Estimation of Smooth Functionals with Interval Censored Data

and something completely different

Estimation of Smooth Functionals with Interval Censored Data

and something completely different

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. J. Blaauwendraad,
in het openbaar te verdedigen ten overstaan van een commissie,
door het College van Dekanen aangewezen,
op dinsdag 11 februari 1997 te 13.30uur

door

Ronald Bertus GESKUS

doctorandus in de wiskunde

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. P. Groeneboom

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof. dr. P. Groeneboom, Technische Universiteit Delft, promotor
Prof. dr. R.D. Gill, Universiteit Utrecht
Prof. dr. C.L. Scheffer, Technische Universiteit Delft
Prof. dr. Ph.P.J.E. Clément, Technische Universiteit Delft
Dr. S. van de Geer, Rijksuniversiteit Leiden
Dr. P. Sasieni, Imperial Cancer Research Fund Laboratories, London
Dr. H.P. Lopuhaä, Technische Universiteit Delft

De publicatie van dit proefschrift is mede mogelijk gemaakt door een subsidie van de Stichting Sarphati Amsterdam.

Geskus, Ronald

Estimation of Smooth Functionals with Interval Censored Data, and something completely different.
Ronald Geskus
Thesis Delft University of Technology. - With ref. - With summary in English and in Dutch.
Subject headings: interval censoring / information lower bounds / nonparametric maximum likelihood / Gibbs models

Dedicated to the memory of
Nico Willems and Peter Sandtke

Contents

Preface

The life of a commuter, living in Amsterdam and working in Delft, can be hard sometimes. He chooses to travel by train, since 78 kilometers are too many to go by bicycle every day. Upon arrival at the railway station, problems may already present themselves. He may receive conflicting information as to which train will be the first to leave in the desired direction. Having interpreted the information obtained from the guard and the traffic-control as well as he can, he ends up in a train having a seat with a very good view of the train that leaves in the desired direction before his own. Of course, we are only at the beginning of the information era and no one can expect the Dutch railways to make use of all the new possibilities offered.

Other inconveniences may be his lot as well. He may have to fight his way to the last empty seat, the driver may forget to make a stop in Delft, or the train may arrive at the commuter's destination with a rather serious delay. Of course, just like the weather, these inconveniences are somewhat compensated by their being a rich source of conversation.

For a statistician travelling by train, however, a commuting life is not bad at all. For every inconvenience is alleviated by the prospect of being able to add another element to his data set of train delays, thus yielding the possibility of applying his statistical techniques in order to enlarge his knowledge about train delays between Amsterdam and Delft.

Of course he has to be careful in his study design, but he has plenty of time to think about it during his 55 minute's journey. In fact, he is in a very fortunate situation, for he is able to observe his data with an inaccuracy of at most a few seconds. Moreover, the data can be collected in such a way that the data can be viewed as a realization of a sample of independent observations, or "draws", from the same "population" of train delays. For a practically inclined statistician, this is the most ideal situation, since a lot of theory has been developed to deal with such data. The statistical models treated in this thesis deal with data sets that are not nearly as ideal. A lot of theory still has to be developed, and this thesis is a small contribution to the understanding of these models.

The first part of this thesis (chapters 1 and 2) deals with a situation in which the data can still be seen as a realization of a sample of independent observations from the same population. However, this sample does not consist of the data in which we are interested. The data of interest are *censored* to a certain extent. Nonetheless, clever use of the censored information still tells us a lot. This work is a continuation of a study that resulted in my Master's thesis (GESKUS (1989)). A few parts of it will be repeated in this Ph.D. thesis.

The second part of this thesis (chapter 3) deals with a completely different model, used to describe the amount of repulsion between elements in some homogeneous area, such as goshawks occupying territory in a homogeneous forest. This side-track was taken as a means to give new inspiration during a period, in which my main area of research was in an impasse.

Later, when new progress was being made, my main attention switched back to the interval censoring model. Therefore, it covers the larger part of my thesis.

Chapter 1

Interval censoring and lower bounds

Censoring models are used in situations in which inaccuracy in part, or even all of the data is not due to some small measurement error, but is of another order of magnitude and often of another origin as well.

We will start with a short characterization of some frequently encountered censored data structures in survival analysis problems. They are mostly exemplified by situations from HIV and AIDS related research. This is followed by an overview of the relevant notions and theorems from efficiency theory. At the beginning we will take a rather intuitive approach. A rigorous and much more general treatment of the subject can be found elsewhere, e.g. in Bickel *et al.* (1993) and part I of Groeneboom and Wellner (1992). We will become more formal when we get nearer to the application of this theory to the interval censoring model. The chapter is closed by the mathematically rigorous lower bound calculations for the different interval censoring cases. Since the reference Groeneboom and Wellner (1992) will be used frequently, it will be abbreviated by GW in the sequel.

1.1 Some censoring models

In survival analysis, the data of interest are time points $x_1, \ldots, x_n$, such as time of death, time of failure or, more generally, event times of some elements of the population studied. Usually, the observation of the event times is censored to some degree.

The most well-known type of censoring is *right censoring.* An event time is right censored because it has not yet occurred at the end of the study or because its observation is made impossible due to the occurrence of some other competing event. In HIV/AIDS cohort studies, right censored incubation times of HIV infected persons occur frequently because these studies have not yet been going on long enough in order to observe long incubation times (study cutoff). Moreover, the AIDS diagnosis may not be observed since persons left the study before they developed AIDS (loss to follow-up) or were subject to some other event, preventing the occurrence of AIDS (competing risk, such as pre-AIDS mortality). Ideally, these censoring mechanisms are unrelated to disease progression, in which situation they can all be treated in the same way. For censoring due to study cutoff, this independence assumption is a reasonable one. For censoring due to loss to follow-up, and especially censoring due to competing risks, one should be very careful. For example, pre-AIDS mortality is a frequently occurring phenomenon among injection drug users, and may very well be related with progression to AIDS (see e.g. VAN HAASTRECHT *et al.* (1996)). If the censoring mechanism is not unrelated to the event time mechanism, model building is much harder. In the sequel, we will assume both mechanisms to act independently.

Left censoring is the equivalent of right censoring at the left side of the time axis. For example, in estimation of the HIV seroconversion distribution over calendar time, one may have persons that were already positive at their first test.

When modelling left and right censoring, one usually assumes a sample of censoring times $t_1, \ldots, t_n$ along with the sample of event times. In case of right censored data, if the event time x_j occurs before the censoring time t_j, it is observed; otherwise the event time is censored. In order to get in correspondence with the terminology used in the interval censoring model, the censoring times are called the *observation times*. There is a wealth of literature on all kinds of aspects of left and right censoring. See ANDERSEN *et al.* (1993) for a recent overview. The estimation of the distribution function in case of independent censoring is usually based on the Kaplan-Meier estimator. This estimator strongly depends on the uncensored event times. If all event times are censored, it cannot be used.

Combinations of left and right censored data are usually called *doubly censored*. See e.g. CHANG AND YANG (1987) and CHANG (1990), who deal with estimation of the distribution function. It can be modelled by introducing two observation times for each event time. If the event occurs between both censoring times, it is observed. Otherwise it is censored. The terminology "doubly censored data" is a somewhat unfortunate one, since another kind of censored data, to be treated below, is called doubly censored as well. It may be better to call this censoring mechanism *two-sided censoring*.

Left, right and two-sided censoring have in common the prerequisite that at least part of the event times can be observed themselves, without censoring. Often, we have to deal with situations in which direct observation of the event times is impossible altogether.

In its simplest form, one has one observation time for each element in the study population, and all one knows about the event time is whether the event has occurred before this observation time or not. An example of this type of censoring is treated by KEIDING (1991) and GW: the occurrence of hepatitis A infection as a function of age is investigated through data from a cross-sectional study among the Bulgarian population. The data used consist of the age of the persons in the study (the observation times), and the information whether they have been infected with the hepatitis A virus. This kind of censoring is called *interval censoring case 1*. Such data are called *current status data*. From a theoretical point of view, it is the most tractable kind of interval censoring. Quite a lot is known about the estimation of the distribution function via the nonparametric maximum likelihood estimator when the observation time distribution is independent of the event time distribution. From now on the nonparametric maximum likelihood estimator will be abbreviated as NPMLE. See GW for an overview of results on the NPMLE, in which estimation of the NPMLE of the mean is treated as well. In HUANG AND WELLNER (1995A) the NPMLE of a wider class of estimands than the mean is treated. That article is much related to the contents of this thesis.

Interval censored data can be summarized by a set of intervals $J_1, \ldots, J_n$ in which the event times $x_1, \ldots, x_n$ are known to be contained. In case 1, all intervals have the form $[\tau_0, t_i]$

or $(t_i, \tau_M]$, with τ_0 the smallest possible event time (often zero) and τ_M the largest possible event time (infinity allowed). In the more general interval censoring situation, the intervals can be arbitrary, say of the form $(u_i, v_i]$. For example, in HIV/AIDS cohort studies, people are tested for HIV antibodies at several, more or less regular, time intervals. So we have a *set* of observation times, with at most two of them delimiting the relevant event time. These are the only observation times that are needed for the computation of the NPMLE if the observation times are independent of the event times. For asymptotic results, however, the distribution of the whole set of observation times is important. The number of observation times per element of the study population gives a further subspecification in cases. If one has two observation times for each unobservable event time, one usually speaks of interval censoring case 2. GW, WELLNER (1995) and GROENEBOOM (1996) deal with estimation of the distribution function for case 2. The main subject of *this* thesis will be the estimation of "smooth" functionals of the distribution function in case 2. With interval censored data, the mean is an example of a smooth functional, whereas the median is a counterexample. Of course, the restriction to two observation times is quite stringent. Extensions to more than two, and a variable number of observation times per unobservable event time are usually treated separately for reasons of ease of notation (see WELLNER (1995) and this thesis). However, as far as the methods and techniques used are concerned, the main distinction is between case 1 and case 2. The situation with data having partly one and partly two observation times per unobservable event time will be treated in this thesis together with case 2. The rest will be treated separately and summarized as interval censoring case k. Note that the main distinction between two-sided censoring and interval censoring case 2 (and case k) is the fact that with interval censoring, the event time is always unobservable, even if the event has occurred between the two observation times u_i and v_i.

With respect to asymptotic considerations, another distinction exists within case 2 (and case k). This distinction is determined by how close the observation times can come to each other. These cases will be called 2A and 2B. We will come back to this later.

All the above censored data structures have both an initiating event and a terminating event, but only the last one is censored. The other one determines the origin of the time axis. For data that may have both events censored within the same individual, DE GRUTTOLA AND LAGAKOS (1989) introduced the term *double censoring*. For example, in research estimating the time from seroconversion to AIDS (the latency period), the date of seroconversion is almost always interval censored, whereas the date of AIDS diagnosis may be right censored, and in some studies interval censored as well. Of course, this problem may be treated univariately, transforming the data to information on the, possibly right censored, length of the period from seroconversion to AIDS, but this leads to loss of information. A better way is to treat the problem bivariately, incorporating the information on both the seroconversion distribution over calendar time and the latency period. Two other examples of doubly censored data are investigated in JEWELL, MALANI AND VITTINGHOFF (1994). For one of them the matter of interest is the estimation of the distribution of the length of the period between becoming infectious and seroconversion, based on data from blood transfusions. The most important observation times are the time of the last seronegative donation and the time of the first seropositive donation. The donor's moment of seroconversion is always in this interval. Compared to the previous example of double censoring, the loss of information by the censoring mechanism is more considerable. Not only are both the initiating event (the donor becomes infectious) and the terminating event (the donor seroconverts) always censored, but also the moment the donor's blood becomes infectious may be in the same interval as his moment of seroconversion. The blood recipients in the study had no other risk factor to HIV-infection prior to transfusion, so their time of seroconversion determines the position of the donor's switch to infectivity relative to the last seronegative donation.

We will only be concerned with estimation of the distribution function and functionals thereof, based on a sample of independent, identically distributed (i.i.d.) interval censored random variables. Models incorporating covariates will not be treated in this thesis, but some things have been done in this field with respect to interval censoring. See e.g. Huang and Wellner (1995b), in which the Cox proportional hazards model for case 2 is considered. References to other regression models with interval censored data can be found in Huang and Wellner (1996).

1.2 Efficiency theory for smooth functionals

In this section the event times $x_1, \ldots, x_n$ are assumed to be observable themselves. We assume the data to be a realization of a sample $X_1, \ldots, X_n$ of i.i.d. random variables. Although this may not always hold in practice, often one can design one's experiment in such a way that it is at least approximately true. We will restrict ourselves to the estimation of some real-valued aspect of the distribution, like the mean. Formally, the general set-up is:

- We have an i.i.d. sample $X_1, \ldots, X_n \sim P$. An arbitrary element of this sample is denoted by X.
- P is an unknown distribution, to be estimated from some collection of distributions $\mathcal{P}$.
- We are interested in the estimation of $K(P)$, which is performed via the estimator $T_n = t(X_1, \ldots, X_n)$.

Under these assumptions, criteria have been set up and methods have been developed in order to evaluate the quality of T_n. We will restrict our attention to the Cramér-Rao information lower bound and its generalization to infinite-dimensional parameter spaces.

First we consider the *parametric* model $\mathcal{P} = \{P_\theta \,|\, \theta \in \Theta \subset \mathbb{R}^n\}$, with P_θ having density p_θ with respect to some dominating measure μ. One approach to obtaining a good estimator is to look for an unbiased estimator such that its variance has the smallest possible value for each value of θ. The Cramér-Rao information inequality gives a lower bound for the variance of unbiased estimators. In its simplest form, for a one-dimensional parametric model with $\Theta \subset \mathbb{R}$, it says that any unbiased estimator of the estimand $K(\theta) \in \mathbb{R}$ satisfies

$$\mathrm{var}_\theta(T_n) \geq \frac{[K'(\theta)]^2}{n\, I(\theta)} \qquad \text{for each } \theta \in \Theta.$$

The *information function* $I(\theta)$ is given by

$$I(\theta) = E\left[\frac{d}{d\theta} \log p_\theta(X)\right]^2.$$

The function $\frac{d}{d\theta} \log p_\theta(x) =: \dot{l}_\theta(x)$ is called the *score function*. In the above formulation, both K and $\log p_\theta$ are implicitly assumed to be differentiable in θ, the latter for each x. These smoothness conditions, together with some minor extra conditions, are often summarized as the *regularity* conditions.

Example: For $\mathcal{P} = \{N(\theta, 1) \,|\, \theta \in \mathbb{R}\}$, we have $\dot{l}_\theta(x) = x - \theta$ and $I(\theta) \equiv 1$, yielding $\text{var}_\theta(T_n) \geq 1/n$ for any unbiased estimator of $K(\theta) = \theta$. Hence $\bar{X}_n$ can be seen as best in this sense.

However, biased estimators may perform as well as or even better than unbiased ones with respect to mean squared error, and, especially in nonparametric models, estimators that are unbiased for each possible $P \in \mathcal{P}$ often do not even exist. Then the Cramér-Rao theorem loses its value, but the same bound turns up in asymptotic considerations, when the sample size goes to infinity. Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution. Apart from a slight but necessary modification, and under the same kind of regularity conditions as above, the following holds. If

$$\sqrt{n}\,[\,T_n - K(\theta)\,] \xrightarrow{\mathcal{D}} N(0, \text{v}(\theta)), \qquad \text{as } n \to \infty$$

with $\text{v}(\theta) > 0$, then

$$\text{v}(\theta) \geq \frac{[K'(\theta)]^2}{I(\theta)} \qquad \text{for each } \theta \in \Theta.$$

Example (continued)**:** By the central limit theorem, it is seen immediately that $\bar{X}_n$, as estimator of the mean, asymptotically attains the lower bound 1.

The modification that is needed to make things rigorous is motivated by a counterexample which is named after its discoverer:

Hodges' counterexample: In the example above, consider the following estimator

$$T_n' = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| \geq n^{1/4} \\ a\bar{X}_n & \text{if } |\bar{X}_n| < n^{1/4} \end{cases}$$

Then

$$\text{v}(\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ a^2 & \text{if } \theta = 0 \end{cases}$$

Hence, for $0 < a < 1$, T_n' violates the uniform lower bound at $\theta = 0$.

One way to cope with this phenomenon of *superefficiency* is to prove the lower bound to hold for almost all (with respect to Lebesgue measure) parameter values (see e.g. LEHMANN (1983), chapter 6, for a short discussion on this topic).

Another solution, which has been extended to nonparametric models, is to formulate theorems in a *minimax* form. Using some form of bowl-shaped loss function, such theorems say that the supremum of the expected loss, over a collection of parameter values in a neighbourhood of the parameter of interest, is always larger than some lower bound value (see e.g. chapter 2 in part I of GW).

Yet another approach, which can also be extended to nonparametric models, is to look at a way to exclude T_n' as an acceptable estimator. It is suggested by the following observation. Suppose $\theta = 0$, the point where things go wrong, and consider a sequence θ_n, converging to 0 as $n \to \infty$. Let $X_1, \dots, X_n \sim N(\theta_n, 1)$. We look at the limit behaviour of $\sqrt{n}\,[\,T_n' - \theta_n\,]$. Since

$$\begin{aligned}\sqrt{n}\,[\,a\bar{X}_n - \theta_n\,] &= a\sqrt{n}\,[\,\bar{X}_n - \theta_n\,] + (a-1)\sqrt{n}\theta_n \\ &\sim N((a-1)\sqrt{n}\theta_n, a^2),\end{aligned}$$

we have:

I: If $\theta_n = o(n^{-1/2})$:

$$\sqrt{n}\,[\,T_n' - \theta_n\,] \xrightarrow{\mathcal{D}} N(0, a^2)$$

II: If $\theta_n = c_n n^{-1/2}$ with $c_n \to c \neq 0$:

$$\sqrt{n}\,[\,T_n' - \theta_n\,] \xrightarrow{\mathcal{D}} N((a-1)c, a^2)$$

III: If $\theta_n = c_n n^{-1/2}$ with c_n going to infinity, a limit distribution does not exist.

A difference with the mean $\bar{X}_n$ is that for T_n' the limit distribution in case II depends on the "direction of approach" of θ_n to 0, represented by c. We can also say that we have subcollections of distributions $\{P_{\theta_n}\}$, with $P_{\theta_n} \sim N(\theta_n, 1)$ and θ_n as in II, called *submodels*, for which the limit distribution of $\sqrt{n}\,[\,T_n' - \theta_n\,]$ under P_{θ_n} is not equal for all directions c. This leads to the consideration of estimators, which do not have this directional dependence (see the definition on page 11).

We will formalize these observations. In order to be able to extend the theory beyond parametric models, we use a different definition of the regularity conditions. Let our unknown distribution P, defined on $(\mathcal{Y}, \mathcal{B})$, be contained in some class of probability measures $\mathcal{P}$, which is dominated by a σ-finite measure μ. Let P have density p with respect to μ.

Definition: Let, for some $\delta > 0$, the collection $\{P_t\}$ with $t \in (0, \delta)$ be a one-dimensional parametric submodel. Such a submodel is called *regular* or *Hellinger differentiable* at P, if the following holds:

$$\int \left[t^{-1}(\sqrt{p_t} - \sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\right]^2 d\mu \to 0 \quad \text{as } t \downarrow 0, \text{ for some } a \in L_2(P).$$

This property can be seen as a, more generally applicable, L_2-version of the differentiability, in t, of $\log p_t(x)$ at $t = 0$ for each point x (with $p_0 = p$). The function $a(\cdot)$ plays the role of the score function $\frac{d}{dt} \log p_t(\cdot)\big|_{t=0}$. For we have

$$\lim_{t \downarrow 0} \frac{\sqrt{p_t} - \sqrt{p_0}}{t} = \frac{1}{2\sqrt{p_0}}\, \frac{d}{dt} p_t\bigg|_{t=0} = \frac{1}{2}\left(\frac{d}{dt} \log p_t\bigg|_{t=0}\right)\sqrt{p_0}.$$

Therefore, a is called the *score function* as well, sometimes abbreviated to *score*. It is well-known that each score belonging to some Hellinger differentiable submodel integrates to zero: $\int a\,dP = 0$. This follows from

$$\int adP = \lim_{t \downarrow 0} t^{-1} \int (p_t - p) d\mu = 0,$$

using the following proposition.

Proposition 1.2.1 *Suppose*

$$t^{-1}(\sqrt{p_t} - \sqrt{p}) \rightarrow \tfrac{1}{2} a \sqrt{p} \qquad in\ L_2(\mu).$$

Then

$$t^{-1}(p_t - p) \rightarrow ap \qquad in\ L_1(\mu).$$

Proof: Let $\|\cdot\|_1$ denote the $L_1(\mu)$-norm, and let $\|\cdot\|_2$ denote the $L_2(\mu)$-norm. Then we have:

$$\begin{aligned} \|t^{-1}(p_t - p) - ap\|_1 &\leq \|2\sqrt{p}\{t^{-1}(\sqrt{p_t} - \sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\}\|_1 + \|t^{-1}(\sqrt{p_t} - \sqrt{p})^2\|_1 \\ &\leq 2\,\|\sqrt{p}\|_2\,\|t^{-1}(\sqrt{p_t} - \sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\|_2 \\ &\qquad + t\left(\|t^{-1}(\sqrt{p_t} - \sqrt{p})\|_2\right)^2 \end{aligned}$$

The first term converges to zero by Hellinger differentiability, the second term is bounded by

$$2t\left(\|t^{-1}(\sqrt{p_t} - \sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\|_2\right)^2 + 2t\left(\|\tfrac{1}{2}a\sqrt{p}\|_2\right)^2,$$

which also tends to zero as $t \downarrow 0$, using Hellinger differentiability and $a \in L_2(P)$.

□

Notation: The subspace of $L_2(P)$-functions satisfying $\int a dP = 0$ will be denoted as $L_2^0(P)$.

Example (continued): Let $P = N(2,1) \in \mathcal{P} = \{N(\theta, 1) \,|\, \theta \in \mathbb{R}\}$. Many submodels that are Hellinger differentiable at $\theta = 2$ can be found. The most obvious is $P_t = N(2+t, 1)$, yielding $a(x) = x - 2$ as score function. One can as well approach $N(2,1)$ from the other side via $N(2-t, 1)$, or, more generally, one can take $P_t = N(2 + c_t t, 1)$, with $c_t \rightarrow c$, giving $a(x) = c.(x-2)$. Other possible candidates are $P_t = N(2 + t^2, 1)$ and $P_t = N(2 + \sqrt{t}, 1)$. The first one has $a(x) \equiv 0$, whereas the last one is not a Hellinger differentiable submodel.

Definition: The linear span of all possible scores $\{a\}$ is called the *tangent space*, denoted as $T(P)$.

Usually, the collection of scores is a linear space itself. In the example we have $T(N(2,1)) = \{\lambda a \,|\, \lambda \in \mathbb{R}\}$, with $a(x) = x - 2$.

The above example was based on a small one-dimensional parametric model. The extension to finite dimensional parametric models is straightforward, leading to finite dimensional tangent spaces. When the collection $\mathcal{P}$ no longer has a finite dimensional parametrization, one considers all one-dimensional sub-parametrizations. On the other side of the spectrum is the situation with the tangent space as large as possible. The following proposition is proved in BICKEL *et al.* (1993), example 3.2.1 and GESKUS (1989), proposition 3.2.

Proposition 1.2.2 *The model*

$$\mathcal{P}_\mu = \{\text{all } P \text{ on } (\mathcal{Y}, \mathcal{B}) \text{ with } P \ll \mu\}$$

has tangent space $T(P) = L_2^0(P)$ *at* P.

In the proof in BICKEL *et al.* (1993) the score function $a \in L_2^0(P)$ is yielded by the submodel

$$p_t = p \frac{2(1 + e^{-2ta})^{-1}}{\int 2(1 + e^{-2ta})^{-1}\, dP}.$$

In Geskus' proof it is yielded by

$$p_t = \frac{p(1 + ta)1_{\{1+ta \geq 0\}}}{\int p(1 + ta)1_{\{1+ta \geq 0\}} d\mu}.$$

For our interval censoring model, the tangent space is somewhere between a finite dimensional space and the L_2^0-space. In asymptotic lower bound theory, models beyond the parametric domain with a tangent space that is not the whole L_2^0-space are usually denoted as *semi-parametric* models, whereas models with tangent space as large as possible are called *nonparametric*.

The definition of differentiability of the estimand $K(\theta)$ is also extended to fit into this new set-up. From now on $K(\theta)$ is seen as a functional $K: \mathcal{P} \to \mathbb{R}$.

Definition: The functional $K: \mathcal{P} \to \mathbb{R}$ is *pathwise differentiable* at P if there exists a continuous linear map $K'_P: T(P) \to \mathbb{R}$ such that for each Hellinger differentiable path $\{P_t\} \subset \mathcal{P}$, with score a, we have

$$\lim_{t \downarrow 0} t^{-1}[K(P_t) - K(P)] = K'_P(a).$$

K'_P can be written in an inner product form. Since $T(P)$ is a subspace of the Hilbert-space $(L_2(P), \langle \cdot\,, \cdot \rangle)$, the continuous linear functional K'_P can be extended to a continuous linear functional $\overline{K}'_P$ on $L_2(P)$. By the Riesz representation theorem, to $\overline{K}'_P$ belongs a unique $\kappa_P \in L_2(P)$, called the *gradient* or *influence function*, satisfying

$$\overline{K}'_P(h) = \langle \kappa_P, h \rangle_P \text{ for all } h \in L_2(P).$$

Instead of $L_2(P)$, any closed subspace H between $\overline{T(P)}$ and $L_2(P)$ can be chosen as space to which to extend $\overline{K}'_P$ and on which to apply the Riesz representation theorem. Note that κ_P

is uniquely determined once the extension of K'_P has been made. However, many continuous linear extensions of K'_P may be possible, so generally the gradient is not unique. One gradient is playing a special role, which is obtained by making the smallest extension, to $\overline{T(P)}$. Then the extension of K'_P *is* unique, yielding the *canonical gradient* or *efficient influence function* $\tilde{\kappa}_P \in \overline{T(P)}$. The orthogonal projection of any gradient κ_P, obtained after extension of K'_P, into $\overline{T(P)}$ yields the same canonical gradient. Hence we have

$$\|\kappa_P\|_P^2 = \|\tilde{\kappa}_P\|_P^2 + \|\kappa_P - \tilde{\kappa}_P\|_P^2 \geq \|\tilde{\kappa}_P\|_P^2,$$

so the canonical gradient has the smallest norm among all gradients.
For a Hellinger differentiable submodel at P with score a, the lower bound becomes

$$\frac{[K'_P(a)]^2}{\langle a, a\rangle_P} = \frac{[\langle \tilde{\kappa}_P, a\rangle_P]^2}{\langle a, a\rangle_P} = \left[\langle \tilde{\kappa}_P, \frac{a}{\|a\|}\rangle_P\right]^2.$$

Since $\langle \tilde{\kappa}_P, a\rangle_P = \langle \kappa_P, a\rangle_P$ for each $a \in T(P)$, this expression does not depend on the choice of the gradient. Each one-dimensional Hellinger differentiable submodel has a lower bound. The lower bound for $\mathcal{P}$ is obtained by taking the supremum of all these lower bounds. We have

$$\sup_{a\in T(P)} \left[\langle \tilde{\kappa}_P, \frac{a}{\|a\|}\rangle_P\right]^2 = \left[\langle \tilde{\kappa}_P, \frac{\tilde{\kappa}_P}{\|\tilde{\kappa}_P\|}\rangle_P\right]^2 = \|\tilde{\kappa}_P\|_P^2. \tag{1.1}$$

If $\tilde{\kappa}_P$ is contained in $T(P)$ itself, the one-dimensional submodel with score function $\tilde{\kappa}_P$ is called the *least favourable* submodel.

To define *regularity* of the estimator, we go back to the one-dimensional subcollections $\{P_{\theta_n}\}$, with $\theta_n = \theta_0 + c_n n^{-1/2}$ and $c_n \to c$, as in Hodges' counterexample.

Definition: An estimator T_n of $K(P)$ is a regular estimator at $P \in \mathcal{P}$ if for every Hellinger differentiable (or regular) submodel $\{P_t\} \subset \mathcal{P}$ and every sequence $\{\theta_n\}$, with $\theta_n = \mathcal{O}(n^{-1/2})$, $\sqrt{n}\,[T_n - K(P_{\theta_n})]$ converges in distribution, under $X_1, \ldots, X_n \sim P_{\theta_n}$, to the same random variable Z.

The term regularity refers both to smooth submodels as well as to estimators that behave neatly within such models. Both aspects are needed as regularity conditions in the following important theorem, called the *convolution theorem*:

Theorem 1.2.1 *Suppose that:*

(i) K is pathwise differentiable at $P \in \mathcal{P}$ along all regular submodels.

(ii) T_n is a regular estimator, with limit random variable Z.

(iii) The set of all directions $\{a\}$ is a linear space.

Then there exist random variables Z_0 and Δ_0 such that

A. Z has the same distribution as $Z_0 + \Delta_0$.

B. Z_0 *and* Δ_0 *are independent.*

C. $Z_0 \sim N(0, \|\tilde{\kappa}_P\|_P^2)$.

Proof: See e.g. Theorem 3.11.2, p. 414 of VAN DER VAART AND WELLNER (1996)

More general formulations of this theorem exist, but this one is sufficient for the scope of this thesis. The theorem says that the limiting distribution of any regular estimator of $K(P)$ is more spread out than the distribution of Z_0. Hence the smallest asymptotic variance for a regular estimator of $K(P)$ is $\|\tilde{\kappa}_P\|^2$. An *asymptotically efficient* estimator is a regular estimator for which the limiting distribution equals the distribution of Z_0. However, regularity of the estimators to be considered will be ignored in this thesis. The only topic of interest will be to show that $\sqrt{n}\,[T_n - K(P)]$, *with* P *fixed*, converges to a $N(0, \|\tilde{\kappa}_P\|_P^2)$ distribution. Such estimators will be called *optimal* instead of efficient.

So in order to prove optimality of some estimator T_n, two aspects have to be considered. First, one should find out what the lower bound looks like, which is mostly a functional analytic problem. Secondly, one should investigate the limit behaviour of T_n, if $X_1, \ldots, X_n \sim P$, which is mostly a probabilistic problem, using techniques from empirical process theory. In this chapter, attention is paid to the structure of the lower bound. In the next chapter, we will consider the distributional aspects.

1.2.1 Lower bound computations for the nonparametric model

We first consider pathwise differentiability and computation methods for the canonical gradient in the nonparametric setting, thus when $T(P) = L_2^0(P)$. Results for the semi-parametric interval censoring model are related to the gradient structure of the nonparametric model, as will be clarified in the next section.

For the nonparametric situation, once we have proved differentiability and found a gradient $\kappa_P \in L_2(P)$, determining the canonical gradient is an easy task: just subtract $\int \kappa_P dP$ to find the projection into $L_2^0(P)$.

An important class of functionals are the *linear* functionals

$$K(P) = \int c(x)\, dP(x).$$

If P is a one-dimensional distribution, examples of linear functionals are the moment functionals $K(P) = \int x^k\, dP(x)$. Estimation of the distribution function at a fixed point t_0 concerns a linear functional as well: for $K(P) = P(X \leq t_0)$ we have $c(x) = 1_{[0,t_0]}(x)$. For the nonparametric model $\mathcal{P}$, any linear functional with $\sup_{P \in \mathcal{P}} E_P c(X)^2 < \infty$ is pathwise differentiable at any $P \in \mathcal{P}$, with canonical gradient

$$\tilde{\kappa}_P(x) = c(x) - \int c(x)\, dP(x),$$

yielding the information lower bound

$$\|\, c(X) - E_P[c(X)]\, \|_P^2.$$

See proposition A.5.2 in BICKEL *et al.* (1993) for a proof. For nonlinear functionals, there is no general method that immediately establishes differentiability and supplies the formula for the canonical gradient. For two nonlinear functionals the canonical gradient is given below. Moreover, the proof gives information on how to perform the calculations. It is partly similar to the proof of proposition A.5.2 in BICKEL *et al.* (1993) for linear functionals. However, the notation used is different. Moreover, the conditions are stronger than theirs, thus slightly simplifying part of the proof. In view of the next section, attention will be restricted to one dimensional, absolutely continuous distributions having compact support and bounded density. Extensions are possible, but will not be considered here. Also in correspondence with the next section, the distribution will be denoted by its distribution function F instead of P. The class of distributions $\mathcal{P}$ will be assumed to satisfy the following uniformity property: the support of each distribution is contained in the same bounded interval S and the densities are uniformly bounded. Let $\mathcal{F}_S$ denote this nonparametric class of distributions. Restriction to this class does not change the maximality of the tangent space, i.e. we still have $T(F) = L_2^0(F)$ for each $F \in \mathcal{F}_S$. This can be seen by the choice of submodels in proposition 1.2.2 yielding the tangent space: they are still contained in $\mathcal{F}_S$ for t small enough.

Proposition 1.2.3 *Let $F \in \mathcal{F}_S$ have density f and bounded support $[\tau_0, \tau_M] \subset \mathbb{R}$. Consider the functionals*

$$K_1(F) = \int_{\tau_0}^{\tau_M} [f(x)]^2\, w(x)\, dx$$

and

$$K_2(F) = \int_{\tau_0}^{\tau_M} [F(x)]^2\, w(x)\, dx$$

with $w(x)$ a bounded weight function.
Then both functionals are pathwise differentiable, with gradient

$$\kappa_F(x) = 2\, f(x)\, w(x)$$

and

$$\kappa_F(x) = 2 \int_{s=x}^{\tau_M} F(s)\, w(s)\, ds$$

respectively.

Proof: All norms in the proof denote either $L_1(\lambda)$ or $L_2(\lambda)$-norms, with λ denoting Lebesgue measure.

Both functionals can be written in the form

$$K(F) = \int_{\tau_0}^{\tau_M} f(x)\, c_F(x)\, dx,$$

with $c_F(x) = f(x)w(x)$ and $c_F(x) = \int_x^{\tau_M} F(s)\, w(s)\, ds$ respectively. By the conditions on w and $\mathcal{F}_S$, c_F is bounded for both functionals, uniformly over $F \in \mathcal{F}_S$.

Let a be the score function of the regular submodel $\{F_t\} \subset \mathcal{F}_S$ at F. We have

$$\begin{aligned} t^{-1}[K(F_t) - K(F)] &= t^{-1} \int_{\tau_0}^{\tau_M} c_F(x)[f_t(x) - f(x)]\, dx \qquad (1.2) \\ &\quad + t^{-1} \int_{\tau_0}^{\tau_M} f_t(x)[c_{F_t}(x) - c_F(x)]\, dx \end{aligned}$$

Both terms converge to $\langle c_F, a\rangle_F$.

For the first term this is proved in an almost similar way as in BICKEL *et al.* (1993). We have

$$\begin{aligned} \|c_F[t^{-1}(f_t - f) - af]\|_1 &\le \|c_F\left(\sqrt{f_t}+\sqrt{f}\right)\left[t^{-1}\left(\sqrt{f_t}-\sqrt{f}\right) - \tfrac{1}{2}a\sqrt{f}\right]\|_1 \\ &\quad + \|c_F\, \tfrac{1}{2}a\sqrt{f}\left(\sqrt{f_t}-\sqrt{f}\right)\|_1 \\ &= A(t) + B(t). \end{aligned}$$

For $t \downarrow 0$, $|A(t)|$ is bounded by

$$\left[\left\{\int_{\tau_0}^{\tau_M} c_F^2\, dF_t\right\}^{1/2} + \left\{\int_{\tau_0}^{\tau_M} c_F^2\, dF\right\}^{1/2}\right] o(1),$$

whereas $|B(t)|$ can be bounded by

$$\left\{\int_{\tau_0}^{\tau_M} \tfrac{1}{4}c_F^2\, a^2\, dF\right\}^{1/2} \|\sqrt{f_t}-\sqrt{f}\|_2.$$

Using $a \in L_2(F)$, boundedness of c_F, and $\|\sqrt{f_t}-\sqrt{f}\|_2 = o(1)$ as $t \downarrow 0$, one obtains that $A(t) + B(t) \to 0$.

The second term in (1.2) can be transformed into a form, similar to the first term, but with c_F replaced by c_{F_t}. For K_1 this transformation is immediate; for K_2 this follows from

$$\int_{\tau_0}^{\tau_M} f_t\, c_F\, d\lambda = \int_{\tau_0}^{\tau_M} w\, F_t\, F\, d\lambda = \int_{\tau_0}^{\tau_M} f\, c_{F_t} d\lambda.$$

Thus the argument for the first term can be repeated, using uniform boundedness of $\{c_{F_t}\}$.

□

Remarks.

I.) Pathwise differentiability certainly holds for other nonlinear functionals as well. The problem in the proof is determining to what expression the second term in (1.2) converges. For the functionals considered here, we have the advantage that the second term is similar in structure to the first term.

II.) The functionals K_1 and K_2 are the same as the ones considered in HANSEN (1991). She considers the asymptotic distribution of these functionals for interval censoring case 1. Theoretical results are obtained for K_2 only. The limit distribution of the nonparametric maximum likelihood estimator (NPMLE), as obtained by her, is optimal as will be shown in subsection 1.4.1. This result can be extended to the other interval censoring cases, as will be shown in this thesis. The functional K_1 has the disadvantage that some smoothing technique has to be applied. The distributional results for the NPMLE that will be obtained in the next chapter do not apply to K_1. However, as far as lower bound considerations are concerned, both functionals can be treated equivalently.

1.3 Lower bounds with interval censoring

The interval censoring model is an example of a model with information loss. This information loss can be expressed by saying that the distribution of the sample is induced by another distribution, on which we only obtain partial information. The functional of interest is a function of the inducing distribution, but is defined implicitly via the sample distribution. The lower bound theory for such implicitly defined functionals is treated in VAN DER VAART (1991) and BICKEL *et al.* (1993). This theory boils down to solving an operator equation, relating the inducing distribution to the induced one. In this section it will be shown how this operator equation is obtained. The theory needed is illustrated by case 2 of the interval censoring model. However, the derivation of the operator equation for the other interval censoring cases goes in a similar way.

We start with the formulation of the model for case 2. The loss of information is expressed by the fact that, instead of $(X_1, \ldots, X_n)$, we observe $(U_1, V_1, \Delta_1, \Gamma_1), \ldots, (U_n, V_n, \Delta_n, \Gamma_n)$ with $\Delta_i = 1_{\{X_i \le U_i\}}$ and $\Gamma_i = 1_{\{U_i < X_i \le V_i\}}$. The following modelling assumptions are made:

(M1) X is an absolutely continuous random variable with distribution function F. Let $K > 0$ and let S be a bounded interval $\subset \mathbb{R}$. F is contained in the class

$$\mathcal{F}_S := \{F \mid \text{support}(F) \subset S, ; \ F \text{ absolutely continuous}, \ \sup_x |f(x)| \le K\}.$$

F is the distribution on which we want to obtain information; however, we do not observe X directly. Let τ_0 and τ_M be the lower bound and the upper bound of the support of F.

(M2) We observe the pairs (U, V), with simultaneous distribution function H. H is contained in $\mathcal{H}$, the collection of all two-dimensional distributions on $\{(u, v) \mid u < v\}$, absolutely continuous with respect to two-dimensional Lebesgue measure and such that (U, V) is independent of X for each choice of H and F. Let h denote the density of (U, V), with marginal densities and distribution functions h_1, H_1 and h_2, H_2 for U and V respectively. We let H have its mass concentrated on $\{ (u, v) \mid \eta_0 \le u < v \le \eta_M \}$.

(M3) If both H_1 and H_2 put zero mass on some set A, then F has zero mass on this set as well, so $F \ll H_1 + H_2$.

Condition (M3) precludes observation time distributions that are purely discrete, implying that deterministic observation times are not allowed. It will be needed to ensure consistency of the NPMLE with respect to the supremum norm on its support. Moreover, without this assumption the functionals we are interested in are not well-defined. (M3) also implies that $\eta_0 \le \tau_0$ and $\tau_M \le \eta_M$.
The model formulation for the other interval censoring cases is similar in essence. (M1) has nothing to do with the observation times and is similar for all cases; (M2) says that the observation times and the event times should be independent; (M3) says that F cannot have mass on sets in which no observations can occur.

What we *do* observe can be seen as a measurable transformation S of what we *would* observe if there would be no censoring:

$$S(x,u,v) = (u,v,\delta,\gamma).$$

The domain $\{(x,u,v) \mid \tau_0 \le x \le \tau_M,\, \eta_0 \le u < v \le \eta_M\}$ will be called the hidden space, and the image space will be called the observation space. In our model, P is induced by F and H, and is from now on written as $Q_{F,H}$, having density

$$q_{F,H}(u,v,\delta,\gamma) = h(u,v)F(u)^\delta(F(v)-F(u))^\gamma(1-F(v))^{1-\delta-\gamma}$$

with respect to $\lambda_2 \times \nu_2$. Here λ_2 denotes two-dimensional Lebesgue measure and ν_2 denotes counting measure on the set $\{(0,1),(1,0),(0,0)\}$.

We first take a look at the Hellinger differentiable paths. All Hellinger differentiable submodels at $Q_{F,H}$ that can be formed, together with the corresponding score functions, are induced by the Hellinger differentiable paths of densities on the hidden space, according to the following theorem:

Theorem 1.3.1 *Let $\mathcal{P} \ll \mu$ be a class of probability measures on the hidden space $(\mathcal{Y},\mathcal{B})$. Let $P \in \mathcal{P}$ be induced by the random vector Y. Suppose that the path $\{P_t\}$ to P satisfies*

$$\int \left[t^{-1}(\sqrt{p_t}-\sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\right]^2 d\mu \to 0 \quad as \;\; t \downarrow 0$$

for some $a \in L_2^0(P)$.
Let $S\colon (\mathcal{Y},\mathcal{B}) \to (\mathcal{Z},\mathcal{C})$ be a measurable mapping. Let μS^{-1} be σ-finite. Suppose that the induced measures $Q_{P_t} = P_t S^{-1}$ and $Q_P = PS^{-1}$ on $(\mathcal{Z},\mathcal{C})$ are absolutely continuous with respect to μS^{-1}, with densities q_t and q. Then the path $\{Q_{P_t}\}$ is also Hellinger differentiable, satisfying

$$\int \left[t^{-1}(\sqrt{q_t}-\sqrt{q}) - \tfrac{1}{2}\overline{a}\sqrt{q}\right]^2 d\mu S^{-1} \to 0 \quad as \;\; t \downarrow 0$$

with $\overline{a}(z) = E_P[\,a(Y) \mid S = z\,]$.

Proof: See BICKEL *et al.* (1993), proposition A.5.5.

The relation between the scores a in the hidden tangent space $T(P)$ and the induced scores $\overline{a}$ is expressed by the mapping

$$A_P\colon a \mapsto E_P[\,a(Y) \mid S\,].$$

Definition: The mapping $A_P\colon a \mapsto E_P[\,a(Y) \mid S\,]$ is called the *score operator.*

The score operator is continuous and linear. Its range is the induced tangent space, which is contained in $L_2^0(Q_P)$. For the interval censoring model it will turn out to be a proper subspace of $L_2^0(Q_P)$.

Since F and H are assumed to be independent, the one-dimensional submodels in the hidden space are formed by first looking at the classes $\mathcal{F}_S$ and $\mathcal{H}$ separately. By proposition

1.2.2, our assumptions on $\mathcal{F}_S$ and $\mathcal{H}$ make the tangent spaces $T(F)$ and $T(H)$ as large as possible: $T(F) = L_2^0(F)$ and $T(H) = L_2^0(H)$. Let $\{F_t\}$ be a regular submodel at F with score function a, and let $\{H_t\}$ be a regular submodel at H with score function e. Joining these paths gives, using independence,

$$\int \left[t^{-1}(\sqrt{f_t\, h_t} - \sqrt{f\, h}) - \frac{1}{2}(a+e)\sqrt{f\, h}\right]^2 d\lambda_3 \to 0 \quad \text{as } t \downarrow 0.$$

$L_2^0(F)$ and $L_2^0(H)$ are orthogonal subspaces of $L_2^0(F \times H)$. The tangent space $L_2^0(F) + L_2^0(H)$ is a proper subspace of $L_2^0(F \times H)$. Thus, due to the independence assumption, we have left the nonparametric model. Indeed, the construction of proposition 1.2.2 can no longer be used to obtain the whole $L_2^0(F \times H)$-space, since any function $b(x, u, v)$ that cannot be split into a direct sum of $a(x)$ and $e(u, v)$ yields a submodel in which $p_t(x, u, v)$ is the density of *dependent* X and (U, V).

Now theorem 1.3.1 is applied, with $Y = (X, U, V)$, $P = F \times H$ and $\mu = \lambda_3$. The tangent space $T(Q_{F,H})$ of the induced Hellinger differentiable paths is yielded by the score operator A: $L_2^0(F) + L_2^0(H) \to T(Q_{F,H})$ given by:

$$[A_{F,H}(a+e)](u, v, \delta, \gamma) = E_{F,H}\{a(X) + e(U, V) \mid (U, V, \Delta, \Gamma) = (u, v, \delta, \gamma)\}$$

We are interested in estimation of some aspect $K(F)$ of F. However, due to the censoring mechanism, $K(F)$ can only be accessed indirectly through the observation space via the functional $\Theta(Q_{F,H})$, with H acting as a nuisance parameter. Thus we define

$$\Theta(Q_{F,H}) := K(F).$$

Note that $\Theta(Q_{F,H})$ is defined unambiguously by condition (M3).

Having specified the Hellinger differentiable paths in the observation space, differentiability of the functional $\Theta(Q_{F,H})$ in the observation space will now be investigated. Differentiability of implicitly defined functionals $\Theta(Q_P) = K(P)$ can be proved by looking at the structure of the adjoint A_P^* of the map A_P according to theorem 1.3.2 below, which was first proved in VAN DER VAART (1991) in a more general setting, allowing for Banach space-valued functions as estimand. Then the proof is slightly more elaborate. The proof in case of real-valued functionals is very simple and is given below.

The adjoint of a continuous linear mapping A: $G \to H$, with G and H Hilbert-spaces, is the unique continuous linear mapping A^*: $H \to G$ satisfying

$$\langle Ag, h\rangle_H = \langle g, A^*h\rangle_G \quad \forall g \in G, h \in H. \tag{1.3}$$

Any Hilbert space that contains $\overline{\mathcal{R}(A)}$ can be chosen as the image space H, creating a different adjoint A_H^*. However, this does not complicate things: each adjoint A_H^* has the same behaviour on $\overline{\mathcal{R}(A)}$ and its behaviour on $\overline{\mathcal{R}(A)}$ determines A_H^* completely, since $A_H^* h = A_H^*(\Pi(h)) = A^*_{\overline{\mathcal{R}(A)}}(\Pi(h))$, with Π denoting orthogonal projection into $\overline{\mathcal{R}(A)}$. We do not specify the image space chosen and write A^* instead of A_H^*.

