Nonparametric Regression and Influence Curves.

# Contents

# Chapter 1

# Estimation in Semiparametric Models.

## 1.1 Outline and objectives.

The purpose of this chapter is to give you methods for attacking statistical problems which might arise in applications without falling back to known models right away, restricting to the i.i.d. setting with covariates, and mostly censored data examples. This course should give you tools to attack the problem independently and come up with your own proposals in situations where standard models are not applicable. We try to achieve this stepwise by covering the following concepts. On our way we present real life examples to illustrate results and these concepts: univariate right-censoring model, singly and doubly censored current status data, random truncation, bivariate right-censoring model, singly censored current status with high-dimensional covariates, univariate right-censored with high-dimensional covariates. A key reference is Bickel, Klaassen, Ritov and Wellner (1993); it contains a describtion of essentially all well known models and a full historic account of work carried out. The extension of the current status model to high-dimensional covariates is carried out in van der Laan (1995), inspired by work of Robins (1993), Robins and Rotznitzky (1992) and Gill, van der Laan, Robins (1995).

Firstly, understand that the empirical data, which can be summarized by the so called empirical probability distribution, represents all the data is telling you. Hence the first question which arises is how much does this empirical probability distribution tell you. This raises the issues of uniform consistency and a uniform central limit theorem for the empirical distribution, which will be discussed in our first section.

Secondly, in the process of understanding the data structure a parametric or semiparametric model will often arise. Then one will often be concerned with estimation of a (or a function of) parameter indexing the distribution of the data. This raises the issue of identifiability; can we recover the parameter from the distribution of the data, given the model? Showing identifiability comes down to showing that the parameter can be represented as a function of the distribution of the data. It is important to practice the construction of these representations in a number of examples or exercises (which are given).

Such a representation implies an estimator by simple substitution of the empirical probability distribution or a smoothed version of it. Most estimators can be represented in this way. Are these estimators consistent? In other words, if the number of observations is large, is the

estimator close to the truth? It appears that continuity of the representation as a mapping from empirical distributions to the estimators answers this question.

For construction of confidence intervals for the parameter of interest it is important to have a limit distribution (for a standardized difference between the estimator and the parameter) for obtaining approximate quantiles for the difference between the estimator and the parameter. For this purpose we discuss and apply the so called delta-method, generalized to any type of estimator. The delta-method is a technique for linearizing (approximating) the estimator as a sum of i.i.d. random variables $f(X_i)$ which are functions of the observed $X_i$, here $f$ being the so called influence curve since it measures the influence of one observation $X$ on the estimator. Once one succeeds in the linearization, application of the central limit theorem provides us with a normal mean zero limiting distribution

This limit distribution depends only on the variance of the influence curve so that estimation of the influence curve will provide us with approximate confidence intervals. The influence curve is even of more interest. It can be used to judge robustness of the estimator and comparison of asymptotically linear estimators can be done by comparing the variances of their influence curves. Hereby the question arises of how to find the influence curve with minimal variance and its corresponding efficient estimator. We describe a method for finding the efficient influence curve. The efficient influence curve can be used to construct efficient or locally efficient estimators.

## 1.2 The empirical probability distribution

Let $X$ be a observable random variable; in other words, a potential outcome of an experiment. In many applications $X$ is a random draw from a population. For example, $X$ might be the time between infection with the HIV-virus and the onset of the disease AIDS for a randomly drawn individual of a population of HIV-infected patients. We do not restrict ourselves to real valued random variables; $X$ might for example be a vector $X = (X_1, \ldots, X_d)$ of measurements on a randomly drawn individual. The observable random variable might have a discrete and a continuous component. For example, if $T$ is a survival time of interest and $C$ is an irrelevant variable which censors $T$ in the sense that if $C$ occurs before $T$, then we observe $C$ and otherwise we observe $T$ (e.g. $C$ is the time at a car-accident or the time at which the patient emigrates etc.). How would you represent the observable random variable? So here $X = (T \wedge C, \Delta = I(T \leq C))$.

$X$ will follow a certain probability distribution, say $P_0$, which means that for a given set $A$ of possible outcomes of $X$ we have that the probability that $X$ falls in $A$ is given by $P_0(A)$; in notation we write $\Pr(X \in A) = P_0(A)$. The right heuristic way to think about a probability, say $P_0(A) = 0.1$, is that if we repeat the experiment (e.g. randomly drawing a person from a population) which generates $X$ a large number of times, then the fraction of times that $X$ falls in $A$ is close to 0.1.

Let now $X_1, \ldots, X_n$ be $n$ *independent* copies of the random variable $X$ which are obtained by repeating the experiment which generates $X$.

**Exercise 1.** You do not have any knowledge about $P_0$. How would you estimate $P_0(A)$ for some set $A$.

**Exercise 2.** Let $f(X)$ be a real valued function of $X$ with finite expectation; i.e.

$\int f(x)dP_0(x) < \infty$. We introduce here the following notation:

$$P_0 f \equiv \int f(x)dP_0(x),$$

i.e. $P_0 f$ stands for the expectation of the random variable $f(X)$. How would you estimate $P_0 f$? So in particular how would you estimate the moments of $X$?

**Exercise 3.** Show that one can write $P_0(A)$ as $P_0 f$ for some $f$. In other words, estimation of $P_0(A)$ is a special case of estimation of $P_0 f$.

$P_0$ can be estimated with the empirical probability measure

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A).$$

$P_n$ is a measure which puts mass $1/n$ on each observable random variable $X_i$. In words this says that $P_n(A)$ equals the fraction of the observations which have fallen in $A$. We can estimate $P_0 f = \int f(x)dP_0(x)$ with

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

By the Glivenko-Cantelli theorem we have that if $f(X)$ has finite expectation (i.e. $Pf < \infty$), then

$$P_n f - Pf \to 0 \text{ a.s.}$$

Almost surely means here that for each sequence of observations $(X_1, X_2, \ldots)$ randomly drawn from $P$ we have that $P_n f - Pf \to 0$ when $n \to \infty$.

### 1.2.1 Uniform consistency of the empirical probability distribution.

In statistics we are concerned with estimation of certain parameters. An estimator of such a parameter is by definition a function of the observations $(X_1, \ldots, X_n)$ or equivalently a function of the empirical measure $P_n$. Therefore convergence of $P_n$ to $P_0$ in an appropriate sense usually translates in a convergence result for the estimator. In parametric models estimators of parameters are often functions of a finite number of $P_n f_i$, which is basically due to the fact that the parameter to be estimated is of low dimension. For example, a method of moments estimator depends on the empirical moments $\frac{1}{n}\sum_{i=1}^{n} X_i^k = P_n X^k$ for some values of $k$. For example, consider the sample variance which can be written as:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}^2 = \mu_{2n} - \mu_{1n}^2,$$

where $\mu_{1n}$ and $\mu_{2n}$ are the empirical mean and empirical second moment, respectively. Hence convergence results for $P_n f_i$ for a finite number of $f_i$'s will suffice for such estimators; in this example, convergence of $\mu_{1n}$ to $\mu_1$ and $\mu_{2n}$ to $\mu_2$ suffices for convergence of $S^2$ to $\sigma^2$. However, suppose now that $X$ is real valued and that $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ is the empirical cumulative distribution and we are concerned with the question if

$$\|F_n - F\|_\infty \equiv \sup_x \mid F_n(x) - F(x) \mid \to 0 \text{ in probability or almost surely.}$$

Here we see that the dependence of the estimator $F_n$ as an estimator of $F$ goes beyond dependence through a finite collection of $P_n f_i$, $i = 1, \ldots, k$. Therefore it is often of interest to have a consistency result for $P_n f$ uniformly in a collection of $f$'s.

Let $\mathcal{F}$ be a set of functions $f$ from $X$ to $\mathbb{R}$ with finite expectation.. We say that $\mathcal{F}$ is a *Glivenko-Cantelli class* if

$$\sup_{f \in \mathcal{F}} \mid P_n f - P_0 f \mid \to 0 \text{ a.s. or in probability.} \tag{1.1}$$

We should distinguish between Glivenko in probability or Glivenko almost surely, hereby refering to the type of convergence in (1.1). Our default will be almost surely.

**Example 1.2.1 (Indicators of Lower rectangles).** Let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathbb{R}^d$ and let $\mathcal{F}$ be the collection of all indicator functions of lower rectangles $\{I_{(-\infty,t]} : t \in \mathbb{R}^d\}$. It is a classical result that $\mathcal{F}$ is indeed a Glivenko-Cantelli class. If you are interested in a proof of this result I refer for example to section 2.4 of Sen and Singer.

It should be said that many more interesting Glivenko-Cantelli classes exist. It is a good exercise to convince yourself that $\mathcal{F} = L^2(P_0)$ is not a Glivenko-Cantelli class. So $\mathcal{F}$ cannot be too large. In the literature Glivenko-Cantelli classes have been charactarized by an entropy measure which is a measure about the size of the class and how many functions can be approximated by it. If $X \in \mathbb{R}$, then a non-trivial Glivenko-Cantelli class is given by all functions with variation smaller than a universal constant $M < \infty$:

$$\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} : \int \mid df(x) \mid < M\}.$$

Another one is the class of general indicators instead of only lower rectangles:

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

You could as well have replaced $(a, b]$ by $(a, b)$ or $[a, b]$ or $[a, b)$. If $X$ is a bivariate random variable, then we could wonder if indicators of all kinds of shapes form Glivenko-Cantelli classes. If we restrict the shapes to some finite dimensional parametrization, then one will have a Glivenko-Cantelli class. For example,

$$\mathcal{F} = \{I_{A(m,r)} : A(m, r) \text{ a circle with midpoint } m = (m_1, m_2), \text{ radius } r\},$$

indicators of circles, form a Glivenko-Cantelli class. So this teaches us that the fraction of bivariate observations $X_i$ which fall in a circle converges to the probability that $X$ falls in this circle, *uniformly* over all possible circles.

### 1.2.2 Uniform central limit theorem for the empirical probability distribution.

In the following the set of all random variable $f(X)$ with finite second moment (hence with finite variance) is denoted with

$$L^2(P_0) = \{f : \int f^2(x) dP_0(x) < \infty\}.$$

By the central limit theorem we have that for any $f \in L^2(P_0)$

$$G_n f \equiv \sqrt{n}(P_n f - P_0 f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_i) - Ef \stackrel{D}{\Longrightarrow} N(0, Pf^2 - P^2 f).$$

With $\stackrel{D}{\Longrightarrow}$ we mean convergence in distribution.

**Exercise.** Construct an approximate confidence interval for $\int f(X) dP(x)$ for a $f(X)$ with finite variance, using the CLT-result.

More general, by the multivariate central limit theorem we have for any finite set of functions $\{f_1, f_2, \ldots, f_k\}$ with $f_i \in L^2(P_0)$, $i = 1, \ldots, k$:

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

where the $k \times k$ matrix $\Sigma$ has $(i, j)$'th element $P(f_i - Pf_i)(f_j - Pf_j)$, i.e. it equals the covariance between $f_i(X)$ and $f_j(X)$.

Again, for obtaining asymptotic normality of an estimator these results will often not suffice since the estimator often depends on $P_n$ through more than a finite collection of $f$'s. Let $\mathcal{F} \subset L^2(P_0)$ be a class of real valued functions satisfying

$$\sup_{f \in \mathcal{F}} \mid f(X) - Pf \mid < \infty \text{ for all } X.$$

Under this condition on $\mathcal{F}$ the empirical process $\{G_n f : f \in \mathcal{F}\}$ can be viewed as random element of

$$\ell^\infty(\mathcal{F}) \equiv \{H : \mathcal{F} \to \mathbb{R} : \sup_{f \in \mathcal{F}} \mid H(f) \mid < \infty\}.$$

Consequently, it makes sense to investigate conditions under which $G_n$ converges in distribution to $G$, as random elements of $\ell^\infty(\mathcal{F})$, where $G$ is a Gaussian process (a random element of $\ell^\infty(\mathcal{F})$) determined by its finite dimensional distributions; for any finite set$\{f_1, f_2, \ldots, f_k\}$ with $f_i \in L^2(P_0)$, $i = 1, \ldots, k$ we have

$$(Gf_1, \ldots, Gf_k) \stackrel{D}{\Rightarrow} N(0, \Sigma),$$

where $\Sigma$ has been defined above.

It can be shown that indeed such a Gaussian process $G$ is uniquely determined by its finite dimensional distributions. We did not formally define what we mean with convergence in distribution of random elements in large function spaces, but the way to think about it is that the probability that $G_n$ falls in a certain subset of $\ell^\infty(\mathcal{F})$ converges to the probability that $G$ falls in this same subset, where we should remark that this only has to hold for subsets for which the probability that $G$ falls on the boundary of this set equals zero.

We say that $\mathcal{F}$ is a *Donsker class* if the empirical process $G_n \in \ell^\infty(\mathcal{F})$ converges in distribution to $G$ as random elements of $\ell^\infty(\mathcal{F})$. In particular we have that if $\mathcal{F}$ is Donsker, then

$$\|P_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} \mid \frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} f(X_i) - Ef(X) \right)$$

is bounded in probability. In notation we say $\sup_{f\in\mathcal{F}} \mid P_n f - Pf \mid = O_P(1/\sqrt{n})$ which means that for all $\epsilon > 0$, there exists a $M < \infty$ so that $P(\sqrt{n}\|P_n - P\|_{\mathcal{F}} > M) < \epsilon$. In other words, the empirical mean $P_n f$ converges with a rate $1/\sqrt{n}$ to the mean $Pf$, *uniformly* in $f \in \mathcal{F}$.

The mentioned Glivenko-Cantelli classes are also Donsker classes. Could you give an example of a class which is Glivenko, but not Donsker? If a class $\mathcal{F}$ is Donsker, then it follows trivially that it is Glivenko in probability.

**Example 1.2.2 (Indicators of Lower rectangles).** Let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathbb{R}^d$ and let $\mathcal{F}$ be the collection of all indicator functions of lower rectangles $\{I_{(-\infty,t]} : t \in \mathbb{R}^d\}$. The empirical process $G_n$ indexed by $\mathcal{F}$ can be identified with

$$t \to \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(X_i \leq t) - F(t).$$

In this case it is natural to identify $f = I_{(-\infty,t]}$ with $t \in \mathbb{R}^d$ and the space $\ell^\infty(\mathcal{F})$ with $\ell^\infty(\mathbb{R}^d)$, i.e. we consider $G_n(\cdot) \equiv \sqrt{n}(F_n(\cdot) - F(\cdot))$ as a random element of the space of real valued functions on $\mathbb{R}^d$ with finite supnorm. We wonder if this random function $G_n(\cdot)$ converges in distribution to the Gaussian process $G$ as identified above.

It is known from classical results that the class of lower rectangles is Donsker for any underlying $P_0$ of $X$; so the answer is yes. As a result we have that

$$\sup_{t\in\mathbb{R}^d} \mid F_n(t) - F(t) \mid = O_P(1/\sqrt{n}).$$

**Example 1.2.3 (functions of bounded variation).** Let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathbb{R}$ and let $\mathcal{F}$ be a class of real valued functions which have their variation bounded by a universal constant:

$$\mathcal{F} \equiv \{f : \mathbb{R} \to \mathbb{R} : \int \mid df(x) \mid < M\}$$

Also $\mathcal{F}$ is Donsker.

## 1.3 Semiparametric models and identifiability.

In statistics it is often known or assumed that the probability distribution of $X$ lies in a certain set of possible probability distributions.

**Definition 1.3.1** *A model for $X$ is a collection of possible probability measures of $X$.*

A model can always be represented by $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for some index set $\Theta$. If $\Theta \subset \mathbb{R}^k$ for some integer $k$, then we call $\mathcal{M}$ a parametric model and if $\Theta$ is not finite dimensional, then we call $\mathcal{M}$ a semiparametric model.

Suppose that we are interested in estimation of $\Psi(\theta) \in D$, where $D$ is some linear space. For example, $D = \mathbb{R}^k$ for some integer $k$ or $D$ might be the cadlag function space consisting of right-continuous real valued functions which have left hand limits. There is only hope to be able to estimate $\Psi(\theta)$ if it is identifiable from the distribution of the data; remember that the data is all we have, so the data should be able to help us to identify $\Psi(\theta)$.

