that $h_{nt} \equiv I_{V_n}^-(\kappa_{nt})$ is of uniformly (in $V_n$ and $t$) bounded variation and thereby, using (5.11), that $\widetilde{I}(W_n,t)I(D=d)$ is of bounded variation for $d=0,1,2$ uniformly in $n$.

### 5.4.1 Solving for the hardest submodel.

We split the inversion $I_V(h_t)(x) = (\tau - x)I_{(0,t]}(x)$ in two parts, namely we consider this equation for $x \geq \tau$ and for $x \leq \tau$.

**Inversion for $x \geq \tau$.** For $x \geq \tau$ we have $\kappa_t(x) = (\tau - x)I_{(0,t]}(x) = 0$. Recall the definition $g(y) = \int_y^\infty 1/(\tau + u)dV(u)$. We define also:

$$C(h)(\tau) \equiv \frac{\int_\tau^\infty \frac{(u-\tau)h(u)}{\tau+u}dV(u)}{\int_\tau^\infty \frac{u-\tau}{\tau+u}dV(u)}.$$

For $x \geq \tau$ $h_t$ (see (5.16)) is determined by the equation:

$$2\int_0^\tau \frac{\int_y^\infty \frac{h(u)}{\tau+u}dV(u)}{g(y)}dy + (x-\tau)C(h)(\tau) = 0. \tag{5.17}$$

Consequently, we have that $h_t$ solves:

$$\int_0^\tau \frac{\int_y^\infty \frac{h_t(u)}{\tau+u}dV(u)}{g(y)}dy = 0 \tag{5.18}$$

$$C(h_t)(\tau) = 0. \tag{5.19}$$

**Inversion for $x \in [0, \tau)$ using (5.18).** We use that $h_t$ solves (5.18). For $x \leq \tau$ the third term in $I_V(h_t)(x)$ equals zero. Consequently, for $x \leq \tau$ we have (at the second equality we use (5.18) and at the third equality we split up the integration area $[y, \infty) = [y, \tau) \cup [\tau, \infty)$):

$$\begin{aligned} I_V(h_t)(x) &= \frac{\tau - x}{\tau + x}h_t(x) + \frac{2}{\tau + x}\int_0^x \frac{\int_y^\infty \frac{h_t(u)}{\tau+u}dV(u)}{g(y)}dy \\ &= \frac{\tau - x}{\tau + x}h_t(x) - \frac{2}{\tau + x}\int_x^\tau \frac{\int_y^\infty \frac{h_t(u)}{\tau+u}dV(u)}{g(y)}dy \\ &= \frac{\tau - x}{\tau + x}h_t(x) - \frac{2}{\tau + x}\int_x^\tau \frac{\int_y^\tau \frac{h_t(u)}{\tau+u}dV(u)}{g(y)}dy \\ &\quad - \frac{2}{\tau + x}\int_x^\tau \frac{1}{g(y)}dy \times \int_\tau^\infty \frac{h_t(u)}{\tau + u}dV(u). \end{aligned} \tag{5.20}$$

Now, we will use (5.18) in order to express $\int_\tau^\infty \frac{h_t(u)}{\tau+u}dV(u)$ in an integral which only uses the values $h_t(x)$, $x \in [0, \tau]$. By using Fubini at the second equality below, we have:

$$\begin{aligned} 0 &= \int_0^\tau \frac{1}{g(y)}\int_y^\infty \frac{h_t(u)}{\tau + u}V(du)dy \\ &= \int_0^\infty \left(\int_0^{u\wedge\tau} \frac{1}{g(y)}dy\right)\frac{h_t(u)}{\tau + u}V(du) \\ &= \int_0^\tau \left(\int_0^u \frac{1}{g(y)}dy\right)\frac{h_t(u)}{\tau + u}V(du) + \int_0^\tau \frac{1}{g(y)}dy \times \int_\tau^\infty \frac{h_t(u)}{\tau + u}V(du), \end{aligned}$$

and consequently

$$\int_\tau^\infty \frac{h_t(u)}{\tau+u}V(du) = -\frac{\int_0^\tau \left(\int_0^u \frac{1}{g(y)}dy\right)\frac{h_t(u)}{\tau+u}V(du)}{\int_0^\tau \frac{1}{g(y)}dy}. \tag{5.21}$$

Substitute this in (5.20) and we multiply both sides with $(\tau+x)/(\tau-x)$. Before we write down the obtained expression we make some definitions. Define the operator $B : (D[0,\tau], \|\cdot\|_\infty) \to (D[0,\tau], \|\cdot\|_\infty)$ by

$$B(h)(x) = \frac{2}{\tau-x}\int_x^\tau \frac{\int_y^\tau \frac{h(u)}{\tau+u}dV(u)}{g(y)}dy \text{ for } x \in [0,\tau) \text{ and } B(h)(\tau) = 0.$$

Notice that $\lim_{x\uparrow\tau} B(h)(x) = 0 = B(h)(\tau)$ and therefore this operator indeed maps cadlag functions on $[0,\tau]$ to cadlag functions on $[0,\tau]$. Moreover, define:

$$\begin{aligned}
\alpha_{1g}(x) &\equiv \frac{2}{\tau-x}\frac{\int_x^\tau \frac{1}{g(y)}dy}{\int_0^\tau \frac{1}{g(y)}dy} \\
\alpha_2(h)(\tau) &\equiv \int_0^\tau \left(\int_0^u \frac{1}{g(y)}dy\right)\frac{h(u)}{\tau+u}V(du).
\end{aligned}$$

Substituting (5.21) into (5.20) and multiplying both sides of $I_V(h_t) = \kappa_t$ with $(\tau+x)/(\tau-x)$ tells us that $h_t$ solves the following equation for $x \in [0,\tau)$:

$$(I-B)(h_t)(x) = -\alpha_{1g}(x)\alpha_2(h_t)(\tau) + I_{(0,t]}(x)\cdot(\tau+x) \tag{5.22}$$

and for $x \geq \tau$ $h_t$ has to satisfy (5.18) and (5.19).

***B* is a Volterra operator.** Notice that (by using Fubini):

$$\begin{aligned}
B(h)(x) &= \int_x^\tau \left(\frac{2}{\tau-x}\int_{y=x}^u \frac{1}{g(y)}dy\right)\frac{h(u)}{\tau+u}dV(u) \\
&= \int_{u=x}^\tau K_g(x,u)h(u)dV_1(u),
\end{aligned}$$

where

$$\begin{aligned}
K_g(x,u) &\equiv \frac{2}{\tau-x}\int_{y=x}^u \frac{1}{g(y)}dy, \; u \in (x,\tau] \\
dV_1(u) &\equiv \frac{1}{\tau+u}dV(u).
\end{aligned}$$

If $g(\tau) > 0$, then

$$\sup_{x\in[0,\tau),u\in(x,\tau]} K_g(x,u) \leq \frac{2}{g(\tau)} < \infty.$$

By the Volterra structure of $B$ (Kantorovich and Akilov, 1982, Gill and Johansen, 1990, see also Gill, van der Laan, Wellner, 1993), we have:

$$\|B^k(h)\|_\infty \leq \frac{1}{k!}\left(\frac{2}{g(\tau)}V_1(0,\tau]\right)^k \|h\|_\infty.$$

Therefore for any $h \in D[0,\tau]$ we have

$$\|\sum_{k=0}^{\infty} B^k(h)\|_\infty \leq \exp\left(V_1[0,\tau)\frac{2}{g(\tau)}\right)\|h\|_\infty.$$

This proves the following lemma:

**Lemma 5.1** *Assume that $g(\tau) > 0$. We have the following results:*

1. *$I - B : (D[0,\tau], \|\cdot\|_\infty) \to (D[0,\tau], \|\cdot\|_\infty)$ has a bounded inverse given by $\sum_{k=0}^{\infty} B^k(h)$.*
2. *If $h \geq 0$, then $B(h) \geq 0$, and consequently $B^k(h) \geq 0$, $k = 1, 2, \ldots$.*
3. *For any $h \in D[0,\tau]$*

$$\|B^k(h)\|_\infty \leq \frac{1}{k!}\left(\frac{2}{g(\tau)}V_1(0,\tau]\right)^k \|h\|_\infty.$$

Notice that $\alpha_{1g} \in D[0,\tau]$. Now, we can apply the Neumann series (i.e. the inverse of $I - B$) to the left and right-hand side of (5.22). This provides us with the following equation for $h_t \in D[0,\tau]$:

$$\begin{aligned} h_t &= \sum_{k=0}^{\infty} B^k\left(I_{(0,t]}(\cdot)(\tau + \cdot)\right) - \alpha_2(h_t)(\tau)\sum_{k=0}^{\infty} B^k(\alpha_{1g}) \\ &\stackrel{\text{def}}{\equiv} f_{1t} - \alpha_2(h_t)(\tau) f_2. \end{aligned} \tag{5.23}$$

It remains to find $\alpha_2(h_t)(\tau)$. $\alpha_2$ is a linear mapping in $h$ and therefore it makes sense to apply $\alpha_2$ to the left and right-hand side of (5.23). This provides us with:

$$\alpha_2(h_t)(\tau) = \alpha_2(f_{1t})(\tau) - \alpha_2(h_t)(\tau)\alpha_2(f_2)(\tau).$$

By lemma 5.1 $\alpha_{1g} \geq 0$ implies that $f_2 = \sum_{k=0}^{\infty} B^k(\alpha_{1g}) \geq 0$. Consequently $\alpha_2(f_2)(\tau) \geq 0$ and therefore

$$\alpha_2(h_t)(\tau) = \frac{\alpha_2(f_{1t})(\tau)}{1 + \alpha_2(f_2)(\tau)} \equiv \alpha_{3t}.$$

This proves that $h_t$ is uniquely determined on $[0,\tau)$. So $h_t(x)$ is uniquely determined on $[0,\tau)$ and for $x \geq \tau$ it only has to satisfy (5.18) and (5.19) (several solutions are possible). From now on $h_t \in D[0,\tau]$ (from which its restriction to $[0,\tau)$ is the solution we were looking for) will be this unique part which lives on $[0,\tau]$ and $h_t^\tau$ will be that part on $[\tau,\infty)$ for which $h_t + h_t^\tau$ solves (5.18) and (5.19). By substituting this $h_t$ in (5.18) we find that $h_t^\tau(x)$, $x \geq \tau$ has to satisfy

$$\int_\tau^\infty \frac{h_t^\tau(u)}{\tau+u} dV(u) = -\frac{\int_0^\tau \frac{\int_y^\tau \frac{h_t(u)}{\tau+u} dV(u)}{g(y)} dy}{\int_0^\tau \frac{1}{g(y)} dy} \equiv C(h_t, V).$$

Moreover it has to satisfy $C(h_t)(\tau) = 0$. This proves the following lemma.

**Lemma 5.2** *Assume $g(\tau) > 0$. Define (they are all well defined)*

$$\begin{aligned} f_{1t} &\equiv \sum_{k=0}^\infty B^k \left( I_{(0,t]}(\cdot)(\tau + \cdot) \right) \in D[0,\tau] \\ f_2 &\equiv \sum_{k=0}^\infty B^k(\alpha_{1g}) \in D[0,\tau] \\ \alpha_{3t} &\equiv \frac{\alpha_2(f_{1t})(\tau)}{1 + \alpha_2(f_2)(\tau)} \in \mathbb{R}. \end{aligned}$$

*We have: $I_V(h_t)(x) = \kappa_t(x)$ for all $x \in [0,\infty]$ if and only if $h_t(x) = f_{1t}(x) - \alpha_{3t} f_2(x)$ for $x \in [0,\tau)$ and if $h_t^\tau \equiv h_t(x) I(x \geq \tau)$ satisfies*

$$\int_\tau^\infty \frac{h_t^\tau(u)}{\tau+u} dV(u) = C(h_t, V) \ \wedge \ C(h_t^\tau)(\tau) = 0. \tag{5.24}$$

*Moreover, $\|h_t\|_\infty \leq C/g(\tau)$ for a $C < \infty$ which does not depend on $V$ and $t$.*

The last statement is a trivial consequence of lemma 5.1. If one wants to have the solution of $I_V(h) = \kappa$ for a general $\kappa$, then one replaces in this lemma $I_{(0,t]}(x)$ by $\kappa(x)(\tau+x)/(\tau-x)$, where it is required that $\kappa(x)(\tau+x)/(\tau-x) \in D[0,\tau]$. Lemma 5.2 tells us that $I_V(h) = \kappa_t$ is solved by $h(x) = h_t(x) I_{[0,\tau)}(x) + h_t^\tau(x) I_{[\tau,\infty)}$. Consequently, we have $\widetilde{I}(W,t) = A_V(h_t) + A_V(h_t^\tau)$. In the next section we will show the Donsker class condition for $A_V(h_t)$. Before we do this we will show that the Donsker class condition certainly holds for $A_V(h_t^\tau)$ and that in fact the $L^2$-consistency and Donsker class analysis of $A_V(h_t)$ provides us certainly with these results for $A_V(h_t^\tau)$.

### 5.4.2 The information after $\tau$.

Recall the score operator $A_V$. We have that $A_V(h_t^\tau)I(D=2) = C(h_t^\tau)(\tau)$ and consequently this equals zero because $h_t^\tau$ satisfies (5.24). $A_V(h_t^\tau)(\widetilde{x},0) = h_t^\tau(\widetilde{x})$ and because $P_V(\cdot,0)$ puts only mass on $[0,\tau)$ this equals 0 $P_V$ a.e. Finally, we have by using that $h_t^\tau(x) = 0$ for $x < \tau$ and by (5.24) that:

$$\begin{aligned} A_V(h_t^\tau)(\widetilde{x},1) &= \frac{\int_x^\infty \frac{h_t^\tau(u)}{\tau+u} dV(u)}{g(\widetilde{x})} \\ &= \frac{\int_\tau^\infty \frac{h_t^\tau(u)}{\tau+u} dV(u)}{g(\widetilde{x})} \\ &= \frac{C(h_t,V)}{g(\widetilde{x})}. \end{aligned}$$

Let $\|\cdot\|_v$ denote the variation norm. Notice that $g$ is a monotone function. Therefore for showing that $\|A_{V_n}(h_{nt}^\tau)I(D=1)\|_v < M$ for certain $M < \infty$ it suffices to have that $g_n(\tau) > \delta > 0$ for certain $\delta > 0$ and that $\|h_{nt}\|_\infty < M_1$ for certain $M_1 < \infty$. By the last statement of lemma 5.2 the latter follows from $g_n(\tau) > \delta > 0$ for some $\delta > 0$. Using that $g_n(\tau)$ is consistent, this follows from $g(\tau) > 0$. This proves the $P$-Donsker class condition for $A_{V_n}(h_t^\tau)$.

Because convergence of $C(h_{nt}, V_n)$ to $C(h_t, V)$ follows from convergence of $V_n$ to $V$ on $[0,\tau]$ and $h_{nt}$ to $h_t$ it is now also clear that the proved convergence of $h_n^t$ in the $P$-consistency-analysis in section 6 provides us also with the required convergence for $h_{nt}^\tau$ and hence for the $P$-consistency condition for $A_{V_n}(h_{nt}^\tau)$.

### 5.4.3 Estimation of the efficient influence curve in practice.

In this section we explain how $A_{V_n}(h_{nt} + h_{nt}^\tau)$ is computed, as needed for computation of the estimated variance (5.15). Firstly, as shown above we have that $A_{V_n}(h_{nt}^\tau) = I(D=1)C(h_{nt},V)/g_n(\widetilde{x})$. Hence once we have a quick way of computing $h_{nt}$ from $V_n, g_n(\tau)$ we are done. Lemma 5.2 tells us that this is established once we can quickly invert $(I - B_{V_n})(h_n^t)(x) = f(x)$ for some given function $f(x)$. The lemma suggest to invert this equation by applying the Neumann series of $B_{V_n}$ to $f$, which is indeed a quick method since $k$ iterations is already enough to have reduced the error to the order $1/k!$. However, here we show what inversion of the Volterra operator $I - B$ really stands for in the case that $V_n$ is discrete. Suppose $V_n$ puts mass on $k$ points $x_1, \ldots, x_k$, namely the uncensored $\widetilde{X}_i$. Then the only values of $h_t^n$ which count are $h_t^n(x_i)$, $i = 1, \ldots, k$. Hence we can represent $h_t^n$ as $\vec{h}_{nt} \in \mathbb{R}^k$. Similarly, the relevant

values for the kernel $K_{g_n}$ are $K_{g_n}(x_i, x_j)$, $i = 1, \ldots, k$, $j = 1, \ldots, k$, and also $B_{V_n}(h)(x)$ only jumps at $x_1, \ldots, x_k$. Consequently, we can represent $B_{V_n}$ as a matrix operator $(B^n_{ij}) : \mathbb{R}^k \to \mathbb{R}^k$, where

$$B^n_{ij} = 0 \text{ if } i > j \text{ and } B_{ij} = K_{g_n}(x_i, x_j)v_n(x_j) \text{ if } j \geq i.$$

This is due to the fact that the Volterra integral at $x = x_i$ only integrates (sums) over the values $h^t_n(x_j)$ with $j \geq i$. So inverting $(I - B_{V_n})(h^t_n)(x_i) = f(x_i)$, $i = 1, \ldots, k$, comes now down to solving $(\delta_{ij} - B^n_{ij})(\vec{h}_{nt}) = \vec{f}$, where $\delta_{ij} = 1$ if $i = j$ and zero elsewhere. So $\delta_{ij} - B^n_{ij}$ is an upper triangular matrix so that $\vec{h}_{nt}$ is found by first obtaining $h^t_n(x_k)$ from the last equation, then obtain $h^t_n(x_{k-1})$ from the $k-1$'th equation using $h^t_n(x_k)$ and go on like this till you arrive at the first equation. In this way we find in $k$ simple steps $\vec{h}_{nt}$. This shows that inverting the information operator at a discrete $V_n$ is not computer intensive at all and can be easily implemented. This provides us with a quick way of estimating the limit variance of $\sqrt{n}(W_n(t) - W(t))$ and hence for constructing pointwise confidence intervals.

### 5.4.4 Bounded variation of the hardest underlying scores.

In this section we will show that $\sup_{t\in[0,\tau]} \|h_t\|_v = O(1)$ and $\limsup_{n\to\infty} \sup_{t\in[0,\tau]} \|h_{tn}\|_v = O(1)$ a.s. Lemma 5.2 tells us that $h_{tn} = f_{1tn} + \alpha_{3tn} f_{2n}$. We have that $\alpha_{3tn} \leq \alpha_{2n}(f_{1tn}) \leq \|f_{1tn}\|_\infty \alpha_{2n}(1)$. By lemma 5.1 we have that $\|f_{1tn}\|_\infty \leq \exp(2V_n[0,\tau]/g_n(\tau))\,\tau$. Therefore for showing that $\alpha_{3tn} \leq M < \infty$ for some $M < \infty$ for $n$ large enough we need that $g_n(\tau) > \delta > 0$ (some $\delta > 0$) for $n$ large enough, which follows from consistency of $g_n(\tau)$ to $g(\tau) > 0$. This shows that $\limsup_{n\to\infty} \alpha_{3n}$ is bounded a.s.

Now, lemma 5.2 tells us that for showing $\limsup_{n\to\infty} \sup_t \|h_{tn}\|_v = O(1)$ a.s. it suffices to show that (a.s.) for some $M < \infty$

$$\limsup_{n\to\infty} \sum_{k=0}^{\infty} \sup_t \|B^k_n\left(I_{(0,t]}(\cdot)(\tau + \cdot)\right)\|_v \leq M < \infty \tag{5.25}$$

$$\limsup_{n\to\infty} \sum_{k=0}^{\infty} \|B^k_n(\alpha_{1g_n})\|_v \leq M < \infty. \tag{5.26}$$

We will prove (5.26) (the hardest of the two). For proving (5.25) one just substitutes $I_{(0,t]}(\cdot)/(\tau + \cdot)$ for $\alpha_{1g_n}$.

Consider the first term $\alpha_{1g_n}$ of (5.26). Define

$$h_{1n} \equiv \frac{1}{\tau - x} \int_x^\tau \frac{1}{g_n(y)} dy,$$

and notice that $\alpha_{1g_n} = c_n h_{1n}$, where $c_n = 2/\int_0^\tau (1/(g_n(y))dy \leq 2g_n(0)/\tau$ and therefore $\limsup_{n\to\infty} c_n < \infty$. Now $\sum_{k=0}^\infty \|B_n^k(\alpha_{1g_n})\|_v \leq c_n \sum_{k=0}^\infty \|B_n^k(h_{1n})\|_v$. So it suffices to study the series $\sum_{k=0}^\infty \|B_n^k(h_{1n})\|_v$.

We have

$$\frac{1}{\tau - x}\frac{1}{g_n(x)} = \frac{1}{(\tau - x)^2}\int_x^\tau \frac{1}{g_n(x)}dy.$$

Using this it follows that

$$\begin{aligned} \frac{d}{dx}h_{1n}(x) &= \frac{1}{(\tau - x)^2}\int_x^\tau \left(\frac{1}{g_n(y)} - \frac{1}{g_n(x)}\right) dy \\ &\geq 0, \end{aligned} \tag{5.27}$$

by the fact that $g_n$ is monotone. Consequently, $h_{1n}$ is monotone increasing and larger than or equal to zero and therefore $\|h_{1n}\|_v = \|h_{1n}\|_\infty \leq \frac{1}{g_n(\tau)}$, which is uniformly bounded by the preceding argument. This proves:

**Lemma 5.3** *We have that* $\|h_{1n}\|_v = \|h_{1n}\|_\infty \leq 1/g_n(\tau)$. *If* $g(\tau) > 0$, *then* $g_n(\tau) > \delta > 0$ *for certain* $\delta > 0$ *with probability tending to 1.*

Consider now the term $m(\cdot) \equiv B_n^{k+1}(h_{1n})(\cdot)$. Define for a function $f$

$$a_n^k(f)(y) \equiv \int_y^\tau \frac{B_n^k(f)(u)}{\tau + u}dV_n(u).$$

We have:

$$\begin{aligned} \frac{d}{dx}B_n^{k+1}(h_{1n})(x) &= \frac{d}{dx}\left(\frac{1}{\tau - x}\int_x^\tau \frac{a_n^k(h_{1n})(y)}{g_n(y)}dy\right) \\ &= \frac{1}{(\tau - x)^2}\int_x^\tau \frac{a_n^k(h_{1n})(y)}{g_n(y)}dy - \frac{1}{\tau - x}\frac{a_n^k(h_{1n})(x)}{g_n(x)} \\ &= \frac{1}{(\tau - x)^2}\int_x^\tau \left(\frac{a_n^k(h_{1n})(y)}{g_n(y)} - \frac{a_n^k(h_{1n})(x)}{g_n(x)}\right) dy. \end{aligned}$$

Consider a point $x$ where this derivative is larger than or equal to zero. By lemma (5.1) $\|a_n^k(f)\|_\infty \leq \frac{C_n^k}{k!}\|f\|_\infty$ and if $f \geq 0$, then $a_n^k(f)$ is decreasing, where $C_n$ is bounded by $1/(\tau g_n(\tau))$. Therefore, if $x \leq y$, then $a_n^k(h_{1n})(y) \geq a_n^k(h_{1n})(x)$. Hence we have:

$$\begin{aligned} 0 &\leq \frac{1}{(\tau - x)^2}\int_x^\tau \left(\frac{a_n^k(h_{1n})(y)}{g_n(y)} - \frac{a_n^k(h_{1n})(x)}{g_n(x)}\right) dy \\ &\leq a_n^k(h_{1n})(x)\frac{1}{(\tau - x)^2}\int_x^\tau \left(\frac{1}{g_n(y)} - \frac{1}{g_n(x)}\right) dy \\ &\leq a_n^k(h_{1n})(x)\frac{d}{dx}h_{1n}(x) \text{ by (5.27)} \\ &\leq \frac{C_n^k}{k!}\|h_{1n}\|_\infty \frac{d}{dx}h_{1n}(x). \end{aligned} \tag{5.28}$$

We can write $m = m_1 + m_2$ where $m_1$ is increasing and $m_2$ is decreasing. (5.28) and lemma 5.3 show that

$$\|m_1\|_v \leq \frac{C_n^k}{k!}\|h_{1n}\|_v\|h_{1n}\|_\infty = \frac{C_n^k}{k!}\|h_{1n}\|_\infty^2.$$

We also know by lemma 5.1 that $\|m\|_\infty \leq \frac{C_n^k}{k!}\|h_{1n}\|_\infty$. This tells us that

$$\begin{aligned}\|m_2\|_{\mathrm{v}} &\leq 2\|m_2\|_\infty \leq 2\left(\|m\|_\infty + \|m_1\|_\infty\right) \\ &\leq 2\left(\frac{C_n^k}{k!}\|h_{1n}\|_\infty + \frac{C_n^k}{k!}\|h_{1n}\|_\infty^2\right).\end{aligned}$$

Consequently,

$$\|m\|_{\mathrm{v}} \leq \|m_1\|_{\mathrm{v}} + \|m_2\|_{\mathrm{v}} \leq 2\frac{C_n^k}{k!}\|h_{1n}\|_\infty + 3\frac{C_n^k}{k!}\|h_{1n}\|_\infty^2 \leq 5\frac{C_n^k}{k!}\|h_{1n}\|_\infty^2,$$

assuming that $\|h_{1n}\|_\infty > 1$, if not the bound is even better. This proves that

$$\begin{aligned}\sum_{k=0}^{\infty}\|B_n^k(\alpha_{1g_n})\|_{\mathrm{v}} &\leq c_n\sum_{k=0}^{\infty}\|B_n^k(h_{1n})\|_{\mathrm{v}} \\ &\leq 5c_n\|h_{1n}\|_\infty^2\exp(C_n) \\ &\leq M\frac{\exp\left(1/g_n(\tau)\right)}{g_n(\tau)^2},\end{aligned}$$

where we used lemma 5.3 and $C_n \leq 1/(\tau g_n(\tau))$ at the last inequality. This proves (5.26). This provides us with the following result.

**Lemma 5.4** *If $g(\tau) > 0$, then*

$$\limsup_{n\to\infty}\sup_t\|h_{tn}\|_{\mathrm{v}} < M \text{ a.s. and } \sup_t\|h_t\|_{\mathrm{v}} \leq M \text{ for certain } M < \infty.$$

### 5.4.5 Donsker class condition.

Recall $\widetilde{I}(W_n,t)(\cdot) = A_{V_n}(h_{tn} + h_{tn}^\tau)(\cdot) - W_n(t)$. Lemma 5.4 tells us that $h_{tn}$ is of bounded variation uniformly in $V_n$ and $t$. We also have that $A_{V_n}I(D=d)$ : $(D[0,\tau], \|\cdot\|_{\mathrm{v}}) \to (D[0,\tau], \|\cdot\|_{\mathrm{v}})$ is a bounded linear mapping for $d \in \{0,1,2\}$. This follows from the fact that $A_{V_n}I(D=d)$ maps monotone functions to monotone functions, which is in fact a quite general fact which holds in missing data models. Recalling the subsection about $A_{V_n}(h_{tn}^\tau)$, it follows that this provides us with the following corollary of lemma 5.4:

**Corollary 5.1** (Uniform bound on variation of efficient influence curve). *If $g(\tau) > 0$, then*

$$\limsup_{n\to\infty}\sup_t\|\widetilde{I}(W_n,t)I(D=d)\|_{\mathrm{v}} < M \text{ a.s.}$$

*and*

$$\sup_t \|\widetilde{I}(W,t)I(D=d)\|_v \leq M$$

*for* $d = 0, 1, 2$ *and some* $M < \infty$.

This proves the $P_0$-Donsker class condition.

## 5.5 Consistency condition for the efficient influence function.

As shown in the general proof, for verifying this condition we can use that $\sup_{x\in[0,\tau]} \mid (V_n - V_0)(x) \mid \to 0$ and $g_n(\tau) \to g(\tau)$, both in probability.

By lemma 5.2 the hardest score is given by $I_{V_n}^{-1}(\kappa_{nt} - W_n(t)) = \sum_{k=0}^{\infty} B_{V_n}^k(f_{nt}) + h_{nt}^\tau - W_n(t)$, where $f_{nt}(\cdot) \equiv I_{(0,t]}(\cdot)(\tau + \cdot) - \alpha_{3n}\alpha_{1g_n}(\cdot)$. The efficient influence function at $V_n$ is given by $A_{V_n}(I_{V_n}^{-1}(\kappa_{nt} - W_n(t)))$. By lemma 5.2 and the subsection 5.4.2 it follows that the convergence of $A_{V_n}(h_{nt}^\tau)$ follows easily from the convergence of $A_{V_n}(h_{nt})$ shown below.

Denote $J_n = I - B_{V_n}$ and $J_n^{-1} \equiv \sum_{k=0}^{\infty} B_{V_n}^k$. Because $W_n(t) \to W(t)$ we only have to consider:

$$\begin{aligned} A_{V_n}J_n^{-1}(f_{nt}) - A_V J^{-1}(f_t) &= (A_{V_n} - A_V)J^{-1}(f_t) + A_{V_n}J_n^{-1}(f_{nt} - f_t) \\ &\quad + A_{V_n}J_n^{-1}(J_n - J)J^{-1}(f_t). \end{aligned} \tag{5.29}$$

We have to prove that the $L^2(P_V)$-norm of these terms converge to zero uniformly in $t \in [0, \tau]$ in probability. Firstly, we will study the third term. $f_{nt} - f_t$ involves the difference

$$\alpha_{1g_n}(x) - \alpha_{1g}(x) = c\frac{1}{\tau - x}\int_x^\tau \left(\frac{1}{g_n(y)} - \frac{1}{g(y)}\right) dy. \tag{5.30}$$

We have

$$g(y) = \int_y^\tau \frac{1}{\tau + x} dV(x) + g(\tau).$$

Consequently, the uniform consistency of $V_n$ on $[0, \tau]$ and the consistency of $g_n(\tau)$ provides us with uniform consistency of $g_n(\cdot)$ on $[0, \tau]$. The fact that $g_n(\tau) > \delta > 0$ with probability tending to 1 tells us now that the integrand converges uniformly to zero in probability. This proves the convergence of (5.30) to zero in probability. By using the same arguments as we do below for the first and second term we can show that $\alpha_{3n} - \alpha_3 \to_P 0$. This proves that

$\|f_{nt} - f_t\|_\infty \to_P 0$. By lemma 5.1 $J_n^{-1}$ is a bounded (uniformly in $n$) operator on $(D[0,\tau], \|\cdot\|_\infty)$. This proves the convergence to zero of the third term.

Let's now consider the first term of (5.29). Define $h_t \equiv J^{-1}(f_t)$. Recall the score operator $A_V$, where we only have to consider the first two terms because $h_t$ lives on $[0,\tau)$. By telescoping, the first term of $(A_{V_n} - A_V)(h)$ can be written as a sum of two similar terms and one is given by:

$$\frac{\int_y^\tau \frac{h_t(x)}{\tau+x}(V_n - V)(dx)}{g_n(y)}.$$

Applying integration by parts gives:

$$\frac{1}{g_n(y)}\left(-\int_y^\tau (V_n - V)(x)d(\frac{h_t}{\tau+\cdot})(x) - (V_n - V)(y)\frac{h_t(y)}{\tau+y}\right). \tag{5.31}$$

Using $g_n > \delta > 0$ with probability tending to 1, we can bound this by the supnorm of $V_n - V$ and the variation norm of $h_t$. Lemma 5.4 tells us that $h_t$ is of bounded variation uniformly in $t$. This proves the uniform convergence of the first term.