For the score operator from theorem 1.3.1, the adjoint is a conditional expectation operator as well: if $Z \sim Q_P$, and $b \in H \subset L_2(Q_P)$, then

$$[A^*\, b](y) = E_P[b(Z)|Y = y] - E_P[b(Z)] \quad \text{a.e.-}[P],$$

which is seen immediately by definition 1.3 of the adjoint. Note that the term $E_P[b(Z)]$ vanishes if $b \in L_2^0(Q_P)$

Theorem 1.3.2 *Let $\mathcal{Q} = \mathcal{P}S^{-1}$ be a class of probability measures on the image space under the measurable transformation S. Suppose the functional Θ: $\mathcal{Q} \to \mathbb{R}$ can be written as $\Theta(Q_P) = K(P)$. Suppose that K is pathwise differentiable at P in the hidden space, having canonical gradient $\tilde{\kappa}_P$.*
Then Θ is differentiable at $Q_P \in \mathcal{Q}$ along the induced paths if and only if

$$\tilde{\kappa}_P \in \mathcal{R}(A^*) \tag{1.4}$$

If (1.4) holds, then the canonical gradients $\tilde{\theta}_{Q_P}$ of Θ and $\tilde{\kappa}_P$ of K are related via

$$\tilde{\kappa}_P = A^* \tilde{\theta}_{Q_P} \tag{1.5}$$

Proof: We have

$$\begin{aligned} \lim_{t \downarrow 0} t^{-1} [\Theta(Q_{P_t}) - \Theta(Q_P)] &= \lim_{t \downarrow 0} t^{-1} [K(P_t) - K(P)] \\ &= \langle \tilde{\kappa}_P, g \rangle_P \end{aligned} \tag{1.6}$$

Suppose Θ is pathwise differentiable at Q_P. So, for any Hellinger differentiable path $\{P_t\}$, with score-function $g \in T(P)$, we have

$$\begin{aligned} \lim_{t \downarrow 0} t^{-1} [\Theta(Q_{P_t}) - \Theta(Q_P)] &= \langle Ag, \tilde{\theta} \rangle_{Q_P} \\ &= \langle g, A^* \tilde{\theta} \rangle_P, \end{aligned}$$

where $\tilde{\theta} = \tilde{\theta}_{Q_P} \in \overline{\mathcal{R}(A)}$ is uniquely determined. Combining this with (1.6) gives $\tilde{\kappa}_P = A^* \tilde{\theta}_{Q_P}$, hence $\tilde{\kappa}_P \in \mathcal{R}(A^*)$.

Conversely, suppose $\tilde{\kappa}_P = A^* b$ for some b in the domain of A^*. Then we have, for any $\{P_t\}$, having score g,

$$\begin{aligned} \lim_{t \downarrow 0} t^{-1} [\Theta(Q_{P_t}) - \Theta(Q_P)] &= \langle \tilde{\kappa}_P, g \rangle_P \\ &= \langle A^* b, g \rangle_P = \langle b, Ag \rangle_{Q_P} \end{aligned}$$

Hence Θ is pathwise differentiable with gradient b.

□

Equation (1.5) is called the *score equation.* This theorem is applied to the interval censoring model. First, the score operator $A(a+e)$ is split in two parts, one related to the unobservable event times and one related to the observation times:

$$A_{F,H}(a+e) = L_1 a + L_2 e,$$

with

$$L_1 a = E_{F,H}[a(X) \mid (U, V, \Delta, \Gamma)]$$

and

$$L_2 e = E_{F,H}[e(U,V) \mid (U,V,\Delta,\Gamma)].$$

Note that $\mathcal{R}(L_1)$ and $\mathcal{R}(L_2)$ are orthogonal subspaces in $L_2^0(Q_{F,H})$, as is easily shown using independence of X and (U,V), and $Ea(X) = Ee(U,V) = 0$. Due to the fact that the observation times are not censored, L_2 is simply the identity operator. L_1 is given by

$$[L_1 a](u,v,\delta,\gamma) = \frac{\delta \int_{\tau_0}^{u} a\,dF}{F(u)} + \frac{\gamma \int_{u}^{v} a\,dF}{F(v)-F(u)} + \frac{(1-\delta-\gamma)\int_{v}^{\tau_M} a\,dF}{1-F(v)} \quad \text{a.e.} - Q_{F,H} \tag{1.7}$$

Note that $\int_{\eta_0}^{\tau_0} a dF = \int_{\tau_M}^{\eta_M} a dF = 0$.

Now we apply theorem 1.3.2. Since K does not depend on H, the (canonical) gradient of K is a function in the $L_2(F)$-subspace of the $L_2(F \times H)$-space. In fact, with respect to K, we are in the nonparametric model $\mathcal{F}_S$ with tangent space $L_2^0(F)$. So the gradient calculations of subsection 1.2.1 can be used, and we write $\tilde{\kappa}_F$ for the canonical gradient. Now theorem 1.3.2 says that Θ is pathwise differentiable if and only if

$$A_{F,H}^* \tilde{\theta} = \tilde{\kappa}_F, \tag{1.8}$$

for some $\tilde{\theta} \in \overline{\mathcal{R}(L_1) + \mathcal{R}(L_2)}$. The adjoint can be written as

$$A^* b = L_1^* b + L_2^* b. \tag{1.9}$$

For we have, for any $a \in L_2^0(F)$, $e \in L_2^0(H)$ and $b \in L_2(Q_{FH})$,

$$\begin{aligned} \langle A^* b, a+e\rangle &= \langle b, A(a+e)\rangle \\ &= \langle b, L_1 a\rangle + \langle b, L_2 e\rangle \\ &= \langle L_1^* b, a\rangle + \langle L_2^* b, e\rangle \\ &= \langle L_1^* b, a+e\rangle + \langle L_2^* b, a+e\rangle \\ &= \langle L_1^* b + L_2^* b, a+e\rangle, \end{aligned}$$

using $\langle L_1^* b, e\rangle = \langle L_2^* b, a\rangle = 0$. By (1.9), $\mathcal{R}(L_2^*) \perp \tilde{\kappa}_F$ and $\mathcal{R}(L_1) \perp \mathcal{R}(L_2)$, (1.8) is equivalent to

$$L_1^* \tilde{\theta} = \tilde{\kappa}_F \tag{1.10}$$

for some $\tilde{\theta} \in \overline{\mathcal{R}(L_1)}$. So only the adjoint of L_1 plays a role in the score equation. The adjoint of L_1 is given by the formula

$$\begin{aligned} [L_1^* b](x) \;=\; & \int_{u=x}^{\eta_M}\int_{v=u}^{\eta_M} b(u,v,1,0)\,h(u,v)\,dv\,du\; + \\ & \int_{u=\eta_0}^{x}\int_{v=x}^{\eta_M} b(u,v,0,1)\,h(u,v)\,dv\,du\; + \\ & \int_{u=\eta_0}^{x}\int_{v=u}^{x} b(u,v,0,0)\,h(u,v)\,dv\,du \qquad \text{a.e.-}[F]. \end{aligned} \tag{1.11}$$

Many functionals that are pathwise differentiable in the model without censoring, lose this property in the interval censoring model. Due to the smoothness of the adjoint operator, any functional K with a canonical gradient that is not a.e. equal to a continuous function cannot be obtained under L_1^*. So not all linear functionals remain pathwise differentiable. For example, $K(F) = F(t_0)$, with canonical gradient $1_{[0,t_0]}(\cdot) - F(t_0)$, loses this property. Hence the above lower bound theory can no longer be applied -one may also say that the lower bound has an infinite value- and $F(t_0)$ is not estimable at $\sqrt{n}$-rate. In the next chapter, the convergence rate and limit distribution of the NPMLE of $F(t_0)$ is briefly discussed.

However, functionals of the form $K(F) = \int c_F(x)\,dF(x)$, with c_F sufficiently smooth, will be shown to remain differentiable under censoring. Hence for these functionals the above information lower bound theory does apply. This will be the subject of the next sections.

We first state one more general result. The information lower bound $\|\tilde\theta\|^2_{Q_P} = \int \tilde\theta^2 dQ_P$ can also be written as an inner product with respect to the hidden probability P, instead of Q_P, according to the following theorem:

Theorem 1.3.3 *i): Let $\tilde\theta = \tilde\theta_{Q_P}$ satisfy $\tilde\kappa_P = A^*\tilde\theta$ and assume that $\tilde\theta$ is contained in $\mathcal{R}(A)$: $\tilde\theta = Aa_0$ for some $a_0 \in \overline{T(P)}$. Then we have*

$$\|\tilde\theta\|^2_{Q_P} = \langle a_0, \tilde\kappa_P\rangle_P.$$

ii): If moreover P is a one-dimensional continuous distribution with support contained in $[a,b]$, and $\tilde\kappa_P$ can be written as $\tilde\kappa_P(t) = \int_a^t \tilde\kappa'_P(x)dx + \tilde\kappa_P(a)$, with $\tilde\kappa'_P$ bounded, we have

$$\|\tilde\theta\|^2_{Q_P} = \int_a^b \tilde\kappa'_P(x)\phi_0(x)dx$$

with $\phi_0(x) = \int_x^b a_0(t)\,dP(t)$.

iii): If $\tilde\theta \in \overline{\mathcal{R}(A)} \setminus \mathcal{R}(A)$, say $\tilde\theta = \lim_{n\to\infty} Aa_n$, we have

$$\|\tilde\theta\|^2_{Q_P} = \lim_{n\to\infty}\langle a_n, \tilde\kappa_P\rangle_P,$$

which, under the conditions in ii), can be rewritten as

$$\|\tilde\theta\|^2_{Q_P} = \lim_{n\to\infty}\int_a^b \tilde\kappa'_P(x)\phi_n(x)dx \tag{1.12}$$

with $\phi_n(x) = \int_x^b a_n(t)\,dP(t)$.

Proof:

i): We have

$$\begin{aligned}\|\tilde\theta\|^2_{Q_P} &= \langle Aa_0, \tilde\theta\rangle_{Q_P}\\ &= \langle a_0, A^*\tilde\theta\rangle_P\\ &= \langle a_0, \tilde\kappa_P\rangle_P\end{aligned}$$

ii): Under the extra conditions, this can be rewritten as

$$\begin{aligned}\langle a_0, \tilde{\kappa}_P \rangle_P &= \int_a^b a_0(t) \left[\int_a^t \tilde{\kappa}'_P(x) dx \right] dP(t) + \int_a^b a_0(t) \tilde{\kappa}_P(a) dP(t) \\ &= \int_a^b \tilde{\kappa}'_P(x) \phi_0(x) dx\end{aligned}$$

using $\int a_0 dP = 0$.

iii): Similar to i) and ii).

□

Remarks.

I): Although, in iii), the sequence $\{a_n\}$ cannot converge if $\tilde{\theta} \neq \mathcal{R}(A)$, it may happen that ϕ_n does converge, say to ϕ_0. Then (1.12) becomes

$$\|\tilde{\theta}\|^2_{Q_P} = \int_a^b \tilde{\kappa}'_P(x) \phi_0(x) dx$$

This ϕ_0 may fail to be continuous. Examples will be given in the next section.

II): In the interval censoring model, the function

$$\phi(x) := \int_x^{\tau_M} a(t)\, dF(t) \text{ with } a \in L_2^0(F).$$

appears explicitly in the score operator L_1. Therefore it will play an important role. It will be called the *integrated score function.* From its definition we know that ϕ satisfies $\phi(\tau_0) = \phi(\tau_M) = 0$ and that ϕ is continuous for $F \in \mathcal{F}_S$.

1.4 Lower bound computations

In this section, the lower bound computations will be given for the different interval censoring cases. Actually, "computations" may not be the right description of what will be done. Apart from case 1 and some special choices of F, K and h in case 2, the computations do not lead to a formula for the lower bound. Solvability of the score equation will be the main topic, and the structure of the lower bound will be investigated. The derivation for case 1 also appeared in e.g. VAN DER VAART (1991), GW, BICKEL *et al.* (1993) and HUANG AND WELLNER (1995A). However, some extra remarks will be made that have not been made in these references, especially with respect to the situation with $\tilde{\theta} \in \mathcal{R}(L_1)$. The lower bound theory for case 2 is new. A subdivision is made. In case 2A, the (U, V)-distribution has no mass around the diagonal, meaning that U_i and V_i cannot come arbitrarily close. This case is also treated in GESKUS AND GROENEBOOM (1995A,1995B,1996A). In case 2B, the observation times are allowed to get arbitrarily close, implying that, asymptotically, part of the event times get arbitrarily close to being direct observations. This case is also treated in Geskus and Groeneboom (1996b). Case A and case B are treated separately, since the techniques used are much simpler in case A. Both lower bound calculations are given in the next subsections. The distribution theory for case 2A will be treated in detail in the next chapter. The distribution theory for case 2B will only be given partially. The many technical details that are needed can be found in Geskus and Groeneboom (1996b). With respect to the asymptotic behaviour of the NPMLE of the distribution function, case 2A is in many respects more similar to case 1 than to case 2B. In the next chapter, more will be said about this distinction between case 2A and case 2B. For case k, which covers everything not covered by case 1 and case 2, the operator equation is completely similar in structure to the one for case 2. In section 1.5, the explicit form of the solution is given in case 2, for $K(F) = E_F(X)$ and some specific choices of F and H.

For all interval censoring cases, the basic ingredients are the model assumptions (M1) to (M3), and the score equation

$$L_1^*\tilde{\theta} = \tilde{\kappa}_F,$$

with $\tilde{\theta} \in \overline{\mathcal{R}(L_1)}$. Most attention will be given to the situation

$$L_1^* L_1 a = \tilde{\kappa}_F.$$

The operator L_1^* is given by

$$L_1^* b = E[\, b(\text{observables}) \mid X\,].$$

It is the adjoint of the operator L_1, given by

$$L_1 a = E[\, a(X) \mid \text{observables}\,].$$

The combination $L_1^* L_1$ is called the *information operator*.

Fredholm integral equations

An important theory for the cases beyond case 1 is the theory on *Fredholm integral equations of the second kind.* A Fredholm integral equation of the second kind has the general form

$$\phi(x) - \int_a^b K(x,t)\,\phi(t)\,dt = r(t), \qquad x \in [a,b].$$

Because of the free occurrence of $\phi(x)$ it is called an equation of the second kind; the fixed range of integration makes it a Fredholm integral equation. In 1903, I. Fredholm investigated the solvability of such equations. Later, in 1918, F. Riesz extended the results to general operator equations of the form $\phi - A\phi = r$, with A being a compact operator. This theory can be found in many textbooks. In Kress (1989), chapters 1 to 4, a general account on the theory is given. We restrict ourselves to giving the most important results for our situation. One theorem from Reed and Simon (1972) is used as well.

A linear operator $A\colon X \to Y$ from a normed space X into a normed space Y is called *compact* if it maps each bounded set in X into a relatively compact set in Y. An equivalent condition is to say that for each bounded sequence $\{x_n\}$ in X, the sequence $\{y_n\} = \{Ax_n\}$

contains a convergent subsequence in Y. A compact operator with values in an infinite dimensional space in some respects almost behaves as a finite dimensional operator. It shares the property with finite dimensional linear operators that each bounded set is mapped to a totally bounded one. If we have a compact operator from a Banach space X into itself, we know that the set of eigenvalues has at most one limit point, namely 0, and for any $\varepsilon > 0$, the number of eigenvalues λ with $|\lambda| > \varepsilon$ is finite. The important property for us that is shared with finite dimensional linear operators is the following one:

Theorem 1.4.1 *Let X be a normed linear space, and let A: $X \to X$ be a compact linear operator.*
If $I - A$ is injective, then the inverse operator $(I - A)^{-1}$: $X \to X$ exists and is bounded.

Proof: See KRESS (1989), theorem 3.4

So if the homogeneous equation

$$\phi - A\phi = 0. \tag{1.13}$$

only has the trivial solution $\phi = 0$, then for each $r \in X$ the inhomogeneous equation

$$\phi - A\phi = r$$

has a unique solution $\phi \in X$ and this solution depends continuously on r, with respect to the norm of X.

Note that hardly any restrictions are imposed on the space X. For example, it need not be a complete space.
In the next chapter, an extension of theorem 1.4.1 will be used, stating what happens if the homogeneous equation has a nontrivial solution. We will only formulate it for the case $X = C([a, b])$, but more general formulations are possible. As in section 4.1 in KRESS (1989), the system $\langle C([a, b]), C([a, b])\rangle$ is a dual system with the bilinear form

$$\langle \phi, \psi \rangle = \int_a^b \phi(x)\, \psi(x)\, dx, \qquad \phi,\ \psi \in C([a, b]).$$

If the kernel K is continuous, the integral operators

$$[A\phi](x) \ := \ \int_a^b K(x, y)\, \phi(y)\, dy$$

$$[B\phi]\,(x) \ := \ \int_a^b K(y, x)\, \phi(y)\, dy$$

are adjoint.

Theorem 1.4.2 *Let $X = C([a, b])$. Consider the Fredholm integral equation*

$$\phi - A\phi = r,$$

with A: $X \to X$ being a compact linear operator.

If the homogeneous equation $\phi - A\phi = 0$ has a nontrivial solution, two possibilities arise, depending on the structure of the homogeneous adjoint equation

$$\psi - B\psi = 0. \tag{1.14}$$

I) If

$$\int_a^b r(x)\psi(x)\,dx = 0 \text{ for all } \psi \text{ satisfying (1.14)}, \tag{1.15}$$

then the inhomogeneous equation

$$\phi - A\phi = r$$

is solvable, and its general solution is of the form

$$\phi = \tilde{\phi} + \sum_{k=1}^{m} \alpha_k \phi_k$$

where $\phi_1, \ldots, \phi_m$ are linearly independent solutions of the homogeneous equation, $\tilde{\phi}$ denotes a particular solution of the inhomogeneous equation and $\alpha_1, \ldots, \alpha_m$ are arbitrary real or complex numbers. Moreover

$$m = \dim(\mathcal{N}(I - A)) = \dim(\mathcal{N}(I - A^*))$$

II) If (1.15) is not satisfied, then the inhomogeneous equation is unsolvable.

Proof: See Kress (1989), corollary 3.7, theorem 4.3, theorem 4.15 and theorem 4.17.

For integral equations of the form $A\phi = r$ (these are called integral equations of the first kind), possible existence of a solution is much harder to prove. It is the structure $(I - A)$ that makes the problem tractable.

In order to establish solvability of Fredholm integral equations of the second kind, the first thing to do is to prove compactness of the operator

$$A\colon \phi(\cdot) \longmapsto \int_a^b K(\cdot, t)\,\phi(t)\,dt. \tag{1.16}$$

Therefore, we have to specify X and its norm. Note that the solvability condition (1.13) is not related to any norm. So, after X has been determined, its norm is of minor importance with respect to showing solvability, as long as the norm makes the integral operator into a compact one. More attention should be given to X. X is determined by the kind of solution we expect, which in turn is determined by the structure of the kernel and the function r on the right-hand side. At the same time, it should be chosen such that $I - A$ is injective, hence it should not be chosen too large. In our situation, the integrated score function plays the role of ϕ in the integral equation. As long as we look at the situation with $\tilde{\theta} \in \mathcal{R}(L_1)$, ϕ should be continuous. If the conditions on the distributions are such that the kernel $K(x, t)$ is continuous, then $A\phi$ is a continuous function for each continuous ϕ. So, if r is continuous as well, X can be chosen to be $C([a, b])$, the space of continuous functions on $[a, b]$. If this space is equipped with the supremum norm, we can use theorem 2.20 in Kress (1989):

Theorem 1.4.3 *Let $X = C([a,b])$, and let K be a continuous kernel. Then the integral operator is compact with respect to the supremum norm.*

An application of the Arzelà-Ascoli theorem is the main step in the proof of this theorem. Note that compactness also holds if we supply $C([a,b])$ with the L_2-norm. See KRESS (1989), problem 2.5.
We will also allow for kernels having the following property.

(C) For each x, $K(x,\cdot)$ and $K(\cdot,x)$ are bounded real-valued functions, right-continuous with left limits (cadlag). The points of jump belong to a finite (possibly empty) set E, independent of x.

With such a kernel, the space X has to be extended. Each cadlag function is mapped to a cadlag function, so X can be taken to be the space of cadlag functions $D([a,b])$. Then compactness of the integral operator can be proved with respect to the supremum norm, using an extension of the Arzelà-Ascoli theorem. Note that a cadlag function on a compact set is bounded.

Theorem 1.4.4 *The operator A as defined in (1.16), with a kernel K satisfying (C), is a compact operator on the space of cadlag functions $(D([a,b]), \|\cdot\|_\infty)$.*

Proof: If the set E of points of discontinuity τ_i of the functions $K(x,\cdot)$ and $K(\cdot,x)$ is empty, the result is an easy consequence of the Arzelà-Ascoli theorem. So suppose that $\tau_i \in (a,b)$ is a discontinuity point and that τ_{i-1} is the preceding discontinuity point or is equal to a if there is no such point. Since the functions $K(\cdot,x)$ have left-hand limits for each $x \in [a,b]$, we can modify these functions on $[\tau_{i-1},\tau_i]$ by making them left-continuous at τ_i. Let (f_n) be a bounded sequence in $D([a,b])$. Then the sequence of functions

$$x \mapsto (Af_n)(x),\ x \in [\tau_{i-1},\tau_i], n = 1,2\ldots,$$

for the modified kernel K is equicontinuous and hence has a convergent subsequence in $(D([\tau_{i-1},\tau_i]), \|\cdot\|_\infty)$, converging to a continuous function $g : [\tau_{i-1},\tau_i] \mapsto I\!R$. Since the same subsequence of functions, restricted to $[\tau_{i-1},\tau_i)$, converges to the restriction of g to $[\tau_{i-1},\tau_i)$, which obviously has a left-hand limit at τ_i, we get that the sequence of functions

$$x \mapsto (Af_n)(x),\ x \in [\tau_{i-1},\tau_i), n = 1,2\ldots,$$

restricted to the half-open interval $[\tau_{i-1},\tau_i)$, has a uniformly convergent sequence, converging to a function which is continuous on $[\tau_{i-1},\tau_i)$ and has a left-hand limit at τ_i.

Since we can repeat the argument for the other (at most finitely many) intervals of continuity of the (non-modified) functions $K(\cdot,x)$, $x \in [a,b]$, we get that the sequence (Af_n) has a uniformly convergent subsequence, converging to a function g which is right continuous and has left-hand limits at the points τ_i.

□

The following theorem on compactness in L_2-spaces will be used as well.

Theorem 1.4.5 *Let $X = L_2([a,b],\mu)$ and let K be a kernel satisfying*

$$\int_a^b \int_a^b [K(x,t)]^2 \, d\mu \times \mu < \infty.$$

Then the integral operator A is compact with respect to the L_2-norm.

Proof: See section VI.6 in Reed and Simon (1972). The proof of this theorem uses techniques different from an application of the Arzelà-Ascoli theorem.

Remark: The operator from theorem 1.4.5 is a Hilbert-Schmidt operator.

1.4.1 Case 1

In case 1, we have one observation time T_i for each unobservable event time X_i. Suppose the observation times T_i to have an absolutely continuous distribution function G with a density g, and a support $[\eta_0, \eta_M]$. Let the event times X_i have an absolutely continuous distribution function F with a density f and a support $[\tau_0, \tau_M]$. We assume X_i and T_i to be independent (compare assumption (M2)) and F to be dominated by G (compare assumption (M3)). Note that strict inequality $\eta_0 < \tau_0$ and/or $\tau_M < \eta_M$ implies that on part of the event times no information at all may be obtained, since the corresponding observation time is outside the support of F. The score operator L_1 has the form

$$\begin{aligned} [L_1 a](t,\delta) &= E[a(X) | T = t, \Delta = \delta] \\ &= \frac{\delta \int_{\tau_0}^t a\, dF}{F(t)} + \frac{(1-\delta)\int_t^{\tau_M} a\, dF}{1-F(t)} \quad \text{a.e.} - [Q_{F,G}] \end{aligned}$$

with adjoint

$$\begin{aligned} [L_1^* b](x) &= E[b(T,\Delta) | X = x] \\ &= \int_{t=x}^{\eta_M} b(t,1)\, g(t)\, dt + \int_{t=\eta_0}^{x} b(t,0)\, g(t)\, dt \quad \text{a.e.-}[F]. \end{aligned}$$

First consider the case $\tilde{\theta} \in \mathcal{R}(L_1)$. Then the score equation

$$L_1^* L_1 a = \tilde{\kappa}_F$$

has to be solved in $a \in Ł_2^0(F)$. $L_1^* L_1$ has the form

$$[L_1^* L_1 a](x) = \int_{u=\tau_0}^{\tau_M} K(x,u)\, a(u)\, dF(u),$$

with kernel

$$K(x,u) = \int_{t=\eta_0}^{x \wedge u} \tfrac{1}{1-F(t)} dG(t) + \int_{t=x\vee u}^{\eta_M} \tfrac{1}{F(t)} dG(t).$$

Since $\int\int K^2(x,u)\,dF(u)\,dF(x)$ can be shown to be finite (see also section 3.2 in part I of GW), we can apply theorem 1.4.5 to obtain compactness of $L_1^*L_1$ with respect to the $L_2(F)$-norm. The composition $A \circ B$ of two operators is compact whenever one of them is compact, unless the other one is unbounded. Since the identity operator is not compact in any infinite dimensional space, $(L_1^*L_1)^{-1}$ does not exist as a bounded operator, and existence of a solution cannot be shown directly.

We will follow another approach. $L_1^*L_1 a = \tilde{\kappa}_F$ can be written as an equation in the integrated score function $\phi(x) = \int_x^{\tau_M} a dF$, having the form

$$\int_{t=\tau_0}^{x} \frac{\phi(t)}{1-F(t)}\, g(t)\, dt - \int_{t=x}^{\tau_M} \frac{\phi(t)}{F(t)}\, g(t)\, dt = \tilde{\kappa}_F(x) \qquad \text{a.e.-}[F].$$

Since $\phi(t) = 0$ if $t \le \tau_0$ and $t \ge \tau_M$, integration starts at τ_0 and ends at τ_M.
We assume

(G1) g is continuous, with $g(x) > 0$ for all $x \in [\tau_0, \tau_M]$

(K1) $\tilde{\kappa}_F$ is continuously differentiable

By taking derivatives on both sides, we get,

$$\frac{\phi(x)}{1-F(x)} g(x) + \frac{\phi(x)}{F(x)} g(x) = \tilde{\kappa}'_F(x), \tag{1.17}$$

from which the following expression for ϕ is obtained

$$\phi(x) = \tilde{\kappa}'_F(x) \frac{F(x)[1-F(x)]}{g(x)}. \tag{1.18}$$

Thus the canonical gradient is

$$\begin{aligned} \tilde{\theta}_F(t,\delta) &= -\delta \frac{\phi(t)}{F(t)} + (1-\delta)\frac{\phi(t)}{1-F(t)} \\ &= \begin{cases} -\tilde{\kappa}'_F(t)\frac{1-F(t)}{g(t)} & \text{if } \delta = 1 \\ \tilde{\kappa}'_F(t)\frac{F(t)}{g(t)} & \text{if } \delta = 0 \end{cases} \end{aligned}$$

and the information lower bound is, using theorem 1.3.3.ii)

$$\begin{aligned} \|\tilde{\theta}\|^2_{Q_{F,G}} &= \int_{\tau_0}^{\tau_M} \tilde{\kappa}'_F(x)\phi(x)\, dx \\ &= \int_{\tau_0}^{\tau_M} [\tilde{\kappa}'_F(x)]^2\, \frac{F(x)[1-F(x)]}{g(x)}\, dx. \end{aligned}$$

This is subject to the condition that $\phi(x)$ can be obtained as the integral $\int_x^{\tau_M} a(t)\, dF(t)$ over some $L_2^0(F)$-function a. Since the derivative $a = d\phi/dF$ is equal to

$$\frac{\tilde{\kappa}'}{g}(1-2F) + F(1-F)\frac{d}{dF}\left[\frac{\tilde{\kappa}'}{g}\right],$$

the condition

(C1) $(\tilde{\kappa}'/g) \circ F^{-1}$ is Lipschitz on $[0, 1]$

is sufficient to make a square integrable. The same Lipschitz condition (C1) is used by HUANG AND WELLNER (1995A) to prove asymptotic optimality of the NPMLE for linear functionals in case 1. I will come back to this proof in the next chapter. The condition $\int a dF = 0$ is fulfilled because $\phi(\tau_0) = 0$.

If we use the somewhat stronger assumptions that both f and g are bounded away from zero on $[\tau_0, \tau_M]$, with g having bounded derivative g', and that $\frac{d}{dx}\tilde{\kappa}'$ is bounded, we obtain that

$$\frac{d\phi}{dF} = \frac{\tilde{\kappa}'}{g}(1 - 2F) + \frac{F(1-F)}{g}\left[\frac{d\tilde{\kappa}'}{dF} - \frac{\tilde{\kappa}'}{g}\frac{dg}{dF}\right]$$

is an $L_2(F)$-function. In GW, the NPMLE of the mean is proved asymptotically to attain the efficiency bound under these stronger conditions, using a method of proof which is different from the one in HUANG AND WELLNER (1995A).

It may happen that $\tilde{\theta}$ is contained in $\overline{\mathcal{R}(L_1)}\backslash\mathcal{R}(L_1)$, as is illustrated by the following examples.

Example 1. Take $F(x) = x$ on $[0, 1]$, $g(x) = \frac{3}{2}\sqrt{x}$ on $[0, 1]$ and $K(F) = E_F(X)$. Then we have $\tilde{\kappa}_F(x) = x - \frac{1}{2}$, and $\phi(x) = \frac{2}{3}\sqrt{x}(1-x)$, the latter implying $a(x) = \phi'(x) = \sqrt{x} - 1/(3\sqrt{x})$, which is not a square integrable function. However, if we take the sequence $\{a_n\} \subset L_2^0(F)$, given by

$$a_n(x) = \begin{cases} \sqrt{x} - \frac{1}{3}\sqrt{n} & \text{if } 0 \le x \le 1/n \\ \sqrt{x} - 1/(3\sqrt{x}) & \text{if } 1/n \le x \le 1, \end{cases}$$

each a_n is continuous. $\{a_n\}$ does not converge, but we have, pointwise,

$$\lim_{n\to\infty} \int_x^1 a_n(t)\, dt = \phi(x),$$

and, in L_2-norm,

$$\tilde{\theta}_F = \lim_{n\to\infty} L_1 a_n$$

Example 2. Another class of examples for which $\tilde{\theta}$ is contained in $\overline{\mathcal{R}(L_1)} \setminus \mathcal{R}(L_1)$ arises if g or $\tilde{\kappa}'$ has jumps. As long as $g(x) \ge c > 0$ for each x, we can define a function ϕ as in (1.18). Since this ϕ is not continuous, $\tilde{\theta}$ cannot be contained in $\mathcal{R}(L_1)$. However, at least if $f \ge c' > 0$ at these points of discontinuity, ϕ can be approached by continuous ϕ_n that satisfy $\phi_n(x) = \int_x^{\tau_M} a_n(t)\, dF(t)$, with $a \in L_2(F)$. For example, one can take $\phi_n = F(1-F)\gamma_n$, with γ_n linear with slope of order n near the jump, and otherwise equal to $\tilde{\kappa}'/g$. When ϕ_n is chosen in this way, we again have $\tilde{\theta}_F = \lim_{n\to\infty} L_1 a_n$ with respect to the L_2-distance.

1.4.2 Case 2

Functionals that are pathwise differentiable in case 1, are likely to be the same in case 2, since it is a more informative situation. Indeed, the following function solves $L_1^* b = \tilde{\kappa}_F$:

$$b(u,v,\delta,\gamma) = \begin{cases} -\tilde{\kappa}'_F(u)\frac{1-F(u)}{h_1(u)} & \text{if } \delta = 1 \\ 0 & \text{if } \gamma = 1 \\ \tilde{\kappa}'_F(v)\frac{F(v)}{h_2(v)} & \text{if } \delta = \gamma = 0 \end{cases} \tag{1.19}$$

Since this function neglects the middle part, $\gamma = 1$, one may already suspect that it is not the canonical gradient $\tilde{\theta}$, meaning that is contained in $L_2^0(Q_{F,H}) \setminus \overline{\mathcal{R}(L_1)}$. We are left with the problem to project this gradient into $\overline{\mathcal{R}(L_1)}$ in order to obtain the information lower bound. However, outside finite dimensional spaces, there is no standard way to perform projections. So we leave this approach, and try to solve the operator equation $L_1^* \tilde{\theta} = \tilde{\kappa}_F$ directly.

We will mainly consider the case $\tilde{\theta} \in \mathcal{R}(L_1)$. So solvability of the equation

$$\tilde{\kappa}_F = L_1^* L_1 a$$

in the variable $a \in L_2^0(F)$ will be the point of interest. We follow the same approach as in case 1. $\tilde{\kappa}_F = L_1^* L_1 a$ can be reformulated as an equation in ϕ:

$$\begin{aligned} \tilde{\kappa}_F(x) \;\; = \;\; & - \int_{u=x}^{\tau_M} \int_{v=u}^{\eta_M} \tfrac{\phi(u)}{F(u)}\, h(u,v)\, dv\, du \\ & - \int_{u=\eta_0}^{x} \int_{v=x}^{\eta_M} \tfrac{\phi(v)-\phi(u)}{F(v)-F(u)}\, h(u,v)\, dv\, du \\ & + \int_{v=\tau_0}^{x} \int_{u=\eta_0}^{v} \tfrac{\phi(v)}{1-F(v)}\, h(u,v)\, du\, dv \qquad \text{a.e.-}[F]. \end{aligned} \tag{1.20}$$

The support of F may consist of a finite number of disjoint intervals. However, (1.20) is an equation a.e.-F, so we need not worry about intervals where F does not put mass. Without loss of generality we assume the support of F to consist of one interval $[\tau_0, \tau_M]$.

Differentiating equation (1.20) on both sides, and writing $k(x)$ instead of $\tilde{\kappa}'_F(x)$, yields the following equation:

$$\phi(x) + d(x)\left[\int_{t=\eta_0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\, h(t,x)\, dt - \int_{t=x}^{\eta_M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\, h(x,t)\, dt\right] = k(x)d(x) \tag{1.21}$$

with $d(x)$ being the function

$$d(x) = \frac{F(x)[1-F(x)]}{h_1(x)[1-F(x)] + h_2(x)F(x)}. \tag{1.22}$$

Here h_1 and h_2 denote the marginal densities of U and V respectively.

Unlike case 1, taking derivatives does not yield an explicit formula for ϕ. Also, further differentiation does not simplify things. We will investigate the solvability of equation (1.21), and whether the canonical gradient $\tilde{\theta}$ obtained via the solution ϕ is in $\mathcal{R}(L_1)$. Moreover, solvability of (1.21), and properties of this solution, will be investigated for convex combinations

$$F = (1-\alpha)F_0 + \alpha \hat{F}_n,$$

where $F_0 \in \mathcal{F}_S$ (the unknown distribution) is continuous and $\hat{F}_n$ (the NPMLE of F_0) is purely discrete. This will be needed in the next chapter on the NPMLE. The function k, however, remains completely determined by the underlying distribution F_0 (so $k = \tilde{\kappa}'_{F_0}$). Therefore we write k instead of $\tilde{\kappa}'_F$.

If (1.21) is solvable, its solution ϕ can be shown to contain a factor $F(1-F)$, just like in case 1. The structure of d already suggests this factor to be present. We will essentially need this property in the next chapter. Validity of the factorization is shown by inserting

$$\phi = F(1-F)\,\xi$$

in (1.21). Some reordering yields an integral equation in ξ:

$$\xi(x) + c(x)\left[\int_{t=\tau_0}^{x} \tfrac{\xi(x)-\xi(t)}{F(x)-F(t)}\, h^\circ(t,x)\,dt - \int_{t=x}^{\tau_M} \tfrac{\xi(t)-\xi(x)}{F(t)-F(x)}\, h^\circ(x,t)\,dt\right] = k(x)c(x) \tag{1.23}$$

with $c(x)$ given by

$$\begin{aligned} 1/c(x) &= \int_{t=\eta_0}^{x} [1-F(t)]\, h(t,x)\,dt + \int_{t=x}^{\eta_M} F(t)\, h(x,t)\,dt \\ &= h_2(x)\, E[1-F(U)|V=x] + h_1(x)\, E[F(V)|U=x] \end{aligned} \tag{1.24}$$

and

$$\begin{cases} h^\circ(t,x) = F(t)[1-F(t)]\, h(t,x) & \text{if } t \le x \\ h^\circ(x,t) = F(t)[1-F(t)]\, h(x,t) & \text{if } x \le t \end{cases} \tag{1.25}$$

Note that, on $[\eta_0, \tau_0]$ and $[\tau_M, \eta_M]$, $h^\circ(\cdot, x)$ and $h^\circ(x, \cdot)$ are zero, so the domain of integration in (1.23) can be restricted to $[\tau_0, \tau_M]$. In c^{-1}, however, integration over the whole interval $[\eta_0, \eta_M]$ is performed.
If (1.23) is solvable as well, then the factorization indeed holds. The lemmas and theorems in the rest of this section apply to both the ϕ-equation (1.21) and the ξ-equation (1.23), since they are very similar in structure.

We briefly pay some attention to the situation with a combination of type 1 and type 2 censoring. The most natural way to model such a combination, is to split the observation time distribution into a univariate case 1 part and a bivariate case 2 part. If $h^\oplus$ denotes the density of the combined distribution, we can write

$$h^\oplus = \beta_1 h^{(1)} + \beta_2 h^{(2)},$$

with $h^{(1)}$ and $h^{(2)}$ denoting the densities, conditioned on the univariate and bivariate part respectively, and β_1 and β_2 the fractions of the combined observation time distribution yielded by both parts. We have $\beta_1 + \beta_2 \le 1$, with equality if to each event time corresponds at least one observation time on $[\tau_0, \tau_M]$. Note that $h^{(1)}$ and $h^{(2)}$ have their mass concentrated on

$[\tau_0, \tau_M]$ and $\{(u,v) | \tau_0 \le u < v \le \tau_M\}$ respectively. We obtain an integral equation which is similar in structure to (1.21), but with $d(x)$ replaced by

$$d^{\oplus}(x) = \frac{F(x)[1-F(x)]}{\beta_1 h^{(1)}(x) + \beta_2 \left\{h_1^{(2)}(x)[1-F(x)] + h_2^{(2)}(x)F(x)\right\}}. \tag{1.26}$$

The function $d^{\oplus}$ can also be obtained from (1.20) directly. For if we split off the part of h in (1.20) outside $\{\tau_0 \le u < v \le \tau_M\}$, we obtain (1.21), with η_0 replaced by τ_0, η_M replaced by τ_M, and the denominator of $d(x)$ given by

$$\left[\int_{u=\eta_0}^{\tau_0} h(u,x)\,du + \int_{v=\tau_M}^{\eta_M} h(x,v)\,dv\right] \\ + \left[\int_{v=x}^{\tau_M} h(x,v)\,dv\right][1-F(x)] + \left[\int_{u=\tau_0}^{x} h(u,x)\,du\right]F(x).$$

So $\beta_1 h^{(1)}$ can be obtained by considering the mass of the bivariate observation time distribution h outside the support of F, namely on $[\eta_0, \tau_0] \times [\tau_0, \tau_M]$ and on $[\tau_0, \tau_M] \times [\tau_M, \eta_M]$. $\beta_2 h^{(2)}$ is the mass of this observation time distribution on $\{(u,v) | \tau_0 \le u < v \le \tau_M\}$. The remaining part of $\{(u,v) | \eta_0 \le u < v \le \eta_M\}$ contains the observation time mass that corresponds to the situation with no information at all. From these observations, and especially formula (1.26), we see that restricting to the strict case 2 situation does not lead to loss of generality. So integration in (1.21) and (1.24) will be considered from τ_0 to τ_M, and formula (1.22) will be used.

Apart from the model conditions $(M1)$ to $(M3)$, some extra conditions will have to be introduced in order to make the proofs in this section possible. We suppose

(H1) h_1 and h_2 are continuous, with $h_1(x) + h_2(x) > 0$ for all $x \in [\tau_0, \tau_M]$

(H2) $h(u,v)$ is continuous

(F1) The density of F_0 is bounded away from zero, say $f_0(x) \ge c$ for all $x \in [\tau_0, \tau_M]$

(K1) $\tilde{\kappa}'_{F_0} = k$ is continuous

Of course, (H2) implies continuity of h_1 and h_2, which is also stated in (H1). However, (H2) can be relaxed (see remark II after theorem 1.4.7). (H1) is the analogue of $g > 0$ in case 1. Note that (H1) implies that the functions c and d are bounded. The above conditions are sufficient to prove solvability of the integral equation for continuous ϕ in case 2A. Showing solvability in case 2B, and, for both case 2A and 2B, showing that ϕ is an integrated score function, i.e. $\phi(x) = \int_x^{\tau_M} a(t)\,dF(t)$ for some $a \in L_2^0(F)$, requires some more conditions. These are:

(H3) The partial derivatives $\Delta_t^1(x) = \frac{\partial}{\partial x} h(x,t)$ and $\Delta_t^2(x) = \frac{\partial}{\partial x} h(t,x)$ exist, except for at most a finite number of points x, where left and right derivatives with respect to x do exist for each t. The derivatives are bounded, uniformly in t and x.

(K2) k is differentiable, except for at most a finite number of points x, where left and right derivatives exist. The derivative is bounded, uniformly in x.

With respect to the NPMLE $\hat{F}_n$ we assume

(CF) $\hat{F}_n$ is a non-degenerate, non-defective, piecewise constant distribution function with at most finitely many points of jump. Let $D = \{\tau_0, \tau_1, \ldots, \tau_m, \tau_M\}$ denote the ordered set of jump points of $\hat{F}_n$, augmented with the endpoints of its support. Moreover we assume $\hat{F}_n$ to satisfy

$$\sup_{x \in [\tau_0, \tau_M]} |\hat{F}_n(x) - F_0(x)| < \epsilon$$

for some ϵ to be determined. The class of functions thus obtained is denoted by $\mathcal{F}$. Note that ϵ has to be uniform over $\hat{F}_n \in \mathcal{F}$.

Note that (CF) does not hold for all possible realizations of $\hat{F}_n$. However, by the strong uniform consistency of $\hat{F}_n$ (see the next chapter), together with condition (F1), it always holds for n sufficiently large.

If the integral equation (1.21) has solution ϕ, the canonical gradient $\tilde{\theta}_F \in \mathcal{R}(L_1)$ has the form:

$$\tilde{\theta}_F(u, v, \delta, \gamma) = -\delta\,\frac{\phi_F(u)}{F(u)} - \gamma\,\frac{\phi_F(v) - \phi_F(u)}{F(v) - F(u)} + (1 - \delta - \gamma)\,\frac{\phi_F(v)}{1 - F(v)} \tag{1.27}$$

Now one can see that the gradient b, defined at the beginning of this section and solving $L_1^* b = \tilde{\kappa}_F$, is indeed not the canonical gradient since the middle part of the canonical gradient is not zero.

A: Observation times bounded away

For this case, we only have to look at the situations $\alpha = 0$ ($F = \hat{F}_n$) and $\alpha = 1$ ($F = F_0$). When looking at equations (1.21) and (1.23), we see a singularity of the form $1/(F(v) - F(u))$ appearing in the kernel, implying that it does not belong to one of the standard integral equations. However, in case A the singularity vanishes. Formally we suppose:

(H4) h does not have mass close to the diagonal, i.e. $\text{Prob}\{V - U \le \epsilon_0\} = 0$ for some $\epsilon_0 > 0$.

The ϵ in condition (CF), determining $\mathcal{F}$, is chosen in such a way that $\hat{F}_n(v) - \hat{F}_n(u)$ remains bounded away from zero as long as $v - u > \epsilon_0$. This implies that (1.21) can be rewritten as a Fredholm integral equation of the second kind. The kernel is equal to

$$K(x, t) = \frac{d(x)\,D(x, t)}{1 + d(x)\int_{\tau_0}^{\tau_M} D(x, t)\,dt}, \tag{1.28}$$

with $D(x, t)$ defined as

$$D(x, t) := \begin{cases} \frac{h(t,x)}{F(x) - F(t)} & \text{if } t \le x \\ \frac{h(x,t)}{F(t) - F(x)} & \text{if } t \ge x \end{cases} \tag{1.29}$$

The function r in the general Fredholm form $\phi - A\phi = r$ is in our situation

$$r(x) = \frac{k(x) d(x)}{1 + d(x)\int_{\tau_0}^{\tau_M} D(x, t)\,dt}. \tag{1.30}$$

For $F = F_0$ the kernel, d and r are all three continuous. So we can take $X = C([\tau_0, \tau_M])$ and we have compactness of the integral operator by theorem 1.4.3. For $F = \hat{F}_n \in \mathcal{F}$ the kernel obtained satisfies condition (C) on page 25. By theorem 1.4.4 we have compactness of the integral operator on $D([\tau_0, \tau_M])$. Now we are ready to apply theorem 1.4.1, establishing solvability of both integral equations (1.21) and (1.23). The conditions stated are slightly more general.

Theorem 1.4.6 *Consider the integral equation*

$$\phi(x) + \int_{\tau_0}^{\tau_M} \left[\frac{d(x)\, D(x,t)}{1 + d(x) \int_{\tau_0}^{\tau_M} D(x,t)\, dt} \right] \phi(t)\, dt = r(x)$$

with $D \geq 0$ satisfying condition (C) on page 25. r and d are cadlag functions having at most finitely many jumps, with d being nonnegative.
This equation has a unique solution in $D([\tau_0, \tau_M])$.

Proof:
Theorem 1.4.1 will be used. So consider the homogeneous equation

$$\phi(x) = \int_{\tau_0}^{\tau_M} K(x,t)\, \phi(t)\, dt \qquad \text{for all } x \in [\tau_0, \tau_M]$$

This equation is equivalent to

$$\phi(x) + \left[d(x) \int_{\tau_0}^{\tau_M} D(x,t)\, dt \right] \phi(x) = d(x) \int_{\tau_0}^{\tau_M} D(x,t)\, \phi(t)\, dt, \tag{1.31}$$

for all $x \in [\tau_0, \tau_M]$. Suppose $\phi(x) \neq 0$ for some x. Without loss of generality, we may assume $\phi(x) > 0$.
If the supremum is attained, say at s, we get, since $D(s,t) \geq 0$ and $d(x) \geq 0$,

$$d(s) \int_{\tau_0}^{\tau_M} D(s,t)\, \phi(t)\, dt \leq \left[d(s) \int_{\tau_0}^{\tau_M} D(s,t)\, dt \right] \phi(s).$$

The right-hand side is strictly smaller than

$$\phi(s) + \left[d(s) \int_{\tau_0}^{\tau_M} D(s,t)\, dt \right] \phi(s)$$

which contradicts equation (1.31).
It may happen that ϕ jumps downward just before the supremum is attained:

$$\sup_{x \in [\tau_0, \tau_M]} \phi(x) = \phi(s-) > \phi(s)$$

Then one can find a $\delta > 0$ such that $\phi(s-\delta) > 0$ and

$$d(s-\delta) \int_{\{t: \phi(t) > \phi(s-\delta)\}} D(s-\delta, t)\, \phi(t)\, dt \leq \tfrac{1}{2}\, \phi(s-\delta)$$

Hence

$$d(s-\delta)\int_{\tau_0}^{\tau_M} D(s-\delta,t)\,\phi(t)\,dt \le \left[d(s-\delta)\int_{\tau_0}^{\tau_M} D(s-\delta,t)\,dt\right]\phi(s-\delta)+\tfrac{1}{2}\,\phi(s-\delta),$$

again contradicting (1.31).

□

If $F = F_0$, and d, h and r are all continuous, the solution ϕ is contained in $C([\tau_0, \tau_M])$. However, by (H3) and (K2), it may not be differentiable everywhere. At some points it may only have separate left and right derivatives. The next theorem proves that the derivatives are bounded, uniformly in x, which yields $\tilde{\theta} \in \mathcal{R}(L_1)$.