**Definition 1.3.2** *$\Psi(\theta)$ is called identifiable if $\Psi(\theta) = V(P_\theta)$ for all $\theta \in \Theta$ and a fixed mapping $V : \mathcal{M} \to D$. The mapping $V : \mathcal{M} \to D$ is called a parameter.*

In words, this says that $\Psi(\theta)$ is identifiable if you have a recipe for computing $\Psi(\theta)$ from $P_\theta$. For example, let $\mathcal{M} = \{N(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{\geq 0}\}$ and $\Psi(\mu, \sigma^2) = \mu$. Then $\Psi(\mu, \sigma^2)$ is identifiable since $\mu = \int x f_{\mu,\sigma^2}(x)dx$; i.e. if I give you $f_{\mu,\sigma^2}$, then you just compute this integral in order find $\mu$. Notice that $\Psi(\theta)$ can be identifiable, while $\theta$ is not.

There are two methods for proving identifiability of $\Psi(\theta)$. One could prove it indirectly by showing that

$$P_{\theta_1} = P_{\theta_2} \Rightarrow \Psi(\theta_1) = \Psi(\theta_2),$$

or one could prove it by constuction of the mapping $V{:}\mathcal{M} \to D$, i.e. $\Psi(\theta) = V(P_\theta)$ for all $\theta \in \Theta$. The latter method is preferable since the mapping $V$ can be used to obtain an estimator, e.g. $V(P_n)$, where $P_n$ is the empirical distribution of the observations, or if $V$ is not defined on a discrete $P_n$, then we could take for $P_n$ a smoothed empirical distribution.

### 1.3.1 Proving identifiability by explicit construction in semiparametric examples.

You will be familiar with many parametric models. We will now mention a number of examples of semiparametric models which naturally arise in applications and show identifiability of the parameters of interest by explicit construction of the mapping $V$.

**Example 1.3.1 Density estimation.** Suppose that $X \in \mathbb{R}$ is a continuous random variable with a density $f \in C^{(k)}(0, \tau)$. We want to estimate $f$. $f$ is identifiable since $f = dF/dx$, where $F$ is the cumulative distribution of $X$. In other words, if I give you the probability distibution of $X$, say $P$, then you obtain $f$ by first determinig $F(x) = P(X \leq x)$ and then you differentiate $F(x)$ with respect to $x$.

**Example 1.3.2 (Nonparametric) regression.** Suppose that we observe $n$ independent copies of a joint random variable $(X, Y)$. We assume that $Y = m(X) + e$, where $Ee = 0$ and $m \in \mathcal{R}$, where $\mathcal{R}$ is a class or real valued functions. For example, $\mathcal{R}$ might be the class of all monotone functions or one could think about parametric representations of $m$ as $\mathcal{R} = \{m(x) = a_0 + a_1 x + a_2 x^2 + \ldots a_k x^k : (a_0, a_1, \ldots, a_k) \in \mathbb{R}^k\}$. It is quite common to assume that $e \sim N(0, \sigma^2)$ which can be weakened by making $\sigma^2$ depend on $x$. We want to estimate the unknown function $m$. Identifiability of $m$ follows from the fact that $m(x) = E(Y \mid X = x) = \int y f(y \mid x)dy$, assuming that $m$ is such that $Y$, given $X$, has a continuous density.

We will now give a few examples of nonparametric missing data models which are of high importance in biostatistical applications.

**Example 1.3.3 Current status data.** Consider a carcinogenicity experiment with $n$ rats. Let $T$ be the time of onset of the tumor in a rat. $T$ is a random variable with an unknown distribution $F$. Let $C$ be a time at which the rat is sacrificed or at which the rat naturally dies. We observe $(C, \Delta \equiv I(T \leq C))$; in other words, we observe if the tumor is present at time $C$ or not. In order to have identifiability of $F$ it is often assumed that $T$ and $C$ are independent and that $C$ is a continuous random variable with density $g$. Our goal is to estimate $F(t)$.

Let's write down the density of the data:

$$p(c,\delta) = p(\delta = 1 \mid c)g(c)\delta + p(\delta = 0 \mid c)g(c)(1-\delta) = F(c)g(c)\delta + (1-F(c))g(c)(1-\delta).$$

It follows that

$$F(c) = \frac{p(c,1)}{p(c,1)+p(c,0)}.$$

This proves identifiability of $F$ since $F$ is a mapping from the density $p(c,\delta)$ of the data $(C,\Delta)$ to $F$. Notice that in this application assuming independence between $T$ and $C$ is only justified if all the rats are sacrificed since if $C$ is time of natural death it will be correlated with the time $T$ of onset of the tumor. We will consider interesting extensions of this current status model to the case where we also observe covariates related to $T$ and/or $C$, which will allow us to weaken the independence assumption essentially.

**Example 1.3.4 Univariate right censored data.** Let $T$ be a survival time of interest with distribution $F$. Assume that we observe

$$Y = (\widetilde{T}, \Delta) = (T \wedge C, I(T \leq C)),$$

where $C$ is a censoring variable with distribution $G$. In order to obtain identifiability of $F$ it is often assumed that $T$ and $C$ are independent.

In order to show identifiability we will again construct explicitly a mapping from $P_Y$ to $F$. Let $0 = t_0 < t_1 < \ldots < t_m = t$ be a partition of $[0,t]$. Notice that

$$S(t) = \frac{S(t_m)}{S(t_{m-1})}\frac{S(t_{m-1})}{S(t_{m-2})}\cdots\frac{S(t_1)}{S(t_0)}.$$

We have that

$$\frac{S(t_m)}{S(t_{m-1})} = 1 - P(T \in (t_{m-1}, t_m] \mid T > t_{m-1}) \approx 1 - \lambda(dt_{m-1}),$$

where $\lambda(dt) = P(T \in dt \mid T > t)$. Hence by making the partition finer and finer it follows that

$$S(t) = \prod_{(0,t]} (1 - \lambda(ds)).$$

Finally, notice that

$$\lambda(ds) = P(T \in ds \mid T > s) = \frac{P(T \in ds, C > s)}{P(T > s, C > s)} = \frac{P(\widetilde{T} \in ds, \Delta = 1)}{P(\widetilde{T} > s)}$$

and hence $S$ is an explicit mapping from the distribution of $(\widetilde{T}, \Delta)$ to $S$.

### 1.3.2 Practice examples.

It will be a good exercise for you to construct such explicit mappings yourself for a parameter of interest. Firstly, you can find plenty of such examples in parametric models; just take an introduction book and look at the section "method of moments estimator" in which one chooses for representing the parameter of interest as a function of moments of the distribution of the data. Here are a few important semiparametric examples you should try to solve:

**Example 1.3.5** (Doubly censored current status data). Consider a population of partners from which it is that one of them is infected with the HIV-virus. Let $I$ be the chronological time at infection of the index partner and let $J$ be the chronological time at infection of its sexual parnter. Let $T = J - I$ be the random variable of interest. Individual observations are sampled conditional on $A \leq I \leq B$, and instead of complete observation of $(I, J)$, only current status information on $J$ is available at time $B$. That is, in addition to $A$ and $B$, we observe $\Delta \equiv I(J \leq B)$. To avoid unidentifiability, we assume that the conditional distribution of $I$ on $[A, B]$ is known, and, following Jewell, Malani and Vittinghoff (1994), we take this distribution to be the uniform distribution on $[A, B]$. Moreover, it is assumed that $T$ is independent of $I, A, B$. We refer to this data structure as *doubly censored current status data* since observation of both $I$ and $J$ are interval censored.

If we define $C = B - A$, then we have (at the third equality we use that $T$ and $I$ are independent, given $A, B$, and that $T$ is independent of $A, B$):

$$\begin{aligned} P(\Delta = 1 \mid A, B) &= P(J \leq B \mid A, B) && (1.2)\\ &= P(T \leq B - I \mid A, B) && (1.3)\\ &= \int_0^{B-A} P(I < B - t \mid A, B) dG(t) && (1.4)\\ &= \int_0^{B-A} \frac{B - A - t}{B - A} dG(t) && (1.5)\\ &= F_G(C) \equiv \int_0^C \frac{C - t}{C} dG(t) = G(C) - \frac{1}{C} \int_0^C t dG(t). && (1.6) \end{aligned}$$

Let $h$ be the unknown sampling density of $C = B - A$, supported on $[0, \tau]$. Then the density of the data $Y = (C, \Delta)$ is

$$p_G(c, \Delta) = F_G(c)h(c)\Delta + (1 - F_G(c))h(c)(1 - \Delta).$$

**Exercise 1.** Verify for yourself that $F_G$ is a cumulative distribution function. This is not surprising since if $B - A$ increases, then it gets more and more likely that $I, J \in [A, B]$, i.e. that the partner becomes infected before time $B$. Notice that because $F_G$ is a distribution function this is just a *submodel* of the current status data model. This means that we can estimate $F_G$ using the estimator as proposed for current status data based on the representation:

$$F_G(c) = E(\Delta \mid C = c) = P(\Delta = 1 \mid C = c).$$

Hence estimating $F_G(c)$ is like estimating a monotonic regression which can be carried out with the "Pool Adjacent Violator Algorithm". I would like you to show that $G$ is identifiable; tell me how you can find $G$ if I give you $p_G(c, \Delta)$. Since you already know how to go from $p_G$ to $F_G$ it remains to show how $F_G$ identifies $G$.

**Answer**: For simplicity, define $F = F_G$ and note that

$$F(c + h) - F(c) = G(c + h) - G(c) - \left( \int_c^{c+h} \frac{t}{c} dG(t) - \frac{1}{c^2} \int_0^c t dG(t) h + O(h^2) \right).$$

Thus

$$f(c) \equiv \lim_{h \to 0} \frac{F(c + h) - F(c)}{h} = \frac{1}{c^2} \int_0^c t dG(t). \quad (1.7)$$

We conclude that $F$ is differentiable with derivative $f$, independently of the smoothness of $G$. If $G$ is differentiable at 0 with derivative $g(0)$, then $f(0+) = g(0)/2$. So in this case $f$ is uniformly bounded on $[0, \infty)$. Moreover, for $k = 1, 2, \ldots$, if $g \in C^{(k-1)}[0, \tau]$ and $g(c)/c^k$ is bounded at 0+, then $f \in C^{(k)}[0, \tau]$; so, roughly speaking, $F$ is one degree smoother than $G$.

We are concerned with the question of how well $G$ or functionals of $G$ can be estimated from the data, assuming that $G$ is completely unknown. As indicated by Jewell, Malani, Vittinghoff (1994), and simply following from (1.38) and (1.7), we have the following direct relation between $F$ and $G$:

$$G(c) = \frac{d}{dc}\left(cF(c)\right) = F(c) + cf(c). \tag{1.8}$$

Since the doubly censored current status model represents a submodel of the singly censored current status case, with $F = F_G$ replacing $G$, it follows that we can estimate $F_G$ directly using current status methods. From current status data we saw that estimation of $F_G$ is like estimation of a density and hence this relation shows that, for doubly censored current status, data estimation of $G$ is like estimation of a derivative of a density.

This suggests estimating $G$ with:

$$G_n(c) = \tilde{F}_n(c) + cf_n(c), \tag{1.9}$$

where $f_n$ is the derivative of a smooth $\widetilde{F}_n$. It follows that we should smooth $\widetilde{F}_n$ so that we obtain an optimal rate of convergence for $f_n$.

Suppose now that we are interested in estimating $\int R(x)dG(x)$ for some known function $R(x)$. If $R_1(x) = R(x) - xr(x) \in L^2(F_G)$ and $\lim_{x\to\infty} R(x)/x = 0$, then we have:

$$\begin{aligned}
\mu(G, R) &\equiv \int RdG = \int \frac{R(t)}{t} tdG(t) \\
&= \int_0^\infty \left( \int_t^\infty \frac{R(x) - xr(x)}{x^2} dx \right) tdG(t) \\
&= \int R_1(x) \left( \frac{1}{x^2} \int_0^x tdG(t) \right) dx \\
&= \int R_1(x) dF_G(x).
\end{aligned} \tag{1.10}$$

At the second line we applied Fubini's theorem using that $\int_t^\infty R_1(x)/x^2 dx$ is finite and $\lim_{x\to\infty} R(x)/x = 0$. $\int_t^\infty R_1(x)/x^2 dx < \infty$ is shown as follows. Because $f_G(x) = 1/x^2 \int_0^x tdG(t)$ we have

$$\int_t^\infty \frac{R_1(x)}{x^2} dx = \int_t^\infty \frac{R_1(x)}{\int_0^x tdG(t)} f_G(x) dx.$$

Using that $\int_0^x tdG(t) > 0$ $F_G$ a.e. and $\int R_1^2 dF_G < \infty$, it follows that the last term is finite.

Suppose that $G$ has support $[0, \tau]$. Then $\int RdG = \int R_\tau dG$, for any function $R_\tau$ which equals $R$ on $[0, \tau]$, but which satisfies the constraint $R_\tau(x)/x \to 0$ if $x \to \infty$. In other words, we have

$$\int R(x)dG(x) = \int R_{1\tau}(x)dF_G(x),$$

where $R_{1,\tau}(x) = R_\tau(x) - xr_\tau(x)$. It is a known fact that if $F_n$ is a current status data estimator of $F_G$, then $\sqrt{n}(\int R_{1,\tau} d(F_n - F)$ is asymptotically normal with mean zero and variance

$$\int \frac{F_G(x)(1-F_G(x))}{h(x)} r_{1,\tau}^2(x)dx < \infty, \tag{1.11}$$

where $r_{1,\tau}$ is the derivative of $R_{1,\tau}$. Hence it is of interest to determine a $R_{1,\tau}$, satisfying the condition above, so that (1.11) is mininal. Derick Petersen showed in class using variational analysis that there indeed exists such an optimal solution, which of course depends on $G, h$. In practice one should use an estimate of this optimal solution. In this way one obtains a locally efficient estimator; it is always asymptotically normal and if we estimate the optimal $R_{1,\tau}$ consistently, then the estimator is efficient.

**Example 1.3.6** (Random Truncation Model). Suppose one is interested in estimating the survival time $T$ of a diabetes patient measured from the onset of the disease. For this purpose one follows all the diabetes patients entering a certain hospital during the period 1973-1994. Let $U$ be the chronological time of the onset of the disease for a patient. Then the patient will only be part of the sample if $U + T > 1973$, i.e. if $T > C \equiv 1973 - U$. In other words, our study will provide us with n i.i.d. observations $(T_i, C_i)$ from the conditional distribution of $(C, T)$, given $C \leq T$. It is assumed that $T$ and $C$ are independent. So the density of the data is:

$$p(c,t) = \frac{1}{\alpha} f_T(t) g_C(c) I(c \leq t),$$

where $\alpha = P(C \leq T) = \int G(t)dF(t)$.

**Exercise 2.** The question is if $F$ is identifiable. In other words, can we express $F$ as a function of the distribution of the data. The proof is of the same nature as the proof we gave in the univariate censoring model. Let me give some hints. Firstly show that

$$\frac{P(T \in dt)}{P(T > t)} = \frac{P(T \in dt, C < t)}{P(T > t, C < t)}$$

can be expressed in the distribution of $(C, T)$, given $T \geq C$ (the data). Secondly, use the already derived result that the survival function $S(t)$ is a product integral of the hazard $P(T \in ds)/P(T > s)$. Write down the estimator resulting from this representation of the survival function in terms of the distribution of the data. Can you generalize this estimator to right-censored truncated data?

## 1.4 Ad hoc method for estimation.

Suppose that we are interested in estimation of $\Psi(\theta) \in D$, where $D$ is some linear space. Moreover, assume that we have been able to construct the mapping $V : \mathcal{M} \to D$ so that $\Psi(\theta) = V(P_\theta)$ for all $\theta \in \Theta$. In this section we consider estimators of $\Psi(\theta)$ which are obtained by substitution of the empirical distribution or a smoothed empirical distribution (integrated density estimator) into the *representation* $V$ of $\Psi(\theta)$. In other words, we estimate $\Psi(\theta)$ with

$$V(P_n) \text{ or with } V(\tilde{P}_n),$$

where $\tilde{P}_n$ is a smoothed empirical distribution function.

Firstly, one should notice that in order to do this one needs to be able to define $V$ on $P_n$ or $\tilde{P}_n$. Since $V$ is in principle only defined on the model $\mathcal{M}$ this is not necessarily possible.

Fortunately, it is often possible. For example, if one is concerned with estimation of the mean $\mu = \int x dP_0(x) = V(P_0)$, then $V(P_n) = \bar{X}$ is well defined. Of course, more general this is true for estimation of any parameter $V(P_0) = \int f(x) dP_0(x)$ with $V(P_n) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. Before we construct estimators in this way in the examples above we want to make clear that the choice of representation $V$ plays a role in the behavior of the estimator.