Let's now consider the second term of (5.29). In $(J_n - J)(h_t) = (B_n - B)(h_t)$ we have to deal with the following kind of terms:

$$\frac{1}{\tau - x}\int_x^\tau \int_y^\tau h_t(u)d(V_n - V)(u)dy. \tag{5.32}$$

We can bound, by integration by parts, $\int_y^\tau h_t(u)d(V_n - V)(u)$ by the supnorm of $V_n - V$ and the variation norm of $h_t$, where the latter is bounded uniformly in $t$. This proves the uniform convergence of $(J_n - J)(h_t)$ and the boundedness of the operator $J^{-1}$ finishes the proof of the convergence of the second term of (5.29). This proves:

**Lemma 5.5** ($\rho$-consistency condition). *If $V_n$ is uniformly consistent for $V$ on $[0,\tau]$ and $g_n(\tau)$ is consistent for $g(\tau) > 0$, then*

$$\sup_{t\in[0,\tau]} \|\widetilde{I}(W_n,t) - \widetilde{I}(W,t)\|_{P_V} \to 0 \text{ in probability.}$$

## 5.5.1 The identity condition

For $\alpha \in [0,1]$ we define the line $V_n(\alpha) = \alpha V + (1-\alpha)V_n$ of distributions. Then $V \ll V_n(\alpha)$. This shows that the identity (5.12) holds for $P_{V_n(\alpha),g_n(\alpha(\tau)}$:

$$W_n(\alpha)(t) - W(t) = -P_{V,h}\widetilde{I}(W_n(\alpha),t). \tag{5.33}$$

We have $V_n(\alpha) - V_n = \alpha(V_n - V)$. Hence for $\alpha \to 0$ $V_n(\alpha)$ converges uniformly to $V_n$ on $[0,\tau]$ and the corresponding $g_n(\alpha)(\tau)$ converges to $g_n(\tau)$. Hence if $\alpha \to 0$,

then the left-hand side of (5.33) converges to $W_n(t) - W(t)$. Since $V_n(\alpha) - V_n$ converges to zero uniformly on $[0, \tau]$, by imitating the steps in the proof for the $\rho$-consistency condition (using the bounded invertibility of $I - B_V$ w.r.t. the variation and supnorm, uniformly in $V$) it follows that $\widetilde{I}(W_n(\alpha), t)(\cdot, d)$ converges uniformly to $\widetilde{I}(W_n, t)(\cdot, d)$, which shown that in particular

$$P_{V,h}\widetilde{I}(W_n(\alpha), t) \to P_{V,h}\widetilde{I}(W_n, t).$$

This proves that we have the wished identity:

$$W_n(t) - W(t) = -P_{V,h}\widetilde{I}(W_n, t). \tag{5.34}$$

This proves the lemma as needed in the general proof in section 3.

**Lemma 5.6** *Let* $V_n(\alpha) = \alpha V + (1 - \alpha)V_n$. *Then for* $d \in \{0, 1, 2\}$

$$\|\widetilde{I}(W_n(\alpha), t)(\cdot, d) - \widetilde{I}(W_n, t)(\cdot, d)\|_\infty \to 0. \tag{5.35}$$

This completes the proof of all conditions we needed in the proof of theorem 5.1: the uniform bound on the variation of $\widetilde{I}(W_n, t)$, the efficient score-equation, the identity for the Sieved-NPMLE and the $\rho_P$-consistency of $\widetilde{I}(W_n, t)$.

## 5.6 Discussion.

Our results used that $g_n(\tau)$, or equivalently $V_n(\tau-)$, is consistent, which has not been proved. The reason why the original proof of Wijers (1991) broke down at $\tau$ was the fact that one could not establish $g_n(\tau) > \delta > 0$ for some $\delta > 0$; a condition we also needed in our analysis. Wijers (1993, page 136) proposed a slight extra censoring of the data to $[0, \tau - \epsilon]$ which guarantees that $g_n(\tau - \epsilon) > \delta > 0$. Hence our results apply to the Sieved-NPMLE based on the slightly transformed data; $W_n$ is an almost efficient estimator on $[0, \tau - epsilon]$ for every $\epsilon > 0$.

For $T_i \leq 0$ we observe pairs $(Z_i, D_i)$

$$Z_i = \min(T_i + X_i, \tau - \epsilon),\ D_i = \begin{cases} 1 & T_i + X_i \leq \tau - \epsilon \\ 2 & T_i + X_i > \tau - \epsilon \end{cases}$$

and for the $T_i \in (0, \tau)$ we observe

$$Z_i = \min(X_i, \tau - T_i, \tau - \epsilon),\ D_i = \begin{cases} 0 & T_i + X_i \leq \tau,\ X_i \leq \tau - \epsilon \\ 1 & T_i + X_i > \tau,\ T_i > \epsilon \\ 3 & X_i > \tau - \epsilon,\ T_i \leq \epsilon \end{cases}$$

$D_i = 0$ corresponds with an uncensored line-segment on $[0, \tau - \epsilon]$, $D_i = 1$ corresponds with a singly-left or right-censored line-segment on $[0, \tau - \epsilon]$, but

right-censored so that it would also have been right-censored when using the window $[0,\tau]$, $D_i = 2$ corresponds with a doubly censored line-segment on $[0,\tau-\epsilon]$ and $D_i = 3$ corresponds with a singly-right censored line-segment on $[0,\tau-\epsilon]$, but which would have uncensored on $[0,\tau]$. Define $h_\epsilon \equiv h + \int_0^\epsilon g(\tau - x)dx$. Now, the distributions of the data are:

$$\begin{aligned} P(dz,0) &= I_{[0,\tau-\epsilon)}(z)\frac{\tau-u}{\tau+u}dV(u) \\ P(dz,1) &= I_{[0,\tau-\epsilon)}(z)g(z)dz \\ P(D=2) &= h_\epsilon \\ P(D=3) &= \epsilon g(\tau-\epsilon). \end{aligned}$$

The equivalent of the relation (5.6) is

$$\bar{V}(\tau-\epsilon) = (2\tau-\epsilon)g(\tau-\epsilon) + h_\epsilon. \tag{5.36}$$

This shows that the distribution of the data is indexed by $V$ on $[0,\tau-\epsilon)$ and $h_\epsilon$. The EM-equations are the same: just replace $h$ by $h_\epsilon$ and (5.6) by (5.6). For this case it can be shown (see Wijers, 1993, page 139) that $V_n$, the sieved-NPMLE for this transformed data, is uniformly consistent on $[0,\tau-\epsilon)$ and hence $g_n(\tau-\epsilon)$ is consistent so that our analysis can be imitated for this case. The reason that his proof now works is that $P_n(D=3)$ is uniformly (in $n$) bounded away from zero which by the loglikelihood principle enforces the NPMLE $P_{V_n,h_n}(D=3)$ of $P(D=3) = \epsilon g(\tau-\epsilon)$ to be uniformly bounded away from zero and hence that $g_n(\tau-\epsilon) = P_{V_n,h_n}(D=3)/\epsilon$ is bounded way from zero. In order to obtain a fully efficient estimator $W_n^\epsilon$ of $W$ one should let $\epsilon_n \to 0$ slowly enough when the number of observations converge to infinity. A practical suggestion is given in Wijers (1993).

### 5.6.1 Inhomogeneous Poisson Process.

Suppose that the starting points $T$ of the line-segments follow a inhogeneous Poisson process with intensity measure $\lambda(t)dt$. By going through the same steps as in beginning of section 1 we show that after having conditioned on the observed line-segments each line-segment corresponds with an i.i.d. observation $(X_i,T_i)$:

$$X \sim dV(x) \equiv \frac{S(x)}{S}dF(x),$$

where

$$S(x) = \int_{-x}^{\tau}\lambda(t)dt \text{ and } S = \int_0^\infty\int_{t=-x}^{\tau}\lambda(t)dtdF(x),$$

and $T$, given $X = x \in [0, \tau)$, have conditional distribution

$$dA(t \mid X = x) \equiv I_{(-x,\tau)}(t)\frac{\lambda(t)dt}{S(x)}.$$

This means that in the distribution of the data we replace the uniform on $(-x, \tau)$ by $dA(t \mid x)$ and similarly one rewrites the EM-equations in this way. We can estimate $\lambda(t)$ on $[0, \tau)$. Hence by assuming a certain parametric shape this can provide us with an estimate of $\lambda(t)$ on $(-\tau, tau)$ which is all we need. In this way our estimation method is easily generalized to the case where $\lambda(t)$ is not constant, but is replaced by an estimate. Also our efficiency results did not rely on the fact that $T \mid X = x$ was uniform, but go through for any other known distribution.

# Part II
# Inefficient Estimation in Semiparametric Models

# Chapter 6

# Inefficient Estimators of the Bivariate Survival Function in the Bivariate Censoring Model

## 6.1 Three approaches to estimation.

In this section we will describe three *representations* of the bivariate survival function, as maps from the distribution function of the data, on which three estimators can be based. The estimators are obtained by substituting the empirical distribution of the data into the representation.

Our aim is to prove that these estimators are *uniformly consistent* and that the estimators *converge weakly* as random elements in the bivariate cadlag function space $D[0,\tau]$ endowed with the *supremum norm* at root-$n$ rate to a Gaussian process. Moreover, we also want to show that the *bootstrap* can be used to estimate the variance of these estimators and we obtain some local *efficiency* results for these estimators.

The weak convergence and bootstrap results can be proved by applying the *functional delta-method* (see theorem 1.6). This means that we have to verify the required differentiability of the representation and the weak convergence of the empirical process which we plug into the representation. We are able to verify these conditions under essentially no conditions on the model. For a formal statement of our results see our final theorem in section 5. We also succeeded in proving that the Dabrowska and Prentice-Cai estimator are efficient under

independence. Practical simulations show that the asymptotic distribution is closely approached for surprisingly small samples (Bakker, 1990, Prentice and Cai, 1992a and chapter 8 van der Laan, 1993d).

The organization of the chapter is as follows. In section 2 we will give the basic techniques as lemmas for obtaining the required *differentiability* result for the representations and illustrate how these lemmas lead to the required convergence of the hardest terms which appear in our differentiability proofs. In section 3 we will prove the differentiability results by applying these lemmas. In section 4 we will see how each representation leads to an estimator by just substituting the empirical distribution of the data. In section 5 we verify the weak convergence of these empirical processes which provide us, by application of the functional delta method, with results which are summarized in our final theorem. Finally, in section 6 we prove that for the bivariate censoring model the Dabrowska and Prentice-Cai estimator are efficient under independence. For the sake of completeness we describe the bivariate censoring model once more:

**Model.** *Bivariate random censoring.*
Suppose that $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ are independent and identically distributed copies of $T = (T_1, T_2)$ which has distribution function $F$ on $\mathbb{R}^2_{\geq 0} \equiv [0, \infty)^2$, that $(C_{11}, C_{21}), \ldots, (C_{1n}, C_{2n})$ are independent and identically distributed with distribution function $G$ on $\mathbb{R}^2_{\geq 0}$ independent of all of the $(T_1, T_2)$'s, and that we observe $n$ i.i.d. copies

$$\left(\widetilde{T}_{1i}, \widetilde{T}_{2i}, D_{1i}, D_{2i}\right) \equiv (T_{1i} \wedge C_{1i}, T_{2i} \wedge C_{2i}, I(T_{1i} \leq C_{1i}), I(T_{2i} \leq C_{2i})).$$

*Problem*: Use the observed data to estimate $F$.

We assumed that we have observations in $\mathbb{R}^2_{\geq 0}$. The estimators we propose are invariant under monotone transformations. Therefore our results can be generalized to data on $\mathbb{R}^2$.

The analyzed estimators have natural generalizations to the $k$-variate case, and the $k$-variate analysis can be done by simply using $k$-variate analogues of the ingredients we use in the analysis for the case $k = 2$; for some of these, see Gill (1992). (He shows that no further ingredients are needed for general $k$.)

We refer to the bibliography on the bivariate random censoring model at the end of chapter 4. In chapter 4 we proposed a modified maximum likelihood estimator, a SOR-MLE, which depends on a grid-width $h_n$ ($n$ is the number of observations), which is proved to be efficient for $h_n \to 0$ slowly enough. The

choice of the grid-width in practice is as with density estimation a problem. Moreover, we needed an additional smoothness assumption on $F$ (though a rather weak one). This estimator had to be computed with the EM-algorithm, which is quite computer intensive. We also saw that the efficient influence function is implicitly defined. Only in the special case of independence will we (see section 6 of this chapter) succeed in obtaining an explicit expression for the information bound.

These difficulties in constructing efficient estimators; that they only seem to work under additional regularity assumptions and/or reduction of the data; and that they are computer intensive; are a motivation for considering inefficient estimators.

In this chapter we focus on three inefficient estimators, but estimators (except the Volterra estimator) which have been shown to have good practical performance. We included the Volterra representation because it is included in the Prentice-Cai representation, and the analysis of the Dabrowska and Volterra estimator gives the analysis of the Prentice-Cai estimator for free. The estimators are explicit and easy (quickly) to compute. The estimators are very smooth functions of the observations and therefore they are very robust: i.e. insensitive to small changes of the underlying distributions. Moreover, the only condition we need for obtaining consistency, weak convergence and bootstrap results on $[0, \tau]$ is that there is mass on $[\tau, \infty)$ in $F$ and $G$. Also the last two properties are certainly not shared with efficient estimators: in chapter 4 we needed beyond the grid-reduction of the singly censored observations to reduce the data to $[0, \tau]$ before we were able to prove these results.

Our approach to estimation of $F$ in these three models is as follows: we find representations of $F$ as maps $\Phi$ from the distribution of the data, which can be estimated from the observed data, to $F$. The three particular representations which we study here are given by:

**A.** Dabrowska's (1988) representation.

**B.** The Volterra equation.

**C.** Prentice and Cai (1992a) representation.

We give a new proof of the Prentice and Cai (1992a) representation.

**Notation and Definition of $[0,\tau]$.** If we write $\leq, \geq, <, >$ then this should hold componentwise for both components: so if $x \in \mathbb{R}^2$ then $x \leq y \Leftrightarrow x_1 \leq y_1, x_2 \leq y_2$. We often will not use a special notation for the *bivariate* time-vector; if we do not mean a vector this will be made clear. If $F(t) = P(X \leq t)$ is a distribution function we will denote its survival function with $S(t) = P(X > t)$. All functions we encounter are defined on a rectangle $[0,\tau] \subset \mathbb{R}^2_{\geq 0}$ where $\tau$ can be chosen arbitrarily large except that $S(\tau-) > 0$ and $H(\tau-) > 0$ is required. Finally, we define for a bivariate function $f : \mathbb{R}^2 \to \mathbb{R}$ $\|f\|_\infty \equiv \sup_{x \in [0,\tau]} | f(x) |$.

**A. Dabrowska's Representation.** The representation, the estimator and $L$-measure were all introduced by Dabrowska (1988), but in a rather different way than we do here; we take the representation in terms of product-integrals as done in Gill, 1990 (see also Andersen, Borgan, Gill, Keiding, 1993). We define the following three hazard measures with their heuristic interpretation:

$$\begin{aligned}
\Lambda_{10}(du, v-) &= P\left(T_1 \in [u, u+du) \mid (T_1, T_2) \geq (u, v)\right), \\
\Lambda_{01}(u-, dv) &= P\left(T_2 \in [v, v+dv) \mid (T_1, T_2) \geq (u, v)\right), \\
\Lambda_{11}(du, dv) &= P\left(T_1 \in [u, u+du),\ T_2 \in [v, v+dv) \mid T_1 \geq u,\ T_2 \geq v\right).
\end{aligned}$$

Formally, we introduce a vector *hazard* function $\vec{\Lambda} : [0,\tau] \subset \mathbb{R}^2_{\geq 0} \to \mathbb{R}^3_{\geq 0}$ as follows: $\vec{\Lambda}(t) \equiv (\Lambda_{10}(t), \Lambda_{01}(t), \Lambda_{11}(t))$, $t \in \mathbb{R}^2_{\geq 0}$, where

$$\begin{aligned}
\Lambda_{10}(t) &= \int_{[0,t_1]} \frac{1}{S(u-, t_2)} F(du, t_2) \qquad (6.1) \\
\Lambda_{01}(t) &= \int_{[0,t_2]} \frac{1}{S(t_1, v-)} F(t_1, dv) \\
\Lambda_{11}(t) &= \int_{[0,t]} \frac{1}{S(u-, v-)} F(du, dv).
\end{aligned}$$

One of the main advantages of model building in terms of hazards is that they are undisturbed by censoring and therefore we can get natural estimates of the integrated hazards by replacing them by their natural empirical counterparts (see section 4).

For a bivariate distribution $M$ (i.e. measure) $\prod_{(0,t]}(1 + dM)$ is the bivariate product integral over the rectangle $[0,t]$ (see Gill, Johansen, 1990, or our section 3.2). It is just like the univariate product integral defined as the limit of finite products over finite rectangular partitions of $[0,t]$. Now, the following

representation can be proved:

$$\begin{aligned} S(t) &= \prod_{[0,t_1]}(1-\Lambda_{10}(du,0)) \prod_{[0,t_2]}(1-\Lambda_{01}(0,dv)) \prod_{[0,t]}(1-L(du,dv)) \quad (6.2) \\ &\equiv \Gamma_1\left(\Lambda_{10}(\cdot,0),\Lambda_{01}(0,\cdot),\Gamma_2(L)\right), \end{aligned}$$

$L$ is defined by

$$\begin{aligned} L(t) &\equiv \int_{[0,t]} \frac{\Lambda_{10}(du,v-)\Lambda_{01}(u-,dv)-\Lambda_{11}(du,dv)}{(1-\Lambda_{10}(\Delta u,v-))(1-\Lambda_{01}(u-,\Delta v))} \quad (6.3) \\ &\equiv \Gamma_3(\Lambda_{10},\Lambda_{01},\Lambda_{11}), \end{aligned}$$

and $\Gamma_2$ represents the bivariate product-integral mapping. With $\Lambda_{10}(\Delta u, v-) = \Lambda_{10}(u,v-) - \Lambda_{10}(u-,v-)$ we denote the jump of $s \to \Lambda_{10}(s,v-)$ at $u$. Assume that for each $v$ $(u \mapsto F(u,v)) \ll \mu_1$ and for each $u$ $(v \mapsto F(u,v)) \ll \mu_2$ for certain (signed) measures $\mu_1$ and $\mu_2$. We define

$$\int F(du,v)G(u,dv) \equiv \int \frac{dF}{d\mu_1}(u,v)\frac{dG}{d\mu_2}(u,v)d\mu_1(u)d\mu_2(v), \quad (6.4)$$

where the Radon-Nykodim derivatives are taken in $u$ for fixed $v$ and in $v$ for fixed $u$, respectively. These assumptions are easily verified for the hazard measures by choosing $\mu_1 = S_1$ and $\mu_2 = S_2$, the marginals of $S$. We will see in section 4 that the empirical counterpart $\vec{\Lambda}_n$ of $\vec{\Lambda}$ is obtained by replacing in the representation of $\vec{\Lambda}$ $S$ by an empirical survival function. Therefore, the assumptions are verified in exactly the same way by choosing $\mu_1$ and $\mu_2$ the marginals of this empirical survival function. We will do this in the proof of the final theorem in section 5.

Note that by (6.2) and (6.3) this gives a map $\Gamma$ such that

$$\begin{aligned} S &= \Gamma(\vec{\Lambda}) \equiv \Gamma_1\left(\Lambda_{10}(\cdot,0),\Lambda_{01}(0,\cdot),\Gamma_2(L)\right) \\ &= \Gamma_1\left(\Lambda_{10}(\cdot,0),\Lambda_{01}(0,\cdot),\Gamma_2\Gamma_3(\vec{\Lambda})\right). \quad (6.5) \end{aligned}$$

This representation can quite easily be heuristically verified, basically using the same idea as with the one-dimensional Kaplan-Meier product-integral.

In section 4 we will give natural empirical estimators of $\vec{\Lambda}$ which generalize the famous *Nelson-Aalen* estimator from the one dimensional case.

If we denote the estimate of $\vec{\Lambda}$ with $\vec{\Lambda}_n$, then the estimate of $S$ based on Dabrowska's representation is simply

$$S_n = \Gamma(\vec{\Lambda}_n).$$

This estimator was studied by Dabrowska (1988, 1989). Gill (1990) generalized the representation to dimension $k \geq 2$ and analyzed the estimator by applying the functional delta-method.

**B. The Volterra equation.** This equation is derived by extending the following argument for $k = 1$: let

$$\Lambda(t) \equiv \int_{[0,t]} \frac{1}{S_-} dF \text{ for } t \geq 0$$

be the cumulative hazard function corresponding to $F$. Then

$$F(t) = \int_{[0,t]} S_- d\Lambda$$

and consequently

$$S(t) = 1 - \int_{[0,t]} S_- d\Lambda.$$

For a given function $\Lambda$, this is a *homogeneous Volterra equation* for $S$, where the solution is given by the *Peano series* (a special Neumann series) $\sum_{i=1}^{\infty} A^i(1)$, where $A(S) = \int_{[0,\cdot]} S_- d\Lambda$. In this case, $k = 1$, this is solved explicitly by the product-integral of $\Lambda$:

$$S(t) = \prod_{[0,t]} (1 - d\Lambda(s)).$$

For theory about the univariate product-integral and in particular the equivalence between the univariate Peano series and the univariate product-integral we refer to Gill and Johansen (1990).

For $k = 2$ the argument generalizes as follows. For $F$ on $[0, \infty)^2$, we define as above

$$\Lambda_{11}(t) \equiv \int_{[0,t]} \frac{1}{S_-} dF, \text{ where } S(x) \equiv P(X > x).$$

This implies that

$$F(t) = \int_{[0,t]} S_- d\Lambda_{11}. \tag{6.6}$$

It remains only to relate $F$ to $S$ and the marginal distributions: let $F_1$ and $F_2$ denote the marginal distributions of $F$. Then since

$$F_1(t_1) + F_2(t_2) - F(t) + S(t) = 1,$$

(6.6) yields

$$\begin{aligned} S(t) &= 1 - F_1(t_1) - F_2(t_2) + \int_{[0,t]} S_- d\Lambda_{11} \\ &\equiv \Psi(t) + \int_{[0,t]} S_- d\Lambda_{11}, \end{aligned} \tag{6.7}$$

where $\Psi(t) = 1 - F_1(t_1) - F_2(t_2)$ involves only the marginal distributions $F_1$ and $F_2$. Regarded as an equation for $S$ given fixed functions $\Psi$ and $\Lambda$, (6.7) is an *inhomogeneous Volterra equation* with a unique solution $\Phi_1(\Psi, \Lambda_{11})$ (Gill and Johansen (1990), Kantorovich and Akilov (1982), page 396). This can be seen as follows. Represent the equation as $(I - A_\Lambda)(S) = \Psi$ where $A_\Lambda(S)(t) = \int_{[0,t]} S_- d\Lambda$. It is easy to check that this structure takes care that

$$\|A_\Lambda^k(\Psi)\|_\infty \leq \frac{\|\Psi\|_\infty \|\Lambda\|_\infty^k}{k!},$$

where one has to notice that by definition of $\tau$ the supremum norm (over $[0, \tau]$) $\|\Lambda\|_\infty$ is bounded.

Consequently, $\sum_{k=0}^\infty A_\Lambda^k$ is a bounded operator:

$$\|\sum_{k=0}^\infty A_\Lambda^k(h)\|_\infty \leq \|h\|_\infty \exp(\|\Lambda\|_\infty).$$

This proves that $S$ is given by the Neumann series of $A_{\Lambda_{11}}$:

$$S = \sum_{k=0}^\infty A_{\Lambda_{11}}^k(\Psi). \tag{6.8}$$

Because $A_{\Lambda_{11}}$ depends only on $\Lambda_{11}$ and

$$\begin{aligned} \Psi(t) &= 1 - \prod_{(0,t_1]} (1 - \Lambda_{10}(ds_1, 0)) - \prod_{(0,t_2]} (1 - \Lambda_{10}(0, ds_2)) \\ &\equiv \Phi_2(\Lambda_{10}(\cdot, 0), \Lambda_{01}(0, \cdot)) \end{aligned}$$

(6.8) defines a map

$$S = \Phi(\vec{\Lambda}) \equiv \Phi_1(\Psi, \Lambda_{11}) = \Phi_1(\Phi_2(\Lambda_{10}(\cdot, 0), \Lambda_{01}(0, \cdot)), \Lambda_{11}). \tag{6.9}$$

It is not clear from (6.8) that $F$ depends continuously on $\Lambda_{11}$, but we will prove in section 3, as in Gill and Johansen (1990), that the bivariate Peano series, and thereby also $\Phi_1$, satisfies the characterization of *weak continuous compact differentiability* at $(\Psi, \Lambda)$ as stated in theorem 1.6.

Finally, it should be noticed that because of the exponential convergence of the terms $A_{\Lambda_{11}}^k$ to zero, (6.8) provides us also with an exponentially fast algorithm for finding a solution of the Volterra equation for known $(\Psi, \Lambda_{11})$. Finally, the Volterra estimator of $S$ is given by:

$$S_n \equiv \Phi(\vec{\Lambda}_n).$$

**C. Prentice and Cai's representation.** We give a new proof of the Prentice-Cai representation (see also Prentice and Cai, 1992a). For still another proof, see Wellner (1993). For this we need the following differentiability rules for $U : \mathbb{R} \to \mathbb{R}$ and $V : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} d(UV) &= U_- dV + dUV \\ d(\frac{1}{U}) &= \frac{dU}{UU_-}. \end{aligned}$$

If we apply these one dimensional rules to the sections $u \to F(u,v)$ and $v \to F(u,v)$ of a bivariate function $F$, then we denote these with $d_1$ and $d_2$, respectively. We apply these two one dimensional rules to each of the two variables of $R \equiv S/S_1S_2$ in turn in order to express $dR = d_{12}R = d_1(d_2(R))$ as follows:

$$dR = R_- d\widetilde{L} \text{ for certain measure } \widetilde{L}.$$

Define the familiar univariate hazards $\Lambda_1(ds_1) \equiv \Lambda_{10}(ds_1, 0)$, $\Lambda_2(ds_2) \equiv \Lambda_{01}(0, ds_2)$. The reader can easily verify the following formulas (when applying the product rule to $S/S_i$ we give the left continuous version to $S$ instead of one of the $S_i$, $i = 1, 2$, and we denote $F_{(-1)}(s_1, s_2) \equiv F(s_1-, s_2)$, $F_{(-2)}(s_1, s_2) \equiv F(s_1, s_2-)$):

$$\begin{aligned} & dR = d_{12}R \\ &= \frac{d_{12}S}{S_1S_2} - \frac{d_2S_2 d_1S_{(-2)}}{S_2S_{2-}S_1} - \frac{d_1S_1 d_2S_{(-1)}}{S_1S_{1-}S_2} + \frac{d_1S_1 d_2S_2 S_-}{S_1S_{1-}S_2S_{2-}} \\ &= R_- \frac{S_{1-}S_{2-}}{S_1S_2}\left(\frac{d_{12}S}{S_-} - \frac{d_2S_2}{S_{2-}}\frac{d_1S_{(-2)}}{S_-} - \frac{d_1S_1}{S_{1-}}\frac{d_2S_{(-1)}}{S_-} + \frac{d_1S_1}{S_{1-}}\frac{d_2S_2}{S_{2-}}\right) \\ &= R_-\left(\frac{\Lambda_{11}(ds) - \Lambda_2(ds_2)\Lambda_{10}(ds_1, s_2-) - \Lambda_1(ds_1)\Lambda_{01}(s_1-, ds_2) + \Lambda_1(ds_1)\Lambda_2(ds_2)}{(1-\Lambda_1(\Delta s_1))(1-\Lambda_2(\Delta s_2))}\right) \\ &\equiv R_- d\widetilde{L}. \end{aligned}$$

At the third equality notice that $1/(1-\Lambda_1(\Delta s_1))(1-\Lambda_2(\Delta s_2)) = S_{1-}S_{2-}/S_1S_2$. Integrating the left and right-hand side over the rectangle $(0,t]$ provides us with:

$$R(t) = 1 + \int_{(0,t]} R(s-)\widetilde{L}(ds). \tag{6.10}$$

Here one has to notice that $R(t_1, 0) = R(0, t_2) = 1$ for all $t$. This is a homogeneous Volterra equation with a unique solution given by the Peano series of $\widetilde{L}$ which we will write out below.

By definition of $R$ and the well known product integral representation of the univariate $S_i$, $i = 1, 2$, this provides us with the following representation for the bivariate survival function:

$$\begin{aligned} S(t) &= \prod_{[0,t_1]} (1 - \Lambda_{10}(du, 0)) \prod_{[0,t_2]} (1 - \Lambda_{01}(0, dv))\, R(t) \\ &\equiv \Theta_1 (\Lambda_{10}, \Lambda_{01}, R), \end{aligned} \tag{6.11}$$

where $R$ is the unique solution of (6.10), just the Neumann series $\sum_{k=0}^{\infty} A_{\widetilde{L}}^{k}(1)$ as given in (6.8), given by the Peano series:

$$\begin{aligned} R &= 1 + \sum_{n=1}^{\infty} \int \dots \int_{0 \le u^1 < u^2 < \dots < u^n \le t} \prod_{j=1}^{n} \widetilde{L}(du^j) \\ &\equiv \Theta_2(\widetilde{L}). \end{aligned}$$

Define

$$\beta(s) \equiv (1 - \Lambda_{10}(\Delta s_1, 0))\,(1 - \Lambda_{01}(0, \Delta s_2)).$$

Above, we derived the following representation of $\widetilde{L}$ in terms of the hazards $\vec{\Lambda}$

$$\begin{aligned} \widetilde{L}(t) &= \int_{(0,t]} \frac{1}{\beta(s)} \{\Lambda_{11}(ds) - \Lambda_{10}(ds_1, s_2-)\Lambda_{01}(0, ds_2) \\ &\quad -\Lambda_{10}(ds_1, 0)\Lambda_{01}(s_1-, ds_2) + \Lambda_{10}(ds_1, 0)\Lambda_{01}(0, ds_2)\} \\ &\equiv \Theta_3 \left(\vec{\Lambda}\right). \end{aligned}$$

Note that this gives a map

$$\begin{aligned} S &= \Theta(\vec{\Lambda}) = \Theta_1 \left(\Lambda_{10}, \Lambda_{01}, \Theta_2(\widetilde{L})\right) \\ &= \Theta_1 \left(\Lambda_{10}, \Lambda_{01}, \Theta_2\Theta_3 \left(\vec{\Lambda}\right)\right). \end{aligned} \tag{6.12}$$

Again, the estimate of $S$ based on the Prentice-Cai representation is simply

$$S_n = \Theta(\vec{\Lambda}_n).$$

Prentice and Cai (1992a) motivated this representation through a connection between $\widetilde{L}$ and the covariance of univariate counting process martingales. Moreover, they proved almost sure consistency of the resulting estimator via continuity of $\Theta$.

**Remark.** Firstly, the Volterra estimator is based on the idea to express $dS$ in $S\_dA$ for a certain measure $A$ which makes $S$ a solution of an *inhomogeneous* Volterra equation, while in Prentice-Cai's representation we do the same with

$d(S/S_1S_2)$ which leads in this case to a *homogeneous* Volterra equation. The Volterra estimator uses an estimate of only one bivariate hazard $\Lambda_{11}$, while the Dabrowska and the Prentice-Cai representations involve other functions $L$ and $\widetilde{L}$ which describe the covariance structure. Furthermore, notice the similarity in the structure of $L$ and $\widetilde{L}$; this will save work in the differentiability proofs.

**The functional delta-method.** Our approach to studying the estimators, which we will denote by $S_n^V, S_n^D, S_n^{PC}$ will be to study the maps $\Phi$, $\Gamma$ and $\Theta$ which define them (analytically). In sections 2 and 3 we show that these satisfy weak continuous Hadamard differentiability with respect to the supremum norm-metric for the sequences which can occur in practice. In section 4 we represent $\vec{\Lambda}$ as differentiable maps from the distribution function of the data to $\vec{\Lambda}$, which, by the chain-rule gives us differentiable mappings $\Phi \circ \vec{\Lambda}$, $\Gamma \circ \vec{\Lambda}$ and $\Theta \circ \vec{\Lambda}$. Application of the functional delta-method theorem 1.6 to these representations in the distribution of the data provides us with consistency, weak convergence, and asymptotic validity of the bootstrap for $S_n^V, S_n^D, S_n^{PC}$. These results are summarized in theorem 5.1.

This approach provides us with optimal results for the estimators in the sense that we essentially do not need any conditions. The only improvements can be made by extending these results to the whole plane and by investigating the rate at which the normalized estimators converge to its linearization in terms of the empirical processes we plug in.

Also, the analysis has been separated into a purely analytical part (differentiability of $\Phi$) and a purely probabilistic part (weak convergence of $Z_n = \sqrt{n}(P_n - P)$), where the latter is well known in our case. After establishing these purely analytical properties of components of $\Phi$ one may also conclude similar results for different sampling methods or models (e.g. models 1–3 in Gill, van der Laan and Wellner, 1993) without repeating the analysis. The supremum norm might be considered as a quite naive choice in order to get an optimal weak convergence result, but the supremum norm is easy to use, to interpret, and has an easy generalization to higher dimensions.

After establishing the differentiability of the functionals which appear in the representation $\Gamma$ and $\Phi$ we will get the differentiability of the Prentice-Cai representation for free: by the chain-rule a differentiability result for a functional can be used for establishing differentiability for any composition of several mappings involving this functional. Because of this property, other proposed explicit estimators can immediately be put into our framework.