Theorem 1.4.7 *Let the conditions (M1) to (M3) on page 15, (H1) to (H4), (F1) and (K1) and (K2) be satisfied. Then the score equation $\tilde{\kappa}_{F_0} = L_1^* L_1 a$ is solvable.*

Proof:
Taking derivatives on both sides of the integral equation, using left and right derivatives if necessary, existence of, possibly different, left and right derivatives ϕ' is shown. Then we obtain, writing F and f instead of F_0 and f_0:

$$\begin{aligned}
\phi'(x) &= d(x)k'(x)\\
&\quad +d'(x)\phi(x)/d(x)\\
&\quad -d(x)\Big\{\int_{\tau_0}^{x} \tfrac{\phi'(x)}{F(x)-F(t)}\,h(t,x)\,dt+\int_{x}^{\tau_M} \tfrac{\phi'(x)}{F(t)-F(x)}\,h(x,t)\,dt\Big\}\\
&\quad +d(x)\left\{\int_{\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{(F(x)-F(t))^2}\,f(x)h(t,x)\,dt+\int_{x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{(F(t)-F(x))^2}\,f(x)h(x,t)\,dt\right\}\\
&\quad -d(x)\Big\{\int_{\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\,\tfrac{\partial}{\partial x}h(t,x)\,dt-\int_{x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\,\tfrac{\partial}{\partial x}h(x,t)\,dt\Big\}.
\end{aligned}$$

We have $\phi/d = \xi\,[h_1(1-F)+h_2F]$. Bringing everything containing ϕ' to the left-hand side, we obtain

$$\begin{aligned}
&\phi'(x)\Big\{1+d(x)\int_{\tau_0}^{x} \tfrac{1}{F(x)-F(t)}\,h(t,x)\,dt+d(x)\int_{x}^{\tau_M} \tfrac{1}{F(t)-F(x)}\,h(x,t)\,dt\Big\}\\
&= d(x)k'(x)\\
&\quad +d'(x)\,\xi(x)\,[h_1(x)[1-F(x)]+h_2(x)F(x)]\\
&\quad +d(x)\left\{\int_{\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{(F(x)-F(t))^2}\,f(x)h(t,x)\,dt+\int_{x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{(F(t)-F(x))^2}\,f(x)h(x,t)\,dt\right\}\\
&\quad -d(x)\Big\{\int_{\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\,\tfrac{\partial}{\partial x}h(t,x)\,dt-\int_{x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\,\tfrac{\partial}{\partial x}h(x,t)\,dt\Big\}.
\end{aligned}$$

The right-hand side is bounded, uniformly over x. Since the part between curly brackets on the left-hand side is bounded away from zero, we get boundedness of ϕ'. Using $f \ge c$, this implies $d\phi/dF \in L_2^0(F)$.

□

Remarks:

I): The conditions (F1), (H3) and (K2) can be weakened, more in the line of $(\tilde{\kappa}'_{F_0}/g) \circ F_0^{-1}$ being Lipschitz on $[0,1]$ for case 1. Instead of (F1), (H3) and (K2), it is sufficient to suppose

$$\frac{dk}{dF_0} \quad \text{is bounded}$$

and

$$\left|\tfrac{\partial}{\partial F_0} h(t,x)\right| \leq K g_1(t) \qquad \text{and} \qquad \left|\tfrac{\partial}{\partial F_0} h(x,t)\right| \leq K g_2(t)$$

with g_i independent of x and satisfying

$$\int_{\tau_0}^{\tau_M} g_i(t)\,dt < \infty.$$

II): From theorem 1.4.6 we see that solvability of the integral equation also holds if we allow for discontinuities in the simultaneous observation time density h or in k. The function ϕ thus obtained is in general no longer continuous, but will be contained in $\overline{\mathcal{R}(L_1)} \setminus \mathcal{R}(L_1)$. An example in which (H2) is violated, but continuity of the solution does hold, is: h is constant, and zero on the set $0 \leq y - x \leq \epsilon_0$ along the diagonal.

When $F = \hat{F}_n$, the same kind of boundedness property can be proved for ϕ' and ξ', uniformly over $\hat{F}_n \in \mathcal{F}$. Therefore we introduce the class of integral equations

$$\{IQ_F | \, F \in \mathcal{F}\},$$

given by

$$\phi_F(x) \;+\; d_F(x)\left[\int_{t=\tau_0}^{x} \tfrac{\phi_F(x)-\phi_F(t)}{F(x)-F(t)}\, h_F(t,x)\,dt - \int_{t=x}^{\tau_M} \tfrac{\phi_F(t)-\phi_F(x)}{F(t)-F(x)}\, h_F(x,t)\,dt\right]$$
$$= k(x) d_F(x).$$

Here $h_F \geq 0$ satisfies condition (C) on page 25, $\{d_F | \, F \in \mathcal{F}\}$ is uniformly bounded and nonnegative, and k is a cadlag function having at most finitely many jumps. Let ϕ_F be the solution to IQ_F.

First we prove a uniform boundedness property of ϕ_F itself.

Lemma 1.4.1 *The class $\{\phi_F | \, F \in \mathcal{F}\}$ is uniformly bounded.*

Proof:
Let $F \in \mathcal{F}$. Define

$$I_F(x) := \int_{t=\tau_0}^{x} \tfrac{\phi_F(x)-\phi_F(t)}{F(x)-F(t)}\, h_F(t,x)\,dt$$

and

$$J_F(x) := -\int_{t=x}^{\tau_M} \tfrac{\phi_F(t)-\phi_F(x)}{F(t)-F(x)}\, h_F(x,t)\,dt$$

So we have

$$\phi_F = d_F[k - I_F - J_F]$$

The proof is based on the observation that I_F and J_F have a reducing influence on the value of the extremum.
First suppose that the minimum and the maximum of ϕ_F are attained. Let $m = \arg\min(\phi_F)$ and $M = \arg\max(\phi_F)$.
Since ϕ_F reaches its minimum at m, both $I_F(m)$ and $J_F(m)$ should be ≤ 0. Hence, for each $x \in [\tau_0, \tau_M]$,

$$\begin{aligned} \phi_F(x) &\geq \phi_F(m) = d_F(m)k(m) - d_F(m)[I_F(m) + J_F(m)] \\ &\geq d_F(m)k(m). \end{aligned}$$

Likewise, from $I_F(M) \geq 0$ and $J_F(M) \geq 0$ we derive

$$\phi_F(x) \leq d_F(M)k(M)$$

for every x.
If the maximum is not attained, say $\sup_{x\in[\tau_0,\tau_M]} \phi_F(x) = \phi_F(M-) > \phi_F(M)$, we have

$$\phi_F(x) \leq k(M-)\, d_F(M-) \text{ for all } x.$$

If the minimum is not attained, we have $\phi_F(x) \geq k(m-)\, d_F(m-)$ for all x.
From boundedness of k and uniform boundedness of $\{d_F | F \in \mathcal{F}\}$, uniform boundedness of $\{\phi_F | \, F \in \mathcal{F}\}$ follows.

□

Remark:
From the proof we see that, if k is nonnegative, ϕ_F is nonnegative as well; likewise $k \leq 0$ implies $\phi_F \leq 0$. This also holds if $F = F_0$. So, for example, when the functional K is the mean, with $k \equiv 1$, we have $\phi \geq 0$.

The proof of the following lemma is very similar to the proof of theorem 1.4.7.

Lemma 1.4.2 *Let ϕ_F and ξ_F denote the solutions to the equations (1.21) and (1.23) respectively. The following holds:*

I. The derivative of ϕ_F at the points of continuity is bounded, uniformly over $F \in \mathcal{F}$ and the points of continuity, implying

$$|\phi_F(y) - \phi_F(x)| \leq K_1 \, |y - x|$$

if y and x are in the same interval between jumps. Here K_1 is independent of F and x and y.
The same holds for ξ_F.

II: The jumps satisfy

$$|\phi_F(x) - \phi_F(x-)| \leq K_2 \, |F(x) - F(x-)|,$$

with K_2 independent of x and F.
The same holds for ξ_F.

Proof: The denominator of d_F satisfies

$$\inf_{F\in\mathcal{F}} \inf_{x\in[\tau_0,\tau_M]} [h_1(x)[1-F(x)]+h_2(x)F(x)] > 0. \tag{1.32}$$

For let $x \in [\tau_0, \tau_M]$ be arbitrary.
If x satisfies $\tau_0 + \epsilon_0 \le x \le \tau_M - \epsilon_0$ we have, using (H4) and $F \in \mathcal{F}$,

$$h_1(x)\,[1-F(x)] + h_2(x)\,F(x) \ge c'(h_1(x)+h_2(x)) > 0.$$

If $\tau_0 \le x < \tau_0 + \epsilon_0$, we have $h_2(x) = 0$. Hence $h_1(x) > 0$ by (H1), implying $h_1(x)\,[1-F(x)] > 0$. The argument for $\tau_M \ge x > \tau_M - \epsilon_0$ runs in a similar way.
(For the denominator of c_F the argument is almost the same.)
Now the proof is almost similar to the proof of theorem 1.4.7. We only give it for ϕ_F.

I: At each continuity point x of F we have, by taking derivatives and some reordering:

$$\begin{aligned}
&\phi_F'(x)\Big\{1 + d_F(x)\int_{\tau_0}^{x} \tfrac{1}{F(x)-F(t)}\,h(t,x)\,dt + d_F(x)\int_{x}^{\tau_M} \tfrac{1}{F(t)-F(x)}\,h(x,t)\,dt\Big\}\\
&= d_F(x)\tilde{\kappa}_{F_0}'(x) + d_F'(x)\,[h_1(x)[1-F(x)]+h_2(x)F(x)]\\
&\quad -d_F(x)\Big\{\int_{\tau_0}^{x} \tfrac{\phi_F(x)-\phi_F(t)}{F(x)-F(t)}\,\tfrac{\partial}{\partial x}h(t,x)\,dt - \int_{x}^{\tau_M} \tfrac{\phi_F(t)-\phi_F(x)}{F(t)-F(x)}\,\tfrac{\partial}{\partial x}h(x,t)\,dt\Big\}.
\end{aligned}$$

By lemma 1.4.1, using (1.32), $\{\phi_F | F \in \mathcal{F}\}$ is uniformly bounded. Hence, again using (1.32), the right-hand side is bounded, uniformly over x and F. Since the part between brackets on the left-hand side is bounded away from zero, we get uniform boundedness of ϕ_F'.

II: At the points of jump x of F we get a similar expression. Define $\Delta g(x) := g(x) - g(x-)$. Then we have

$$\begin{aligned}
&\tfrac{\Delta\phi_F(x)}{\Delta F(x)}\Big\{1 + d_F(x-)\int_{\tau_0}^{x} \tfrac{1}{F(x)-F(t)}\,h(t,x)\,dt + d_F(x-)\int_{x}^{\tau_M} \tfrac{1}{F(t)-F(x)}\,h(x,t)\,dt\Big\}\\
&= \tfrac{\Delta d_F(x)}{\Delta F(x)}\xi_F(x)\,[h_1(x)[1-F(x)]+h_2(x)F(x)]\\
&\quad + d_F(x-)\int_{\tau_0}^{x} \tfrac{\phi_F(x-)-\phi_F(t)}{\{F(x)-F(t)\}\{F(x-)-F(t)\}}\,h(t,x)\,dt\\
&\quad + d_F(x-)\int_{x}^{\tau_M} \tfrac{\phi_F(t)-\phi_F(x-)}{\{F(t)-F(x)\}\{F(t)-F(x-)\}}\,h(x,t)\,dt,
\end{aligned}$$

with $\Delta d_F(x)/\Delta F(x)$ given by

$$\frac{\Delta d_F(x)}{\Delta F(x)} = \frac{[1-F(x)][1-F(x-)]h_1(x) \;-\; F(x)F(x-)h_2(x)}{\{h_1(x)[1-F(x)]+h_2(x)F(x)\}\{h_1(x)[1-F(x-)]+h_2(x)F(x-)\}}$$

Using boundedness of $\Delta d_F(x)/\Delta F(x)$, boundedness of $\Delta\phi_F(x)/\Delta F(x)$, uniformly over the points of jump, is obtained.

□

B: Observation times arbitrarily close

We now allow the observation time density to have mass around the diagonal. So condition (H4) is no longer imposed. The approach for case A can no longer be used directly. We first have to change the integral equation to make it into a Fredholm integral equation by "desingularization". The change we make is replacing $(F(v) - F(u))$ by $(F(v) - F(u)) \vee \epsilon$ for some $\epsilon \in (0,1)$. This equation is similar in structure to the one from case A, so it has a unique solution by theorem 1.4.6.

What remains to be proved is the convergence of ϕ_ϵ, as $\epsilon \downarrow 0$, to some function ϕ in $C([\tau_0, \tau_M])$ or $D([\tau_0, \tau_M])$ with respect to the supremum norm $\|\,.\,\|_\infty$. Moreover, this ϕ has to satisfy the original equation (1.21). Finally, for $F = F_0$, ϕ needs to be Lipschitz, implying that $d\phi/dF$ is an $L_2^0(F)$-function. Boundedness of ϕ_ϵ as well as $\frac{\phi_\epsilon(x)-\phi_\epsilon(t)}{(F(x)-F(t))}$, uniformly in ϵ, is needed.

The case $F = \hat{F}_n$ will not be considered in this section. We will look at convex combinations $F = (1-\alpha)F_0 + \alpha\hat{F}_n$, with $\alpha \in [0,1)$. These combinations have the advantage that they do not have intervals of constancy. If F has jumps, the solution of the integral equation will in general also have jumps. However, the key observation in analyzing the integral equation and in proving the efficiency of the NPMLE is that, even when F has discontinuities, we can make a change of scale in such a way that the solution of the integral equation can be extended to a Lipschitz function in the transformed scale.

We first introduce some notation. Let $G(y) = F^{-1}(y)$, $y \in [0,1]$, with a derivative g which exists except for at most a finite number of points, where, however, G has left and right derivatives. Furthermore, let $\bar{k}(y) = k(G(y))$, $\bar{H}(u,v) = H(G(u), G(v))$ and likewise $\bar{h}(u,v) = h(G(u), G(v))$, and let $\bar{d}_F$ be defined by

$$\bar{d}_F(y) = \frac{y(1-y)}{(1-y)\bar{h}_1(y) + y\,\bar{h}_2(y)}, \tag{1.33}$$

where $\bar{h}_i = h_i \circ G$, $i = 1,2$. Note that, if F has jumps, $\bar{d}_F \neq d_F \circ G$. Also note that $\bar{k}$, $\bar{d}$ and $\bar{h}$ are continuous. In a similar way, we define

$$\bar{c}_F(y) = \int_0^y (1-s)\,\bar{h}(s,y)\,dG(s) + \int_y^1 s\,\bar{h}(y,s)\,dG(s)$$

and $\bar{h}^\circ(s,y) = s(1-s)\bar{h}(s,y)$ and $\bar{h}^\circ(y,s) = s(1-s)\bar{h}(y,s)$.

Lemma 1.4.3 *(i) The integral equation*

$$\bar{\phi}_\epsilon(y) = \bar{d}_F(y)\Big\{\bar{k}(y) - \int_0^y \frac{\bar{\phi}_\epsilon(y)-\bar{\phi}_\epsilon(s)}{(y-s)\vee\epsilon}\,\bar{h}(s,y)\,dG(s) + \int_y^1 \frac{\bar{\phi}_\epsilon(s)-\bar{\phi}_\epsilon(y)}{(s-y)\vee\epsilon}\,\bar{h}(y,s)\,dG(s)\Big\} \tag{1.34}$$

has a unique continuous solution $\bar{\phi}_\epsilon$, satisfying

$$\inf_{x\in[\tau_0,\tau_M]} d_F(x)k(x) \le \bar{\phi}_\epsilon(y) \le \sup_{x\in[\tau_0,\tau_M]} d_F(x)k(x), \tag{1.35}$$

for all $y \in [0,1]$ and $\epsilon > 0$.
For points y in the range of F, say $y = F(x)$, we have $\bar{\phi}_\epsilon(y) = \phi_\epsilon(x)$

(ii) The integral equation

$$\bar{\xi}_\epsilon(y) = \bar{c}_F(y)\Big\{\bar{k}(y) - \int_0^y \tfrac{\bar{\xi}_\epsilon(y)-\bar{\xi}_\epsilon(s)}{(y-s)\vee\epsilon}\,\bar{h}^\circ(s,y)\,dG(s) + \int_y^1 \tfrac{\bar{\xi}_\epsilon(s)-\bar{\xi}_\epsilon(y)}{(s-y)\vee\epsilon}\,\bar{h}^\circ(y,s)\,dG(s)\Big\} \quad (1.36)$$

has a unique continuous solution $\bar{\xi}_\epsilon$, satisfying

$$\inf_{x\in[\tau_0,\tau_M]} c_F(x)k(x) \le \bar{\xi}_\epsilon(y) \le \sup_{x\in[\tau_0,\tau_M]} c_F(x)k(x), \quad (1.37)$$

for all $y \in [0,1]$ and $\epsilon > 0$.
For points y in the range of F, say $y = F(x)$, we have $\bar{\xi}_\epsilon(y) = \xi_\epsilon(x)$

Proof:
ad (i) By theorem 1.4.6, the $\bar{\phi}_\epsilon$-equation (1.34) has a unique continuous solution, for each $\epsilon > 0$. Note that the integration in (1.34) is only with respect to $dG(s)$ and therefore only involves values belonging to the range of F. So for points y in the range of F we have

$$\bar{\phi}_\epsilon(y) = \phi_\epsilon(G(y)).$$

The proof of the bounds in (1.35) is completely similar to the proof in lemma 1.4.1
ad (ii) The argument is completely similar to the argument given for (i).

□

The following lemma is the crux of the proof of the existence of the solution to the original integral equation.

Lemma 1.4.4 *The functions $\bar{\phi}_\epsilon$ are Lipschitz on $[0,1]$, uniformly in $\epsilon > 0$.*

Proof: As before, let $\tau_1, \ldots, \tau_m$ denote the points of jump of F. Furthermore, let $z_i = F(\tau_i)$, $i = 0, \ldots, m, M$. The interval $[z_i, z_{i+1}]$ can be divided into two parts:

(1) the interval $[z_i, z_i')$, where $z_i' = F(\tau_{i+1}-)$. The interval $[z_i, z_i')$ corresponds to the interval $[\tau_i, \tau_{i+1})$ in the original scale. The function G is strictly increasing and differentiable on the interval (z_i, z_i'), and is right and left differentiable at z_i and z_i' respectively.

(2) the interval $[z_i', z_{i+1}]$. This interval corresponds to the jump of F at τ_{i+1}. Here the function G is constant, again having right and left derivatives at the respective endpoints.

If $i = m$, the second interval only consists of the point 1. Let

$$\begin{aligned} D' \;=\; & \{z_0, \ldots, z_M\} \cup \{z_0', \ldots, z_m'\} \\ & \cup \{\text{discontinuity points of } \bar{k}'(y),\ \bar{d}_F'(y), \\ & \quad \Delta^1(y) = \tfrac{\partial}{\partial y}\bar{h}(y,s) \text{ for y} \le \text{s, and } \Delta^2(y) = \tfrac{\partial}{\partial y}\bar{h}(s,y) \text{ for y} \ge \text{s}\}. \end{aligned}$$

Then $\bar\phi_\epsilon(y)$ is differentiable for $y \notin D'$, and has left and right derivatives for $y \in D'$, satisfying

$$\begin{aligned}\bar\phi'_\epsilon(y) = \quad & \bar d'_F(y)\,\bar\xi_\epsilon(y)\,\Big[(1-y)\bar h_1(y) + y\,\bar h_2(y)\Big] \\ & + \bar d_F(y)\Big\{\bar k'(y) - \int_0^y \tfrac{\bar\phi_\epsilon(y)-\bar\phi_\epsilon(s)}{(y-s)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar h(s,y)\,dG(s) \\ & \qquad + \int_y^1 \tfrac{\bar\phi_\epsilon(s)-\bar\phi_\epsilon(y)}{(s-y)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar h(y,s)\,dG(s)\Big\} \\ & - \bar d_F(y)\Big\{\int_{s:y-s>\epsilon}\Big\{\tfrac{\bar\phi'_\epsilon(y)}{y-s} - \tfrac{\bar\phi_\epsilon(y)-\bar\phi_\epsilon(s)}{(y-s)^2}\Big\}\,d\bar H(s,y) \\ & \qquad + \int_{s:s-y>\epsilon}\Big\{\tfrac{\bar\phi'_\epsilon(y)}{s-y} - \tfrac{\bar\phi_\epsilon(s)-\bar\phi_\epsilon(y)}{(s-y)^2}\Big\}\,d\bar H(y,s)\Big\} \\ & - \bar d_F(y)\,\bar\phi'_\epsilon(y)\,\epsilon^{-1}\Big\{\int_{y-\epsilon}^y \bar h(s,y)g(s)\,ds + \int_y^{y+\epsilon} \bar h(y,s)g(s)\,du\Big\}. \qquad (1.38)\end{aligned}$$

Note that $\frac{\partial}{\partial y}\bar H(y,s) = \bar h(y,s)g(y)$ and similarly for the other partial derivative of $\bar H$. Moving the terms containing $\bar\phi'_\epsilon$ to the left-hand side of (1.38), shows that $\bar\phi'_\epsilon(y)$ has a finite upper bound, using lemma 1.4.3. Moreover, $\bar\phi'_\epsilon$ is piecewise continuous on the closed intervals from one point in D' to the subsequent one. So $\bar\phi'_\epsilon$ attains a maximum value, which may be a right or left derivative.

The rest of the proof is devoted to showing that this maximum value is uniform in ϵ. Let $M_\epsilon := \sup_{s\in[0,1]} \bar\phi'_\epsilon(y)$ and suppose that $\bar\phi'_\epsilon$ attains its supremum at a point M. Note that $M_\epsilon \geq 0$, since $\bar\phi_\epsilon(0) = \bar\phi_\epsilon(1) = 0$ and $\bar\phi_\epsilon$ is continuous. Then, if $0 < s < M - \epsilon$,

$$\frac{\bar\phi'_\epsilon(M)}{M-s} - \frac{\bar\phi_\epsilon(M)-\bar\phi_\epsilon(s)}{(M-s)^2} \geq \frac{\int_s^M \{\bar\phi'_\epsilon(M) - \bar\phi'_\epsilon(u)\}\,du}{(M-s)^2} \geq 0.$$

Likewise, if $1 > s > M + \epsilon$, we get

$$\frac{\bar\phi'_\epsilon(M)}{s-M} - \frac{\bar\phi_\epsilon(s)-\bar\phi_\epsilon(M)}{(s-M)^2} \geq 0.$$

So these parts work in the opposite direction, and are harmless in (1.38).

Now let $K_\epsilon(y)$ be defined by

$$\begin{aligned}K_\epsilon(y) \;:=\; & \bar d_F(y)\bar k'(y) + \bar d'_F(y)\,\bar\xi_\epsilon(y)\,\Big[(1-y)\bar h_1(y) + y\,\bar h_2(y)\Big] \\ & - \bar d_F(y)\Big\{\int_0^y \tfrac{\bar\phi_\epsilon(y)-\bar\phi_\epsilon(s)}{(y-s)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar h(s,y)\,dG(s) - \int_y^1 \tfrac{\bar\phi_\epsilon(s)-\bar\phi_\epsilon(y)}{(s-y)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar h(y,s)\,dG(s)\Big\}\end{aligned}$$

and let $C_\epsilon(y)$ be defined by

$$C_\epsilon(y) := 1 + \bar d_F(y)\,\epsilon^{-1}\Big\{\int_{y-\epsilon}^y \bar h(s,y)g(s)\,ds + \int_y^{y+\epsilon} \bar h(y,s)g(s)\,ds\Big\},\; y \in [0,1]. \qquad (1.39)$$

Then we have

$$\bar\phi'_\epsilon(M)\,C_\epsilon(M) \leq K_\epsilon(M), \qquad (1.40)$$

implying

$$M_\epsilon \leq \sup_{s\in[0,1]} K_\epsilon(s)/C_\epsilon(s). \tag{1.41}$$

In a similar way, if $m_\epsilon := \inf_{s\in[0,1]} \phi'_\epsilon(s)$, we get

$$m_\epsilon \geq \inf_{s\in[0,1]} K_\epsilon(s)/C_\epsilon(s). \tag{1.42}$$

Let the function A_δ be defined by

$$A_\delta(y) := \bar{d}_F(y)\left\{\int_{y-\delta}^{y} |\tfrac{\partial}{\partial y}\bar{h}(s,y)|\, dG(s) + \int_{y}^{y+\delta} |\tfrac{\partial}{\partial y}\bar{h}(y,s)|\, dG(s)\right\},\ y\in[0,1]$$

Fix $\delta > 0$ such that, for all $y \in [0,1]$,

$$A_\delta(y)/C_\epsilon(y) \leq \frac{1}{2}. \tag{1.43}$$

Note that $\delta > 0$ can be chosen independently of $\epsilon > 0$, since

$$\lim_{\epsilon\downarrow 0} C_\epsilon(y) = 1 + 2\bar{d}_F(y)\bar{h}(y,y)g(y),\ y\in(0,1).$$

Then we get from (1.43), for each $y \in [0,1]$, by applying the mean value theorem on the ratios $\{\bar{\phi}_\epsilon(v) - \bar{\phi}_\epsilon(u)\}/(v-u)$,

$$\begin{aligned}
&\frac{\bar{d}_F(y)}{C_\epsilon(y)}\left\{\int_{y-\delta}^{y}\left|\tfrac{\bar{\phi}_\epsilon(y)-\bar{\phi}_\epsilon(s)}{(y-s)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar{h}(s,y)\right| dG(s) + \int_{y}^{y+\delta}\left|\tfrac{\bar{\phi}_\epsilon(s)-\bar{\phi}_\epsilon(y)}{(s-y)\vee\epsilon}\,\tfrac{\partial}{\partial y}\bar{h}(y,s)\right| dG(s)\right\}\\
\leq\ & A_\delta(y)\max\{M_\epsilon,|m_\epsilon|\}/C_\epsilon(y) \leq \tfrac{1}{2}\max\{M_\epsilon,|m_\epsilon|\}.
\end{aligned}$$

Defining $B_\delta(y)$ by

$$\begin{aligned}
B_\delta(y) \ :=\ & \bar{d}_F(y)|\bar{k}'(y)| + |\bar{d}'_F(y)|\left[(1-y)\bar{h}_1(y) + y\,\bar{h}_2(y)\right]\sup_{s\in[0,1]}\{\bar{c}_F(s)|\bar{k}(s)|\}\\
&+\frac{2\bar{d}_F(y)}{\delta}\sup_{s\in[0,1]}\{\bar{d}_F(s)|\bar{k}(s)|\}\left\{\sup_{s\in[0,y]}|\tfrac{\partial}{\partial y}\bar{h}(s,y)| + \sup_{s\in[y,1]}|\tfrac{\partial}{\partial y}\bar{h}(y,s)|\right\}
\end{aligned}$$

we get, for $y \in [0,1]$,

$$\begin{aligned}
&\bar{d}_F(y)|\bar{k}'(y)| + |\bar{d}'_F(y)\,\bar{\xi}_\epsilon(y)|\left[(1-y)\bar{h}_1(y) + y\,\bar{h}_2(y)\right]\\
&+\bar{d}_F(y)\left\{\int_0^{y-\delta}\tfrac{|\bar{\phi}_\epsilon(y)|+|\bar{\phi}_\epsilon(s)|}{y-s}\,|\tfrac{\partial}{\partial y}\bar{h}(s,y)|\, dG(s) + \int_{y+\delta}^{1}\tfrac{|\bar{\phi}_\epsilon(y)|+|\bar{\phi}_\epsilon(s)|}{y-s}\,|\tfrac{\partial}{\partial y}\bar{h}(y,s)|\, dG(s)\right\}\\
\leq\ & \bar{d}_F(y)|\bar{k}'(y)| + |\bar{d}'_F(y)\,\bar{\xi}_\epsilon(y)|\left[(1-y)\bar{h}_1(y) + y\,\bar{h}_2(y)\right]\\
&+\frac{2\bar{d}_F(y)}{\delta}\sup_{s\in[0,1]}|\bar{\phi}_\epsilon(s)|\left\{\int_0^{y-\delta}|\tfrac{\partial}{\partial y}\bar{h}(s,y)|\, dG(s) + \int_{y+\delta}^{1}|\tfrac{\partial}{\partial y}\bar{h}(y,s)|\, dG(s)\right\}\\
\leq\ & B_\delta(y) \leq c,
\end{aligned} \tag{1.44}$$

for some constant c, independent of ϵ and y. Hence, for each $y \in [0,1]$,

$$|\bar\phi_\epsilon'(y)| \le A_\delta(y)/C_\epsilon(y) + B_\delta(y)/C_\epsilon(y) \le \tfrac{1}{2}\max\{M_\epsilon, |m_\epsilon|\} + B_\delta(y)/C_\epsilon(y),$$

implying

$$\tfrac{1}{2}\max\{M_\epsilon, |m_\epsilon|\} \le \sup_{s\in[0,1]} B_\delta(s)/C_\epsilon(s) \le \sup_{s\in[0,1]} c/C_\epsilon(s) \le c', \tag{1.45}$$

for some constant c' independent of ϵ.
Hence $\bar\phi_\epsilon'(y)$ is bounded on $[0,1]$, uniformly in ϵ and y, implying that $\bar\phi_\epsilon$ is Lipschitz, uniformly in $\epsilon > 0$.

□

We now have the following theorem.

Theorem 1.4.8 *Let $G(y) = F^{-1}(y)$, $y \in [0,1]$, with a derivative g which exists except for at most a finite number of points, where G has left and right derivatives. Furthermore, let $\bar k(y) = k(G(y))$, $\bar H(u,v) = H(G(u),G(v))$, $\bar h(u,v) = h(G(u),G(v))$, and let $\bar d_F$ be defined by*

$$\bar d_F(y) := \frac{y(1-y)}{(1-y)\bar h_1(y) + y\,\bar h_2(y)}, \tag{1.46}$$

where $\bar h_i = h_i \circ G$, $i = 1,2$. Then

(i) The integral equation

$$\bar\phi(y) = \bar d_F(y)\Big\{\bar k(y) - \int_0^y \tfrac{\bar\phi(y)-\bar\phi(s)}{y-s}\,d\bar H(s,y) + \int_y^1 \tfrac{\bar\phi(s)-\bar\phi(y)}{s-y}\,d\bar H(y,s)\Big\},\ y \in [0,1], \tag{1.47}$$

has a unique solution which is Lipschitz on $[0,1]$.

(ii) The Lipschitz norm in (i) has the following upper bound. Let $C(y)$ be defined by

$$C(y) := 1 + 2\bar d_F(y)g(y)\bar h(y,y). \tag{1.48}$$

Moreover, let $A_\delta(y)$ and $B_\delta(y)$ be defined by

$$A_\delta(y) := \bar d_F(y)\Big\{\int_{y-\delta}^y |\tfrac{\partial}{\partial y}\bar h(s,y)|\,dG(s) + \int_y^{y+\delta} |\tfrac{\partial}{\partial y}\bar h(y,s)|\,dG(s)\Big\}, \tag{1.49}$$

and

$$\begin{aligned} B_\delta(y) \;:=\; & \bar d_F(y)|\bar k'(y)| \\ & + |\bar d_F'(y)|\,\Big[(1-y)\bar h_1(y) + y\,\bar h_2(y)\Big] \sup_{s\in[0,1]}\{\bar c_F(s)|\bar k(s)|\} \\ & + \frac{2\bar d_F(y)}{\delta}\sup_{s\in[0,1]}\{\bar d_F(s)|\bar k(s)|\} \times \\ & \times \Big\{\sup_{s\in[0,y]}|\tfrac{\partial}{\partial y}\bar h(s,y)| + \sup_{s\in[y,1]}|\tfrac{\partial}{\partial y}\bar h(y,s)|\Big\} \end{aligned} \tag{1.50}$$

At the points in

$$\begin{aligned} D' &= \{\text{discontinuity points of } g(y)\text{, augmented with 0 and 1}\} \\ &\cup \{\text{discontinuity points of } \bar{k}'(y),\ \bar{d}'_F(y), \\ &\qquad \Delta^1(y) = \tfrac{\partial}{\partial y}\bar{h}(y,s) \text{ for y} \le \text{s, and } \Delta^2(y) = \tfrac{\partial}{\partial y}\bar{h}(s,y) \text{ for y} \ge \text{s}\}, \end{aligned}$$

A_δ and B_δ have two versions, one corresponding to taking left derivatives and one corresponding to taking right derivatives.

Then there exists a $\delta > 0$ such that

$$\sup_{s\in[0,1]} A_\delta(s)/C(s) \le 1/2$$

and we have

$$|\bar{\phi}(v) - \bar{\phi}(u)| \le c(v-u),\ 0 \le u < v \le 1, \tag{1.51}$$

where c is given by

$$c = 2\sup_{s\in[0,1]} B_\delta(s)/C(s). \tag{1.52}$$

(iii) The integral equation (1.21) has a unique solution ϕ.

Proof:
ad (i) By the preceding two lemma's, the set $\{\bar{\phi}_\epsilon : \epsilon \le \epsilon_0\}$ (for some $\epsilon_0 > 0$) is bounded and equicontinuous. Hence, by the Arzelà-Ascoli theorem, each sequence $\bar{\phi}_{\epsilon_n}$, $\epsilon_n \downarrow 0$, has a subsequence $(\bar{\phi}_{\epsilon_m})$, converging in the supremum metric to a continuous function $\bar{\phi}$ on $[0,1]$. By Lebesgue's dominated convergence theorem we get, for such a subsequence $(\bar{\phi}_{\epsilon_m})$,

$$\begin{aligned} \bar{\phi}(y) &= \lim_{m\to\infty} \bar{\phi}_{\epsilon_m}(y) \\ &= \bar{d}_F(y)\Big\{\bar{k}(y) - \int_0^y \tfrac{\bar{\phi}(y)-\bar{\phi}(s)}{y-s}\,\bar{h}(s,y)\,dG(s) + \int_y^1 \tfrac{\bar{\phi}(s)-\bar{\phi}(y)}{s-y}\,\bar{h}(y,s)\,dG(s)\Big\}. \end{aligned} \tag{1.53}$$

Uniqueness of the solution follows by applying the same kind of supremum argument as in lemma (1.4.1) on the difference of two solutions of equation (1.53).
ad (ii) It was shown in (1.45) in the proof of lemma 1.4.4 that

$$\sup_{s\in[0,1]} |\phi'_\epsilon(s)| \le 2\sup_{s\in[0,1]} B_\delta(s)/C_\epsilon(s),$$

where C_ϵ is defined by (1.39). But since

$$\lim_{\epsilon\downarrow 0} C_\epsilon(y) = 1 + 2\bar{d}_F(y)\bar{h}(y,y)g(y),$$

for $y \in [0,1]$, (1.51) now follows.
ad (iii) We define ϕ by $\phi(x) = \bar{\phi}(F(x))$. If $y = F(x)$, we get, by a change of variables,

$$\begin{aligned} \phi(x) &= \bar{\phi}(y) \\ &= \bar{d}_F(y)\big\{\bar{k}(y) - \int_0^y \tfrac{\bar{\phi}(y)-\bar{\phi}(s)}{y-s}\,d\bar{H}(s,y) + \int_y^1 \tfrac{\bar{\phi}(s)-\bar{\phi}(y)}{s-y}\,d\bar{H}(y,s)\big\} \\ &= d_F(x)\big\{k(x) - \int_0^x \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\,dH(t,x) + \int_x^M \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\,dH(x,t)\big\}, \end{aligned}$$

and hence ϕ satisfies the original integral equation. Uniqueness of ϕ follows from uniqueness of $\bar{\phi}$ (since a solution ϕ conversely defines a solution $\bar{\phi}$ on the inverse scale).

□

Remark. The same arguments can be applied to prove existence of a solution to the ξ-equation. Hence ϕ can be written as

$$\phi = F(1-F)\,\xi.$$

Solvability of $\tilde{\kappa}_F = L_1^* L_1 a$ can now immediately be seen.

Corollary 1.4.1 *Let the conditions (M1) to (M3), (H1) to (H3), (F1) and (K1) and (K2) be satisfied. Then the equation $\tilde{\kappa}_F = L_1^* L_1 a$ is solvable.*

Proof: By the Lipschitz property of $\bar{\phi}$ we have, for any $0 \le u < v \le M$,

$$\frac{|\phi(v)-\phi(u)|}{F(v)-F(u)} = \frac{|\bar{\phi}(F(v))-\bar{\phi}(F(u))|}{F(v)-F(u)} \le K,$$

for some constant K. Thus the Radon-Nikodym derivative $d\phi/dF$ is a.e.-$[F]$ bounded by K.

□

Remark: Again the conditions (F1), (H3) and (K2) can be weakened, this time to:

$$\frac{dk}{dF},\ \frac{\partial}{\partial F(x)}h(t,x) \text{ and } \frac{\partial}{\partial F(x)}h(x,t)$$

exist, possibly at some points only as separate left and right derivatives, and are bounded.

1.4.3 Case k

Consider the case with exactly k observation times per unobservable event time. The observation time distribution becomes a higher dimensional distribution, so the $L_2(Q_{F_0,H})$-space changes. This also has consequences for the score operator and the score equation $\tilde{\kappa}_{F_0} = L_1^* L_1 a$. However, taking derivatives in this equation, we turn up with an equation which is similar in structure to the ϕ-equation (1.21).

Let the ordered observation times $(T_1, T_2, \ldots, T_k)$ replace (U, V). Let the simultaneous density function of these observation times be denoted by $h(t_1, t_2, \ldots, t_k)$. Moreover, let the simultaneous density of (T_i, T_{i+1}) be denoted by $h_{i,i+1}$, and let h_1 and h_k denote the density of the first and last observation times respectively. Then we get as integral equation

$$\phi(x) + d(x)\left[\int_{t=\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\,\tilde{h}(t,x)\,dt - \int_{t=x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\,\tilde{h}(x,t)\,dt\right] = k(x)d(x), \tag{1.54}$$

with

$$\tilde{h}(t,x) = \sum_{i=1}^{k-1} h_{i,i+1}(t,x) \qquad \tilde{h}(x,t) = \sum_{i=1}^{k-1} h_{i,i+1}(x,t)$$

and d given by

$$d(x) = \frac{F(x)[1-F(x)]}{h_1(x)[1-F(x)] + h_k(x)F(x)}.$$

The situation with a varying number of observation times can be treated in the same way.

1.5 Some special choices of F, h and the functional

In this section, some choices of F_0, h and $\tilde{\kappa}$ will be treated, for which we have been able to find a more or less explicit solution to the integral equation in case 2. In all cases we take $\tilde{\kappa}'_{F_0} \equiv 1$ (estimation of the mean). An explicit solution to the integral equation exists in the following situations:

I) Let $F(x) = x$ on $[0,1]$ and $h(u,v) = 4$ on the square $\{0 \le u \le 1/2,\ 1/2 \le v \le 1\}$. Then we have the solution

$$\phi_0(x) = \tfrac{2}{3}x(1-x)$$

II) Let

$$h(u,v) = C_F\,(F(v) - F(u))$$

with $1/C_F = \{(\tau_M - \tau_0)\int x dF(x) - \int x^2 dF(x)\}$, and with marginal densities

$$\begin{aligned} h_1(u) &= C_F \int_u^{\tau_M} f(s)[(\tau_M - \tau_0) - s]\,ds \\ h_2(v) &= C_F \int_{\tau_0}^{v} s f(s)\,ds. \end{aligned}$$

Now the singularity is wiped away by h, and the integral part reduces to

$$\int_{t=\tau_0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\, h(t,x)\,dt - \int_{t=x}^{\tau_M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\, h(x,t)\,dt =$$

$$C_F\left\{(\tau_M - \tau_0)\phi(x) - \int_{\tau_0}^{\tau_M} \phi(s)\,ds\right\}$$

We arrive at the solution

$$\phi(x) = \frac{c(x)}{1 - C_F\int_{\tau_0}^{\tau_M} c(x)dx}$$

with

$$c(x) = \frac{F(x)(1-F(x))}{C_F(\tau_M - \tau_0)F(x)[1-F(x)] + F(x)h_2(x) + [1-F(x)]\,h_1(x)}$$

Hence, the lower bound is given by

$$\int_{\tau_0}^{\tau_M} \phi(x)\,dx = \frac{\int_{\tau_0}^{\tau_M} c(x)dx}{1 - C_F\int_{\tau_0}^{\tau_M} c(x)dx}.$$

In the next situation the solution is not given by an explicit formula.

1.5.1 Uniform distributions on $[0,1]$

In this subsection, the solution will be given if X is uniform on $[\tau_0, \tau_M]$ and we have two independent observation times T_1 and T_2, also uniformly distributed on $[\tau_0, \tau_M]$. Letting $U = \min\{T_1, T_2\}$ and $V = \max\{T_1, T_2\}$, (U,V) is uniformly distributed on the triangle

$\tau_0 \leq u < v \leq \tau_M$. This situation is the case 2 analogue of $g \equiv 1/(\tau_M - \tau_0)$ in case 1. Having more information we may expect a smaller lower bound

$$\int_{\tau_0}^{\tau_M} \phi(x)\, dx$$

in case 2.

No explicit solution is available. We give the solution with respect to a basis of Legendre polynomials. Legendre polynomials are most commonly considered for the interval $[-1,1]$. However, we first solve the problem for the interval $[0,1]$. So we have $F(x) = x$ and $h \equiv 2$.

The integral equation can be written in the form:

$$\frac{2 - 4x(1-x)}{x(1-x)}\, \phi(x) + [A\phi](x) = 1, \tag{1.55}$$

with the operator A defined by:

$$[A\phi](x) := 2\left\{ \int_0^x \tfrac{\phi(x)-\phi(t)}{x-t}\, dt - \int_x^1 \tfrac{\phi(t)-\phi(x)}{t-x}\, dt \right\}. \tag{1.56}$$

First the structure of A is investigated.

The operator A

In this subsection it will be proved that the Legendre polynomials on [0,1] are the eigenfunctions of A, and the corresponding eigenvalues will be given.

The Legendre polynomials $\{P_n\}$ on $[0,1]$ are defined as a complete orthogonal basis of the space $L_2([0,1])$ with respect to the inner product induced by the standard norm and having the extra restriction degree$(P_n) = n$. This last condition makes them uniquely determined up to multiplying constants. We choose the constants such that $P_n(1) = 1$. Then $\{P_n\}$ can be obtained via

$$P_n(x) = \frac{d^n}{dx^n} \frac{[x(x-1)]^n}{n!} \tag{1.57}$$

or via the recurrent relation

$$(n+1)P_{n+1}(x) = (2n+1)(2x-1)P_n(x) - nP_{n-1}(x), \tag{1.58}$$

with $P_0(x) = 1$ and $P_1(x) = 2x - 1$ as starting values. The properties that are needed in the sequel are given in the next proposition. Their proofs are based on the above procedure (1.57).

Proposition 1.5.1 *The polynomials as given by (1.57) or (1.58) satisfy:*

i: degree$(P_n)=n$.

ii: $P_k \perp P_l$ *if* $k \neq l$.

iii: $\langle P_n, P_n \rangle = \frac{1}{2n+1}$.

iv: $P_n(0) = (-1)^n$, $P_n(1) = 1$.

v: $\int_x^1 P_n(t)dt = -\int_0^x P_n(t)\,dt = \frac{1}{4n+2}(P_{n-1}(x) - P_{n+1}(x))$.

vi: The coefficient of the leading term x^n of P_n is $\frac{(2n)!}{n!n!}$.

Proof.

i: Trivial.

ii: Suppose $k > l$. Then

$$\begin{aligned}\int_0^1 P_k(x)P_l(x)dx &= \int_0^1 \tfrac{P_l(x)}{k!} d([x(x-1)]^k)^{(k-1)} \\ &= -\int_0^1 \tfrac{([x(x-1)]^k)^{(k-1)}}{k!} P_l^{(1)}(x)dx \\ &= \ldots = (-1)^k \int_0^1 \tfrac{[x(x-1)]^k}{k!} P_l^{(k)}(x)dx = 0,\end{aligned}$$

since $P_l^{(k)} = 0$.

iii: Since $[x(x-1)]^n = x^{2n}$ + lower order terms, the coefficient of x^n in the n-th derivative of $[x(x-1)]^n$ is $\frac{(2n)!}{n!}$, hence $P_n^{(n)}(x) = \binom{2n}{n} n!$. Thus we have

$$\begin{aligned}\int_0^1 P_n(x)P_n(x)dx &= (-1)^n \int_0^1 \tfrac{[x(x-1)]^n}{n!} P_n^{(n)}(x)dx \\ &= \binom{2n}{n} \int_0^1 x^n(1-x)^n dx = \tfrac{1}{2n+1}.\end{aligned}$$

iv: Differentiating $\frac{[x(x-1)]^n}{n!}$ n times yields

$$P_n(x) = (2x-1)^n + \text{terms with at least one factor } x(1-x).$$

v: Since

$$\int_0^x P_n(t)dt = \tfrac{d^{n-1}}{dx^{n-1}} \tfrac{[x(x-1)]^n}{n!}$$

and

$$P_{n+1}(x) = \tfrac{d^{n-1}}{dx^{n-1}} \left\{ \tfrac{[x(x-1)]^{n-1}}{n-1!} (2x-1)^2 \right\} + 2 \tfrac{d^{n-1}}{dx^{n-1}} \tfrac{[x(x-1)]^n}{n!},$$

we have

$$\begin{aligned}P_{n+1}(x) - P_{n-1}(x) &= \tfrac{d^{n-1}}{dx^{n-1}} \left\{ \tfrac{[x(x-1)]^{n-1}}{n-1!} [(2x-1)^2 - 1] \right\} + 2 \int_0^x P_n(t)dt \\ &= 4 \tfrac{d^{n-1}}{dx^{n-1}} \tfrac{[x(x-1)]^n}{n-1!} + 2 \int_0^x P_n(t)dt \\ &= (4n+2) \int_0^x P_n(t)dt\end{aligned}$$

vi: Trivial.

As an illustration, we give the first seven Legendre polynomials on $[0,1]$ with $P_n(1)=1$:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= 2x-1 \\ P_2(x) &= 6x^2-6x+1 \\ P_3(x) &= 20x^3-30x^2+12x-1 \\ P_4(x) &= 70x^4-140x^3+90x^2-20x+1 \\ P_5(x) &= 252x^5-630x^4+560x^3-210x^2+30x-1 \\ P_6(x) &= 924x^6-2772x^5+3150x^4-1680x^3+420x^2-42x+1 \end{aligned}$$

Let $I\!P^n$ be the $(n+1)$-dimensional space of real-valued polynomials on $[0,1]$ of maximal degree n. The operator A maps polynomials of degree n to polynomials of the same degree. For, using $x^n - t^n = (x-t)\sum_{j=0}^{n-1} x^{n-j-1}t^j$, we obtain

$$A[x^n](x) = 4S_n x^n + q_{n-1}(x),$$

with $S_n = \sum_{j=1}^n \frac{1}{j}$ and $q_{n-1} \in I\!P^{n-1}$. Moreover, constant functions are mapped to the zero function.
Now we have:

Proposition 1.5.2 *The Legendre polynomials $\{P_0,\dots,P_n\}$ are an orthogonal basis of eigenfunctions for the operator $A: I\!P^n \to I\!P^n$, with corresponding eigenvalues $\lambda_k = 4S_k$.*

Proof: One easily shows that the operator A, defined on $I\!P^n$, is symmetric with respect to the standard L_2 inner-product. So there exists an orthogonal basis of eigenfunctions $\{p_0,\dots,p_n\}$ on $I\!P^n$. This holds for any $n \in I\!N$. Together with the fact that A preserves the degree of polynomials, it follows by induction that degree$(p_n) = n$. So this orthogonal basis of eigenfunctions consists of the Legendre polynomials up to degree n.

Expanding x^n along the Legendre basis, $x^n = \sum_{k=0}^n \gamma_k P_k(x)$, we get

$$\begin{aligned} A\gamma_n P_n &= Ax^n - A\sum_{k=0}^{n-1}\gamma_k P_k \\ &= 4S_n x^n + q_{n-1} - \sum_{k=0}^{n-1}\gamma_k A P_k \\ &= 4S_n\gamma_n P_n + \tilde{q}_{n-1}, \end{aligned}$$

with q_{n-1} and $\tilde{q}_{n-1} \in I\!P^{n-1}$. Since $\tilde{q}_{n-1}$ is of lower degree than P_n and since P_n is an eigenfunction, we have $\tilde{q}_{n-1} = 0$, and thus $AP_n = 4S_nP_n$.

□

Remark: Since the eigenvalues converge to ∞, A is an unbounded operator.

Solution with respect to the Legendre basis

We turn back to the integral equation (1.55). Since $L_2([0,1])$ can be split up in the orthogonal subspaces $L_2^0([0,1])$ and span$\{P_0\}$, any $a \in L_2^0([0,1])$ can be written as $a = \sum_{k=1}^{\infty} \alpha_k P_k$. Let

$$a_0 = \sum_{k=1}^{\infty} \beta_k P_k$$

denote the solution to $\tilde{\kappa}_F = L_1^* L_1 a$ with respect to the Legendre basis. By proposition 1.5.1.v, we have

$$\phi_0 = \sum_{k=1}^{\infty} \tfrac{\beta_k}{4k+2}(P_{k-1} - P_{k+1}).$$

Using orthogonality of the Legendre polynomials, the lower bound is equal to

$$\langle \phi_0, 1\rangle = \langle \sum_{k=1}^{\infty} \tfrac{\beta_k}{4k+2}(P_{k-1} - P_{k+1}), P_0 \rangle = \frac{\beta_1}{6}.$$

The coefficients β_k can be found by taking the inner product of the Legendre polynomials $P_1, P_2, \ldots$ with both sides of $L_1^* L_1 a = \tilde{\kappa}_F$. This yields an infinite set of linear equations in β_k, which will turn out to be easily solvable.