### 1.4.1 Representations.

Above we defined $\Psi(\theta)$ to be identifiable if it can be represented as a function $V$ of the distribution $P_\theta$ of the data for all $\theta \in \Theta$. Therefore $V$ is often called a *representation* of $\Psi(\theta)$. A representation $V$ is often not unique once you define this function $V$ on distributions which do not belong to the model $\mathcal{M}$ and hence the choice of the representation determines often the optimality of the estimator $V(P_n)$. This can already be demonstrated with a simple parametric example.

**Example 1.4.1** The angle $\theta$ at which electrons are emitted in muon decay has a distribution with the density

$$f(x \mid \alpha) = \frac{1+\alpha x}{2},\ -1 \le x \le 1 \text{ and } -1 \le \alpha \le 1,$$

where $x = \cos(\theta)$. Notice that $\alpha = 3\mu$, where $\mu$ is the mean of $X$. This shows that $\alpha$ is identifiable and if we base our estimator on this representation we obtain $\alpha_n = 3\bar{X}$, i.e. this is the method of moments estimator for this problem. We could also represent $\alpha$ as the following function $V$ of the distribution of $X$, say $F$:

$$\alpha = \max^{-1} \int \log(f(x \mid \alpha)) dF(x).$$

By the well known Jensen inequality (i.e. concavity of the log) it follows that if $f = f(\cdot \mid \alpha_0)$, then $\alpha_0 = V(F)$, i.e. indeed each $F$ corresponding with the parametric family $f(x \mid \alpha)$ leads to a unique $\alpha$. If we now base our estimator of $\alpha$ on this representation we obtain $\alpha_n = V(F_n)$, where $F_n$ is the empirical distribution function of $X_1, \ldots, X_n$ and hence that $\alpha_n$ is the maximizer of the empirical loglikelihood (i.e. it is the MLE):

$$\alpha_n = \max^{-1} \frac{1}{n} \sum_{i=1}^n \log(f(X_i \mid \alpha)).$$

It is well known that MLE are efficient and that method of moments estimators are often not efficient.

It is not always that obvious that one representation leads to a better estimator than the other. A detailed comparison of estimators is often necessary and one often concludes that the differences are strongly dependent on the actual distibution we are sampling from. Unfortunately, but certainly not always, (there are very nice examples of simple estimators which perform better than complicated efficient estimators!) the best estimators are complicated functions of the data (this is already shown by the example above). We can already say here that the estimators based on the representations derived in the univariate censoring, current status data, and doubly censored current status data, random truncation, are all asymptotically optimal, which might be considered as a surprise because of our ad hoc way we derived these representations.

### 1.4.2 Examples.

**Example 1.4.2 density estimation.** In this example we are concerned with estimation of $V(F) = f = dF/dx$, where $F$ is the cumulative distribution function of $X$ and $f$ its derivative. It is clear that $V$ cannot be defined on the empirical cumulative distribution $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$. However, it is well defined on a smoothed $\tilde{F}_n(x) = \int_0^x f_n(x)dx$, where $f_n$ is a density estimator of $f$.

**Example 1.4.3 nonparametric regression.** Here we are concerned with estimation of $m(x) = V(F_{X,Y})(x) = \int yf(y \mid x)dy$. The same remark as above holds here and we can naturally estimate $m(x)$ with $\hat{m}(x) = \int yf_n(y \mid x)dy$, where $f_n(y \mid x)$ is a density estimator of $f(y \mid x)$.

**Example 1.4.4 Current status data.** Here we are concerned with estimation of

$$F(x) = V(P_{F,G})(x) \equiv \frac{p(x,1)}{p(x,1)+p(x,0)},$$

where $p(x,\delta)$ is the derivative (density) of $x \to P(C \leq x, \delta)$. Again, the same remark as above holds here and we can naturally estimate $F(x)$ with $V(\tilde{P}_n)$, where $\tilde{P}_n(x,\delta) = \int_0^x p_n(u,\delta)du$ with $p_n(x,\delta)$ a density estimator of $p(x,\delta)$:

$$F_n(x) = \frac{p_n(x,1)}{p_n(x,1)+p_n(x,0)}.$$

**Example 1.4.5 Univariate censoring.** Substitution of the empirical cumulative distribution function in the representation for $\lambda(ds)$ provides us with

$$\lambda_n(ds) = \frac{\sum_{i=1}^n I(\widetilde{T}_i \in ds, \Delta_i = 1)}{\sum_{i=1}^n I(\widetilde{T}_i \geq s)},$$

which provides us with $S_n(t) = \prod_{(0,t]}(1-\lambda_n(ds))$ as an estimator of $S(t)$. $S_n(t)$ is the well known Kaplan-Meier estimator.

**Exercise 1.** How would you estimate $G$ based on the representation of $G$ you found in the doubly censored current status model example?
**Exercise 2.** Similarly, how do you estimate $F$ in the random truncation model?

## 1.5 Current Status Model with (time-dependent)-covariates.

In this section we will extend the current status model to the case where time-dependent and or time-independent covariates are present and we will present an ad hoc estimator which might be a good candidate in practice. We will consider a practical example. On our way we discuss some important identifiability issues in censored data models (censoring should be noninformative in the sense as described below)

Consider a population of persons who developed lung cancer in their live. For a randomly drawn person we let $T$ be the time of onset of lung-cancer. Moreover, let $L(\cdot) = (L_1(\cdot), L_2(\cdot))$ be a vector of two functions. $L_1$ keeps track of the amount of cigarettes smoked. So $L_1$ is an

increasing function, though it can be flat during a period where the person does not smoke. Let $L_2$ measure the stage in the developement of the cancer. So $L_2(t) = 0$ for $t \leq T$ and $L_2$ is increasing for $t > T$. As one can expect the time of onset $T$ will never be exactly observed. Similarly, the function $L$ will not be completely observed. Let $C$ be a time at which the person is tested on the presence of lung cancer. This time is often refered to as a monitoring time. We will observe $C$, we observe if $T$ already occurred (in other words, if cancer is present or not) and we observe $L_2(C)$ (in other words, how far the cancer has developed) and we observe the function $L_1$ up till time $c$ (i.e. we observe his smoking history). We will denote the smoking history up till time $c$ with $\bar{L}_1(c)$. So in symbols we can represent our observation as follows:

$$Y = (C, \Delta = I(T \leq C), L_C), \tag{1.12}$$

where $L_C = (\bar{L}_1(C), L_2(C))$. Using $n$ of i.i.d. copies of $Y$ we are concerned with estimation of $F_T$, the marginal distribution of $T$, or a certain characteristic of $F_T$ represented by $\int R(t)dF(t)$ for some function $R$; e.g. if $R(t) = t^k$, we obtain the moments of $F_T$. Notice that if $r(x) \equiv d/dxR(x)$, then by integration by parts it follows that:

$$\int_0^\infty r(t)(1 - F(t))dt = R(t)(1 - F(t) \mid_0^\infty + \int_0^\infty R(t)dF(t).$$

We will assume that $\lim_{x\to\infty} R(x)(1 - F(x)) = 0$ (which is practically always true). Then it follows that we have

$$\int R(t)dF(t) = \int_0^\infty r(t)(1 - F(t))dt + R(0).$$

Therefore it suffices to construct an estimator for $\int r(t)(1 - F(t))dt$.

### 1.5.1 Censored data model, Coarsening at random assumption.

A useful way to think about this data structure is as censored or missing data. We call $(C, T, L)$ the complete data. The complete data has a distribution which is determined by the joint distribution $F_{T,L}$ of $(T, L)$ and of the conditional distribution $G(\cdot \mid (T, L))$ of $C$ given $(T, L)$. We only observe a many to one mapping of the complete data, namely the one described by (1.12). Data structures which can be represented in this way are called *censored data* or *missing data*.

Each observation $Y = y = (c, \delta, l_c, z)$ tells us that $(t, l)$ lies in a certain region, namely in $\{(T, L) : T \leq c, L_C = l_c\}$ if $\Delta = 1$ and $\{(T, L) : T \geq c, L_C = l_c\}$ if $\Delta = 0$. The distribution of $(T, L)$ will in general be unidentifiable from the observations $Y$ if the fact that $C = c$ is informative about where $(T, L)$ lies in these regions (except if we make assumptions on $C$, given $(T, L)$). In other words, if $C = T + e$ for some error term $e$ and we do not know this, then we are in deep trouble. If $C$ is noninformative about the place of $T$ in the implied region, then one says that the conditional distribution of $C$, given $T, L$, implies a *coarsening at random* of $(T, L)$.

Let $g(c \mid T = t, L = l)$ be the density w.r.t. the Lebesque measure of the conditional distribution of $C$, given $(T, L)$. Then the fact that $C = c$ in only noninformative about the location of $(T, L)$ in the by $Y = y$ implied region if $(t, l) \to g(c \mid (t, l))$ is constant over this region. In other words, we need

$$g(c \mid T = t, L = l) = h(c, \delta, l_c), \tag{1.13}$$

for some function $h$ of the data $y$.

It is not hard to verify heuristically that this assumption implies that the density of the data can be represented as:

$$p(y) = p_{F_{T,L}}(y)h(y),$$

where $p_{F_{T,L}}$ does only depend on $F_{T,L}$. In words, this says that the density of the data factorizes in the parameter of interest $F_{T,L}$ and the nuisance parameter (this is a name for the parameter which is not of interest) $g(c \mid T, L)$. It is the factorization of the likelihood which makes $p_{F_{T,L}}(y)$ identifiable in the case that $h$ can be recovered from the data, since

$$p_{F_{T,L}}(y) = p(y)/h(y).$$

In this example, we could assume that the decision to monitor the patient at a time $c$ depends on the amount he has smoked so far by for example modelling the hazard with the Cox-proportional hazards model:

$$\lambda(c \mid (T, L)) = \lambda_0(c) \exp(-\beta l_1(c)).$$

Recall now that a hazard is a ratio of $g(c \mid (T, L))$ and the survival function $\bar{G}(c \mid (T, L)) = P(C > c \mid (T, L))$ and recall that

$$\bar{G}(c \mid (T, L)) = -\exp\left(-\int_0^c \lambda(u \mid (T, L))du\right).$$

This implies that

$$g(c \mid (T, L) = \lambda(c \mid (T, L)) \exp(-\int_0^c \lambda(u \mid (T, L))du,$$

which is indeed only a function of $c, \bar{L}_1(c)$ and hence satisfies (1.13). Notice that $g(c \mid (T, L))$ uses the whole smoking history of the patient. Notice also that now $g(c \mid (T, L))$ can be estimated from the data by using the (standard) Cox-partial likelihood estimators of $\beta$ and $\lambda_0$. Hence $g$ is identifiable.

### 1.5.2 A simple estimator.

Let's now propose an estimator for $\mu = \int r(t)(1 - F(t))dt$. Suppose for the moment that $g(c \mid (T, L))$ is known. Consider the following estimator of $\mu$:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(\Delta_i = 0)r(C_i)}{g(C_i \mid (T_i, L_i))}. \tag{1.14}$$

Because $g(C_i \mid (T_i, L_i)) = h(C_i, \Delta_i, L_{i,C_i})$ we have that we actually observe $g(C_i \mid (T_i, L_i))$ which shows that (1.14) is indeed an estimator. Notice now that (for convenience we set $X = (T, L)$)

$$\begin{aligned} E\left(\frac{I(\Delta = 0)r(C)}{g(C \mid X)}\right) &= EE\left(\frac{I(\Delta = 0)r(C)}{g(C \mid X)} \mid X\right) \\ &= E\left(\int_0^\tau I(T \geq c)r(c)dc\right) \\ &= \int_0^\tau r(c)(1 - F(c))dc. \end{aligned}$$

This proves that (1.14) is an unbiased estimator.

This suggest to take the following estimator of $\int r(t)(1-F(t))dt$:

$$\mu_n^0 = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\Delta_i = 0)r(C_i)}{g_n(C_i \mid X_i)}. \tag{1.15}$$

The estimator $\mu_n^0$ is asymptotically equivalent with $\int r(t)(1-F_n)(t)dt$, where

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i \frac{1}{h} K((C_i - t)/h)}{g_n(C_i \mid X_i)}, \tag{1.16}$$

where $K$ is a kernel and $h$ a bandwidth which converges quickly enough to zero.

**Exercise.** Try to imitate this section, but now for data of the form $(C, T \wedge C, L_c)$. In other words, we now have right censored data on $T$ and we observe a part $L_c$ of a covariate process $L$. For example, let $L_c$ be a relevant history of the patient up till time $C$, and let $T$ be a survival time of interest.

### 1.5.3 Some interesting remarks about this simple estimator.

Suppose that in the preceding example $C$ is independent of $T, L$. Then it might seem natural to estimate $g(c \mid (T, L)) = g(c)$ with a kernel density estimator based on $C_1, \ldots, C_n$; i.e. we just set the coefficients in the Cox-proportional hazards model equal to zero since we know that they are zero. This might seem appropriate, but their is clear theoretical evidence that one should (if $n$ large enough) put all relevant covariates from $L$ for $T$ into $g(c \mid (T, L))$ and just *ignore* the information that the coefficients corresponding with the covariates are zero! We can summarize this statement in a slogan: "In models where the likelihood factorizes in the part of interest and the part containing the nuisance parameter, *ignoring information about the nuisance parameter will improve asymptotic efficiency of the estimator*".

Let's illustrate this with an example. Consider the simple current status data model; we observe $(C, I(T \leq C))$ and $C$ and $T$ are independent. Assume that we know the density $g$ of $C$. Then a natural estimator of $F$ based on the representation $F(c) = p(c, 1)/g(c)$ seems:

$$F_n(c) = \frac{p_n(c, 1)}{g(c)}.$$

This estimator is inefficient (when used for estimating $\int R(t)dF(t)$) while if we estimate $g$ (just ignoring that we know $g$) with a kernel density estimator $g_n$, then the estimator is efficient.

### 1.5.4 Heuristic explanation of why ignoring information about nuisance parameter improves efficiency.

At a later stage we will give an algebraic proof or our slogan. In this section we give a heuristic explanation. We start with a simple example.

**Example 1.5.1** Let $(X, Y) \in \mathbb{R}^2$ be a joint random variable. Assume that $Y = \beta X + e$, where $e \sim N(0, \sigma^2)$. We observe $n$ i.i.d. copies of $(X, Y)$. We are concerned with estimation of $\beta$. This is clearly a case where the density of the data $(X, Y)$ factorizes since

$$f(y, x) = f_\beta(y \mid x) f_X(x),$$

where $f_\beta(y \mid x)$ is the part of the likelihood of interest and $f_X$ is the nuisance parameter.

Let $\mu_2^n \equiv 1/n \sum_{i=1}^n X_i^2$, $\mu_{xy}^n = 1/n \sum_{i=1}^n X_i Y_i$, $\mu_2 = EX^2$ and $\mu_{xy} = EXY$. Suppose that we know $f_X$ and hence in particular we know $\mu_2$. It seems now reasonable to estimate $\beta$ with

$$\beta_n^* = \frac{\mu_{xy}^n}{\mu_2}.$$

We could now ignore the knowledge of $\mu_2$ and estimate it with $\mu_2^n$. Then we get a new estimator:

$$\beta_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2}.$$

It is well known and easy to verify that $\beta_n$ is the maximum likelihood estimator of $\beta$ and it equals the least squares estimator. In other words,

$$\beta_n = \overset{-1}{\min} \sum_{i=1}^n (Y_i - \beta X_i)^2.$$

Notice that $\beta_n$ is a real average of the $Y_i$'s since the weights in front of each $Y_i$ add up till one. On the other hand, $\beta_n^*$ is not an average which causes a bias; $E\beta_n^* \neq \beta$. A deviation of $\mu_2^n$ from $\mu_2$ causes variability in the estimator $\beta_n^*$.

In a way we could say that the estimator $\beta_n^*$ does not really use that $Y_i$ belongs to $X_i$, but instead it replaces the role of $X_i$ by some characteristic of the whole $X$ population. In other words, we are not allowed to forget that $Y_i$ is really coming from the conditional distribution of $Y$, given $X = X_i$!