## 6.2 Lemmas.

In this section we give lemmas containing the basic analytic techniques for establishing the required differentiability property. Moreover, we will give two illustrations which show how these techniques can be used for this purpose.

**Definitions and Notation.** All functions we encounter are considered as elements of the space of bivariate cadlag functions on $[0,\tau]$ which is denoted by $D[0,\tau]$, endowed with the *supremum norm* (see definition 1.2). If $F_n$ converges in supremum norm to $F$, then we will denote this with $F_n \to F$ (no other types of convergence are needed here). Furthermore, we refer to some basic facts mentioned in section 2.3: If a cadlag function is of bounded variation, then it generates a signed measure and we refer to the telescoping lemma 1.6. If for $F$ we write $F(du,v), F(u,dv), F(du,dv)$, we mean the one dimensional measures generated by the sections $u \mapsto F(u,v), v \mapsto F(u,v)$ and the two dimensional measure generated by $(u,v) \mapsto F(u,v)$, respectively, and it will be automatically assumed that these sections and the function itself are of bounded variation.

In our applications we want to be able to define integrals $\int F dH$ and $\int F(du,v)H(u,dv)$, when $H$ is of *unbounded* variation and $F$ is of bounded uniform sectional variation. This can be done by applying *integration by parts* so that $H$ appears as function. We want to have an integration by parts formula which takes account of mass given to the edge of the rectangle $[0,\tau]$. The following does this (here $\int_0^s = \int_{(0,s]}$, the first formulas is the same as in lemma (1.3)):

**Lemma 6.1** (Integration by Parts).

$$\begin{aligned}\int_0^s\int_0^t F(u,v)H(du,dv) &= \int_0^s\int_0^t H\left([(u,s)\times(v,t)]\right)F(du,dv)\\ &+\int_0^s H\left([u,s]\times(0,t]\right)F(du,0)+\int_0^t H\left((0,s]\times[v,t]\right)F(0,dv)\\ &+F(0,0)H\left((0,s]\times(0,t]\right).\end{aligned}$$

$$\int_0^s\int_0^t F(du,v)H(u,dv) = \int_0^s\int_0^t H(u,[v,t])F(du,dv)+\int_0^s H(u,(0,t])F(du,0).$$

Notice that with these formulas we can also define these integrals for $H$ of unbounded variation. We can bound both integrals by $16\|H\|_\infty\|F\|_v^*$. Notice that if $F$ is zero at the bottom edges of $[0,\tau]$, then only the first term on the right hand-sides is non-zero.

**Proof.** We refer to (Gill, 1990) for the general $\mathbb{R}^k$ case. It works as follows. For the first integral, substitute

$$F(u,v) = \int_{(0,u]\times(0,v]} F(du',dv') + \int_{(0,u]} F(du',0) + \int_{(0,v]} F(0,dv') + F(0,0)$$

and for the second integral substitute

$$F(du,v) = \int_{(0,v]} F(du,dv') + F(du,0)$$

and apply Fubini's theorem. □

Assume that $F_1, F_2$ are of bounded variation and that $F_i(s_1,s_2)$ or $F_i(s_1-,s_2)$ or $F_i(s_1,s_2-)$ or $F_i(s_1-,s_2-)$ is cadlag, $i = 1,2$. Then $F_1$ and $F_2$ generate signed measures and we have

$$\int F_1(u)F_2(u)dH(u) = \int F_1(u)d\left(\int_0^u F_2(v)dH(v)\right). \tag{6.13}$$

So by twice applying lemma 6.1 to $\int F_1(u)F_2(u)dH(u) = \int F_1(u)d\left(\int_0^u F_2(v)dH(v)\right)$ we can do integration by parts so that $H$ appears as function and $F_1, F_2$ as measures.

The following lemma is trivially checked, but useful.

**Lemma 6.2** ($d$-$\Delta$ interchange). *We have:*

$$\begin{aligned}\int\int F(\Delta s,\Delta t)G(ds,dt) &= \int\int F(ds,dt)G(\Delta s,\Delta t)\\ \int\int F(ds,t)G(\Delta s,dt) &= \int\int F(\Delta s,t)G(ds,dt).\end{aligned}$$

Recall the denominator in the mappings $L$ and $\widetilde{L}$ which appear in $\Gamma$ and $\Theta$, which are of the form $1/\{(1-a)(1-b)\}$, where $a, b$ are are only nice functions in one coordinate, and therefore certainly do not generate a measure. Therefore it is not clear how we can integrate w.r.t. this denominator. The following lemma will take care of this problem.

**Lemma 6.3** (Denominator splitting). *Let $a_1, a_2$ be real numbers.*

*Then the following holds:*

$$\frac{1}{(1-a_1)(1-a_2)} = \frac{1}{1-a_1} + \frac{a_2}{(1-a_1)(1-a_2)}$$

*or*

$$\frac{1}{(1-a_1)(1-a_2)} = 1 + \frac{a_1}{1-a_1} + \frac{a_2}{1-a_2} + \frac{a_1a_2}{(1-a_1)(1-a_2)}.$$

In general, we have:

$$\frac{1}{\prod_{i=1}^{n}(1-a_i)} = 1 + \sum_i \frac{a_i}{1-a_i} + \sum_{i,j,i\neq j} \frac{a_i a_j}{(1-a_i)(1-a_j)} + \ldots + \frac{\prod_{i=1}^{n} a_i}{\prod_{i=1}^{n}(1-a_i)}.$$

This follows from the identity $1/(1-a_i) = 1 + a_i/(1-a_i)$.

Now, we are able to define the following terms with integration by parts as follows:

**Corollary 6.1** *Define* $\beta(u,v) \equiv (1-\Lambda_{10}(\Delta u, v))(1-\Lambda_{01}(u,\Delta v))$. *We have:*

$$\int\int \frac{H(du,v)\Lambda(u,dv)}{\beta(u,v)} = \int\int H(du,v)\frac{\Lambda(u,dv)}{1-\Lambda_{01}(u,\Delta v)} + \int\int \frac{H(\Delta u,v)\Lambda_{10}(du,v)\Lambda(u,dv)}{\beta(u,v)}$$

$$\int\int \frac{H(du,dv)}{\beta(u,v)} = \int\int H(du,dv) + \int\int H(du,\Delta v)\frac{\Lambda_{01}(u,dv)}{1-\Lambda_{01}(u,\Delta v)} + \int\int \frac{H(\Delta u,\Delta v)\Lambda_{10}(du,v)\Lambda_{01}(u,dv)}{\beta(u,v)} + \int\int H(\Delta u,dv)\frac{\Lambda_{10}(du,v)}{1-\Lambda_{10}(\Delta u,v)}.$$

$H$ plays the role of a function of unbounded variation (Brownian bridge) and $\Lambda, \Lambda_{10}, \Lambda_{01}$ are cadlag functions of bounded uniform sectional variation. Notice that all terms on the right-hand side of the equalities where $H$ appears as measure are of the form $\int F dH$ where $F$ generates a finite measure. Therefore, for all these terms we can apply the integration by parts formulas of lemma 6.1 in order to make $H$ appear as function.

Again, this corollary is simple to prove by applying denominator splitting and $d$–$\Delta$-interchange. In the differentiability proof of the $L$-mapping we have to be able to bound the terms above in the supremum norm of $H$. It is now clear that this can be done with the integration by parts formulas. We will see that this is the whole story of the differentiability proofs: we use denominator-splitting and $d$–$\Delta$-interchange in order to produce an integral $\int F dH$, where $F$ generates a measure and is of bounded uniform sectional variation. Then we apply integration by parts in order to bound these terms in the supremum norm of $H$ and the uniform sectional variation of $F$.

We did not deal, yet, with an integral of the form $\int H dF_n$, $\|F_n\|_\infty \to 0$, $\|H\|_v = \infty$, which we want to show to converge to zero. Since $H$ is not of finite variation one cannot do integration by parts in order to bound this in the supremum norm of $F_n$. The next ingredient takes care of this, the so called Helly-Bray lemma:

**Lemma 6.4** (Helly-Bray lemma). *If $H \in (D[0,\tau], \|\cdot\|_\infty)$ is of unbounded variation, then we can approximate $H$ with a sequence $H_m$ where $\|H_m\|_v^* \leq M(m) < \infty$ and $\|H - H_m\|_\infty \to 0$. This gives us the following bound:*

$$\|\int H dF\|_\infty \leq \|H - H_m\|_\infty \|F\|_v^* + 16\|F\|_\infty M(m).$$

For $H_m$ one can (e.g.) take the step function equal to $H$ on a grid $\pi^m$. We did the substitution $H = (H - H_m) + H_m$, integration by parts and bounding terms like $\int F dH_m$ by $\|F\|_\infty \|H_m\|_v^*$. The bound in lemma 6.4 is useful because it proves that integrals of the form $\int H dF_n$ converge to zero when $\|F_n\|_\infty \to 0$, even if $\|H\|_v = \infty$, provided that $\|F_n\|_v^* < \infty$ (just let $m \to \infty$ slowly enough).

**Illustration 1.** We will illustrate how these lemmas easily provide us with compact differentiability (see theorem 1.5) of $\Phi : (F,G) \to \int F dG$ at a point $(F,G)$ with $F$ and $G$ cadlag functions of bounded uniform sectional variation for sequences $F_n, G_n$ of uniformly (in $n$) bounded uniform sectional variation: if $Y_n \equiv \sqrt{n}(F_n - F) \to Y$, $Z_n \equiv \sqrt{n}(G_n - G) \to Z$, then

$$\sqrt{n}(\Phi(F_n, G_n) - \Phi(F,G)) - d\Phi(F,G)(Y,Z) \to 0$$

for a certain continuous linear mapping $d\Phi(F,G) : (D[0,\tau])^2 \to \mathbb{R}$. We have by telescoping:

$$\sqrt{n}(\Phi(F_n, G_n) - \Phi(F,G)) = \int Y_n dG + \int F_n dZ_n.$$

So if we subtract from this its supposed limit $d\Phi(F,G)(Y,Z) = \int Y dG + \int F dZ$, then we obtain by telescoping:

$$\int (Y_n - Y) dG + \int (F_n - F) dZ + \int F_n d(Z_n - Z),$$

where the last two integrals are defined by integration by parts (lemma 6.1). The first integral can immediately be bounded by $\|Y_n - Y\|_\infty \|G\|_v \to 0$. The second integral converges to zero by the Helly-Bray lemma 6.4. For the third integral we can do integration by parts with respect to $F_n$ and thereby bound this term by $c\|Z_n - Z\|_\infty \|F_n\|_v^* \to 0$.

**Illustration 2.** We will give an illustration of how these lemmas are used to prove convergence to zero of quite complicated terms which we will encounter in our analysis of Dabrowska's estimator. Consider the term

$$\int \frac{1}{\beta_n(u,v)} - \frac{1}{\beta(u,v)} H(du,dv), \tag{6.14}$$

where $H$ is of unbounded variation. $\beta_n(u,v)$ $(\beta(u,v))$ is the denominator of $L$ as defined above corresponding to $(\Lambda_{10}^n, \Lambda_{01}^n)$ $((\Lambda_{10}, \Lambda_{01}))$ and $(\Lambda_{10}^n, \Lambda_{01}^n)$ converges in supremum norm to $(\Lambda_{10}, \Lambda_{01})$. We will show that this term converges to zero if $\Lambda_{10}, \Lambda_{01}, \Lambda_{10}^n, \Lambda_{01}^n$ have the following four properties :

1) $\beta > \delta > 0$ on $[0,\tau]$ for certain $\delta > 0$.

2) There exists a sequence of uniformly in $n$ finite (signed) measures $\mu_{2n}$ so that $\Lambda_{10}^n(u,dv) \ll \mu_{2n}(dv)$ for all $u$. Similarly for $\Lambda_{10}, \Lambda_{01}, \Lambda_{01}^n$.

3) There exists a sequence of uniformly in $n$ finite (signed) measures $\mu_{1n}$ so that $\Lambda_{10}^n(du,v) \ll \mu_{2n}(du)$ for all $v$. Similarly for $\Lambda_{10}, \Lambda_{01}, \Lambda_{01}^n$.

4) $\|\Lambda_{10}^n(du,v)/\mu_{1n}(du)\|_\infty < M$ and $\|\Lambda_{10}^n(u,dv)/\mu_{2n}(dv)\|_\infty < M$ for certain $M < \infty$ (uniform boundedness of the Radon-Nykodym derivatives). Similarly for $\Lambda_{10}, \Lambda_{01}, \Lambda_{01}^n$.

In our applications the assumptions 2–4 are easily verified by a simple choice of $\mu_{1n}, \mu_1, \mu_{2n}, \mu_2$ and by choice of $[0,\tau]$ assumption 1 will hold trivially. This will be done in the proof of the final theorem in section 5.

The term (6.14) involves all the above mentioned techniques. Apply the *denominator splitting* lemma to rewrite $1/\beta_n(u,v) - 1/\beta(u,v)$. This gives

$$\begin{aligned}
&\frac{1}{\beta_n(u,v)} - \frac{1}{\beta(u,v)} \\
&\quad = \left(\frac{\Lambda_{10}^n(\Delta u,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(\Delta u,v)}{1-\Lambda_{10}(\Delta u,v)}\right) + \left(\frac{\Lambda_{01}^n(u,\Delta v)}{1-\Lambda_{01}^n(u,\Delta v)} - \frac{\Lambda_{01}(u,\Delta v)}{1-\Lambda_{01}(u,\Delta v)}\right) \\
&\qquad + \left(\frac{\Lambda_{10}^n(\Delta u,v)\Lambda_{01}^n(u,\Delta v)}{\beta_n(u,v)} - \frac{\Lambda_{10}(\Delta u,v)\Lambda_{01}(u,\Delta v)}{\beta(u,v)}\right).
\end{aligned}$$

Then the integral (6.14) is the sum of three integrals which we will denote with $A, B$ and $C$ respectively. The first term $A$ is given by:

$$\begin{aligned}
&\int \left(\frac{\Lambda_{10}^n(\Delta u,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(\Delta u,v)}{1-\Lambda_{10}(\Delta u,v)}\right) H(du,dv) \\
&= \int \left(\frac{\Lambda_{10}^n(\Delta u,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(\Delta u,v)}{1-\Lambda_{10}(\Delta u,v)}\right) (H-H_m)(du,dv) \\
&\quad + \int \left(\frac{\Lambda_{10}^n(\Delta u,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(\Delta u,v)}{1-\Lambda_{10}(\Delta u,v)}\right) H_m(du,dv) \\
&= \int \left(\frac{\Lambda_{10}^n(du,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(du,v)}{1-\Lambda_{10}(\Delta u,v)}\right) (H-H_m)(\Delta u,dv) \\
&\quad + \int \left(\frac{\Lambda_{10}^n(\Delta u,v)}{1-\Lambda_{10}^n(\Delta u,v)} - \frac{\Lambda_{10}(\Delta u,v)}{1-\Lambda_{10}(\Delta u,v)}\right) H_m(du,dv).
\end{aligned}$$

We did the substitution $H = H - H_m + H_m$ (Helly-Bray) and applied $d$–$\Delta$-interchange. Consider the first term, say $A1$.

**A1.** Here, we can apply the second integration by parts formula of lemma 6.1. Then one of the terms is given by:

$$\begin{aligned}
&\int (H - H_m)(\Delta u, [v, \tau_2]) \frac{1}{(1 - \Lambda_{10}^n(\Delta u, v))^2} \Lambda_{10}^n(du, v)\Lambda_{10}^n(\Delta u, dv) \\
&\leq \frac{4}{\left(\inf_{(u,v)\in[0,\tau]} \mid 1 - \Lambda_{10}^n(\Delta u, v) \mid\right)^2} \|H - H_m\|_\infty \int \mid \Lambda_{10}^n(du, v)\Lambda_{10}^n(\Delta u, dv) \mid \\
&\leq C\|H - H_m\|_\infty \int \mid \Lambda_{10}^n(du, v)\Lambda_{10}^n(\Delta u, dv) \mid,
\end{aligned}$$

where we used assumption $\beta_n > \delta > 0$ on $[0, \tau]$ for certain $\delta > 0$ in the last line, which follows from assumption 1 and the uniform convergence of $\beta_n$ to $\beta$. The other terms which one gets after applying integration by parts are dealt in the same way. By assumption 2-4 we have:

$$\begin{aligned}
\int \mid \Lambda_{10}^n(du, v)\Lambda_{10}^n(\Delta u, dv) \mid &= \int \left| \frac{\Lambda_{10}^n(du, v)}{\mu_{1n}(du)} \frac{\Lambda_{10}^n(u, dv)}{\mu_{2n}(dv)} \right| |\mu_{1n}(du)\mu_{2n}(dv)| \\
&\leq M^2 \|\mu_{1n}\|_v \|\mu_{2n}\|_v \leq M',
\end{aligned}$$

for some $M' < \infty$. So if $\Lambda_{10}^n$ satisfies assumptions 1–4, then $C\|H - H_m\|_\infty \int \mid \Lambda_{10}^n(du, v)\Lambda_{10}^n(\Delta u, dv) \mid \to 0$ for $m \to \infty$. The other terms are dealt similarly using the assumptions 1–4 for $\Lambda_{10}$ and $\Lambda_{10}^n$.

**A2.** The second term can be bounded by the supremum norm of $\Lambda_{10}^n(\Delta u, v)/(1 - \Lambda_{10}^n(\Delta u, v)) - \Lambda_{10}(\Delta u, v)/(1 - \Lambda_{10}(\Delta u, v))$ (which converges to zero) times the variation norm of $H_m$.

So if we let $m = m(n) \to \infty$ slowly enough for $n \to \infty$, then both terms $A1$ and $A2$ converge to zero.

The second term $B$ is dealt similarly. Now, we will deal with the third term $C$. Firstly, by *telescoping* we can rewrite:

$$\begin{aligned}
&\frac{\Lambda_{10}^n(\Delta u, v)\Lambda_{01}^n(u, \Delta v)}{\beta_n(u, v)} - \frac{\Lambda_{10}(\Delta u, v)\Lambda_{01}(u, \Delta v)}{\beta(u, v)} \\
&= \left(\frac{1}{\beta_n(u, v)} - \frac{1}{\beta(u, v)}\right) \Lambda_{10}(\Delta u, v)\Lambda_{01}(u, \Delta v) \\
&\quad + \frac{1}{\beta_n(u, v)} \left(\Lambda_{10}^n(\Delta u, v) - \Lambda_{10}(\Delta u, v)\right) \Lambda_{01}(u, \Delta v) \\
&\quad + \frac{1}{\beta_n(u, v)} \Lambda_{10}^n(\Delta u, v) \left(\Lambda_{01}^n(u, \Delta v) - \Lambda_{01}(u, \Delta v)\right).
\end{aligned}$$

We have to integrate these terms with respect to $H$. We set $H = (H - H_m) + H_m$ (here an application of the *Helly-Bray*-lemma starts). By using the $d$-$\Delta$-

*interchange* trick we can transform all three terms with $H - H_m$ into integrals where $H - H_m$ appears as a function: e.g.

$$\int (H - H_m)(du, dv) \frac{1}{\beta_n(u, v)} \left(\Lambda_{10}^n(\Delta u, v) - \Lambda_{10}(\Delta u, v)\right) \Lambda_{01}(u, \Delta v)$$
$$= \int (H - H_m)(\Delta u, \Delta v) \frac{1}{\beta_n(u, v)} \left(\Lambda_{10}^n(du, v) - \Lambda_{10}(du, v)\right) \Lambda_{01}(u, dv).$$

So if assumption 1–4 holds, then as we did above we can bound this term by

$$c\|H - H_m\|_\infty M^2 \left(\|\mu_{1n}\|_v \|\mu_2\|_v + \|\mu_1\|_v \|\mu_2\|_v\right) \leq M' \|H - H_m\|_\infty.$$

Similarly, we have this bound for the other terms with $H - H_m$. The three terms with $H_m$ we can directly bound by $\|(1/\beta_n(u, v) - 1/\beta(u, v))\|_\infty M(m)$ , $\|(\Lambda_{10}^n(\Delta u, v) - \Lambda_{10}(\Delta u, v))\|_\infty M(m)$, $\|(\Lambda_{01}^n(u, \Delta v) - \Lambda_{01}(u, \Delta v))\|_\infty M(m)$, where $M(m)$ stands for a constant times the variation norm of $H_m$. So we can conclude that we have the following bound:

$$\| \int H(du, dv) \left( \frac{\Lambda_{10}^n(\Delta u, v) \Lambda_{01}^n(u, \Delta v)}{\beta_n(u, v)} - \frac{\Lambda_{10}(\Delta u, v) \Lambda_{01}(u, \Delta v)}{\beta(u, v)} \right) \|_\infty$$
$$\leq c\|H - H_m\|_\infty + \epsilon_n M(m),$$

where $\epsilon_n$ converges to zero. Let now $m \to \infty$ slowly enough to obtain that the left-hand side bound converges to zero. This proves the convergence of (6.14).

In general all terms we will encounter in the differentiability proofs are dealt in the following way:

**Telescoping.** Step 1. Firstly, we do telescoping in order to rewrite a difference of two products as a sum of single differences: $\int A_n B_n - \int AB = \int (A_n - A)B + \int A_n(B_n - B)$. Consider one term (e.g.) $\int (A_n - A)B$. Here, we know that $A_n \to A$, but $A_n$ can appear as a measure in one or two coordinates: $\int (A_n - A)(du, dv)B(u, v)$ or $\int (A_n - A)(du, v)B(u, dv)$ or the easiest case $\int (A_n - A)(u, v)B(du, dv)$.

**Goal.** Step 2. We want to bound the term $\int (A_n - A)B$, where we usually have that $A_n - A$ appears as a measure, in the supremum norm of $A_n - A$ which is known to converge to zero. Therefore if $A_n - A$ does not appear as a function, then our goal is to get this term in a form so that we can apply integration by parts with respect to $B$.

**Denominator-splitting, $d$–$\Delta$-interchange.** Step 3.

**Case 0** If $A_n - A$ appears as function we can immediately bound $\int (A_n -$

$A)dB$ by the supnorm of $A_n - A$.

**Case 1** If $B$ is of bounded uniform sectional variation or is it a product of such functions (of bounded uniform sectional variation but some left and some right continuous) we can bound the term in the supremum norm of $A_n - A$ by applying the integration by parts formula of lemma 6.1.

**Case 2** If $B$ is of unbounded variation, we substitute $B = (B - B_m) + B_m$ and we now want to bound the term with $B - B_m$ in the supremum norm of $B - B_m$ and the term with $B_m$ in the uniform sectional variation norm of $B_m$ (Helly-Bray lemma 6.4). We go back to step 3.

**Case 3** If $B$ involves the denominator $\beta$ we firstly apply the denominator trick lemma 6.3 and $d - \Delta$-interchange lemma 6.2 as in corollary 6.1 in order to rewrite the term to a term of Case 0 or 1.

## 6.3 Differentiability of the Dabrowska, Volterra, and Prentice and Cai representations of $F$.

In this section our goal is to establish weak continuous Hadamard differentiability (see theorem 1.6) of the Volterra, Dabrowska, and Prentice-Cai representations of $F$, thereby paving the way for validity of the bootstrap in each case.

**Notation and assumptions on sequences.** For any symbol which occurs as argument of the analyzed mapping, say $\Lambda$, $\Lambda_n$ and $\Lambda_n^\#$ are sequences which both converge in supremum norm to $\Lambda$ and moreover it will be automatically assumed that they are of bounded uniform sectional variation uniformly in $n$. The latter can be done by choosing $D_n$ in theorem 1.6 appropriately and because these properties hold for the estimators we plug in.

$\Lambda_n$ plays the role of the estimator of $\Lambda$ using the original data and $\Lambda_n^\#$ plays the role of the same estimator, but using a bootstrap sample of the original data.

### 6.3.1 The Volterra Representation.

We give the proof of the Volterra representation before the proof of the bivariate product integral (as part of the Dabrowska representation), because the proof is easier to generalize from the univariate case and for the Dabrowska representation we will be able to refer to the main lines of the differentiability proof given here.

Consider the inhomogeneous Volterra equation

$$S(t) = \Psi(t) + \int_{[0,t]} S(s-)d\Lambda_{11}(s). \tag{6.15}$$

We consider this equation as an implicit equation for $S$ for given functions $\Psi$ and $\Lambda_{11}$. For any measure $\alpha$ on $\mathbb{R}^2$ set:

$$\mathcal{P}((s,t];\alpha) = 1 + \sum_{n=1}^{\infty} \int_{s\le u^1<\ldots<u^n\le t} \alpha(du^1)\ldots\alpha(du^n). \tag{6.16}$$

$\mathcal{P}_\alpha \equiv \mathcal{P}(\cdot;\alpha)$ is the Peano series corresponding to $\alpha$. The following propositions will be proved below in a separate subsection. The proofs are similar to the proofs given in Gill and Johansen (1990) as they already remarked on page 1531. The inhomogeneous Volterra equation has a unique solution in terms of $\mathcal{P}(\cdot;\Lambda_{11})$:

**Proposition 6.1** *If $S$ satisfies (6.15), then*

$$S(t) = \Psi(t) + \int_{0<s\le t} \Psi(s-)\mathcal{P}((s,t];\Lambda_{11})\,d\Lambda_{11}(s).$$

Repeated substitution of the Volterra equation into itself and interchange of the order of integration make the claim intuitively clear. Here are two propositions giving useful properties of the Peano series $\mathcal{P}$.

**Proposition 6.2** (Kolmogorov equations). *The Peano series $\mathcal{P} \equiv \mathcal{P}_\alpha$ defined by (6.16) satisfies*

$$\begin{aligned}\mathcal{P}_\alpha(s,t] &= 1 + \int_{s<u\le t} \mathcal{P}_\alpha(s,u)\alpha(du) \\ &= 1 + \int_{s<u\le t} \mathcal{P}_\alpha(u,t]\alpha(du).\end{aligned}$$

**Proposition 6.3** (Duhamel equation). *If $\alpha$ and $\beta$ are two measures on $\mathbb{R}^2$ with corresponding Peano series $\mathcal{P}_\alpha$ and $\mathcal{P}_\beta$, then*

$$\mathcal{P}_\beta(s,t] - \mathcal{P}_\alpha(s,t] = \int_{s<u\le t} \mathcal{P}_\alpha(s,u)\mathcal{P}_\beta(u,t](\beta-\alpha)(du). \tag{6.17}$$

With the Duhamel equation one can show the following differentiability result for the Peano series. For all propositions and theorems recall our assumptions on the sequences $\Lambda_n$.

**Proposition 6.4** (weak continuous compact differentiability of $\mathcal{P}_\alpha$ in supremum norm). *Assume*

$$h_n \equiv \sqrt{n}(\alpha_n^\# - \alpha_n) \to h \text{ in } D[0,\tau].$$

*Then, with* $\mathcal{P}_n^\# \equiv \mathcal{P}(\cdot;\alpha_n^\#)$, $\mathcal{P}_n \equiv \mathcal{P}(\cdot;\alpha_n)$

$$\sqrt{n}\left(\mathcal{P}_n^\# - \mathcal{P}_n\right) \to \dot{\mathcal{P}}h \text{ in } D[0,\tau], \tag{6.18}$$

*where* $\dot{\mathcal{P}}$ *is given by*

$$\dot{\mathcal{P}}h(s,t] = \int_{s<u\le t} \mathcal{P}(s,u)\mathcal{P}(u,t]dh(u). \tag{6.19}$$

*If h is of unbounded variation this is defined by (repeated) integration by parts (see lemma 6.13).*

**Consistency.** In general, notice that this differentiability result for a mapping $A$ certainly implies continuity of $A$; if $F_n \to F$ then $A(F_n) \to A(F)$. Therefore our differentiability results will also provide us with almost sure uniform consistency of our estimators.

Now, we have the tools to prove the weak continuous compact differentiability property of the Volterra representation $\Phi_1(\Psi, \Lambda_{11})$.

**Theorem 6.1** (weak continuous compact differentiability of $\Phi_1$). *Suppose that*

$$\begin{aligned} t_n^{-1}(\Psi_n^\# - \Psi_n) &\to \alpha \text{ in } D[0,\tau] \\ t_n^{-1}(\Lambda_{11}^{n\#} - \Lambda_{11}^n) &\to \beta \text{ in } D[0,\tau]. \end{aligned}$$

*Then*

$$t_n^{-1}\left(\Theta(\Psi_n^\#, \Lambda_{11}^{n\#}) - \Theta(\Psi_n, \Lambda_{11}^n)\right) \to d\Theta(\Psi,\Lambda_{11})(\alpha,\beta) \text{ in } D[0,\tau], \tag{6.20}$$

*where* $d\Theta(\Psi,\Lambda_{11})(\cdot,\cdot)$ *is a continuous linear functional defined on* $(D[0,\tau], \|\cdot\|_\infty)^2$.

**Proof of theorem 6.1.** For convenience denote $\Lambda_{11}$ with $\Lambda$.

$$\begin{aligned} \mathcal{P}_n^\#(s,t] &\equiv \mathcal{P}((s,t];\Lambda_n^\#) \\ \mathcal{P}_n(s,t] &\equiv \mathcal{P}((s,t];\Lambda_n), \end{aligned}$$

and write $S_n^\# = \Phi_2(\Psi_n^\#, \Lambda_n^\#)$, $S_n = \Phi_2(\Psi_n, \Lambda_n)$ and $S = \Phi_2(\Psi, \Lambda)$. By equation (6.15)

$$S_n^\#(t) = \Psi_n^\#(t) + \int_{s \le t} \Psi_n^\#(s-) \mathcal{P}_n^\#(s,t] d\Lambda_n^\#(s) \tag{6.21}$$

and

$$S_n(t) = \Psi_n(t) + \int_{s \le t} \Psi_n(s-) \mathcal{P}_n(s,t] d\Lambda_n(s) \tag{6.22}$$

so that subtraction yields (by telescoping)

$$\begin{aligned} t_n^{-1}\left(S_n^\#(t) - S_n(t)\right) &= t_n^{-1}\left(\Psi_n^\#(t) - \Psi_n(t)\right) \\ &\quad + \int_{s \le t} t_n^{-1}(\Psi_n^\# - \Psi_n)(s-) \mathcal{P}_n^\#(s,t] d\Lambda_n^\#(s) \\ &\quad + \int_{s \le t} \Psi_n(s-) t_n^{-1} \left(\mathcal{P}_n^\# - \mathcal{P}_n\right)(s,t] d\Lambda_n^\# \\ &\quad + \int_{s \le t} \Psi_n(s-) \mathcal{P}_n(s,t] t_n^{-1}\left(d\Lambda_n^\# - d\Lambda_n\right)(s) \\ &= \mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n + \mathrm{IV}_n. \end{aligned}$$

$\mathrm{I}_n \to \mathrm{I}$ by hypothesis. Our goal is to show that $\mathrm{II}_n, \mathrm{III}_n, \mathrm{IV}_n$ converge to their supposed limits II, III, IV. Firstly, one should notice that the supposed limits are well defined: for example IV$= \int_{s \le t} \Psi(s-) \mathcal{P}(s,t] d\beta(s)$ is defined by repeated integration by parts (lemma 6.13). Here we need that $s \to \mathcal{P}(s,t]$ is of bounded uniform sectional variation, which follows from the bounded uniform sectional variation of $\Lambda_n^\#$ as shown in the proof of proposition 6.4. By telescoping we have:

$$\begin{aligned} \mathrm{II}_n - \mathrm{II} &= \int_{s \le t} \left(\alpha_n^\# - \alpha\right)(s-) \mathcal{P}(s,t) d\Lambda(s) \\ &\quad + \int_{s \le t} \alpha_n^\#(s-) \left(\mathcal{P}_n^\# - \mathcal{P}\right)(s,t] d\Lambda(s) \\ &\quad + \int_{s \le t} \alpha(s-) \mathcal{P}_n^\#(s,t] d(\Lambda_n^\# - \Lambda)(s). \\ &\quad + \int_{s \le t} (\alpha_n^\# - \alpha)(s-) \mathcal{P}_n^\#(s,t] d(\Lambda_n^\# - \Lambda)(s). \end{aligned}$$

Because $\Lambda$ is of bounded variation the first two terms can directly be bounded by a constant times the supremum norm of $(\alpha_n^\# - \alpha)$ and $(\mathcal{P}_n^\# - \mathcal{P})(s,t]$, respectively. $(\alpha_n^\# - \alpha)$ converges to zero by hypothesis and $(\mathcal{P}_n^\# - \mathcal{P})(s,t]$ converges to zero by proposition 6.4. Similarly, using that $\Lambda_n, \Lambda_n^\#$ are of bounded variation uniformly in $n$, we prove that the fourth term converges to zero by

bounding it in the supremum norm of $(\alpha_n^{\#} - \alpha)$. For the third term we write $\mathcal{P}_n^{\#} = \mathcal{P} + (\mathcal{P}_n^{\#} - \mathcal{P})$ and for the non-trivial term with $\mathcal{P}$ we apply the Helly-Bray lemma 6.4 with $H(s) = \alpha(s)\mathcal{P}(s,t]$ and $F(s) = \Lambda_n^{\#} - \Lambda$, because $\alpha$ is of unbounded variation.