For the right-hand side we get, since $\tilde{\kappa}(x) = \frac{1}{2} P_1(x)$,

$$\langle P_j, \tilde{\kappa}\rangle = \begin{cases} 1/6 & \text{for } j = 1 \\ 0 & \text{for } j = 2, \ldots . \end{cases}$$

For the left-hand side we make use of the structure of the operator A, appearing in the derivative of $L_1^* L_1 a$. For our choice of F, h and $\tilde{\kappa}$, we have

$$\frac{d}{dx}[L_1^* L_1 a](x) = [(A-4)]\phi(x) + \frac{2}{x(1-x)}\,\phi(x).$$

with $\phi(x) = \int_x^1 a(x)\,dx$. Define

$$\xi_{k-1} := \frac{\beta_k}{4k+2}.$$

Then we get, using continuity of $L_1^* L_1$, property 1.5.1.v and proposition (1.5.2),

$$\begin{aligned}
\langle P_j, L_1^* L_1 a_0\rangle &= \sum_{k=1}^{\infty} \beta_k \langle P_j, L_1^* L_1 P_k\rangle \\
&= \sum_{k=1}^{\infty} \xi_{k-1} \int_{x=0}^{1} P_j(x) \times \\
&\qquad \left\{ \int_{t=0}^{x} \Big([(A-4)(P_{k-1} - P_{k+1})](t) + \tfrac{2}{t(1-t)}[(P_{k-1} - P_{k+1})](t) \Big)\, dt \right\} dx \\
&\quad + \sum_{k=1}^{\infty} \xi_{k-1} \langle P_j, C_k\rangle
\end{aligned}$$

$$= \sum_{k=1}^{\infty} \xi_{k-1} \left\{ \tfrac{4(S_{k-1}-1)}{4k-2} \langle P_j, P_k - P_{k-2} \rangle - \tfrac{4(S_{k+1}-1)}{4k+6} \langle P_j, P_{k+2} - P_k \rangle \right\}$$
$$+ \sum_{k=1}^{\infty} \xi_{k-1} \int_{t=0}^{1} 2 \, \tfrac{1}{t(1-t)} \, [P_{k-1} - P_{k+1}](t) \, \frac{[P_{j-1} - P_{j+1}](t)}{4j+2} \; + \; 0$$

We have to pay some extra attention to the part with the factor $1/[t(1-t)]$. By property 1.5.1.iv we have $P_k(1) = (-1)^k P_k(0) = 1$, so $P_{k-1} - P_{k+1}$ is divisible by the factor $t(1-t)$. Hence we can write

$$P_{k-1}(t) - P_{k+1}(t) = (4k+2) \, t(1-t) \, Q_{k-1}(t), \tag{1.59}$$

with degree$(Q_{k-1}) = k - 1$.

The relevant properties of Q_k are given in the following proposition:

Proposition 1.5.3 *In the space $(L_2([0,1]), [\cdot\,,\cdot])$, with inner product*

$$[f, g] = \int_{x=0}^{1} x(1-x) \, f(x) \, g(x) \, dx,$$

the polynomials $\{Q_j\}$ are a complete orthogonal basis for which the following relations hold:

$$\langle f, P_j - P_{j+2} \rangle = (4j+6) \, [f, Q_j]. \tag{1.60}$$

and

$$[Q_j, Q_j] = \frac{1}{(j+1)(j+2)(2j+3)}.$$

Proof:
Relation (1.60) is a direct consequence of (1.59).
Since $P_{j-1} - P_{j+1}$ is orthogonal to $I\!P^{j-2}$ with respect to the standard inner product $\langle \cdot\,, \cdot \rangle$, Q_{j-1} is orthogonal to $I\!P^{j-2}$ with respect to $[\cdot\,, \cdot]$. By theorem 3.1.5. in Szegö (1978) the Q-polynomials form a complete basis.

The inner product expression can be found by using the relations between P_j and Q_j. Let $Q_j = \sum_{k=0}^{j} \gamma_k P_k$. Then we have

$$[Q_j, Q_j] = \frac{\langle P_j - P_{j+2}, Q_j \rangle}{4j+6} = \frac{\gamma_j}{(2j+1)(4j+6)}.$$

In order to determine the factor γ_j, let c_j and d_j denote the coefficients of the leading term of P_j and Q_j respectively. Then, on one hand

$$P_j(x) - P_{j+2}(x) = (4j+6)(x - x^2)(d_j x^j + \text{ polynomial of degree } \le j-1),$$

whereas on the other hand

$$P_j(x) - P_{j+2}(x) = -c_{j+2} x^{j+2} + \text{ polynomial of degree } \le j+1.$$

Hence $c_{j+2} = (4j+6) d_j$ and $\gamma_j = d_j / c_j = [c_{j+2}/c_j] \times [1/(4j+6)]$. From property 1.5.1.vi we know that $c_j = [(2j)!]/[j!\, j!]$, yielding the result. □

The first five Q-polynomials have the following form:

$$\begin{aligned} Q_0(x) &= 1 \\ Q_1(x) &= 2x-1 \\ Q_2(x) &= 5x^2-5x+1 \\ Q_3(x) &= 14x^3-21x^2+9x-1 \\ Q_4(x) &= 42x^4-84x^3+56x^2-14x+1 \end{aligned}$$

Using this proposition, we get

$$\int_{t=0}^{1} 2\,\tfrac{1}{t(1-t)}\,[P_{k-1}-P_{k+1}](t)\,\tfrac{[P_{j-1}-P_{j+1}](t)}{4j+2}\,dt = 2(4k+2)[Q_{k-1},Q_{j-1}],$$

which is equal to $4/[j(j+1)]$ if $k=j$ and zero otherwise.
So, finally, multiplying both sides by $2\,(2j+1)$ and shifting the index by one, we get, for $j=0,1,\dots$

$$-4\tfrac{S_j-1}{2j+1}\,\xi_{j-2} + 4\big(\tfrac{S_j-1}{2j+1}+\tfrac{S_{j+2}-1}{2j+5}\big)\,\xi_j - 4\tfrac{S_{j+2}-1}{2j+5}\,\xi_{j+2} + \tfrac{8(2j+3)}{(j+1)(j+2)}\,\xi_j = 1_{\{j=0\}}$$

with $\xi_{-1}=\xi_{-2}=0$ by definition. This is essentially a tridiagonal system:

$$\begin{array}{llllllll}
\alpha_{0,0}\,\xi_0 & & +\,\alpha_{0,2}\,\xi_2 & & & & & =1 \\
& \alpha_{1,1}\,\xi_1 & & +\,\alpha_{1,3}\,\xi_3 & & & & =0 \\
\alpha_{2,0}\,\xi_0 & & +\,\alpha_{2,2}\,\xi_2 & & +\,\alpha_{2,4}\,\xi_4 & & & =0 \\
& \alpha_{3,1}\,\xi_1 & & +\,\alpha_{3,3}\,\xi_3 & & +\,\alpha_{3,5}\,\xi_5 & & =0 \\
& & \alpha_{4,2}\,\xi_2 & & +\,\alpha_{4,4}\,\xi_4 & & +\,\alpha_{6,4}\,\xi_6 & =0 \\
& & & \ddots & & \ddots & \ddots &
\end{array}$$

with coefficients $\alpha_{j,k}$ given by:

$$\begin{aligned} \alpha_{j,j} &= b_j+b_{j+2}+\sigma_j && \text{for } j\ge 0 \\ \alpha_{j,j+2} &= -b_{j+2} && \text{for } j\ge 0 \\ \alpha_{j,j-2} &= -b_j && \text{for } j\ge 2 \\ \alpha_{j,k} &= 0 && \text{for } |j-k|>2 \text{ or } |j-k|=1. \end{aligned} \tag{1.61}$$

Here b_j and σ_j have the following values:

$$\begin{aligned} b_j &= \tfrac{4(S_j-1)}{2j+1} \\ \sigma_j &= \tfrac{8(2j+3)}{(j+1)(j+2)}. \end{aligned}$$

By corollary 1.4.1 we already know the solution to this system of equations to be unique. We immediately see that the ξ-values with odd index are zero. So from now on we will only look at the ξ-values with even index. Note that, in principle, any value ξ_0 can be inserted, but once ξ_0 is determined, all the other values $\xi_2, \xi_4, \ldots$ are fixed. Uniqueness of ξ_0 follows from the extra restriction that $a_0 = \sum_{k=1}^{\infty} \beta_k P_k$ has to be an L_2-function. Thus

$$\langle a_0, a_0 \rangle = \sum_{k=1}^{\infty} \frac{(\beta_k)^2}{2k+1} = \sum_{k=1}^{\infty} 4(\xi_{k-1})^2(2k+1) < \infty. \tag{1.62}$$

It will be shown that this leaves exactly one solution for ξ_0, which can be written as a continued fraction expansion, arising naturally from the tridiagonal system of linear equations.

Starting with ξ_0, ξ_{2k} can be obtained as:

$$\xi_{2k} = A_k \xi_0 + B_k.$$

Equivalently we have

$$\xi_0 = \frac{\xi_{2k}}{A_k} - \frac{B_k}{A_k}.$$

It will be shown that A_k goes to infinity very quickly as $k \to \infty$, leaving

$$\xi_0 = \lim_{k\to\infty} \frac{B_k}{A_k}$$

as the only possible solution satisfying (1.62).

Proposition 1.5.4 *The following holds:*

i: $1/A_k = o(k^{-\alpha})$ *as* $k \to \infty$ *for any* $\alpha \in \mathbb{R}$.

ii: $\lim_{k\to\infty} \frac{B_k}{A_k}$ *exists.*

Proof:

i): Define

$$p_k := -\frac{\alpha_{2k,2k}}{\alpha_{2k,2(k+1)}} = \frac{b_{2k}}{b_{2(k+1)}} + 1 + \frac{\sigma_{2k}}{b_{2(k+1)}} \qquad k = 0, 1, 2, \ldots$$

$$q_k := -\frac{\alpha_{2k,2(k-1)}}{\alpha_{2k,2(k+1)}} = -\frac{b_{2k}}{b_{2(k+1)}} \qquad k = 1, 2, \ldots$$

$$q_0 := -\frac{1}{b_2} = -2.5$$

The following relation holds for $\{A_k\}$ and $\{B_k\}$:

$$\begin{aligned} A_{k+1} &= p_k A_k + q_k A_{k-1} \\ B_{k+1} &= p_k B_k + q_k B_{k-1}, \end{aligned}$$

with $A_{-1} = 0$, $A_0 = 1$, $B_{-1} = 1$ and $B_0 = 0$, which can be proved by induction. Subtracting A_k in the first relation yields:

$$\begin{aligned} A_{k+1} - A_k &= \frac{S_{2k} - 1}{S_{2(k+1)} - 1} \frac{4k+5}{4k+1} (A_k - A_{k-1}) \\ &+ \frac{2}{S_{2(k+1)} - 1} \frac{(4k+3)(4k+5)}{(2k+1)(2k+2)} A_k. \end{aligned} \tag{1.63}$$

Since the second term on the right-hand side is ≥ 0, we immediately see

$$\begin{aligned} A_{k+1} - A_k &\geq \frac{S_{2k} - 1}{S_{2(k+1)} - 1} \frac{4k+5}{4k+1} (A_k - A_{k-1}) \\ &\geq \dots \\ &\geq \frac{4k+5}{S_{2(k+1)} - 1} \frac{S_2 - 1}{5} (A_1 - A_0) = 2 \frac{4k+5}{S_{2(k+1)} - 1} \end{aligned}$$

Hence, using $S_{2k} - 1 \leq \log(2k)$,

$$A_k \geq C_1 \frac{2k}{\log(2k)} + A_{k-1} \quad \text{for all } k \geq 0 \text{ and some fixed } C_1 > 0,$$

which implies $A_k \geq C_1 \sum_{j=1}^{k} \frac{2j}{\log(2j)} \geq C_2 \frac{(2k)^2}{\log(2k)}$ for some $C_2 > 0$. Using the second term on the right-hand side of (1.63), we have

$$A_{k+1} - A_k \geq C_2 \left(\frac{2k}{\log(2k)} \right)^2 \quad \text{for all } k > 0,$$

implying

$$A_k \geq C_2 \sum_{j=1}^{k} \left(\tfrac{2j}{\log(2j)} \right)^2 \geq C_3 \frac{(2k)^3}{(\log(2k))^2}$$

and

$$A_{k+1} - A_k \geq C_3 \left(\frac{2k}{\log(2k)} \right)^3 \text{ for some } C_3 > 0.$$

Repeating the same argument again and again, we find, for any $l \in \mathbb{N}$ and for all $k \in \mathbb{N}$:

$$A_k \geq C_{l+1} \, 2k \left(\frac{2k}{\log(2k)} \right)^l \quad \text{with } C_l > 0.$$

Of course $1/A_k = o(k^{-\beta})$ for some $\beta \in \mathbb{R}$ implies $1/A_k = o(k^{-\alpha})$ for all $\alpha \leq \beta$.

ii): To prove the second part of the proposition, define

$$D_k := \begin{pmatrix} 0 & q_k \\ 1 & p_k \end{pmatrix}$$

Then

$$D_0 D_1 \dots D_k = \begin{pmatrix} B_k & B_{k+1} \\ A_k & A_{k+1} \end{pmatrix}$$

Hence, computing determinants left and right:

$$-\prod_{j=0}^{k} \frac{b_{2k}}{b_{2(k+1)}} = B_k A_{k+1} - A_k B_{k+1},$$

implying

$$\frac{B_k}{A_k} - \frac{B_{k+1}}{A_{k+1}} = \frac{4k+5}{S_{2(k+1)} - 1} \frac{1}{A_k A_{k+1}}. \tag{1.64}$$

Since $1/A_k = o(k^{-\alpha})$ as $k \to \infty$ for all $\alpha > 0$, the sequence $\left\{\frac{B_k}{A_k}\right\}$ is a Cauchy sequence, hence convergent. (Notice that the approximated solution $-\frac{B_k}{A_k}$ is monotonously increasing in k.)

□

Although we already know that at least one solution should exist, it can also be shown directly that the solution $\{\xi_k\}$ obtained in this way indeed satisfies (1.62). It is sufficient to show

$$\lim_{k\to\infty} \xi_{2k} k^{\alpha_0} = 0 \quad \text{for some } \alpha_0 > 1. \tag{1.65}$$

This holds for any $\alpha \in \mathbb{R}$. For, writing $\xi_{2k} = A_k \xi_0 + B_k$, with $\xi_0 = \lim_{k\to\infty} -\frac{B_k}{A_k}$, and using (1.64), we obtain

$$\begin{aligned} |\xi_{2k} k^\alpha| &= |\{A_k(\xi_0 + \frac{B_k}{A_k})\} k^\alpha| \\ &= \sum_{j=k}^{\infty} \frac{4j+5}{S_{2(j+1)}} \frac{A_k}{A_j} \frac{k^\alpha}{A_{j+1}} \end{aligned}$$

This sum converges to zero as $k \to \infty$, since $A_k \le A_j$ and $k^\alpha A_{j+1}^{-1}$ can be made smaller than Cj^{-3} for some $C > 0$.
What remains to be done is computing $\lim_{k\to\infty} -\frac{B_k}{A_k}$. Since the countable system of equations is tridiagonal, one of the possible solutions ξ_0 from the system of equations can be represented as a continued fraction expansion:

$$\cfrac{1}{\alpha_{0,0} - \cfrac{\alpha_{0,2}\alpha_{2,0}}{\alpha_{2,2} - \cfrac{\alpha_{2,4}\alpha_{4,2}}{\alpha_{4,4} - \ddots}}} \tag{1.66}$$

This is easily seen by rewriting:

$$\alpha_{0,0}\,\xi_0 + \alpha_{0,2}\,\xi_2 = 1 \iff \xi_0 = \frac{1}{\alpha_{0,0} + \alpha_{0,2}\frac{\xi_2}{\xi_0}}$$

and

$$\begin{aligned} \alpha_{2k,2(k-1)}\,\xi_{2(k-1)} + \alpha_{2k,2k}\,\xi_{2k} + \alpha_{2k,2(k+1)}\,\xi_{2(k+1)} &= 0 \iff \\ \frac{-\alpha_{2k,2(k-1)}}{\alpha_{2k,2k} + \alpha_{2k,2(k+1)}\frac{\xi_{2(k+1)}}{\xi_{2k}}} &= \frac{\xi_{2k}}{\xi_{2(k-1)}} \end{aligned}$$

It turns out to be the same solution as the one obtained by taking $\lim_{k\to\infty} -\frac{B_k}{A_k}$.

Proposition 1.5.5 $\lim_{k\to\infty} -\frac{B_k}{A_k}$ *is the same as the continued fraction expansion (1.66).*

Proof: Define the k^{th} approximand of the continued fraction expansion as:

$$P_k(w) := \cfrac{1}{\alpha_{0,0} - \cfrac{\alpha_{0,2}\alpha_{2,0}}{\alpha_{2,2} - \ddots \cfrac{\ddots}{\cfrac{\alpha_{2(k-1),2k}\,\alpha_{2k,2(k-1)}}{\alpha_{2k,2k} + \alpha_{2k,2(k+1)}\,w}}}}$$

So $P_k(\frac{\xi_{2(k+1)}}{\xi_{2k}})$ yields the solution ξ_0, based on the first k equations, as a function of $\frac{\xi_{2(k+1)}}{\xi_{2k}}$. Furthermore define

$$\begin{aligned} T_0(w) &:= \frac{1}{\alpha_{0,0} + \alpha_{0,2}\,w} \\ T_k(w) &:= \frac{-\alpha_{2k,2(k-1)}}{\alpha_{2k,2k}\;\alpha_{2k,2(k+1)}\,w} \end{aligned}$$

Then

$$P_k(w) = T_0 \ldots T_k(w).$$

By induction, it is easily shown that

$$T_0 \ldots T_k(w) = -\frac{B_{k+1} - B_k\,w}{A_{k+1} - A_k\,w}.$$

Hence

$$\lim_{k\to\infty} -\frac{B_{k+1}}{A_{k+1}} = \lim_{k\to\infty} T_0 \ldots T_k(0) = \lim_{k\to\infty} P_k(0).$$

□

Computing the continued fraction expansion gives

$$\xi_0 = 0.1194623\ldots$$

The continued fraction expansion converges very quickly. For the 0^{th} approximand, we already have

$$P_0(0) = \frac{1}{\alpha_{0,0}} = \frac{5}{42} = 0.1190476\ldots$$

and each next approximand increases accuracy by about two digits.

For case 1, with a uniform observation time distribution on $[0,1]$, we have $\phi_0(x) = x(1-x)$. On the other hand, in the model without censoring, the score operator L reduces to the identity operator, yielding $a_0(x) = x - 1/2$, hence $\phi_0(x) = 1/2\,x(1-x)$. Since all even coefficients ξ_{2k} in the expansion are nonzero, ϕ_0 is definitely not a polynomial in case 2. However, since (1.65) holds for any $\alpha > 0$, convergence of the coefficients to zero goes very fast, faster than $k^{-\alpha}$ for any $\alpha > 0$. In figure 1.1, the first order and third order expansion of

$$\phi_0(x) = \sum_{k=0}^{\infty} \xi_{2k}\,(8k+6)\,x(1-x)\,Q_{2k}(x)$$

are given by the middle dashed curve and the solid curve respectively.

$$\begin{aligned} \phi_0^{(1)}(x) &= \xi_0\,6\,x(1-x)\,Q_0(x) \;=\; 0.71677\,x(1-x) \\ \phi_0^{(3)}(x) &= x(1-x)(7.1677\,10^{-1}\,Q_0(x) + 1.2192\,10^{-1}\,Q_2(x) + 2.4321\,10^{-2}\,Q_4(x)) \end{aligned}$$

A further expansion does not give any visible change in this plot. The upper dashed curve is the corresponding solution for case 1, $\phi_0(x) = x(1-x)$, whereas the lower dashed curve is the function $1/2\,x(1-x)$. Note that in case 2, being a situation between case 1 and the uncensored model, ϕ_0 is enclosed by $x(1-x)$ and $1/2\,x(1-x)$.

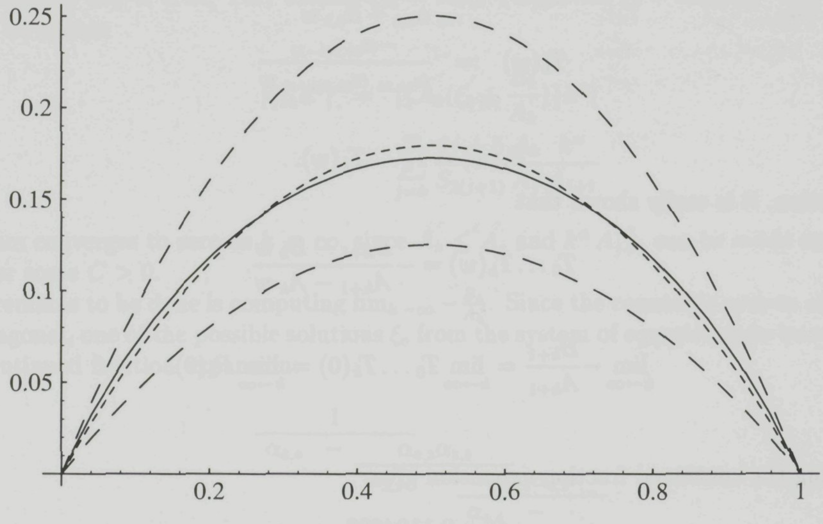


Figure 1.1: The functions ϕ, with first order approximation for case 2.

Remark: For the special case II from the beginning of this subsection, taking $F(x) = x$, we can derive the information bound in a similar way, since the integral operator A is very simple. The derivative of the operator equation becomes

$$\left\{\frac{3}{x(1-x)} - 3\right\}\phi(x) - 6\int_0^1 \phi(s)ds = 1.$$

Again, we get a tridiagonal system in ξ, with

$$\tfrac{42}{5}\xi_0 + \tfrac{3}{5}\xi_2 = 1,$$

$$\tfrac{32}{7}\xi_1 + \tfrac{3}{7}\xi_3 = 0,$$

$$\tfrac{-3}{2j+1}\xi_{j-2} + \left\{\tfrac{3}{2j+1} + \tfrac{3}{2j+5} - \tfrac{12(2j+3)}{(j+1)(j+2)}\right\}\xi_j - \tfrac{3}{2j+5}\xi_{j+2} = 0.$$

Choose ξ_{2k+1} to be zero and write $\xi_{2k} = A_k\xi_0 + B_k$ and $\xi_0 = \frac{\xi_{2k}}{A_k} - \frac{B_k}{A_k}$. Proceeding in the same way, we have $1/|A_k| = o(k^{-\alpha})$ as $k \to \infty$ and we get

$$\lim_{k\to\infty} -\frac{B_k}{A_k} = 0.1198987\ldots$$

which is indeed the same number as

$$\frac{\int_0^1 c(x)dx}{1 - 6\int_0^1 c(x)dx} = \frac{\int_0^1 \frac{x(1-x)}{3(1-x(1-x))}dx}{1 - 6\int_0^1 \frac{x(1-x)}{3(1-x(1-x))}dx}.$$

Uniform distributions on $[\tau_0, \tau_M]$

For uniform distributions on the interval $[\tau_0, \tau_M]$, the solution of the integral equation for the $[0,1]$-case can be used. The integral equation in the transformed scale is similar to equation (1.55), with the right-hand side replaced by $(\tau_M - \tau_0)$. This implies that all coefficients ξ_k are multiplied by a factor $(\tau_M - \tau_0)$, and the transformed-scale solution can be written as

$$\overline{\phi}_{\tau_0,\tau_M} = (\tau_M - \tau_0)\,\phi_{0,1}$$

with $\phi_{0,1}$ the solution for the $[0,1]$-case. The lower bound becomes

$$\int_{\tau_0}^{\tau_M} \phi_{\tau_0,\tau_M}(x)\,dx = (\tau_M - \tau_0)^2 \int_0^1 \phi_{0,1}(x)\,dx = (\tau_M - \tau_0)^2\, 0.1194623$$

Chapter 2

Interval censoring: the NPMLE

The aim of this chapter is to show that the lower bound for estimation of smooth functionals with interval censored data is reached asymptotically by the nonparametric maximum likelihood estimator $\hat{F}_n$. Just like in the preceding chapter, we assume the support of F_0 to consist of one interval $[\tau_0, \tau_M]$. However, we may allow for intervals of constancy of F_0 without changing the proofs. Let $K(F_0)$ be the smooth, real valued aspect of F_0 in which we are interested. From chapter 1 we know that the lower bound is determined by the canonical gradient $\tilde{\theta}_{F_0}$ of the functional $\Theta(Q_F) := K(F)$, defined on the observation space. (In the sequel we neglect the dependence on the observation time distribution in the notations. So we write Q_{F_0} instead of $Q_{F_0,H}$ etc.) We will show that the NPMLE $\hat{\Theta}_n = K(\hat{F}_n)$ of $\Theta(Q_{F_0}) = K(F_0)$ satisfies

$$\sqrt{n}(\hat{\Theta}_n - \Theta(Q_{F_0})) \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{F_0}\|^2_{Q_{F_0}}) \quad \text{as } n \to \infty. \tag{2.1}$$

One further specification is made to the kind of functionals that are allowed. Let $\tilde{\kappa}_{F_0}$ be the canonical gradient at F_0 of the functional $K(F)$, defined on the hidden event time space. We assume:

(K3)
$$K(F) - K(F_0) = \int \tilde{\kappa}_{F_0}(x)\, d(F - F_0)(x) + \mathcal{O}(\|F - F_0\|^2_\lambda),$$

for all distributions F with support contained in $[\tau_0, \tau_M]$. The norm $\|F - F_0\|_\lambda$ is the L_2-distance between the distribution functions F and F_0 w.r.t. Lebesgue measure on $\mathbb{R}$.
For linear functionals

$$G \mapsto \int_{\tau_0}^{\tau_M} c(x)\, dG(x),$$

we have $\tilde{\kappa}_{F_0}(x) = c(x) - \int c\, dF_0$, and (K3) even holds without the $\mathcal{O}$-term. However, condition (K3) is satisfied by a wider class of functionals than the linear ones. For example, the functional

$$K_2(F) = \int F^2(x)\, w(x)\, dx$$

from subsection 1.2.1, with gradient $\kappa_F(x) = 2\int_{s=x}^{\tau_M} F(s)\, w(s)\, ds$, also satisfies (K3), using boundedness of the weight function w. For we have

$$\int_{\tau_0}^{\tau_M} G^2(x)\, w(x)\, dx - \int_{\tau_0}^{\tau_M} F_0^2(x)\, w(x)\, dx =$$

$$\int_{x=\tau_0}^{\tau_M} \left[2\int_{s=x}^{\tau_M} F_0(s)\, w(s)\, ds\right] d(G - F_0) + \int_{\tau_0}^{\tau_M} [G(s) - F_0(s)]^2\, w(s)\, ds$$

We start with an overview of known results of the NPMLE $\hat{F}_n$, with emphasis on case 2. Most results for case 1 are similar in essence. Then we will sketch the proof for case 1 as it appears in Huang and Wellner (1995a). Since the canonical gradient has an explicit expression in case 1, it is the simplest case. Our summary of their proof will serve to illustrate what are the main ingredients needed to show optimality in case 2. A short overview of empirical process theory will be given, since this will be needed in the proof. Throughout this chapter, we use "Prob" to denote the probability measure needed for the asymptotic considerations. More specifically, Prob is the product measure on the sample space of all infinite sequences (X_1, U_1, V_1), $(X_2, U_2, V_2), \ldots$ (in case 2), endowed with the Borel σ-algebra which is generated by the product topology.

2.1 Some known results

Based on the sample of observations $(U_1, V_1, \Delta_1, \Gamma_1), \ldots, (U_n, V_n, \Delta_n, \Gamma_n)$, the NPMLE $\hat{F}_n$ is the (sub)distribution function that maximizes the likelihood

$$\prod_{i=1}^{n} F(U_i)^{\Delta_i}(F(V_i) - F(U_i))^{\Gamma_i}(1 - F(V_i))^{1-\Delta_i-\Gamma_i}\, h(U_i, V_i) \tag{2.2}$$

over the class of non-decreasing cadlag functions F with values in $[0, 1]$. The factor $\prod h(U_i, V_i)$ is of no importance in the maximization procedure with respect to F and can be neglected. First note that only the values of F at the observation times occur explicitly in the likelihood, and even not all of them. If $\Delta_i = 1$, i.e. $X_i \le U_i$, the corresponding V_i does not play any role. Likewise, if $X_j > V_j$, we can throw away the corresponding U_j. The remaining observation points are called the *relevant* observation points. The order restriction on F causes the NPMLE to be a function that is piecewise constant and uniquely defined on large parts of its domain. Generally these intervals contain several observation times. The only places where $\hat{F}_n$ is not uniquely defined is between two consecutive ordered relevant observation times for which $\hat{F}_n$ has a different value. Here $\hat{F}_n$ can be chosen freely. However, how $\hat{F}_n$ is chosen there does not influence the asymptotic properties that follow, since the total length of these intervals shrinks to zero as the sample size goes to infinity. So, without loss of generality, we impose $\hat{F}_n$ to be piecewise constant everywhere, and only to have jumps at (a subset of) the observation points. As before, let τ_i, $i = 1, \ldots, m$ denote the points of jump of $\hat{F}_n$, and define

$$J_i := [\tau_i, \tau_{i+1}),\ i = 0, \ldots, m-1 \qquad \text{and} \qquad J_m := [\tau_m, \tau_M]. \tag{2.3}$$

Then $\hat{F}_n$ is uniquely determined from τ_1 up to and including τ_m. Except for the rare case that all Δ_i's are one, we always have $\hat{F}_n(\tau_0) = 0$ and $\hat{F}_n$ is also uniquely determined from τ_0 to τ_1. At the other end we may end up with a degenerate distribution, having $\hat{F}_n(t) < 1$ at all observation points. This occurs when the largest relevant observation time corresponds with an event time beyond that observation time. Then the largest relevant observation time is equal to some V_j, and $[1 - F(V_j)]$ in the likelihood formula is larger than zero. The NPMLE is not determined beyond this V_j. The asymmetry between the left-hand side and the right-hand side of $[\tau_0, \tau_M]$ is due to the right continuity of the NPMLE. For estimation of smooth functionals based on a finite sample, we have to specify where to put the remaining mass. However, for properties concerning the limit behaviour, this question does not play any role, since the probability to obtain a defective distribution function tends to zero as $n \to \infty$, as long as $F_0 \ll H_1 + H_2$ (condition (M3) on page 15).

Proposition 2.1.1

$$\lim_{n\to\infty} \text{Prob}\{\ \hat{F}_n \text{ is defective }\} = 0$$

Proof: Let $X_{(n)}$ denote the (unobservable) event time corresponding to the largest relevant observation time. Let v_0 denote the left limit of the support of V, with V having distribution function H_2. Then we have, using independence of X and V and integration by parts,

$$\begin{aligned}
\text{Prob}\{\ \hat{F}_n \text{ is defective }\} &= \int_{v_0}^{\tau_M} \text{Prob}\{X_{(n)} > v\}\, nh_2(v)H_2(v)^{n-1}\, dv \\
&= \int_{v_0}^{\tau_M} \{1 - F_0(v)\}\, nh_2(v)H_2(v)^{n-1}\, dv \\
&= \int_{v_0}^{\tau_M} [H_2(v)]^n\, dF_0(v)
\end{aligned}$$

Using Lebesgue's dominated convergence theorem finishes the proof.

□

Given a sample $(U_1, V_1, \Delta_1, \Gamma_1), \ldots, (U_n, V_n, \Delta_n, \Gamma_n)$, let Q_n denote the corresponding empirical probability measure. If we let the NPMLE have its mass restricted to the observation times, proposition 1.3 in GW gives an alternative criterion which is necessary and sufficient for a function to be the NPMLE. Consider the random class $\mathcal{G}_n$ of distribution functions F satisfying

$$\begin{cases} F(U_i) > 0 & , \text{ if } X_i \le U_i, \\ F(V_i) - F(U_i) > 0 & , \text{ if } U_i < X_i \le V_i, \\ 1 - F(V_i) > 0 & , \text{ if } X_i > V_i, \end{cases}$$

and having mass concentrated on the set of observation points augmented with an extra point bigger than all observation points. It is easily seen that $\hat{F}_n$ belongs to this class. For distribution functions in this class, the following process $t \mapsto W_F(t)$ is properly defined:

$$\begin{aligned}
W_F(t) = \quad & \int_{u\in[\tau_0,t]} \delta\, F(u)^{-1} dQ_n(u,v,\delta,\gamma) \\
& - \int_{u\in[\tau_0,t]} \gamma\,\{F(v) - F(u)\}^{-1} dQ_n(u,v,\delta,\gamma) \\
& + \int_{v\in[\tau_0,t]} \gamma\,\{F(v) - F(u)\}^{-1} dQ_n(u,v,\delta,\gamma) \qquad (2.4) \\
& - \int_{v\in[\tau_0,t]} (1-\delta-\gamma)\,\{1 - F(v)\}^{-1} dQ_n(u,v,\delta,\gamma), \\
& \text{for } t \ge \tau_0.
\end{aligned}$$

Proposition 1.3 in GW and the discussion preceding it say:

Proposition 2.1.2 *Let $T_{(n)}$ denote the largest relevant observation time, and let $X_{(n)}$ denote the corresponding (unobservable) event time X_i. Then* *(i) If $X_{(n)} \le T_{(n)}$, $\hat F_n$ maximizes (2.2) over all $F \in \mathcal{G}_n$ if and only if*

$$\int_{[t,\tau_m)} dW_{\hat F_n}(t') \le 0, \qquad \forall t \ge \tau_1, \tag{2.5}$$

and

$$\int_{[\tau_1,\tau_m)} \hat F_n(t)\, dW_{\hat F_n}(t) = 0. \tag{2.6}$$

Moreover, $\hat F_n$ is uniquely determined by (2.5) and (2.6) and is non-defective.

(ii) If $X_{(n)} > T_{(n)}$, $\hat F_n$ maximizes (2.2) over all $F \in \mathcal{G}_n$ if and only if

$$\int_{[t,T_{(n)}]} dW_{\hat F_n}(t') \le 0, \qquad \forall t \ge \tau_1, \tag{2.7}$$

and

$$\int_{[\tau_1,T_{(n)}]} \hat F_n(t)\, dW_{\hat F_n}(t) = 0. \tag{2.8}$$

Moreover, $\hat F_n$ is uniquely determined by (2.7) and (2.8) and is defective.

The following corollary is an immediate consequence.

Corollary 2.1.1 *Any function σ that is constant on the same intervals as $\hat F_n$ satisfies*

$$\begin{aligned}\int_{J_i} \sigma(u)\, dW_{\hat F_n}(u) &= \int_{u\in J_i} \sigma(u)\left\{\frac{\delta}{\hat F_n(u)} - \frac{\gamma}{\hat F_n(v)-\hat F_n(u)}\right\} dQ_n(u,v,\delta,\gamma)\\ &\quad + \int_{v\in J_i} \sigma(v)\left\{\frac{\gamma}{\hat F_n(v)-\hat F_n(u)} - \frac{1-\gamma-\delta}{1-\hat F_n(v)}\right\} dQ_n(u,v,\delta,\gamma)\\ &= 0,\end{aligned}$$

for $i = 1,\ldots,m-1$ (under the conditions of proposition 2.1.2.(i)) or $i = 1,\ldots,m$ (under the conditions of proposition 2.1.2.(ii)).

Proof: Suppose $\hat F_n$ is non-defective. Then we have case (i) of proposition 2.1.2. We now use: if $0 = a_0 < \ldots < a_{m-1}$, $\sum_{j=i}^{m-1} x_j \le 0$ for $i \in \{1,\ldots,m-1\}$, and $\sum_{j=1}^{m-1} a_i x_i = 0$, then $x_1 = \ldots = x_{m-1} = 0$. This easily follows by rewriting

$$\sum_{i=1}^{m-1} a_i x_i = \sum_{i=1}^{m-1} (a_i - a_{i-1}) \sum_{j=i}^{m-1} x_j.$$

Taking $x_i = \int_{J_i} dW_{\hat F_n}$ and $a_i = \hat F_n(\tau_i)$, and using proposition 2.1.2, we derive:

$$\int_{J_i} dW_{\hat F_n} = 0,\ i = 1,\ldots,m-1. \tag{2.9}$$

The proof of (2.9) is completely similar for case (ii) of proposition 2.1.2, and then also holds for the interval J_m. The result now follows, since σ is constant on the intervals J_i. □

Remark. The case 1 analogue of corollary 2.1.1 is derived and used in the proof of proposition 2.1.2, as it appears in GW. The corollary follows from Fenchel duality theory (see e.g. ROCKAFELLAR (1970), theorem 28.3). However, once we have proposition 2.1.2 it can also be used to derive corollary 2.1.1.

Proposition 2.1.2 characterizes maximization of the likelihood, in contrast with the so-called "self-consistency equation" which only yields a necessary but not a sufficient condition. If the points of jump of the NPMLE, and hence the intervals of constancy, were known, the problem would be reduced to a normal maximization problem without order restrictions. Then equations (2.6) and (2.8), or rather corollary 2.1.1, having the partial derivatives of the loglikelihood appearing in the integrand, characterizes the maximization procedure. Equations (2.5) and (2.7) serve to take account of the order restrictions. The fact that only the interval $[\tau_1, \tau_m)$ is playing a role in (2.5) and (2.6) is caused by the extra restriction that the solution should have values between zero and one. In case of the situation in proposition 2.1.2.(ii), $\hat{F}_n(\tau_M) \leq 1$ is fulfilled automatically.

The above characterization of the NPMLE also plays an important role when the NPMLE has to be computed. Contrary to case 1, for which the NPMLE $\hat{F}_n$ can be computed via a one-step procedure, only iterative procedures are available for computation of the NPMLE in case 2. A slight modification of the *iterative convex minorant algorithm*, as introduced in part II of GW, is shown always to converge to the maximizing value in JONGBLOED (1995). See also GROENEBOOM (1996). Computer experiments show that convergence is generally quite fast. Since we only consider theoretical aspects, we do not go into this any further.

Before we give the asymptotic results for the NPMLE, we first group together all conditions that are needed for showing the asymptotic optimality of the NPMLE of smooth functionals in case 2.

Conditions on $X \sim F_0$.

(M1) F_0, with support $[\tau_0, \tau_M]$, is unknown and contained in the class

$$\mathcal{F}_S := \{F \mid \text{support}(F) \subset S;\ F \text{ absolutely continuous, } \sup_x |f(x)| \leq K\},$$

for a fixed $K > 0$ and a fixed bounded interval $S \subset \mathbb{R}$.

(F1) The density satisfies $f_0(x) \geq c_1$ for some $c_1 > 0$ and for all $x \in [\tau_0, \tau_M]$.

Conditions on $(U, V) \sim H$.

(M2) H, with support $\{(u, v) \mid \eta_0 \leq u < v \leq \eta_M\}$, is unknown and contained in $\mathcal{H}$, the collection of all two-dimensional distributions on $\{(u, v) \mid u < v\}$, absolutely continuous with respect to two-dimensional Lebesgue measure and such that (U, V) is independent of X for each choice of $H \in \mathcal{H}$ and $F \in \mathcal{F}_S$.

(H2) The density $h(u, v)$ is continuous.

(H4) **Case A.**

$$h(u, v) = 0 \text{ whenever } |u - v| \leq \epsilon_0 \text{ for some } \epsilon_0 > 0$$

Case B.

$$h(x,x) = \lim_{v \downarrow x} h(x,v) \geq c_2 > 0$$

for all $x \in [\tau_0, \tau_M]$ and some $c_2 > 0$.

(H1) The marginal densities h_1 and h_2 of U and V are continuous, satisfying $h_1(x)+h_2(x) > 0$ for all $x \in [\tau_0, \tau_M]$.

(H3) The partial derivatives $\frac{\partial}{\partial x}h(x,t)$ and $\frac{\partial}{\partial x}h(t,x)$ exist, except for at most a finite number of points x, where left and right derivatives with respect to x do exist for each t. The derivatives are bounded, uniformly in t and x.

(M3) If both H_1 and H_2 put zero mass on some set A, then $F \in \mathcal{F}_S$ has zero mass on A as well, so $F \ll H_1 + H_2$.

Conditions on the functional K.

(K1) The hidden-space canonical gradient $\tilde{\kappa}_{F_0}$ and its derivative $\tilde{\kappa}'_{F_0} \equiv k$ are continuous.

(K2) k is differentiable, except for at most a finite number of points x, where left and right derivatives exist. k' is bounded, uniformly over $x \in [\tau_0, \tau_M]$.

(K3) For all distributions F with support contained in $[\tau_0, \tau_M]$ we have

$$K(F) - K(F_0) = \int \tilde{\kappa}_{F_0}(x)\, d(F - F_0)(x) + \mathcal{O}(\|F - F_0\|_\lambda^2).$$

For case 1, (M1), (K1), (K3) and the one-dimensional analogues of (M2), (M3), (H1) and (H2) are needed as well. (F1), (K2), and the one-dimensional analogue of (H3), together, are replaced by the combined condition

(C) $(k/g) \circ F_0^{-1}$ is Lipschitz on $[0,1]$,

with g denoting the density of the observation time distribution.

For case 2, mixtures of A and B may occur as well, meaning that there is positive mass along part of the diagonal. We make a short remark about this when case 2B is treated.

Asymptotic results

We have uniform consistency of the NPMLE of F_0:

Proposition 2.1.3

$$\text{Prob}\left\{\lim_{n\to\infty} \|\hat{F}_n - F_0\|_\infty = 0\right\} = 1$$

Proof: See GW, part II, sections 4.1 (case 1) and 4.3 (case 2).

A rate of convergence result that will be needed can be deduced from VAN DE GEER (1993) (case 1) and VAN DE GEER (1996) (case 2). For case 2, define the densities $q_{\hat{F}_n}$ and q_{F_0}, with respect to $H \otimes \nu_2$, by

$$q_{\hat{F}_n}(u, v, \delta, \gamma) := \delta \hat{F}_n(u) + \gamma\{\hat{F}_n(v) - \hat{F}_n(u)\} + (1 - \gamma - \delta)\{1 - \hat{F}_n(v)\}, \tag{2.10}$$

and

$$q_{F_0}(u, v, \delta, \gamma) := \delta F_0(u) + \gamma\{F_0(v) - F_0(u)\} + (1 - \delta - \gamma)\{1 - F_0(v)\}. \tag{2.11}$$

Similar definitions hold for case 1. Define the Hellinger distance $h(q_{F_1}, q_{F_2})$ by

$$h(q_{F_1}, q_{F_2}) = \left[\frac{1}{2}\int \left(\sqrt{q_{F_1}} - \sqrt{q_{F_2}}\right)^2 dH \otimes \nu_2\right]^{1/2}$$

Proposition 2.1.4 *Let G denote the distribution function of the observation times in case 1 and let H_1 and H_2 denote the marginal distribution functions of respectively the first and the second observation time of the pair of observation times (U, V) in case 2. Then*

(i) The Hellinger distance $h(q_{\hat{F}_n}, q_{F_0})$ satisfies

$$h(q_{\hat{F}_n}, q_{F_0}) = \mathcal{O}_p(n^{-1/3}) \quad \text{as } n \to \infty \qquad \text{for case 1 and 2A}$$

and

$$h(q_{\hat{F}_n}, q_{F_0}) = \mathcal{O}_p(n^{-1/3}(\log n)^{1/6}) \quad \text{as } n \to \infty \qquad \text{for case 2B.}$$

(ii) Similar rates hold for the L_2-distance $\|\hat{F}_n - F_0\|_G$ and $\|\hat{F}_n - F_0\|_{H_i}$, $i = 1, 2$.

Proof:
ad (i): For case 1, the result is proved in example 4.8(a) in VAN DE GEER (1993). The result for case 2 is proved in example 3.2 in VAN DE GEER (1996) for a particular choice of F_0 and H, belonging to subcase B. It is accompanied by the remark that the result also holds if both h and f_0 remain bounded away from zero. These conditions can be relaxed to $H(F_0^{-1}(u), F_0^{-1}(v))$ being Lipschitz in both variables. For case 2A, the convergence can be shown to be a little faster, since her truncation devices

$$\int_{\{q_{F_0} > \sigma_n\}} \frac{1}{q_{F_0}}\, d\mu$$

and

$$\int_{\{q_{F_0} \leq \sigma_n\}} q_{F_0}\, d\mu$$

are not needed. A self-contained proof for case 2A, using the general theory in VAN DE GEER (1996), can be found in GESKUS AND GROENEBOOM (1996A).
ad (ii): We use part (i) and

$$(\hat{F}_n - F_0)^2 \leq 4\left(\sqrt{\hat{F}_n} - \sqrt{F_0}\right)^2 \text{ and } (\hat{F}_n - F_0)^2 \leq 4\left(\sqrt{1 - \hat{F}_n} - \sqrt{1 - F_0}\right)^2.$$

In case 2, considering the parts $\delta = 1$ and $\delta = \gamma = 0$ separately, we obtain the desired results for H_1 and H_2. The result for G in case 1 is obtained in a similar way. □

Although not used in the sequel, we spend some words on the asymptotic distribution of $\hat{F}_n(t_0)$, for fixed $t_0 \in [\tau_0, \tau_M]$. Contrary to smooth functionals, $K(F) = F(t_0)$ cannot be estimated at $\sqrt{n}$-rate. Moreover, the limit distribution is no longer normal. The limit distribution is determined by a random variable Z, defined as the last time where standard two-sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

For case 1 we have theorem 5.1 in GW:

Theorem 2.1.1 *Let t_0 be such that $0 < F_0(t_0), G(t_0) < 1$, and suppose that f_0 and g_0 are continuous at t_0 and strictly positive. Then we have, as $n \to \infty$,*

$$n^{1/3} \left[\frac{\hat{F}_n(t_0) - F_0(t_0)}{\left[\frac{1}{2} f_0(t_0)/c_1(t_0)\right]^{1/3}} \right] \xrightarrow{\mathcal{D}} 2Z, \tag{2.12}$$

with

$$c_1(t_0) = \frac{g(t_0)}{F_0(t_0)[1 - F_0(t_0)]}.$$

For case 2A we have:

Theorem 2.1.2 *Let $h_1(\tau_0) > 0$ and $h_2(\tau_M) > 0$. Moreover let the conditions (F1), (H1), (H2), (H3) and (H4).Case A be satisfied.*
Let t_0 be such that $0 < F_0(t_0), H(t_0, t_0) < 1$.
Define

$$k_1(u) := \int_u^{\tau_M} \frac{h(u,v)}{F_0(v) - F_0(u)}\, dv$$

and

$$k_2(v) := \int_{\tau_0}^{v} \frac{h(u,v)}{F_0(v) - F_0(u)}\, du.$$

Then

$$n^{1/3} \left[\frac{\hat{F}_n(t_0) - F_0(t_0)}{\left[\frac{1}{2} f_0(t_0)/c_2(t_0)\right]^{1/3}} \right] \xrightarrow{\mathcal{D}} 2Z, \tag{2.13}$$

with

$$c_2(t_0) = \frac{h_1(t_0)}{F_0(t_0)} + k_1(t_0) + k_2(t_0) + \frac{h_2(t_0)}{1 - F_0(t_0)}.$$

Proof: See GROENEBOOM (1996).