Let's try to determine the loss in variance of $\beta_n^*$ relative to the variance of $\beta_n$. It is a well known fact that $\beta_n$ is asymptotically efficient among the class of all asymptotically normal estimators. Firstly, since $\beta_n$ is a maximum likelihood estimator its limiting variance equals the information bound which equals 1 over the variance of the score w.r.t. $\beta$. The score for $\beta$ is given by:

$$S_\beta(X, Y) = \frac{1}{\sigma^2}(Y - \beta X)X.$$

Since a score has mean zero this equals

$$ES_\beta(X, Y)^2 = \frac{1}{\sigma^4} E(X^2(Y - \beta X)^2).$$

By first taking a conditional expectation w.r.t. $X$ we obtain:

$$E(X^2(Y - \beta X)^2) = E(X^2 E(Y - \beta X)^2 \mid X)) = E(X^2\sigma^2) = \sigma^2\mu_2.$$

Hence the information bound is given by:

$$\frac{\sigma^2}{\mu_2}.$$

Let's now compute the limiting variance of $\sqrt{n}(\beta_n^* - \beta)$. Let $g(z) = z/\mu_2)$. Then $\beta = g(\mu_{xy}) = \mu_{xy}/\mu_2$ and $\beta_n^* = g(\mu_{xy}^n)$. Notice that $d/dz g(z) = 1/\mu_2$. So by the delta-method we have

$$\begin{aligned} \beta_n^* - \beta &= g(\mu_{xy}^n) - g(\mu_{xy}) \\ &\approx \frac{1}{\mu_2}(\mu_{xy}^n - \mu_{xy}). \end{aligned}$$

So the limiting variance of $\beta_n^*$ is given by

$$\frac{1}{\mu_2^2}\text{VAR}(XY).$$

We have

$$E(XY) = E(XE(Y \mid X)) = E(\beta X^2) = \beta\mu_2.$$

Furthermore we have

$$E(X^2Y^2) = E(X^2E(Y^2 \mid X)) = E(X^2(\sigma^2 + \beta^2X^2)) = \sigma^2\mu_2 + \beta^2E(X^4).$$

So we conclude that

$$\text{VAR}(XY) = E(X^2Y^2) - E^2(XY) = \sigma^2\mu_2 + \beta^2\text{VAR}(X^2).$$

So the limiting variance of $\beta_n^*$ is given by:

$$\frac{\sigma^2}{\mu_2} + \frac{\beta^2}{\mu_2^2}\text{VAR}(X^2).$$

Hence we conclude that in this example ignoring information about the nuisance parameter provides us with a gain in efficiency given by:

$$\frac{\beta^2}{\mu_2^2}\text{VAR}(X^2).$$

It is interesting to see that exactly the variance of $\mu_n^2$ is causing the efficiency loss. This corresponds with our heuristic explanation.

We can extend this example by making it nonparametric. Suppose we are concerned with estimation of $m(x) = E(Y \mid X = x) = \int yf(y \mid x)dy$ nonparametrically. based on $(X_i, Y_i)$, $i = 1, \ldots, n$. Assume that $f(x)$ is known. Then using $f(y \mid x) = f(y, x)/f(x)$ it might seem natural to estimate $f(y \mid x)$ with

$$m_n^*(x) \equiv \frac{\frac{1}{nh}\sum_{i=1}^n Y_iK((X_i - x)/h))}{f(x)}.$$

Notice that if $f(x)$ deviates from $f_n(x) = 1/nh\sum K((X_i - x)/h)$, then $m_n^*(x)$ is not an average anymore of $Y_i$'s with $X_i$ in the neighborhood of $x$ since the coefficients in front of the $Y_i$'s do not add up till one anymore. In other words, just as in the simple linear regression example, using $f(x)$ instead of $f_n(x)$ creates a bias (extra variability). So also here we see that $m_n^*$ does not really use the knowledge that $Y_i$ is drawn from $f(y \mid X = X_i)$ (i.e. $X_i$ and $Y_i$ belong to each other). On the other hand

$$m_n(x) = \frac{\frac{1}{nh}\sum_{i=1}^n Y_iK((X_i - x)/h))}{f_n(x)}$$

is a perfectly natural estimator. This can be made rigorous by comparing approximate variances of the two estimators as in our simpel example above.

We could give this example a practical implementation. Suppose we have a population of man. Let $X$ be age and $Y$ be a measure for baldness of a randomly drawn person from this population. We want to estimate $E(Y \mid X = x)$; i.e. given a man's age what is the expectation of the measure for baldness. It is intuitively already clear that if we have drawn the sample $(X_i, Y_i)$, then it is absolutely irrelevant for estimation of $E(Y \mid X = x)$ if the $X_i$'s are a representative sample for the population or not. All what matters is that there are a number of $X_i$'s close to $x$, because these persons have to provide us with information about $E(Y \mid X = x)$. So for example, assume that the probability $P(X \in dx)$ is small, but by randomness it happens that we obtain a lot of $X_i$ close to $x$. Then that is just great; in other words, in this example, we hope that $f_n(x)$ deviates from $f(x)$. It is clear that the estimator which uses $f(x)$ in the denominator would have been a very bad one to use.

So let's now go back to our current status data example. We observe $(C, \Delta = I(T \leq C))$. Notice that

$$F(c) = E(\Delta \mid C = c).$$

In other words, this example fits perfectly in our nonparametric example above; here $X = C$ and $Y = \Delta$ and $m(x) = F(x)$ is known to be monotone. Hence by the same reason we do not care if the $C_i$'s form a representative sample of $g(c)$, but we really want to use the $C_i$'s as a given set of covariates on which we condition in the sense that $\Delta_i$ is drawn from $P(\Delta \mid C = c_i)$.

Finally, let's now generalize our insight to a general CAR-missing data model. So suppose that we have $n$ i.i.d. observations $(C_i, \Phi(X_i, C_i))$ for some known $\Phi$ and assume that $X \sim F$ and $g(c \mid X = x) = h(c, \Phi(c, x))$ for some function $h$. In such missing data models the density of $(C, \Phi(C, X))$ factorizes in a $F$-part and a $g$-part. This is seen as follows (heuristically). In such missing data models the observation $\Phi(X, C) = Y$ tells us that $X \in D_1(Y) = \{x : \Phi(x, c) = Y\}$ and $C \in D_2(Y)$ (what are $D_1(y)$ and $D_2(y)$ in the current status example, and univariate censoring example?). Now it follows for any $c \in D_2(y)$:

$$\begin{aligned} P(C \in dc, \Phi(X, C) \in dy) &= \int_{D_1(y)} g(c \mid x) dF(x) \\ &= \int_{D_1(y)} dF(x) h(c, y). \end{aligned}$$

Here $h$ is the nuisance parameter and $\int_{D_1(y)} dF(x)$ is the part of interest. Since by CAR $C = c$ is non informative about the exact location of $x \in D_1(y)$ and the $F$-part of the likelihood is really

$$P(\Phi(X, c) = y \mid C = c).$$

In other words, we want to estimate functionals of the conditional density of the random variable $Y = \Phi(X, C)$, given $C = c$. Hence we have put the problem in the same framework as in our nonparametric regression problem so that the same arguments tell us that for optimal estimation one really wants to condition on the observed $C_i$'s instead of using characteristics of the distribution of the $C_i$'s.

Finally, we will explain why our slogan applies to any model with an orthogonal parametrization, where orthogonal means that information on the nuisance parameter does not provide one with more information on the parameter of interest. Suppose the model is indexed by two real valued parameters $\theta$ and $\eta$. We are concerned with estimation of $\theta$. It is known that the scores for $\theta$ and $\eta$ are orthogonal. In this case we know that the information bound for estimation of $\theta$ does not depend on knowing $\eta$. This is seen as follows:

**Example 1.5.2** $X \sim p_{\theta,\eta}$. Let $S_\theta = \frac{d}{d\theta}\log(p_{\theta,\eta})(X)$ and $S_\eta = \frac{d}{d\eta}\log(p_{\theta,\eta})(X)$ be the scores obtained by differentiating with respect to $\theta$ and $\eta$, respectively. Let $L_0^2(P_{\theta,\eta})$ be the Hilbert space of real valued functions of $X$ with mean zero and finite variance. Notice that $S_\theta, S_\eta \in L_0^2(P_{\theta,\eta})$. A Hilbert space is a linear space with an inner product. In this case the natural inner product is given by:

$$\langle f, g\rangle_{P_{\theta,\eta}} \equiv \int f(x)g(x)dP_{\theta,\eta}(x) = \mathrm{E}_{P_{\theta,\eta}} fg = \mathrm{Cov}(f(X), g(X).$$

This inner product implies a norm by

$$\|f\|_{P_{\theta,\eta}} \equiv \sqrt{\langle f, f\rangle_{P_{\theta,\eta}}}.$$

An inner product defines an orthogonality relation by: $f \perp g$ if and only if $\langle f, g\rangle = 0$. In Hilbert spaces one can work as if one works in $\mathbb{R}^3$. The projection of $f$ on a function $g$ is given by:

$$\Pi(f \mid g) = \frac{\langle f, g\rangle_{P_{\theta,\eta}}}{\langle g, g\rangle_{P_{\theta,\eta}}} g.$$

This is simply verified by checking that $f - \Pi(f \mid g)$ is orthogonal to $g$.

Suppose one is interested in estimation of $\Phi(\theta)$ for some function $\Phi$. If $\eta$ is known, then the Cramér-Rao lower bound for the variance of an unbiased or for the asymptotic variance of an asymptotically unbiased estimator is given by:

$$\frac{\Phi'^2(\theta)}{\|S_\theta\|^2_{P_{\theta,\eta}}},$$

while if $\eta$ is unknown, then it is given by:

$$\frac{\Phi'^2(\theta)}{\|S_\theta - \Pi(S_\theta \mid S_\eta)\|^2_{P_{\theta,\eta}}}.$$

We call a parametrization orthogonal if $\Pi(S_\theta \mid S_\eta) = 0$ which is equivalent with $S_\theta \perp S_\eta$ which is equivalent $\mathrm{Cov}(S_\theta(X), S_\eta(X)) = 0$.

The orthogonality means that knowledge about $\eta$ is (asymptotically) irrelevant for estimation of $\theta$. Suppose that $\theta_n^0$ is an estimator of $\theta$ which relies on the known fact that $\eta = \eta_0$. In other words, if the sample of observations has features which correspond with a value $\eta \approx \eta_0$, then the estimator is having a good performance, but if the sample happens to have features which are not representative for the actual $\eta_0$ value, then the estimator will have a bad performance. Hence this estimator responds on absolute irrelevant features in the data. So this estimator will have extra variability caused by the fact that samples will not always exactly represent the $\eta = \eta_0$ value; it has extra variability caused by irrelevant variability of the samples. Hence this estimator can be improved by making him not respond anymore on these irrelevant features in the data.

As an example consider estimation of $\mu$ from an i.i.d. sample on $X \sim N(\mu, \sigma^2)$. The score for $\mu$ and the score for $\sigma^2$ have covariance zero. Hence any feature of the data representing $\sigma^2$, i.e. the value of the sample variance, is asymptotically irrelevant for estimation of $\mu$. Therefore an estimator using the known value of $\sigma^2$, say $\mu_n = S^2/\sigma^2\bar{X}$, has an extra variability causes

by irrelevant features in the data and hence is inefficient. An even more obvious example is the following; suppose that we observe n i.i.d. observations on $(X, Y)$, where $X$ and $Y$ are independent. We want to estimate the mean of $X$ and assume that we have an estimator of this mean which uses the $Y$ sample. It will be clear that the estimator can be improved by making him ignore the $Y$-sample; the $Y$-sample creates unnecessary extra variability in the estimator.

## 1.6 Estimation of the bivariate survival function.

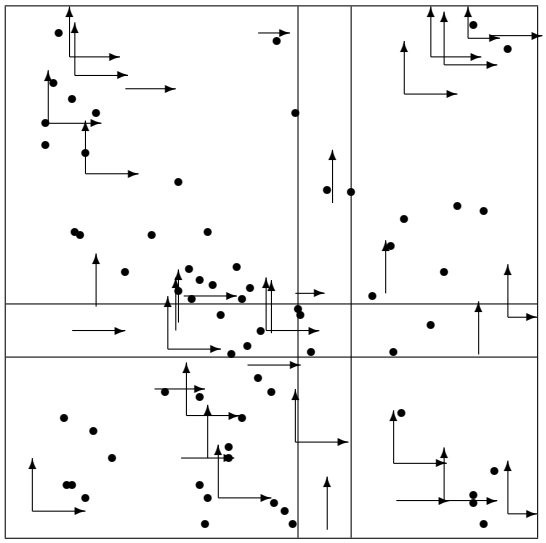
Assume that we would like to get some knowledge about the bivariate lifetime distribution $S_0$ of a population of twins with a certain disease. Let $T = (T_1, T_2)$ be the corresponding bivariate survival time of a randomly drawn twin from this population and assume that each twin is subject to right-random censoring by an irrelevant censoring vector $C = (C_1, C_2)$. A censoring time represents the length of time we are able to observe survival; $C_1$ might represent the end of an observation study or experiment, it might be the age of twin1 at which he/she dies by an irrelevant factor (e.g. car accident), or the age of twin1 at which he/she leaves the study (e.g. emigration). The i.i.d. observations on $n$ twins are now

$$Y_i \equiv (T_i \wedge C_i, I(T_i \leq C_i)) \,,$$

with components given by:

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

In other words, for twin1 we observe the minimum of censoring and survival and we observe if this minimum is the actual survival time of interest, and similarly for twin2. Each bivariate randomly right-censored observation tells us that $T = (T_1, T_2)$ has fallen in a region in the plane where this region is a dot if both $T_1$ and $T_2$ are observed (uncensored), it is a half-line if only one of the survival times $T_i$ is right-censored (singly-censored) and it is a right-upper quadrant if both $T_1$ and $T_2$ are right-censored (doubly-censored). Therefore the data can be nicely presented in a picture: $\bullet$ =uncensored, $\rightarrow$=censored. (disregard the strips, here)



Right Censored Bivariate Data

In this paper the region for $T$ implied by observation $Y$ will be denoted with $B(Y)$. The estimation problem is to estimate the bivariate survival function $S_0(t_1, t_2) \equiv P(T_1 > t_1, T_2 > t_2)$ of the survival times $T_1$ and $T_2$ of twin1 and twin 2, using the $n$ i.i.d. observations; i.e. the information $T_i \in B(Y_i)$, $i = 1, \ldots, n$.

The last decennia there has been a lot of interest in using semiparametric models which means that one does not want to restrict $S_0$ to a parametric class of models, but that one is willing to make some less stringent assumptions like, for example, symmetry in $T_1$ and $T_2$. If one has a lot of data, then this certainly seems to be the safest approach for estimation. In this paper we study the model where one does not assume anything at all about the shape of $S_0$. This is also the model which has been extensively studied in the literature.

It is often a too conservative model for practical purposes. However, a nonparametric model for $S_0$ is the right model for studying the phenomenon of bivariate right-randomly censoring and the discussed estimators can be easily extended to models where one makes semiparametric assumptions on $S_0$ and the same conclusions as we will make in this paper about the estimators and the NPMLE can be expected to hold in smaller (but still infinite dimensional) models.

The usual NPMLE in this model is not consistent for continuous data. As a consequence, there has been paid a lot of attention to constructing ad hoc explicit estimators in the literature. In the next section we restrict our attention to the two explicit estimators with best practical performance (see simulation results of Bakker, 1990, Pruitt, 1993, and Prentice and Cai, 1992a,b), namely Dabrowska's estimator (Dabrowska, 1988, 1989) and Prentice and Cai's estimator (Prentice and Cai, 1992a,b). These estimators are based on the following representation: $S(t_1, t_2) = S_1(t_1)S_2(t_2)R(t_1, t_2)$, where $S_1, S_2$ are the marginal survival functions of the lifetimes $T_1$ and $T_2$, respectively, and $R(t_1, t_2) = S(t_1, t_2)S(0, 0)/S_1(t_1)S_2(t_2)$, using that $S(0, 0) = 1$. Now, the marginals are naturally estimated by the well known Kaplan-Meier estimator using the marginal samples for $T_1$ and $T_2$.

Dabrowska considers $R$ as a cross-ratio of $S$ over the corners of the rectangle $[0, t_1] \times [0, t_2]$; in two by two tables the odds ratio is defined by the same cross product of the four numbers and we know that they can be used to find any correlation between the column and row variable; if the odds ratio is 1, then the row and column effect are independent and a deviation from 1 indicates positive or negative correlation. Notice that if one would observe the empirical fractions of $T_i$'s corresponding with $S(t_1, t_2), S(0, 0), S(t_1, 0)$ and $S(0, t_2)$, then the cross-ratio over these four counts is just as the odds ratio in two by two tables a measure of dependence between the first column ($T_1 > 0$) and the second column $T_1 > t_1$ or between the first row ($T_2 > 0$) and the second row ($T_2 > t_2$). This shows that $R$ is a measure for dependence between the events $T_1 > t_1$ and $T_2 > t_2$; a deviation from 1 being an indication that there is dependence.