The convergence of $\mathrm{III}_n, \mathrm{IV}_n$ to their supposed limits is proved, similarly: only integration by parts and Helly-Bray are needed. This completes the proof. □

**Proofs of propositions.**

**Proof of proposition 6.2** (Kolmogorov equations). For convenience, we define the region which appears in each term of the Peano series: $B_n(s,t] \equiv \{(u^1, \ldots, u^n) \in (\mathbb{R}^2)^n : s < u^1 < \ldots < u^n \leq t\}$. Now,

$$\mathcal{P}_\alpha(s,u) = 1 + \sum_{n=1}^{\infty} \int_{B_n(s,u)} \alpha(du^1) \cdots \alpha(du^n), \tag{6.23}$$

so

$$\begin{aligned} \int_{s<u\leq t} \mathcal{P}_\alpha(s,u)\alpha(du) &= \int_{s<u\leq t} 1\alpha(du) + \sum_{n=1}^{\infty} \int_{B_{n+1}(s,t]} \alpha(du^1)\cdots\alpha(du^n)\alpha(du) \\ &= \sum_{n=1}^{\infty} \int_{B_n(s,t]} \alpha(du^1)\cdots\alpha(du^n) \\ &= \mathcal{P}_\alpha(s,t] - 1. \end{aligned}$$

The backward equation is similarly proved. □

**Proof of proposition 6.3** (Duhamel equation). Consider the following $m+n$-fold integral:

$$\int_{B_{m+n}(s,t]} \alpha(du^1)\cdots\alpha(du^m)\beta(du^{m+1})\cdots\beta(du^{m+n}). \tag{6.24}$$

By splitting the integration on $u^m$ we can write this as:

$$\int_s^t \left\{ \int_{B_{m-1}(s,u^m)} \alpha(du^1)\cdots\alpha(du^{m-1}) \right\} \left\{ \int_{B_n(u^m,t]} \beta(du^{m+1})\cdots\beta(du^{m+n}) \right\} \alpha(du^m).$$

Similarly, splitting the integration on $u^{m+1}$, we can also write it as:

$$\int_s^t \left\{ \int_{B_m(s,u^{m+1})} \alpha(du^1)\cdots\alpha(du^m) \right\} \left\{ \int_{B_n(u^{m+1},t]} \beta(du^{m+2})\cdots\beta(du^{m+n}) \right\} \beta(du^{m+1}).$$

Since these two integrals are equal to each other for all $m$ and $n$, we can sum up the resulting identity on $m$ and $n$ to obtain

$$\int_{s<u\leq t} \mathcal{P}_\alpha(s,u)\left\{\mathcal{P}_\beta(u,t]-1\right\}\alpha(du)=\int_{s<u\leq t}\left\{\mathcal{P}_\alpha(s,u)-1\right\}\mathcal{P}_\beta(u,t]\beta(du) \tag{6.25}$$

Combining (6.25) with the Kolmogorov equations yields the Duhamel equation. □

**Proof of proposition 6.4** (weak continuous compact differentiability of $\mathcal{P}_\alpha$). By the Duhamel equation we have:

$$t_n^{-1}\left(\mathcal{P}_n^{\#}-\mathcal{P}_n\right)(s,t]=\int_{s<u\leq t}\mathcal{P}_n(s,u)\mathcal{P}_n^{\#}(u,t]dh_n^{\#}(u). \tag{6.26}$$

The difference with its supposed limit is given by (telescoping)

$$\int_s^t\left(\mathcal{P}_n-\mathcal{P}\right)(s,u)\mathcal{P}(u,t]dh(u)+\int_s^t\mathcal{P}_n(s,u)\left(\mathcal{P}_n^{\#}-\mathcal{P}\right)(u,t]dh(u)$$
$$+\int_{s<u\leq t}\mathcal{P}_n(s,u)\mathcal{P}_n^{\#}(u,t]d(h_n^{\#}-h)(u).$$

Firstly, notice that all three terms are defined by repeated integration by parts (corollary 6.13), which can be done because $s\to\mathcal{P}_n(s,t]$ (and $\mathcal{P}_n^{\#},\mathcal{P}$) are of bounded uniform sectional variation uniformly in $n$ (see below). The first and second term converge to zero by the Helly-Bray lemma 6.4 and the third term can be bounded by the supremum norm of $h_n^{\#}-h$ by applying integration by parts (lemma 6.1). In all three bounds the uniform sectional variation norm of $\mathcal{P}_n,\mathcal{P}_n^{\#},\mathcal{P}$ considered as functions $s\to\mathcal{P}_n(s,t]$ appear which are uniformly bounded. This is seen as follows. From the definition (6.16) of the Peano series it follows directly that $\|\mathcal{P}_\alpha\|_\infty\leq\exp(\|\alpha\|_\infty)$ (see (6.8). Then by the Kolmogorov equation we have:

$$\|\mathcal{P}_\alpha\|_v^*\leq\|\mathcal{P}_\alpha\|_\infty\|\alpha\|_v^*\leq\exp(\|\alpha\|_\infty)\|\alpha\|_v^*.$$

So if $\|\alpha\|_v^*<M$, then $\|\mathcal{P}_\alpha\|_v^*$ is bounded. This proves the bounded uniform sectional variation property of $\mathcal{P}_n,\mathcal{P}_n^{\#},\mathcal{P}$, by assumption on $\alpha_n$. This completes the proof. □

## 6.3.2 The Dabrowska representation.

### The covariance-mapping.

We will state the differentiability result for the by far most complicated mapping $L$ in the Dabrowska representation.

**Proposition 6.5** *Denote* $\vec{\Lambda} = (\Lambda_{10}, \Lambda_{01}, \Lambda_{11})$. *Let* $\Gamma_3 = \Gamma_{31} - \Gamma_{32}$, *where*

$$\Gamma_{31}(\vec{\Lambda}) = \int_{[0,t]} \frac{\Lambda_{10}(du, v-)\Lambda_{01}(u-, dv)}{\beta(u,v)} \text{ and } \Gamma_{32}(\vec{\Lambda}) = \int_{[0,t]} \frac{\Lambda_{11}(du, dv)}{\beta(u,v)}.$$

*Assume that* $\|(\Lambda_{11}^n, \Lambda_{11}^{n\#})\|_v < M < \infty$ *and*

1. $\beta > \delta > 0$ *on* $[0, \tau]$ *for certain* $\delta > 0$.
2. *There exists a sequence of uniformly in n finite (signed) measures* $\mu_{2n}$ *so that* $\Lambda_{10}^n(u, dv) \ll \mu_{2n}(dv)$ *for all u. Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$.
3. *There exists a sequence of uniformly in n finite (signed) measures* $\mu_{1n}$ *so that* $\Lambda_{10}^n(du, v) \ll \mu_{2n}(du)$ *for all v. Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$.
4. $\|\Lambda_{10}^n(du, v)/\mu_{1n}(du)\|_\infty < M$ *and* $\|\Lambda_{10}^n(u, dv)/\mu_{2n}(dv)\|_\infty < M$ *for certain* $M < \infty$ *(uniform boundedness of the Radon-Nykodym derivatives). Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$

*If* $\vec{h}_n^\# \equiv \sqrt{n}(\vec{\Lambda}_n^\# - \vec{\Lambda}_n) \to \vec{h}$, *then we have:*

$$\sqrt{n}(\Gamma(\vec{\Lambda}_n^\#) - \Gamma(\vec{\Lambda}_n)) - d\Gamma(\vec{\Lambda})(\vec{h}_n^\#) \to 0, \tag{6.27}$$

*for a certain continuous linear map* $d\Gamma(\Lambda) : (D[0,\tau], \|\cdot\|_\infty)^3 \to (D[0,\tau], \|\cdot\|_\infty)$.

**Proof.** We will give the proof of ordinary compact differentiability, i.e. we replace $\Lambda_n$ by $\Lambda$ and $\Lambda_n^\#$ by $\Lambda_n$ in (6.27). The reader can easily verify that the proof goes through when we do not do this. We have by telescoping:

$$\begin{aligned}
&\sqrt{n}\,(\Gamma_{31}(\Lambda_{01}^n, \Lambda_{10}^n) - \Gamma_{31}(\Lambda_{01}, \Lambda_{10})) \\
&= \sqrt{n}\int\int \frac{\beta(u,v)\,(\Lambda_{10}^n(du, v-)\Lambda_{01}^n(u-, dv) - \Lambda_{10}(du, v-)\Lambda_{01}(u-, dv))}{\beta_n(u,v)\beta(u,v)} \\
&\quad + \sqrt{n}\int\int \frac{(\beta - \beta_n)(u,v)\Lambda_{10}(du, v-)\Lambda_{01}(u-, dv)}{\beta_n(u,v)\beta(u,v)} \\
&= \int\int \frac{h_{10}^n(du, v-)\Lambda_{01}(u-, dv) + h_{01}^n(u-, dv)\Lambda_{10}^n(du, v-)}{\beta_n(u,v)} \\
&\quad + \int\int \sqrt{n}\frac{\beta - \beta_n}{\beta_n\beta}(u,v)\Lambda_{10}(du, v-)\Lambda_{01}(u-, dv).
\end{aligned}$$

It is easy to check that $\sqrt{n}(\beta - \beta_n)/(\beta_n\beta) \to H(u,v)$ for a fixed function $H(h_{10}, h_{01})$ linear in $(h_{10}, h_{01})$ which we will not write down. So the last term converges in $\|\cdot\|_\infty$ to $\int\int H(h_{10}, h_{01})(u,v)\Lambda_{10}(du, v-)\Lambda_{01}(u-, dv)$. Notice that the supposed limit $d\Phi(\vec{\Lambda})$ is a continuous linear map because all terms can be defined by integration by parts with lemma 6.1. We only consider the second integral. The first is dealt similarly. The difference between the second integral

and its supposed limit can by telescoping be rewritten as the following sum of terms:

$$\int\int \frac{(h_{01}^n - h_{01})(u-,dv)\Lambda_{10}^n(du,v-)}{\beta_n(u,v)} + \int\int \frac{(\Lambda_{10}^n - \Lambda_{10})(du,v-)h_{01}(u-,dv)}{\beta_n(u,v)}$$
$$+ \int\int \left(\frac{1}{\beta_n} - \frac{1}{\beta}\right)(u,v) + h_{01}(u-,dv)\Lambda_{10}(du,v-).$$

**Term i.** Use corollary 6.1 with $H = h_{01}^n - h_{01}$, $\Lambda = \Lambda_{10}^n$, $\beta = \beta_n$. Then apply integration by parts (the second part of lemma 6.1) and bound this term by the supremum norm of $h_{10}^n - h_{10}$ times integrals like $\int \mid 1/\beta(\Lambda_{10}(du,v-)\Lambda_{01}(u-,dv)) \mid$. For the rest we refer to the techniques in illustration 2 where we show, by using the assumptions 1–4, that this variation is bounded.

**Term ii.** Substitute $h_{01} = (h_{01} - h_{01}^m) + h_{01}^m$. Now bound the term with $(h_{01} - h_{01}^m)$ in the supremum norm of $(h_{01} - h_{01}^m)$ times a constant, and bound the term with $h_{01}^m$ in the supremum norm of $\Lambda_{10}^n - \Lambda_{10}$ times the uniform sectional variation of $h_{01}^m$, both in exactly the same way as we did in term i. Now, let $m \to \infty$ slowly enough (Helly-Bray lemma 6.4).

**Term iii.** Similar to our illustration II with $h(du,dv)$ replaced by $h_{01}(u-,dv)\Lambda_{10}(du,v-)$.

The proof for $\Gamma_{32}$ is similar, but easier.□

**Bivariate product-integral.**

The essential ingredient for establishing differentiability results for the product-integral is the Duhamel equation. For the univariate product integral theory we refer to Gill and Johansen (1990). They also sketch how the proofs can be generalized to the multivariate product-integral. Here, we will present and prove the bivariate analogues of the Kolmogorov equations and Duhamal equation and finally state the differentiability result for the bivariate product-integral. For any signed measure $L$ on $\mathbb{R}^2$ set

$$\prod\!\!\!\!\prod((s,t],L) = \prod_{(s,t]}\!\!\!\!\prod(1 + L(du,dv)), \tag{6.28}$$

where the bivariate product-integral $\prod\!\!\!\!\prod_L(s,t] \equiv \prod\!\!\!\!\prod((s,t],L) = \prod\!\!\!\!\prod_{(s,t]}(1 + L(du,dv))$ is defined as the limit of finite products of $\prod_{i,j=1}^m(1 + L((u_{i-1},v_{j-1}),(u_i,v_j)]))$ over partition-elements $J_{i,j} \equiv ((u_{i-1},v_{j-1}),(u_i,v_j)]$ with $\max_{i,j}\{\mid (u_{i-1},v_{j-1}) - (u_i,v_j) \mid\}$ converging to zero. The ordering (specifying in what way we multiply over the elements of the partition) of this product

is not relevant by the commutativity of multiplication in $\mathbb{R}$, but for our proofs we choose the video ordering (left under to right under then to left under one strip higher etc.). The proof that this product-integral is uniquely defined (that each sequence of partitions of rectangles with mesh converging to zero has the same limit) is exactly the same as the proof for the univariate product-integral as given in Gill and Johansen (1990), page 1515 (see Gill, 1993a).

**Remark.** In one dimension the product integral equals the Peano-series. In two dimensions the same properties (Kolmogorov equations, Duhamel equation) for both can be proved. By using the total ordering in $\mathbb{R}^2$ we can obtain all one dimensional results and we can go back and forth from total ordering to partial ordering.

**Property 6.1** $\prod_{(0,t]}(1 + L(du, dv)) \leq \exp(\|L\|_v)$ *So if $L$ is of bounded variation, then $t \to \prod((0,t], L)$ is bounded in supremum norm.*

This follows immediately from $1+ \mid L(J_{i,j}) \mid \leq \exp(\mid L(J_{i,j}) \mid)$. We will see that we can easily get generalizations of the Kolmogorov and Duhamel equation of the univariate case by replacing univariate intervals by rectangles with respect to the total (video) ordering. That is indeed what we will do. Then we will show that we can rewrite the obtained results in terms of rectangles with respect to the usual partial ordering.

**Lemma 6.5** *Write $(0,t] = \{x \in \mathbb{R}^2 : 0 < x \leq t\}$ for an interval with respect to the partial ordering on $\mathbb{R}^2$. Denote $]]0,t]]$ for an interval with respect to the total (video) ordering on $\mathbb{R}^2$: $(x,y) \in ]]0,t]] \Leftrightarrow 0 < y < t_2$ or $y = t_2, x \leq t_1$. Then*

$$\begin{aligned}(0,t] \cap ]]0,s]] &= ((0,t_1] \times (0,s_2)) \cup ((0,s_1] \times \{s_2\}) \\ (0,t] \cap ]]s,\infty]] &= ((s_1,t_1] \times \{s_2\}) \cup ((0,t_1] \times (s_2,t_2])\,.\end{aligned}$$

The lemma says that we can describe these intersections as the union of one two dimensional rectangle and a one-dimensional line segment, both with respect to the partial ordering. The proof is trivial.

For the next proposition and lemma it should be remarked that the Kolmogorov and Duhamel equations are certainly not true if the rectangles w.r.t. the total ordering are replaced by a partial ordering; by the total ordering, if we walk in video ordering from left under to right above the region grows monotonically to the total rectangle and that makes the identities essentially the same as the univariate identities.

**Proposition 6.6** (Kolmogorov equations). *Denote* $(0,t]$ *for an interval with respect to the partial ordering on* $\mathbb{R}^2$ *and denote* $]]0,t]]$ *for an interval with respect to the total ordering on* $\mathbb{R}^2$. *The bivariate product-integral* $\prod\!\!\!\!\!\!\int \equiv \prod\!\!\!\!\!\!\int_L$ *satisfies:*

$$\begin{aligned}\prod_L\!\!\!\!\!\!\int (s,t] &= 1+\int_{(s,t]}\prod_L\!\!\!\!\!\!\int ((s,t]\cap]]s,u[[)\,L(du)\\ &= 1+\int_{(s,t]}\prod_L\!\!\!\!\!\!\int ((s,t]\cap]]u,t[[)\,L(du).\end{aligned}$$

**Proof.** We prove the first equality. Consider a finite partition $\pi_{h_m}$ of $(s,t]$ of rectangles with diameter smaller than $h_m$. Replace the product-integrals by a finite product in video ordering over this partition. Then the integral is an integral of a simple function with respect to the measure $L$. Because of the identity $\prod_{i=1}^m(1+a_i) = 1+\sum_{i=1}^m\prod_{j=1}^{i-1}(1+a_j)a_i$ it follows that the equality holds for this finite partition. By the convergence of this product to the product integral for $h_m \to 0$ (see definition of product integral) the left-hand side $\prod\!\!\!\!\!\!\int_L^m(s,t]$ and the integrand on the right-hand side $\prod\!\!\!\!\!\!\int_L^m((s,t]\cap]]s,u[[)$ converge to $\prod\!\!\!\!\!\!\int_L(s,t]$ and $\prod\!\!\!\!\!\!\int_L((s,t]\cap]]s,u[[)$, respectively. The dominated convergence theorem tells us that the right hand side converges for this sequence of partitions to $1+\int_{(s,t]}\prod\!\!\!\!\!\!\int_L((s,t]\cap]]s,u[[)\,L(du)$. □

**Corollary 6.2** *If* $L$ *is of bounded variation, then* $t \to \prod\!\!\!\!\!\!\int_L(0,t]$ *is of bounded variation. Similarly, for bounded uniform sectional variation.*

**Proof.** This follows straightforwardly from property 6.1 and the Kolmogorov equations. For the precise argument see the proof of proposition 6.4 □

**Lemma 6.6** (Duhamel equation with total ordering). *We have:*

$$\prod_\alpha\!\!\!\!\!\!\int (0,t] - \prod_\beta\!\!\!\!\!\!\int (0,t] = \int_{(0,t]}\prod_\alpha\!\!\!\!\!\!\int ((0,t]\cap]]0,s[[)\;d(\alpha-\beta)(s)\;\prod_\beta\!\!\!\!\!\!\int ((0,t]\cap]]s,\infty[[).$$

**Proof.** The proof is the same as the proof for the Kolmogorov equations except that we now have to use the telescoping-identity $\prod_{i=1}^n a_i - \prod_{i=1}^n b_i = \sum_{i=1}^n\prod_{j=1}^{i-1}a_j(a_i-b_i)\prod_{j=i+1}^n b_j$. □

Now, with lemmas 6.5 and 6.6 we are able to write down a Duhamel equation which involves product-integrals over rectangles or lines with lower and upper corner chosen out of the corners of $(s,t]$. We can simplify this to only product-integrals over rectangles and lines with lower corner at $(0,0)$ as follows.

**Lemma 6.7**

$$\prod_{\alpha}(s,t] = \left(\prod_{\alpha}(0,t]\prod_{\alpha}(0,s]\right) / \left(\prod_{\alpha}(0,(s_1,t_2)]\prod_{\alpha}(0,(t_1,s_2)]\right),$$

*which is the generalized ratio of the product-integral over a rectangle with lower corner at* $(0,0)$ *and upper corner at one of the four corners of* $(s,t]$.

**Proof.** The proof follows straightforwardly from the multiplicativity of the product-integral. □

**Proposition 6.7** (Duhamel equation). *Define*

$$V(s,t) \equiv \prod_{\alpha}((0,t_1]\times(0,s_2))\prod_{\alpha}((0,s_1]\times\{s_2\})$$
$$\times\prod_{\beta}((s_1,t_1]\times\{s_2\})\prod_{\beta}((0,t_1]\times(s_2,t_2]),$$

*where* $\prod_{\beta}((s_1,t_1]\times\{s_2\})$ *and* $\prod_{\beta}((0,t_1]\times(s_2,t_2])$ *can be written as a generalized ratio of product-integrals over rectangles with lower corners at* $(0,0)$ *and upper corners with coordinates taken from* $s$ *and* $t$ *(see lemma 6.7). Then*

$$\prod_{\alpha}(0,t] - \prod_{\beta}(0,t] = \int_{(0,t]} V(s,t)d(\alpha-\beta)(s). \tag{6.29}$$

All these product-integrals are of bounded (uniformly in $t$) variation in $s$ by application of the property 6.1. So by our repeated integration by parts formula (6.13) we can do integration by parts so that $\alpha-\beta$ appears as a function and thereby bound it in the supremum norm of $\alpha-\beta$ and the uniform sectional variation norm of $V$.

**Proof.** (Duhamel equation). Firstly, apply lemma 6.6. Then by lemma 6.5 and the multiplicativity of $\prod_{\alpha}$ we can write the product-integrals as a product over product-integrals over rectangles and hyperplanes with respect to the partial ordering. Finally apply lemma 6.7. □

**Theorem 6.2** (Weak continuous compact differentiability of the bivariate product integral). *The bivariate product-integral* $\prod : (D[0,\tau],\|\cdot\|_\infty) \to (D[0,\tau],\|\cdot\|_\infty)$*:*

$$L \mapsto \prod_{[0,t]}(1-L(du,dv))$$

*satisfies differentiability property (6.27) for sequences* $\|L_n\|_v^* < C, \|L_n^\#\|_v^* < C$ *converging to a signed measure* $L$.

It is already known that it holds for the univariate product-integral (Gill and Johansen, 1990).

**Proof.** For this we refer to the differentiability proof of the bivariate Peano series in the preceding section: the same ingredients (Kolmogorov equations, Duhamel, repeated integration by parts) have to be used in the same way.□

### 6.3.3 The Prentice-Cai representation.

Recall the Prentice-Cai representation

$$\begin{aligned} S(t) &= \Theta_1\left(\Lambda_{10}(\cdot,0), \Lambda_{01}(0,\cdot), R\right) \\ &= \Theta_1\left(\Lambda_{10}(\cdot,0), \Lambda_{01}(0,\cdot), \Theta_2(\widetilde{L})\right) \\ &= \Theta_1\left(\Lambda_{10}(\cdot,0), \Lambda_{01}(0,\cdot), \Theta_2(\Theta_3(\vec{\Lambda})\right) \\ &= \Theta(\vec{\Lambda}), \end{aligned}$$

where $\Theta_1$ is a product of two univariate product integrals w.r.t. $\Lambda_{10}(\cdot,0)$ and $\Lambda_{01}(0,\cdot)$, respectively, times $R$; $\Theta_2 = \Phi_2$ is the Volterra representation; $\Theta_3$ is the $\widetilde{L}$ mapping which has the same structure (slightly easier) as the $\Gamma_3 = L$ mapping of Dabrowska's representation. So the weak continuous differentiability has been proved for $\Theta_1$ in Gill and Johansen (1990), for $\Theta_2$ in theorem 6.1, for $\Theta_3$ by copying the proof of proposition 6.5. The chain rule provides us now with the weak continuous differentiability of $\Theta$.

### 6.3.4 Differentiability theorem for the three representations.

**Theorem 6.3** *All three representations are defined in section 1. Let* $\Gamma$ *be the Dabrowska representation and* $\vec{\Lambda}$ *the vector of hazard measures corresponding with* $S$ *as defined in section 1:* $S = \Gamma(\vec{\Lambda})$.

**Dabrowska representation.**

**Assumptions.** *Assume that* $\|\Lambda_{11}^n\|_v < M < \infty$, $\|\Lambda_{11n}^{\#}\|_v < M < \infty$ *and*

**1.** $S(\tau) > 0$.

**2.** *There exists a sequence of uniformly in n finite (signed) measures* $\mu_{2n}$ *so that* $\Lambda_{10}^n(u,dv) \ll \mu_{2n}(dv)$ *for all u. Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$.

**3.** *There exists a sequence of uniformly in n finite (signed) measures* $\mu_{1n}$ *so that* $\Lambda_{10}^n(du,v) \ll \mu_{2n}(du)$ *for all v. Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$.

**4.** $\|\Lambda_{10}^n(du,v)/\mu_{1n(du)}\|_\infty < M$ *and* $\|\Lambda_{10}^n(u,dv)/\mu_{2n}(dv)\|_\infty < M$ *for some* $M < \infty$ *(uniform boundedness of the Radon-Nykodym derivatives). Similarly for* $\Lambda_{10}, \Lambda_{10}^{n\#}, \Lambda_{01}, \Lambda_{01}^n, \Lambda_{01}^{n\#}$.

*If* $Z_{\vec{\Lambda}}^{n\#} \equiv \sqrt{n}(\vec{\Lambda}_n^{\#} - \vec{\Lambda}_n) \to Z_{\vec{\Lambda}}$, *then*

$$\sqrt{n}(\Gamma(\vec{\Lambda}_n^{\#}) - \Gamma(\vec{\Lambda}_n)) - d\Gamma(\vec{\Lambda})(Z_{\vec{\Lambda}}^{n\#}) \to 0,$$

*for a continuous linear map* $d\Gamma(\Lambda) : (D[0,\tau], \|\cdot\|_\infty)^3 \to (D[0,\tau], \|\cdot\|_\infty)$.

**Prentice-Cai representation.** *The same statement holds for the Prentice-Cai representation* $S = \Theta(\vec{\Lambda})$.

**Volterra representation.** *The same differentiability result holds for* $S = \Phi(\vec{\Lambda})$ *with the assumptions 2,3 and 4 replaced by:* $\|(\Lambda_{10}^n, \Lambda_{01}^n)\|_v < M < \infty$ *and* $\|(\Lambda_{10}^{n\#}, \Lambda_{01}^{n\#})\|_v < M < \infty$.

**Proof.** This differentiability property has been proved for the univariate product integral in Gill and Johansen (1990) (so this gives it for $\Gamma_1$, $\Theta_1$, $\Phi_2$), for the bivariate product integral in theorem 6.2 (so this gives it for $\Gamma_2$), for the bivariate Volterra representation (bivariate Peano series) in theorem 6.1 (so this gives it for $\Theta_2, \Phi_1$), for the $L$ mapping in proposition 6.5 (here we need the denominator assumptions) (so this gives it for $\Gamma_3, \Theta_3$), where one has to notice that assumption 1 tells us that $\beta > 0$ (denominator in $L$ and $\widetilde{L}$). Now, the theorem follows from the chain rule. □

## 6.4 The estimators.

Let $\Phi, \Gamma$ and $\Theta$ denote the Volterra, Dabrowska and Prentice-Cai representation, respectively, which were defined and studied in sections 2 and 3. Now, we will construct the estimators which are based on these representations. From now everything indexed by $n$ is random.

**Estimators for the hazards.** First define the following subdistributions of the data corresponding with the four kinds of censoring which can occur.

$$P_{ij}(t) \equiv P\left(\widetilde{T}_1 \le t_1, \widetilde{T}_2 \le t_2, D_1 = i, D_2 = j\right) \text{ for } i,j \in \{0,1\}, \quad t \in \mathbb{R}^2_{\ge 0}$$

and

$$P(t) \equiv P\left(\widetilde{T}_1 \le t_1, \widetilde{T}_2 \le t_2\right) = \sum_{i,j} P_{ij}(t).$$

Then, on $[0,\tau]$ with $\overline{P}(\tau-) = SH(\tau-) > 0$,

$$\Lambda_{11}(t) = \int_{[0,t]} \frac{H(s-)}{S(s-)H(s-)} dF(s) = \int_{[0,t]} \frac{1}{\overline{P}(s-)} dP_{11}(s) \tag{6.30}$$

$$\Lambda_{10}(t) = \int_{[0,t_1]} \frac{H(u-,t_2)}{S(u-,t_2)H(u-,t_2)} F(du,t_2) \tag{6.31}$$

$$
\begin{aligned}
&= \int_{[0,t_1]} \frac{1}{\overline{P}(u-,t_2)}(P_{11}+P_{10})(du,t_2) \\
\Lambda_{01}(t) &= \int_{[0,t_2]} \frac{H(t_1,v-)}{S(t_1,v-)H(t_1,v-)}F(t_1,dv) \qquad (6.32) \\
&= \int_{[0,t_2]} \frac{1}{\overline{P}(t_1,v-)}(P_{11}+P_{01})(t_1,dv).
\end{aligned}
$$

If we define $\vec{P} = (P_{10}, P_{01}, P_{00}, P_{11})$, then $\vec{\Lambda} = \vec{\Lambda}(\vec{P})$. Let

$$P_{nij} \equiv \frac{1}{n}\sum_{k=1}^{n} I(\widetilde{T}_{1k} \leq t_1, \widetilde{T}_{2k} \leq t_2, D_{1k} = i, D_{2k} = j)$$

be the empirical distribution of $P_{ij}$ for $i,j \in \{0,1\}$ and $\vec{P}_n = (P_{10}^n, P_{01}^n, P_{00}^n, P_{11}^n)$. We estimate $\vec{\Lambda}$ with the Nelson-Aalen estimator

$$\vec{\Lambda}_n = \vec{\Lambda}(\vec{P}_n) \text{ for } t \in W_n^+ \equiv \{t : \overline{P}(t) > 0\}.$$

In other words $\vec{\Lambda}_n$ is given by the formulas above with $P_{ij}$ replaced by $P_{ij}^n$.

**The Dabrowska estimator.** Recall the representation $S = \Gamma(\vec{\Lambda})$. We have

$$S_n^D(t) = \Gamma(\vec{\Lambda}_n)(t) \text{ for } t \in W_n^+, \qquad (6.33)$$

which equals the product $S_{1n}S_{2n}\prod(1 - L(\vec{\Lambda}_n))$ where $S_{1n}$, $S_{2n}$ are the univariate Kaplan-Meier estimators of the marginals $S_1, S_2$, respectively.

**The Volterra estimator.** Recall the representation $S = \Phi(\vec{\Lambda})$. We have

$$S_n^V = \Phi(\vec{\Lambda}_n) \text{ for } t \in W_n^+, \text{ where } \Phi_2(\vec{\Lambda}_n) = 1 - F_{1n} - F_{2n}. \qquad (6.34)$$

**The Prentice-Cai estimator.** Recall the representation $S = \Theta(\vec{\Lambda})$. So

$$S_n^{PC} = \Theta(\vec{\Lambda}_n) \text{ for } t \in W_n^+, \qquad (6.35)$$

which is equal to the product $F_{1n}F_{2n}\Theta_2(\Theta_3(\vec{\Lambda_n}))$.

## 6.5 Asymptotic properties of the estimators.

We will use the results of section 3 to establish a functional central limit theorem for the estimators defined in section 4. As outlined in section 1, we do this by applying the functional delta-method theorem 1.6 to the representations $\Phi \circ \vec{\Lambda}$, ,$\Gamma \circ \vec{\Lambda}$ and $\Theta \circ \vec{\Lambda}$ as functionals in $\vec{P}$. Since weak continuous Hadamard differentiability of $\Phi$, $\Gamma$ and $\Theta$ has been established in section 3, by the chain

rule, the remaining differentiability result for the delta-method which we need to verify is the weak continuous Hadamard differentiability for $\vec{P} \to \vec{\Lambda}(P)$. The weak convergence hypothesis of the delta-method requires that the bootstrap works for the empirical process $\vec{P}_n$. $\vec{P}_n$ is the usual empirical process indexed by the indicators $I_{[0,t]}$ and therefore its bootstrap result is well known; let $\vec{P}_n^{\#}$ be the bootstrapped empirical process obtained by resampling from the empirical $\vec{P}_n$, then $Z_n \equiv \sqrt{n}(\vec{P}_n - \vec{P}) \stackrel{D}{\Longrightarrow} Z$, where $Z$ is a Gaussian process with the same covariance structure as the left-hand side, and $Z_n^{\#} \equiv \sqrt{n}(\vec{P}_n^{\#} - \vec{P}_n) \stackrel{D}{\Longrightarrow} Z$ given $\vec{P}_n$.

The following lemma provides us easily with the weak continuous differentiability of the representation $\vec{\Lambda}$.

**Lemma 6.8** *The functional*

$$A : (F, G) \mapsto \int F(s) dG(s)$$

*satisfies the differentiability property (6.27) at any point $(F, G)$ where $F$ and $G$ are of bounded uniform sectional variation for sequences $(F_n, G_n), (F_n^{\#}, G_n^{\#})$ of bounded uniform sectional variation uniformly in $n$.*

The proof is a copy of illustration I and this mapping is also contained in the mapping $L$ and $\tilde{L}$: the integration by parts lemma 6.1 and the Helly-Bray lemma 6.4 are the only ingredients we need.