If the relative amount of mass of the (U, V)-distribution near the diagonal point (t_0, t_0), compared to the amount of mass of F near t_0 is very small we are in a case 1-type situation and we have a $n^{1/3}$-convergence rate. Although rate and limit distribution are different, the norming constant in the cases 1 and 2A shows some similarities to the integrated score function for smooth functionals. The conditions needed are comparable as well, the main difference being that here most of them only have to hold at the point of interest t_0.

For case 2B, the limit distribution of $\hat{F}_n(t_0)$ has not been established yet. It is conjectured to have the same asymptotic distribution as the "toy estimator" $F_n^{(1)}(t_0)$, which is obtained by doing one step in the iterative convex minorant algorithm, with the true underlying distribution F_0 as starting value. Of course, this procedure, which does not lead to an estimator in the strict sense, has no practical value. However, the asymptotic distribution of $F_n^{(1)}(t_0)$ is known. For case 2A, the same working hypothesis was originally used in WELLNER (1995), proving the above limit behaviour to hold for $F_n^{(1)}(t_0)$ as well in case 2A. For $F_n^{(1)}(t_0)$, the convergence rate in case 2B increases to $(n \log n)^{1/3}$. Here the norming constant is completely different from the one in case of smooth functionals. So either the limit distribution of $\hat{F}_n(t_0)$ is different from that of $F_n^{(1)}(t_0)$, or in case 2B the norming constant is quite different from the one in case 1 and case 2A, and has no similarities with the integral equation for smooth functionals. For case 2B we have:

Theorem 2.1.3 *Let $0 < F_0(t_0), H(t_0, t_0) < 1$. Let f_0 be continuous at t_0, with $f_0(t_0) > 0$. Suppose that the density $h(u, v)$ is continuous at (u, v) if (u, v) is sufficiently close to (t_0, t_0). Let $h(t_0, t_0) > 0$ and suppose that $h(t, t)$, defined by*

$$h(t,t) = \lim_{v \downarrow t} h(t,v),$$

is continuous in t, for t in a neighbourhood of t_0.
Then

$$(n \log n)^{1/3} \left[\frac{F_n^{(1)}(t_0) - F_0(t_0)}{\left[\frac{1}{2} f_0(t_0)/c_3(t_0)\right]^{1/3}} \right] \xrightarrow{\mathcal{D}} 2Z, \tag{2.14}$$

with

$$c_3(t_0) = \tfrac{2}{3} h(t_0, t_0)/f_0(t_0).$$

Proof: See GW, theorem 5.3.

2.2 Case 1: the main ingredients

The following theorem shows asymptotic optimality of the NPMLE of smooth functionals in case 1. It is a slight modification of theorem 5.1 in HUANG AND WELLNER (1995A). The basic ingredients of the proof serve as an introduction to the techniques used in case 2.

Theorem 2.2.1 *Let F_0 have a bounded support $[\tau_0, \tau_M]$, with $F_0 \in \mathcal{F}_S$. Let the observation time distribution G satisfy $F_0 \ll G$. Let G have a continuous density g, satisfying $g(t) > 0$ for all $t \in [\tau_0, \tau_M]$. Let the functional K satisfy (K1) and (K3) on page 64. Moreover suppose that $(\tilde{\kappa}'_{F_0}/g) \circ F_0^{-1}$ is a bounded Lipschitz function on $[0,1]$. Then we have*

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{F_0}\|^2_{Q_{F_0}}) \quad as \quad n \to \infty$$

Basic ingredients of the proof:
We may assume $\hat{F}_n$ to be piecewise constant. Moreover, by proposition 2.1.1, we may assume

$\hat{F}_n(\tau_M) = 1$. Let Q_n denote the empirical measure of the observations $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$. It is sufficient to show the following:

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] = \sqrt{n} \int \tilde{\theta}_{F_0}\, d(Q_n - Q_{F_0}) + o_p(1). \tag{2.15}$$

Then an application of the central limit theorem finishes the proof. The proof of (2.15) consists of the following steps.

I: The nonlinear aspect of the functional is negligible.
By condition (K3), we have

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] = \sqrt{n} \int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) + o_p(1),$$

if we can show

$$\|\hat{F}_n - F_0\|_\lambda = o_p(n^{-1/4}). \tag{2.16}$$

This follows from proposition 2.1.4.(ii), using $d\lambda = (1/g)\,dG$ and $g > 0$.

II: Transformation to observation space measure.
The expression $\sqrt{n} \int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0)$ is an integral with respect to the measure $(\hat{F}_n - F_0)$ in the hidden event time space. (Note that $\int \tilde{\kappa}_{F_0} dF_0 = 0$.) We now show that it can be rewritten as an integral with respect to the probability measure of the observations Q_{F_0}. Define the function $\tilde{\theta}_{\hat{F}_n}$ by

$$\tilde{\theta}_{\hat{F}_n}(t, \delta) = \begin{cases} -[1 - \hat{F}_n(t)]\,[\tilde{\kappa}'_{F_0}(t)/g(t)] & \text{if } \delta = 1 \\ \hat{F}_n(t)\,[\tilde{\kappa}'_{F_0}(t)/g(t)] & \text{if } \delta = 0 \end{cases}$$

If $\hat{F}_n$ is replaced by F_0, this is the canonical gradient formula for case 1. Note that $\tilde{\theta}_{\hat{F}_n}$ no longer has an interpretation as canonical gradient, since $\hat{F}_n$, being a discrete distribution function, is not dominated by G. The following holds:

$$\int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) = -\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0}. \tag{2.17}$$

This is easily seen by writing out the definition of $\tilde{\theta}_{\hat{F}_n}$:

$$\begin{aligned} -\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0} &= \int \left\{ \left([1 - \hat{F}_n]\,\frac{\tilde{\kappa}'_{F_0}}{g}\, F_0 \right) - \left(\hat{F}_n\, \frac{\tilde{\kappa}'_{F_0}}{g}\, [1 - F_0] \right) \right\} dG \\ &= \int (\hat{F}_n - F_0)\, d\tilde{\kappa}_{F_0} \\ &= \int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0), \end{aligned}$$

using integration by parts and $\hat{F}_n(\tau_M) = 1$.

III: NPMLE condition; inserting of empirical measure.
Now we will use the fact that we deal with the NPMLE. The case 1 equivalent of corollary 2.1.1 says that any function σ that is constant on the same intervals J_i as $\hat{F}_n$ satisfies

$$\sum_{\delta=0}^{1} \int_{t \in J_i} \sigma(t) \left\{ \tfrac{\delta}{\hat{F}_n(t)} - \tfrac{1-\delta}{1-\hat{F}_n(t)} \right\} dQ_n(t, \delta) = 0$$

Let $\overline{\theta}_{\hat{F}_n}$ denote the function $\tilde{\theta}_{\hat{F}_n}$, but with $\xi := \tilde{\kappa}'_{F_0}/g$ replaced by a function that is constant on the intervals J_i. The following is an obvious choice:

$$\overline{\xi}(t) := [\xi \circ F_0^{-1}](\hat{F}_n(t))$$

Then we obtain, using $\sigma = \hat{F}_n[1 - \hat{F}_n]\,\overline{\xi}$,

$$\int \overline{\theta}_{\hat{F}_n}\, dQ_n = 0.$$

So (2.17) can be rewritten as

$$-\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0} = \int \overline{\theta}_{\hat{F}_n}\, d(Q_n - Q_{F_0}) + \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n})\, dQ_{F_0}. \tag{2.18}$$

Using the formulas for $\overline{\theta}_{\hat{F}_n}$ and $\tilde{\theta}_{\hat{F}_n}$, we get

$$\int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n})\, dQ_{F_0} = \int \left[(\hat{F}_n - F_0)\,(\overline{\xi} - \xi)\right] dG.$$

By applying Cauchy-Schwarz we see that we have to prove

$$\|\hat{F}_n - F_0\|_G \times \|\overline{\xi} - \xi\|_G = o_p(n^{-1/2}).$$

Since $F_0^{-1}(F_0(x)) = x$ a.e.-F_0 and $\tilde{\kappa}'_{F_0}$ can be taken zero at places where F_0 does not have mass, $\xi(t)$ can be replaced by $\xi(F_0^{-1}(F_0(t))$ in the $L_2(G)$-norm. Using the Lipschitz condition for $\xi \circ F_0^{-1}$, we see that it is sufficient to prove

$$\|\hat{F}_n - F_0\|_G = o_p(n^{-1/4}) \tag{2.19}$$

which again follows from proposition 2.1.4.(ii).

IV: Closeness in empirical process.
The first term in (2.18) is further split into

$$\begin{aligned}\int \overline{\theta}_{\hat{F}_n} d(Q_n - Q_{F_0}) \;&=\; \int \tilde{\theta}_{F_0} d(Q_n - Q_{F_0}) \\ &\quad + \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}) d(Q_n - Q_{F_0}).\end{aligned}$$

The last term is $o_p(n^{-1/2})$. To show this we need to use some empirical process theory.

2.3 Empirical processes

We need to show that

$$\sqrt{n} \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}) d(Q_n - Q_{F_0}) = o_p(1).$$

This is performed by considering the empirical process

$$\nu_{n,F_0}(t_F) = \sqrt{n}(Q_n - Q_{F_0})(t_F),$$

indexed by the set $\{t_F\} = \{\overline{\theta}_F : F \in \mathcal{F}\} \cup \{\tilde{\theta}_{F_0}\}$. $\mathcal{F}$ should be such that it contains all possible realizations $\hat{F}_n$, for each n, or at least a subset of all possible realizations occurring with a probability tending to one as $n \to \infty$. So the defective distribution functions may be excluded beforehand. The process $\nu_{n,F_0}(\cdot)$ has to converge to a tight Gaussian process $G_{F_0}(\cdot)$. If this holds, then $\nu_{n,F_0}(\overline{\theta}_F)$ and $\nu_{n,F_0}(\tilde{\theta}_{F_0})$ are close with high probability whenever the indices are. Closeness of $\overline{\theta}_F$ and $\tilde{\theta}_{F_0}$ is shown to hold for $F = \hat{F}_n$ with probability tending to one, using convergence of $\hat{F}_n$ to F_0 and corresponding convergence of $\overline{\theta}_{\hat{F}_n}$ to $\tilde{\theta}_{F_0}$.

In the general setting, let $X_1, \ldots, X_n$ be a sample of i.i.d. observations, each with distribution P on the sample space $(\mathcal{X}, \mathcal{A})$. $P_n = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i\}}$ is the empirical measure, based on this sample. Let $\mathcal{F}$ denote a class of functions, being a subset of $L_2(P)$. Define the empirical process

$$\nu_{n,P}(\cdot) := \sqrt{n}(P_n - P)(\cdot)$$

as a process on the index set $\mathcal{F}$. Assume that

$$\sup_{f\in\mathcal{F}} |f(x) - P(f)| < \infty \qquad \text{for every } x. \tag{2.20}$$

For each finite subset $\{f_1, \ldots, f_k\} \subset \mathcal{F}$ we have, by the multivariate central limit theorem,

$$(\nu_{n,P}(f_1), \ldots, \nu_{n,P}(f_k)) \xrightarrow{\mathcal{D}} N(0, \Sigma)$$

with the matrix Σ having coefficients $\alpha_{i,j} = P[f_i - P(f_i)][f_j - P(f_j)]$. For the limit distribution of the empirical process over $\mathcal{F}$, we have to define a space in which $\{\nu_{n,P}(f) : f \in \mathcal{F}\}$ takes its values. By (2.20), this space can be taken to be the Banach space of all bounded functions B from $\mathcal{F}$ to $\mathbb{R}$:

$$\ell^\infty(\mathcal{F}) := \{B : \mathcal{F} \to \mathbb{R} : \|B\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |B(f)| < \infty\}.$$

Convergence in distribution of $\nu_{n,P}$ to a Borel measurable process G_P is defined as

$$E^*(h(\nu_{n,P})) \to E(h(G_P))$$

for all bounded $\|\cdot\|_{\mathcal{F}}$-continuous real-valued functions h on $\ell^\infty(\mathcal{F})$, using outer expectations E^* whenever $\nu_{n,P}$ is not Borel measurable.

We say that the uniform central limit theorem holds at P if G_P is a tight Borel measurable element in $\ell^\infty(\mathcal{F})$. A class $\mathcal{F}$ for which the uniform central limit theorem holds at P is called a *P-Donsker class*.

Consider the following semi-metric on $\mathcal{F}$:

$$\rho_P(f, g) := \left\{\int [(f - g) - P(f - g)]^2 \, dP\right\}^{1/2}.$$

$\{\nu_{n,P}\}$ is called *asymptotically uniformly equicontinuous in probability* on $\mathcal{F}$ with respect to ρ_P if for every $\epsilon, \eta > 0$ there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} \text{Prob}^*\left(\sup_{\rho_P(f,g)<\delta} |\nu_{n,P}(f) - \nu_{n,P}(g)| > \epsilon\right) < \eta$$

with Prob* denoting outer probability on the relevant product sample space. We have the following theorem (see e.g. VAN DER VAART AND WELLNER (1996), example 1.4.9).

Theorem 2.3.1 *The sequence $\{\nu_{n,P}\}$ converges in distribution to a tight Gaussian process G_P if and only if the following three properties hold*

- $(\nu_{n,P}(f_1), \ldots, \nu_{n,P}(f_k)) \xrightarrow{\mathcal{D}} N(0, \Sigma)$
- *$\{\nu_{n,P}\}$ is asymptotically uniformly ρ_P-equicontinuous in probability on $\mathcal{F}$.*
- *$\mathcal{F}$ is totally bounded for ρ_P*

Remark. If $\sup_{f \in \mathcal{F}} |P(f)| < \infty$, the ρ_P metric can be replaced by the $L_2(P)$ metric (see VAN DER VAART AND WELLNER (1996), problem 2.1.2).

The stochastic equicontinuity property is the one that is needed. However, this is by far the strongest property of the three; hence showing the P-Donsker class property is almost similar. So our question is: when is a class of functions a P-Donsker class? We restrict ourselves to classes of cadlag functions. For our purpose, the definition of univariate cadlag functions is extended to bivariate functions in the following way. A function f defined on (a subset of) $\mathbb{R}^2$ is called cadlag if

$$f(x,y) = f(x+, y) = f(x, y+) = f(x+, y+)$$

Sufficient for $\mathcal{F}$ to be a P-Donsker class is $\mathcal{F}$ to be of uniformly bounded variation (see e.g. VAN DER LAAN (1993)). This can be characterized via the *variation norm* $\|\cdot\|_V^*$. For the one-dimensional case, if f is a cadlag function on $[b_0, b_M] \subset \mathbb{R}$, its variation norm is defined as

$$\|f\|_{V_1}^* := \max\{\|f\|_\infty, \|f\|_{V_1}\},$$

with the norm $\|\cdot\|_{V_1}$ defined by

$$\|f\|_{V_1} := \sup_{I_1,\ldots,I_M \in \mathcal{I}} \sum_{j=1}^{M} |f(b_j) - f(b_{j-1})|.$$

Here $\mathcal{I}$ is the set of all disjoint partitions $I_j = (b_{j-1}, b_j]$ of $(b_0, b_M]$.

For a bivariate real-valued cadlag function f on $[a_1, a_2] \times [b_1, b_2] \subset \mathbb{R}^2$, the variation norm is defined as the maximum of four norms:

$$\|f\|_{V_2}^* := \max\left\{\|f\|_\infty,\ \|f\|_{V_2},\ \sup_x \|f(x,\cdot)\|_{V_1},\ \sup_y \|f(\cdot,y)\|_{V_1}\right\}$$

The norm $\|f\|_{V_2}$ is defined as

$$\|f\|_{V_2} := \sup_{A_{ij}} \sum_{i,j} |f(A_{ij})|,$$

where the supremum is taken over all finite rectangular partitions $\{A_{ij}\}$ of $(a_1, a_2] \times (b_1, b_2]$. If A_{ij} is of the form $(s,x] \times (t,y]$, then $f(A_{ij})$ is defined as

$$f(A_{ij}) := f(x,y) - f(s,y) - f(x,t) + f(s,t).$$

IV: Closeness in empirical process, continued.
Now we return to case 1. The index set is formed by the class

$$\mathcal{T} = \{\overline{\theta}_F \mid F \in \mathcal{F}\} \cup \{\tilde{\theta}_{F_0}\}$$

with $\mathcal{F}$ being the class of piecewise constant non-defective distribution functions with mass contained in $[\tau_0, \tau_M]$. By the Lipschitz condition on $\xi \circ F_0^{-1}$ this class is easily shown to have a uniformly bounded variation norm, hence it is a Q_{F_0}-Donsker class. Now we use theorem 2.3.1 and the subsequent remark. Again by the Lipschitz property of $\xi \circ F_0^{-1}$, we have

$$\begin{aligned} \|\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}\|_{Q_{F_0}} &\leq C\|\hat{F}_n - F_0\|_{Q_{F_0}} \\ &= C\|\hat{F}_n - F_0\|_G \end{aligned}$$

Using convergence of $\hat{F}_n$ to F_0, e.g. in $L_2(G)$-norm (proposition 2.1.4 once more) or in supremum norm (see GW, section II.4.1), we obtain the desired result, which ends the proof for case 1

□

In the next subsection, on case 2A, we more or less repeat the above proof. However, some things are slightly different. Moreover, the proof is more complicated since we lack an explicit formula for the canonical gradient. For case 2B the main difficulty is in part IV, closeness in empirical process. Standard results on Donsker classes cannot be used, due to the singularity in $\gamma.h(u,v)/[F(v) - F(u)]$.

2.4 Case 2A: observation times bounded away

The following theorem will be shown to be valid.

Theorem 2.4.1 *Let the conditions on page 63 be satisfied, except for (H4).Case B. Then we have*

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{F_0}\|^2_{Q_{F_0}}) \quad as \quad n \to \infty$$

Proof:
Again it is sufficient to show the following:

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] = \sqrt{n}\int \tilde{\theta}_{F_0}\, d(Q_n - Q_{F_0}) + o_p(1). \tag{2.21}$$

By the strong consistency of the NPMLE (proposition 2.1.3), $f_0 \geq c_1$ and condition H4.Case A, there exists a constant c', such that

$$\hat{F}_n(u) - \hat{F}_n(t) \geq c', \text{ if } u - t \geq \epsilon_0, \tag{2.22}$$

with probability tending to one if n is sufficiently large. This c' determines the ϵ in condition (CF) on page 32, giving the definition of the class $\mathcal{F}$, consisting of the non-defective, piecewise constant distribution functions that have enough increase to prevent occurrence of a singularity in $h(u, v)/(F(v) - F(u))$:

$$\mathcal{F} := \{F : F \text{ is non-defective and satisfies (CF) on page 32}\}.$$

Using proposition 2.1.1 and property (2.22) we see that $\hat{F}_n \in \mathcal{F}$, with probability tending to one as $n \to \infty$. Hence we may restrict ourselves to $\hat{F}_n \in \mathcal{F}$.
Now we take the same steps as in the proof of theorem 2.2.1.

I: Functional almost linear.

From $h_1 + h_2 > 0$, (K3) and proposition 2.1.4 we derive

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] = \sqrt{n}\int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) + o_p(1)$$

II: Transformation to observation space.

In case 1, $\tilde{\theta}_{\hat{F}_n}$ was defined as

$$\tilde{\theta}_{\hat{F}_n}(t, \delta) = -\delta\frac{\phi_{\hat{F}_n}(t)}{\hat{F}_n(t)} + (1-\delta)\frac{\phi_{\hat{F}_n}(t)}{1-\hat{F}_n(t)},$$

with $\phi_{\hat{F}_n}$ the solution of the equation

$$\frac{\phi_{\hat{F}_n}(x)}{1-\hat{F}_n(x)}g(x) + \frac{\phi_{\hat{F}_n}(x)}{\hat{F}_n(x)}g(x) = \tilde{\kappa}'_{F_0}(x).$$

This equation has its analogue in the integral equation

$$\phi_F(x)/d_F(x) + \left[\int_{t=\eta_0}^{x} \tfrac{\phi_F(x)-\phi_F(t)}{F(x)-F(t)}\, h(t,x)\, dt - \int_{t=x}^{\eta_M} \tfrac{\phi_F(x)-\phi_F(t)}{F(t)-F(x)}\, h(x,t)\, dt\right] = $$
$$= \tilde{\kappa}'_{F_0}(x) \qquad (2.23)$$

with $F = \hat{F}_n$. So we look at this equation and the corresponding ξ_F-equation for $F \in \mathcal{F}$. Note that ξ_F is only defined on the interval from the first point of jump of $\hat{F}_n$ to its last one, say on $[\tau_1(\hat{F}_n), \tau_m(\hat{F}_n)]$. By theorem 1.4.6, we know both equations to have a unique solution in $D([\tau_0, \tau_M])$ and $D([\tau_1(\hat{F}_n), \tau_m(\hat{F}_n)])$ respectively.

Contrary to case 1, where ξ was given by $\xi = \tilde{\kappa}'_{F_0}/g$, the ξ-function for case 2 depends on F. Moreover, this ξ-function is no longer continuous everywhere. It is a cadlag function instead, having jumps at the same points as F. The solution ϕ_F can be used to extend the definition of $\tilde{\theta}_F$ to $F \in \mathcal{F}$:

$$\tilde{\theta}_F(u, v, \delta, \gamma) := -\delta\,\frac{\phi_F(u)}{F(u)} - \gamma\,\frac{\phi_F(v) - \phi_F(u)}{F(v) - F(u)} + (1-\delta-\gamma)\,\frac{\phi_F(v)}{1-F(v)}, \qquad (2.24)$$

where $\phi_F(u)/F(u)$ and $\phi_F(v)/(1 - F(v))$ are defined to be zero if $F(u) = 0$ or $F(v) = 1$, respectively. At points where the denominator in the γ-part of (2.24) is zero, we have $h(u, v) = 0$ as well. There we need not define $\tilde{\theta}_F$, since the integral on the right-hand side of lemma 2.4.1 can be restricted to $\{\eta_0 \le u \le v - \epsilon_0 \le \eta_M\}$.

Lemma 2.4.1 *For any $F \in \mathcal{F}$ we have*

$$\int \tilde{\kappa}_{F_0}\, d(F - F_0) = -\int \tilde{\theta}_F\, dQ_{F_0}$$

Proof:
Let, for any distribution function $F \in \mathcal{F}$, L_F: $L_2(F) \rightarrow L_2(Q_F)$ denote the conditional expectation operator

$$\begin{aligned}[L_F a](u, v, \delta, \gamma) &= E_F[a(X)|U = u, V = v, \Delta = \delta, \Gamma = \gamma] \\ &= \frac{\delta \int_{\tau_0}^{u} a\, dF}{F(u)} + \frac{\gamma \int_{u}^{v} a\, dF}{F(v) - F(u)} + \frac{(1 - \delta - \gamma) \int_{v}^{\tau_M} a\, dF}{1 - F(v)} \quad \text{a.e.-}[Q_F],\end{aligned}$$

with adjoint given by

$$[L^* b](x) = E[b(U, V, \Delta, \Gamma)|X = x] \quad \text{a.e.-}F.$$

Since the adjoint is an expectation, conditionally on the value of the random variable $X \sim F$, its structure does not depend on F. F only determines where it has to be defined (the a.e.-F part). Still $a \in L_2^0(F)$ implies $L_F(a) \in L_2^0(Q_F)$.

Note that $\tilde{\theta}_F \in L_2(Q_{F_0})$: for $\delta = 1$ and $\delta = \gamma = 0$ we use boundedness of ξ_F, for $\gamma = 1$ we use boundedness of ϕ_F, together with condition (H4).Case A. Let $1 \in L_2(F)$ denote the constant function $1(x) = 1$, $x \in \mathbb{R}$. Under L_F this transforms into the constant function $1^\circ(u, v, \delta, \gamma) = 1$ on $L_2(Q_F)$. Now we have,

$$\begin{aligned}\int \tilde{\theta}_F\, dQ_{F_0} &= < \tilde{\theta}_F, 1^\circ >_{Q_{F_0}} \\ &= < \tilde{\theta}_F, L_{F_0}(1) >_{Q_{F_0}} \\ &= < L^*(\tilde{\theta}_F), 1 >_{F_0} \\ &= \int L^*(\tilde{\theta}_F)\, dF_0.\end{aligned}$$

If we can prove

$$L^*(\tilde{\theta}_F) = \tilde{\kappa}_{F_0} - \int \tilde{\kappa}_{F_0}\, dF \quad \text{a.e.-}F_0$$

we are done.
This is shown as follows:
The integral equation was obtained by taking derivatives in the equation $\tilde{\kappa}_{F_0}(x) = [L^* \tilde{\theta}_{F_0}](x)$ for all $x \in [\tau_0, \tau_M]$. Now we will go the other way, integrate, but replace $\tilde{\theta}_{F_0}$ by $\tilde{\theta}_F$, obtaining

$$[L^* \tilde{\theta}_F](x) = [\tilde{\kappa}_{F_0}](x) + C \quad \text{for all } x \in [\tau_0, \tau_M].$$

For the constant C we have, using that F is non-defective,

$$\begin{aligned}C &= \int C\, dF \\ &= \int L^*(\tilde{\theta}_F)\, dF - \int \tilde{\kappa}_{F_0}\, dF \\ &= < L^*(\tilde{\theta}_F), 1 >_F - \int \tilde{\kappa}_{F_0}\, dF\end{aligned}$$

It is easily shown that $\tilde{\theta}_F$ is contained in $L_2^0(Q_F)$. (However, it is *not* contained in $\mathcal{R}(L_F)$, since ϕ_F is not piecewise constant.) Now we have

$$\begin{aligned} < L^*(\tilde{\theta}_F), 1 >_F &= < \tilde{\theta}_F, L_F(1) >_{Q_F} \\ &= < \tilde{\theta}_F, 1^\circ >_{Q_F} \\ &= 0 \end{aligned} \tag{2.25}$$

□

Remarks.

I. This result can also be proved by writing out the integrals using definition (2.24).

II. For case 1, the equality

$$\int \tilde{\kappa}_{F_0}\, d(F - F_0) = -\int \tilde{\theta}_F\, dQ_{F_0}$$

was first shown to hold in GW. In VAN DER LAAN (1993), this equality is derived for a general class of missing data models which allow for complete observations. The interval censoring models do not belong to this class, however, since direct observations do not occur. The above proof suggests that the equality holds more generally in missing data models, also when direct observations do not occur. Basically, what is needed is:

- F is non-defective
- $[L^*\tilde{\theta}_F](x) = [\tilde{\kappa}_{F_0}](x) + C$ for all $x \in [\tau_0, \tau_M]$
- $\tilde{\theta}_F \in L_2(Q_{F_0})$
- $\tilde{\theta}_F \in L_2^0(Q_F)$

Note that $\tilde{\theta}_{\hat{F}_n}$ does not belong to the range (nor the closure of the range) of the score operator $L_{\hat{F}_n}$. A modification $\overline{\phi}_{\hat{F}_n}$ is introduced below, which does belong to the range of $L_{\hat{F}_n}$.

III. Validity of the lemma is not restricted to this choice of $\tilde{\theta}_F$. The same result holds if we had based $\tilde{\theta}_F$ on the non-canonical gradient at the beginning of section 1.4.2, i.e.

$$\tilde{\theta}_F(u, v, \delta, \gamma) := -\delta\, \tilde{\kappa}'_{F_0}(u)\, \frac{1 - F(u)}{h_1(u)} + (1 - \delta - \gamma)\, \tilde{\kappa}'_{F_0}(v)\, \frac{F(v)}{h_2(v)}.$$

Since this definition is given by an explicit formula, the lemma can be proved by a simple direct computation, like in case 1. Of course, this cannot lead to the optimality result, so one of the next two steps should go wrong. Indeed, part III does not hold for this choice of $\tilde{\theta}_F$: if we insert the empirical measure, the correction we have to make to get a piecewise constant function is not negligible in the limit.

III: Inserting of empirical measure.

Now we will use the NPMLE characterization, corollary 2.1.1. Since $\phi_{\hat{F}_n}$ and $\xi_{\hat{F}_n}$ are not piecewise constant, we introduce the functions $\overline{\phi}_{\hat{F}_n}$ and $\overline{\xi}_{\hat{F}_n}$. These functions are constant on the same intervals $J_i = [\tau_i, \tau_{i+1})$ as $\hat{F}_n$. The value of $\overline{\xi}_{\hat{F}_n}$ on J_i is defined to be:

$$\begin{array}{lll} \overline{\xi}_{\hat{F}_n}(J_i) := & \xi_{\hat{F}_n}(s) & \text{if there exists a point } s \in J_i \text{ with } \hat{F}_n(s) = F_0(s) \\ \overline{\xi}_{\hat{F}_n}(J_i) := & \xi_{\hat{F}_n}(\tau_i) & \text{if } F_0(x) > \hat{F}_n(\tau_i) \text{ for all } x \in J_i \\ \overline{\xi}_{\hat{F}_n}(J_i) := & \xi_{\hat{F}_n}(\tau_{i+1}-) & \text{if } F_0(x) < \hat{F}_n(\tau_i) \text{ for all } x \in J_i \end{array}$$

The function $\overline{\phi}_{\hat{F}_n}$ is defined as

$$\overline{\phi}_{\hat{F}_n}(x) := \hat{F}_n(x)[1 - \hat{F}_n(x)]\,\overline{\xi}_{\hat{F}_n}(x)$$

Let $\overline{\theta}_{\hat{F}_n}$ denote the function defined in (2.24), but with $\phi_{\hat{F}_n}$ replaced by $\overline{\phi}_{\hat{F}_n}$. Now corollary 2.1.1 says

$$\int \overline{\theta}_{\hat{F}_n}\, dQ_n = 0,$$

yielding

$$-\sqrt{n} \int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0} = \sqrt{n} \int \overline{\theta}_{\hat{F}_n}\, d(Q_n - Q_{F_0}) + \sqrt{n} \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n})\, dQ_{F_0}$$

The last term will be shown to be $o_p(1)$ in lemma 2.4.2. Note that the area of integration of Q_n can be taken to be $\{\tau_0 \le u \le v - \epsilon_0 \le \tau_M\}$ as well, since points (U_i, V_i) with $V_i - U_i < \epsilon_0$ do not occur.

Lemma 2.4.2

$$\sqrt{n} \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n})\, dQ_{F_0} = o_p(1)$$

Proof:
Let the function ψ_n be defined by

$$\begin{aligned} \psi_n(u,v) = &- [\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n}](u,v,1,0)\, F_0(u) \; - \; [\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n}](u,v,0,1)\, [F_0(v) - F_0(u)] \\ &+ [\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n}](u,v,0,0)\, [1 - F_0(v)]. \end{aligned}$$

Using the decomposition $\phi_F = F(1-F)\,\xi_F$, and

$$F(v) - F(u) = -[(1 - F(v)) - (1 - F(u))],$$

we get

$$\begin{aligned} \psi_n(u,v) \;=\; & \frac{1 - \hat{F}_n(u)}{\hat{F}_n(v) - \hat{F}_n(u)}\, [\overline{\xi}_{\hat{F}_n}(u) - \xi_{\hat{F}_n}(u)] \times \\ & \times [F_0(v)\,(\hat{F}_n(u) - F_0(u)) + F_0(u)\,(F_0(v) - \hat{F}_n(v))] \\ & - \frac{\hat{F}_n(v)}{\hat{F}_n(v) - \hat{F}_n(u)}\, (\overline{\xi}_{\hat{F}_n}(v) - \xi_{\hat{F}_n}(v)) \times \\ & \times [(1 - F_0)(v)\,(\hat{F}_n(u) - F_0(u)) + (1 - F_0)(u)\,(F_0(v) - \hat{F}_n(v))]. \end{aligned}$$

Applying Cauchy-Schwarz we obtain:

$$\begin{aligned}&\left|\sqrt{n}\int(\overline{\theta}_{\hat{F}_n}-\tilde{\theta}_{\hat{F}_n})\,dQ_{F_0}\right|\\ &\leq \sqrt{n}\,K\,\|\overline{\xi}_{\hat{F}_n}-\xi_{\hat{F}_n}\|_{H_1}\times\left[\|\hat{F}_n-F_0\|_{H_1}+\|\hat{F}_n-F_0\|_{H_2}\right]\\ &\qquad+\sqrt{n}\,K\,\|\overline{\xi}_{\hat{F}_n}-\xi_{\hat{F}_n}\|_{H_2}\times\left[\|\hat{F}_n-F_0\|_{H_1}+\|\hat{F}_n-F_0\|_{H_2}\right]\end{aligned}$$

By part I of lemma 1.4.2 on page 36 and $f_0 \geq c_1 > 0$ we find

$$|\overline{\xi}_{\hat{F}_n}(u)-\xi_{\hat{F}_n}(u)| \leq K\,|\hat{F}_n(u)-F_0(u)|. \tag{2.26}$$

An application of proposition 2.1.4 on page 65 finishes the proof. Property (2.26) is seen as follows.
For example, if the interval $J_i \ni u$ has a point s where $\hat{F}_n$ and F_0 have equal value, we have

$$\begin{aligned}|\overline{\xi}_{\hat{F}_n}(u)-\xi_{\hat{F}_n}(u)| = |\xi_{\hat{F}_n}(s)-\xi_{\hat{F}_n}(u)| &\leq K_1\,|s-u|\\ &\leq (K_1/C)\,|F_0(s)-F_0(u)|\\ &= (K_1/C)\,|\hat{F}_n(s)-F_0(u)|\\ &= (K_1/C)\,|\hat{F}_n(u)-F_0(u)|\end{aligned}$$

The same argument is used for the other two situations, with s replaced by τ_i or τ_{i+1}- and one =-sign replaced by a <-sign.

□

IV: Closeness in empirical process.

The first term is further split into

$$\begin{aligned}\sqrt{n}\int\overline{\theta}_{\hat{F}_n}d(Q_n-Q_{F_0}) &= \sqrt{n}\int\tilde{\theta}_{F_0}d(Q_n-Q_{F_0})\\ &\quad+\sqrt{n}\int(\overline{\theta}_{\hat{F}_n}-\tilde{\theta}_{F_0})d(Q_n-Q_{F_0}).\end{aligned}$$

Again the last term will be shown to be $o_p(1)$:

Lemma 2.4.3

$$\sqrt{n}\int(\overline{\theta}_{\hat{F}_n}-\tilde{\theta}_{F_0})d(Q_n-Q_{F_0}) = o_p(1)$$

Proof:
Consider the class of functions

$$\mathcal{K} = \{\overline{\theta}_F : F \in \mathcal{F}\} \cup \{\tilde{\theta}_{F_0}\}.$$

We show the class $\mathcal{K}$ to be a Q_{F_0}-Donsker class by showing the variation norm to be uniformly bounded.

The parts with $\delta = 1$ and $\delta = \gamma = 0$ are essentially one-dimensional. For example, for $\delta = 1$, we only have to consider the one-dimensional variation of

$$\overline{\theta}_F(u, v, 1, 0) = -(1 - F(u))\,\overline{\xi}_F(u)$$

From lemma 1.4.2 on page 36 we derive,

$$|\xi_F(y) - \xi_F(x)| \leq K(|y - x| + |F(y) - F(x)|).$$

From this one easily derives the variation of $\overline{\xi}_F$ to be bounded, uniformly over $F \in \mathcal{F}$. For the part with $\gamma = 1$, we have

$$f(x, y) = \frac{\overline{\phi}_F(y) - \overline{\phi}_F(x)}{F(y) - F(x)}.$$

Note that f is a function that is constant on rectangles of the form $[\tau_i, \tau_{i+1}) \times [\tau_j, \tau_{j+1})$, with τ_i and τ_j being points of jump of F. Let $A_{ij} = (s, x] \times (t, y]$. $f(A_{ij})$ can be rewritten as

$$f(A_{ij}) = \frac{N_1 + N_2 - N_3}{D}$$

with

$$\begin{aligned}
N_1 &= [F(t) - F(x)]\,[F(t) - F(s)] \times [F(x) - F(s)]\,[\overline{\phi}_F(y) - \overline{\phi}_F(t)] \\
N_2 &= [F(y) - F(x)]\,[F(t) - F(x)] \times [F(y) - F(t)]\,[\overline{\phi}_F(x) - \overline{\phi}_F(s)] \\
N_3 &= [F(y) + F(t) - F(x) - F(s)]\,[\overline{\phi}_F(t) - \overline{\phi}_F(x)] \times [F(y) - F(t)]\,[F(x) - F(s)] \\
D &= [F(y) - F(x)]\,[F(y) - F(s)]\,[F(t) - F(x)]\,[F(t) - F(s)]
\end{aligned}$$

The denominator D remains larger than $(c')^4$. For N_1, N_2 and N_3 only the parts after the $\times$-sign are important. Again using lemma 1.4.2, one obtains

$$|\phi_F(y) - \phi_F(x)| \leq K(|y - x| + |F(y) - F(x)|)$$

implying boundedness of $\|f\|_{V_2} = \sup_{A_{ij}} \sum_{i,j} |f(A_{ij})|$.

With respect to the one-dimensional variation for $\gamma = 1$ we have, for $x < t < y$,

$$f(x, y) - f(x, t) = \frac{[F(y) - F(x)]\,[\overline{\phi}_F(y) - \overline{\phi}_F(t)] - [\overline{\phi}_F(y) - \overline{\phi}_F(x)]\,[F(y) - F(t)]}{[F(y) - F(x)][F(t) - F(x)]}$$

implying boundedness of $\sup_x \|f(x, \cdot)\|_{V_1}$. Boundedness of variation in the other variable is shown in a similar way. The same arguments apply to the function $\tilde{\theta}_{F_0} \in \mathcal{K}$.

From the Q_{F_0}-Donsker class property for $\mathcal{K}$, we derive asymptotic uniform equicontinuity over $\mathcal{K}$, with respect to $L_2(Q_{F_0})$-norm, of the empirical process $\sqrt{n}(Q_n - Q_{F_0})$ (theorem 2.3.1). Finally we have to show

$$\|\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}\|_{Q_{F_0}} = o_p(1).$$

For this we use (2.26), i.e. $|\overline{\xi}_{\hat{F}_n}(u) - \xi_{\hat{F}_n}(u)| \leq K\,|\hat{F}_n(u) - F_0(u)|$, together with convergence of $\xi_{\hat{F}_n}$ to ξ_{F_0}, which is shown in lemma 2.4.4 below. Using L_2-consistency of $\hat{F}_n$ with respect to Lebesgue measure, we are done.

□

Lemma 2.4.4 *Let $\|\cdot\|_\lambda$ denote L_2-norm with respect to Lebesgue measure. Then*

$$\|F_n - F\|_\lambda \to 0 \qquad \textit{implies} \qquad \|\phi_{F_n} - \phi_F\|_\lambda \to 0.$$

The same holds for ξ_F.

Proof:
The following holds:

Let A_n: $X \to X$, $n = 1, 2, \ldots$ and A: $X \to X$ be compact linear operators on the normed space $(X, \|\cdot\|)$. Let ϕ_n be the solution to $(I - A_n)\phi_n = f_n$ and let ϕ satisfy $(I - A)\phi = f$. Then we can write

$$\phi - \phi_n = \left[(I-A)^{-1}\right](f - f_n) - \left[(I-A)^{-1}(A_n - A)\right]\phi_n$$

We apply this, with $X = [\tau_0, \tau_M]$ and the $L_2(\lambda)$-norm and A being our integral operator. Boundedness of $(I - A)^{-1}$ (theorem 1.4.1) and uniform boundedness of $\{\phi_{F_n}\}$ is used, together with the following inequalities:

$$\|f_{\hat{F}_n} - f_{F_0}\|_\lambda \le K_1 \|\hat{F}_n - F_0\|_\lambda.$$

and

$$\|A_{\hat{F}_n} - A\|_\lambda \le K_2 \|\hat{F}_n - F_0\|_\lambda.$$

These inequalities can be proved by repeatedly using

$$\frac{a_1}{b_1} - \frac{a_2}{b_2} = \frac{1}{b_1}(a_1 - a_2) + \frac{a_2}{b_1 b_2}(b_2 - b_1).$$

□

2.5 Case 2B: observation times arbitrarily close

Again, the following theorem will be shown to be valid.

Theorem 2.5.1 *Let the conditions on page 63 be satisfied, except for (H4).Case A. Then we have*

$$\sqrt{n}\,[\,K(\hat{F}_n) - K(F_0)\,] \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{F_0}\|^2_{Q_{F_0}}) \quad \textit{as} \quad n \to \infty$$

Proof:
We again go through the successive steps. To prove that the functional is almost linear (part I), we refer to case 2A.

II: Transformation to observation space.

Like in the previous cases, our definition of the canonical gradient $\tilde{\theta}_F$ will be extended to piecewise constant distribution functions F with finitely many discontinuities. However, since $F(v) - F(u)$ no longer remains bounded away from zero on the region where H has mass, the situation is quite different from case 2A. One may guess what will happen from the following observations.

On one hand, the quotient

$$\frac{\phi_F(v) - \phi_F(u)}{F(v) - F(u)},$$

for u and v in the same interval of constancy of F, can only be defined correctly if ϕ_F is constant on the same interval. On the other hand, d_F, h and $\tilde{\kappa}'_{F_0}$ in general are not constant on these intervals, making a completely discrete version of the integral equation impossible. The integral equation for discrete F is a compromise between these two conflicting demands.

Instead of *one* function ϕ_F we have a *pair* of functions (ϕ_F, ψ_F), satisfying

$$\phi_F(x) = d_F(x)\Big\{k(x) - \int_{\tau_0}^{x} r_F(t,x)\,h(t,x)\,dt + \int_{x}^{\tau_M} r_F(x,t)\,h(x,t)\,dt\Big\}, \tag{2.27}$$

where $r_F(u,v)$, for $F(u) < F(v)$, is defined by

$$r_F(u,v) = \frac{\phi_F(v) - \phi_F(u)}{F(v) - F(u)}. \tag{2.28}$$

If u and v are on the same interval of constancy of F, we have some freedom defining r_F. Two versions will be considered. First we use

$$r_F(u,v) = \psi_F(v) - \psi_F(u) \qquad \text{on } \{(u,v) \mid F(u) = F(v)\}, \tag{2.29}$$

for which it is rather easily shown that (2.27) has a solution, nicely using theory on Fredholm integral equations. This choice has the disadvantage that $\{\psi_F \mid F \in \mathcal{F}\}$ is not uniformly Lipschitz, as will be shown in section 2.6. Moreover, the way of proof given does not show $\bar{\phi}_F$ to be uniformly Lipschitz in the inverse scale. These uniform Lipschitz properties will be used when showing the Donsker property for $\{\tilde{\theta}_F \mid F \in \mathcal{F}\}$. Another version is

$$r_F(u,v) = \frac{\psi_F(v) - \psi_F(u)}{F_0(v) - F_0(u)} \qquad \text{on } \{(u,v) \mid F(u) = F(v)\}, \tag{2.30}$$

for which ψ_F will be shown to be uniformly Lipschitz. The Fredholm technique used to prove version (2.29) to be a valid one cannot be used for version (2.30), due to the fact that a singularity is introduced via the quotient. A different approach showing that (2.28) and (2.30) lead to a solvable equation (2.27) will be given in theorem 2.5.2, in which the uniform Lipschitz property for both $\{\bar{\phi}_F \mid F \in \mathcal{F}\}$ on the inverse scale and $\{\psi_F \mid F \in \mathcal{F}\}$ is shown as well.

The definition of the function $\tilde{\theta}_F$ is extended to piecewise constant distribution functions F by defining, for the pair (ϕ_F, ψ_F) solving equation (2.27),

$$\tilde{\theta}_F(u,v,\delta,\gamma) = -\delta\,\frac{\phi_F(u)}{F(u)} - \gamma\, r_F(u,v) + (1-\delta-\gamma)\,\frac{\phi_F(v)}{1-F(v)},$$

where $\phi_F(u)/F(u)$ and $\phi_F(v)/(1-F(v))$ are defined to be zero if $F(u)=0$ or if $F(v)=1$, respectively.

Since ϕ_F is constant, the only real integral part is the ψ-part; the rest of the integral can be written as a summation. As in chapter 1, we let $z_i = F(\tau_i)$; moreover we let $y_i = \phi_F(\tau_i)$. We define

$$\Delta_i(g) := \int_{\tau_i}^{\tau_{i+1}} g(t)\,dt, \tag{2.31}$$

$$\Delta_{ij}(h) := \int_{u=\tau_i}^{\tau_{i+1}} \int_{v=\tau_j}^{\tau_{j+1}} h(u,v)\,dv\,du, \quad \text{for } i<j \tag{2.32}$$

and

$$\tilde{d}_i := \frac{z_i(1-z_i)}{\Delta_i(h_1)\,(1-z_i) + \Delta_i(h_2)\,z_i}. \tag{2.33}$$

We now start with the first choice (2.29). For the following proposition, condition (H4).CaseB can be slightly weakened.

Using the Fredholm theorem 1.4.2 we obtain

Proposition 2.5.1 *Let F be a piecewise constant distribution function, having a finite number of jumps. Instead of (H4).CaseB, the following is supposed to hold:*

$$\int_{u=x}^{t} h(u,t)\,du + \int_{v=t}^{y} h(t,v)\,dv > 0 \text{ for each } x<t<y \tag{2.34}$$

Then a pair of functions (ϕ_F, ψ_F), solving equation (2.27), exists, with r_F defined by (2.28) and (2.29). ϕ_F is a piecewise constant function, constant on intervals of constancy of F.

The vector $y = (y_1, \ldots, y_m)'$, with $y_i = \phi_F(\tau_i)$, is the unique solution of the set of linear equations

$$\begin{aligned} &y_i\Big\{\tilde{d}_i^{-1} + \sum_{j<i} \tfrac{\Delta_{ji}(h)}{z_i - z_j} + \sum_{j>i} \tfrac{\Delta_{ij}(h)}{z_j - z_i}\Big\} \\ &= \Delta_i(k) + \sum_{j<i} \tfrac{\Delta_{ji}(h)}{z_i - z_j}\, y_j + \sum_{j>i} \tfrac{\Delta_{ij}(h)}{z_j - z_i}\, y_j\,, \quad i = 1, \ldots, m-1. \end{aligned} \tag{2.35}$$

Proof: Define

$$h^*(t,x) := \begin{cases} h(t,x) & \text{if } t \le x \\ h(x,t) & \text{if } t \ge x \end{cases}$$

Split $[\tau_0, \tau_M]$ into the intervals $J_i = [\tau_i, \tau_{i+1})$, $i = 0, \ldots, m-1$ and $J_m = [\tau_m, \tau_M]$. We have $\phi_F(J_0) = y_0 = 0$ and $\phi_F(J_m) = y_m = y_M = 0$. On these intervals we can choose $\psi_F(x) = 0$.

Assume ϕ_F to be constant on the same intervals as F. Let $x \in J_i$ for some index $i \in \{1, 2, \ldots, m-1\}$. Rewriting equation (2.27) as an integral equation in ψ on $[\tau_i, \tau_{i+1}]$, we obtain:

$$\psi_F(x) \int_{t\in J_i} h^*(t,x)\,dt - \int_{t\in J_i} h^*(t,x)\,\psi_F(t)\,dt = r(x), \tag{2.36}$$

with

$$\begin{aligned} r(x) &= k(x) - \frac{y_i}{z_i(1-z_i)}\left[h_1(x)(1-z_i) + h_2(x)z_i\right] \\ &\quad -\sum_{j=0}^{i-1} \tfrac{y_i - y_j}{z_i - z_j}\Delta_j(h(\cdot,x)) + \sum_{j=i+1}^{m} \tfrac{y_j - y_i}{z_j - z_i}\Delta_j(h(x,\cdot)). \end{aligned} \tag{2.37}$$

The homogeneous equation corresponding to (2.36) is solved by the constant functions. So we are in the situation of theorem 1.4.2 on page 23. Part I of this theorem is applied. Defining

$$\chi_F(x) := \psi_F(x)\int_{J_i} h^*(t,x)\,dt,$$

we look at the adjoint homogeneous equation of the χ_F-equation, which is given by

$$\sigma(x)\int_{J_i} h^*(t,x)\,dt - \int_{J_i} h^*(t,x)\,\sigma(t)\,dt = 0. \tag{2.38}$$

Note that this is the same equation as the homogeneous part of (2.36). By a supremum argument as in theorem 1.4.6 on page 33, using (2.34), one can show that *only* the constant functions solve this homogeneous adjoint equation. Thus equation (2.36) is solvable if and only if

$$\int_{J_i} C\,r(x)\,dx = 0 \qquad \text{for all } C \in \mathbb{R}.$$

This is equivalent to the condition

$$\frac{y_i}{z_i(1-z_i)}\left[\Delta_i(h_1)(1-z_i) + \Delta_i(h_2)\,z_i\right] =$$

$$\Delta_i(k) - \sum_{j=0}^{i-1} \tfrac{y_i - y_j}{z_i - z_j}\Delta_{ji}(h) + \sum_{j=i+1}^{m} \tfrac{y_j - y_i}{z_j - z_i}\Delta_{ij}(h) \tag{2.39}$$

Note that this should hold for any interval J_i, so the *set* of linear equations $Ay = b$ should be solvable, with:

$$\begin{aligned} \alpha_{ii} &= \tilde{d}_i^{-1} + \textstyle\sum_{j<i} \frac{\Delta_{ji}(h)}{z_i - z_j} + \sum_{j>i} \frac{\Delta_{ij}(h)}{z_j - z_i} \\ \alpha_{ij} &= -\tfrac{\Delta_{ij}(h)}{z_i - z_j} \\ b_i &= \Delta_i(k) \qquad \text{for } i,\, j = 1, \ldots, m-1 \end{aligned}$$

Again we can use theorem 1.4.6 on page 33 to show that this equation has a unique solution.