The cross-ratio of $S$ has the following multiplicative property: the cross-ratio over the union of two adjacent rectangles equals the product of the cross-ratios of the rectangles. This property tells us that $R(t_1, t_2)$ can be computed iteratively by multiplying the cross-ratios of $S$ over the rectangles of a lattice partition of $[0, t_1] \times [0, t_2]$. As we will see the cross-ratios over infinitesimal small rectangles are insensitive to censoring, just like the univariate hazard in the univariate censoring model, and hence they can be naturally estimated from the observed counts. This will provide us with the Dabrowska estimator.

We can also represent $R(t_1, t_2) = e^{\log(R(t_1, t_2))}$. Noticed that $\log(R)$ is an additive measure $\log(S)$ over the rectangle $(0, t_1] \times (0, t_2]$ giving mass $\log(S)(s_1, s_2) - \log(S)(s_1, 0) - \log(S)(0, s_2) + \log(S)(0, 0)$ to a rectangle $(0, s_1] \times (0, s_2]$, which equals the logarithm of the odds-ratio of $S$ over

this rectangle. So we can compute $\log(R(t_1,t_2))$ by computing $\log(R)$ over small rectangles of a partition of $(0,t_1]\times(0,t_2]$; in other words, we have $\log(R)(t_1,t_2)=\int_{(0,t_1]\times(0,t_2]} d\log(R)(s_1,s_2)$. Notice here the similarity with Dabrowska's approach; Dabrowska uses the multiplicative measure property of $R$ to get grip on estimation of $R$, while here we use the additive property of $\log(R)$. The left-hand side and right-hand side in $R=e^{\log(R)}$ give the same measure to a rectangle. So with some handwaving we have:

$$dR(s_1,s_2)=e^{\log(R(s_1,s_2))}d\log(R)(s_1,s_2)=R(s_1,s_2)d\log(R)(s_1,s_2).$$

By integrating the left and right-hand side over $[0,t_1]\times[0,t_2]$ and taking care of the edge terms ($R(0,0)=R(t_1,0)=R(0,t_2)=1$) we obtain:

$$R(t_1,t_2)=1+\int_{[0,t_1]\times[0,t_2]} R(s_1-,s_2-)d\log(R)(s_1,s_2). \tag{1.17}$$

This is essentially the approach followed by Prentice and Cai. The measure $\log(R)$ over a small rectangle equals the logarithm of the cross-ratio over this small rectangle which could be naturally estimated from the data as mentioned above. Hence $d\log(R)(s_1,s_2)$ can be estimated by plugging in the estimates for the cross ratios used for the Dabrowska estimator. Once one has plugged in this discrete estimator of $d\log(R)$ we can solve for $R(t_1,t_2)$ from (1.17) iteratively over a lattice partition starting from the edges of $[0,t_1]\times[0,t_2]$.

We refer to Gill and Johansen (1990) for the general equivalence between additive (like $\log(R)$) and multiplicative measures (like $R$). It is not surprising that both estimators have a very similar practical behavior. In the next section we show how these explicit estimators are computed and give more on the intuitive background of them.

A NPMLE solves the so called self-consistency equation which can be solved with the EM-algorithm. Each step in the EM-algorithm works in the following nice heuristic manner: give each observation in the picture above mass $1/n$ and the censored observations (lines and quadrants) have to redistribute this mass over their associated region $B(Y)$ for $(T_1,T_2)$ (line or quadrant) according to an estimate of the conditional distribution over this region which is obtained by "listening" to the other observations in the region. If the data is continuous, then the lines do not contain any uncensored observations and hence singly-censored observations do not get information about how to redistribute their mass over their associated lines from the uncensored observations. This explains why the NPMLE is not consistent for continuous data. This problem would not occur if we had chosen a parametric model for $S$ because then the model in combination with the data will tell how the singly-censored observations will have to redistribute their mass $1/n$ over the half-line.

Pruitt (1991) noticed this fact and came up with a modification of the EM-algorithm by telling *himself* how the mass $1/n$ corresponding with singly-censored observations is redistributed, by using kernel-density estimators.

Van der Laan (1992,1995b) proposed an NPMLE based on interval censored singly-censored observations so that the lines in the plane become strips around these lines (as indicated in the picture above), which now contain uncensored observations which will tell how to redistribute the mass $1/n$ over such a strip. This estimator is shown to be asymptotically efficient (van der Laan, 1995b) if one lets the reduction of the data converge to zero if the number of observations converges to infinity. This estimator has the nice heuristic that it is is asymptotically unbiased even if the width of the strips is fixed.

In section 3 we discuss the EM-algorithm in detail and show how to compute Pruitt's modified EM-estimator and the reduced data NPMLE. Also the intuitive background is discussed in more detail.

Finally, we compared the practical performance of these estimators under various levels of censoring and dependence between $T_1$ and $T_2$. It appeared that the reduced data NPMLE works well if one takes care that the strips around the lines are small (width of 0.02 for $n = 200$). The Prentice-Cai and Dabrowska estimator have an excellent practical performance if $T_1$ and $T_2$ happen to be independent or weakly dependent which is due the fact that their representations are directly linking the representation under independence and the representation under dependence. In fact, it has been proved in Gill, van der Laan and Wellner (1995) that these estimators are efficient under complete independence and the simulations show that the Cramér-Rao lower bound is already achieved for samples of $n = 100$.

Then if the dependence or censoring level increases the reduced data NPMLE has a lower variance than the other estimators in the inner area where most data is concentrated, but not at the edge-area. The Prentice-Cai and Dabrowska estimator appear to be hardly distinguishable and both have a good and stable practical performance. Unexpectedly, Pruitt's estimator appeared not to be better than these two explicit estimators and is clearly worse if we are close to independence.

*Practical Advice.* In practice, one should first estimate the correlation coefficient between $T_1$ and $T_2$; for this one could use Dabrowska's estimator for estimation of $E(T_1T_2)$ and the marginal Kaplan-Meier estimators for estimation of the variances of $T_1$ and $T_2$. If the correlation coefficient is larger than 0.2, then it is worthwhile to use the reduced-data NPMLE at areas with a reasonable amount of uncensored observations and one of the explicit estimators at the tail-areas where the NPMLE is unstable. This might cause a strongly non-smooth behavior of the resulting estimator at the boundary of the region where we go from the NPMLE to the explicit estimator. This could be solved by smoothing the estimator at this boundary. Because of the explicitness and robustness of Dabrowska's estimator one can easily (and quickly) construct confidence intervals for this estimator by estimating its limiting variance or using the bootstrap (see Gill, van der Laan and Wellner, 1995, for the explicit limiting distribution and for the bootstrap results). These confidence intervals can now be used for constructing conservative confidence intervals for the reduced data NPMLE because its variance will not be larger than the variance of Dabrowska's estimator. In the last section we discuss the simulation-results in more detail.

### 1.6.1 Direct Estimators.

In this section we cover the estimator introduced by Dabrowska (1988, 1989) and the estimator introduced by Prentice and Cai (1992a,b).

### 1.6.2 The Dabrowska Estimator.

The following derivation of the Dabrowska representation of the bivariate survival function is from Gill (1992). It works as follows. We have

$$S(t_1, t_2) = S(t_1, 0)S(0, t_2)\frac{S(t_1, t_2)S(0, 0)}{S(t_1, 0)S(0, t_2)}.$$

$S(t_1,t_2)S(0,0)/S(t_1,0)S(0,t_2)$ is called the cross-ratio over the four corners of the rectangle $(0,t_1]\times(0,t_2]$. The marginal survival functions $S_1(t_1)$ and $S_2(t_2)$ can be estimated by using their corresponding marginal samples with the well known Kaplan-Meier estimator. We have

$$S_1(t_1) = \prod_{(0,t_1]} (1-\Lambda_1(ds)),$$

where

$$\Lambda_1(ds) = \frac{F_1(ds)}{S_1(s)} = \frac{P(T_1 \in ds, C_1 \geq s)}{P(T_1 \geq s, C_1 \geq s)} = \frac{P(\widetilde{T_1} \in ds, D_1 = 1)}{P(\widetilde{T_1} \geq s)}. \tag{1.18}$$

Here $\Lambda_1(ds_1)$ is the well known univariate hazard representing the conditional probability to die the coming moment given that you are alive right now. Here $\Lambda_1$ is estimated by the well known Nelson-Aalen estimator. If we set

$$\begin{aligned} N_{1n}(ds) &= \sum_{i=1}^{n} I(\widetilde{T}_{i1} \in ds, D_{1i} = 1) \\ Y_{1n}(s) &= \sum_{i=1}^{n} I(\widetilde{T}_{i1} \geq s), \end{aligned}$$

then the Nelson-Aalen estimator is defined by

$$\Lambda_{1n}(ds) = \frac{N_{1n}(ds)}{Y_{1n}(s)}, \tag{1.19}$$

and the Kaplan-Meier estimator is now given by

$$S_{1n}(t_1) = \prod_{[0,t_1]} (1-\Lambda_{1n}(ds)) = \prod_{\widetilde{T}_{1i} \leq t_1, D_{1i}=1} (1-\Lambda_{1n}(\Delta\widetilde{T}_{1i})).$$

Similarly we can estimate $\Lambda_2(ds)$ with $\Lambda_{2n}(ds)$ in order to obtain the Kaplan-Meier estimator for $S_2(t_2)$ based on the marginal sample for $T_2$.

Let $A_{ij}$, $i = 1,\ldots,k$, $j = 1,\ldots,k$ form a lattice partition of $[0,t_1]\times[0,t_2]$:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $A_{2,1}$ | | | | | | | | | |
| $A_{1,1}$ | $A_{1,2}$ | $A_{1,3}$ | $A_{1,4}$ | $A_{1,5}$ | $A_{1,6}$ | $A_{1,7}$ | $A_{1,8}$ | $A_{1,9}$ | $A_{1,10}$ |

If we multiply two cross ratios over adjacent rectangles, then we obtain the odds ratio over the large rectangle given by the union of the two adjacent rectangles: i.e. if we denote the cross ratio of $S$ over $A_{i,j}$ with $\mathrm{Odd}_S(A_{i,j})$ (from Odds-ratio), then $\mathrm{Odd}_S(A_{i,j}\cup A_{i,j+1}) = \mathrm{Odd}_S(A_{i,j})\mathrm{Odd}_S(A_{i,j+1})$. Consequently with $(0,t] \equiv (0,t_1]\times(0,t_2]$

$$\mathrm{Odd}_S((0,t]) = \prod_{i=1}^{k}\prod_{j=1}^{k} \mathrm{Odd}_S(A_{i,j}).$$

This holds for each lattice partition which makes it plausible that:

$$\mathrm{Odd}_S((0,t]) = \prod_{(0,t]} \frac{S(s_1+ds_1, s_2+ds_2)S(s_1,s_2)}{S(s_1+ds_1,s_2)S(s_1,s_2+ds_2)}, \tag{1.20}$$

where $\prod_{(0,t]}$ is the so called product integral (Gill and Johansen, 1990) and stands for a limit of approximating products over lattice partitions of $(0,t]$ as the partitions become finer.

Denote the four corners of $(s_1, s_1+ds_1] \times (s_2, s_2+ds_2]$ by $c_i(s)$, $i = 1, \ldots, 4$. Because the rectangle $(s, s+ds]$ is infinitely small we know for each censored observation with $\widetilde{T}_i > s$ if $T_i > c_i(s)$ or not. Hence we can estimate

$$P(T_i \geq c_i(s) \mid \widetilde{T}_i > s) = \frac{S(c_i(s))}{S(s)}$$

by

$$\frac{L_n(c_i(s))}{\sum_j I(\widetilde{T}_j \geq s)} \equiv \frac{\sum_{j:\widetilde{T}_j \geq s} I\{T_j \text{ is known to be larger than } c_i(s)\}}{\sum_j I(\widetilde{T}_j \geq s)},\ i = 1,2,3,4.$$

We conclude that the odds ratio can be naturally estimated by replacing the four factors $S(c_i(s))$ corresponding with corner $c_i(s)$ by $L_n(c_i(s))$. For example, we replace $S(s_1, s_2+ds_2)$ by summing up the number of uncensored observations larger than $(s_1, s_2+ds_2)$, the number of vertical lines with startpoint larger than $(s_1, s_2)$, the number of horizontal lines with startpoint larger than $(s_1, s_2+ds_2)$ and the number of doubly censored observations with startpoint larger than $(s_1, s_2)$.

In order to define $L_n(c_i(s))$ explicitly we define the following counting processes:

$$\begin{aligned}
N_{10}^n(ds_1, s_2) &\equiv \sum_{i=1}^n I(\widetilde{T}_1 \in ds_1, \widetilde{T}_2 \geq s_2, D_1 = 1) \\
N_{01}^n(s_1, ds_2) &\equiv \sum_{i=1}^n I(\widetilde{T}_1 \geq s_1, \widetilde{T}_2 \in ds_2, D_2 = 1) \\
N_{11}^n(ds_1, ds_2) &\equiv \sum_{i=1}^n I(\widetilde{T}_1 \in ds_1, \widetilde{T}_2 \in ds_2, D_1 = 1, D_2 = 1) \\
Y_n(s_1, s_2) &\equiv \sum_{i=1}^n I(\widetilde{T}_1 \geq s_1, \widetilde{T}_2 \geq s_2).
\end{aligned}$$

$Y_n(s)$ is the number of observations in $[s, \infty)$, $N_{11}^n(ds)$ is the number of $\bullet$'s in $[s, s+ds]$, $N_{10}^n(ds_1, s_2)$ is the number $\bullet$'s and vertical lines in the strip $(s_1, s_1+ds_1] \times [s_2, \infty)$, $N_{01}(s_1, ds_2)$ is the number $\bullet$'s and horizontal lines in the strip $(s_1, \infty) \times (s_2 + ds_2]$.

So we can estimate the four factors in the odds ratio over $(s, s+ds]$ as follows:

$$\begin{aligned}
L_n(s_1, s_2) &= Y_n(s_1, s_2) \\
L_n(s_1+ds_1, s_2+ds_2) &= Y_n(s_1, s_2) - N_{01}(s_1, ds_2) - N_{01}(ds_1, s_2) + N_{11}(ds_1, ds_2) \\
L_n(s_1+ds_1, s_2) &= Y_n(s_1, s_2) - N_{10}(s_1, ds_2) \\
L_n(s_1, s_2+ds_2) &= Y_n(s_1, s_2) - N_{01}(ds_1, s_2).
\end{aligned}$$

Substituting $L_n$ in (1.20) provides us with an estimator $R_n$ of $R$. Notice that the product integral becomes now a finite product over the lattice partition spanned by the marginal samples $\widetilde{T}_{1i}$ and $\widetilde{T}_{2j}$, $i = 1, \ldots, n$, $j = 1, \ldots, n$. Hence $R_n$ is given by:

$$R_n(t) = \prod_{i=1, \widetilde{T}_{1i} \leq t_1}^{n} \prod_{j=1, \widetilde{T}_{2j} \leq t_2}^{n} \frac{L_n(\widetilde{T}_{1i}, \widetilde{T}_{2j}) L_n(\widetilde{T}_{1,i+1}, \widetilde{T}_{2,j+1})}{L_n(\widetilde{T}_{1,i+1}, \widetilde{T}_{2j}) L_n(\widetilde{T}_{1i}, \widetilde{T}_{2,j+1})}, \tag{1.21}$$

where in terms of the counts $Y_n, N^n$ $R_n(t)$ equals:

$$\prod_{\widetilde{T}_{1i} \leq t_1}^{n} \prod_{\widetilde{T}_{2j} \leq t_2}^{n} \frac{Y_n(\widetilde{T}_{1i}, \widetilde{T}_{2j}) \Big( Y_n(\widetilde{T}_{1i}, \widetilde{T}_{2j}) - N_{01}(\widetilde{T}_{1i}, \Delta\widetilde{T}_{2j}) - N_{01}(\Delta\widetilde{T}_{1i}, \widetilde{T}_{2j}) + N_{11}(\Delta\widetilde{T}_{1i}, \Delta\widetilde{T}_{2j}) \Big)}{\Big( Y_n(\widetilde{T}_{1i}, \widetilde{T}_{2j}) - N_{10}(\widetilde{T}_{1i}, \Delta\widetilde{T}_{2j}) \Big) \Big( Y_n(\widetilde{T}_{1i}, \widetilde{T}_{2j}) - N_{01}(\Delta\widetilde{T}_{1i}, \widetilde{T}_{2j}) \Big)}.$$

So we conclude that the Dabrowska estimator can be computed as follows;
1) Compute the Kaplan-Meier estimators $S_{1n}(t_1)$ and $S_{2n}(t_2)$ of $S_1(t_1)$ and $S_2(t_2)$, respectively.
2) Compute $L_n(s_1, s_2)$ at $(s_1, s_2) = (\widetilde{T}_{1i}, \widetilde{T}_{2j})$, $i, j \in \{1, \ldots, n\}$.
3) Now, compute $R_n(t)$ (see 1.21)) and set

$$S_n^D(t_1, t_2) = S_{1n}(t_1) S_{2n}(t_2) R_n(t_1, t_2).$$

Notice that $S_n^D$ is a functional of the empirical distributions $N_{10}^n$, $N_{01}^n$, $N_{11}^n$, $Y_n$, $N_{1n}$, $N_{2n}$, $Y_{1n}$, $Y_{2n}$. It can be is shown that if we replace these empirical distributions in this functional by their Glivenko- Cantelli limits, then we obtain $S(t)$ (Gill, 1992). Hence, the functional delta-method (see Gill, 1989), which comes down to verifying the required differentiability of this functional, is applicable and leads to uniform consistency of $S_n^D$ and weak convergence of $\sqrt{n}(S_n^D - S)$ to a Gaussian process. Moreover, it provides us with asymptotic validity of the bootstrap. For the derivation of these theoretical results by applying the functional delta-method we refer to Gill, van der Laan, Wellner (1995).