Recall the representation $\vec{\Lambda}$: it is a composition of $Y \to 1/Y$ and $A$. So the weak continuous differentiability of $\vec{\Lambda}$ follows directly by the weak continuous differentiability of $Y \to 1/Y$ at a $Y > \delta > 0$ on $[0, \tau]$ for some $\delta > 0$ and application of lemma 6.8 and the chain rule, using the fact that, by lemma 1.5, the uniform sectional variation of $1/Y$ is bounded by the uniform sectional variation norm of $Y$. Here $SH$ plays the role of $Y$. So we need that $S(\tau)H(\tau) > 0$.

### 6.5.1 Final results.

**Theorem 6.4** (Functional central limit theorems for the estimators $S_n^D, S_n^{PC}$ and $S_n^V$). *Suppose that*

$$S(\tau)H(\tau) > 0.$$

*Recall the definitions of $Z \in D[0, \tau]^4$ and the representations $\vec{P} \to \vec{\Lambda}(\vec{P})$, $\vec{\Lambda} \to \Gamma(\vec{\Lambda})$, $\vec{\Lambda} \to \Phi(\vec{\Lambda})$, $\vec{\Lambda} \to \Theta(\vec{\Lambda})$ as given in section 1. We denote the derivatives*

*with* $d\Gamma, d\Phi$ *and* $d\Theta$.

**The Dabrowska estimator.**

$$S_n^D \to S \text{ a.s.}$$

*and*

$$\sqrt{n}(S_n^D - S) \xRightarrow{D} \left(d\Gamma(\vec{\Lambda}) \circ d\vec{\Lambda}(F)\right)(Z) \text{ in } (D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$$

*for a continuous linear map* $d\Gamma(\vec{\Lambda}) \circ d\vec{\Lambda}(F) : (D[0,\tau], \|\cdot\|_\infty)^4 \to (D[0,\tau], \|\cdot\|_\infty)$. *Moreover,*

$$\sqrt{n}(S_n^{\#D} - S_n^D) \xRightarrow{D} d\Gamma(\vec{\Lambda})(Z) \text{ a.s. in } (D[0,\tau], \|\cdot\|_\infty).$$

*So this estimator is consistent, its normalized version converges weakly to a Gaussian process and the bootstrap is asymptotically valid.*

**The Prentice-Cai estimator.** *The same statement holds for* $S_n^{PC}$ *with* $\Gamma$ *replaced by* $\Theta$ *everywhere.*

**The Volterra estimator.** *The same statement holds for* $S_n^V$ *with* $\Gamma$ *replaced by* $\Phi$ *everywhere.*

**Proof.** We have to verify the conditions of the functional delta-method theorem 1.6 and apply it to $\Gamma \circ \vec{\Lambda}$, $\Theta \circ \vec{\Lambda}$ and $\Phi \circ \vec{\Lambda}$ all three considered as functionals in $\vec{P}$. The weak convergence of $\sqrt{n}(\vec{P}_n^{\#} - \vec{P}_n)$ (a.s.) and of $\sqrt{n}(\vec{P}_n - \vec{P})$ has already been established above. Because we already verified the differentiability condition for $\vec{\Lambda}$ it remains (by the chain-rule) to verify the conditions of theorem 6.3. Assumption 1 in theorem 6.3 is $S(\tau) > 0$. For the other assumptions it suffices to show that $\vec{\Lambda}$, $\vec{\Lambda}_n$ and $\vec{\Lambda}_n^{\#}$ satisfy the assumptions 2–4 stated in theorem 6.3 (the bounded Radon-Nykodym derivatives assumptions). Here, one has to notice that assumption 2–4 for $\Lambda_{11}^n$, $\Lambda_{11}^{n\#}$ are stronger than the requirement of theorem 6.3 that these functions are of bounded variation uniformly in $n$.

**Verification of assumptions 2–4 of theorem 6.3.** We will prove these conditions for $\Lambda_{10}(t) = -\int S(du, t_2)/S(u-, t_2)$. It will be clear that the proof for $\Lambda_{01}$ and $\Lambda_{11}$ is similar. We have

$$\Lambda_{10}(du, v) = -\frac{S(du, t_2)}{S(u-, t_2)} \leq \frac{1}{S(\tau)} S(du, 0).$$

Therefore we have $\Lambda_{10}(du, v) \ll S(du, 0)$ and $\Lambda_{10}(du, v)/S(du, 0) \leq 1/S(\tau)$ (i.e. Radon-Nykodym derivative is bounded). Furthermore we have:

$$\begin{aligned} \Lambda_{10}(u, dv) &= -\int_{(0,u]} \frac{S(ds, dv)}{S(s-, v)} + \int_{(0,u]} \frac{S(ds, v)}{S(s-, v)^2} S(s-, dv) \\ &\leq \frac{1}{S(\tau)} S(0, dv) + \frac{1}{S(\tau)^2} S(0, dv). \end{aligned}$$

Therefore we also have $\Lambda_{10}(du, v) \ll S(0, dv)$ and $\Lambda_{10}(u, dv)/S(du, 0) \leq 1/S(\tau) + 1/S(\tau)^2$. This proves conditions 2–4 for $\Lambda_{10}$ by setting $\mu_1 = S_1$ and $\mu_2 = S_2$ (the marginals of $S$). The formulas (6.30) tell us that $\Lambda_{10}^n(t) = -\int P_n'(du, t_2)/\overline{P}_n(u-, t_2)$ for $P_n' = P_{11}^n + P_{10}^n$ and $\overline{P}_n = \sum_{i,j} \overline{P}_{i,j}^n$, where the latter converges a.s. to $SH > \delta > 0$. So by copying the proof of assumptions 2–4, above, we obtain bounds $1/\overline{P}_n(\tau)$ and $1/\overline{P}_n(\tau) + 1/\overline{P}_n(\tau)^2$ for the Radon-Nykodym derivatives. By the almost sure convergence of $\overline{P}_n$, these bounds are bounded uniformly in $n$. Similarly, for $\Lambda_{10}^{n\#}(t)$. Therefore the same proof works for all hazard measures. This completes the verification of the assumption 2–4.

We can now apply theorem 6.3 and thereby we can apply the functional delta-method theorem 1.6. This proves the weak convergence and bootstrap results of the theorem.

The consistency follows from the continuity of the representations $\Gamma \circ \vec{\Lambda}, \Theta \circ \vec{\Lambda}, \Phi \circ \vec{\Lambda}$ in $\vec{P}$ and the almost sure consistency of $\vec{P}_n$ to $\vec{P}$ in supremum norm (Glivenko-Cantelli). □

So far we did not write down the influence curves (derivatives) $d\Gamma(\vec{\Lambda}) \circ d\vec{\Lambda}(\vec{P})(Z)$, $d\Theta(V) \circ d\vec{\Lambda}(\vec{P})(Z)$ and $d\Phi(\vec{\Lambda}) \circ d\vec{\Lambda}(\vec{P})(Z)$ of the estimators because these formulas are large and not necessary for this work. The variance of these influences curves equal the variance of the limiting distributions of the estimators. Therefore, the influence curves become useful if one wants to estimate the variance of the limiting distribution or in any other efficiency analysis. Below we will write down the proof of efficiency of the Dabrowska and Prentice-Cai estimator in case of independence, and thereby also give an illustration of how an influence curve can be fairly easily obtained.

## 6.6 Influence curves.

If an estimator is a compactly differentiable function of the empirical distribution of an i.i.d. sample $X_1, \ldots, X_n \sim P$, then it is asymptotically linear by application of the functional delta-method theorem 1.5; one can write

$$\Theta_n = \Theta + \frac{1}{n} \sum_{i=1}^{n} I(P, \Theta)(X_i) + o_P(n^{-\frac{1}{2}}),$$

where $I(P, \Theta)(X_i)$, called the influence curve at the point $X_i$, is the derivative of the function in question applied to the centred empirical process, at sample size 1, based on the single observation $X_i$. This follows from linearity of the derivative and the fact that an empirical distribution function is a sample average. One has $E_\Theta(I(P, \Theta)(X_i)) = 0$, while $\mathrm{Var}(I(P, \Theta)(X_i))$ is the asymptotic

variance of $\sqrt{n}(\Theta_n - \Theta)$. It is not surprising that the influence curve plays an important role in efficiency and robustness studies.

We discuss here computation of the influence curves of our three estimators $S_n^D(t)$, $S_n^V(t)$, $S_n^{PC}(t)$, for given $t$, as function of a bivariate censored observation $(\widetilde{T}_1, \widetilde{T}_2, D_1, D_2)$. The form of the influence curve also depends on the point at which we make the calculations, i.e. on the assumed 'true' values of $F$ and $G$.

In principle, using the chain rule, one can write down formulas by applying the derivative of each composing mapping in turn. The resulting formulas are very large and not very illuminating. The procedure can be speeded up by noting the following algorithm for computing the derivative of our mappings, applied to any function: consider integrals and product-integrals as ordinary sums and products, consider differentials $dF, dh$ etc. as ordinary variables indexed by (e.g.) $t$; apply the usual rules of algebra, and then convert back to a proper mathematical expression by replacing sums and products involving differentials by the 'obvious' integrals or product integrals. This also applies to the Peano series since it is an infinite sum of multiple integrals.

The above statement is trivially true if the distributions involved are discrete. By approximating the continuous distributions by discrete distributions and using that the algorithm is correct for discrete distributions, the result for continuous distributions follows straightforwardly from appropriate continuity of the compact derivative in the sense that $I(P_n, \Theta_n)(X) \to I(P, \Theta)(X)$, $\Theta_n = \Theta(P_n)$, for sequences $P_n \to P$. So the idea which makes this algorithm work is that by appropriate continuity of the derivative one can determine the derivative at a general point from the derivative at a discrete approximation and the derivative at a discrete approximation is obtained by applying the usual rules of algebra (i.e. the algorithm is then trivially correct). This is proved for the Dabrowska representation in van der Laan (1990).

We will compute the influence curve by direct formal algebraic manipulation of the representations of the estimators. We will use the chain rule in the sense that we will decompose the calculation in two steps: from the empirical distributions to the empirical hazards, and from the empirical hazards to the survival functions.

Also we will only compute the influence curve at a special point: namely $F$ is continuous, $F = F_1F_2$, and $G = G_1G_2$. We call this 'complete independence' (of all survival and all censoring variables), and continuity of survival. The simplification caused by independence of the survival variables is obvious. Continuity of survival means that all unpleasant terms like $1/(1 - \Delta\Lambda)$, both

arising as derivatives and as part of the representations themselves (the $\beta$ function in the Dabrowska and Prentice-Cai representations) disappear completely. That the terms arising from the derivatives of $\beta$ disappear, is a more subtle point (this is shown by using the $d-\Delta$-interchange lemma and that by continuity the underlying hazards have no jumps), but fortunately true. Finally, independence of censoring makes the probabilistic structure of the influence curves easier still and also allows optimality calculations (computation of the efficient influence curve) to be done explicitly.

The finding will be: at complete independence the Dabrowska and the Prentice-Cai estimators are efficient. We prove this 'at continuity' and conjecture it is also true without this restriction. The Volterra estimator is not efficient at this point. We will not write down the influence curve of Volterra's estimator, but refer to Gill, van der Laan and Wellner (1993). The result means that the Dabrowska and Prentice-Cai are very similar under complete independence and close to efficient under weak dependence, while the Volterra estimator is much inferior. This finding has been supported by extensive simulations (Bakker, 1990, Prentice-Cai, 1992a, Pruitt, 1992, and chapter 8).

### 6.6.1 Computation of the influence curves.

We do not go through the computation in detail but just make the remark that each step is made rigorous by application of our differentiability results for all mappings which occur. Since we are going to suppress $T_1, T_2$ etc. a different notation is more convenient. We replace $n$ by $\widehat{\ }$ and use $1, 2$ to indicate functions only depending on the first or second variable. In particular we use:

$$\Lambda,\ \Lambda_1,\ \Lambda_2,\ \Lambda_{1|2} \text{ and } \Lambda_{2|1}$$

instead of

$$\Lambda_{11}(\cdot,\cdot),\ \Lambda_{10}(\cdot,0),\ \Lambda_{01}(0,\cdot),\ \Lambda_{10}(\cdot,\cdot) \text{ and } \Lambda_{01}(\cdot,\cdot).$$

The influence curves for $\widehat{\Lambda}, \widehat{\Lambda}_1$ etc. are very simple and are given by:

$$\begin{aligned} d\widehat{\Lambda} - d\Lambda &\approx \frac{dM}{y} \\ d\widehat{\Lambda}_i - d\Lambda_i &\approx \frac{dM_i}{y_i}\ i = 1 \text{ or } 2 \\ d\widehat{\Lambda}_{i|j} - d\Lambda_{i|j} &\approx \frac{dM_{i|j}}{y}\ i, j = 1,2 \text{ or } 2,1. \end{aligned}$$

Here for one bivariate censored observation $(\widetilde{T}_1, \widetilde{T}_2, D_1, D_2)$,

$$\begin{aligned}
M(s,t) &= I\left(\widetilde{T}_1 \le s, \widetilde{T}_2 \le t, D_1 = 1, D_2 = 1\right) \\
&\quad - \int_0^s \int_0^t I\left(\widetilde{T}_1 \ge u, \widetilde{T}_2 \ge v\right) \Lambda(du, dv) \\
M_1(s) &= I\left(\widetilde{T}_1 \le s, D_1 = 1\right) - \int_0^s I\left(\widetilde{T}_1 \ge u\right) \Lambda_1(du) \\
M_2(t) &= I\left(\widetilde{T}_2 \le t, D_2 = 1\right) - \int_0^t I\left(\widetilde{T}_2 \ge v\right) \Lambda_2(dv) \\
M_{1|2}(s,t) &= I\left(\widetilde{T}_1 \le s, \widetilde{T}_2 \ge t, D_1 = 1\right) - \int_0^s I\left(\widetilde{T}_1 \ge u, \widetilde{T}_2 \ge t\right) \Lambda_{1|2}(ds, t) \\
y(s,t) &= P(\widetilde{T}_1 \ge s, \widetilde{T}_2 \ge t) \\
y_1(s) &= P(\widetilde{T}_1 \ge s).
\end{aligned}$$

Using $\prod(1+dL)$ for the product integral of $L$ and $\mathcal{P}_{[0,t]}(L)$ for the Peano series $\mathcal{P}([0,t] : L)$ of $L$, we note that

$$\begin{aligned}
\prod(1 + d\widehat{L}) - \prod(1 + dL) &\approx \prod(1 + dL) \int \frac{d\widehat{L} - dL}{(1 + \Delta L)} \\
\mathcal{P}_{[0,t]}(\widehat{L}) - \mathcal{P}_{[0,t]}(L) &\approx \int \mathcal{P}_{[0,s)}(L)(\widehat{L} - L)(ds)\mathcal{P}_{(s,t]}(L). \quad (6.36)
\end{aligned}$$

The two representations for Dabrowska and Prentice-Cai are:

$$\prod_{[0,t_1]}(1 - d\Lambda_1) \prod_{[0,t_2]}(1 - d\Lambda_2) \prod_{[0,t]}\left(1 + \frac{d\Lambda - d\Lambda_{1|2} d\Lambda_{2|1}}{(1 - \Delta\Lambda_{1|2})(1 - \Delta\Lambda_{2|1})}\right)$$
$$\mathcal{P}_{[0,t_1]}(\Lambda_1)\mathcal{P}_{[0,t_2]}(\Lambda_2)\mathcal{P}_{[0,t]}\left(\int\int \frac{d\Lambda - d\Lambda_{1|2} d\Lambda_2 - d\Lambda_{2|1} d\Lambda_1 + d\Lambda_1 d\Lambda_2}{(1 - \Delta\Lambda_1)(1 - \Delta\Lambda_2)}\right).$$

This gives us then, by inspection (just notice that the denominator of $L$ and $\widetilde{L}$ do not contribute to the influence curve by the $d - \Delta$ interchange lemma and noting that $f(\Delta s) = 0$ if $f$ is continuous) the following influence curve for Dabrowska

$$S\left\{-\int \frac{dM_1}{y_1} - \int \frac{dM_2}{y_2} + \int\int \left(\frac{dM - dM_{1|2} d\Lambda_{2|1} - dM_{2|1} d\Lambda_{1|2}}{y}\right)\right\}$$

and for Prentice-Cai we have

$$S\left\{-\int \frac{dM_1}{y_1} - \int \frac{dM_2}{y_2}\right\} + S_1 S_2 \left\{\int \mathcal{P}_{[0,s)}(L)(\widehat{L} - \widetilde{L})(ds)\mathcal{P}_{(s,t]}(L)\right\}$$

where

$$(d\widehat{L} - d\widetilde{L})$$
$$\approx \frac{dM}{y} - d\Lambda_{1|2}\frac{dM_2}{y_2} - dM_{1|2}\frac{d\Lambda_2}{y} - d\Lambda_{2|1}\frac{dM_1}{y_1} - dM_{2|1}\frac{d\Lambda_1}{y}$$
$$+\tfrac{dM_1}{y_1}d\Lambda_2 + \tfrac{dM_2}{y_2}d\Lambda_1.$$

Next, simplification arises on assuming independence in $F$ and $G$. Then $\Lambda_{1|2} = \Lambda_1$, $\Lambda = \Lambda_1\Lambda_2$, $y = y_1y_2$, $\mathcal{P}(\widetilde{L}) = 1$ ($\widetilde{L} = 0$) and $L = 0$.

$$S\left\{-\int\frac{dM_1}{y_1} - \int\frac{dM_2}{y_2} + \int\int\left(\frac{dM - dM_{1|2}d\Lambda_2 - dM_{2|1}d\Lambda_1}{y_1y_2}\right)\right\}$$

and notice that by cancellation of terms $PC$ simplifies to exactly the same influence curve as Dabrowska's! Now, let $dN_1$, $dN_2$, $Y_1, Y_2$ be defined by $dM = dN - Yd\Lambda$, $dM_{1|2} = dN_1Y_2 - Y_1Y_2d\Lambda_1$ etc. Then we obtain for Dabrowska and Prentice-Cai

$$S\left\{-\int\frac{dM_1}{y_1} - \int\frac{dM_2}{y_2} + \int\int\left(\frac{dN_1dN_2 - Y_1Y_2d\Lambda_1d\Lambda_2}{y_1y_2}\right)\right.$$
$$\left.+\int\int\left(\frac{-dN_1Y_2d\Lambda_2 + Y_1Y_2d\Lambda_1d\Lambda_2 - dN_2Y_1d\Lambda_1 + Y_1Y_2d\Lambda_1d\Lambda_2}{y_1y_2}\right)\right\}$$
$$= S\left\{-\int\frac{dM_1}{y_1} - \int\frac{dM_2}{y_2} + \int\frac{dM_1}{y_1}\int\frac{dM_2}{y_2}\right\}.$$

We will now show that this is also the optimal influence curve.

### 6.6.2 Optimal influence curve under complete independence.

Denote the bivariate censored data with $Y$: so $Y = (\widetilde{T}_1, \widetilde{T}_2, D_1, D_2)$. The score operator for $S$ is given by:

$$A_F : L^2(F) \to L^2(P_{F,G}) : A_F(h)(Y) = E_F(h(T_1, T_2) \mid Y).$$

The information operator $A_F^\top A_F : L^2(F) \to L^2(F)$ is given by

$$A_F^\top A_F(h)(T_1, T_2) = E_{P_{F,G}}(E_F(h(T_1, T_2) \mid Y) \mid (T_1, T_2)).$$

Define $\kappa_t \equiv I_{(t,\infty)} - S(t) \in L^2(F)$. Then the efficient influence curve for estimating $S(t)$ is given by (see corollary 2.1):

$$\widetilde{I}(F, t) = A_F(A_F^\top A_F)^{-1}(\kappa_t) \in L^2(P_{F,G}).$$

Assume now complete independence. Let $t = (t_1, t_2)$, $\kappa_{t_1} \equiv I_{(t_1,\infty)} - S_1(t_1)$, $\kappa_{t_2} \equiv I_{(t_2,\infty)} - S_2(t_2)$. Define $h_1$ (univariate function in $T_1$) by $A_F^\top A_F(h_1) = \kappa_{t_1}$

and $h_2$ (univariate function in $T_2$) by $A_F^\top A_F(h_2) = \kappa_{t_2}$. Then by complete independence (notice that $A_F^\top A_F(h_1 h_2) = A_F^\top A_F(h_1) A_F^\top A_F(h_2)$) we have

$$\begin{aligned} A_F^\top A_F\left(h_1 h_2 + h_1 S_2(t) + h_2 S_1(t)\right) &= A_F^\top A_F(h_1) A_F^\top A_F(h_2) + S_2(t) A_F^\top A_F(h_1) \\ &\quad + S_1(t) A_F^\top A_F(h_2) \\ &= \kappa_{t_1}\kappa_{t_2} + S_2(t)\kappa_{t_1} + S_1(t)\kappa_{t_2} \\ &= \kappa_t. \end{aligned}$$

So under complete independence we have:

$$\widetilde{I}(F,t) = A_F(h_1 h_2 + h_1 S_2(t) + h_2 S_1(t)).$$

Again, by complete independence we have $A_F(h_1 h_2) = A_{F_1}(h_1) A_{F_2}(h_2)$ where $A_{F_1}(h_1) = E(h_1(T_1) \mid (\widetilde{T}_1, D_1))$ and $A_{F_2}(h_2) = E(h_2(T_2) \mid (\widetilde{T}_2, D_2))$. $A_{F_1}(h_1)$ is the efficient influence curve for estimating $S_1(t_1)$ for the univariate censoring model where we only observe $(\widetilde{T}_1, D_1)$ and we have a same statement for $A_{F_2}(h_2)$. So $A_{F_i}(h_i)$, $i = 1, 2$, equals the influence curve of the Kaplan-Meier estimator for estimating $F_i$ which is given by: $IC_i(t_i) \equiv -S_i(t_i) \int dM_i / y_i$, $i = 1, 2$. So under complete independence we have

$$\begin{aligned} \widetilde{I}(F,t) &= IC_1(t_1) IC_2(t_2) + IC_1(t_1) S_2(t_2) + IC_2(t_2) S_1(t_1) \qquad (6.37) \\ &= S(t) \left\{ \int \frac{dM_1}{y_1} \int \frac{dM_2}{y_2} - \int \frac{dM_1}{y_1} - \int \frac{dM_2}{y_2} \right\} \end{aligned}$$

and this is exactly the influence curve of the Dabrowska and Prentice-Cai estimator under complete independence. This proves that the Dabrowska and Prentice-Cai estimator are efficient under complete independence. Finally notice that (6.37) provides us with a nice and simple formula for the variance of the efficient influence curve:

$$\mathrm{Var}(IC_1(t_1))\mathrm{Var}\left(IC_2(t_2)\right) + S_1^2(t_1)\mathrm{Var}(IC_2(t_2)) + S_2^2(t_2)\mathrm{Var}(IC_1(t_1)).$$

For example, in the case that $T_1, T_2, C_1, C_2$ are all four independent and uniform$(0,1)$, the reader can easily verify that this variance equals:

$$\frac{1}{4}\left(1 + (1-t_1)^2 + (1-t_2)^2 - 3(1-t_1)^2(1-t_2)^2\right). \qquad (6.38)$$

Computer simulations for the Prentice-Cai and Dabrowska estimator show that this limiting variance is already closely approximated for $n = 100$ (see Bakker, 1990, Prentice and Cai, 1992a,1992b and chapter 8 of van der Laan, 1993).

# Chapter 7

# Modified EM-Estimator of the Bivariate Survival Function

## 7.1 Introduction.

In chapter 4 we proposed a SOR-MLE based on a modification of the data and explained why a solution of the self-consistency equation, computed by the EM- equations (iterating the self-consistency equation), will not be consistent for continuous data; the singly censored observations are not told how to redistribute their mass $1/n$ over their associated lines.

Pruitt (1991b) proposed an interesting estimator which is a solution of a modification of the self-consistency equation. Pruitt modifies the self-consistency equation by replacing the singly censored terms by ad hoc estimates (which are fixed in the subsequent EM-iterations) and thereby the singly censored observations are now told how to redistribute their mass. So it is not an NPMLE, and not efficient, but it shares several of the appealing properties of a self-consistent estimator and hence (see theorem 3.1) of a NPMLE.

Each observation in the bivariate censoring model (doubly, singly censored and uncensored) tells us that the survival time has fallen in a certain region: points for uncensored, lines for singly censored and quadrants for doubly censored. Iterating Pruitt's modified self-consistency equation tells us now that his estimator works as follows: each observation gets mass $1/n$ which it has to redistribute over its associated region for the survival time. By using kernel density estimators the singly censored observation are told how to redistribute their

mass $1/n$ over their associated lines. The uncensored observations give mass $1/n$ to the observed survival time. By solving the modified EM-equations the mass $1/n$ of the doubly censored observations is redistributed self-consistently over their associated quadrants: i.e. a point $t$ in the quadrant gets mass $1/n$ times the conditional density under the estimator, given the survival time lies in the quadrant. Consequently, the estimator is a distribution function and the mass $1/n$ corresponding with each observation is redistributed over the region where it belongs in a self-consistent (for the uncensored and doubly censored) or consistent way by listening to the other observations. The singly censored redistribution is estimated with product limit estimators of univariate kernel density estimators. Pruitt (1991b) makes his estimator intuitively clear and proves its self-consistency properties. Uniform consistency, asymptotic normality and asymptotic validity of the bootstrap has not yet been proved, and is done in this chapter (based on van der Laan, 1991).

We consider a slightly different version of his estimator: we use edge corrected bivariate kernel density estimators while he smoothes in one direction.

There are several motivations for being interested in Pruitt's estimator. Simulations (Pruitt, 1991b, and chapter 8 van der Laan, 1993d) show that his estimator is competitive with Dabrowska's and Prentice and Cai's estimators, while his estimator does not put negative mass on points (which is not true for Dabrowska's and Prentice and Cai's estimators). His idea of telling the singly censored observations how to redistribute its mass invites for less ad hoc estimators by not using kernel density estimators, but using the SOR-MLE in order to obtain an estimate of the conditional density over the lines (see van der Laan, 1993f, for practical results with these estimators).

Pruitt's estimator uses kernel density estimators and therefore also depends on a bandwidth, but simulations show that his estimator is *less sensitive* to the choice of the bandwidth than the SOR-MLE of chapter 4 to the choice of the grid-width. This is intuitively clear because the bandwidth of Pruitt's estimator influences only the redistribution of mass $1/n$ over lines, while a change of the grid-width in the SOR-MLE changes all interactions between the regions generated by the observations and hence the estimator might change at all its support points. Pruitt's estimator is also *less computer-intensive* than the SOR-MLE.

In the special submodel of the bivariate censoring model where one of the two survival times is always uncensored and the other is randomly censored, Pruitt's estimator is explicitly known and it is a NPMLE which by our results converges at root-$n$ rate. Gill and van der Vaart (1993) have a general theory

which shows efficiency of NPMLE which are known to be root-$n$ consistent. Unfortunately, their theory requires one cumbersome regularity condition which is expected to hold but which is hardly verifiable. All other conditions hold trivially. This submodel has an important application in regression analysis. Ritov (1992) proposes an efficient estimator for this submodel. For this submodel Pruitt's estimator is similar to Ritov's estimator. By explicitly writing out the influence curve of Pruitt's estimator one should be able to check that its asymptotic distribution is indeed the optimal one as given in Ritov (1992), but this goes beyond the scope of this chapter.

Finally, we have some remarks on points of technique: we use some novel methods which may well be useful in other analyses of $M$-estimators and analyses which involve density estimators. Pruitt's estimator is analyzed by applying the *implicit function theorem.* The implicit function theorem requires *invertibility of a derivative* of the modified self-consistency equation solved by Pruitt's estimator and a strong *differentiability* condition. We apply a general trick in order to get an equation with the required smoothness, see section 3. The invertibility proof (section 4) is highly non-trivial and might give techniques for proving invertibility of quite complicated operators of the form $I - A$ where $A$ has a norm larger than 1: so where it is certainly non-trivial that the Neumann series $\sum_{i=1}^{\infty} A^i(h)$ converges. We also formulate a functional delta-method for functionals like $\int \phi(f_n)d\mu$, where $f_n$ is a density estimator of $f_0$.

In this chapter we will prove (beyond existence of Pruitt's estimator $S_n$) strong uniform consistency of $S_n$ and weak convergence of the normalized difference $\sqrt{n}(S_n - S)$. The main work consists of proving weak convergence of the singly-censored terms in Pruitt's modified self-consistency equation (7.4), below, which involve density estimators, and proving the necessary conditions for the implicit function theorem for Banach spaces (Hildebrandt and Graves, 1927, Flett, 1980), to take care of the implicit character of equation (7.4) (its fourth term).

The *organisation* of this chapter is as follows. In section 2 we define the estimator and the modified self-consistency equation which is solved by it. We also define that part of the equation which is explicitly known and denote it by $\Psi_n$. In section 3 we state the consistency and weak convergence theorem and give the outline of the proof which is based on the implicit function theorem and the functional delta-method. The ingredients we need to verify will be formulated (like weak convergence and consistency of the explicit term $\Psi_n$ and invertibility of the derivative of the modified self-consistency equation). In section 4 we prove the first two ingredients; in particular the required invert-

ibility of the derivative of the modified self-consistency-equation. It remains to cover the analysis of $\Psi_n$. This requires a functional delta-method for density estimators as proved in section 5. In section 6 this functional delta-method is applied in the analysis of $\Psi_n$. The probabilistic conditions of this delta-method are covered in generality in section 6.1 and 6.2 and the result is summarized in lemma 7.4. In section 6.3 we make clear how to apply this delta method to our specific term and in section 6.4 the differentiability condition is proved, which completes the proof of all four ingredients.

## 7.2 A Modified EM-estimator (Pruitt).

For the description of the bivariate censoring model and notation we refer to chapter 4 and 6. In our analysis we need the following assumptions on $F_0$ and $G_0$:

**Assumptions.**
**1.** We restrict functions to a rectangle $[0,\tau] \subset [0,\infty)^2$, $\tau = (\tau_1,\tau_2)$, where $\tau$ is chosen so that $G_0(\tau-) = \delta_1 < 1$, $G_0(\tau) = 1$ and $G_0$ has an atom at $\tau$: $G_0(\{\tau\}) = \delta > 0$.
**2.** We assume that $G_0$ has a density $g_0$ w.r.t. the Lebesque measure on $[0,\tau)$ and that $F_0$ has a density $f_0$ w.r.t. the Lebesque measure on $[0,\tau+\epsilon]$ for certain $\epsilon > 0$. Furthermore we assume that $f_0, g_0 \in C^3[0,\tau]$.
**3.** Moreover, we assume that for some $\epsilon = (\epsilon_1,\epsilon_2)$ $f_0$ is strictly positive on $[0,\tau+\epsilon]\backslash[0,\tau]$.

Assumption 1 can be accomplished by censoring observations which do not fall in the rectangle $[0,\tau]$ at the edge of the rectangle. In real life this means a small loss of information, but a gain in stability of the estimator.

Let $P_n$ be the empirical distribution function of $Y_i$, $i = 1,\ldots,n$. The EM-algorithm finds a solution of the self-consistency equation:

$$S_n(t) = \int P_{S_n}(T > t \mid y) dP_n(y). \tag{7.1}$$

We refer to chapter 3 for a discussion on the EM-algorithm and its heuristics. The integral w.r.t. $P_n$ in equation (7.1) can be written as a sum of four integrals, namely w.r.t. the empirical distribution $P_{11}^n$ of the uncensored observations, two with respect to the empirical distributions $P_{10}^n$ and $P_{01}^n$ of the singly censored observations and one with respect to the empirical distribution $P_{00}^n$ of the doubly censored observations. Pruitt's estimator is the solution

of the equation obtained by replacing in (7.1) the integrands $P_{S_n}(T > t \mid y)$ in the two singly censored terms, which involve the unknown $S_n$, by explicit estimators.

Let's write down the modified self-consistency equation which is solved by $S_n$ (Pruitt's estimator). The two conditional densities over the lines corresponding with the singly censored observations which appear in the self-consistency equation are given by:

$$\begin{aligned} W_{1F_0}(t_1, y_1, y_2) &\equiv P_{F_0}(T_1 > t_1 \mid T_1 > y_1, T_2 = y_2) \\ W_{2F_0}(t_2, y_1, y_2) &\equiv P_{F_0}(T_2 > t_2 \mid T_2 > y_2, T_1 = y_1). \end{aligned} \tag{7.2}$$

Pruitt estimates them with two weighted product limit estimators $\widehat{W}_1, \widehat{W}_2$, respectively. We will define these product limit estimators in section 6 (see (7.17)).