Since $y_i = \phi_F(J_i)$, this solution specifies ϕ_F. Moreover, this specification is not in conflict with the integral equation (2.27). (2.27) is the same as (2.36), and integrating both sides of (2.36) from τ_i to τ_{i+1} and applying Fubini's theorem cancels the left-hand side. Hence we have shown the existence of a solving pair (ϕ_F, ψ_F).

Note that ψ_F is only determined up to a constant, since only the *difference* between two values of ψ_F occurs in equation (2.27). (This also follows from the Fredholm theory: $\dim(\mathcal{N}(I-A)) = \dim(\mathcal{N}(I-A^*)) = 1$.) We can ensure uniqueness by defining $\psi_F(\tau_i) = 0$.

□

Remarks.

I. The same method can be used to prove existence of a comparable (ξ, ζ) equation. Let $\xi(\tau_i) = w_i$. Then the matrix equation $Aw = b$ has coefficients:

$$\begin{aligned}\alpha_{ii} &= 1 + \sum_{j<i} \tilde{c}_i \tfrac{\Delta_{ji}(h)}{z_i - z_j} z_j(1-z_j) + \sum_{j>i} \tilde{c}_i \tfrac{\Delta_{ij}(h)}{z_j - z_i} z_j(1-z_j) \\ \alpha_{ij} &= -\tilde{c}_i \tfrac{\Delta_{ij}(h)}{z_i - z_j} z_j(1-z_j) \\ b_i &= \tilde{c}_i \Delta_i(k) \qquad \text{for } i,\, j = 1, \dots, m-1,\end{aligned}$$

where $\tilde{c}_i$ is defined by

$$1/\tilde{c}_i := \sum_{j=0}^{i-1} [(1-z_j)\Delta_{ji}(h)] + \int_{u=\tau_i}^{\tau_{i+1}} \int_{v=u}^{\tau_{i+1}} h(u,v)\, dv\, du + \sum_{j=i}^{m} z_j \Delta_{ij}(h).$$

II. We may have a situation in which no mass is present along part of the diagonal. If, for each $t \in [\tau_k, \tau_{k+1}]$,

$$\int_{u=\tau_k}^{t} h(u,t)\, du + \int_{v=t}^{\tau_{k+1}} h(t,v)\, dv = 0,$$

equation 2.38 is solved by any function σ, hence the approach of lemma 2.5.1 fails. This is to be expected, since this is situation 2A, in which ϕ_F cannot be constant.

However, if, for each $t \in [\tau_l, \tau_{l+1}]$,

$$\int_{u=x}^{t} h(u,t)\, du + \int_{v=t}^{y} h(t,v)\, dv > 0, \text{ for each } x < t < y,$$

lemma 2.5.1 can still be applied on this part of $[\tau_0, \tau_M]$. $r(x)$ has a form which is slightly different from the full case 2B. If $[\tau_k, \tau_{k+1}]$ belongs to a section of the diagonal where mass is absent, ϕ_F is no longer constant on these intervals, implying that e.g.

$$\frac{y_l - y_k}{z_l - z_k} \Delta_{kl}(h)$$

should be replaced by

$$\int_{\tau_l}^{\tau_{l+1}} \int_{\tau_k}^{\tau_{k+1}} \tfrac{y_l - \phi_F(t)}{z_l - z_k} h(t,x)\, dt\, dx.$$

Of course, an interval $[\tau_k, \tau_{k+1}]$ may also partly belong to a part of the diagonal where mass is present and partly to a part were mass is absent. Then, ϕ_F should be constant on the part where mass is present, with a ψ-function needed in the integral equation for this section of $[\tau_k, \tau_{k+1}]$. On the rest, ϕ_F is not constant, and no compensating ψ is needed there.

For the second choice of r_F we use a representation of the equation for ϕ_F on the inverse scale and the construction of a continuous extension of the equation for ϕ_F on this inverse scale (similar techniques were used in subsection 1.4.2). Using a similar notation, we denote by G the inverse of F, where, for purely discrete distribution functions F, we take the right-continuous version of the inverse, defined by

$$G(y) = \inf\{x \in [\tau_0, \tau_M] : F(x) > y\}, 0 \le y \le 1.$$

Similarly, we define

$$\bar{k}_F = k \circ G,\ \bar{h}_{1,F} = h_1 \circ G,\ \bar{h}_{2,F} = h_2 \circ G,$$
$$\bar{d}_F(y) = \frac{y(1-y)}{(1-y)\bar{h}_{1,F}(y) + y\,\bar{h}_{2,F}(y)},$$
$$\bar{H}(u,v) = H(G(u),G(v)),\ 0 \le u \le v \le 1,$$

and

$$\bar{c}_F(y) = \int_0^y (1-s)\,d\bar{H}(s,y) + \int_y^1 s\,d\bar{H}(y,s).$$

We again have to restrict ourselves to the class

$$\mathcal{F} := \{F : F \text{ is non-defective and satisfies (CF) on page 32}\},$$

which was also used incase 2A. The choice of ϵ in (CF) is given in the proof of the theorem below. We have

Theorem 2.5.2 *The following holds:*

(i) There exists a unique Lipschitz function $\bar{\phi}_F : [0,1] \to I\!R$ such that, for $y \in [0,1] \setminus D$,

$$\bar{\phi}_F(y) = \bar{d}_F(y)\Big\{\bar{k}_F(y) - \int_{s\in[0,y)} \frac{\bar{\phi}_F(y)-\bar{\phi}_F(s)}{y-s}\,d\bar{H}(s,y) + \int_{s\in(y,1]} \frac{\bar{\phi}_F(s)-\bar{\phi}_F(y)}{s-y}\,d\bar{H}(y,s)\Big\}, \quad (2.40)$$

where D is the (finite) set of discontinuities of the right-continuous inverse $G = F^{-1}$ in $(0,1)$, augmented with 0 and 1. The function $\bar{\phi}_F$ is Lipschitz, uniformly for $F \in \mathcal{F}$.

(ii) There exists a pair (ϕ_F, ψ_F), solving the integral equation (2.27), with r_F defined by (2.28) and (2.30). ϕ_F is absolutely continuous with respect to F and the function ψ_F is Lipschitz on each interval between jumps of F, uniformly for $F \in \mathcal{F}$, with a Lipschitz norm not depending on the interval.

(iii) The vector $y = (y_1, \ldots, y_m)'$, with $y_i = \phi_F(\tau_i)$, is the unique solution of the set of linear equations

$$y_i\Big\{\tilde{d}_i^{-1} + \sum_{j<i} \frac{\Delta_{ji}(h)}{z_i - z_j} + \sum_{j>i} \frac{\Delta_{ij}(h)}{z_j - z_i}\Big\} = \Delta_i(k) + \sum_{j<i} \frac{\Delta_{ji}(h)}{z_i - z_j}\,y_j + \sum_{j>i} \frac{\Delta_{ij}(h)}{z_j - z_i}\,y_j\,, \quad i = 1, \ldots, m. \quad (2.41)$$

Theorem 2.5.2 will be proved by approximating the purely discrete distribution function F by the function $F_\alpha = (1-\alpha)F_0 + \alpha F$, which was considered in section 1.4.2, and by studying the behaviour of the corresponding function ϕ_{F_α}, as $\alpha \uparrow 1$.

Proof:
ad (i) Let $F_\alpha = (1-\alpha)F_0 + \alpha F$. For $\alpha \in [0,1)$, the function F_α is strictly increasing and continuous between jumps and hence the solution ϕ_{F_α} to the integral equation exists by theorem 1.4.8 in section 1.4.2. For simplicity of notation, we will denote ϕ_{F_α} by ϕ_α. Moreover we let $G_\alpha = F_\alpha^{-1}$, with derivative g_α. Furthermore, we write $\bar{k}_\alpha$ instead of $\bar{k}_{F_\alpha}$, and use the same notation for the other functions in the inverse scale. By theorem 1.4.8, $\bar\phi_\alpha$ is the unique solution of the integral equation

$$\bar\phi_\alpha(y) = \bar d_\alpha(y)\Big\{\bar k_\alpha(y) - \int_{s\in[0,y)} \tfrac{\bar\phi_\alpha(y)-\bar\phi_\alpha(s)}{y-s}\, d\bar H_\alpha(s,y) + \int_{s\in(y,1]} \tfrac{\bar\phi_\alpha(s)-\bar\phi_\alpha(y)}{s-y}\, d\bar H_\alpha(y,s)\Big\},\ y\in[0,1].$$

Let the set D_α be defined by

$$\begin{aligned} D_\alpha &= \{\text{discontinuity points of } g_\alpha(y), \text{ augmented with 0 and 1}\} \\ &\quad \cup \{\text{discontinuity points of } \bar k'_\alpha(y),\ \bar d'_\alpha(y), \\ &\qquad \Delta^1(y) = \tfrac{\partial}{\partial y}\bar h_\alpha(y,s) \text{ for } y \le s, \text{ and } \Delta^2(y) = \tfrac{\partial}{\partial y}\bar h_\alpha(s,y) \text{ for } y \ge s\}, \end{aligned}$$

and let $A_{\alpha,\delta}(y)$ and $B_{\alpha,\delta}(y)$ be defined by

$$A_{\alpha,\delta}(y) = \bar d_\alpha(y)\left\{\int_{y-\delta}^{y} |\tfrac{\partial}{\partial y}\bar h_\alpha(s,y)|\, dG_\alpha(s) + \int_y^{y+\delta} |\tfrac{\partial}{\partial y}\bar h_\alpha(y,s)|\, dG_\alpha(s)\right\}, \tag{2.42}$$

and

$$\begin{aligned} B_{\alpha,\delta}(y) &= \bar d_\alpha(y)|\bar k'_\alpha(y)| + |\bar d'_\alpha(y)|\,[(1-y)\bar h_{1,\alpha}(y) + y\,\bar h_{2,\alpha}(y)] \sup_{s\in[0,1]}\{\bar c_\alpha(s)|\bar k_\alpha(s)|\} \\ &\quad + \frac{2\bar d_\alpha(y)}{\delta} \sup_{s\in[0,1]}\{\bar d_\alpha(s)|\bar k_\alpha(s)|\} \times \\ &\qquad \times \Big\{\sup_{s\in[0,y]} |\tfrac{\partial}{\partial y}\bar h_\alpha(s,y)| + \sup_{s\in[y,1]} |\tfrac{\partial}{\partial y}\bar h_\alpha(y,s)|\Big\}. \end{aligned} \tag{2.43}$$

Moreover let

$$C_\alpha(y) = 1 + 2\bar d_\alpha(y) g_\alpha(y) \bar h_\alpha(y,y). \tag{2.44}$$

As in theorem 1.4.8 we have that at points of D_α the functions $A_{\alpha,\delta}$ and $B_{\alpha,\delta}$ have two versions, one corresponding to taking left derivatives and one corresponding to taking right derivatives. By theorem 1.4.8, there exists a $\delta > 0$ such that

$$\sup_{s\in[0,1]} A_{\alpha,\delta}(s)/C_\alpha(s) \le 1/2,$$

and we have

$$|\bar{\phi}_\alpha(v) - \bar{\phi}_\alpha(u)| \le K_\alpha(v-u),\ 0 \le u < v \le 1,$$

where K_α is given by

$$K_\alpha = 2 \sup_{s\in[0,1]} B_{\alpha,\delta}(s)/C_\alpha(s), \tag{2.45}$$

for $\delta > 0$ such that $\sup_{s\in[0,1]} A_{\alpha,\delta}(s)/C_\alpha(s) \le 1/2$. We need to show that we can choose δ and K_α independently of α and F in a small (supremum distance) neighbourhood of F_0.

If (using the same notation as in the proof of lemma 1.4.4) y belongs to an interval (z_i, z_i'), on which G_α increases, then, going back to the original scale, we get

$$A_{\alpha,\delta}(y)/C_\alpha(y)$$
$$\le \sup_{x\in(\tau_0,\tau_M)} \frac{\int_{t:F_\alpha(x)-\delta<F_\alpha(t)<F_\alpha(x)} |\frac{\partial}{\partial x}h(t,x)|\,dt + \int_{t:F_\alpha(x)<F_\alpha(t)<F_\alpha(x)+\delta} |\frac{\partial}{\partial x}h(x,t)|\,dt}{2h(x,x)}$$

The essential observation here is that, although $A_{\alpha,\delta}(y)$ tends to ∞, as $\alpha \uparrow 1$, for points y in the range of F_α, the ratio $A_{\alpha,\delta}(y)/C_\alpha(y)$ stays bounded, since the factor $g_\alpha(y)$, causing the steep increase of $A_{\alpha,\delta}(y)$ via $\frac{\partial}{\partial y}\bar{h}_\alpha$, also occurs in the denominator $C_\alpha(y)$.

If, on the other hand, y belongs to an interval (z_i', z_{i+1}), on which G_α is constant, then $A_{\alpha,\delta}(y) = 0$, since $g_\alpha(y) = 0$ on such an interval. Hence we can choose $\delta > 0$ such that

$$\sup_{s\in[0,1]} A_{\alpha,\delta}(s)/C_\alpha(s) \le 1/2,$$

for all $\alpha \in [0,1)$ and all F such that $\sup_{x\in[\tau_0,\tau_M]} |F(x) - F_0(x)| \le \epsilon$, for a fixed suitably chosen $\epsilon > 0$. (Note that here the ϵ in condition (CF) on page 32 is determined.)

In a similar way we get, using (1.50) in section 1.4.2, if F is close enough to F_0,

$$
\begin{aligned}
&B_{\alpha,\delta}(y)/C_\alpha(y)\\
&\le \sup_{x\in(\tau_0,\tau_M)} \frac{|k'(x)|}{2h(x,x)}\\
&+ \sup_{x\in(\tau_0,\tau_M)} \frac{1+|h_1(x)-h_2(x)|+|h_1'(x)|+|h_2'(x)|}{h(x,x)} \sup_{x\in(\tau_0,\tau_M)} \frac{k(x)}{\epsilon_1\, h(x,x)}\\
&+ \frac{1}{\delta \inf_{x\in(\tau_0,\tau_M)} h(x,x)} \sup_{x\in(\tau_0,\tau_M)} \frac{2k(x)}{h_1(x)+h_2(x)} \times\\
&\qquad \times \Big\{ \sup_{x,y\in(\tau_0,\tau_M)} |\tfrac{\partial}{\partial x}h(x,y)| + \sup_{x,y\in(\tau_0,\tau_M)} |\tfrac{\partial}{\partial y}h(x,y)| \Big\},
\end{aligned}
\tag{2.46}$$

for some $\epsilon_1 > 0$, uniform over α and F, implying that $\sup_{s\in[0,1]} B_{\alpha,\delta}(s)/C_\alpha(s)$ and hence also K_α in (2.45) has a finite upper bound (given by the right-hand side of (2.46)) which is independent of α and F.

It follows that the sequence $(\bar{\phi}_\alpha)$ is equicontinuous and hence has a subsequence, converging to a function $\bar{\phi}_F$ which is Lipschitz on $[0,1]$. Let $(\alpha_n)_{n=1,2,\dots}$ be a sequence of numbers such that $\alpha_n \uparrow 1$ and $\bar{\phi}_{\alpha_n} \to \bar{\phi}_F$ in the supremum distance. Define

$$\phi_F(x) := \bar{\phi}_F(F(x)). \tag{2.47}$$

Then, by the equicontinuity of the sequence $(\bar{\phi}_{\alpha_n})$, we obtain, for each $x \in [\tau_0, \tau_M]$,

$$\begin{aligned}\phi_F(x) = \bar{\phi}_F(F(x)) &= \lim_{n\to\infty} \bar{\phi}_{\alpha_n}(F_{\alpha_n}(x)) \\ &= \lim_{n\to\infty} \phi_{\alpha_n}(x). \end{aligned} \tag{2.48}$$

Now let $y \in [0,1] \setminus D$. Then y is a point of continuity of G and does not belong to the range of F. We have

$$\begin{aligned}\bar{\phi}_F(y) &= \lim_{n\to\infty} \bar{\phi}_{\alpha_n}(y) \\ &= \lim_{n\to\infty} \bar{d}_{\alpha_n}(y)\Big\{\bar{k}_{\alpha_n}(y) - \int_{s\in[0,y)} \tfrac{\bar{\phi}_{\alpha_n}(y)-\bar{\phi}_{\alpha_n}(s)}{y-s}\, d\bar{H}_{\alpha_n}(s,y) \\ &\qquad + \int_{s\in(y,1]} \tfrac{\bar{\phi}_{\alpha_n}(s)-\bar{\phi}_{\alpha_n}(y)}{s-y}\, d\bar{H}_{\alpha_n}(y,s)\Big\} \\ &= \bar{d}_F(y)\Big\{\bar{k}_F(y) - \lim_{n\to\infty}\int_{s\in[0,y)} \tfrac{\bar{\phi}_F(y)-\bar{\phi}_{\alpha_n}(s)}{y-s}\, d\bar{H}_{\alpha_n}(s,y) \\ &\qquad + \lim_{n\to\infty}\int_{s\in(y,1]} \tfrac{\bar{\phi}_{\alpha_n}(s)-\bar{\phi}_F(y)}{s-y}\, d\bar{H}_{\alpha_n}(y,s)\Big\}. \end{aligned} \tag{2.49}$$

Suppose $F(\tau_i) < y < F(\tau_{i+1})$. Hence $G(y) = \tau_{i+1}$. Then (2.49) can be written as

$$\begin{aligned}\bar{\phi}_F(y) &= \frac{y(1-y)}{h_1(\tau_{i+1})(1-y) + h_2(\tau_{i+1})y} \times \\ &\quad \times\Big\{k(\tau_{i+1}) - \lim_{n\to\infty}\int_{t\in[\tau_0, G_{\alpha_n}(y))} \tfrac{\bar{\phi}_F(y)-\phi_{\alpha_n}(t)}{y-F_{\alpha_n}(t)}\, dH(t, G_{\alpha_n}(y)) \\ &\qquad + \lim_{n\to\infty}\int_{t\in(G_{\alpha_n}(y),\tau_M]} \tfrac{\phi_{\alpha_n}(t)-\bar{\phi}_F(y)}{F_{\alpha_n}(t)-y}\, dH(G_{\alpha_n}(y), t)\Big\} \end{aligned} \tag{2.50}$$

and by the dominated convergence theorem and (2.48), we get

$$\begin{aligned}\bar{\phi}_F(y) &= \frac{y(1-y)}{h_1(\tau_{i+1})(1-y) + h_2(\tau_{i+1})y} \times \\ &\quad \times\Big\{k(\tau_{i+1}) - \int_{t:F(t)<y} \tfrac{\bar{\phi}_F(y)-\phi_F(t)}{y-F(t)}\, dH(t, \tau_{i+1}) \\ &\qquad + \int_{t:F(t)>y} \tfrac{\phi_F(t)-\bar{\phi}_F(y)}{F(t)-y}\, dH(\tau_{i+1}, t)\Big\} \\ &= \frac{y(1-y)}{\bar{h}_1(y)(1-y) + \bar{h}_2(y)y} \times \\ &\quad \times\Big\{\bar{k}_F(y) - \int_{s\in[0,y)} \tfrac{\bar{\phi}_F(y)-\bar{\phi}_F(s)}{y-s}\, d\bar{H}(s,y) + \int_{s\in(y,1]} \tfrac{\bar{\phi}_F(s)-\bar{\phi}_F(y)}{s-y}\, d\bar{H}(y,s)\Big\}. \end{aligned} \tag{2.51}$$

Uniqueness of $\bar{\phi}_F$ will be proved below in part (iii).

ad (ii) We define, for $\alpha \in (0,1)$, the functions $\psi_\alpha : [\tau_0, \tau_M] \to I\!R$ by

$$\psi_\alpha(x) = \frac{\phi_\alpha(x) - \phi_\alpha(\tau_i)}{1-\alpha},\ x \in J_i,\ i = 0, \dots, m. \tag{2.52}$$

Recall that $J_i = [\tau_i, \tau_{i+1})$ and $J_m = [\tau_m, \tau_M]$. Using the uniform Lipschitz property of $\{\bar\phi_\alpha\}$, we get, for x, y in the same interval J_i,

$$\begin{aligned} |\psi_\alpha(y) - \psi_\alpha(x)| &= \frac{|\phi_\alpha(y) - \phi_\alpha(x)|}{1-\alpha} = \frac{|\bar\phi_\alpha(F_\alpha(y)) - \bar\phi_\alpha(F_\alpha(x))|}{1-\alpha} \\ &\le c\frac{|F_\alpha(y) - F_\alpha(x)|}{1-\alpha} = c|F_0(y) - F_0(x)|, \end{aligned} \tag{2.53}$$

where $c > 0$ is independent of α and $F \in \mathcal{F}$. By the continuity of $\bar\phi_\alpha$, we can extend the function ψ_α, restricted to an interval $[\tau_i, \tau_{i+1})$ to a continuous function $\psi_{\alpha,i}$, defined on the closed interval $[\tau_i, \tau_{i+1}]$. The functions $\psi_{\alpha,i}$ are equicontinuous in α on the intervals $[\tau_i, \tau_{i+1}]$ and hence have a convergent subsequence, converging (in the supremum metric for functions defined on $[\tau_i, \tau_{i+1}]$) to a continuous function $\tilde\psi_i$, defined on $[\tau_i, \tau_{i+1}]$. Let $\psi_F : [\tau_0, \tau_M] \to I\!R$ be the function, such that

$$\psi_F(x) = \tilde\psi_i(x),\ x \in [\tau_i, \tau_{i+1}),\ i = 0, \dots, m,\ \psi_F(\tau_M) = \tilde\psi_m(\tau_M),$$

and let $(\psi_{\alpha_n})_{n=1,2,\dots}$ be a sequence such that the restriction of ψ_{α_n} to $[\tau_i, \tau_{i+1})$, $i = 0, \dots, m-1$, or to $[\tau_m, \tau_M]$ converges to ψ_F in the supremum metric for continuous functions on such an interval. Since the sequence $(\bar\phi_{\alpha_n})$ is also equicontinuous, we can also assume (by switching to a further subsequence) that $\bar\phi_{\alpha_n}$ converges in the supremum metric to a Lipschitz function $\bar\phi_F$, as in part (i). Then we have

$$\begin{aligned} \frac{\psi_F(y) - \psi_F(x)}{F_0(y) - F_0(x)} &= \lim_{n\to\infty} \frac{\phi_{\alpha_n}(y) - \phi_{\alpha_n}(x)}{(1-\alpha_n)\{F_0(y) - F_0(x)\}} \\ &= \lim_{n\to\infty} \frac{\phi_{\alpha_n}(y) - \phi_{\alpha_n}(x)}{F_{\alpha_n}(y) - F_{\alpha_n}(x)}, \end{aligned} \tag{2.54}$$

for $\tau_i \le x < y < \tau_{i+1}$, $i = 0, \dots, m$. Since, by theorem 1.4.8, part (iii), on page 42, ϕ_α satisfies the integral equation

$$\phi_\alpha(x) = d_\alpha(x)\Big\{k(x) - \int_{t\in[\tau_0,x)} \tfrac{\phi_\alpha(x)-\phi_\alpha(t)}{F_\alpha(x)-F_\alpha(t)}\, dH(t,x) + \int_{t\in(x,\tau_M]} \tfrac{\phi_\alpha(t)-\phi_\alpha(x)}{F_\alpha(t)-F_\alpha(x)}\, dH(x,t)\Big\},$$

we now get, by (2.48), (2.54) and the dominated convergence theorem,

$$\phi_F(x) = d_F(x)\Big\{k(x) - \int_{t\in[\tau_0,x)} r_F(t,x)\, dH(t,x) + \int_{t\in(x,\tau_M]} r_F(x,t)\, dH(x,t)\Big\}, \tag{2.55}$$

where r_F is defined by (2.28) and (2.30).

The function ϕ_F is absolutely continuous with respect to F, since, by the Lipschitz property of $\bar\phi_F$,

$$\begin{aligned} |\phi_F(y) - \phi_F(x)| &= |\bar\phi_F(F(y)) - \bar\phi_F(F(x))| \\ &\le c|F(y) - F(x)|, \qquad x, y \in [\tau_0, \tau_M]. \end{aligned}$$

This shows in particular that constancy of F on an interval implies constancy of ϕ_F on that same interval. Moreover, by (2.53) and the bounded differentiability of F_0, we have that ψ_α is Lipschitz on each interval $[\tau_i, \tau_{i+1})$, and hence ψ_F is also Lipschitz on such an interval.

ad (iii) Multiplying both sides of (2.55) by $d_F(x)^{-1}$, and integrating from τ_i to τ_{i+1}, the ψ_F-part cancels and we get a finite set of linear equations $Ay = b$ for $y_i = \phi_F(\tau_i)$, given by (2.41). This matrix equation was already shown to have a unique solution in the proof of proposition 2.5.1, using theorem 1.4.6 on page 33. The unicity of $\bar{\phi}_F$ is easily obtained from this, since the integral parts of the equation for $\bar{\phi}_F$ are with respect to a measure that has mass restricted to the values $z_i = F(\tau_i)$.

□

Remark. The proof of theorem 2.5.2 crucially uses $h(x, x) > 0$, whereas case 2A assumes $h(u, v)$ to be zero in a neighbourhood of the diagonal. For the situation in-between, a uniform Lipschitz property has not been established yet.

Now we have for both versions of r_F, similarly to case 2A:

Lemma 2.5.1

$$\int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) = -\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0}$$

Proof: The proof is similar to the proof of lemma 2.4.1 on page 74. The basic properties needed in the proof were given in the remark following lemma 2.4.1. $\tilde{\theta}_F \in L_2(Q_{F_0})$ follows from boundedness of the ratios occurring in $\tilde{\theta}_F$. $\int \tilde{\theta}_F dQ_F = 0$ is easily shown to hold by writing out the definition of $\tilde{\theta}_F$.

□

III: Inserting of empirical measure.

Unlike case 2A, $\phi_{\hat{F}_n}$ and $\xi_{\hat{F}_n}$ are constant on the same intervals as $\hat{F}_n$. $\psi_{\hat{F}_n}$ is not constant. However, we always have $\gamma = 0$ if $\hat{F}_n(v) = \hat{F}_n(u)$. So proposition 2.1.1 can be used directly to obtain

$$\int \tilde{\theta}_{\hat{F}_n}\, dQ_n = 0,$$

yielding

$$-\sqrt{n}\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0} = \sqrt{n}\int \tilde{\theta}_{\hat{F}_n}\, d(Q_n - Q_{F_0}).$$

IV: Closeness in empirical process.

Again we write

$$\begin{aligned}\sqrt{n}\int \tilde{\theta}_{\hat{F}_n} d(Q_n - Q_{F_0}) &= \sqrt{n}\int \tilde{\theta}_{F_0} d(Q_n - Q_{F_0}) \\ &\quad + \sqrt{n}\int (\tilde{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}) d(Q_n - Q_{F_0}),\end{aligned}$$

and the last term will be shown to be $o_p(1)$. We will not give the proof in full detail, but restrict ourselves to the basic ideas. The full proof is rather extended and can be found in GESKUS AND GROENEBOOM (1996B). We use the second version of r_F, given by (2.28) and (2.30), since for this version we have proved the uniform Lipschitz properties to hold. However, the other version, given in (2.29), leads to the same result. For we have

$$\begin{aligned}
&\int_{u=\tau_i}^{\tau_{i+1}} \int_{v=u}^{\tau_{i+1}} [\psi(v)-\psi(u)]\, d(Q_n(u,v,0,1) - Q_{F_0}(u,v,0,1)) \\
&\quad= \int_{u=\tau_i}^{\tau_{i+1}} \int_{v=u}^{\tau_{i+1}} [\psi(v)-\psi(u)]\, h(u,v)\, [F_0(v)-F_0(u)]\, dv\, du \\
&\quad= \int_{v=\tau_i}^{\tau_{i+1}} F_0(v) \left[\int_{u=\tau_i}^{v} [\psi(v)-\psi(u)]\, h(u,v)\, du\right] dv \\
&\qquad - \int_{u=\tau_i}^{\tau_{i+1}} F_0(u) \left[\int_{v=u}^{\tau_{i+1}} [\psi(v)-\psi(u)]\, h(u,v)\, dv\right] du \\
&\quad= \int_{v=\tau_i}^{\tau_{i+1}} F_0(v) \left[r(v) + \int_{u=v}^{\tau_{i+1}} [\psi(u)-\psi(v)]\, h(v,u)\, du\right] dv \\
&\qquad - \int_{u=\tau_i}^{\tau_{i+1}} F_0(u) \left[\int_{v=u}^{\tau_{i+1}} [\psi(v)-\psi(u)]\, h(u,v)\, dv\right] du \\
&\quad= \int_{v=\tau_i}^{\tau_{i+1}} F_0(v)\, r(v)\, dv,
\end{aligned}$$

with r defined by (2.37) on page 82. A similar computation for version (2.30) also leads to $\int F_0(v)\, r(v)\, dv$, implying that both versions lead to $\sqrt{n}\int(\tilde\theta_{\hat F_n} - \tilde\theta_{F_0})d(Q_n - Q_{F_0})$.

So we see that the ψ-function is only playing a minor role. It is needed for a correct definition of the integral equation, and should occur in such a way that lemma 2.5.1 holds. However, this still leaves some freedom in choosing ψ.

The main extra difficulty compared to case 2A is the fact that the denominator in $\tilde\theta_F$ for $\gamma = 1$ can be arbitrarily close to zero and is not compensated by h being zero. The parts $\delta = 1$ and $\delta = \gamma = 0$ do not lead to extra difficulties, since ϕ_F contains a factor $F(1-F)$. But for $\gamma = 1$ we can no longer neglect the denominator when trying to compute the two-dimensional Lipschitz norm. In fact, computer simulations strongly suggest the quotient $\gamma.r_F$ *not* to be of uniformly bounded variation. We have

$$\begin{aligned}
\frac{\phi_{\hat F_n}(\tau_{i+1}) - \phi_{\hat F_n}(\tau_i)}{\tau_{i+1}-\tau_i} &= \frac{\int_{z_i}^{z_{i+1}} \bar\phi'_{\hat F_n}(s)\, ds}{\tau_{i+1}-\tau_i} \\
&\le K\, \frac{z_{i+1}-z_i}{\tau_{i+1}-\tau_i} = K\, \frac{\hat F_n(\tau_{i+1}) - \hat F_n(\tau_i)}{\tau_{i+1}-\tau_i}.
\end{aligned}$$

The latter quotient does not seem to show less variation with increasing number of points.

This difficulty is faced by considering three regions of integration:

$$C_{n,\eta}(F) = \{w : q_F(w) > \eta\, q_{F_0}(w),\ q_{F_0}(w) > n^{-1/3}\}, \tag{2.56}$$

$$D_\eta(F) = \{w : q_F(w) \le \eta q_{F_0}(w)\}, \tag{2.57}$$

and

$$C_n(F_0) = \{w : q_{F_0}(w) \le n^{-1/3}\}, \tag{2.58}$$

for some $\eta \in (0,1)$, where the elements w are of the form $w = (t, u, \delta, \gamma)$. On the region $C_{n,\eta}(F)$, $\tilde{\theta}_F$ has a behaviour which is comparable to the behaviour of $\tilde{\theta}_{F_0}$; on the other regions we use the uniform boundedness of $\tilde{\theta}_F$ and the fact that the integrals over these regions become sufficiently small.

First we state two lemma's, which are proved in GESKUS AND GROENEBOOM (1996B).

Lemma 2.5.2 *(i) Let the function a_n be defined by*

$$a_n = 1_{\{q_{F_0} > n^{-1/3}\}} \; / \; q_{F_0}^2.$$

Then

$$Q_n a_n = \mathcal{O}_p(\log n) \tag{2.59}$$

(ii) Let the function b_n be defined by

$$b_n = 1_{\{q_{F_0} \le n^{-1/3}\}}.$$

Then

$$Q_n b_n = \mathcal{O}_p(n^{-2/3}) \tag{2.60}$$

For the next lemma we have to exclude part of the possible outcomes $\hat{F}_n$, occurring with small probability. By proposition 2.1.4, if $\mathcal{F}_n$ is the set of distribution functions $F \in \mathcal{F}$, satisfying

$$h^2(q_F, q_{F_0}) \le n^{-2/3} \log n, \tag{2.61}$$

with h denoting Hellinger distance (see page 65), we have:

$$\text{Prob}\{\hat{F}_n \in \mathcal{F}_n\} \to 1, \text{ as } n \to \infty.$$

In fact, the upper bound $n^{-2/3} \log n$, defining the class $\mathcal{F}_n$, can be replaced by

$$c_n n^{-2/3} (\log n)^{1/3},$$

where we only need $c_n \to \infty$, as $n \to \infty$. However, being a little bit wasteful with powers of $\log n$ avoids an accumulation of constants in the upper bounds.

Now we have

Lemma 2.5.3 *Let, for $\eta \in (0,1)$, the set $D_\eta(F)$ be defined by (2.57). Then*

$$\sup_{F \in \mathcal{F}_n} Q_n D_\eta(F) = \mathcal{O}_p(n^{-2/3} \log n). \tag{2.62}$$

Lemma 2.5.2.(i) is not directly related to one of the areas of integration $C_{n,\eta}(F)$, $D_\eta(F)$ or $C_n(F_0)$. However, it is used in the entropy calculations in lemma 2.5.3 and below.

Using the notation of POLLARD (1984), page 150, we let E_n denote the empirical process $\sqrt{n}(Q_n - Q_{F_0})$ and let E_n^0 denote the symmetrized empirical process. Fix an (arbitrary) $\epsilon > 0$. Restricting to the most difficult part with $\gamma = 1$, we have by the symmetrization lemma

$$\text{Prob}\Big\{|E_n(r_F - r_{F_0})\gamma| > \epsilon \text{ for some } F \in \mathcal{F}_n\Big\} \le$$
$$\le 4\,\text{Prob}\Big\{|E_n^0(r_F - r_{F_0})\gamma| > \tfrac{1}{4}\epsilon \text{ for some } F \in \mathcal{F}_n\Big\}.$$

Let $\eta \in (0,1)$ be fixed. The rest of the proof consist in showing that

$$Pr\Big\{|E_n^0(r_F - r_{F_0})\gamma| > \tfrac{1}{4}\epsilon \text{ for some } F \in \mathcal{F}_n \mid \xi_n\Big\} \to 0, \text{ as } n \to \infty, \tag{2.63}$$

for all

$$\xi_n = ((T_1, U_1, \Delta_1, \Gamma_1), \ldots, (T_n, U_n, \Delta_n, \Gamma_n)),$$

such that

$$\int_{q_{F_0} \le n^{-1/3}} dQ_n \le n^{-2/3} \log n, \tag{2.64}$$

$$\int_{q_{F_0} > n^{-1/3}} q_{F_0}^{-2}\, dQ_n \le (\log n)^2, \tag{2.65}$$

and

$$\sup_{F \in \mathcal{F}_n} Q_n D_\eta(F) = \sup_{F \in \mathcal{F}_n} \int_{q_F \le \eta q_{F_0}} dQ_n \le n^{-2/3} (\log n)^2, \tag{2.66}$$

are satisfied for the empirical measure Q_n, corresponding to ξ_n. By the preceding lemmas, the probability that these conditions are *not* satisfied for the sample ξ_n tends to zero, as $n \to \infty$. In (2.64) to (2.66) we again use the method of absorbing constants into extra powers of $\log n$. For the entropy calculations, ratios $r_{F_k, G_k, \bar\phi_k}$ of the form

$$\frac{\bar\phi_k(G_k(u)) - \bar\phi_k(F_k(t))}{G_k(u) - F_k(t)}$$

are used, where F_k and G_k are distribution functions such that $F_k \le F \le G_k$ ((F_k, G_k) is a "bracket" for F) and where $\bar\phi_k$ is a Lipschitz function approximating $\bar\phi_F$. In this way the good behaviour of the ratios r_F on the region $C_{n,\eta}(F)$ is preserved on the same region by the approximating ratio $r_{F_k, G_k, \bar\phi_k}$. Note that the approximating ratios are outside the original class of ratios r_F. The basic remaining part of the proof consists of a chaining argument, somewhat along the lines of proof on page 161 in POLLARD (1984).

2.6 Some simulations

2.6.1 A computation of $\phi_{\hat F_n}$, $\bar\phi_{\hat F_n}$ and $\psi_{\hat F_n}$

Let F be a discrete distribution function, belonging to the class $\mathcal{F}$ on page 73. For case 2A, we have a Fredholm integral equation, and we can use the algorithms from the Numerical Recipes book (PRESS *et al.* (1992)). We will only give the computation for an example belonging to case 2B. For case 2B, computation is even easier. We know from theorem 2.5.2 that ϕ_F, as given by equation 2.27, is a piecewise constant function as well. In this equation, we do not need the ψ_F-part in order to obtain the ϕ_F-solution. We know from part (iii) of theorem 2.5.2 that the values of ϕ_F can be found from a finite set of linear equations $Ay = b$. The matrix A has positive diagonal elements and non-positive off-diagonal elements (such a matrix is called an *M-matrix*). It is a *strictly diagonally dominant* matrix, meaning that each diagonal element is strictly bigger than the sum of the absolute values of the off-diagonal elements in the same row. In BERMAN AND PLEMMONS (1979) it is shown that a symmetric, strictly diagonally dominant M-matrix (also called a Stieltjes matrix) is positive definite. So Cholesky decomposition can be used, which is a fast algorithm and numerically stable.

The solution of the integral equation in the transformed scale is easily obtained from this, since the integral parts are with respect to a measure that has mass restricted to the values $z_i = \hat{F}_n(\tau_i)$. In figure 2.1 we give a picture of the NPMLE and in figures 2.2 to 2.4 we give the solutions $\bar{\phi}_{\hat{F}_n}$, $\bar{\xi}_{\hat{F}_n}$ and $\bar{\phi}'_{\hat{F}_n}$ respectively, based on a random sample of size $n = 300$ from a uniform distribution on $[0,1]$, censored by two uniformly distributed observation times, where $k \equiv 1$. This is the case considered in section 1.5.1. Hence these solutions can be compared with the solution $\bar{\phi}_{F_0}$, which is equal to ϕ_{F_0} since F_0 is the uniform distribution on $[0,1]$.

The number of jumps of the NPMLE was 15 and the locations of the jumps are indicated by small vertical bars (slightly smaller than the tickmarks at 0.25, etc.) on the x-axis in figure 2.1. On the other hand, in figures 2.2 to 2.4 the small vertical bars on the x-axis denote the values of $\hat{F}_n$ at these points of jump.

There are some interesting things to notice from these figures. The derivative $\bar{\phi}'_{\hat{F}_n}$ is continuous (this does not hold in general!). Moreover, it has cusps at the points $\hat{F}_n(\tau_i)$, which seem to be located on a curve. Indeed, we will show the cusps to be located on the curve $t \mapsto \frac{1}{2}(1-2t)\bar{\xi}_{\hat{F}_n}(t)$, $t \in (0,1)$.

Proposition 2.6.1 *Let $U = \min(T_1, T_2)$, $V = \max(T_1, T_2)$, with T_1 and T_2 uniformly distributed on $[0,1]$. Let $K(F_0) = \int x\,dF_0(x)$.*
Then $\bar{\phi}'_{\hat{F}_n}$ is continuous and

$$\bar{\phi}'_{\hat{F}_n}(\tau_i) = \frac{1}{2}\,\xi_{\hat{F}_n}(\tau_i)\,[1 - 2F(\tau_i)]. \tag{2.67}$$

Proof:
We turn back to the integral equation (2.40) on the inverse scale, as given in part (i) of theorem 2.5.2 on page 84. This equation was only defined at the points that do not belong to the range of $\hat{F}_n$. However, letting $y \downarrow F(\tau_i)$, we find

$$\begin{aligned} &\phi_F(\tau_i) = \bar{\phi}_F(F(\tau_i)) \\ = \;& \frac{F(\tau_i)(1-F(\tau_i))}{h_1(\tau_{i+1})(1-F(\tau_i)) + h_2(\tau_{i+1})F(\tau_i)} \times \\ &\times\Big\{k(\tau_{i+1}) - \int_{t:F(t)<F(\tau_i)} \tfrac{\phi_F(\tau_i)-\phi_F(t)}{F(\tau_i)-F(t)}\,dH(t,\tau_{i+1}) \\ &-\bar{\phi}'_F(F(\tau_i)+)\int_{t:F(t)=F(\tau_i)} dH(t,\tau_{i+1}) + \int_{t:F(t)>F(\tau_i)} \tfrac{\phi_F(t)-\phi_F(\tau_i)}{F(t)-F(\tau_i)}\,dH(\tau_{i+1},t)\Big\}, \end{aligned} \tag{2.68}$$

and letting $y \uparrow F(\tau_i)$, we get

$$\begin{aligned} &\phi_F(\tau_i) = \bar{\phi}_F(F(\tau_i)) \\ = \;& \frac{F(\tau_i)(1-F(\tau_i))}{h_1(\tau_i)(1-F(\tau_i)) + h_2(\tau_i)F(\tau_i)} \times \end{aligned}$$

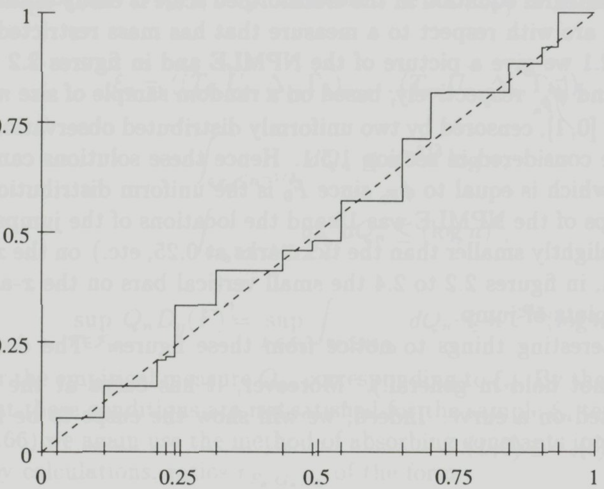


Figure 2.1: $\hat{F}_n$, based on sample size 300, and F_0 (dashed)

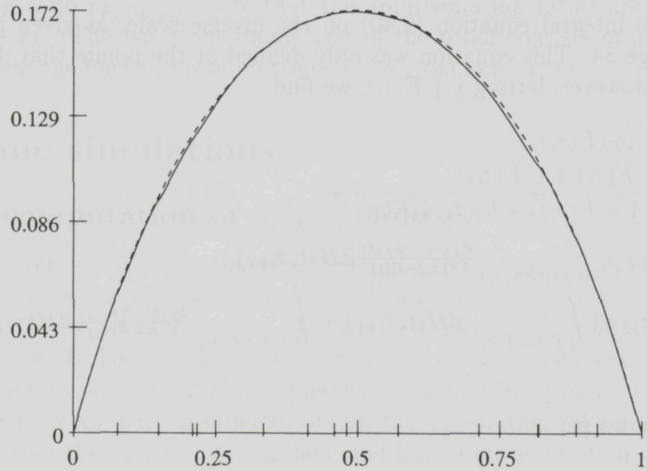


Figure 2.2: $\bar{\phi}_{\hat{F}_n}$ and $\bar{\phi}_{F_0}$(dashed)

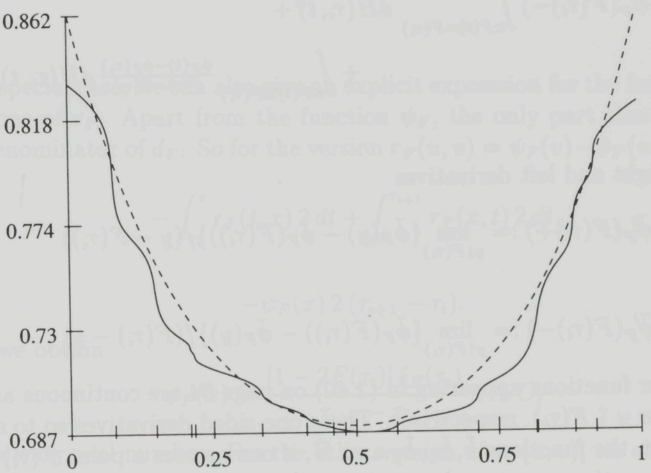


Figure 2.3: $\bar{\xi}_{\hat{F}_n}$ and $\bar{\xi}_{F_0}$(dashed)

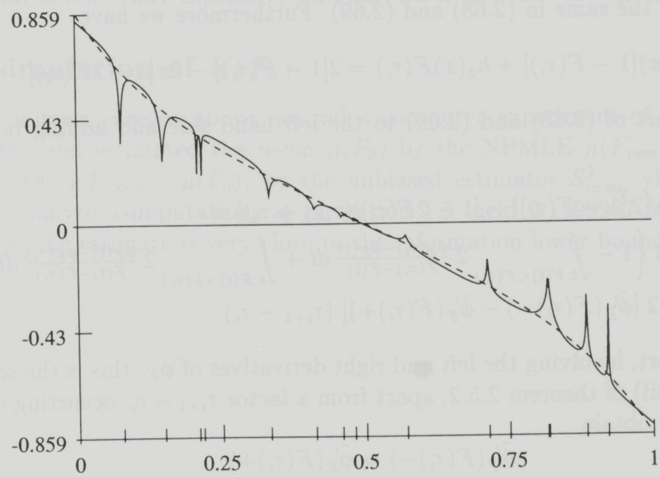


Figure 2.4: $\bar{\phi}'_{\hat{F}_n}$ and $\bar{\phi}'_{F_0}$(dashed)

$$\times\Big\{k(\tau_i)-\int_{t:F(t)<F(\tau_i)}\tfrac{\phi_F(\tau_i)-\phi_F(t)}{F(\tau_i)-F(t)}\,dH(t,\tau_i)$$
$$+\bar\phi'_F(F(\tau_i)-)\int_{t:F(t)=F(\tau_i)}dH(\tau_i,t)+$$
$$+\int_{t:F(t)>F(\tau_i)}\tfrac{\phi_F(t)-\phi_F(\tau_i)}{F(t)-F(\tau_i)}\,dH(\tau_i,t)\Big\}. \tag{2.69}$$

Note that the right and left derivatives

$$\bar\phi'_F(F(\tau_i)+):=\lim_{y\downarrow F(\tau_i)}\{\bar\phi_F(y)-\bar\phi_F(F(\tau_i))\}/(y-F(\tau_i))$$

and

$$\bar\phi'_F(F(\tau_i)-):=\lim_{y\uparrow F(\tau_i)}\{\bar\phi_F(F(\tau_i))-\bar\phi_F(y)\}/(F(\tau_i)-y)$$

exist, since the other functions appearing in (2.40) on page 84 are continuous and have finite limits as $y\downarrow F(\tau_i)$ or $y\uparrow F(\tau_i)$, respectively. These one-sided derivatives so to speak "catch" the discontinuities in the functions $\bar k,\bar h_1,\bar h_2$ and $\bar H$, if one crosses a point $F(\tau_i)$ in the range; the function $\bar\phi_F$ is continuous at such a point and can be defined there by either taking the left-hand limit (involving $h_i(\tau_i),k(\tau_i)$, etc. at $F(\tau_i)$), or the right-hand limit (involving $h_i(\tau_{i+1}),k(\tau_{i+1})$, etc. at $F(\tau_i)$).
Also note that the integrals are just summations.