### 1.6.3 The Prenctice-Cai estimator.

This estimator had been introduced by Prentice and Cai (1992a,b). The following derivation of the Prentice-Cai representation is from Gill, van der Laan, Wellner (1995). We will use $t = (t_1, t_2)$. As in the derivation of Dabrowska's representation we have $S(t) = S_1(t_1) S_2(t_2) R(t)$, where $R \equiv S/S_1 S_2$. In the Dabrowska representation we noticed that $R(t_1, t_2)$ can be considered as a multiplicative measure by defining $R(t) = R((0, t]) \equiv S(t_1, t_2) S(0, 0) / S(t_1, 0) S(0, t_2)$ and noticing that then $R(A \cup B) = R(A) R(B)$ (i.e. then $R$ is a multiplicative measure). Hence we could compute $R((0, t])$ by computing a product integral of $R((s, s + ds])$ over infinitesimal rectangles $(s, s+ds]$ and $R((s, s+ds])$ appeared to be a quantity which can be naturally estimated because of its insensitivity to censoring.

Now, we will consider $R$ as an additive measure; $R((0, t]) = R(t_1, t_2) - R(t_1, 0) - R(0, t_2) + R(0, 0)$. Using that $R$ equals 1 if $t_1$ and or $t_2$ equals zero it follows that:

$$R(t_1, t_2) = 1 + \int_{(0,t]} R(ds), \tag{1.22}$$

where as reasoned in the introduction we expect to find that $R(ds) = R(s-) d\widetilde{L}(s)$ for certain measure $\widetilde{L}$ whose increments can be estimated just as the cross ratios in the Dabrowska representation.

In order to express $R(ds)$ we need the following differentiability rules for $U : \mathbb{R} \to \mathbb{R}$ and $V : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} d(UV) &= U_- dV + (dU)V \\ d\left(\frac{1}{U}\right) &= \frac{dU}{UU_-}. \end{aligned}$$

If we apply these one dimensional rules to the sections $u \to F(u,v)$ and $v \to F(u,v)$ of a bivariate function $F$, then we denote these with $d_1$ and $d_2$, respectively. We apply these two one dimensional rules to each of the two variables of $R$ in turn in order to compute $dR = d_{12}R = d_1(d_2(R))$.

When applying the product rule to $S/S_i$ we give the left continuous version to $S$ instead of giving it to one of the $S_i$, $i = 1, 2$, and we denote $F_{(-1)}(s_1, s_2) \equiv F(s_1-, s_2)$, $F_{(-2)}(s_1, s_2) \equiv F(s_1, s_2-)$). We find

$$\begin{aligned} dR &= d_{12}R \\ &= \frac{d_{12}S}{S_1S_2} - \frac{d_2S_2d_1S_{(-2)}}{S_2S_{2-}S_1} - \frac{d_1S_1d_2S_{(-1)}}{S_1S_{1-}S_2} + \frac{d_1S_1d_2S_2S_-}{S_1S_{1-}S_2S_{2-}} \\ &= R_- \frac{S_{1-}S_{2-}}{S_1S_2}\left(\frac{d_{12}S}{S_-} - \frac{d_2S_2}{S_{2-}}\frac{d_1S_{(-2)}}{S_-} - \frac{d_1S_1}{S_{1-}}\frac{d_2S_{(-1)}}{S_-} + \frac{d_1S_1}{S_{1-}}\frac{d_2S_2}{S_{2-}}\right) \\ &\equiv R_- \frac{S_{1-}S_{2-}}{S_1S_2}\left(\Lambda_{11}(ds) - \Lambda_2(ds_2)\Lambda_{10}(ds_1, s_2) - \Lambda_1(ds_1)\Lambda_{01}(s_1, ds_2) + \Lambda_1(ds_1)\Lambda_2(ds_2)\right) \\ &= R_-\left(\frac{\Lambda_{11}(ds) - \Lambda_2(ds_2)\Lambda_{10}(ds_1, s_2) - \Lambda_1(ds_1)\Lambda_{01}(s_1, ds_2) + \Lambda_1(ds_1)\Lambda_2(ds_2)}{(1-\Lambda_1(\Delta s_1))(1-\Lambda_2(\Delta s_2))}\right) \\ &\equiv R_- d\widetilde{L}. \end{aligned} \tag{1.23}$$

The bivariate hazards $\Lambda_{10}$, $\Lambda_{01}$, $\Lambda_{11}$ represent bivariate analoques of the univariate hazards. At the fifth equality notice that $S_{1-}S_{2-}/S_1S_2 = 1/(1-\Lambda_1(\Delta s_1))(1-\Lambda_2(\Delta s_2))$.

Substituing $R = R_- d\widetilde{L}$ in (1.22) provides us with:

$$R(t) = 1 + \int_{(0,t]} R(s-)\widetilde{L}(ds). \tag{1.24}$$

The hazards $\Lambda_1$ and $\Lambda_2$ in the representation of $\widetilde{L}$ are estimated by the Nelson-Aalen estimators $\Lambda_{1n}$ and $\Lambda_{2n}$ as given by (1.19). The bivariate analoques $\Lambda_{10}, \Lambda_{01}, \Lambda_{11}$ are estimated by the straightforward generalizations of the univariate Nelson-Aalen estimators:

$$\begin{aligned} \Lambda^n_{10}(ds_1, s_2) &= \frac{N^n_{10}(ds_1, s_2)}{Y_n(s_1, s_2)} \\ \Lambda^n_{01}(s_1, ds_2) &= \frac{N^n_{01}(s_1, ds_2)}{Y_n(s_1, s_2)} \\ \Lambda^n_{11}(ds_1, ds_2) &= \frac{N^n_{11}(ds_1, ds_2)}{Y_n(s_1, s_2)}. \end{aligned} \tag{1.25}$$

Substitution of these hazard estimates in the representation (1.23) of $\widetilde{L}$ provides us with an estimator $\widetilde{L}_n$ of $\widetilde{L}$. (1.24) suggests now to estimate $R$ with $R_n$ where $R_n$ is the solution of:

$$R_n(t_1, t_2) = 1 + \int_{(0,t]} R_n(s-)\widetilde{L}_n(ds). \tag{1.26}$$

$R_n$ can be computed iteratively by starting the recursion in $(0,0)$ and working up rowwise from left to right (so according to the so called video ordering).

It works as follows: Let $\widetilde{T}_{1(i)}$, $i = 1, \ldots, n$, be the ordered $\widetilde{T}_{1i}$ and similarly let $\widetilde{T}_{2(j)}$, $j = 1, \ldots, n$, be the ordered $\widetilde{T}_{2j}$. Let $z_{i,j} \equiv (x_i, y_j) \equiv (\widetilde{T}_{1(i)}, \widetilde{T}_{2(j)})$ and $x_0 = y_0 = 0$. We can compute $\widetilde{L}_n(\Delta z_{i,j})$ for all grid points $z_{i,j} = (x_i, y_j)$ as explained above. The relation (1.26) is given by:

$$R_n(z_{i,j}) = 1 + \sum_{i_1 \leq i, j_1 \leq j} R_n(z_{i_1-1,j_1-1}) \widetilde{L}_n(\Delta z_{i_1,j_1}). \tag{1.27}$$

Because $R_n(t_1, 0) = R_n(0, t_2) = 1$ we know that $R_n(z_{i,0}) = R_n(z_{0,j}) = 1$. (1.27) tells us that for computing $R_n(x_i, y_1)$, $i = 1, \ldots, n$ we only need to know $R_n(x_1, 0)$, $\ldots$, $R_n(x_n, 0)$ and we know now that $R_n$ equals 1 here. So

$$R_n(z_{1,1}) = 1 + 1L_n(\Delta z_{1,1})$$

and

$$R_n(z_{1,2}) = 1 + 1L_n(\Delta z_{1,1}) + 1L_n(\Delta z_{1,2})$$

and so on

$$R_n(z_{1,j}) = 1 + \sum_{j_1 \leq j} L_n(\Delta z_{1,j_1}),\ j = 1, \ldots, n.$$

Again, (1.27) tells us that for computing $R_n(x_i, y_2)$, $i = 1, \ldots, n$, we only need to know $R_n(0, y_1), \ldots, R_n(x_n, y_1)$: e.g.

$$R_n(z_{2,2}) = 1 + 1L_n(\Delta z_{1,1}) + 1L_n(\Delta z_{1,2}) + 1L_n(\Delta z_{2,1}) + R_n(z_{1,1})L_n(\Delta z_{2,2}).$$

Hence we can compute all values $R_n(z_{2,j})$, $j = 1, \ldots, n$. With these values we can compute all values $R_n(z_{3,j})$, $j = 1, \ldots, n$ and in this way we can proceed till the $n$-th final row which determines $R_n$ completely.

The Prentice-Cai estimator is given by:

$$S_n^{PC}(t) = S_{1n}(t_1) S_{2n}(t_2) R_n(t).$$

Also here the functional delta-method is applicable and leads to uniform consistency, weak convergence and asymptotic validity of the bootstrap for $S_n^{PC}$ (Gill, van der Laan, Wellner, 1995).

### 1.6.4 Estimators based on the EM-algorithm

Assume that we would have observed all $T_i$. Then the NPMLE $S_n(t_1, t_2)$ of $S(t_1, t_2)$ equals the fraction of the $T_i$ which are larger than $(t_1, t_2)$. In other words we would give each observation $T_i$ weight $1/n$ and sum up the weights of the $T_i$'s with $T_i > t$.

In our case we can still give all uncensored observations weight $1/n$. The censored observations tell us only that $T_i \in B(Y_i)$; so we want to give the mass $1/n$ to $B(Y_i)$ in an appropriate way. Assume that by using the observations in $B(Y_i)$ we are able to obtain a good estimator $P_{F_n^0}(T \in \cdot \mid T \in B(Y_i))$ of the conditional distribution $P(T \in \cdot \mid T \in B(Y_i))$ of $T_i$ given that $T_i \in B(Y_i)$. Then a natural thing to do is to redistribute the mass $1/n$ corresponding with the censored observation $Y_i$ over $B(Y_i)$ as follows (assume for convenience that the estimate is

discrete): a point $s > t$ gets the following fraction of the mass $1/n$: $P_{F_n^0}(T = s \mid T \in B(Y_i))$. If we do the redistribution for all censored observations, then we obtain a new estimator $F_n^1$, which might be an improvement.

This suggests the following algorithm:
1) Let $\{s_1, \ldots, s_k\}$ be a set of points in the plane which contains all uncensored $T_i$ and it is such that each $B(Y_i)$ (lines and quadrants) contains at least one of these $s_i$'s.
2) Give each $s_i$ a weight $f_n^0(s_i) > 0$ s.t. $\sum_{i=1}^k f_n^0(s_i) = 1$. Set the count $M = 0$.
3) Compute a new estimator $f_n^{M+1}$, as follows:

$$f_n^{M+1}(s_i) = \sum_{j=1}^n P_{f_n^M}(T = s_i \mid T \in B(Y_j)_1)\frac{1}{n},\ i = 1, \ldots, k. \tag{1.28}$$

In other words, a point $s_i$ gets from each observation $Y_j$ mass $1/nP_{f_n^M}(T = s_i \mid T \in B(Y_j)_1)$, which is zero if $s_i \notin B(Y_j)_1$ and it is 1 if $s_i$ equals the observed $T_j$.
4) Replace $M$ by $M + 1$ and go to step 3.

This is the EM-algorithm (Turnbull, 1976, Dempster, Laird, Rubin, 1977). $f_n^M$ can be shown to converge to a solution $f_n$ of (1.28) with $f^M = f^{M+1} = f_n$. The equation (1.28) in $f_n$ is the well known self-consistency equation of Efron (1967) which is solved by NPMLE.

If an uncensored observation gets mass from a censored observation at step $M$ of the EM-algorithm, then this influences the conditional probabilities $P_{f_n^{M+1}}(T = s \mid T \in B(Y_j)_1)$ of each region $B(Y_j)_1$ which contains this uncensored observation. If the underlying $F$ is continuous, then the half-lines $B(Y_j)_1$ corresponding with singly-censored observations will with probability one not contain any uncensored observations. So then the conditional probabilities over lines do not change at a step of the EM-algorithm by mass given to the uncensored observations: the singly-censored observations do not listen to information given by uncensored observations, but they might change by mass given by doubly censored and other singly censored observations. Since uncensored observations around a half- line give a lot of information about the distribution over the half-line, there is no reason to expect any good performance of $S_n$. Indeed $S_n$ is not consistent for continuous data (Tsay, Leurgans and Crowley, 1986).

Because the idea behind the EM-algorithm is very natural, in each application, where it fails, it seems to be worthwhile to think about a simple repair of the algorithm.

**Pruitt's Modification of the EM-algorithm.**

The self-consistency equation (1.28) in $f_n$ can by simple integration be written in terms of $S_n(t)$:

$$S_n(t) = \frac{1}{n}\sum_{i=1}^n P_{F_n}(T > t \mid T \in B(Y_i)).$$

The idea of Pruitt's estimator (Pruitt, 1991) is that in the EM-algorithm *we* should tell the singly-censored observations how to redistribute their mass $1/n$ over the half-lines instead of letting the EM-algorithm do the work.

This means that for each observation $(\widetilde{T}, D = (0,1))$ we need to estimate the conditional probabilities $P_F(T > t \mid T_1 \geq \widetilde{T}_1, T_2 = \widetilde{T}_2)$ and for each observation $(\widetilde{T}, D = (1,0))$ we need to estimate $P_F(T > t \mid T_1 = \widetilde{T}_1, T_2 \geq \widetilde{T}_2)$ over lines which can then be plugged in the EM-algorithm; so at each step in the EM-algorithm, where we have to redistribute mass over a line, we use these fixed estimates.

If we define

$$S_1(y_1, y_2) \equiv P(T_1 > y_1, T_2 = y_2),$$

then

$$P_F(T > t \mid T_1 \geq y_1, T_2 = y_2) = I(y_2 \geq t_2)\frac{S_1(t_1 \vee y_1, y_2)}{S_1(y_1, y_2)}. \tag{1.29}$$

$S_1$ is just a survival function in $T_1$ and hence we can use an equivalent of the Kaplan-Meier estimator by using all uncensored observations and singly-censored (horizontal) observations in a strip around the half-line $(y_1, \infty) \times \{y_2\}$ which can be weighted by their distance from this half-line. In order to define this Kaplan-Meier estimator we define for a univariate kernel $K$ and bandwidth $h$ :

$$\begin{aligned} N_{1n}(dy_1, y_2) &\equiv \sum_{i=1}^{n} I(D_{1i} = 1, D_{2i} = 1, \widetilde{T}_{1i} \in dy_1) K\left(\frac{y_2 - \widetilde{T}_{2i}}{h}\right) \\ Y_{1n}(y_1, y_2) &\equiv \sum_{i=1}^{n} I(\widetilde{T}_{1i} \geq y_1, D_{2i} = 1) K\left(\frac{y_2 - \widetilde{T}_{2i}}{h}\right). \end{aligned}$$

Now, the equivalent of Kaplan-Meier is given by:

$$S_{1n}(y_1, y_2) = \prod_{\widetilde{T}_{1i} \leq y_1, D_{2i}=1} \left(1 - \frac{N_{1n}(\Delta\widetilde{T}_{1i}, y_2)}{Y_{1n}(\widetilde{T}_{1i}, y_2)}\right).$$

If we take $K(x) \equiv I(x \in [-0.5, 0.5])$, then we are really just computing the Kaplan-Meier estimator based on the uncensored and horizontal singly-censored within a distance $0.5h$ from $(y_1, \infty) \times \{y_2\}$. If we take higher order kernels, then we are weighting the observations by their distance from the line $(y_1, \infty) \times \{y_2\}$.