We have the equation $S(t) = \int P_S(T > t \mid y) dP_{F,G}(y)$. In formulas, using at the second equality that $G$ has support on $[0, \tau]$, it is given by:

$$\begin{aligned} S(t) &= \int_{t_1}^{\tau_1} \int_{t_2}^{\tau_2} dP_{11}(y_1, y_2) \\ &\quad + \int_0^{\tau_1} \int_{t_2}^{\tau_2} W_{1F}(t_1, y_1, y_2) dP_{01}(y_1, y_2) \\ &\quad + \int_{t_1}^{\tau_1} \int_0^{\tau_2} W_{2F}(t_2, y_1, y_2) dP_{10}(y_1, y_2) \\ &\quad + \int_0^{\tau_1} \int_0^{\tau_2} \frac{S(t_1 \vee y_1, t_2 \vee y_2)}{S(y_1, y_2)} dP_{00}(y_1, y_2) \\ &\equiv \Psi(t) + \int_0^{\tau_1} \int_0^{\tau_2} \frac{S(t_1 \vee y_1, t_2 \vee y_2)}{S(y_1, y_2)} dP_{00}(y_1, y_2), \end{aligned} \tag{7.3}$$

where $S(t_1 \vee y_1, t_2 \vee y_2)/S(y_1, y_2) = P(T_1 > t_1, T_2 > t_2 \mid T_1 > y_1, T_2 > y_2)$ and

$$\Psi(t) \equiv \overline{P}_{11}(t) + \int_0^{\tau_1} \int_{t_2}^{\tau_2} W_1(t_1, y) dP_{01}(y) + \int_{t_1}^{\tau_1} \int_0^{\tau_2} W_2(t_2, y) dP_{10}(y).$$

If we replace $P_{F,G}$ by $P_n$, then we obtain the self-consistency equation. $S_n$ (Pruitt's estimator) solves

$$\begin{aligned} S_n(t) &= \tfrac{1}{n} \sum_{i=1}^{n} I(\widetilde{T}_i > t, D_{i1} = D_{i2} = 1) + \int_0^{\tau_1} \int_{t_2}^{\tau_2} \widehat{W}_1(t_1, y) dP_{01}^n(y) \\ &\quad + \int_{t_1}^{\tau_1} \int_0^{\tau_2} \widehat{W}_2(t_2, y) dP_{10}^n(y) + \int_0^{\tau_1} \int_0^{\tau_2} \frac{S_n(t \vee y)}{S_n(y)} dP_{00}^n(y) \\ &\equiv \Psi_n(t) + \int_0^{\tau_1} \int_0^{\tau_2} \frac{S_n(t \vee y)}{S_n(y)} dP_{00}^n(y). \end{aligned} \tag{7.4}$$

The only difference with the self-consistency equation is that in the self-consistency equation we have $W_{1F_n}$, $W_{2F_n}$, self-consistent redistribution of mass $1/n$, instead of $\widehat{W}_1$, $\widehat{W}_2$, redistribution of the mass $1/n$ according to a predetermined estimate. $\Psi_n$ represents the empirical counterpart of $\Psi$.

## 7.3 Outline of proof of consistency and weak convergence.

We will prove the following theorem.

**Theorem 7.1** *Assume that the underlying densities $f, g$ satisfy assumptions* **1**, **2** *and* **3** *made in the introduction. Assume that for $\widehat{W}_1, \widehat{W}_2$ we use kernel density estimators with a kernel $K$ satisfying the assumptions as stated in lemma 7.4 and bandwidth $h_n = n^{-1/7}$.*

*Then $\|(S_n - S)\|_\infty \to 0$ in supnorm a.s. and $\sqrt{n}(S_n - S)$ converges weakly in $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$.*

*Outline of Proof.* Equation (7.3) is given by:

$$S(t) = \Psi(t) + \int_0^{\tau_1}\int_0^{\tau_2} \frac{S(t \vee y)}{S(y)} dP_{00}(y). \tag{7.5}$$

If we consider $P_{00}$ as fixed (known), then $S$ can be considered as a solution of $K(S, \Psi) = 0$, where

$$K(S,\Psi)(t) \equiv \Psi(t) - S(t) + \int_0^{\tau_1}\int_0^{\tau_2} \frac{S(t \vee y)}{S(y)} dP_{00}(y). \tag{7.6}$$

Pruitt's estimator $S_n$ is a solution of (7.4) which is the same equation but where $P_{11}, P_{01}, P_{10}, P_{00}$ are replaced by their empirical distributions and the singly censored conditional probabilities $W_1$ and $W_2$ are replaced by the weighted product limit estimators $\widehat{W}_1$ and $\widehat{W}_2$, respectively, defined by equation (7.17) in section 6. In formulas we have:

$$\begin{aligned} S_n(t) &= \Psi_n(t) + \int_0^{\tau_1}\int_0^{\tau_2} \frac{S_n(t \vee y)}{S_n(y)} dP^n_{00}(y) \\ &= \Psi^*_n(t) + \int_0^{\tau_1}\int_0^{\tau_2} \frac{S_n(t \vee y)}{S_n(y)} dP_{00}(y), \end{aligned}$$

where

$$\Psi^*_n(t) \equiv \Psi_n(t) + \int_0^{\tau_1}\int_0^{\tau_2} \frac{S_n(t \vee y)}{S_n(y)} d(P^n_{00} - P_{00})(y).$$

Consequently, $S_n$ is an implicit solution of $K(S_n, \Psi^*_n) = 0$.

We will apply the implicit function theorem for Banach spaces (Hildebrandt and Graves, 1927, Flett, 1980, p. 205) to

$$K : (D([0,\tau]), \|\cdot\|_\infty)^2 \to (D([0,\tau]), \|\cdot\|_\infty) : \ (S,\Psi) \to K(S,\Psi).$$

It says:

**Theorem 7.2** (Implicit function theorem). *Assume*

1. *$K$ is a continuously Fréchet differentiable functional from an open subset $W$ of $(D([0,\tau]), \|\cdot\|_\infty)^2$ into $(D([0,\tau]), \|\cdot\|_\infty)$, with $(S,\Psi) \in W$. Continuity of the derivative is defined as continuity with respect to the operator norm: If $\|x_n - x\|_\infty \to 0$, then $\sup_{\|h\|_\infty=1} \|dK(x_n)(h) - dK(x)(h)\|_\infty \to 0$.*

2. *The partial derivative $d_1K(S,\Psi) : (D([0,\tau]), \|\cdot\|_\infty) \to (D([0,\tau]), \|\cdot\|_\infty)$ is invertible, and its inverse is continuous (i.e. it is an isomorphism).*

*Then there are open neighborhoods $U_0$ of $\Psi$ and $V_0$ of $S$ in $(D([0,\tau]), \|\cdot\|_\infty)$ such that for each $\Psi' \in U_0$, there is a unique $S' \in V_0$ such that $K(S',\Psi') = 0$. Moreover, if we define $\Theta$ by $S' = \Theta(\Psi')$, then for $U$ and $V$ small enough, $\Theta(\cdot)$ is a continuously Fréchet differentiable mapping from $U$ into $V$. Its derivative is given by*

$$d\Theta(\Psi) = -(d_1K(\Theta(\Psi),\Psi))^{-1} \circ d_2K(\Theta(\Psi),\Psi).$$

Because of the simple structure of $K$ ($P_{00}$ is fixed), continuous Fréchet differentiability of $K$ is easy to verify, provided that $S > \epsilon > 0$ as is guaranteed by our assumptions.

All the work has to be done in the verification of 2. The partial derivative $d_1K(S,\Psi)$ of $K$ with respect to $S$ is given by: $-(I-A) : (D([0,\tau]), \|\cdot\|_\infty) \to (D([0,\tau]), \|\cdot\|_\infty)$, where

$$(I-A)(h)(t) = h(t) - \int_0^{\tau_1}\int_0^{\tau_2} \frac{h(t\vee y)S(y) - h(y)S(t\vee y)}{S^2(y)} dP_{00}(y).$$

In the next section it will be proved that $I - A$ is invertible and that its inverse $\sum_{n=0}^{\infty} A^n$ is a continuous operator (see theorem 7.3). In this proof it is important to notice that the integrand $(h(t\vee y)S(y) - h(y)S(t\vee y))/S^2(y)$ is zero at point $\tau$. Therefore we only have to integrate over $\widetilde{B} \equiv [0,\tau]\backslash\{\tau\}$, and by assumption 1 we have that

$$\int_{\widetilde{\tau}} \frac{dP_{00}}{S} = G([0,\tau]) - G(\{\tau\}) = 1 - \delta < 1,$$

which we will need in the invertibility proof.

Consequently, we can apply the implicit function theorem. The implicit function theorem tells us that there exists a solution $S_n'$ close to $S$ where $S_n' = \Theta(\Psi_n^*)$. In section 5 we will prove that $\Psi_n$ is uniformly consistent and that $\sqrt{n}(\Psi_n - \Psi)$ converges weakly as elements of $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$. This does not immediately imply the same results for $\Psi_n^*$ because it involves $S_n$. However, the following argument proves it.

The modified self-consistency equation (7.4) tells us that $S_n(t) > \overline{P}_{11}^n(t)$. By assumption 2 and 3 on $f, g$ we have that $\overline{P}_{11}$ is uniformly bounded away from zero on $[0,\tau]$ and we know by Glivenko-Cantelli that $\overline{P}_{11}^n \to \overline{P}_{11}$. Consequently $S_n(t) > \delta > 0$ with probability tending to 1. Moreover, $S_n$ is monotone ($S_n$ only assigns positive mass) and $y \to S_n(t \vee y)/S_n(y)$ is bounded by 1 and, by lemma 1.5, is of uniformly (in $n$ and $t$) bounded uniform sectional variation . Now, by using integration by parts we can bound $\int_0^{\tau_1}\int_0^{\tau_2} S_n(t\vee y)/S_n(y) d(P_{00} - P_{00}^n)(y)$ by a bounded constant (involving the latter variation) times the supremum norm of $P_{00} - P_{00}^n$, and consequently it follows that this term converges uniformly to zero with probability one, independently of the asymptotic behaviour of $S_n$. Therefore, if $\Psi_n$ converges uniformly with probability one to $\Psi$ then $\Psi_n^*$ converges uniformly with probability one to $\Psi$, independently of the asymptotic behaviour of $S_n$.

Consequently the consistency of $\Psi_n$ provides us with consistency of $\Psi_n^*$ and therefore the continuous mapping theorem 1.2 provides us with uniform consistency of $S_n = \Theta(\Psi_n^*)$ ($\Theta$ is Fréchet differentiable). Moreover, the continuous mapping theorem provides us also straightforwardly with the following: if $S_n$ is uniformly consistent and $\sqrt{n}(\Psi_n - \Psi)$ converges weakly to a Gaussian process, then $\sqrt{n}(\Psi_n^* - \Psi)$ converges weakly to a Gaussian process, but another process. Now, the functional delta method theorem 1.5 applied to $\Theta(\Psi_n^*)$ provides us with the weak convergence of $\sqrt{n}(S_n - S)$ to a Gaussian process, namely a linear transformation of the limiting distribution of $\sqrt{n}(\Psi_n^* - \Psi)$.

The implicit function theorem tells us that there exists a solution $S_n'$ close to $S$ where $S_n' = \Theta(\Psi_n^*)$ and the result derived above holds for this $S_n'$. Because $K(S, \Psi_n^*) = 0$ might have several solutions, the $S_n$ which we compute with the EM-algorithm is not necessarily the $S_n' = \Theta(\Psi_n^*)$ given by the implicit function theorem. However, if we prove that each survival function $S_n$ which solves $K(S_n, \Psi_n^*) = 0$ is consistent, then for $n$ large enough we have $S_n = \Theta(\Psi_n^*)$. We will prove this in the next section (lemma 7.1).

We conclude that in the next sections the following four things have to be proved:

- $I - A$ is invertible, and has a continuous inverse (theorem 7.3).
- Each survival function $S_n$ which solves $K(S_n, \Psi_n^*) = 0$ is consistent (lemma 7.1).
- $||\Psi_n - \Psi||_\infty$ converges with probability 1 to zero (section 6).
- $\sqrt{n}(\Psi_n - \Psi)$ converges weakly as elements of $(D[0,\tau], \mathcal{B}, ||\cdot||_\infty)$, jointly with the empirical process $\sqrt{n}(P_n - P_{F,G})$. This is also proved in section 6 by application of the results of section 5.

Notice that $\Psi_n$ involves density estimators so that the second and third point do not follow from empirical process theory and are certainly not trivial. In order to carry through the analysis we need conditions on the kernel and the bandwidth (see theorem 7.1).

**Bootstrap.** We can explicitly write down the linearization of $\sqrt{n}(\Psi_n^* - \Psi)$ in terms of Gaussian processes. Denote this derivative with $d\Psi(Z)$ where $Z$ is a Gaussian process. Then we have that $\sqrt{n}(S_n - S) \xRightarrow{D} \sum_{i=1}^{\infty} A^i(d\Psi(Z))$. It is clear that this is a quite complicated expression which cannot be explicitly written down, just as the efficient influence curve for the bivariate censoring model. Therefore the most one can do is to approach the covariance structure of the limit distribution of $S_n$ numerically (for certain known $F$ and $G$). We can also use a semiparametric bootstrap (sampling from a smoothed $P_{F_n,G_n}$) in order to estimate the variance of $S_n$ (see van der Laan, 1991).

## 7.4 Invertibility of the derivative of the modified self-consistency equation.

Recall $\widetilde{B} \equiv [0,\tau]\backslash\{\tau\}$, and define the operator: $I - A : (D[0,\tau], ||\cdot||_\infty) \rightarrow (D[0,\tau], ||\cdot||_\infty)$ by

$$(I - A)(h)(t) = h(t) - \int_{\widetilde{\tau}} \left\{ \frac{h(t \vee y)S(y) - h(y)S(t \vee y)}{S^2(y)} \right\} dP_{00}(y).$$

As shown in the general proof, in order to apply the implicit function theorem to the equation $K(S, \Psi)$ we need to prove that the linear operator $(I - A)$ is an isomorphism.

**Theorem 7.3** *The linear operator* $I - A : (D[0,\tau], \|\cdot\|_\infty) \to (D[0,\tau], \|\cdot\|_\infty)$ *as defined above is an isomorphism (i.e. a linear invertible mapping with continuous inverse). Its inverse is given by:*

$$(I-A)^{-1} = \sum_{n=0}^{\infty} A^n.$$

**Proof.** Define

$$\begin{aligned} A_1(h) &\equiv \int_{\widetilde{\tau}} \left( \frac{h(t \vee y) S(y)}{S^2(y)} \right) dP_{00}(y) \\ A_2(h) &\equiv \int_{\widetilde{\tau}} \left( \frac{h(y) S(t \vee y)}{S^2(y)} \right) dP_{00}(y) \\ A(h) &\equiv A_1(h) - A_2(h). \end{aligned}$$

One should notice that if for an $h \in D([0,\tau])$ the series $T(h) \equiv \sum_{n=0}^{\infty} A^n(h)$ converges, then $(I-A)\left(\sum_{n=0}^{\infty} A^n\right)(h) = \left(\sum_{n=0}^{\infty} A^n\right)(I-A)(h) = h$. Assume that for $h \geq 0$: $\|T(h)\|_\infty \leq M\|h\|_\infty$. Then the same inequality holds for $h \leq 0$. Consider now a general $h \in D([0,\tau])$ with $h = hI(h>0) + hI(h \leq 0) \equiv h_1 + h_2$, $h_1, h_2 \in D([0,\tau])$. Then

$$\|T(h)\|_\infty \leq \|T(h_1)\|_\infty + \|T(h_2)\|_\infty \leq M(\|h_1\|_\infty + \|h_2\|_\infty) \leq 2M\|h\|_\infty.$$

So, then $T$ is a well defined bounded linear operator, which proves the theorem. So it remains to prove that if $h \geq 0$, then $\|T(h)\|_\infty \leq M\|h\|_\infty$.

Here follows the proof of this. Let $h \geq 0$ be fixed. For a constant $c$ we have that $A_1(c) = \delta c$, where $\delta \equiv \int_{\widetilde{B}} 1/S dP_{00}$. Using this tells us that:

$$A(h) = A_1(h) - A_2(h) = A_1\left(h - \|h\|_\infty\right) - \left(A_2(h) - \delta\|h\|_\infty\right). \tag{7.7}$$

One should notice that (use $S > 0$) $\delta = \int_{\widetilde{B}} 1/S dP_{00} = \int_{\widetilde{B}} P(C_1 \in dy_1, C_2 \in dy_2) = P(C \in \widetilde{B})$. By assumption 1 $G$ has an atom in the point $\{\tau\}$. Therefore we have $\delta < 1$.

We have for each $h \in D([0,\tau])$: $\|A_i(h)\|_\infty \leq \delta\|h\|_\infty$, $i = 1, 2$. If in the sequel we say that $f$ is non-increasing, then we mean: if $t \geq s$, so $t_1 \geq s_1$, $t_2 \geq s_2$, then $f(t) \leq f(s)$. We have that $A_2(h) - \delta\|h\|_\infty \leq 0$ and because $S$ is non-increasing $A_2(h) - \delta\|h\|_\infty$ is non-increasing (recall $h \geq 0$). With the use of this fact we can prove the following property for $h \geq 0$:

$$\|A^n\left(A_2(h) - \delta\|h\|_\infty\right)\|_\infty \leq \delta^{n+1}\|h\|_\infty. \tag{7.8}$$

**Proof of Property (7.8).** Assume that $h \in D([0,\tau])$ is non-decreasing and $h \geq 0$ (we will denote this with $h \geq 0\ \uparrow$). Then it is easy to see that $A(h)$ is also non-decreasing: if $t \geq s$, then $A(h)(t) - A(h)(s)$ equals

$$\int \frac{\left(h(t \vee y) - h(s \vee y)\right) S(y) + h(y)\left(S(s \vee y) - S(t \vee y)\right)}{S^2(y)} dP_{00}(y).$$

Because $h(t \vee y) - h(s \vee y) \geq 0$ and $S(s \vee y) - S(t \vee y) \geq 0$ this term is equal or larger than zero. Now, rewrite the numerator of the integrand of $A$ as follows:

$$h(t \vee y)S(y) - h(y)S(t \vee y) = (h(t \vee y) - h(y))\, S(y) + h(y)\,(S(y) - S(t \vee y))\,.$$

So we have $A(h) \geq 0$. This shows that: if $h \geq 0 \uparrow$, then $A(h) \geq 0 \uparrow$. We also have that if $h \geq 0$, then $\|A(h)\|_\infty \leq \delta \|h\|_\infty$. Therefore, if $h \geq 0 \uparrow$, then $\|A^n(h)\|_\infty \leq \delta^n \|h\|_\infty$. This provides us also with the following result:

$$\text{if } h \leq 0 \downarrow, \text{ then } \|A^n(h)\|_\infty \leq \delta^n \|h\|_\infty. \tag{7.9}$$

Now, by applying (7.9) to $A_2(h) - \delta\|h\|_\infty \leq 0 \downarrow$ we have:

$$\|A^n\,(A_2(h) - \delta\|h\|_\infty)\,\|_\infty \leq \delta^n \|\,(A_2(h) - \delta\|h\|_\infty)\,\|_\infty \leq \delta^{n+1}\|h\|_\infty, \tag{7.10}$$

which proves (7.8).

Notice also that $A_1\,(h - \|h\|_\infty) \leq 0$. Now, we are ready to prove with induction that the following statement $P(n)$ is true for all $n \in \mathbb{N}$:

$$P(n),\ n \in \mathbb{N}: \text{ If } h \geq 0, \text{ then } \|A^n(h)\|_\infty \leq n\delta^n \|h\|_\infty.$$

$P(1)$ is trivially true. Assume $P(n)$ is true. We will prove $P(n+1)$.

$$\begin{aligned}
\|A^{n+1}(h)\|_\infty &\leq \|A^n A_1\,(\|h\|_\infty - h)\,\|_\infty + \|A^n\,(A_2(h) - \delta\|h\|_\infty)\,\|_\infty \\
&\quad (\text{ by (7.7) and the triangle inequality, respectively.}) \\
&\leq n\delta^n \|A_1\,(\|h\|_\infty - h)\,\|_\infty + \delta^{n+1}\|h\|_\infty \\
&\quad (\text{ by } P(n) \text{ and (7.8), respectively}). \\
&\leq n\delta^{n+1}\|h\|_\infty + \delta^{n+1}\|h\|_\infty = (n+1)\delta^{n+1}\|h\|_\infty.
\end{aligned}$$

So with induction we proved: if $h \geq 0$, then $\|A^n(h)\|_\infty \leq n\delta^n \|h\|_\infty$. Consequently, if $h \geq 0$, then

$$\|T(h)\|_\infty \leq \sum_{n=0}^{\infty} \|A^n(h)\|_\infty \leq \|h\|_\infty \sum_{n=0}^{\infty} n\delta^n = \|h\|_\infty \frac{\delta}{(1-\delta)^2}.$$

This completes the proof of theorem 7.3. □

We will now prove consistency for each survival function $S_n$ which solves $K(S_n, \Psi_n^*) = 0$. We have

$$S_n(t) = \Psi_n^*(t) + \int \frac{S_n(t \vee y)}{S_n(y)} dP_{00}(y)$$

and

$$S_0(t) = \Psi_0(t) + \int \frac{S_0(t \vee y)}{S_0(y)} dP_{00}(y).$$

Subtracting these two equations provides us with:

$$\begin{aligned}(S_n - S_0)(t) &= (\Psi_n^* - \Psi_0)(t) + \int (S_n - S_0)(t \vee y) \frac{dP_{00}(y)}{S_0(y)} \\ &\quad - \int (S_n - S_0)(y) \frac{S_n(t \vee y)}{S_n(y) S_0(y)} dP_{00}(y). \qquad (7.11)\end{aligned}$$

Denote

$$\begin{aligned}A_{S_n}(S_n - S_0)(t) &\equiv \int (S_n - S_0)(t \vee y) \frac{dP_{00}(y)}{S_0(y)} \\ &\quad - \int (S_n - S_0)(y) \frac{S_n(t \vee y)}{S_n(y) S_0(y)} dP_{00}(y) \\ &\equiv A_1(S_n - S_0)(t) - A_2^{S_n}(S_n - S_0)(t),\end{aligned}$$

where $A_1$ is the same as defined in the proof of theorem 3.3 and $A_2^{S_n}$ is slighly different from the operator $A_2$. Now, (7.11) reduces to:

$$(I - A_{S_n})(S_n - S_0) = (\Psi_n^* - \Psi_0).$$

Therefore for consistency of $S_n$ it suffices to prove that $\sum_{k=0}^{\infty} A_{S_n}^k$ is a bounded linear operator (uniformly in $n$). However, because $A_2^{S_n}$ has all the properties which we needed from $A_2$ (as the reader can verify for himself) we can do exactly the same proof as the proof of theorem 3.3 and we also get the same bound $\delta/(1-\delta)^2$ of the norm of $(I - A_{S_n})^{-1}$. The only condition we need is that $S_n > \epsilon > 0$ on $[0, \tau]$ which holds for $n$ large enough (because $S_n > \overline{P}_{11}^n$ and $\overline{P}_{11}^n \to \overline{P}_{11} > \delta > 0$, by assumption 1). This proves that $||S_n - S_0||_\infty \to 0$ a.s.

In the same way it is proved that $K(S, \Psi_n^*) = 0$ has a unique solution among the survival functions $S > 0$ on $[0, \tau]$. This provides us with the following lemma:

**Lemma 7.1** *Recall the assumptions on the model. Each survival function $S_n$ which solves $K(S_n, \Psi_n^*) = 0$ is strongly uniformly consistent and if $S_n > 0$ on $[0, \tau]$, then $S_n$ is also the unique survival function solution of $K(S, \Psi_n^*) = 0$ in the class of survival functions $S$ with $S > 0$ on $[0, \tau]$.*

## 7.5 Functional delta-method for functionals of density estimators.

Consider the problem of estimation of a functional $\Phi(F) = \Gamma(f) \in (D[0,\tau], \|\cdot\|_\infty)$, where $f \equiv dF/d\mu : \mathbb{R}^d \to \mathbb{R}$ is the density of a $d$-variate distribution $F$ w.r.t. to the Lebesque measure $\mu$, using i.i.d. observations $X_1, \ldots, X_n$, $X_i \sim F$. We can estimate $\Gamma(f)$ with $\Gamma(f_n)$ where $f_n(x) = (nh^d)^{-1} \sum_{i=1}^n K((x - X_i)/h)$ is the usual $d$-variate kernel density estimator (Silverman, 1986) with a bandwidth $h = h(n) \to 0$. In this section we show how we can use theorem 1.5 in order to obtain a functional delta method theorem for the analysis of functionals of density estimators.

**Lemma 7.2** *Assume that:*

**1.** $\|f_n - f_0\|_\infty \to 0$ *a.s.*

**2.** *Define* $\widetilde{F}_n(x) \equiv \int_0^x f_n(x)dx$ *and let* $F_n$ *be the empirical distribution function of* $X_i$, $i = 1, \ldots, n$. *Denote the limiting distribution of* $\sqrt{n}(F_n - F)$ *with* $Z$ *(i.e. the* $F$*-Brownian bridge), where* $Z$ *is a Borel measurable Gaussian process concentrated on a separable subset* $D_0$ *of* $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$. *Assume that* $\widetilde{Z}_n \equiv \sqrt{n}(\widetilde{F}_n - F) \stackrel{D}{\Longrightarrow} Z$ *in* $(D[0,\tau], \mathcal{B}, \|\cdot\|_\infty)$.

**3.** $\limsup_n \|f_n\|_v^* < M < \infty$ *a.s.*

*Assume now that* $\Phi$ *satisfies the following purely analytical property: For each sequence* $\widetilde{Z}_n \equiv \sqrt{n}(\widetilde{F}_n - F) \to Z$ *in supnorm for* $Z \in D_0$, $\|f_n - f\|_\infty \to 0$ *and* $\|f_n\|_v^* = O(1)$, *we have:*

$$\sqrt{n}(\Phi(\widetilde{F}_n) - \Phi(F)) - d\Phi(F)(Z) \to 0$$

*in supnorm for a continuous linear mapping* $d\Phi(F) : (D[0,\tau], \|\cdot\|_\infty) \to (D[0,\tau], \|\cdot\|_\infty)$.

*Then*

$$\sqrt{n}\left(\Phi(\widetilde{F}_n) - \Phi(F)\right) \stackrel{D}{\Longrightarrow} d\Phi(F)(Z) \text{ in } (D[0,\tau], \mathcal{B}, \|\cdot\|_\infty). \tag{7.12}$$

The proof of this lemma is nothing else than an application of theorem 1.5 applied to $\Phi : (D[0,\tau], \|\cdot\|_\infty) \to (D[0,\tau], \|\cdot\|_\infty)$ with a good choice for $D_n$ so that we only have to verify the differentiability property for sequences $\widetilde{F}_n$ for which $f_n \to f$ and $\|f_n\|_v^* < M < \infty$. Firstly, notice that $\|f_n - f\|_\infty \to 0$ a.e. is equivalent with: for each $\epsilon > 0$ $P(\lim_{N\to\infty} \sup_{n>N} \|f_n - f\|_\infty > \epsilon) = 0$. By Fatou's lemma this implies that for all $\epsilon > 0$ $\lim_{N\to\infty} P(\sup_{n>N} \|f_n - f\|_\infty > \epsilon) = 0$. This implies that there

exist sequences $\epsilon_n \to 0$, $\delta_n \to 0$ and $N(\delta_n) \in \mathbb{N}$ so that

$$P\left(\sup_{n>N(\delta_n)} ||f_n - f||_\infty < \epsilon_n\right) > 1 - \delta_n. \tag{7.13}$$

Now, we will define the $D_n$ in theorem 1.5. Let $\mathcal{F}$ be the set of all distribution functions $F_1 : \mathbb{R}^d \to \mathbb{R}$ which are absolute continuous w.r.t. the Lebesque measure. Define now

$$D_n \equiv \{\sqrt{n}(F_1 - F) : F_1 \in \mathcal{F},\ ||f_1 - f||_\infty < \epsilon_n, ||f_1||_v^* \leq M\}$$

and

$$\widetilde{Z}_n^* \equiv \widetilde{Z}_n I(\widetilde{Z}_n \in D_n),$$

where we mean with $I(\widetilde{Z}_n \in D_n)$ that if $\widetilde{Z}_n \notin D_n$, then $I(\widetilde{Z}_n \in D_n) = 0$. Consequently $\widetilde{Z}_n^* \in D_n$. By $\limsup_n ||f_n||_v^* = O(1)$ and (7.13) we have that for each $\epsilon > 0$, there exists a $N(\epsilon)$ so that $\widetilde{Z}_n \in D_n$ for all $n > N(\epsilon)$ with probability $1 - \epsilon$. Therefore, $\widetilde{Z}_n \xRightarrow{D} Z$ implies $\widetilde{Z}_n^* \xRightarrow{D} Z$. Now, apply theorem 1.5 to $\Phi(\widetilde{F}_n^*)$, where $\widetilde{F}_n^* = F + 1/\sqrt{n}\widetilde{Z}_n^*$. This provides us with:

$$\sqrt{n}\left(\Phi(\widetilde{F}_n^*) - \Phi(F)\right) \xRightarrow{D} d\Phi(F)(Z).$$

Because $Z_n^* = Z_n$ with probability tending to 1 we have that

$$||\sqrt{n}\left(\Phi(\widetilde{F}_n^*) - \Phi(F)\right) - \sqrt{n}\left(\Phi(\widetilde{F}_n) - \Phi(F)\right)||_\infty = o_P(1).$$

The required weak convergence follows now from the general fact that

$$X_n \xRightarrow{D} X,\ Y_n = o_P(1) \Rightarrow Z_n \equiv X_n + Y_n \xRightarrow{D} X,$$

which completes the proof of the lemma.

The lemma can be immediately generalized to all kinds of properties of the sequences $\widetilde{F}_n$ which we plug in, as long as these properties hold with probability tending to 1. The lemma will be applied in the analysis of $\sqrt{n}(\Psi_n - \Psi)$ in the next section. Here, the probabilistic conditions of the lemma will be analyzed in generality.

## 7.6 Weak convergence of the explicit part.

We will apply the refined functional delta-method lemma 7.2 in the analysis of $\sqrt{n}(\Psi_n - \Psi)$. We will see in the next subsection that $\int W_1 dP_{01}$ has a representation in terms of two distribution functions $F_N$ and $F_Y$ of the data, and of course a symmetric version of this statement holds for $\int W_2 dP_{10}$ (say $F_N'$, $F_Y'$).

So we can represent $\Psi$ in terms of distribution functions of the data for which we have a joint weak convergence result for its empirical counterpart, namely $P \equiv (F_N, F_Y, F'_N, F'_Y, P_{11}, P_{01}, P_{10}, P_{00})$. In order to get $\Psi_n$ one replaces these distributions by their empirical versions: so $\Psi = \Psi(P)$ and $\Psi_n = \Psi(P_n)$ where we know that $\sqrt{n}(P_n - P) \stackrel{D}{\Longrightarrow} Z$ for a certain Gaussian process $Z$. The refined delta method lemma 7.2 states now that in order to prove weak convergence of $\sqrt{n}(\Psi_n - \Psi)$ it is enough to show that this representation satisfies the characterization of compact differentiability for all sequences $Z_n = \sqrt{n}(P_n - P) \in D_n$, where $D_n$ is chosen so that the empirical process $Z_n \in D_n$ with probability tending to 1.