First we show the derivative $\bar\phi'_{\hat F_n}$ to be continuous, for the special case under consideration. We have $k(\tau_i)=1$ and $dH(t,\tau_i)=dH(\tau_i,t)=2\,dt$ for all i. Hence the parts involving the integrals and k are the same in (2.68) and (2.69). Furthermore we have

$$h_1(x)[1-F(\tau_i)]+h_2(x)F(\tau_i)=2[1-F(\tau_i)]-2x[1-2F(\tau_i)] \tag{2.70}$$

Bringing the d_F-part of (2.68) and (2.69) to the left-hand side and adding both equations yields

$$2\,\xi_F(\tau_i)\,[2\,[1-F(\tau_i)]-[1-2F(\tau_i)](\tau_{i+1}+\tau_i)]=$$
$$=2\Big\{1-\int_{t:F(t)<F(\tau_i)}2\,\tfrac{\phi_F(\tau_i)-\phi_F(t)}{F(\tau_i)-F(t)}\,dt+\int_{t:F(t)>F(\tau_i)}2\,\tfrac{\phi_F(t)-\phi_F(\tau_i)}{F(t)-F(\tau_i)}\,dt\Big\}$$
$$+2\,[\bar\phi'_F(F(\tau_i)-)-\bar\phi'_F(F(\tau_i)+)]\,(\tau_{i+1}-\tau_i)$$

Without the last part, involving the left and right derivatives of $\bar\phi_F$, this is the same equation as the one in part (iii) of theorem 2.5.2, apart from a factor $\tau_{i+1}-\tau_i$, occurring on both sides of (2.41). Hence we obtain

$$\bar\phi'_F(F(\tau_i)-)=\bar\phi'_F(F(\tau_i)+).$$

To prove the second part of the proposition, we bring the d_F-part of (2.68) and (2.69) to the left-hand side and subtract the first from the second. This yields

$$\phi_F(\tau_i)\,(\tau_{i+1}-\tau_i)\,[1-2F(\tau_i)]=2\,\bar\phi'_F(F(\tau_i))\,F(\tau_i)[1-F(\tau_i)]\,(\tau_{i+1}-\tau_i),$$

from which we derive

$$\bar{\phi}'_{\hat{F}_n}(\tau_i) = \frac{1}{2}\xi_{\hat{F}_n}(\tau_i)\,[1 - 2F(\tau_i)].$$

□

For the above special case, we can also give an explicit expression for the function ψ_F in (2.27), for both versions of r_F. Apart from the function ψ_F, the only part that is not constant in (2.27) is the denominator of d_F. So for the version $r_F(u,v) = \psi_F(v) - \psi_F(u)$, the non-constant part in

$$-\int_{\tau_i}^{x} r_F(t,x)\,2\,dt + \int_{x}^{\tau_{i+1}} r_F(x,t)\,2\,dt$$

is equal to

$$-\psi_F(x)\,2\,(\tau_{i+1} - \tau_i).$$

Using (2.70), we obtain

$$\psi_F(x) = \frac{[1 - 2F(\tau_i)]\,\xi_F(\tau_i)}{\tau_{i+1} - \tau_i}\,x + C,$$

with C an arbitrary real number. For the other version of $r_F(u,v)$, we obtain

$$\psi_F(x) = \lambda\,x + C$$

with

$$\lambda = \frac{1}{2}[1 - 2F(\tau_i)]\,\xi_F(\tau_i).$$

Note that this function is equal to (2.67). Also note that this last version of r_F leads to a ψ_F-function that is uniformly Lipschitz, whereas this is not the case for the first version.

2.6.2 A simulation of $K(\hat{F}_n)$

For the same uniform case as above, we did a computer experiment of 10.000 samples of magnitude 1000, and estimated the mean $\mu(F_0)$ by the NPMLE $\mu(\hat{F}_{1000})$. Estimating the variance of $\sqrt{1000}(\mu(\hat{F}_{1000}) - \mu(F_0))$ by the unbiased estimator $S^2_{10.000}$ yielded the number 0.11917, while analytic computations as in section 1.5 yield 0.1198987 for the information lower bound. So the estimate is very close to the information lower bound.

from which we derive

□

For the above special case, we can also give an explicit expression for the function c_F in (2.27), for depth versions of [illegible]. Apart from the function [illegible], the only part that is not constant in (2.27) is the [illegible] part is

is equal to

Using [illegible] we obtain

with [illegible]

$$+2\,[\dot\phi'_F(F(\tau_i)-)-\dot\phi'_F(F(\tau_i)+)]\,(\tau_{i+1}-\tau_i)$$

Without the last part, involving the left and right derivatives of $\dot\phi_F$, this is the same equation as the one in part (iii) of theorem 2.5.2, apart from a factor $\tau_{i+1}-\tau_i$, occurring on both sides of (2.41). Hence we obtain

$$\dot\phi'_F(F(\tau_i)-)=\dot\phi'_F(F(\tau_i)+).$$

To prove the second part of the proposition, we bring the ϕ_F-part of (2.58) and (2.59) to the left-hand side and subtract the first from the second. This yields

$$\phi_F(\tau_i)\,(\tau_{i+1}-\tau_i)\,[1-2F(\tau_i)]=2\,\dot\phi'_F(F(\tau_i))\,F(\tau_i)[1-F(\tau_i)]\,(\tau_{i+1}-\tau_i).$$

Bibliography

ANDERSEN P.K., BORGAN Ø., GILL R.D. AND KEIDING N. (1993). *Statistical models based on counting processes.* Springer, New York.

BERMAN A. AND PLEMMONS R.J. (1979). *Nonnegative matrices in the mathematical sciences.* Academic Press, New York.

BICKEL P.J., KLAASSEN C.A.J., RITOV Y. AND WELLNER J.A. (1993). *Efficient and adaptive estimation in semiparametric models.* John Hopkins University Press, Baltimore.

CHANG M.N. AND YANG G.L. (1987). *Strong consistency of a nonparametric estimator of the survival function with doubly censored data.* Annals of Statistics, vol. 15, p. 1536-1547.

CHANG M.N. (1990). *Weak convergence of a self-consistent estimator of the survival function with doubly censored data.* Annals of Statistics, vol. 18, p. 391-404.

DE GRUTTOLA V. AND LAGAKOS S.W. (1989). *Analysis of doubly-censored data, with application to AIDS.* Biometrics, vol. 45, p. 1-11.

GESKUS R.B. (1989). *On the differentiability of functionals defined on a class of probability measures.* Master's thesis, University of Amsterdam.

GESKUS R.B. (1992). *Efficient estimation of the mean for interval censoring case II.* Technical Report 92-83, Delft University of Technology.
`ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1992`

GESKUS R.B. AND GROENEBOOM P. (1995a). *Asymptotically optimal estimation of smooth functionals for interval censoring, case 2 and beyond.* Technical Report 95-78, Delft University of Technology.
`ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1995`

GESKUS R.B. AND GROENEBOOM P. (1995b). *Asymptotically optimal estimation of smooth functionals for interval censoring, part 1.* Statistica Neerlandica 50, p. 69-88.

GESKUS R.B. AND GROENEBOOM P. (1996a). *Asymptotically optimal estimation of smooth functionals for interval censoring, part 2.* Statistica Neerlandica, to appear.

GESKUS R.B. AND GROENEBOOM P. (1996b). *Asymptotically optimal estimation of smooth functionals for interval censoring, case 2.* Technical Report 96-36, Delft University of Technology.

GROENEBOOM, P. (1996). *Lectures on inverse problems*, in: Lectures on probability theory, Ecole d'Eté de Probabilités de Saint-Flour XXIV-1994, Editor: P. Bernard. Springer Verlag, Berlin.

GROENEBOOM P. AND WELLNER J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* Birkhäuser Verlag.

HANSEN, B.E. (1991). *Nonparametric estimation of functionals for interval censored observations.* Master's thesis, Delft University of Technology.

HUANG J. AND WELLNER J.A. (1995a). *Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1.* Statistica Neerlandica, vol. 49, p. 153-163.
`http://www.stat.washington.edu:80/jaw/jaw.research.available.html`

HUANG J. AND WELLNER J.A. (1995b). *Efficient estimation for the proportional hazards model with "Case 2" interval censoring.* Technical Report 290, University of Washington, Seattle. Submitted to Biometrika.
`http://www.stat.washington.edu:80/jaw/jaw.research.available.html`

HUANG J. AND WELLNER J.A. (1996). *Interval censored survival data: a review of recent progress.* In: Proceedings of the survival analysis symposium, Seattle.

JEWELL N.P., MALANI H.M. AND VITTINGHOFF E. (1994). *Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS.* Journal of the American Statistical Association, vol. 89, p. 7-18.

JONGBLOED G. (1995). *The iterative convex minorant algorithm for nonparametric estimation.* Technical Report 95-105, Delft University of Technology, submitted.
`ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1995`

KEIDING N. (1991). *Age-specific incidence and prevalence: a statistical perspective (with discussion).* Journal of the Royal Statistical Society A, vol. 154, p. 371-412.

KRESS R. (1989). *Linear integral equations.* Applied Mathematical Sciences, vol. 82, Springer Verlag, New York.

LEHMANN E.L. (1983). *Theory of point estimation.* Wiley, New York.

POLLARD, D. (1984). *Convergence of stochastic processes.* Springer-Verlag.

PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. AND FLANNERY, B.P. (1992). *Numerical recipes in C.* Cambridge University Press.

REED, M. AND SIMON, B. (1972). *Methods of modern mathematical physics I: functional analysis.* Academic Press, New York.

ROCKAFELLAR, R.T. (1970). *Convex analysis.* Princeton University Press.

SZEGÖ. G. (1978). *Orthogonal polynomials.* Colloquium Publications 23, A.M.S., Providence, Rhode Island.

VAN DE GEER S. (1993). *Hellinger consistency of certain nonparametric maximum likelihood estimators.* Ann. Statist., vol. 21, p. 14-44.

VAN DE GEER S. (1996). *Rates of convergence for the maximum likelihood estimator in mixture models.* Nonparametric Statistics, vol. 6, p. 293-310.

VAN HAASTRECHT, H.J.A., VAN AMEIJDEN, E.J.C., VAN DEN HOEK, A., MIENTJES, G.H.C., BAX, J.S. AND COUTINHO, R.A. (1996). *Predictors of mortality in the Amsterdam cohort of HIV-positive and HIV-negative drug users.* Am. J. Epidemiol. vol. 143, p. 380-391.

VAN DER LAAN M.J. (1993). *Efficient and inefficient estimation in semiparametric models.*. Ph.D. Thesis, Utrecht University, The Netherlands.

VAN DER VAART A.W. (1991). *On differentiable functionals.* Annals of Statistics, vol. 19, p. 178-204.

VAN DER VAART A.W., AND WELLNER, J. (1996). *Weak convergence and empirical processes.* Springer Verlag, New York.

WELLNER J. (1995). *Interval censoring case 2: an exploration of alternative hypotheses.* In: Analysis of censored data. IMS Lecture Notes-Monograph Series, vol. 29, Hayword. Editors: H.L. Koul and J.V. Deshpande.
`http://www.stat.washington.edu:80/jaw/jaw.research.available.html`

Huang J. and Wellner J.A. (1996). *Interval censored* [illegible] progress. In: Proceedings of the [illegible]

Jewell N.P., Malani H.M. and Vittinghoff E. (1994). *Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS.* Journal of the American Statistical Association, vol. 89, p. 7-18.

Jongbloed G. (1995). *The iterative convex minorant algorithm for nonparametric estimation.* Technical Report 95-105, Delft University of Technology, submitted.
ftp://ftp.twi.tudelft.nl/pub/publications/tech-reports/1995

Keiding N. (1991). *Age-specific incidence and prevalence: a statistical perspective (with discussion).* Journal of the Royal Statistical Society A, vol. 154, p. 371-412.

Kress R. (1989). *Linear integral equations.* Applied Mathematical Sciences, vol. 82, Springer Verlag, New York.

Lehmann E.L. (1983). *Theory of point estimation.* Wiley, New York.

Pollard, D. (1984). *Convergence of stochastic processes.* Springer-Verlag.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical recipes in C.* Cambridge University Press.

Reed, M. and Simon, B. (1972). *Methods of modern mathematical physics I: functional analysis.* Academic Press, New York.

Rockafellar, R.T. (1970). *Convex analysis.* Princeton University Press.

Szegö, G. (1978). *Orthogonal polynomials.* Colloquium Publications 23, A.M.S., Providence, Rhode Island.

van de Geer S. (1993). *Hellinger consistency of certain nonparametric maximum likelihood estimators.* Ann. Statist., vol. 21, p. 14-44.

van de Geer S. (1995). *Rates of convergence for the maximum likelihood estimator in mixture models.* Nonparametric Statistics, vol. 6, p. 293-310.

Chapter 3

On the NPMLE in repulsive Gibbs models

3.1 Spatial patterns

An illustrative real-life example of a random spatial pattern is the location of nest sites of birds. Depending on habitat and species, several kinds of nest patterns occur. The influence of habitat is clear: no goshawk has ever built its nest on open water, whereas a mute swan does not build its nest on top of a tree. Within a suitable habitat, differences do occur as well. Herring gulls show gregarious nesting behaviour, whereas goshawk and mute swan are more inclined to have a breeding territory, in which no birds of the same species are allowed.

When modelling a random pattern of small objects in a bounded region, say A, the simplest model arises if we suppose *complete spatial randomness.* This means that there is no preference for certain subregions (homogeneity), and that the location of each object is not influenced by the location of the other objects (independence).

If the number of objects is fixed, say n, a realization of such a pattern can be obtained by randomly choosing n points in A. The number of points in a subregion $B \subset A$, $N(B)$, has a binomial distribution with parameters n and $|B|/|A|$. (Here and in the sequel $|\cdot|$ denotes the area of a set.) Often, the number of objects in A is random as well. Then complete spatial randomness can be constructed via a limiting procedure. $n(K)$ points are generated uniformly in a region $K \supset A$, which is expanded to $\mathbb{R}^2$. The number of points in K is made to converge to ∞ in such a way that $n(K)/|K| \to \lambda$ for some $\lambda \in \mathbb{R}$. What we get in the limit is called a *homogeneous Poisson point process* on $\mathbb{R}^2$. λ is called the intensity of the process, the average number of points per unit area. For any $B \in \mathbb{R}^2$ we have

$$P\{N(B) = k\} = \exp(-\lambda|B|)\,\frac{(\lambda|B|)^k}{k!} \tag{3.1}$$

Hence $E(N(B)) = \lambda|B|$. Since the points have been generated independently, the random variables $N(B_i)$ are independent if the B_i's are disjoint. Formula (3.1) completely characterizes a homogeneous Poisson point process.

Although an extension to $\mathbb{R}^2$ was used in the construction of the process in order to allow for a random number of points in A, we can forget everything outside A and use formula (3.1) for all $B \subset A$ to obtain a homogeneous Poisson point process on A. Note that conditionally on $N(A)=n$ we are back in the situation of randomly choosing n points in A.

Complete spatial randomness can be tested; a wide range of possibilities to test on the null hypothesis of a homogeneous Poisson process exists. See STOYAN *et al.* (1987), section 2.7, and RIPLEY (1977) for an overview of testing procedures for the null hypothesis of a homogeneous Poisson process.

With respect to the nest patterns of birds, the homogeneous Poisson model may sometimes be a reasonable model. Some nesting habitats, like the Russian taiga or the desert, are quite homogeneous. However, many habitats show geographical variation, thus causing non-constant nesting intensity over the region. If we abandon the homogeneity assumption, but still stick to an independent choice of nest sites, we get a *nonhomogeneous Poisson point process*, for which (3.1) holds as well, but with $\lambda|B|$ replaced by an intensity measure $\Lambda(B)$. For a nonhomogeneous pattern, the independence assumption can only be tested if one has several independent realizations of the same point process, since any single point pattern on a bounded region can always be fitted in the Poisson model by choosing an appropriate intensity measure. A recent article testing the null hypothesis of independence is MCDONALD (1989). As an example, the hypothesis that redwings choose their nest sites independently in some inhomogeneous region is tested, and rejected.

For points that cannot be seen as generated independently, such as nest sites of birds showing gregarious or repulsive nesting behaviour, several alternative models have been developed (see STOYAN *et al.* (1987), chapter 5, for an overview). Although these models are no longer Poisson processes, the Poisson process is often still playing a role. Estimation procedures usually have to be based on a single instantaneous observation of a point pattern. Then homogeneity of the point pattern has to be taken as an untestable model assumption. Note that outside the realm of the Poisson model homogeneity is no longer common terminology. Stationarity (translation invariance) and isotropy (rotation invariance), together called *motion invariance*, are used instead.

One of these models is the cluster process. This process is obtained by considering two subprocesses. The first generates a random collection of parent points. At each parent point, daughter points are generated according to the second process. In some models the parent points are included in the final realization, in some models they are not. Often, both the parent process and the daughter process are assumed to be homogeneous Poisson processes. The cluster model may be used to describe the choice of nest sites of a gregarious bird species like the herring gull.

A model dealing with repulsive forces between points is the point process with dependent thinning. In this model, a pattern is formed by first generating points through some point process, often a homogeneous Poisson process. Afterwards points are deleted simultaneously with deletion probability of a point depending on the distance to its nearest neighbour. Simultaneous deletion is needed in order to prevent the probability of the configuration to depend on the order of deletion.

We will deal with another class of models, which are called Gibbs point processes. Attractive as well as repulsive forces between points can be modelled by such processes. Usually the strength of the interaction forces is assumed to be determined by all interpoint distances (and not by the distance to the *nearest* neighbouring point only, as in the thinning model).

In CRESSIE (1991), section 8.5, an overview of estimation methods for Gibbs processes is given. The observable region A is usually seen as a subregion of some larger region E. Sometimes this approach may be motivated by the real life situation on which the model is based: the observation window may be part of some larger area with the same characteristics. However, points outside A, which are invisible, exert influence on points inside A, so with this approach edge corrections have to be taken into account.

The amount of repulsion can be modelled by the *nearest-neighbour distribution function* D. For a stationary process, it is defined as one minus the probability that there are no other points in a ball $b(x, r)$ around an arbitrary point x of the process:

$$D(r) = 1 - P\{N(b(x, r)) = 1 | \text{ point at } x\}$$

Equivalently, it is the distribution function of the distance from an arbitrary point of the process to the nearest other point in the process. The *empty space function* F is the distribution function of the distance of an arbitrary point in the area A to the nearest point of the process. For a homogeneous Poisson process on $\mathbb{R}^2$ with intensity λ, we have

$$D(r) = F(r) = 1 - \exp(-\lambda\pi r^2).$$

Several estimators have been proposed for D and F, see for example STOYAN *et al.* (1987), p. 128. In VAN LIESHOUT AND BADDELEY (1996), both functions are combined in the formula

$$J(r) = \frac{1 - D(r)}{1 - F(r)}.$$

They show this function to have some nice characteristics. It is computed rather easily. It is identically one for Poisson processes. Values smaller than one indicate clustering, whereas values larger than one indicate repulsive forces. The function remains constant for values larger than the interaction range.

A nice characteristic of Gibbs point processes is that a formula for the density is available. However, this formula contains a very complex norming constant, making pure maximum likelihood estimation an impossible task. Apart from methods that are not related to the maximum likelihood procedure, like the method based on the function J mentioned above, one can use some approximation of the likelihood function or use a pseudo maximum likelihood approach. In DIGGLE *et al.* (1994) three methods are compared: an approximate maximum likelihood, a pseudo maximum likelihood and the Takacs-Fiksel method, all from a parametric point of view. Only few authors have considered nonparametric methods. In DIGGLE *et al.* (1987) the Percus-Yevick approximation, known from statistical physics, is used to obtain a nonparametric procedure.

We will consider the pure nonparametric maximum likelihood approach under the side condition of the interaction forces being repulsive only, with strength decreasing with increasing interpoint distance. An approximate maximum likelihood estimator is derived. Simulation experiments are performed in order to test the behaviour of the estimator for some choices of repulsive forces.

Repulsive Gibbsian point patterns generally have a more regular structure than Poisson patterns, therefore such patterns are sometimes called *regular* point patterns.

3.2 Gibbs point processes

Gibbs point processes have their origin in statistical physics, describing the behaviour of a particle system (like a gas or a fluid) in a bounded volume $V \subset \mathbb{R}^3$. The particle system is supposed to be in equilibrium, and to have a fixed number of elements, say n. On a microscopic scale, the positions of the n elements with respect to some coordinate system can be seen either as an ordered n-tuple $(x_1, \ldots, x_n)$ or as an unordered set $\{x_1, \ldots, x_n\}$, leading to formulas differing by a factor $n!$. The microscopic behaviour of the particle system is described by a point process Φ, with probability density of the (ordered) configuration $\phi = (x_1, \ldots, x_n)$ given by a function f_n: $V^n \to [0, \infty)$ of the form

$$f_n(\phi) = \exp\{-\frac{U_n'(\phi)}{kT}\}/Z. \tag{3.2}$$

Here T denotes absolute temperature and k is Boltzmann's constant. The function

$$U_n' \colon V^n \to \mathbb{R} \cup \{\infty\}$$

is called the *energy function* or *multiparticle potential.* It is usually written as a sum of *interaction potentials* over all subconfigurations:

$$U_n'(\phi) = \sum_{\psi \subset \phi} W(\psi),$$

thus giving f_n a multiplicative structure:

$$f_n(\phi) = \frac{1}{Z} \prod_{\psi \subset \phi} \exp\{-\frac{W(\psi)}{kT}\}. \tag{3.3}$$

Often, the function W is further specified as

$$\begin{cases} W(\psi) &= 0 & \text{if } N(\psi) \neq 2 \\ W(\{x, y\}) &= \theta(\|x - y\|) \end{cases} \tag{3.4}$$

leading to the formula

$$f_n(\phi) = \frac{1}{Z} \prod_{1 \le i < j \le n} h(\|x_i - x_j\|). \tag{3.5}$$

The function θ is called the *pair potential function*, h is called the *interaction function.* The normalizing factor Z is called the *partition function.* The partition function is a very important quantity in statistical physics, since it describes the macroscopic properties temperature and pressure of the system. Its value is obtained by integration over all possible configurations:

$$Z = \int_{V^n} \exp\{-\frac{U_n'(\phi)}{kT}\} d\phi = \int_V \cdots \int_V \exp\{-\frac{U_n'(x_1, \ldots, x_n)}{kT}\}\, dx_1 \ldots dx_n.$$

A density formula as in (3.2) can be shown to arise using arguments from physics. Starting with an energy function U_n' and a density function g, in equilibrium, the system of particles is required to have a fixed total energy

$$\mathcal{E}_g = \int_{V^n} U_n'(\phi) g(\phi) d\phi.$$

The extra condition of maximal entropy

$$\mathcal{H}_g = -\int_{V^n} g(\phi) \log g(\phi) d\phi$$

leads to the above density formula f_n.

Note that we get a uniform density if no interaction takes place. Then, the energy function U_n' is zero, yielding $f_n(\phi) = 1/|V|^n$.

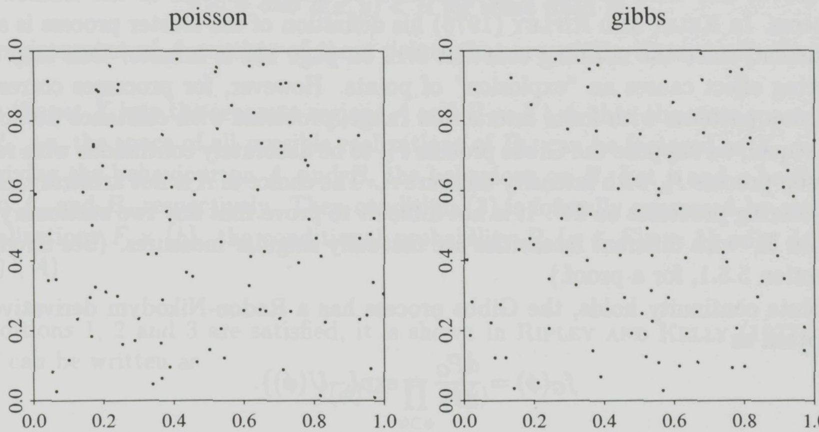


Figure 3.1: Realization of Poisson and Gibbs process, with 100 points

In our model of repulsive point patterns on a bounded region $A \subset \mathbb{R}^2$, the description and terminology from statistical physics is adopted, apart from the terms k and T. So we suppose that the point configuration has arisen after an adjustment of points, such that an equilibrium situation is attained. The density $f_n \colon A^n \to [0, \infty)$ at a configuration $\phi = (x_1, \dots, x_n)$ is assumed to be given by (3.5). Our aim is to estimate the interaction function $h = \exp\{-\theta\}$, by performing the maximum likelihood procedure on the density function f_n. As an illustration, in figure 3.1 two samples of 100 points on the unit square are given. The first is a completely random configuration of points. The second is a realization from a Gibbs distribution with interaction function

$$h(x) = \begin{cases} 1 - (1 - (20x)^2)^2 & \text{if } x \leq 0.05 \\ 1 & \text{if } x > 0.05 \end{cases}$$

Note that the point configuration for the second case is slightly more regular.

Before we continue with the estimation procedure, a section is spent on the justification of the Gibbs model.

3.3 Some justification for the Gibbs distribution

Just like in the case of point configurations without interaction, the total number of points in the bounded region A may either be finite and fixed, or random. For a gas in a closed medium, the number of molecules is clearly finite and fixed, but when modelling nesting behaviour, this assumption is violated. So we have to allow for a varying number of points. We first consider this approach, although finally, for reasons of simplicity, our estimation procedure will be based on the conditional model, with $N(A) = n$.

When considering general Gibbs point processes, problems with existence may arise. STRAUSS (1975) has tried to cast a special kind of cluster process in the framework of a Gibbs process. In KELLY AND RIPLEY (1976) his definition of the cluster process is shown to be non-existent, since the norming constant 3.12 on page 112 is infinite. One may say that the clustering effect causes an "explosion" of points. However, for processes corresponding with repulsion patterns with finite interaction range, problems with existence do not occur.

In the sequel, we suppose the Gibbs process P_G to be absolutely continuous with respect to some Poisson process P_Λ with intensity measure Λ. The choice of Λ is not arbitrary, especially when considering processes on $\mathbb{R}^2$. It is not difficult to prove that any two stationary Poisson processes on $\mathbb{R}^2$ with different intensities are mutually singular measures. (See STOYAN *et al.* (1987), section 5.5.1, for a proof.)

If absolute continuity holds, the Gibbs process has a Radon-Nikodym derivative, which can be written as

$$f_G(\phi) = \frac{dP_G}{dP_\Lambda} = \exp\{-U(\phi)\}. \qquad (3.6)$$

Note that $U(\phi)=\infty$ (configuration impossible under Gibbs process) and $U \equiv 0$ (Gibbs process is Poisson process P_Λ) are allowed.

3.3.1 Markov point processes: product structure

If we impose some extra conditions, a multiplicative structure like in (3.3) can be derived and related to a Markov property in higher dimensions, as is described in RIPLEY AND KELLY (1977). Thereby the *exponential space* approach of point processes is used. Let X be a, not necessarily bounded, region. The exponential space can be seen as a union of classes $X_n, n = 0, 1, \ldots$, where a typical element of X_n consists of n elements from X. So only a finite number of points is allowed to occur. Note that a configuration is seen as an unordered set. A point process P_Φ is a *Markov point process* if the following holds:

1. P_Φ is absolutely continuous with respect to some Poisson process P_Λ with *finite* intensity measure Λ.

2. The set $H = \{f > 0\}$ is hereditary. A set is called *hereditary* if both $\phi \in H$ and $\psi \subset \phi$ imply $\psi \in H$. So, if both $f(\phi) > 0$ and $\psi \subset \phi$, then $f(\psi) > 0$ should hold as well.

3. If the point configuration is known on a subset $B \subset X$, then the behaviour of the process on $A = X \setminus B$, given the configuration on B, only depends on those points in B that are within the interaction range of A.

For a repulsion process, existence of a dominating Poisson process and $\{f > 0\}$ being hereditary are natural assumptions.

Condition 3. can be expressed in a formal way as follows. Suppose we have a measurable symmetric reflexive relation $\sim$ on X. We say that two points $x, y \in X$ are *neighbours* if $x \sim y$. The *environment* $E(A)$ of $A \subset X$ is defined by

$$E(A) = \{x | x \sim y \text{ for some } y \in A\}.$$

An example of a neighbourhood relation is

$$x \sim y \Longleftrightarrow d(x, y) < R \text{ for some fixed } R$$

and the environment of A consists of those points that are within distance R of some point in A.
If we split the set X into the separate regions A and $B = X \setminus A$, then the state space X_e of the process P_Φ, i.e. the space of all possible realizations of P_Φ, can be factored as $X_e = A_e \times B_e$, with A_e giving the behaviour on A, and B_e the behaviour on B. Let p and q be the projections onto A_e and B_e respectively. Then condition (3) is formally expressed by saying that, for all realizations $F \times \{b\}$, the conditional probability $P_\Phi\{p \in F | q = b\}$ only depends on $b \cap (E(A) \setminus A)$.

If conditions 1, 2 and 3 are satisfied, it is shown in RIPLEY AND KELLY (1977) that the density f can be written as

$$f(\phi) = \prod_{\psi \subset \phi} g(\psi). \tag{3.7}$$

Here g is a non-negative function defined on all finite sets of points, including the empty set, having the property that $g(\psi) \neq 1$ implies that all points of ψ are neighbours, i.e. are within each other's interaction range. Hence g: $X_e \to [0, \infty)$ is a function satisfying

$$g(\psi) \neq 1 \implies x \sim y \text{ for all } x, y \in \psi.$$

Our repulsive point patterns are seen as generated by a Markov point process, so, writing $g(\psi) = \exp\{-V(\psi)\}$, we indeed get the multiplicative structure as in (3.3). The function U in (3.6), which is playing the role of the energy function of statistical physics, can be written as a sum of interaction potentials

$$U(\phi) = \sum_{\psi \subset \phi} V(\psi),$$

with $V(\psi) = 0$ if (at least) two points in ψ are outside each other's interaction range.

Note that the intensity measure is supposed to be finite, hence this construction fails if we consider a stationary Poisson process on $\mathbb{R}^2$ as dominating measure of the Markov point process.

3.3.2 Pairwise interaction processes

In a molecular gas, only *mutual* interactions play a role. This means that the amount of repulsion or attraction between, say, 3 points, is the sum of the repulsive or attractive forces between each pair of molecules separately. Of course, birds are not molecules, and the presence of two nests around some location may make it *extra* unlikely for a bird to nest there, if the "threat" of two nests is greater than the sum of the "threats" of each nest separately. For the sake of simplicity, however, it is supposed that a bird is only influenced by each nest separately. This means that only interaction potentials depending on at most 2 points do matter. Hence we can write, for a configuration of points $\phi = \{x_1, \ldots, x_{N(\phi)}\}$,

$$U(\phi) = c + \sum_i w(x_i) + \sum_{i<j} z(x_i, x_j).$$

Now we return to our motion invariance assumption. The nest sites are supposed to be situated in some homogeneous habitat. Expanding this habitat to an area on which a group of translations and rotations can be defined, like $\mathbb{R}^2$, homogeneity can be modelled by assuming the process to be stationary and rotation invariant.

Of course, this means that we suppose the point configuration inside the bounded region A to depend on the location of points outside A, which in reality are not present. So in reality less repulsive influence is expected near the edge, but the absence of repulsive forces from points outside A may partially be compensated by a repulsive influence from the edge itself. For example, there may be a preference for birds not to nest too close to the edge of their habitat, especially if their breeding area is also their food-supply area. So the problem may be treated as if points outside A do occur, but cannot be observed. This means that some edge correction procedure will have to be applied.

When considering a motion invariant Gibbs process, a homogeneous Poisson process P_λ with intensity λ is the most natural dominating measure. However, in the Markov process approach of the previous subsection, the dominating process was assumed to have finite intensity measure. So the Poisson process on $\mathbb{R}^2$ cannot be taken as dominating measure. However, we only consider repulsive forces with finite interaction range. Instead of defining the Gibbs process on $\mathbb{R}^2$, we define it on a large region $E \supset A$. We let the process on E be independent of the choice of the origin and coordinate axes. Near the edge of E, there will be an outward force working on the points. However, this influence will be negligible for points in A, so there the Gibbs process still behaves like a stationary process on $\mathbb{R}^2$.

Taking the homogeneous Poisson process on E as dominating measure, w becomes a constant function. Since each distance preserving transformation is a combination of a translation and a rotation, z only depends on the distance between the points. So, under these assumptions, $f = dP_G/dP_\lambda$ in formula (3.7) can be simplified to:

$$f_G(\emptyset) = g(\emptyset) = C,$$

$$f_G(\{x\}) = g(\emptyset) \times g(\{x\}) = C\beta,$$

and for $\phi = \{x_1, \ldots, x_{N(\phi)}\}$ we have:

$$f_G(\phi) = C\,\beta^{N(\phi)} \prod_{i<j} h(\|x_i - x_j\|). \tag{3.8}$$

The above considerations were used to motivate the structure in (3.8). Depending on the situation under consideration, either we use this construction and see A as an observation window for a process on a larger region E, or we just assume (3.8) and we see A as is. In the rest of this section, the last approach is taken. However, one could as well read E instead of A and see A as an observation window.

C is equal to $P_G\{N(A) = 0\}/P_\lambda\{N(A) = 0\}$. C depends on h via an infinite number of high dimensional integrals. Using $\int f_G(\phi)dP_\lambda(\phi) = 1$, we get, writing b instead of $\beta\lambda$,

$$\exp\{\lambda|A|\}\, C^{-1} = 1 + b|A| + \sum_{n=2}^{\infty} \frac{b^n}{n!} \int_{A^n} \prod_{i<j} h(\|x_i - x_j\|)\, dx_1 \ldots dx_n. \tag{3.9}$$

C should satisfy $0 < C < \infty$. A sufficient condition for this to hold is the following inequality

$$\prod_{1 \le i < j \le n} h(\|x_i - x_j\|) \le \mathrm{e}^{K\,n},$$

which is automatically fulfilled for repulsion processes, but may be problematic for Gibbs models with attractive forces.
Conditioning on $N_G(A) = n$, we get the density

$$f_G(\phi | N(A) = n) = \frac{C\,\beta^n \prod_{i<j} h(\|x_i - x_j\|)}{P_G\{N(A) = n\}}. \tag{3.10}$$

The function h can be seen to express the likelihood of points to be at interpoint distance r, relative to a completely random configuration of points. The stronger the repulsive forces at some interpoint distance, the smaller the value of h. For $h(r) < 1$, we have repulsion at distance r, with total inhibition if $h(r) = 0$. $h(r) = 1$ means no interaction, whereas $h(r) > 1$ means attractive influence. For example, in the so-called hard core inhibition process, we have $h(r) = 0$ for $0 \le r \le R_0$, and $h(r) = 1$ for $r > R_0$. This process can be used to model the configuration of balls, each having diameter R_0, but without any further repulsive or attractive forces among each other.

Up till now we have written the density of the Gibbs process as a Radon-Nikodym derivative with respect to a homogeneous Poisson process with intensity λ. It is also possible to write down the density with respect to Lebesgue measure. Multiplying both numerator and denominator in formula (3.8), using C as in (3.9), by a factor $\lambda/(n!)$, a term $\exp\{-\lambda\,|A|\}\,\lambda/(n!)$ is obtained in the denominator. This term can be seen as the density $f^*(\phi)$ at an ordered point configuration $\phi = (x_1, \ldots, x_n)$ with respect to Lebesgue measure under the Poisson assumption. For

$$f^*(\phi) = P\{N(A) = n\} \times \frac{1}{|A|^n} = \exp\{-\lambda\,|A|\}\,\frac{\lambda}{n!}.$$

The factor λ appearing in the numerator is absorbed into β to obtain b, whereas the factor $1/(n!)$ is absorbed by considering ordered point configurations instead of unordered ones. Now we obtain

$$f_G^*(\phi) = C\, b^{N(\phi)} \prod_{i<j} h(\|x_i - x_j\|), \tag{3.11}$$

with C^{-1} equal to

$$C^{-1} = 1 + b|A| + \sum_{n=2}^{\infty} \frac{b^n}{n!} \int_{A^n} \prod_{i<j} h(\|x_i - x_j\|)\, dx_1 \dots dx_n. \tag{3.12}$$

For the conditional case, b is absorbed by the norming constant and the density can be written as in (3.5).

This density will be used as input for the maximum-likelihood procedure. It has the advantage that we lose the influence of the dominating Poisson measure in the formula of the density: we are no longer confronted with the problem that the value of λ occurs as a free variable in $\exp\{-\lambda\,|A|\}$.

3.4 An approximation to the likelihood function

Using formula (3.11) for the density of a Gibbs process, we can, in principle, perform maximum likelihood estimation in order to estimate the interaction function h, using a single instantaneous spatial realization of points $(x_1, \dots, x_{N(\phi)})$ in equilibrium. In our model we assume only repulsive forces between points to occur, with the amount of repulsion diminishing with increasing distance. Moreover we assume the interaction range to be finite. This results in the following side conditions for the maximization procedure:

(S1): $0 \le h \le 1$, $h(0) = 0$, and $h(R) = 1$ for some R

(S2): h is increasing.

h only depends on the interpoint distances $\|x_i - x_j\|$, which will be denoted by d_{ij}. The ordered interpoint distances are written as $d_{(i)}$ for $i = 1, \dots, M = \binom{N(\phi)}{2}$. Note that $b > 0$. A direct implementation of the maximization procedure is an impossible task, due to the very complicated structure of C as a function of h. However, one thing can be said about the maximum-likelihood estimator under the side conditions (S1) and (S2). From the structure of the likelihood we can derive that the maximizing $\hat{h}$ is a piecewise constant function.

Theorem 3.4.1 *Under the side conditions (S1) and (S2), the density*

$$f_G(\phi) = \frac{b^{N(\phi)} \prod_{i<j} h(d_{ij})}{1 + b|A| + \sum_{n=2}^{\infty} \frac{b^n}{n!} \int_{A^n} \prod_{i<j} h(d_{ij}) dx_1 \dots dx_n}$$

is maximized by a piecewise constant function h, with all jumps contained in a subset of the set of interpoint distances $\{d_{ij}\}$.

Proof:
Maximizing the expression in (b, h) under the side condition $0 \le h(d_{(1)}) \le \dots \le h(d_{(M)}) \le 1$, we have to cope with two opposite effects with respect to h. Maximization of the numerator will make $h(d_{(i)})$ as large as possible, whereas the denominator forces h towards zero. For the numerator, only the values at the points $d_{(i)}$ do matter, so the denominator forces $h(d)$ equal to $h(d_{(i)})$ on $[d_{(i)}, d_{(i+1)})$. Since this holds for any $b > 0$, we obtain the desired result.

Due to the complicated structure of the normalizing constant, attention will be restricted to maximum likelihood estimation, conditionally on the observed number of points $N(A) = n$. If the phenomenon under study has a random number of points, this is only a valid procedure if the number of points $N(A)$ is approximately ancillary for h. We arrive at the following procedure:

Given a realization $x_1, \ldots, x_n$, maximize

$$L_n(h) = \frac{1}{Z_h} \prod_{i<j} h(\|x_i - x_j\|),$$

with Z_h given by

$$Z_h = \int_{A^n} \prod_{i<j} h(\|x_i - x_j\|) dx_1 \ldots dx_n,$$

under the side conditions (S1) and (S2).

The conditional likelihood is also maximized by a stepfunction, with all jumps contained in a subset of the set of interpoint distances $\{d_{ij}\}$.

We are still left with the problem that the partition function Z_h cannot be computed, but some approximation will be used. A frequently used technique, with its origin in statistical physics, is the *Mayer cluster expansion.* In the Mayer cluster expansion the function h is expanded around the value 1. Let $g = h - 1$. Then we obtain

$$\begin{aligned} Z_h &= \int_{A^n} \prod_{i<j} \{1 + g(d_{ij})\} dx_1 \ldots dx_n \\ &= \int_{A^n} \{1 + \sum_{i<j} g(d_{ij}) + \sum_{\substack{i<j \\ (i,j)\neq(k,l)}} \sum_{k<l} g(d_{ij}) g(d_{kl}) + \ldots\} dx_1 \ldots dx_n \end{aligned}$$

The first term gives $|A|^n$. For the second term we obtain

$$\begin{aligned} \sum_{i<j} \int_{A^n} g(d_{ij}) dx_1 \ldots dx_n &= \binom{n}{2} \int_{A^n} g(d_{12}) dx_1 \ldots dx_n \\ &= \binom{n}{2} |A|^{n-2} \int_A \{ \int_A g(d_{12})\, dx_2 \} dx_1 \\ &\approx \binom{n}{2} |A|^{n-1} \int_0^R (h(r) - 1) 2\pi r\, dr \end{aligned}$$

In the last approximation it is assumed that we can take a fixed point, say x_1, from which integration over another variable, say x_2, is performed. This is approximately true if the interaction area πR^2 is of a lower order of magnitude than the area of the total region.

A derivation of higher order terms is given in Ripley (1988). This higher order expansion contains integrals that are hard to compute. For some special parametric models, fairly accurate expressions for these integrals have been obtained. Since our approach is nonparametric, and no general formula exists for the higher order terms, this method is not applicable. Another approach is to use Markov Chain Monte Carlo methods. In OGATA AND TANEMURA (1984) and OGATA AND TANEMURA (1989), some parametric models which are characterized by a scale parameter are considered. A Monte Carlo approximant to the derivative, with respect to the scale parameter, of the logarithm of the norming constant is used. Another option is to use a Monte Carlo approximant to the norming constant itself (GEYER AND THOMPSON (1992)). It may be possible to implement this approach under our order restriction. We did not investigate this, however. We will truncate the expansion at the second term, which may be a reasonable approximation if the process is close to Poisson.

3.5 The nonparametric estimation procedure

We use one further approximation: $\log(1+x) \approx x$. If πR^2 is of lower order of magnitude than $|A|/M$, x is close to zero and the approximation is good. Then the approximate conditional log-likelihood becomes, with $M = \binom{n}{2}$:

$$\log L_n(h) = \sum_{i=1}^{M} \log h(d_{(i)}) - n \log |A| - M \frac{2\pi}{|A|} \int_0^R (h(r) - 1) r \, dr \tag{3.13}$$

Of course we do not know the value of R, but since we assume the interaction area to be of lower order than the total area, we can safely assume $h(d_{(M)}) = 1$. The last term in (3.13) forces $h(r) = 0$ for $r < d_{(1)}$ and theorem 3.4.1 yields $h(r) = h(d_{(i)})$ for $d_{(i)} \leq r < d_{(i+1)}$. So the problem reads, writing $y_i = h(d_{(i)})$:

Maximize

$$\sum_{i=1}^{M-1} \log y_i - M \frac{\pi}{|A|} \sum_{i=1}^{M-1} [y_i - 1] \, [d^2_{(i+1)} - d^2_{(i)}] = \sum_{i=1}^{M-1} \log y_i - \sum_{i=1}^{M-1} (y_i - 1) a_i \tag{3.14}$$

under the side condition $0 \leq y_1 \leq \ldots \leq y_{M-1} \leq 1$.

The maximization problem can be seen as a generalized isotonic regression problem (see ROBERTSON *et al.* (1988), section 1.5). Our problem boils down to maximizing their formula (1.5.4), with $\Phi(y) = y \log y + 1$. As weight function we take $w(x_i) = a_i$ and for g we take $g(x_i) = 1/a_i$. Their theorem 1.5.1 says that maximizing

$$\sum_{i=1}^{M-1} \log y_i - \sum_{i=1}^{M-1} y_i a_i$$

under the isotony restriction $0 \leq y_1 \leq \ldots \leq y_{M-1}$ is the same as minimizing

$$\sum_{i=1}^{M-1} \left\{ y_i - \frac{1}{a_i} \right\}^2 a_i$$

under the same restriction. Isotonic regression theory gives a general procedure for finding the function $\tilde{y}$ that minimizes $\sum_{i=1}^{N}\{y_i - g_i\}^2 w_i$ over a class of isotonic functions on $\{1, 2, \ldots, n\}$. This procedure reads:

Plot the points $P_0 = (0,0)$ and $P_j = (W_j, G_j)$, $j = 1, 2, \ldots, N$, with , $W_j = \sum_{i=1}^{j} w_i$ and $G_j = \sum_{i=1}^{j} w_i g_i$. Connect these points by a piecewise linear function. The function we get is called the cusum-diagram (cumulative sum diagram). Then $\tilde{y}_i$ is equal to the left derivative of the greatest convex minorant of this function at the point i.

In our situation the cusum-diagram is generated by the points

$$(W_j, G_j) = (M\pi \frac{d_{(j+1)}^2 - d_{(1)}^2}{|A|}, j),$$

which gives the same result with respect to $\tilde{y}_i$ as considering the points

$$(\pi \frac{d_{(j+1)}^2 - d_{(1)}^2}{|A|}, \frac{j}{M}). \tag{3.15}$$

So the estimator $\hat{y}_i = \hat{h}(d_{(i)})$ is obtained by computing one over the relative increase in circle area,

$$\left(\frac{\pi d_{(i+1)}^2 - \pi d_{(i)}^2}{|A|} \Big/ \frac{1}{M} \right)^{-1}, \tag{3.16}$$

under the side restriction that this relative increase is isotonic.

If we want to use some edge correction procedure, a natural correction is to discard the amount of circle area not covered by the region A. Let $C(x, r)$ denote the circle with midpoint x and radius r. If d_{ij} is the distance between the points p_i and p_j, πd_{ij}^2 is replaced by

$$0.5\,[\,\text{area}\{C(p_i, d_{ij}) \cap A\} + \text{area}\{C(p_j, d_{ij}) \cap A\}\,] \tag{3.17}$$

as input to the maximum likelihood procedure.

Note that, for indices near M, both coordinates in (3.15) are approximately one. These high indices will have to be neglected in order to get a reasonable procedure. We correct this in the algorithm, by only considering values $d_{(i)}$ which are much smaller than the maximal interpoint distance. This means that we assume R to be smaller than some predetermined value, much smaller than the largest interpoint distance.

3.6 Consistency

We will prove uniform consistency of the approximate NPMLE in the situation that the true process is without interaction. For processes with repulsive forces, the approximate NPMLE is not consistent, since higher order terms have been neglected in the Mayer cluster expansion. The simulations in the next section confirm this. Moreover, the approximation $\log(1+x) \approx x$, made at the beginning of section 3.5, is only reasonable for processes that are close to a Poisson process. In the next section, we provide some simulation results, showing the behaviour of the estimator under a Gibbs process with repulsion.