Substitution of $S_{1n}$ in (1.29) provides us with an estimator of $P_F(T > t \mid T_1 \geq y_1, T_2 = y_2)$:

$$\widehat{P_F}(T > t \mid T_1 \geq y_1, T_2 = y_2) = I(y_2 \geq t_2)\frac{S_{1n}(t_1 \vee y_1, y_2)}{S_{1n}(y_1, y_2)}. \tag{1.30}$$

Replace now in the EM-algorithm for each singly-censored observation $Y_i$ $P_{f_n^N}(T = s \mid T \in B(Y_i))$ by the estimate derived from (1.30) and iterate the EM-algorithm with a certain initial estimator. Then the redistribution over lines is determined by this fixed (in $M$) estimate.

Also here the algorithm converges to a solution $f_n^P$ of an equation which is in terms of $S_n$ given by:

$$\begin{aligned} S_n^P(t) &= \frac{1}{n}\sum_{i=1}^{n} I(D_i = (1,1), \widetilde{T}_i > t) + \frac{1}{n}\sum_{i=1}^{n} I(D_i = (0,1), \widetilde{T}_{2i} \geq t_2)\frac{S_{1n}(t_1 \vee \widetilde{T}_{1i}, \widetilde{T}_{2i})}{S_{1n}(\widetilde{T}_{1i}, \widetilde{T}_{2i})} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} I(D_i = (1,0), \widetilde{T}_{1i} \geq t_1)\frac{S_{2n}(\widetilde{T}_{1i}, t_2 \vee \widetilde{T}_{2i})}{S_{2n}(\widetilde{T}_{1i}, \widetilde{T}_{2i})} + \frac{1}{n}\sum_{i=1}^{n} I(D_i = (0,0))\frac{S_n^P(t \vee \widetilde{T}_i)}{S_n^P(\widetilde{T}_i)}. \end{aligned}$$

Pruitt's estimator $S_n^P$ is the solution of this equation. The EM-algorithm just iterates this equation; at the right hand side plug in an initial estimator $S_n^0$ (only the doubly censored term is unknown) and compute the right hand side which gives us a $S_n^1$ and so on.

In van der Laan (1994) uniform consistency, weak convergence and a bootstrap result have been proved for a slight modification (we need smooth kernels with orthogonality properties

and we use bivariate smoothing) of this estimator, all under smoothness assumptions on the underlying distribution $F$ and $G$ of $T$ and $C$, respectively. The proof is tricky, but essentially based on the implicit function theorem and results about bivariate density estimators.

**Sequence of reductions NPMLE.**

The following estimator is introduced in van der Laan (1992) and theoretically further developed in van der Laan (1995b, and chapter 4 of 1995d). This estimator corresponds with a NPMLE (so it is self-consistent and computable with the EM-algorithm) based on reduced data, where reduced is meant in the sense that the uncensored component of the singly censored observations $Y_i$ are interval censored by using a lattice partition $A_{k,l} = (u_k, u_{k+1}] \times (v_l, v_{l+1}]$, in the sense that (we do as if) it is only known to lie in $(u_k, u_{k+1}]$ (if $T_1$ is uncensored) or in $(v_l, v_{l+1}]$ (if $T_2$ is uncensored). Its name "sequence of reductions MLE" will be abbreviated with SOR-MLE. The interval censored singly-censored observations tell us that $T_i$ has fallen in a strip around the original singly-censored observation, which will contain other uncensored observations, and hence we expect a better result from the EM-algorithm.

However, since the interval censoring means a grouping of singly-censored observations the density of the interval censored singly-censored observations equals now the integral over this interval of the density of the original singly-censored observations Hence the corresponding likelihood does not factorize anymore in a part which only depends on $F$ and a part which only depends on $G$, the distribution of the censoring vector $C$, so that one also has to compute the NPMLE of $G$; now the observation for $C$ gives information for $T$. In more formal terminology, introduced by Heitjan and Rubin (1991), we say that $T$ is not coarsened at random, anymore; we can not consider our observation as $T \in B(Y)$, but we need also to incorporate knowledge on $C$. On the other hand if the censoring variables $C_i$ had a distribution which is purely discrete on the partition points $(u_k, v_l)$, then the likelihood would factorize again so that we can ignore the distribution of $C$ again, which makes the estimator computationally more tractable. The following simulation method simulates such $C_i$ and computes the corresponding interval censored singly censored observations. In experiments where the censoring time is caused by the end of the experiment one will observe all the $C_i$'s and hence in this case the simulation method can be skipped.
1) *Simulation method*. Each observation $Y_i$ tells us that $C_i \in B(Y_i)_2$ (line, point, or quadrant). Given $Y_i$ we estimate $P(C \in dc \mid C \in B(Y_i)_2)$ and we draw a $C_i'$ from this estimate of the conditional distribution over $B(Y_i)_2$. This provides us with $C_i'$, $i = 1, 2, \ldots, n$.
2) Choose a number $h$ and make a lattice partition $A_{k,l}$ of squares with width $h$ and denote the left lower corners of $A_{k,l}$ with $(u_k, v_l)$.
3) Discretize the $C_i'$, $i = 1, \ldots, n$ by pulling them back to the left lower corner $(u_k, v_l)$ of $A_{k,l}$, where $A_{k,l}$ is that rectangle of the lattice partition which contains $C_i'$.
Denote these discretized $C_i'$ with $C_i''$.
4) Compute for each observation $Y_i$ its discretized version

$$Y_i' = (\widetilde{T}_i', D_i') = (\min(T_i, C_i''), I(T_i \leq C_i'')).$$

So a quadrant (corresponding with a doubly censored observation) starts at a $(u_k, v_l)$, a horizontal line starts with a $x$-coordinate equal to a $u_k$, a vertical line starts with a $y$-coordinate equal to a $v_l$, but the uncensored observations are still continuously distributed.

In comparison with the original data set all observed components $C_{1i}$ (corresponding with a singly or doubly censored observation) are drawn backwards to the closest $u_k$ and similarly are all censored observed components $C_{2i}$ drawn backwards to the closest $v_l$. Moreover, some of the uncensored $T_{1i}$ or $T_{2i}$ might have become censored.
5) For each singly censored (reduced) observation with $D_i' = (1,0)$ we let $E_{k,l}(\widetilde{T}_i', D_i')$ be that vertical strip $(u_k, u_{k+1}] \times (v_l, \infty)$ for which $\widetilde{T}_i' \in A_{k,l}$. Similarly, we associate with each singly censored (reduced) observation with $D_i' = (0,1)$ a horizontal strip $E_{k,l}(\widetilde{T}_i', D_i')$.
6) Compute the solution of the following self-consistency equation:

$$\begin{aligned} S_n^E(t) &= \frac{1}{n}\sum_{D_i'=(1,1)}^{n} I(\widetilde{T}_i' > t) + \frac{1}{n}\sum_{D_i'=(1,0)}^{n} P_{F_n^E}(T > t \mid T \in E_{k,l}(\widetilde{T}_i', D_i')) \\ &+ \frac{1}{n}\sum_{D_i'=(0,1)}^{n} P_{F_n^E}(T > t \mid T \in E_{k,l}(\widetilde{T}_i', D_i')) + \frac{1}{n}\sum_{D_i'=(0,0)}^{n} \frac{S_n^E(t \vee \widetilde{T}_i')}{S_n^E(\widetilde{T}_i')}. \end{aligned}$$

The solution is found by iterating this equation with a purely discrete $S_n^0$ which puts mass on all uncensored observations and on at least one point in each strip and quadrant on which we condition. This is equivalent with iterating the EM-algorithm with $S_n^0$ and as data $Y_i'$, but where the lines $B(Y_i)$ are replaced by their corresponding strips $E_{k,l}$.

In van der Laan (1995b) it is proved that if we had observed $C_i$ and hence could set $C_i' = C_i$, then this "reduced data" NPMLE is uniformly consistent, asymptotically Gaussian and that if we let the width $h$ of the partition (used for discretizing $C_i$) converge slowly enough to zero when the number of observations converge to infinity, then this estimator is also asymptotically efficient. The proof is strongly based on an identity for NPMLE in convex linear models as introduced in van der Laan (1995a) which says that the NPMLE $S_n(t)$ minus $S(t)$ equals the empirical mean of the so called efficient influence curve at $S_n$ minus its true mean.

## 1.7 Consistency.

Let $\Psi(\theta) = V(P_\theta) \in (D, \|\cdot\|)$ be an identifiable banachspace valued parameter (so e.g. $D = \mathbb{R}^k$ with the euclidiean norm or $D$ can be the space of continuous functions, or right-continuous functions, endowed with the supremum norm, or $D = L^2$ etc) and assume that we estimate $\Psi(\theta)$ with $V(P_n)$ or $V(\tilde{P}_n)$. For convenience, $P_n$ stands for the smoothed or unsmoothed empirical probability distribution.

As examples showed, $V(P_n)$ can depend on $P_n$ only through its moments or through a finite number of values $P_n f_i$ for a collection of $f_i$'s, or through its complete cumulative distribution function or even through its complete density. Depending on this dependence on $P_n$ there exist usually a more natural representation of $V(P_n)$, say $\Phi(Q_n)$, where $Q_n$ is the relevant part of $P_n$, and $\Phi : (D_1, \|\cdot\|_1) \to (D, \|\cdot\|)$, where $D_1$ is a linear space containing $Q_n$.

For example, if $V(P_n)$ depends on $P_n$ only through its mean $\bar{X}_n$, (or more general, on a finite number of its moments, as is the case for any method of moments estimator in parametric models) then it is more natural to represent $V(P_n)$ as a function of the real valued random variable $\bar{X}_n$ instead of as a function of the infinite dimensional $P_n$.

**Example 1.7.1** For example consider the gamma density. The gamma distribution depends

on two unknown parameters $\alpha$ and $\lambda$:

$$f(x:\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}\exp(-\lambda x).$$

The first two moments of the gamma distribution are

$$\mu_1 = \frac{\alpha}{\lambda} \text{ and } \mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

It follows that

$$\alpha = \frac{\mu_1^2}{\mu_2-\mu_1^2} \text{ and } \lambda = \frac{\mu_1}{\mu_2-\mu_1^2},$$

which shows that $\alpha$ and $\lambda$ are identifiable parameters since we have shown that $(\alpha,\lambda) = V(f(\cdot : \alpha,\lambda))$ for a $V$. Suppose now that we have $n$ i.i.d. observations from $f(\cdot:\alpha,\lambda)$. The method of moments estimator $(\alpha_n,\lambda_n)$ is obtained by estimating $\mu_j$ with $\mu_j^n = \int x^j dP_n(x) = \frac{1}{n}\sum_{i=1}^n X_i^j$, $j=1,2$. It is clear that it is more natural to consider $(\alpha_n,\lambda_n)$ as a function of $Q_n = (\mu_1^n,\mu_2^n)$ instead of as a function of the empirical distribution $P_n$.

Proving *consistency* of $\Phi(Q_n)$, i.e. $\|\Phi(Q_n)-\Phi(Q)\| \to 0$ in probability or almost surely, involves now two steps, one probabilistic step and one analytical step. Firstly, one shows that $\|Q_n - Q\|_1 \to 0$ in probability or almost surely. Secondly, one shows that $\Phi : (D_1, \|\cdot\|_1) \to (D, \|\cdot\|)$ is continuous in the sense that if $\|Q_n - Q\|_1 \to 0$, then

$$\|\Phi(Q_n)-\Phi(Q)\| \to 0.$$

One only has to verify this continuity of $\Phi$ for sequences $Q_n, Q$ which can occur as realizations of the random $Q_n, Q$. For establishing the probabilistic result one can often refer to empirical process theory or density estimation theory, for example if $Q_n = P_n$ and

$$\|Q_n - Q\|_1 = \|Q_n - Q\|_{\mathcal{F}} \equiv \sup_{f\in\mathcal{F}} \mid Q_n f - Qf \mid$$

for a Glivenko-Cantelli class $\mathcal{F}$, then the result is shown.

**Exercise 1.** Prove consistency of $\int f_n^2(x)dx$ as an estimator of $\int f^2(x)dx$ if $f_n$ is a supnorm-consistent estimator of $f(x)$. Can you also prove consistency if $f_n$ is only known to be consistent in $L^2$; i.e. if $\int (f_n - f)^2(x)dx \to 0$.
**Exercise 2.** Prove supnorm-consistency of $F_n$ in the current status model, assuming that $p_n(c,\delta)$ is a uniformly consistent estimator of $p(c,\delta)$ and assuming that $g(c) = p(c,1)+p(c,0) > 0$ for all $c$.

Suppose now that you know that $Q_n$ converges with a rate $r(n)$ to $Q$ in probability: i.e. $\|Q_n - Q\|_1 = O_P(r(n))$ for some rate $r(n)$. Instead of only showing that the estimator $\Phi(Q_n)$ is consistent it is also of interest to know if the estimator is consistent at the same rate as $Q_n$ was. This means that $\|Q_n - Q\|_1 / r(n) = O_P(1)$; i.e. for all $\epsilon > 0$, there exists a $M < \infty$ so that $P(\|Q_n - Q\|_1/r(n) > M) < \epsilon$. Suppose now that $\Phi : (D_1, \|\cdot\|_1) \to (D, \|\cdot\|)$ satisfies: if $\|Q_n - Q\|_1 = O(r(n))$, then

$$\|\Phi(Q_n) - \Phi(Q)\|_1 = O(r_1(n)).$$

Then $\|\Phi(Q_n) - \Phi(Q)\|_1 = O_P(r_1(n))$.

In proving the probabilistic part of the consistency proof one will often refer to results in empirical process theory; e.g. $\|(F_n - F)\|_\infty \to 0$ a.s, where $F_n$ is the empirical cumulative distribution function of $F(t) = P(T \leq t)$, $t \in \mathbb{R}^d$. In proving the analytical part of the consistency proof it is useful to have the following lemmas as tools available. Firstly,

**Lemma 1.7.1** *(Integration by parts.) Let $F$ and $G$ be functions of bounded variation. Then*

$$\int_a^b F(x)dG(x) = GF\mid_a^b - \int_a^b GdF(x).$$

Secondly, we will often bound integrals as follows:

**Lemma 1.7.2** *Let $F$ be any function and $G$ be of bounded variation. Then*

$$\int_a^b F(x)dG(x) \leq \|F\|_{\infty,[a,b]}\|G\|_{v,[a,b]},$$

*where $\|G\|_{v,[a,b]} = \int_a^b \mid dG(x) \mid$.*

Finally, the following so called telescoping trick is very useful.

**Lemma 1.7.3** *(Telescoping.) Let $a_i, b_i$, $i = 1, \ldots, k$, be real valued numbers. We have*

$$\prod_{i=1}^k a_i - \prod_{i=1}^k b_i = \sum_{l=1}^k \prod_{i=1}^l a_i(a_l - b_l) \prod_{i=l+1}^k .$$

Compare this with the differentiation rule for a product of functions: $d/dx(fg) = (d/dxf)g + f(d/dxg)$. In other words, this trick is nothing else then application of the product rule to $\prod_{i=1}^k f_i(x)$ for functions $f_i(x)$. With these tricks in mind we will be able to prove consistency of the Kaplan-Meier estimator.