Define the following normalized estimators:

$$\begin{aligned} U_{01}^n(t) &\equiv \sqrt{n}(P_{01}^n - P_{01})(t) \\ U_{11}^n(t) &\equiv \sqrt{n}(P_{11}^n - P_{11})(t) \\ U_{10}^n(t) &\equiv \sqrt{n}(P_{10}^n - P_{10})(t). \end{aligned}$$

Then we can rewrite $\sqrt{n}(\Psi_n - \Psi)$ in terms of these normalized empirical differences:

$$\begin{aligned} &\sqrt{n}(\Psi_n - \Psi)(t) = U_{11}^n(t) \\ &+ \int_0^{\tau_1} \int_{t_2}^{\tau_2} W_1(t_1, y_1, y_2) dU_{01}^n(y) + \sqrt{n} \int_0^{\tau_1} \int_{t_2}^{\tau_2} \left(\widehat{W}_1 - W_1\right)(t_1, y_1, y_2) dP_{01}(y) \\ &+ \int_{t_1}^{\tau_1} \int_0^{\tau_2} W_2(t_2, y_2, y_1) dU_{uc}^n(y) + \sqrt{n} \int_{t_1}^{\tau_1} \int_0^{\tau_2} \left(\widehat{W}_2 - W_2\right)(t_2, y_2, y_1) dP_{10}(y) \\ &+ \int_0^{\tau_1} \int_{t_2}^{\tau_2} \left(\widehat{W}_1 - W_1\right)(t_1, y_1, y_2) dU_{01}^n(y) + \int_{t_1}^{\tau_1} \int_0^{\tau_2} \left(\widehat{W}_2 - W_2\right)(t_2, y_2, y_1) dU_{10}^n(y). \end{aligned}$$

We will now verify the purely analytical characterization of compact differentiability. Assume that $\sqrt{n}(P_n - P)$ converges in supremum norm to $Z$. In order to prove the characterization of compact differentiability we need to prove that the first, second, third, fourth and fifth term converge in supremum norm, and that the last terms converge to zero in supremum norm. The third term will be analyzed in the next subsection. In that analysis one has to keep continually in mind that if we consider weak convergence of the normalized empirical processes which occur in this term that these should be taken jointly with the other normalized empirical processes! (we will not remind the reader again of this fact). The fifth term is of exactly the same structure. The first term is trivial. For the convergence of the second and fourth term we apply integration by parts lemma 6.1 so that the integrals become integrals with respect to $W_1$

and $W_2$ and that $U_{01}^n$ and $U_{10}^n$ appear as functions. This can be done because $W_1$ and $W_2$ are of bounded uniform sectional variation uniformly in $t \in [0, \tau]$, by assumption 2 about $f, g$ and $S(\tau) > 0$. This proves the convergence of the second and fourth integral. In the next section we will see that $\widehat{W}_1, \widehat{W}_2$ are continuous functionals of strongly uniformly consistent estimators. This gives that $\widehat{W}_1 - W_1$ and $\widehat{W}_2 - W_2$ converge uniformly to zero almost everywhere. Furthermore, we will show that $\widehat{W}_1$ and $\widehat{W}_2$ are of bounded uniform sectional variation uniformly in $n$. Therefore the last two terms are of the form:

**Lemma 7.3** (Helly-Bray). *Let* $f_n, Z_n, Z \in (D[0,\tau], \|\cdot\|_\infty)$. *Assume* $\|f_n\|_\infty \to 0$, $\|f_n\|_v^* < M < \infty$, $\|Z_n - Z\|_\infty \to 0$. *Then* $\int f_n dZ_n \to 0$.

**Proof.** These terms are shown to converge to zero as follows. $\int f_n dZ_n = \int f_n d(Z_n - Z) + \int f_n dZ$. Apply integration by part to the first term so that we can bound it by $C\|Z_n - Z\|_\infty \|f_n\|_v^*$. For the second term we apply the Helly-Bray lemma 6.4. □

This proves the convergence to zero of the last two terms. Now, we have verified the required differentiability of $\Psi(P)$ at $P$. Application of the functional delta method provides us now with weak convergence of $\sqrt{n}(\Psi_n - \Psi)$.

Similarly, but easier, it is shown that the strong uniform consistency of $\widehat{W}_1, \widehat{W}_2$ and $P_{01}^n$, $P_{10}^n$, $P_{11}^n$ provides us with the strong uniform consistency of $\Psi_n$.

It remains to analyze the third term. We will do this by application of the functional delta-method for functionals of density estimators as stated in lemma 7.2 in the preceding section. For this we need uniformly consistent density estimators on $[0, \tau]$ which are of bounded uniform sectional variation and the integrated density estimator should be asymptotically equivalent with the empirical distribution function. These will be constructed in the next two subsections. The uniform consistency on $[0, \tau]$ requires an edge-correction at the edge of $[0, \tau]$. We will study this in the next subsection.

### 7.6.1 Uniformly consistent edge-corrected bivariate density estimators.

If we have a density which is uniformly continuous on $\mathbb{R}^2$, then necessary and sufficient conditions for strong uniform consistency of the kernel density estimator $f_n$ with kernel $K$ and bandwidth $h_n$ are: $h_n \to 0$, $(nh_n^2)/\log n \to \infty$ as $n \to \infty$, $K$ measurable w.r.t. Lebesque measure, $\int \mid K(t) \mid dt < \infty$, $\int K(t)dt = 1$ (see Bertrand-Retali, 1974, 1978). We will also assume that

$K$ has compact support within $[-1,1]^2$. Schuster (1985) shows how to get uniformly consistent estimators of univariate densities with support $[c,d]$, by using symmetrization techniques which brings one back to the problem of estimating a uniformly continuous density on $\mathbb{R}$. In van der Laan (1991) this method is generalized to the two dimensional case. Here we only discuss a method introduced by Richard Gill. From now on we will work with a symmetric kernel which satisfies the conditions mentioned above.

**Gill's method.** This method requires that $f$ also puts mass outside the rectangle $[0,\tau]$. Let $f_n$ be the kernel density estimator with bandwidth $h$. Let $\vec{h} = (h,h)$. Now, define the edge corrected kernel density estimator $f_n^*$ as follows: $f_n^*$ equals $f_n$ on $[\vec{h}, \tau - \vec{h}]$, and $f_n^*(x)$ gets the value of $f_n(x')$, where $x'$ is the closest point to $x$ on the edge of $[\vec{h}, \tau - \vec{h}]$. We will prove that $f_n^*$ is uniformly consistent.

We have:

$$\sup_{x\in[0,\tau]} \mid f_n^*(x) - f(x) \mid \leq \sup_{x\in[\vec{h},\tau-\vec{h}]} \mid f_n(x) - f(x) \mid + \sup_{x\in[\vec{h},\tau-\vec{h}]^c} \mid f_n(x') - f(x) \mid,$$

where the complement is taken within $[0,\tau]$. Consider the first term. $f_n$ uses here only data on $[0,\tau]$. Therefore, if there is mass outside $[0,\tau]$, then the data is indistinguishable from a uniformly continuous density which equals $f$ on $[0,\tau]$, but nicely bends down to zero outside $[0,\tau]$. Now, by Bertrand-Retali's result $f_n$ is uniformly consistent on $[\vec{h}, \tau - \vec{h}]$, which proves that the first term converges to zero.

For the second term we have:

$$\mid f_n(x') - f(x) \mid \leq \mid f_n(x') - f(x') \mid + \mid f(x') - f(x) \mid .$$

So the supremum over $[\vec{h}, \tau - \vec{h}]^c$ of the first term converges to zero by the uniform consistency of $f_n$ on $[\vec{h}, \tau - \vec{h}]^c$ and the supremum of the second term converges to zero by the uniform continuity of $f_\epsilon$.

## 7.6.2 Bivariate kernel density estimators: simultaneous uniform consistency, consistency of derivative and asymptotic normality of the integrated kernel density estimator.

Let $X_i$, $i = 1,\ldots,n$ be $n$ i.i.d. copies of a bivariate random variable $X = (X_1, X_2) \sim f$, where $f$ is a bivariate continuous density on $[0,\tau]$ w.r.t. the

Lebesque measure. Let

$$f_n(x_1, x_2) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h}, \frac{x_2 - X_{2i}}{h}\right)$$

be the bivariate kernel density estimator with kernel $K$.

We will now find conditions on $f$, $K$ and $h$ which provides us with a kernel density estimator which is uniformly consistent, which is of bounded uniform sectional variation uniformly in $n$ and for which the integrated density estimator is asymptotically equivalent with the empirical distribution. We know already the conditions for uniform consistency (see above): for the bandwidth we need $(nh^2)/\log(n) \to \infty$. We will find the conditions for each of the two remaining properties and then combine them in a lemma.

**Pointwise consistency of derivative of density.** Assume that $K$ has compact support and that $K^{1,1}(x) \equiv (d^2/dx_1dx_2)K(x_1, x_2)$ is continuous on $[0, \tau]$. Firstly, we will find conditions which guarantee that $\limsup \|f_n\|_v^* = O_P(1)$. By the triangle inequality we have

$$| f_n^{1,1}(x) - f^{1,1}(x) | \leq | f_n^{1,1}(x) - Ef_n^{1,1}(x) | + | Ef_n^{1,1}(x) - f^{1,1}(x) |, \quad (7.14)$$

where

$$f_n^{1,1}(x_1, x_2) \equiv \frac{d^2}{dx_1dx_2} f_n(x_1, x_2) = \frac{1}{nh^4} \sum_{i=1}^{n} K^{1,1}\left(\frac{x_1 - X_{1i}}{h}, \frac{x_2 - X_{2i}}{h}\right).$$

Firstly, we will study the second term. We have that

$$\begin{aligned} E\left(f_n^{1,1}(x_1, x_2)\right) &= \frac{1}{h^4} E\left(K^{1,1}\left(\frac{x_1 - X_{1i}}{h}, \frac{x_2 - X_{2i}}{h}\right)\right) \\ &= \frac{1}{h^4} \int K^{1,1}\left(\frac{x_1 - y_1}{h}, \frac{x_2 - y_2}{h}\right) f(y_1, y_2) dy_1 dy_2 \\ &= \frac{1}{h^2} \int K^{1,1}(z_1, z_2) f\left(x_1 - hz_1, x_2 - hz_2\right) dz_1 dz_2. \end{aligned}$$

We will say that $f \in C^k[0, \tau]$ if $f$ has $k$ derivatives in both coordinates and $f^{k,k}$ is continuous. If $f \in C^k[0, \tau]$, then we have the following Taylor expansion:

$$f\left(x_1 + hz_1, x_2 + hz_2\right) = \sum_{i \geq 0, j \geq 0, i+j \leq k} \frac{(hz_1)^i (hz_2)^j}{i!j!} f^{i,j}(x_1, x_2) + o(h^k).$$

Assume that $K^{1,1}$ satisfies the following *orthogonality conditions.*

$$\int K^{1,1}(z_1, z_2) z_1^i z_2^j dz_1 dz_2 = 0, \; i \geq 0, \; j \geq 0, \; 1 \leq i + j \leq k. \quad (7.15)$$

Then we have

$$\begin{aligned} Ef_n^{1,1}(x_1,x_2) &= f^{1,1}(x_1,x_2) + o\left(\frac{1}{h^2}h^k \int \mid K^{1,1}(z_1,z_2) \mid dz_1 dz_2\right) \\ &= f^{1,1}(x_1,x_2) + o(h^{k-2}), \end{aligned} \tag{7.16}$$

using that $\int \mid K^{1,1}(z_1,z_2) \mid dz_1 dz_2 < \infty$ because $K \in C^1[0,\tau]$. So if $f \in C^2[0,\tau]$, then $\mid Ef_n^{1,1}(x_1,x_2) - f^{1,1}(x_1,x_2) \mid = o(1)$, uniformly in $x \in [0,\tau]$.

Let's now consider the first term of (7.14). The variance of $f_n^{1,1}(x_1,x_2)$ is given by

$$\frac{1}{nh^8}\mathrm{Var}\left(K^{1,1}\left(\frac{x_1 - X_{1i}}{h}, \frac{x_2 - X_{2i}}{h}\right)\right),$$

where the variance of the $K^{1,1}$ term is bounded from above by

$$E\left(K^{1,1}\left(\frac{x_1 - X_{1i}}{h}, \frac{x_2 - X_{2i}}{h}\right)\right)^2 = O(h^2).$$

Here we use that we only have to integrate over a square with width $h$. So we conclude that if $f \in C^2[0,\tau]$, then

$$\sup_{x\in[0,\tau]} \mathrm{Var}\left(\mid f_n^{1,1}(x) - f^{1,1}(x) \mid\right) = O\left(\frac{1}{nh^6}\right) + o(1).$$

This tells us that if we choose $h$ so that $h_n n^{1/6} \to \infty$, then $f_n^{1,1}(x) \to f^{1,1}(x)$ a.s. for all $x$. Moreover, by the triangle inequality, $\mathrm{E}\left(\int \mid f_n^{1,1} - Ef_n^{1,1} \mid (x)dx\right) \to 0$ and (7.16) it follows also that

$$\limsup_{n\to\infty} \int \mid f_n^{1,1}(x) \mid dx = \int \mid f^{1,1}(x) \mid dx < \infty \text{ a.s.}$$

By doing similar calculation for the sections this provides us with: $\limsup \|f_n\|_v^* < \infty$.

**Weak convergence of the integrated kernel density estimator.** Let $F_n$ be the empirical distribution function and $k_n(z) \equiv 1/h^2 K(z)$. Then

$$\widetilde{F}_n(x) \equiv \int_0^x f_n(y)dy = F_n * k_n(x), \text{ the convolution of } F_n \text{ and } k_n.$$

Now,

$$\sqrt{n}(\widetilde{F}_n - F) = \sqrt{n}\,(F_n * k_n - F * k_n) + \sqrt{n}\,(F * k_n - F).$$

The first term is an empirical process indexed by smoothed indicators which clearly form a Donsker class and therefore empirical process theory provides us

immediately with weak convergence of the first term. The second term is the bias of which we have to take care. We have ($\int K(z)dz = 1$)

$$\begin{aligned}(F * k_n - F)(x) &= \int F(x-y)\frac{1}{h^2}K\left(\frac{y}{h}\right)dy - F(x) \\ &= \int (F(x_1 - hz_1, x_2 - hz_2) - F(x))\,K(z_1, z_2)dz_1dz_2.\end{aligned}$$

Assume $F \in C^k[0,\tau]$ and that $K$ satisfies (7.15). Then we have that

$$(F * k_n - F)(x) = o(h^k).$$

In other words we have to choose $k$ so that $\sqrt{n}o(h^k) \to 0$. However, for the bounded variation condition we needed that $h_n$ converges to zero slower than $n^{-1/6}$. So we need that $F \in C^4[0,\tau]$ and hence that $f \in C^3[0,\tau]$. This proves the following lemma:

**Lemma 7.4** *Let $X_i$ be n i.i.d. copies of a bivariate $X \sim f$, where $f \in C^3[0,\tau]$. Let $f_n$ be a bivariate kernel density estimator with kernel $K$ and bandwidth $h_n$, as defined above and let $\widetilde{F}_n = \int_0^x f_n(y)dy$ be the integrated kernel density estimator.*

*Assume that $K \in C^1[0,\tau]$, $K$ satisfies the orthogonality conditions (7.15) for $k = 4$ and $\int K(t)dt = 1$.*

- *If $h_n \to 0$, $nh_n^2/\log(n) \to \infty$, then $f_n$ is uniformly consistent.*
- *If $h_n n^{1/6} \to \infty$, then $f_n^{1,1}(x) \to f^{1,1}(x)$ a.s. for all $x \in [0,\tau]$ and $\limsup_{n\to\infty} \|f_n\|_v^* \leq M < \infty$ a.s.*
- *If $\sqrt{n}o(h_n^4) \to 0$, then the integrated density estimator is asymptotically equivalent with the empirical distribution function.*

*Consequently, if $h_n = n^{-1/7}$, then $\|f_n - f\|_\infty \to 0$ a.s, $\limsup_n \|f_n\|_v^* = O(1)$ a.s. and $\sqrt{n}(\widetilde{F}_n - F) \stackrel{D}{\Longrightarrow} Z$ where $Z$ is the $F$-Brownian bridge.*

If we choose a $K$ of the form $K(z_1, z_2) = g(z_1)g(z_2)$ for a certain differentiable $g : \mathbb{R} \to \mathbb{R}$, then it is trivial to construct a kernel which satisfies the orthogonality conditions (7.15) for $k = 4$.

**Remark.** In our application we have that $f_n$ is an edge-corrected kernel density estimator. In the preceding section we already showed that under the same assumptions as with the uncorrected kernel density estimator it will be uniformly consistent on $[0,\tau]$. It is straightforward to verify that we can also apply the lemma to the edge corrected $f_n$, we will not go into these technical details.

### 7.6.3 Application of the functional delta-method for density estimators.

Here, we will prove weak convergence of $\sqrt{n}\int_0^{\tau_1}\int_{t_2}^{\tau_2}\left(\widehat{W}_1 - W_1\right)(t_1, y_1, y_2)dP_{01}(y_1, y_2)$ as random elements of the cadlag function space $D([0,\tau])$ endowed with the supremum norm and the Borel sigma-algebra, by application of the delta-method lemma 7.2. Let $K$ be a kernel satisfying the properties as mentioned in Lemma 7.4. We will now define $\widehat{W}_1$. Firstly, we define

$$\begin{aligned} F_N(t_1,t_2) &\equiv P(\widetilde{T}_1 \le t_1, \widetilde{T}_2 \le t_2, D_1 = 1, D_2 = 1) \\ F_Y(t_1,t_2) &\equiv P(\widetilde{T}_1 > t_1, \widetilde{T}_2 \le t_2, D_2 = 1). \end{aligned}$$

So $F_N$ is the subdistribution of the doubly uncensored observations and $F_Y$ is the subdistribution of the in the second coordinate uncensored observations. Moreover, we define their derivatives w.r.t. the second coordinate.

$$\begin{aligned} N(y_1,y_2) &\equiv \frac{d}{dy_2}F_N(y_1,y_2) \\ Y(y_1,y_2) &\equiv \frac{d}{dy_2}F_Y(y_1,y_2). \end{aligned}$$

These are densities which appear in $W_1$. We will estimate them with kernel density estimators. In fact, we will use edge corrected kernel density estimators $N_n, Y_n$ which converges uniformly on the rectangle $[0,\tau]$. This means that we make the usual kernel density estimator constant at distance $h$ from the edge of $[0,\tau]$, as discussed in section 6.1. In order not to complicate the notation we will suppress this fact in the notation. We denote these density estimators with $N_n$ and $Y_n$:

$$\begin{aligned} N_n(y_1,y_2) &\equiv \int_0^{y_1}\frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{u-\widetilde{T}_{1i}}{h}, \frac{y_2-\widetilde{T}_{2i}}{h}\right)I(D_{i1}=1, D_{i2}=1)du \\ Y_n(y_1,y_2) &\equiv \int_{y_1}^{\tau_1}\frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{u-\widetilde{T}_{1i}}{h}, \frac{y_2-\widetilde{T}_{2i}}{h}\right)I(D_{i2}=1)du. \end{aligned}$$

We define $S(\cdot \mid y_2)$ as the survival function of $T_1$ given $T_2 = y_2$. Therefore $S(\cdot \mid y_2)$ has the well known product limit representation in terms of its conditional hazard $\int_0^{t_1} N(du, y_2)/Y(u-, y_2)$ (Gill and Johansen, 1990):

$$S(y_1 \mid y_2) \equiv P(T_1 > y_1 \mid T_2 = y_2)$$

$$= \prod_{(0,y_1]} \left(1 - \frac{P(T_1 \in du \mid T_2 = y_2)}{P(T_1 \geq u \mid T_2 = y_2)}\right)$$
$$= \prod_{(0,y_1]} \left(1 - \frac{N(du, y_2)}{Y(u-, y_2)}\right).$$

Of course, a natural estimator of $S$ is obtained by replacing $N, Y$ by $N_n, Y_n$:

$$S_n(y_1 \mid y_2) \equiv \prod_{(0,y_1]} \left(1 - \frac{N_n(du, y_2)}{Y_n(u-, y_2)}\right).$$

We can now define $\widehat{W}_1$:

$$\begin{aligned} W_1(t_1, y_1, y_2) &\equiv \frac{S(t_1 \vee y_1 \mid y_2)}{S(y_1 \mid y_2)} \\ \widehat{W}_1(t_1, y_1, y_2) &\equiv \frac{S_n(t_1 \vee y_1 \mid y_2)}{S_n(y_1 \mid y_2)}. \end{aligned} \tag{7.17}$$

Because we are also concerned with the probabilistic behaviour of the integrated density estimators, which play the role of $\widetilde{F}_n$ in lemma 7.2, we also need notation for them.

$$\begin{aligned} \widetilde{F}_{Yn}(t_1, t_2) &\equiv \int_0^{t_2} Y_n(t_1, y_2) dy_2 \\ \widetilde{F}_{Nn}(t_1, t_2) &\equiv \int_0^{t_2} N_n(t_1, y_2) dy_2. \end{aligned}$$

So these are smoothed empirical distribution functions $F_{Nn}$, $F_{Yn}$ of $F_N$ and $F_Y$, respectively. Now, we define the following mappings:

$$\begin{aligned} \Phi_1 : D([0,\tau])^2 &\to D([0,\tau]) : \\ (F_N, F_Y) &\mapsto \int_0^{y_1} \frac{N(du, y_2)}{Y(u, y_2)} \equiv \Lambda(y_1 \mid y_2) \\ \Phi_2 : D([0,\tau]) &\to D([0,\tau]) : \\ \Lambda &\mapsto \int_0^{\tau_1} \int_{t_2}^{\tau_2} \prod_{[y_1, t_1 \vee y_1]} (1 - \Lambda(du \mid y_2))\, dP_{01}(y_1, y_2) \\ &= \int_0^{\tau_1} \int_{t_2}^{\tau_2} W_1(t_1, y_1, y_2) dP_{01}(y_1, y_2). \end{aligned}$$

Define $\Lambda_n(y_1 \mid y_2) \equiv \int_0^{y_1} N_n(du, y_2)/Y_n(u-, y_2)$. Finally, we can define the $\Phi$ to which we apply the refined functional delta method lemma 7.2.

$$\Phi \equiv \Phi_2 \circ \Phi_1 \ : D([0,\tau])^2 \to D([0,\tau]). \tag{7.18}$$

We want to prove weak convergence of

$$\sqrt{n}\left(\Phi(\widetilde{F}_{Nn}, \widetilde{F}_{Yn}) - \Phi(F_N, F_Y)\right)(t_1, t_2)$$

which equals

$$\sqrt{n}\int_0^{\tau_1}\int_{t_2}^{\tau_2}\left(\widehat{W}_1 - W_1\right)(t_1, y_1, y_2)dP_{01}(y_1, y_2).$$

We apply lemma 7.2 with ($F_n$ is the empirical df. of $F$):

$$\begin{aligned} F &= (F_N, F_Y) \\ F_n &= (F_{Nn}, F_{Yn}) \\ \widetilde{F}_n &= (\widetilde{F}_{Nn}, \widetilde{F}_{Yn}) \\ f_n &= \left(\frac{d}{dy_1}N_n(y_1, y_2), \frac{d}{dy_1}Y_n(y_1, y_2)\right) \\ f &= \left(\frac{d}{dy_1}N(y_1, y_2), \frac{d}{dy_1}Y(y_1, y_2)\right). \end{aligned}$$

According to lemma 7.2 we can separate the analysis in a purely probabilistic part and a purely analytical part. For the probabilistic conditions we just have to apply lemma 7.4 with the $f, f_n, F, \widetilde{F}_n$ above. This provides us with the following proposition:

**Proposition 7.1** *Assume $f, g \in C^3[0,\tau]$, $K \in C^1[0,\tau]$, $K$ satisfies the orthogonality conditions (7.15) for $k = 4$ and $\int K(t)dt = 1$. Moreover assume that $h_n = n^{-1/7}$. Then*
**1.** *$m_n(y_1, y_2) \equiv (d/dy_1 N_n(y_1, y_2), d/dy_1 Y_n(y_1, y_2))$ is a uniformly consistent estimator of $m \equiv (d/dy_1 N(y_1, y_2), d/dy_1 Y(y_1, y_2))$.*
**2.** *$m_n^{1,1}(x) \to m^{1,1}(x)$ a.s. for all $x \in [0,\tau]$ and $\limsup_{n\to\infty} ||m_n||_v^* = O(1)$ a.s.*
**3.** *$\widetilde{M}_n(\cdot) \equiv \int_0^{(\cdot)} m_n(x)dx$ is asymptotically equivalent with the empirical distribution $M_n = (F_{Nn}, F_{Yn})$ of $M \equiv (F_N, F_Y)$.*

Notice that if $f, g \in C^3[0,\tau]$, then $m \in C^3[0,\tau]$. Moreover notice that $m_n$ is the kernel density estimator of $m$. Therefore, this proposition is an immediate corollary of lemma 7.4.

### 7.6.4 Analytical part of the analysis.

We will now prove the differentiability property for $\Phi$ as stated in lemma 7.2. We can consider sequences $N_n, Y_n$ with: $N_n \to N$, $Y_n \to Y$, $||N_n||_v^* = O(1)$, $||Y_n||_v^* = O(1)$, $H_Y^n \equiv \sqrt{n}(\widetilde{F}_Y^n - F_Y) \to H_Y$, $H_N^n \equiv \sqrt{n}(\widetilde{F}_N^n - F_N) \to H_N$, where

$F_Y(y_1, y_2) = \int_0^v Y(y_1, v)dv$ and similarly for $F_N$, $\widetilde{F}_Y^n$ and $\widetilde{F}_N^n$. Firstly, we will consider the difference $W_{1n} - W_1$.

Recall the definition of $\Lambda(du \mid y_2) = N(du, y_2)/Y(u-, y_2)$ and its empirical version $\Lambda_n(du \mid y_2) = N_n(du, y_2)/Y_n(u-, y_2)$. We can consider $S(\cdot \mid y_2)$ also as a measure: $S((a,b] \mid y_2) \equiv P(T_1 \in (a,b] \mid T_2 = y_2)$. Recalling its product integral representation w.r.t. $\Lambda$ it follows that:

$$\begin{aligned} S((a,b] \mid y_2) &= \prod_{(a,b]} (1 - \Lambda(du \mid y_2)) \\ S_n((a,b] \mid y_2) &= \prod_{(a,b]} (1 - \Lambda_n(du \mid y_2)). \end{aligned}$$

Now, we have that

$$\begin{aligned} (W_{1n} - W_1)(t_1, y_1, y_2) &= \prod_{[y_1, t_1 \vee y_1]} (1 - \Lambda_n(du \mid y_2)) - \prod_{[y_1, t_1 \vee y_1]} (1 - \Lambda_1(du \mid y_2)) \\ &= S_n([y_1, t_1 \vee y_1] \mid y_2) - S([y_1, t_1 \vee y_1] \mid y_2). \end{aligned}$$

Duhamel's equation for the product integral (see Gill and Johansen, 1990, equation 42) tells us that:

$$(S_n - S)([y_1, t_1 \vee y_1] \mid y_2) = \int_{y_1}^{t_1 \vee y_1} (\Lambda_n - \Lambda)(du \mid y_2) S_n([y_1, u) \mid y_2) S((u, t_1 \vee y_1] \mid y_2).$$

So we need to study the difference $(\Lambda_n - \Lambda)(du \mid y_2) = (\Phi_1(N_n, Y_n) - \Phi_1(N, Y))(du, y_2)$. We have that

$$(\Lambda_n - \Lambda)(du \mid y_2) = D\Phi_1^n(du, y_2) + R_1^n(du, y_2),$$

where

$$\begin{aligned} D\Phi_1^n(du, y_2) &\equiv \frac{(N_n - N)(du, y_2)}{Y(u, y_2)} + \frac{N(du, y_2)(Y - Y_n)(u, y_2)}{Y^2(u, y_2)} \\ R_1^n(du, y_2) &\equiv \frac{(N_n - N)(du, y_2)(Y - Y_n)(u, y_2)}{Y^2(u, y_2)} + \frac{N_n(du, y_2)(Y - Y_n)^2(u, y_2)}{Y Y_n^2(u, y_2)}. \end{aligned}$$

This is trivially verified.

Furthermore, we have by using the multiplicativity of the product integral that

$$S_n([y_1, u) \mid y_2) S((u, t_1 \vee y_1] \mid y_2) = (S_n - S)([y_1, u) \mid y_2) S((u, t_1 \vee y_1] \mid y_2)$$

$$
\begin{aligned}
& + S([y_1, u) \mid y_2) S((u, t_1 \vee y_1] \mid y_2) \\
= \; & W_1(t_1, y_1, y_2) + R_2^n(u, t_1, y_1, y_2),
\end{aligned}
$$

where

$$R_2^n(u, t_1, y_1, y_2) \equiv (S_n - S)([y_1, u) \mid y_2) S((u, t_1 \vee y_1] \mid y_2).$$

So we have that $(W_{1n} - W_1)(t_1, y_1, y_2)$ equals

$$\int_{y_1}^{t_1 \vee y_1} (D\Phi_1^n(du, y_2) + R_1^n(du, y_2)) (W_1(t_1, y_1, y_2) + R_2^n(u, t_1, y_1, y_2)).$$

This provides us with:

$$
\begin{aligned}
& \sqrt{n} (\Phi(N_n, Y_n) - \Phi(N, Y))(t) \\
= \; & \sqrt{n} \int_0^{\tau_1} \int_{t_2}^{\tau_2} (W_{1n} - W_1)(t_1, y_1, y_2) dP_{01}(y_1, y_2) \\
= \; & \sqrt{n} \int_0^{t_1} \int_{t_2}^{\tau_2} \int_{y_1}^{t_1} D\Phi_1^n(du, y_2) W_1(t_1, y_1, y_2) dP_{01}(y_1, y_2) + \text{ Rem.} \qquad (7.19)
\end{aligned}
$$

The expression (7.19) consists of four terms of which three involve $R_1^n$ or $R_2^n$ or both. These three terms form the remainder and are shown to converge to zero (below). Firstly, we will be concerned with the main term which is linear in $N_n - N$ and $Y_n - Y$.

Because $N, Y, N_n, Y_n, W_1$ are of bounded uniform sectional variation, integrals with respect to $N, Y, N_n, Y_n, W_1$ are well defined. The terms are of the form $\int (N_n - N) G dx$ for a function $G$ which involves $Y, N$ in numerator and $Y$ in denominator. By using integration by parts they become of the form $\int (\widetilde{F}_N^n - F_N) dG$ (and similar one dimensional integrals over sections). Because $N, Y$ are of bounded uniform sectional variation and $Y$ is uniformly bounded away from zero by assumption 3 (that $f, g$ are strictly positive on $[0, \tau+\epsilon] \backslash [0, \tau]$, see introduction) this $G$ is of bounded uniform sectional variation (lemma 1.5). Therefore, we can linearize the expression above in $H_Y^n, H_N^n$.

After having linearized in $H_Y^n$, $H_N^n$ and using that $H_Y^n$ and $H_N^n$ converge in supremum norm to $H_Y, H_N$, it is trivial to see that this term converges in supremum norm to an expression linear in $H_N$ and $H_Y$: just bound terms of the form $\int (H_N^n - H_N) dG \leq \|H_N^n - H_N\|_\infty \|G\|_v^*$ and use that we know that $\|H_N^n - H_N\|_\infty$ converges to zero and that $G$ is of bounded uniform sectional variation because $N$ and $Y$ are. This expression is written down in van der Laan (1991) and we will denote it with $d\Phi(N, Y)(H_N, H_Y)$.

**Remainder.** Because $N_n, N, Y_n, Y$ are of uniformly bounded uniform sectional variation and $Y_n$ is uniformly bounded away from zero (because $Y$ is uniformly bounded away from zero and $Y_n \to Y$) $y \to R_2^n(u, t_1, y)$ is of bounded uniform sectional variation uniformly in $n$ and $(u, t)$. The product integral is a continuous functional in $(N, Y)$ with respect to the supremum norm (see Gill and Johansen 1990). Therefore $R_2^n$ converges uniformly to zero. Here we use for the first time that $N_n - N \to 0$ and $Y_n - Y \to 0$.

All three terms are in fact dealt with in the same way. These terms have the following structure:

$$\begin{aligned}
&\sqrt{n}\int_0^{\tau_1}\int_{t_2}^{\tau_2}\int_{y_1}^{t_1}(Y_n - Y)^2(s, y_2)m_n(s, y_1, y_2)dsdy_1dy_2 \\
&= \sqrt{n}\int_0^{\tau_1}\int_{y_1}^{t_1}((F_Y^n - F_Y)(s,\cdot)(Y_n - Y)(s,\cdot)m_n(s, y_1, \cdot))((t_2, \tau_2])dsdy_1 \\
&-\sqrt{n}\int_0^{\tau_1}\int_{y_1}^{t_1}\int_{t_2}^{\tau_2}(F_Y^n - F_Y)(s, y_2)((Y_n - Y)(s,\cdot)m_n(s, y_1, \cdot))(dy_2)dsdy_1 \\
&\longrightarrow 0 \text{ w.r.t. the supnorm.}
\end{aligned}$$

Here $m_n$ involves $N_n, N, Y, Y_n$, which are of uniformly bounded uniform sectional variation. The convergence to zero in supnorm of the first term is trivial. The second integral is of the structure $\int Z_n dF_n$, where $\|Z_n - Z\|_\infty \to 0$ (for $\|H_Y^n - H_Y\|_\infty \to 0$) and $\|F_n\|_v^* = O(1)$ (for $\|(Y_n, m_n)\|_v^* = O(1)$) and $\|F_n - F\|_\infty \to 0$ (for $\|Y_n - Y\|_\infty \to 0$). Therefore convergence to zero follows from the Helly-Bray lemma 7.3 with $Z_n \equiv \sqrt{n}(F_Y^n - F_Y)$ and $F_n(\cdot) \equiv (Y_n - Y)(s,\cdot)m_n(s, y_1, \cdot)$.