Let A_m be a window, growing with m, through which the process is seen. Suppose that $N(A_m) = n$. Let $\phi = (x_1, \ldots, x_n)$ and let the empirical measure $1/(n\,(n-1)) \sum_{i \neq j} I_{(x_i, x_j)}$ be denoted by G_n. If we assume R to be smaller than some predetermined value, we can ignore edge effects and edge corrections asymptotically, since the number of points within this predetermined distance from the boundary is asymptotically negligible compared to the total number of points. Then the approximate log-likelihood LL_n, conditionally on $N(A_m) = n$, and divided by $M = n\,(n-1)/2$, can be written as

$$\begin{aligned} LL_n(h) &= \log\, L_n(h) \\ &= \iint \log(\, h(d(x,y)\,)\, dG_n(x,y) - |A_m|^{-2} \iint [h(d(x,y)) - 1]\, dxdy - \frac{n \log |A_m|}{M} \end{aligned}$$

Let $\hat{h}_n$ denote the approximate NPMLE. Since

$$\lim_{\epsilon \downarrow 0} \epsilon^{-1} \left[LL_n((1-\epsilon)\hat{h}_n + \epsilon h_0) - LL_n(\hat{h}_n)\right] \leq 0, \tag{3.18}$$

we obtain

$$\begin{aligned} \iint \frac{h_0(d(x,y))}{\hat{h}_n(d(x,y))}\, dG_n(x,y) + |A_m|^{-2} \iint \hat{h}_n(d(x,y))\, dxdy \\ \leq\; 1 + |A_m|^{-2} \iint h_0(d(x,y))\, dxdy. \end{aligned} \tag{3.19}$$

We now restrict ourselves to the situation without interaction. Conditionally on the number of points $N(A_m) = n$, we have a uniform distribution of n points on A_m. By formula (3.16), we see that, given a realization ϕ, shrinking the area A_m and the distance between the points by a factor $\sqrt{|A_m|}$ leads to exactly the same estimate. Moreover, for the uniform process, such a realization on the shrunken area has the same probability to occur, since, for $B \subset A$,

$$P\{N(B) = k\} = \binom{n}{k} \left(\frac{|B|}{|A|}\right)^k \left(1 - \frac{|B|}{|A|}\right)^{n-k},$$

which only depends on the relative size of B. So, without loss of generality, we may assume the n points to be uniformly distributed on a fixed region A, with the number of points in A increasing to infinity. Let G denote the uniform distribution on A. Formula (3.19) now reads:

$$\iint \frac{1}{\hat{h}_n(d(x,y))}\, dG_n(x,y) + \iint \hat{h}_n(d(x,y))\, dG(x) \times G(y) \leq 2. \tag{3.20}$$

The consistency proof is similar to the method used in the consistency proofs of the NPMLE with interval censored data (chapter 4 in part II of Groeneboom and Wellner (1992)). Let Ω be the sample space of all infinite sequences of points $X_1, X_2, \ldots$, endowed with the Borel σ-algebra and the product measure, which is denoted by $\mathbb{P}$. The following will be shown

Theorem 3.6.1 *For each $\epsilon > 0$*

$$\mathbb{P}\left\{\lim_{n\to\infty}\sup_{t\in[\epsilon,\infty)}|\hat{h}(t)-1|=0\right\}=1$$

Remark. Since the approximate NPMLE always has $\hat{h}_n(0) = 0$, we can only have consistency on an interval $[\epsilon, \infty)$.

Proof:
Let ω denote a point in the sample space. We write $G_n(\cdot,\cdot\,;\omega)$ instead of $G_n(\cdot,\cdot)$, in order to indicate dependence on ω. By the strong law of large numbers,

$$\mathbb{P}\left\{\lim_{n\to\infty}\int f\,dG_n = \int f\,dG\times G\right\}=1,$$

for each bounded continuous function on $A \times A$. By separability of the space of bounded continuous functions on $A \times A$ with respect to the supremum norm, we have that $G_n(\cdot,\cdot\,;\omega)$ converges weakly to $G(\cdot)\times G(\cdot)$, for each ω in a set B, occurring with $\mathbb{P}$-probability one.

Let ω be a realization in this set. By the Helly compactness theorem, the sequence $\{\hat{h}_n(\cdot\,;\omega)\}$ has a weakly converging subsequence, say $\{\hat{h}_{n_k}(\cdot\,;\omega)\}$. Let $h(\cdot\,;\omega)$ denote the right continuous limit. $h(\cdot\,;\omega)$ has its values in $[0,1]$.

Let $\epsilon > 0$. For all n sufficiently large, $1/\hat{h}_n(d;\omega)$ is bounded for $d \geq \epsilon$. This follows from the weak convergence of $G_n(\cdot,\cdot\,;\omega)$ to $G(\cdot)\times G(\cdot)$, together with the inequality (3.20). By the weak convergence of $\hat{h}_{n_k}(\cdot\,;\omega)$ to $h(\cdot\,;\omega)$, $1/h(\cdot\,;\omega)$ is bounded on $[\epsilon,\infty)$ as well. So we may assume, for each $d \geq \epsilon$ and for some $K < \infty$,

$$1/\hat{h}_n(d;\omega) \leq K$$

and

$$1/h(d;\omega) \leq K.$$

Let D_ϵ denote the set $\{\,(x,y)\,|\,d(x,y)\geq\epsilon\,\}$. Now we have

$$\lim_{k\to\infty}\iint_{D_\epsilon}\frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\,dG_{n_k}(x,y;\omega)=\iint_{D_\epsilon}\frac{1}{h(d(x,y);\omega)}\,dG(x)\times G(y).$$

This is shown in essentially the same way as in lemma 4.1 in Groeneboom and Wellner (1992). Fix $0 < \delta < 1$. Take an equidistant grid of points $\epsilon = d_0 < d_1 < \ldots < d_m = R$, such that $m = 1 + [1/\delta^2]$. Without loss of generality, we may assume the d_i to be points of continuity of $h(d;\omega)$. Let I be the set of indices i, $i = 1,\ldots,m$ such that

$$\frac{1}{h(d_{i-1};\omega)}-\frac{1}{h(d_i;\omega)}\geq\delta.$$

The number of such points is not bigger than $1 + [K/\delta]$. Let J denote the remaining set of indices. Let the set $\{\,(x,y)\,|\,d_0 \leq d(x,y) \leq d_1\,\}$ be denoted by D_0 and let the sets $\{\,(x,y)\,|\,d_{i-1} < d(x,y) \leq d_i\,\}$ be denoted by D_i. We have

$$\begin{aligned}\iint_{D_\epsilon} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG_{n_k}(x,y;\omega) &= \sum_{i=0}^{m} \iint_{D_i} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG_{n_k}(x,y;\omega)\\ &= \sum_{i\in I} \iint_{D_i} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG_{n_k}(x,y;\omega)\\ &\quad + \sum_{i\in J} \iint_{D_i} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG_{n_k}(x,y;\omega)\\ &= \iint_{D_\epsilon} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG(x)\times G(y) + r_k(\omega),\end{aligned}$$

with $r_k(\omega) \le c\,\delta$, for a constant $c > 0$. This is because the integrand is of bounded variation on D_i, for $i \in J$, whereas

$$\sum_{i\in I} \iint_{D_i} dG(x)\times G(y) \to 0, \text{ if } \delta \downarrow 0.$$

By dominated convergence, we derive

$$\lim_{k\to\infty} \iint_{D_\epsilon} \frac{1}{\hat{h}_{n_k}(d(x,y);\omega)}\, dG(x)\times G(y) = \iint_{D_\epsilon} \frac{1}{h(d(x,y);\omega)}\, dG(x)\times G(y).$$

By dominated convergence, we moreover have

$$\lim_{k\to\infty} \iint_{D_\epsilon} \hat{h}_{n_k}(d(x,y);\omega))\, dG(x)\times G(y) = \iint_{D_\epsilon} h(d(x,y);\omega)\, dG(x)\times G(y).$$

Combining these results, using (3.20), we obtain

$$\iint_{D_\epsilon} \frac{1}{h(d(x,y);\omega)}\, dG(x)\times G(y) + \iint_{D_\epsilon} h(d(x,y);\omega)\, dG(x)\times G(y) \le 2.$$

By monotone convergence, we derive

$$\iint \frac{1}{h(d(x,y);\omega)}\, dG(x)\times G(y) + \iint h(d(x,y);\omega)\, dG(x)\times G(y) \le 2.$$

The function

$$\frac{1}{y} + y$$

is minimal at $y = 1$, taking the value 2. Using the monotonicity of h, this implies that the inequality can only hold for $h \equiv 1$. Moreover, we have equality in this case. Since, for each subsequence h_{n_k} we have a weakly converging sub-subsequence, all converging to the same limit, we obtain weak convergence of $h_n(\cdot,\cdot;\omega)$ to $h_0 \equiv 1$, for each ω in the set B, occurring with probability one. This is the same as

$$\mathbb{P}\left\{\lim_{n\to\infty} \sup_{t\in[\epsilon,\infty)} |\hat{h}(t) - 1| = 0\right\} = 1,$$

for each $\epsilon > 0$. Note that 0 is not a point of continuity of $h_0 \equiv 1$.

□

Although the approximate NPMLE is not consistent for general Gibbs processes, the complete NPMLE, with the order restrictions, may very well be. A proof may be given along the lines of the above proof, but this will clearly be more complicated. For the conditional situation, with n points, using (3.18) yields the basic inequality:

$$\int\int_{x\neq y} \frac{h_0}{\hat{h}_n}\, dG_n - \frac{B_0}{B} \leq 0$$

with

$$B_0 = \int_{A^n} \prod_{\substack{i<j \\ (i,j)\neq(k,l)}} \hat{h}_n(d_{ij}) \times h_0(d_{kl})\, dx_1 \dots dx_n,$$

for some arbitrary but fixed pair of points kl, and

$$B = \int_{A^n} \prod_{i<j} \hat{h}_n(d_{ij})\, dx_1 \dots dx_n.$$

This inequality will be the basis for the different steps of the proof.

3.7 Simulations

We have performed some simulations in order to test for the behaviour of the approximate NPMLE for a fixed number of points. Consider a rectangular region A, say $A = (0,a)\times(0,b)$. Then an explicit formula exists for the edge correction formula (3.17). Let $p_i = (x_i, y_i)$, $d_1 = \min(x_i, a - x_i)$ and $d_2 = \min(y_i, b - y_i)$. Then we have (see DIGGLE (1983), p. 72):

$$\text{area}(C(p_i, r)\cap A) =$$

$$= \begin{cases} \pi r^2 - r^2[\arccos((d_1 \wedge r)/r) + \arccos((d_2 \wedge r)/r)] & \text{if } r^2 \leq d_1^2 + d_2^2 \\ 0.75\pi r^2 - 0.5 r^2[\arccos(d_1/r) + \arccos(d_2/r)] & \text{if } r^2 > d_1^2 + d_2^2 \end{cases}$$

Realizations of Gibbs point patterns have been obtained via the method described in section 3 of OGATA AND TANEMURA (1989). Choose $\delta > 0$. Starting with some point configuration $\phi(0) = \{x_1(0), \dots, x_n(0)\}$, a Gibbsian pattern with a specified interaction function h is generated via the following iterative procedure:

- **Step k:** We have the pattern $\phi(k) = \{x_1(k), \dots, x_n(k)\}$.
  - Choose one of the points $\{x_1(k), \dots, x_n(k)\}$ at random, say $x_j(k)$.
  - Choose a new point $x_j'(k)$, uniformly on the square with length $2\,\delta$ and midpoint $x_j(k)$, and let $\phi'(k)$ be the point configuration with $x_j(k)$ replaced by $x_j'(k)$.
  - Let f_n be the density of the Gibbs distribution of interest. Then the new configuration is chosen with probability

$$p = \min\left\{1, \frac{f_n(\phi_n'(k))}{f_n(\phi_n(k))}\right\} = \min\left\{1, \frac{\prod_{i\neq j} h(\|x_j'(k) - x_i(k)\|)}{\prod_{i\neq j} h(\|x_j(k) - x_i(k)\|)}\right\}$$

- **Step k+1:** We have the pattern $\phi_n(k+1)$

Note that the normalizing factor is not needed in the procedure, since it cancels in the probability $f_n(\phi'_n(k))/f_n(\phi_n(k))$.
The choice of δ determines how quickly the algorithm converges. Ogata and Tanemura (1989) and Diggle *et al.* (1994) refer to Wood (1968), who found the experimental result that a δ leading to a new point configuration about half of the times, is a reasonable choice. However, such a δ does not exist for sparse configurations.
The following simulations have been performed on the unit square:

1. Poisson process, $h_1 \equiv 1$, with $n = 500$.
2. $$h_2(x) = \begin{cases} 1 - (1 - (x/\alpha)^2)^2 & \text{if } x \le \alpha \\ 1 & \text{if } x > \alpha \end{cases}$$
   with $n = 500$, $\delta = 0.1$ and $\alpha = 0.002,\ 0.008,\ 0.02,\ 0.04$
3. A Strauss process with
   $$h_3(x) = \begin{cases} \exp\{-\alpha\} & \text{if } x \le \beta \\ 1 & \text{if } x > \beta \end{cases}$$
   with $n = 500$, $\delta = 0.1$ and $\alpha = 0.2,\ 1.75$ and $\beta = 0.008,\ 0.04$.

Formulas 2. and 3. have been investigated in Diggle *et al.* (1994) as well, in their comparison of some parametric estimation procedures. The estimates $\hat{h}$ are given in figures 3.2 to 3.6, together with the interaction function by which the points were generated.

In order to investigate the influence of sample size, we did another simulation for the function h_2, with $\alpha = 0.008$, thus

$$h_2(x) = \begin{cases} 1 - (1 - (x/0.008)^2)^2 & \text{if } x \le 0.008 \\ 1 & \text{if } x > 0.008 \end{cases}$$

We looked at a sample size of 100, 500, and 2000 points, on the square with area 0.2, 1 and 4, respectively. The results are in figure 3.7.

We only considered the model in which A is standing on itself, without influence from points outside A. Otherwise simulations can be performed by using a periodic boundary (see Diggle *et al.* (1994)). We ran the above step 2.100.000 times, evaluating the result after every 300.000 steps, starting at step 600.000. In all computations, we only considered values of the interpoint distance which were smaller than 0.126 (corresponding to a circle area with size 0.05).

Some things can be noticed. In general, the estimate is quite far from the interaction function that generated the process. Moreover, the stronger the repulsive forces, the more inaccurate the estimate. The direction of the bias, introduced by not considering higher order terms, seems to depend on the strength of the repulsive forces. In general, there is a tendency to underestimate the repulsive force in case of strong interactions, whereas there is a small tendency to overestimate the repulsive force in case of weak interactions. In figure 3.6, with parameter values $(\alpha, \beta) = (1.75, 0.04)$, all estimates are very close together. It is unclear why this is the case. It may have to do with the fact that the repulsive forces are quite strong with this choice of parameters, leading to a regular pattern, without much variation over the simulations. Note that the total area covered by the interaction forces is $500 \times \pi \times 0.04^2 = 2.512$, which is more than the total area of A.

The results of the simulation with three different sample sizes indicate that the approximate NPMLE converges to some value as the sample size increases. However, the limit is clearly not the true value.

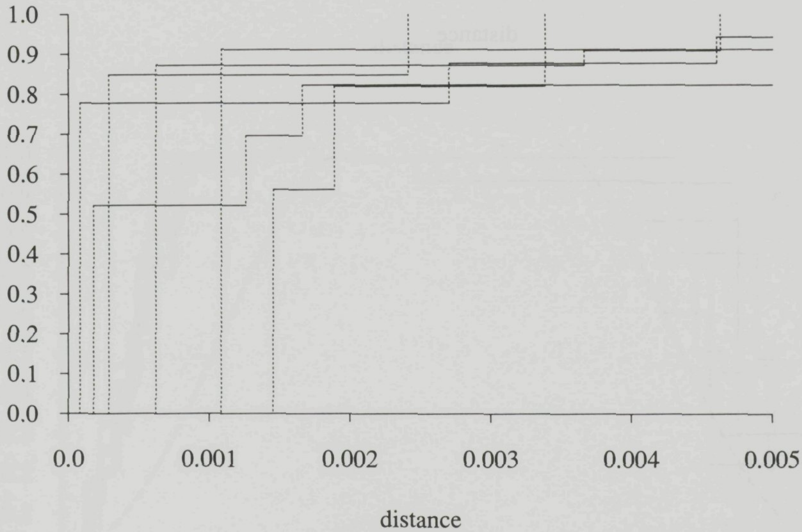


Figure 3.2: Poisson process

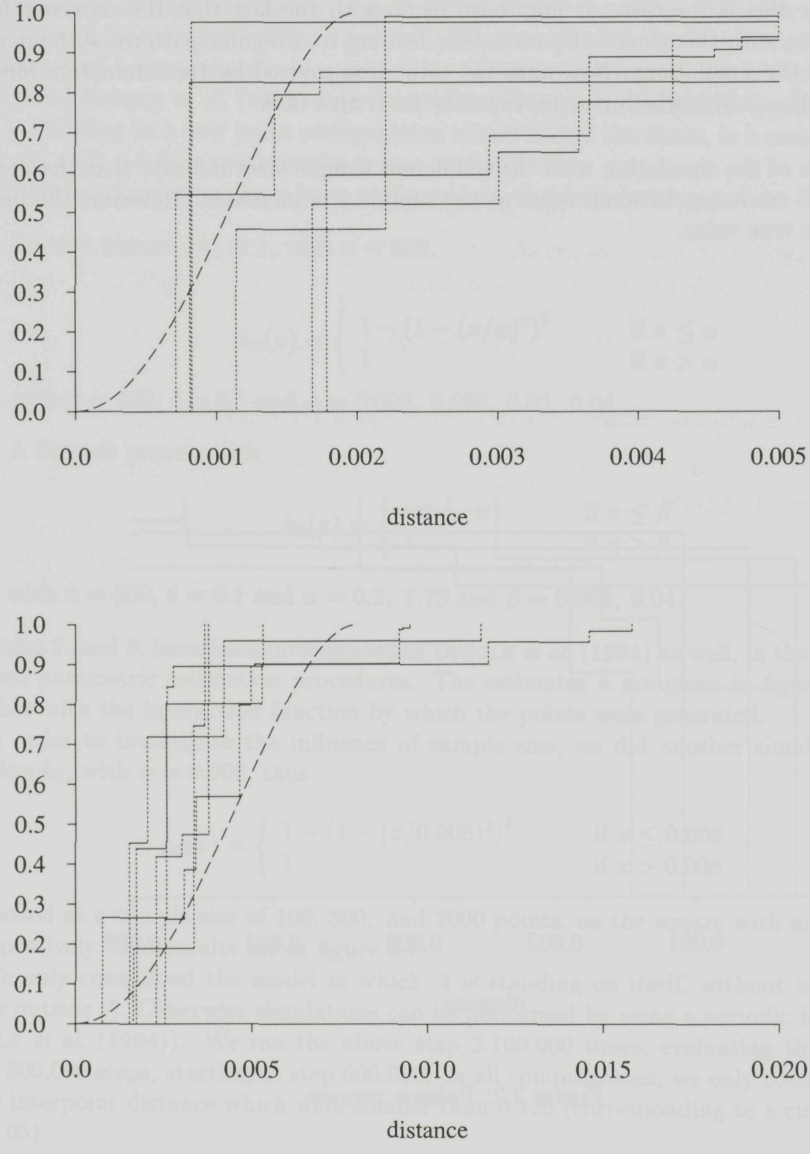


Figure 3.3: h_2, with $\alpha = 0.002$ and $\alpha = 0.008$

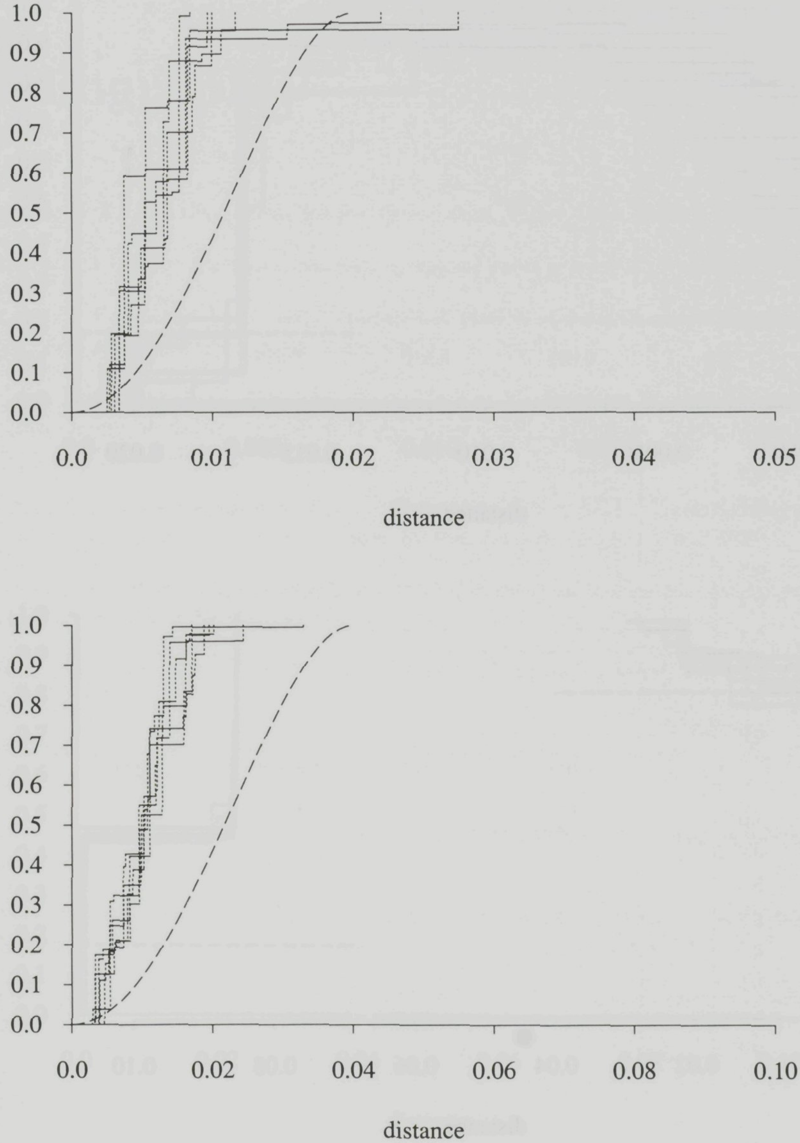
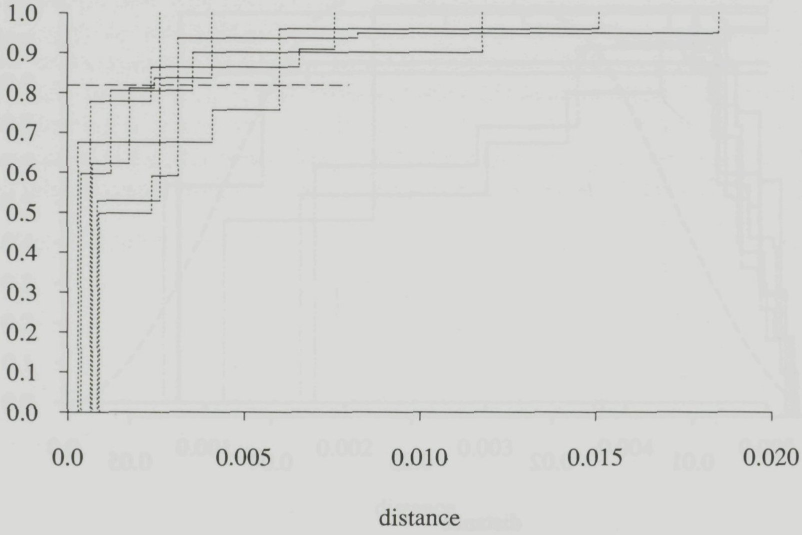


Figure 3.4: h_2, with $\alpha = 0.02$ and $\alpha = 0.04$



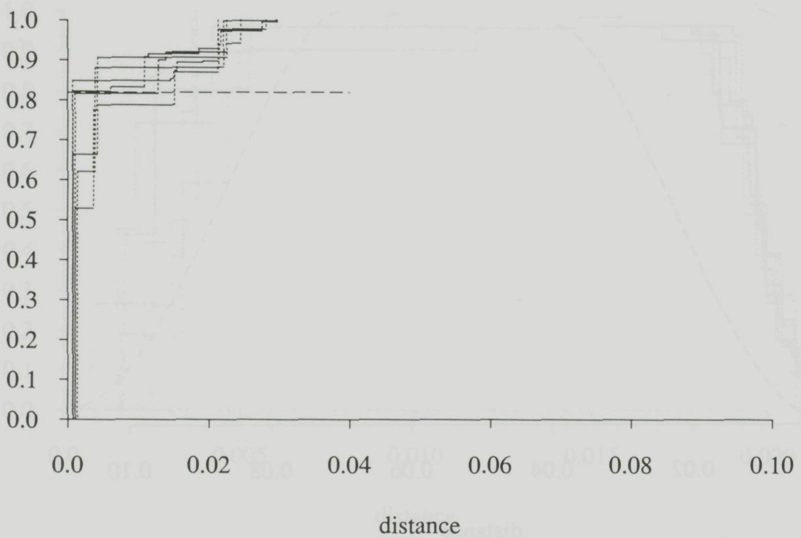
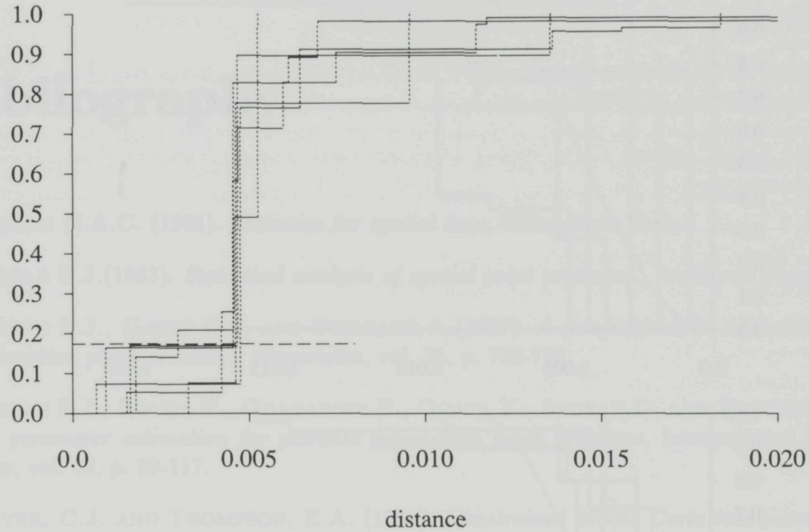


Figure 3.5: Strauss process, with $(\alpha, \beta) = (0.2, 0.008)$ and $(\alpha, \beta) = (0.2, 0.04)$



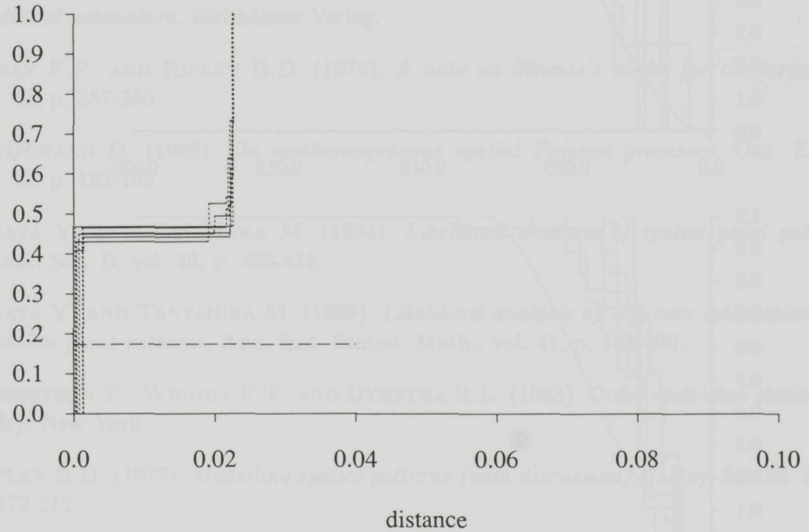


Figure 3.6: Strauss process, with $(\alpha, \beta) = (1.75, 0.008)$ and $(\alpha, \beta) = (1.75, 0.04)$

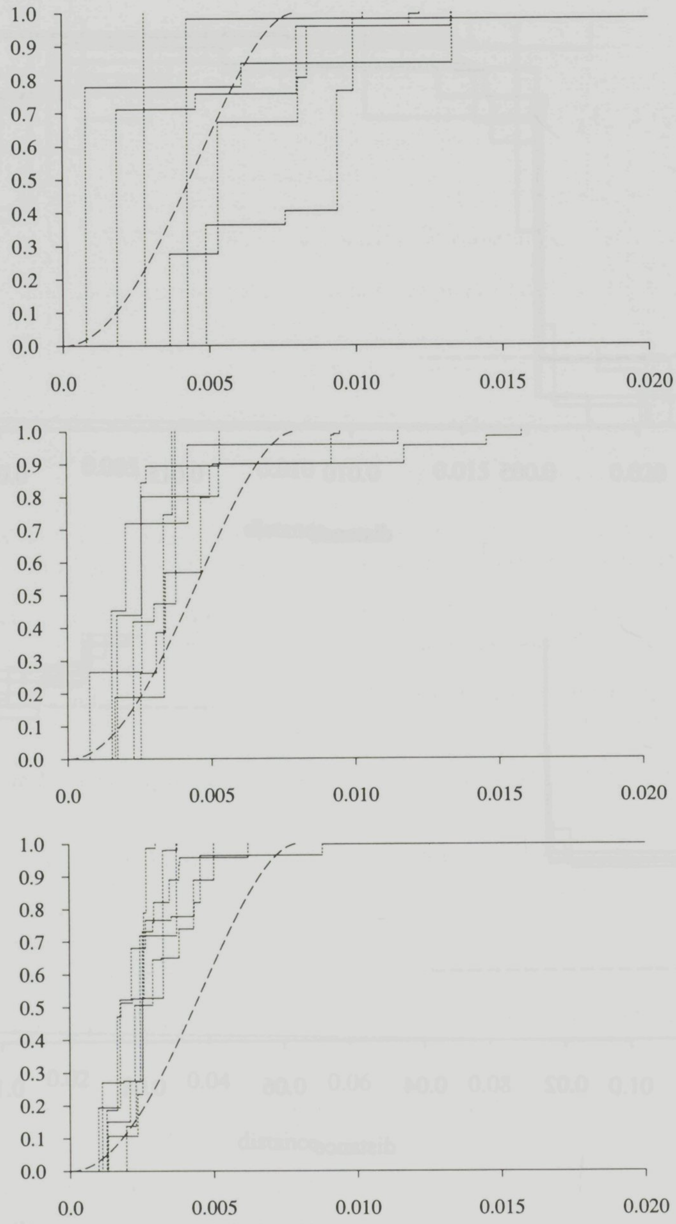


Figure 3.7: h_2, with $\alpha = 0.008$ and with 100, 500 and 2000 points

Bibliography

CRESSIE N.A.C. (1991). *Statistics for spatial data*, Wiley, New York.

DIGGLE P.J.(1983). *Statistical analysis of spatial point processes.*, Academic Press, London.

DIGGLE P.J., GATES D.J. AND STIBBARD A.(1987). *A nonparametric estimator for pairwise-interaction point processes*, Biometrika, vol. 74, p. 763-770.

DIGGLE P.J., FIKSEL T., GRABARNIK P., OGATA Y., STOYAN D. AND TANEMURA M. (1994). *On parameter estimation for pairwise interaction point processes*, International Statistical Review, vol. 62, p. 99-117.

GEYER, C.J. AND THOMPSON, E.A. (1992). *Constrained Monte Carlo maximum likelihood for dependent data (with discussion)*, J. Roy. Statist. Soc. B, vol. 54, p. 657-699.

GROENEBOOM P. AND WELLNER J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* Birkhäuser Verlag.

KELLY F.P. AND RIPLEY B.D. (1976). *A note on Strauss's model for clustering*, Biometrika, vol. 63, p. 357-360.

MCDONALD D. (1989). *On nonhomogeneous spatial Poisson processes*, Can. J. of Statistics, vol. 17, p. 183-195.

OGATA Y. AND TANEMURA M. (1984). *Likelihood analysis of spatial point patterns*, J. Roy. Statist. Soc. B, vol. 46, p. 496-518.

OGATA Y. AND TANEMURA M. (1989). *Likelihood analysis of soft-core interaction potentials for Gibbsian point patterns*, Ann. Inst. Statist. Math., vol. 41, p. 583-600.

ROBERTSON T., WRIGHT F.T. AND DYKSTRA R.L. (1988). *Order restricted statistical inference*, Wiley, New York.

RIPLEY B.D. (1977). *Modelling spatial patterns (with discussion)*, J. Roy. Statist. Soc. B, vol. 39, p. 172-212.

RIPLEY B.D. (1988). *Statistical inference for spatial processes*, Cambridge University Press, Cambridge.

RIPLEY B.D. AND KELLY F.P. (1977). *Markov point processes*, J. Lond. Math. Soc., vol. 15, p. 188-192.

STOYAN D., KENDALL W.S. AND MECKE J. (1987). *Stochastic geometry and its applications*, Wiley, New York.

STRAUSS D.J. (1975). *A model for clustering*, Biometrika, vol. 62, p. 467-475.

VAN LIESHOUT, M.N.M. AND BADDELEY, A.J. (1996). *A nonparametric measure of spatial interaction in point patterns.* Statistica Neerlandica 50, p. 344-361.

WOOD W.W. (1968). *Monte Carlo studies of simple liquid models*, Physics of simple liquids, p. 115-230, Amsterdam.

Summary

Estimation of smooth functionals with interval censored data

and something completely different

Two quite different topics are treated. One is estimation of smooth functionals of the distribution in a situation with interval censored data; the other one is estimation of the amount of repulsion among points in some homogeneous area, based on one spatial realization.

The larger part (chapters 1 and 2) is devoted to the first topic. One has to deal with interval censored data, if one wants to obtain information on some distribution F_0, often representing an event time distribution, without being able to observe the event times $X_1, \ldots, X_n \sim F_0$ directly. One only has a collection of, usually random, observation times, leading to a sample of intervals $J_1, \ldots, J_n$ in which the unobservable X_i are known to be contained. Interval censored data can be subdivided into several categories. In case 1, we have one observation time T_i for each X_i, and we only know whether X_i is smaller or larger than the corresponding observation time T_i. Case 2 is denoted as the situation with two observation times (U_i, V_i) for each unobservable event time, and we only know whether X_i is left of U_i, between U_i and V_i or right of V_i. Situations with more than two observation times, or a variable number of observation times, for each unobservable event time are denoted as the case k situation. This case very much resembles case 2, since only the two observation times immediately surrounding the event time give relevant information.

Typical for interval censored data, contrary to right censored data, is that the event time is never observed itself. This has strong consequences for the asymptotic theory. The distribution function cannot be estimated with the usual $\sqrt{n}$-rate, and the limiting distribution is not normal. However, some aspects of the distribution, such as the mean, remain estimable at $\sqrt{n}$-rate and have a normal limit distribution. Necessary for this to happen is that the functional, representing this aspect of the distribution, is sufficiently "smooth". For such functionals, a general lower bound theory exists, telling us what is the best performance an estimator can have with respect to the variance of the limiting normal distribution. A relation exists with the limit variance in case of direct observable event times, which is expressed by the *score equation*. For smooth functionals, this score equation is solvable, and the squared norm of its solution yields the lower bound in the situation with interval censored data.

In case 1, the score equation is easily shown to be solvable under general conditions, and an explicit formula for the solution is obtained. In case 2, however, we have to solve an integral equation with a Fredholm type structure. A solution is shown to exist under general conditions, but no explicit formula is available. In the last section of chapter 1, for some specific choices of both the observation time distribution and the event time distribution as well as the functional of interest, a more explicit solution to this integral equation is obtained.

In chapter 2, the results of chapter 1 are used to show that the nonparametric maximum likelihood estimator (NPMLE) of a smooth functional asymptotically has the optimal behaviour with respect to the variance of the limit distribution. First the basic ingredients of the proof in case 1 are sketched. In case 2, the proof is similar in essence. Since no explicit expression is available for the lower bound, each part of the proof is considerably more complicated. One of the important parts is the derivation of an equality, reformulating the functional as an integral with respect to the probability measure on the observation

space. The integrand is the solution to a modified integral equation, with F_0 replaced by the NPMLE $\hat{F}_n$. In chapter 1, some smoothness properties of this solution, which hold uniformly over $\hat{F}_n$, are derived. Another important part is a characterization of $\hat{F}_n$ as an integral with respect to the empirical measure on the observation space.

The last chapter deals with Gibbs point processes, which are characterized by the existence of spatial interaction among points. Gibbs processes have a specific structure of the density. The main ingredient of the density is the *interaction function*. Only pairwise interactions are assumed to occur. Moreover, the interaction is assumed to be repulsive in nature, with the amount of repulsion decreasing with increasing interpoint distance. Then the interaction function is monotonically increasing as a function of the interpoint distance, with values between zero and one. The nonparametric maximum likelihood estimator is shown to be a piecewise constant function. Since the density has a very complex norming constant, only approximations to the likelihood can be computed. Attention is restricted to Gibbs processes, conditioned on a fixed number of points. We make a rough approximation, which is only reasonable in case the process is close to Poisson. Consistency of the procedure in case the true process is Poisson is shown, and some simulations are done for other choices of the interaction function. The simulation studies show the estimator to be quite heavily biased.

Samenvatting

Het schatten van gladde functionalen op basis van interval gecensureerde data

en iets geheel anders

Twee verschillende onderwerpen uit de statistiek worden behandeld. Het eerste betreft het schatten van gladde functionalen van de verdelingsfunctie op basis van interval gecensureerde data. Het tweede betreft het schatten van de mate van afstoting tussen elementen op basis van één ruimtelijke configuratie van die elementen in een homogeen gebied.

Het grootste deel (hoofdstuk 1 en 2) is gewijd aan het eerste onderwerp. Men heeft te maken met interval gecensureerde data wanneer men de kansverdeling van de tijd tot het optreden van een gebeurtenis (gerepresenteerd door F_0) wil schatten, zonder dat men de gebeurtenissen direct kan waarnemen. Men heeft slechts de beschikking over een collectie, meestal stochastisch bepaalde, observatietijdstippen, hetgeen leidt tot een steekproef van intervallen $J_1, \ldots, J_n$ waarin de niet direct observeerbare $X_1, \ldots, X_n \sim F_0$ (de tijdstippen van optreden van de gebeurtenis) gelegen zijn. Interval gecensureerde data kunnen in verschillende categorieën ingedeeld worden, afhankelijk van het aantal observatietijdstippen per gebeurtenis. In "case 1" heeft men per onobserveerbare gebeurtenis één observatietijdstip T_i, en over de gebeurtenis is slechts bekend of deze voor of na T_i heeft plaatsgevonden. In "case 2" heeft men twee observatietijdstippen (U_i, V_i) per gebeurtenis, en weet men de locatie van de gebeurtenis ten opzichte van deze twee tijdstippen. Situaties met meer dan twee observatietijdstippen, of een variabel aantal, per gebeurtenis behoren tot "case k". Deze laatste situatie vertoont veel overeenkomsten met geval 2, omdat slechts het laatste observatietijdstip vóór en het eerste observatietijdstip ná de gebeurtenis relevante informatie geven.

Kenmerkend voor interval gecensureerde data, en dit in tegenstelling tot rechts gecensureerde data, is dat de gebeurtenis zelf nooit wordt waargenomen. Dit heeft belangrijke implicaties voor de asymptotiek. De verdelingsfunctie kan niet met $\sqrt{n}$-snelheid geschat worden, en de asymptotische verdeling is niet normaal. Sommige aspecten van de verdeling, zoals de verwachting, blijven echter met $\sqrt{n}$-snelheid schatbaar, met een normale verdeling als limietverdeling. Noodzakelijk hiervoor is dat de functionaal die dit aspect van de verdeling representeert voldoende "glad" is. Voor zulke gladde functionalen kan gebruik gemaakt worden van algemene theorie over informatie-ondergrenzen. De informatie-ondergrens geeft de kleinst mogelijke variantie van de normale limietverdeling die een gestandaardiseerde schatter kan bereiken. Er is een directe relatie met de ondergrens in de situatie met ongecensureerde X_i. Deze relatie wordt uitgedrukt middels de *score vergelijking*. Deze score vergelijking is oplosbaar voor gladde functionalen, en de informatie-ondergrens wordt gegeven door de gekwadrateerde norm van de oplossing.

In geval 1 heeft de score vergelijking een eenvoudige structuur en hebben we een expliciete formule voor de ondergrens. In geval 2, daarentegen, komt het oplossen van de score vergelijking neer op het oplossen van een Fredholm integraalvergelijking. Het bestaan van een oplossing wordt aangetoond onder algemene condities. Een expliciete formule is echter niet beschikbaar. In de laatste paragraaf van hoofdstuk 1 wordt voor een aantal speciale keuzes van de verdeling van X_i en de observatietijdstippen, met als functionaal de verwachting, een min of meer expliciete oplossing afgeleid.

In hoofdstuk 2 worden de resultaten uit hoofdstuk 1 gebruikt om aan te tonen dat de niet-parametrische maximum-likelihood schatter (NPMLE) van een gladde functionaal asymptotisch het optimale gedrag heeft wat betreft de limietvariantie van de normale verdeling. Eerst worden de belangrijke ingredienten van het bewijs in geval 1 geschetst. Het bewijs in geval 2 verloopt in essentie op dezelfde wijze. Echter, omdat er geen expliciete oplossing van de integraalvergelijking is, is iedere stap in het bewijs aanmerkelijk gecompliceerder dan in geval 1. Een belangrijkste stap is de afleiding van een vergelijking die de functionaal herformuleert als een integraal met betrekking tot de kansmaat van de observatieruimte. De integrand is de oplossing van een gemodificeerde integraalvergelijking, waarin F_0 vervangen is door de NPMLE $\hat{F}_n$. In hoofdstuk 1 worden een aantal gladheidseigenschappen bewezen voor deze oplossing, die uniform zijn over alle mogelijke realisaties $\hat{F}_n$. Een andere belangrijke stap is het gebruik van een karakterisatie van $\hat{F}_n$ als een integraal met betrekking tot de empirische kansmaat op de observatieruimte.

Het laatste hoofdstuk gaat over Gibbs puntprocessen, die gebruikt worden om ruimtelijke interactie tussen punten te modelleren. Gibbs processen hebben een dichtheid, waarvan het belangrijkste ingredient een *interactiefunctie* is. Alleen Gibbs processen waarbij alle interacties paarsgewijs zijn worden onderzocht. Bovendien wordt aangenomen dat alle interacties afstotend van aard zijn, waarbij de sterkte van de afstotingskracht afneemt als de afstand tussen de punten groter wordt. Voor de interactiefunctie betekent dit dat deze monotoon niet-dalend is als functie van de afstand tussen punten, met waarden tussen nul en één. De niet-parametrische maximum-likelihood schatter onder de monotoniciteitsrestrictie is een stuksgewijs constante functie. Echter, het uitrekenen hiervan is ingewikkeld, vanwege de aanwezigheid van een gecompliceerde normeringsconstante. Een ruwe benadering van de NPMLE wordt afgeleid voor de situatie waarin het aantal punten vast ligt. Deze benadering is slechts redelijk als er weinig interactie is. Als er geen interactie is, is de schattingsprocedure asymptotisch consistent. Voor een aantal keuzes van de interactiefunctie is een simulatie gedaan. Hieruit blijkt dat de schatter niet erg zuiver is.

Curriculum Vitae

De auteur van dit proefschrift werd geboren op 17 december 1962 in Amsterdam.
Van 1975 tot 1981 bezocht hij het Gymnasium Felisenum in Velsen-Zuid, waar hij eindexamen deed in de vakken nederlands, engels, grieks, natuurkunde, scheikunde, wiskunde en biologie.
In 1981 begon hij met de studie natuurkunde aan de Universiteit van Amsterdam (UvA). Na één jaar en twee maanden stapte hij over naar wiskunde.
In 1986 behaalde hij cum laude het kandidaatsexamen wiskunde, met bijvak filosofie.
In zijn doctoraalfase specialiseerde hij zich in de mathematische statistiek. In 1989 behaalde hij het doctoraalexamen wiskunde (cum laude) bij prof. dr. P. Groeneboom, met een onderwijsbevoegdheid (didactische aantekening) en als bijvak politicologie.
Na een rondfietsend bestaan van drie maanden in Z.O. Azië, vervulde hij van januari 1990 tot en met april 1991 zijn vervangende dienstplicht aan de Technische Universiteit Delft (TU Delft).
In mei 1991 begon hij als assistent in opleiding aan de TU Delft aan zijn promotieonderzoek, onder supervisie van prof. dr. P. Groeneboom.
Sinds oktober 1995 is hij werkzaam als biostatisticus bij de GG&GD in Amsterdam, waar hij zijn opgedane statistische en wiskundige kennis in de praktijk toepast op gegevens verkregen uit de cohort studies naar HIV en AIDS.

Acknowledgements

A few people contributed to the contents of this thesis. I thank Piet for guiding me through all stages of my research. I thank my roommates, Geurt and Peter-Paul, for their advice on mathematical, LaTeX and programming problems and for sharing their different points of view with respect to many things in life. Jon Wellner gave some good suggestions during the DMV course in Günzburg. In the first year, the contribution by Peter Sonneveld to the solving of the integral equation for the special case with uniform distributions was very valuable. It provided the basis for my first results. I thank Carel for clarifying to me some of the theoretical aspects of Gibbs distributions. Philippe Clément explained me some concepts from functional analysis. The assistance by Karma and Rik was important as well. Leo was always willing to provide help when my knowledge of unix was not sufficient to get the workstations doing what I wanted them to do. All the other people in Delft provided a pleasant working environment. I thank Jaap Maas for his assistance with the cover drawing.

During the final stage of my Ph.D. period, I was getting increasingly concerned at not being able to find a new occupation. However, I was fortunate enough to see one of my dreams fulfilled: working as a statistician in an area of applied research that both provides challenging statistical problems and has interested me for years. I thank Roel Coutinho, Frits van Griensven and Anneke van den Hoek for giving me this opportunity.

During my Ph.D. period, all of my friends were important to me, like they had been before and will be in the future. Especially Luc, who came into my life during a research period in which I needed something to look forward to, was of great value. Rather than thanking them all for their presence, I think it is more suitable to apologize for having had increasingly less attention for them as the completion of my thesis was getting closer, and especially during the past one and a half year. I hope I will be able to lead a more social life again.

Stellingen

behorende bij het proefschrift

Het schatten van gladde functionalen op basis van interval gecensureerde data

en iets geheel anders

door

Ronald Geskus

1. *Zij A een niet-singuliere M-matrix. Beschouw de vergelijking $Ax = y$. Als $y_K \neq 0$ voor iedere nucleus K, en als $x_i > 0$ voor iedere index i met $y_i < 0$, dan geldt dat al de coördinaten van x positief zijn.*

   MILASZEWICZ, J.P. AND MOLEDO, L.P. (1993). *On nonsingular M-matrices*, Linear Algebra and its Applications, vol. 195, p. 1-8.

   Een aanmerkelijk eenvoudiger bewijs van deze stelling kan gegeven worden door gebruik te maken van de redenering uit het bewijs van lemma 1.4.1 op pagina 35 van dit proefschrift.

2. Gezien het niet-normale asymptotische gedrag, heeft, bij het niet-parametrisch schatten van niet-gladde functionalen op basis van interval gecensureerde en dubbel gecensureerde data, het gebruik van de Fisher informatie matrix voor het bepalen van betrouwbaarheidsintervallen een twijfelachtige waarde. Alleen wanneer het probleem parametrisch behandeld wordt en een zeer grof vast grid gekozen wordt, geldt het asymptotisch gedrag bij de aanwezige steekproefgrootte bij benadering.

3. Het formuleren van een statistiek opgave via

   Zij $x_1, \ldots x_n$ de realisatie van een steekproef uit een normale verdeling ...

   rechtvaardigt op logische gronden het geven van een onzinnig antwoord.

4. Het gebruik van de archaïsche notatie $\underline{x}$ voor een stochastische variabele heeft een onderwijskundig voordeel. Bij het nakijken van tentamens is veel eenvoudiger te zien of een student het verschil begrijpt tussen een stochastische variabele en een realisatie daarvan.

5. De bewering dat deelnemers aan cohort studies naar HIV en AIDS, die tijdens de studie zijn geseroconverteerd, een gedocumenteerde datum van seroconversie hebben, is vanuit statistisch oogpunt misleidend.

6. De homo-emancipatie is nog verre van voltooid, ook niet bij de homoseksueel zelf. Het is dan ook te prefereren een jongen te vragen of hij een relatie heeft in plaats van te vragen of hij een vriendin heeft.

7. De *straat* van Gibraltar is een eufemisme geworden.

8. Beter één milieu-onvriendelijke white-board stift in de hand, dan tien milieu-vriendelijke in de prullenmand.