Consider the term $\sqrt{n}\int D\Phi_1^n(N_n - N, Y_n - Y)R_2^n dP_{01}$. Just as above we linearize this in $H_N^n$ and $H_Y^n$. Then we obtain a term of the form $\int Z_n dF_n$ where $\|Z_n - Z\|_\infty \to 0$ (for $\|H_N^n - H_N\|_\infty \to 0$), $\|F_n\|_v^* = O(1)$ (for $\|R_2^n\|_v^* = O(1)$) and $\|F_n - F\|_\infty \to 0$ (for $\|R_2^n - R_2\|_\infty \to 0$). The Helly-Bray lemma 7.3 tells us now that this term converges to zero.

Consider now the term $\sqrt{n}\int R_1^n W_1 dP_{01}$. Again, by doing integration by parts we can linearize in $H_N^n$ (or $H_Y^n$) and these will be integrated with respect to a measure which involves $d(Y_n - Y)$ (or $d(N_n - N)$). Therefore this term is again of the form $\int Z_n dF_n$ where $\|Z_n - Z\|_\infty \to 0$ (for $\|H_N^n - H_N\|_\infty \to 0$), $\|F_n\|_v^* = O(1)$ (for $\|(N_n, Y_n, N, Y)\|_v^* = O(1)$) and $\|F_n - F\|_\infty \to 0$ (for $\|(N_n, Y_n) - (N, Y)\|_\infty \to 0$). The Helly-Bray lemma 7.3 tells us now that this term converges to zero.

The term $\sqrt{n}\int R_1^n R_2^n dP_{01}$ is proved to converge to zero in the same way as $\sqrt{n}\int R_1^n W_1 dP_{01}$. This completes the differentiability proof.

We conclude that once we have arranged that we only have to verify the differentiability property in lemma 7.2 for sequences with a consistent density which is of uniformly (in $n$) bounded uniform sectional variation, then integration by parts and the Helly-Bray technique are the only ingredients one needs in order to prove that the remainder converges in supremum norm to zero. We have proved the analytical condition of lemma 7.2. Application of lemma 7.2 provides us now with weak convergence of $\sqrt{n}\int(W_{1n} - W_1)dP_{01}$ to the Gaussian process $d\Phi(N, Y)(H_N, H_Y)$.

This completes the proof of all four ingredients as stated in section 3 and hence the proof of theorem 7.1.

# References


P.K. Andersen, Ø. Borgan, R.D. Gill and N. Keiding (1993),*Statistical models based on counting processes*, Springer, New York.

D.M. Bakker (1990), *Two nonparametric estimators of the survival function of bivariate right censored observations*, Report BS-R9035, Centre for mathematics and computer science, Amsterdam.

M. Bertrand-Retali (1974), Convergence uniforme d'un estimateur d'une densité de probabilité dans $R^s$, *Comptes Rendus de l'Ac. Sci. Paris* **278** 451–453.

M. Bertrand-Retali (1978), Convergence uniforme d'un estimateur de la densité par la méthode du noyau, *Rev. Roumaine Math. Pures et Appl.* **23** 361–385.

P.J. Bickel and D.A. Freedman (1981), Some asymptotic theory for the bootstrap, *Ann.Stat.* **9** 1196–1217.

P.J. Bickel, A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.

P.J. Bickel and Y. Ritov (1992), *Efficient estimation using both direct and indirect observations*, technical report, Department of Statistics, University of California, Berkeley.

P. Billingsley (1968), *Convergence of Probability Measures*, Wiley, New York.

M.D. Burke (1988), Estimation of a bivariate survival function under random censorship, *Biometrika* **75**, 379–382.

G. Campbell, A. Földes (1982), Large-sample properties of non-parametric bivariate estimators with censored data, pp. 103–122 in: *Nonparametric Statistical Inference*, B.V. Gnedenko, M.L. Puri and I. Vincze, eds., North Holland, Amsterdam.

M.N. Chang (1990), Weak convergence of a self-consistent estimator of the survival function with doubly censored data, *Ann. Statist.* **18** 391–404.

M.N. Chang and G. Yang (1987), Strong consistency of a nonparametric estimator of the survival function with doubly censored data, *Ann. Statist.* **15** 1536–1547.

D.M. Dabrowska (1988), Kaplan Meier Estimate on the Plane, *Ann. Statist.* **16**, 1475–1489.

D.M. Dabrowska (1989), Kaplan Meier Estimate on the Plane: Weak Convergence, LIL, and the Bootstrap, *J. Multivar. Anal.* **29**, 308–325.

A.P. Dempster, N.M. Laird and D.B. Rubin (1977), Max-

imum likelihood from incomplete data via the EM-algorithm, *J. Roy. Statist. soc. sec*, **39** 1–38.

R.M. Dudley (1966), Weak convergence of probabilities on nonseparable metric spaces, and empirical measures on Euclidean spaces, *Illinois J.Math.* **10**, 109–126.

R.M. Dudley (1985), An extended Wichura theorem, definitions of Donsker classes, and weighted empirical distributions. *Lecture notes in Math.* **1153**, 141–178. Springer Verlag, New York.

R.M. Dudley
(1989), *Real Analysis and Probability*, Wadsworth & Brooks/Cole, Pacific Grove, California.

B. Efron (1967), The two sample problem with censored data, *Proc. 5th. Berkeley Symp. on Math. Statist. Prob*, 831–853, Berkeley, University of California Press.

J.H.J. Einmahl (1987), *Multivariate empirical processes*, CWI tract **32**, Centre for Mathematics and Computer Science, Amsterdam.

T.M. Flett (1980), *Differential Analysis*, Cambridge University Press, Cambridge.

R.D. Gill (1989), Non-and Semi-parametric Maximum Likelihood Estimators and the von Mises Method (Part 1), *Scand. J. Statist.* **16**, 97–128.

R.D. Gill (1991), *Non-and Semi-parametric Maximum Likelihood Estimators and the von Mises Method (Part 2)*, Revised, joint with A.W. van der Vaart (1993) (submitted to *Scand. J. Statist*, preprint nr. 664, Dept of Mathematics, Utrecht, the Netherlands.

R.D. Gill and S. Johansen (1990), A survey of product integration with a view towards application in survival analysis, *Ann. Statist.* **18**, 1501 - 1555.

R.D. Gill (1992), Multivariate survival analysis, *Theory Prob. Appl.* **37**, 19–36 and 307–328.

R.D. Gill (1993), *Lectures on survival analysis*, In: D. Bakry, R.D. Gill and S. Molchanov, École d'Été de Probabilités de Saint Flour XXII–1992, ed. P. Bernard; Springer Lecture Notes in Mathematics.

R.D. Gill, M.J. van der Laan and J.M. Robins (1995), *Coarsening at Random*, preprint Department of Mathematics, Utrecht, the Netherlands.

R.D. Gill, M.J. van der Laan and J.A. Wellner (1993), *Inefficient estimators of the bivariate survival function for three models*, to appear in *Annales de L'I.H.P. Prob. et Stat.*:

R.D. Gill, M.J. van der Laan & B.J. Wijers (1995), *Laslett's Line-segment Problem*, to appear in *Bernoulli.*

E. Giné and J. Zinn (1990), Bootstrapping general empirical measures, *Ann. Probability* **18**, 851–869.

P.E. Greenwood and W. Wefelmeyer (1991), Efficient estimating equations for nonparametric models, pp. 107–141 in *Statistical Inference in Stochastic Processes*, Edited by N.U. Prabhu and I.V. Basawa. Marcel Dekker.

P. Groeneboom (1991), *Nonparametric maximum likelihood estimators for interval censoring and deconvolution*, Technical report nr. **378**, Department of Statistics Stanford University, California.

P. Groeneboom and J.A. Wellner (1992), *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser verlag.

M.G. Gu and C.H. Zhang (1993), Asymptotic properties of self-consistent estimators based on doubly censored data, *Ann. Math. Statist.* **21** 611–614.

J.A. Hanley and M.N. Parnes (1983), Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics* **39**, 129–139.

D.F. Heitjan (1993), Ignorability and coarse data: some biomedical examples, *Biometrics* **49**, 1099–1109.

D.F. Heitjan and D.B. Rubin (1991), Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.

T.J. Hildebrandt (1963), *Introduction to the theory of integration*, Academic Press, New York and London.

T.H. Hildebrandt and L.M. Graves (1927), Implicit functions and their differentials in general analysis, *Trans. Amer. Math. Soc.* **29** 127–153.

J. Hoffmann-Jørgensen (1984), *Stochastic processes on Polish Spaces*, Unpublished manuscript.

I.A. Ibragimov and R.Z. Has'minskii (1983), On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter, *Lecture Notes in Mathematics* **1021** 195-229, Springer-Verlag, New York. *Theor. Prob. Applic.* **27** 551–562.

M. Jacobsen and N. Keiding (1994) , *Coarsening at random in general sample spaces and random censoring in continuous time*, preprint, Institute of Math. Stat. and Dept. of Biostatistics, University of Copenhagen.

L.V. Kantorovich and G.P. Akilov (1982), *Functional Analysis*, translated by Howard L. Silcock, New York, Permagon Press.

S. Karlin (1981), *A second course in stochastic processes*, Academic Press.

J. Kiefer and J. Wolfowitz (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.* **27**, 887–906.

C.A.J. Klaassen (1987), Consistent estimation of the influence function of lo-

cally asymptotically linear estimators, *Ann. Math. Statist.* **15**, 1548–1562.

G.M. Laslett (1982), The survival curve under monotone density constraints with applications to two-dimensonal line segment processes, *Biometrica* **69** 153–160.

M.J. van der Laan (1990), *Dabrowska's multivariate product limit estimator and the delta-method*, Master's Thesis, Department of Mathematics, Utrecht, the Netherlands.

M.J. van der Laan (1991), *Analysis of Pruitt's estimator of the bivariate survival function*, Preprint nr. 648, Department of Mathematics, Utrecht, the Netherlands.

M.J. van der Laan (1992), *Efficient Estimator of the Bivariate Survival Function for Right Censored Data*, Technical Report No. 337, Department of Statistics, University of California, Berkeley.

M.J. van der Laan (1993a), *General identity for linear parameters in convex models with application to efficiency of the (NP)MLE*, Preprint nr. 765, Department of Mathematics, Utrecht, the Netherlands. To appear in *Ann. Statist.*

M.J. van der Laan (1993b), *Efficient estimator of the bivariate survival function and repairing NPMLE*, Preprint nr. 788, Department of Mathematics, Utrecht, the Netherlands. To appear in *Ann. Statist.*

M.J. van der Laan (1993c), *Efficiency of the NPMLE in the line-segment problem*, Preprint nr. 773, Department of Mathematics, Utrecht, the Netherlands. To appear in *Scand. J. Statist.*

M.J. van der Laan (1993d), *Efficient and Inefficient Estimation in Semiparametric Models*, ISBN 90-393-0339-8, Department of Mathematics, Utrecht, the Netherlands.

M.J. van der Laan (1994a), Modified EM-estimator of the bivariate survival function, *Mathematical Methods of Statistics*, **3**, pp. 213-243.

M.J. van der Laan (1994b), *Proving efficiency of NPMLE and identities*, Technical report **44**, Group in Biostatistics, University of California, Berkeley.

M.J. van der Laan (1995), *Efficiency of the NPMLE in a general class of missing data models*, Submitted to *Mathematical Methods of Statistics.*

N.A. Langberg and M. Shaked (1982), On the identifiability of multivariate life distribution functions, *Ann. Probab.* **10**, 773–779.

M. Loève (1955), *Probability Theory*, van Nostrand, New York.

C.L. Mallows (1972), A note on asymptotic joint normality, *Ann. Stat.* **43** 508–515.

I. Meilijson (1989), A fast improvement to the EM-algorithm. *J. Roy. Statist. Soc. Sec.*, **51** Series B, 127–138.

A. Muñoz (1980), *Non-parametric estimation from censored bivariate observations*, Technical Report, Stanford University.

G. Neuhaus (1971), On weak convergence of stochastic processes with multidimensional time parameter, *Ann. Math. Statist.* **42** 1285–1295.

K.R. Parthasarathy (1967), *Probability Measures on Metric Spaces*, Academic Press, New York.

D. Pollard (1990), *Empirical Processes: Theory and appliacions*, Regional conference series in probability and statistics **2**, Inst. Math. Statist., Hayward, California.

R.L. Prentice and J. Cai (1992a), Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495–512.

R.L. Prentice and J. Cai (1992b), Marginal and conditional models for the analysis of multivariate failure time data. Klein, J.P. and Goel, P.K., editors, *Survival Analysis State of the Art.* Kluwer, Dordrecht.

R.C. Pruitt (1991a), On negative mass assigned by the bivariate Kaplan-Meier estimator, *Ann. Stat.* **19**, 443–453.

R.C. Pruitt (1991b), *Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis*, Technical Report nr. 543, University of Minnesota.

R.C. Pruitt (1993), Small sample comparisons of six bivariate survival curve estimators, *J. Statist. Comput. Simul.*, **45** 147–167.

J.A. Reeds (1976), *On the definition of von Mises functionals* (Ph.D. thesis), Dept. of Statistics, Harvard University, Research Report S-44.

Y. Ritov (1991), *Estimating the survival function in presence of covariates,* to appear in Ann. Statist.

E.F. Schuster (1985), Incorporating support constraints into nonparametric estimators of densities, *Commun. Statist.-Theor. Meth.*, **14(5)** 1123–1136.

G. Shorack and J.A. Wellner (1986), *Empirical processes and its applications*, Wiley, New York.

B.W. Silverman (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

D. Stoyan (1987), *Stochastic geometry and its applications*, New York, Wiley.

W-Y. Tsai, S. Leurgans and J. Crowley (1986), Nonparametric estimation of a bivariate survival function in the presence of censoring, *Ann. Statist.* **14** 1351–1365.

B.W. Turnbull (1976), The empirical distribution with arbitrarily grouped cen-

sored and truncated data, *J.R. Statist. Soc.* **B38**, 290–5.

A.W. van der Vaart (1988), *Statistical Estimation in Large Parameter Spaces.*, CWI tract **44**, Centre for Mathematics and Computer Science, Amsterdam.

A.W. van der Vaart (1991), Efficiency and Hadamard differentiable functionals, *Scand. J. Statist*, **18**, 63-75.

A.W. van der Vaart (1992a), *Maximum likelihood estimation with partially censored data*, technical report, Department of Mathematics, Free University of Amsterdam.

A.W. van der Vaart (1992b), Efficiency of infinite dimensional M-estimators, submitted to *Statistica Neerlandica.*

A.W. van der Vaart, J.A. Wellner (1989), *Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory, with application to convolution and asymptotic minimax theorems*, Tech. Report 157. Dept. of Statistics, University of Washington, Seattle.

A.W. van der Vaart and J.A. Wellner (1995), *Weak convergence and empirical processes.* IMS Lecture Notes-Monograph Series.

Y. Vardi and C. Zhang (1992), Large sample study of empirical distributions in a random multiplicative model, *Ann. Math. Stat.* **20** 1022–1040.

J.A. Wellner (1982), Asymptotic optimality of the product limit estimator, *Ann. Statist.* **10**, 595–602.

J.A. Wellner (1992), Empirical processes in action, *Int. Statist. Rev.* **60**, 247–270.

J.A. Wellner (1994), Covariance formulas via marginal martingales. *Statistica Neerlandica.*

J.A. Wellner (1993), *The delta-method and the bootstrap*, preprint, Department of Statistics, University of Washington.

M.J. Wichura (1968), On the weak convergence of non-Borel probabilities on a metric space, *Ph.D. dissertation*, Columbia University.

B.J. Wijers (1991), *Consistent nonparametric estimation for a one-dimensional line segment process observed in an interval*, preprint Nr. 683, Department of Mathematics, Utrecht, the Netherlands, to appear in *Scand. J. Statist.*.

B.J. Wijers (1994), *Nonparametric estimation for a windowed line-segment process*, thesis, ISBN 90-393-0706-7, Department of Mathematics, Utrecht, the Netherlands.

C.F.J. Wu (1983), On the convergence of the EM-algorithm, *Ann. Stat.* **11** 95–103.

# Notation

$F \ll G$: $F$ is absolutely continuous w.r.t. $G$, page 17.
$F \equiv G$: $F \ll G$ and $G \ll F$.
$\text{supp}(F)$: the support of $F$.
$L_P$: the law under $P$, page 20
$EX$: the expectation of the random variable $X$, page 19
$\text{Var}(X)$: the variance of the random variable $X$, page 19
$f(a,b]$: generalized difference of $f : \mathbb{R}^d \to \mathbb{R}$ over the rectangle $(a,b] \subset \mathbb{R}^d$, page 13
$\|\cdot\|_v$: variation norm, page 13
$\|\cdot\|_v^*$: uniform sectional variation norm, page 13
$D[0,\tau]$: space of multivariate cadlag functions, page 13
$(B(K), \|\cdot\|_\infty)$: space of uniformly bounded functions on $K$, page 68
**Notation for weak convergence theory.**
$Pf = \int f dP$, page 8
$P^*$: outer probability, page 8
$P^*f$: outer expectation, page 8
$C_b(D)$: the class of bounded real valued continuous functions on $D$, page 8
$X_n \xrightarrow{D} X_0$: weak convergence, page 8
**Notation for empirical process theory.**
$P_n$: the empirical distribution, page 4
$\mathcal{F}$: a class of measurable real valued functions, page 8
$\|\cdot\|_{\mathcal{F}}$: supnorm over $\mathcal{F}$, page 8
$l^\infty(\mathcal{F})$: the space of uniformly bounded real valued functions on $\mathcal{F}$, page 8
$\rho_2$: variance metric on $\mathcal{F}$, page 9
$G_n f$: normalized empirical process, page 10
$G_P$: $P$-Brownian-bridge, page 11
$\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_2(f,g) < \delta\}$, page 11
**Notation for efficiency theory.**
$\mathcal{M}$: collection of all possible probability measures of a random variable $X$ (i.e. model), page 17
$\mathcal{M}(\mu)$: measures in $\mathcal{M}$ which are dominated by $\mu$, page 17
$\mathcal{P}(\mu)$: collection of densities w.r.t. $\mu$ corresponding with $\mathcal{M}(\mu)$, page 17
$(D, \|\cdot\|)$: normed vector space, page 18
$\vartheta : \mathcal{M} \to \Theta \subset (D, \|\cdot\|)$: $D$-valued parameter, page 18
$b : (D, \|\cdot\|) \to \mathbb{R}$: a real valued linear mapping on $D$, page 18
$B$: a collection of $b : (D, \|\cdot\|) \to \mathbb{R}$, page 18

$\theta_n$: estimator of $\theta$, page 18
$(L_0^2(P), \|\cdot\|_P)$: Hilbertspace of square integrable functions with mean zero, page 8
$P_{\epsilon,g}$, $(p_{\epsilon,g})$: one dimensional differentiable submodel of measures (densities w.r.t. a certain fixed measure) with score $g$, page 18
$\mathcal{S}(P)$: class of one dimensional submodels, for all our applications it is the class of lines as defined in (2.12), page 18
$S(P)$: tangent cone, page 18
$T(P)$: tangent space, page 19
$b\dot{\vartheta} : T(P) \to \mathbb{R}$: pathwise derivative of $b\vartheta$ at $P$ relative to $S(P)$, page 19
$\tilde{I}(P, b\vartheta)$: efficient influence function for estimating $b\vartheta(P)$, page 19
$I(P, b\vartheta)$: influence curve for estimating $b\vartheta(P)$, page 21
$\|f\|_B \equiv \sup_{b \in B^*} | bf |$, page 21
$A_\theta : T(\theta) \to L_0^2(P_\theta)$: score operator at $P_\theta$, page 36
$A_\theta^\top : L_0^2(P_\theta) \to T(\theta)$: adjoint of $A_\theta$, page 36
$I_\theta = A_\theta^\top A_\theta : T(\theta) \to T(\theta)$: information operator at $P_\theta$, page 38
$I_\theta^-$: the generalized inverse of $I_\theta$.

## CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.
2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.
3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.
4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.
5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.
6 A.P.W. Böhm. *Dataflow computation.* 1984.
7 A. Blokhuis. *Few-distance sets.* 1984.
8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.
9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.
10 C.G. van der Laan, N.M. Temme. *Calculation of special functions: the gamma function, the exponential integrals and error-like functions.* 1984.
11 N.M. van Dijk. *Controlled Markov processes; time-discretization.* 1984.
12 W.H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods.* 1985.
13 D. Grune. *On the design of ALEPH.* 1985.
14 J.G.F. Thiemann. *Analytic spaces and dynamic programming: a measure theoretic approach.* 1985.
15 F.J. van der Linden. *Euclidean rings with two infinite primes.* 1985.
16 R.J.P. Groothuizen. *Mixed elliptic-hyperbolic partial differential operators: a case-study in Fourier integral operators.* 1985.
17 H.M.M. ten Eikelder. *Symmetries for dynamical and Hamiltonian systems.* 1985.
18 A.D.M. Kester. *Some large deviation results in statistics.* 1985.
19 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 1: Philosophy, framework, computer science.* 1986.
20 B.F. Schriever. *Order dependence.* 1986.
21 D.P. van der Vecht. *Inequalities for stopped Brownian motion.* 1986.
22 J.C.S.P. van der Woude. *Topological dynamix.* 1986.
23 A.F. Monna. *Methods, concepts and ideas in mathematics: aspects of an evolution.* 1986.
24 J.C.M. Baeten. *Filters and ultrafilters over definable subsets of admissible ordinals.* 1986.
25 A.W.J. Kolen. *Tree network and planar rectilinear location theory.* 1986.
26 A.H. Veen. *The misconstrued semicolon: Reconciling imperative languages and dataflow machines.* 1986.
27 A.J.M. van Engelen. *Homogeneous zero-dimensional absolute Borel sets.* 1986.
28 T.M.V. Janssen. *Foundations and applications of Montague grammar, part 2: Applications to natural language.* 1986.
29 H.L. Trentelman. *Almost invariant subspaces and high gain feedback.* 1986.
30 A.G. de Kok. *Production-inventory control models: approximations and algorithms.* 1987.
31 E.E.M. van Berkum. *Optimal paired comparison designs for factorial experiments.* 1987.
32 J.H.J. Einmahl. *Multivariate empirical processes.* 1987.
33 O.J. Vrieze. *Stochastic games with finite state and action spaces.* 1987.
34 P.H.M. Kersten. *Infinitesimal symmetries: a computational approach.* 1987.
35 M.L. Eaton. *Lectures on topics in probability inequalities.* 1987.
36 A.H.P. van der Burgh, R.M.M. Mattheij (eds.). *Proceedings of the first international conference on industrial and applied mathematics (ICIAM 87).* 1987.
37 L. Stougie. *Design and analysis of algorithms for stochastic integer programming.* 1987.
38 J.B.G. Frenk. *On Banach algebras, renewal measures and regenerative processes.* 1987.
39 H.J.M. Peters, O.J. Vrieze (eds.). *Surveys in game theory and related topics.* 1987.
40 J.L. Geluk, L. de Haan. *Regular variation, extensions and Tauberian theorems.* 1987.
41 Sape J. Mullender (ed.). *The Amoeba distributed operating system: Selected papers 1984-1987.* 1987.
42 P.R.J. Asveld, A. Nijholt (eds.). *Essays on concepts, formalisms, and tools.* 1987.
43 H.L. Bodlaender. *Distributed computing: structure and complexity.* 1987.
44 A.W. van der Vaart. *Statistical estimation in large parameter spaces.* 1988.
45 S.A. van de Geer. *Regression analysis and empirical processes.* 1988.
46 S.P. Spekreijse. *Multigrid solution of the steady Euler equations.* 1988.
47 J.B. Dijkstra. *Analysis of means in some non-standard situations.* 1988.
48 F.C. Drost. *Asymptotics for generalized chi-square goodness-of-fit tests.* 1988.
49 F.W. Wubs. *Numerical solution of the shallow-water equations.* 1988.
50 F. de Kerf. *Asymptotic analysis of a class of perturbed Korteweg-de Vries initial value problems.* 1988.
51 P.J.M. van Laarhoven. *Theoretical and computational aspects of simulated annealing.* 1988.
52 P.M. van Loon. *Continuous decoupling transformations for linear boundary value problems.* 1988.
53 K.C.P. Machielsen. *Numerical solution of optimal control problems with state constraints by sequential quadratic programming in function space.* 1988.
54 L.C.R.J. Willenborg. *Computational aspects of survey data processing.* 1988.
55 G.J. van der Steen. *A program generator for recognition, parsing and transduction with syntactic patterns.* 1988.
56 J.C. Ebergen. *Translating programs into delay-insensitive circuits.* 1989.
57 S.M. Verduyn Lunel. *Exponential type calculus for linear delay equations.* 1989.
58 M.C.M. de Gunst. *A random model for plant cell population growth.* 1989.
59 D. van Dulst. *Characterizations of Banach spaces not containing $l^1$.* 1989.
60 H.E. de Swart. *Vacillation and predictability properties of low-order atmospheric spectral models.* 1989.

61 P. de Jong. *Central limit theorems for generalized multilinear forms.* 1989.

62 V.J. de Jong. *A specification system for statistical software.* 1989.

63 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part I.* 1989.

64 B. Hanzon. *Identifiability, recursive identification and spaces of linear dynamical systems, part II.* 1989.

65 B.M.M. de Weger. *Algorithms for diophantine equations.* 1989.

66 A. Jung. *Cartesian closed categories of domains.* 1989.

67 J.W. Polderman. *Adaptive control & identification: Conflict or conflux?.* 1989.

68 H.J. Woerdeman. *Matrix and operator extensions.* 1989.

69 B.G. Hansen. *Monotonicity properties of infinitely divisible distributions.* 1989.

70 J.K. Lenstra, H.C. Tijms, A. Volgenant (eds.). *Twenty-five years of operations research in the Netherlands: Papers dedicated to Gijs de Leve.* 1990.

71 P.J.C. Spreij. *Counting process systems. Identification and stochastic realization.* 1990.

72 J.F. Kaashoek. *Modeling one dimensional pattern formation by anti-diffusion.* 1990.

73 A.M.H. Gerards. *Graphs and polyhedra. Binary spaces and cutting planes.* 1990.

74 B. Koren. *Multigrid and defect correction for the steady Navier-Stokes equations. Application to aerodynamics.* 1991.

75 M.W.P. Savelsbergh. *Computer aided routing.* 1992.

76 O.E. Flippo. *Stability, duality and decomposition in general mathematical programming.* 1991.

77 A.J. van Es. *Aspects of nonparametric density estimation.* 1991.

78 G.A.P. Kindervater. *Exercises in parallel combinatorial computing.* 1992.

79 J.J. Lodder. *Towards a symmetrical theory of generalized functions.* 1991.

80 S.A. Smulders. *Control of freeway traffic flow.* 1996.

81 P.H.M. America, J.J.M.M. Rutten. *A parallel object-oriented language: design and semantic foundations.* 1992.

82 F. Thuijsman. *Optimality and equilibria in stochastic games.* 1992.

83 R.J. Kooman. *Convergence properties of recurrence sequences.* 1992.

84 A.M. Cohen (ed.). *Computational aspects of Lie group representations and related topics. Proceedings of the 1990 Computational Algebra Seminar at CWI, Amsterdam.* 1991.

85 V. de Valk. *One-dependent processes.* 1994.

86 J.A. Baars, J.A.M. de Groot. *On topological and linear equivalence of certain function spaces.* 1992.

87 A.F. Monna. *The way of mathematics and mathematicians.* 1992.

88 E.D. de Goede. *Numerical methods for the three-dimensional shallow water equations.* 1993.

89 M. Zwaan. *Moment problems in Hilbert space with applications to magnetic resonance imaging.* 1993.

90 C. Vuik. *The solution of a one-dimensional Stefan problem.* 1993.

91 E.R. Verheul. *Multimedians in metric and normed spaces.* 1993.

92 J.L.M. Maubach. *Iterative methods for non-linear partial differential equations.* 1994.

93 A.W. Ambergen. *Statistical uncertainties in posterior probabilities.* 1993.

94 P.A. Zegeling. *Moving-grid methods for time-dependent partial differential equations.* 1993.

95 M.J.C. van Pul. *Statistical analysis of software reliability models.* 1993.

96 J.K. Scholma. *A Lie algebraic study of some integrable systems associated with root systems.* 1993.

97 J.L. van den Berg. *Sojourn times in feedback and processor sharing queues.* 1993.

98 A.J. Koning. *Stochastic integrals and goodness-of-fit tests.* 1993.

99 B.P. Sommeijer. *Parallelism in the numerical integration of initial value problems.* 1993.

100 J. Molenaar. *Multigrid methods for semiconductor device simulation.* 1993.

101 H.J.C. Huijberts. *Dynamic feedback in nonlinear synthesis problems.* 1994.

102 J.A.M. van der Weide. *Stochastic processes and point processes of excursions.* 1994.

103 P.W. Hemker, P. Wesseling (eds.). *Contributions to multigrid.* 1994.

104 I.J.B.F. Adan. *A compensation approach for queueing problems.* 1994.

105 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 1.* 1994.

106 O.J. Boxma, G.M. Koole (eds.). *Performance evaluation of parallel and distributed systems - solution methods. Part 2.* 1994.

107 R.A. Trompert. *Local uniform grid refinement for time-dependent partial differential equations.* 1995.

108 M.N.M. van Lieshout. *Stochastic geometry models in image analysis and spatial statistics.* 1995.

109 R.J. van Glabbeek. *Comparative concurrency semantics and refinement of actions.* 1996.

110 W. Vervaat, H. Holwerda (ed.). *Probability and lattices.* 1997.

111 I. Helsloot. *Covariant formal group theory and some applications.* 1995.

112 R.N. Bol. *Loop checking in logic programming.* 1995.

113 G.J.M. Koole. *Stochastic scheduling and dynamic programming.* 1995.

114 M.J. van der Laan. *Efficient and inefficient estimation in semiparametric models.* 1995.

115 S.C. Borst. *Polling models.* 1996.

116 G.D. Otten. *Statistical test limits in quality control.* 1996.

117 K.G. Langendoen. *Graph reduction on shared-memory multiprocessors.* 1996.

118 W.C.A. Maas. *Nonlinear $\mathcal{H}_\infty$ control: the singular case.* 1996.

119 A. Di Bucchianico. *Probabilistic and analytical aspects of the umbral calculus.* 1997.

120 M. van Loon. *Numerical methods in smog prediction.* 1997.

121 B.J. Wijers. *Nonparametric estimation for a windowed line-segment process.* 1997.

122 W.K. Klein Haneveld, O.J. Vrieze, L.C.M. Kallenberg (editors). *Ten years LNMB – Ph.D. research and graduate courses of the Dutch Network of Operations Research.* 1997.

## MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print: replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservations of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

37 J.W. de Bakker, G.A. Blaauw, A.J.W. Duijvestijn, E.W. Dijkstra, P.J. van der Houwen, G.A.M. Kamsteeg-Kemper, F.E.J. Kruseman Aretz, W.L. van der Poel, J.P. Schaap-Kruseman, M.V. Wilkes, G. Zoutendijk. *MC-25 Informatica Symposium.* 1971.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

48 F.E.J. Kruseman Aretz, P.J.W. ten Hagen, H.L. Oudshoorn. *An ALGOL 60 compiler in ALGOL 60, text of the MC-compiler for the EL-X8.* 1973.

49 H. Kok. *Connected orderable spaces.* 1974.

50 A. van Wijngaarden, B.J. Mailloux, J.E.L. Peck, C.H.A. Koster, M. Sintzoff, C.H. Lindsey, L.G.L.T. Meertens, R.G. Fisker (eds.). *Revised report on the algorithmic language ALGOL 68.* 1976.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statements.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Empirical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

84 H.L.L. Busard. *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?), books vii-xii.* 1977.

85 J. van Mill. *Supercompactness and Wallmann spaces.* 1977.

86 S.G. van der Meulen, M. Veldhorst. *Torrix I, a programming system for operations on vectors and matrices over arbitrary fields and of variable size.* 1978.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *GO-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewai theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequencies.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

157 H.A.W.M. Kneppers. *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic.* 1983.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.