**Example 1.7.2 Consistency of the integrated hazard in the univariate censoring model.** Let $F_n(x)$ be the empirical distribution based on an i.i.d. sample of $n$ (positive valued) observations $X_i \sim F$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x).$$

By empirical process theory we know that $\|F_n - \|_\infty \to 0$ a.s.. Let $S_n(x) = 1 - F_n(x-)$ be the corresponding empirical survival function. The *cumulative hazard* or *integrated hazard* evaluated at a point $t$ is defined by

$$\Lambda_n(t) \equiv \int_0^t \frac{dF_n(x)}{S_n(x)}.$$

We are concerned with showing that $\Lambda_n(t)$ is a uniformly consistent estimator of $\Lambda(t) = \int_0^t dF(x)/S(x)$. We are only able to show this uniformly over $[0, \tau]$ for any point $\tau$ for which $S(\tau) < 1$. Notice that $\Lambda_n(t) = g(F_n, S_n)$ and $\Lambda(t) = g(F, S)$. We have already shown the probabilisitic part of the consistency proof: $\|F_n - F\|_\infty \to 0$ and $\|S_n - S\|_\infty \to 0$ a.s. By applying $a_nb_n - ab = (a_n - a)b + a_n(b_n - b)$ to $a_n = dF_n(x)$ and $b_n = 1/S_n(x)$ we obtain

$$\int \frac{dF_n}{S_n} - \int \frac{dF}{S} = \int \frac{d(F_n - F)}{S} + \int \frac{(S - S_n)dF_n}{S_nS}.$$

The second term can be bounded by $\|S_n - S\|_{\infty,[0,\tau]} \|F_n\|_v$ and hence converges to zero. Now, apply integration by parts to the first term:

$$\int_0^t \frac{d(F_n - F)}{S_n} = (F_n - F)1/S_n \,|_0^t - \int (F_n - F)(x) \frac{1}{S_n^2(x)} dF_n(x).$$

Now, we can also bound this term in the supnorm of $\|F_n - F\|\infty$ and the variation norm of $F_n$, using that $S_n(\tau) > \delta > 0$ for $n$ large enough and some $\delta > 0$. This proves the analytical part of the consistency proof: $\|\Lambda_n - \Lambda\|_{\infty,[0,\tau]} \to 0$. This proves consistency of $\Lambda_n$ uniformly in $[0, \tau]$.

Recall now the representation for the integrated hazard in the univariate censoring model. Let $P_1(t) = P(\widetilde{T} \leq t, \Delta = 1)$ and $\bar{P}(t) = P(\widetilde{T} \geq t)$ and let $P_{1n}(t)$ and $\bar{P}_n(t)$ be their empirical distributions. Then $\Lambda_n = g(P_{1n}, \bar{P}_n)$ for the same $g$ as above. By empirical process theory we know that $\|P_{1n} - P_1\|_\infty \to 0$ and $\|\bar{P}_n - \bar{P}\|_\infty \to 0$ a.s. Hence the same proof as above shows that $\|\Lambda_n - \Lambda\|_{\infty,\tau} \to 0$ a.s. for any $\tau$ for which $\bar{P}(\tau) = S(\tau)(1 - G(\tau)) > 0$.

We are now ready to prove consistency of the Kaplan-Meier estimator.

**Example 1.7.3 Consistency of the Kaplan-Meier estimator.** We have

$$S(t) = \Phi(\Lambda)(t) \equiv \prod_0^t (1 - d\Lambda(s)).$$

Let $\Lambda_n$ be the integrated hazard as defined in the preceding example. We have that the Kaplan-Meier estimator $S_n(t) = \Phi(\Lambda_n)$. In the preceding example we showed that $\|\Lambda_n - \Lambda\|_{\infty,[0,\tau]} \to 0$ a.s. Hence for proving consistency of $S_n$ it remains to prove the required continuity of $\Phi$: for any sequence $\Lambda_n$ with $\|\Lambda_n - \Lambda\|_{\infty,[0,\tau]} \to 0$ we have $\|\Phi(\Lambda_n) - \Phi(\Lambda)\|_{\infty,[0,\tau]} \to 0$.

A continuous version of the telescoping identity provides us with the so called Duhamel-equation for the product integral:

$$\prod_{(0,t]} (1 - d\Lambda_n(s)) - \prod_{(0,t]} (1 - d\Lambda(s)) = \int_0^t \prod_{(0,s)} (1 - d\Lambda(u)) \prod_{(s,t]} (1 - d\Lambda_n(u)) d(\Lambda_n - \Lambda)(s).$$

The first product integral equals $S(s)$, the second product integral equals the fraction $S_n(t)/S_n(s)$. Hence we can represent the integral as $S_n(t) \int S(s)/S_n(s) d(\Lambda_n - \Lambda)(s)$. By applying integration by parts we can bound this by the supnorm of $\Lambda_n - \Lambda$ and the variation norm of $S/S_n$ which is uniformly bounded, using that $S_n(\tau) > \delta > 0$ for some $\delta > 0$ and $S_n(t) < 1$. This shows that $\Phi(\Lambda_n) - \Phi(\Lambda)$ converges to zero uniformly in $t \in [0, \tau]$. We conclude that the Kaplan-Meier estimator is uniformly consistent on $[0, \tau]$.

## 1.8 Asymptotic linearity and influence curves.

We will now be concerned with proving convergence in distribution of $\sqrt{n}(\Phi(Q_n) - \Phi(Q)) \in \mathbb{R}$ to a normal distribution. If this holds we will say that $\Phi(Q_n)$ is asymptotically normally distributed. Our statements have generalizations to the case where $\Phi(Q_n)$ is a $D$-valued random variable, $D$ being a general linear space. For general $D$-spaces we can not talk about convergence to a normal distribution, but instead one needs to talk about convergence in distribution as random elements of the big space $D$ to a Gaussian process (see our comments about this in the empirical probability distribution section). Here we do not want to state results in this generality and hence we will stick to the real valued case.

*Asymptotic normality* of an estimator is a direct consequence of asymptotic linearity.

**Definition 1.8.1** *Let $X_1, \ldots, X_n$ be n i.i.d. observations from $P \in \mathcal{M}$. An estimator $\theta_n$ is an asymptotically linear estimator of a parameter $\theta$ if*

$$\theta_n - \theta = \frac{1}{n} \sum_{i=1}^{n} IC_P(X_i) + o_P(1/\sqrt{n}),$$

*where the so called* **influence curve** *$IC_P(X)$ has mean zero and has finite variance; i.e. $IC_P \in L_0^2(P)$ and moreover this should hold for all $P \in \mathcal{M}$.*

Notice that the CLT tells us that if $\Phi(Q_n)$ is asymptotically linear, then $\sqrt{n}(\Phi(Q_n) - \Phi(Q))$ converges in distribution to a normal distribution with mean zero and variance $\mathrm{VAR}(IC_P(X))$. Hence explicit knowledge of the influence curve is of great importance for *constructing asymptotic confidence bands* for the unknown parameter $\Phi(Q)$; notice that an estimate of the variance of $IC_P(X)$ provides us with an estimate of the limit distribution of $\sqrt{n}(\Phi(Q_n) - \Phi(Q))$ and hence with an approximate confidence interval for $\Phi(Q)$. The definition of asymptotic linearity says in words that an estimator $\theta_n$ of $\theta$ is asymptotically linear if it can be approximated by an average (sum) of i.i.d. random variables $\theta + IC_P(X_i)$, $i = 1, \ldots, n$.

### 1.8.1 Using transformations to improve the normal approximation.

For a number of parametric models it is well known that some transformation $T$ one will have that $\sqrt{n}(T(\Phi(Q_n)) - T(\Phi(Q)))$ approximates the normal distribution much quicker than $\sqrt{n}(\Phi(Q_n)) - \Phi(Q)))$; for example, the log of multinomial probabilities converges quicker to the normal approximation than the actual probabilities itself. For example, if $\Phi(Q_n)$ has a skewed distribution, then a concave transformation might often be very succesful in making the distribution symmetric and hence making the normal approximation more valid. The succes of the transformation will strongly depend on the actual distribution we are sampling from. So if a log transformation is generally advised for multinomial probabilities, then that insight is based on a simulation study over all possible multinomial distributions and the log should then be succesful at each multinomial distribution. A global success for a particular transformation is only possible in relative small models, at least they should be parametric. However, in general a transformation might have a bad effect for an essential subset of the model.

Since the normal approximation is often used for construction of confidence intervals the following remark is relevant. If one want to map the confidence interval for $T(\Phi(Q))$ based on the normal approximation for $T(\Phi(Q_n))$ to a confidence interval for $\Phi(Q)$ itself, by simply inverting $T$, then one should realize that the transformation $T$ will map a symmetric confidence interval to an asymmetric one. Therefore one should choose such a confidence interval for $T(\Phi(Q))$ so that it is mapped under $T^{-1}$ to a symmetric confidence interval for $\Phi(Q)$.

### 1.8.2 Proving asymptotic linearity.

There are basically two approaches for proving asymptotic linearity of an estimator $\Phi(Q_n)$ of $\Phi(Q)$; a direct approach or one based on verifying the appropriate differentiability (the so called delta-method) of $Q \to \Phi(Q)$. Let's start with an example of a direct proof of asymptotic linearity:

**Example 1.8.1** Consider the sample variance $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ as an estimator of $\sigma^2$. Notice that

$$\begin{aligned} S_n^2 - \sigma^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2 - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 + (\mu - \bar{X})^2 - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left((X_i - \mu)^2 - \sigma^2\right) - (\bar{X} - \mu)^2. \end{aligned}$$

By the CLT we know that $(\bar{X} - \mu) = O_P(1/\sqrt{n})$ and hence the second term is $O_P(1/n)$. This shows that $S_n^2$ is asymptotically linear with *influence curve* $IC_P(X) = (X - \mu)^2 - \sigma^2$ under the assumption that $E(X - \mu)^4 < \infty$. As a consequence we have that $\sqrt{n}(S_n^2 - \sigma^2) \stackrel{D}{\Longrightarrow} N(0, \mathrm{VAR}(IC_P(X))$.

### 1.8.3 The delta-method.

Suppose now that we are interested in the limit distribution of $\sqrt{n}(S_n - \sigma)$. One might hope that in this case asymptotic normality of $S_n$ follows from the asymptotic normality result for $S_n^2$. The idea is the following. Let $g(x) = \sqrt{x}$. Then $S_n = g(S_n^2)$ and $\sigma = g(\sigma^2)$. Since $g$ is differentiable we can approximate $g(S_n^2) - g(\sigma^2)$ by its linear approximation in $S_n^2 - \sigma^2$, namely $g'(\sigma^2)(S_n^2 - \sigma^2)$, where $g'(x) = 1/2\sqrt{x}$. Now, substitute for $S_n^2 - \sigma^2$ its derived (see example above) linear approximation $\frac{1}{n}\sum_{i=1}^n IC_P(X_i)$. This tells us that

$$g(S_n^2) - g(\sigma^2) \approx \frac{1}{n}\sum_{i=1}^n \frac{1}{2\sigma}\left((X - \mu)^2 - \sigma^2\right),$$

which provides us with the required asymptotic linearity of $S_n$ as an estimator of $\sigma$.

Let's now complicate things a little by letting $g$ be a function of two variables. For this purpose, suppose that we are interested in showing asymptotic linearity of $\bar{X}/S_n$ as an estimator of $\mu/\sigma$. Since we have asymptotic linearity of $\bar{X}$ as an estimator of $\mu$ and $S_n$ as an estimator of $\sigma$ we hope that (as above) asymptotic linearity will follow from differentiability of $g(x, s) \equiv x/s$. We have $\bar{X}/S_n = g(\bar{X}, S_n)$ and $\mu/\sigma = g(\mu, \sigma)$. We have the following linear approximation:

$$g(x, y) - g(x_0, y_0) \approx \frac{d}{dx} g(x, y_0) \mid_{x=x_0} (x - x_0) + \frac{d}{dy} g(x_0, y) \mid_{y=y_0} (y - y_0),$$

if $(x, y)$ is close to $(x_0, y_0)$. This tells us that

$$g(\bar{X}, S_n) - g(\mu, \sigma) \approx \frac{1}{\sigma}(\bar{X} - \mu) - \frac{\mu}{\sigma^2}(S_n - \sigma).$$

Now, we can substitute the linear approximations

$$\bar{X} - \mu = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)$$

and

$$S_n - \sigma \approx \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\sigma}\left((X_i-\mu)^2 - \sigma^2\right).$$

This tells us that

$$g(\bar{X}, S_n) - g(\mu,\sigma) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma}(X_i-\mu) + \frac{\mu}{2\sigma^3}\left((X_i-\mu)^2 - \sigma^2\right).$$

### 1.8.4 Generalizing the delta-method.

The abstract way to think about the described method is as follows: One obtains a linear approximation for an estimator $\Phi(Q_n)$ by first approximating $\Phi(Q_n)-\Phi(Q)$ by a linear mapping applied to $Q_n - Q$ and then we plug in the already obtained linear approximation for $Q_n - Q$. Having this description in mind it it easy to generalize the method to a general real valued functional $\Phi$ of a $D$-valued random variable $Q_n$. We will do this by means of an example.

**Example 1.8.2** Let $F_n(x)$ be the empirical distribution based on an i.i.d. sample of $n$ (positive valued) observations $X_i \sim F$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

Let $S_n(x) = 1 - F_n(x-)$ be the corresponding empirical survival function. The *cumulative hazard* or *integrated hazard* evaluated at a point $t$ is defined by

$$\Lambda_n(t) \equiv \int_0^t \frac{dF_n(x)}{S_n(x)}.$$

We are concerned with showing that $\Lambda_n(t)$ is an asymptotically linear estimator of $\Lambda(t) = \int_0^t dF(x)/S(x)$. Notice that $\Lambda_n(t) = g(F_n, S_n)$ and $\Lambda(t) = g(F,S)$ and we have already linear approximations for $F_n - F$ and $S_n - S$, namely

$$F_n(x) - F(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) - F(x)$$

and

$$S_n(x) - S(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \ge x) - S(x).$$

We will determine the linear approximation for $g(F_n, S_n) - g(F,S)$ in $F_n - F$ and $S_n - S$ and then substitute these linear approximations. By applying $a_n b_n - ab = (a_n - a)b + a_n(b_n - b)$ to $a_n = dF_n(x)$ and $b_n = 1/S_n(x)$ we obtain

$$\int \frac{dF_n}{S_n} - \int \frac{dF}{S} = \int \frac{d(F_n - F)}{S} + \int \frac{(S - S_n)dF_n}{S_n S}.$$

The first term is linear in $F_n - F$ and the second term has as linear approximation

$$\int \frac{(S - S_n)dF}{S^2}.$$

Hence we have

$$\Lambda_n(t) - \Lambda(t) \approx \int_0^t \frac{d(F_n - F)(x)}{S(x)} + \int \frac{(S - S_n)(x)dF(x)}{S^2(x)}.$$

Now, we can substitute the linear approximations for $F_n - F$ and $S_n - S$ to obtain that

$$\Lambda_n(t) - \Lambda(t) \approx \frac{1}{n}\sum_{i=1}^n \left( \frac{I(X_i \leq t)}{S(X_i)} + \int_0^{X_i} \frac{dF(x)}{S^2(x)} \right).$$

This shows that $\Lambda_n(t)$ is an asymptotically linear estimator of $\Lambda(t)$ with influence curve

$$IC_F(X_i) = \frac{I(X_i \leq t)}{S(X_i)} + \int_0^{X_i} \frac{dF(x)}{S^2(x)}.$$

One should notice that if one is just interested in deriving the influence curve of the estimator, then one just ignores all remainders as we did above. Of course, the derived linear approximation is only valid if the remainders are $o_P(1/\sqrt{n})$. We will now state the rigorous delta-method for estimators of the form $\Phi(F_n)$, where $F_n$ is the empirical distribution.

**Theorem 1.8.1** *(Delta-Method) Let $X$ be a $\mathbb{R}^d$-valued random variable with cumulative distribution function $F$ and let $F_n$ be the empirical distribution function of $F$ based on $n$ i.i.d. copies $X_1, \ldots, X_n$ of $X$:*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x).$$

*Let $\Phi(F) \in \mathbb{R}$ be a real valued parameter of interest and assume that $\Phi(F_n)$ is well defined.*

*Suppose that for any sequence of distributions $G_n$ for which $\|\sqrt{n}(G_n - G) - Z\|_\infty \to 0$ for some function $Z$ with finite supnorm we have the following differentiability result for $\Phi$:*

$$\Phi(G_n) - \Phi(G) = d\Phi(G)(G_n - G) + o(1/\sqrt{n}), \tag{1.31}$$

*where $d\Phi(G)$ is a linear mapping. Then*

$$\Phi(F_n) - \Phi(F) = \frac{1}{n}\sum_{i=1}^n d\Phi(F)(I(X_i \leq \cdot) - F(\cdot)) + o_P(1/\sqrt{n}).$$

*In other words, then $\Phi(F_n)$ is asymptotically linear with influence curve*

$$IC_F(X_i) = d\Phi(F)(I(X_i \leq \cdot) - F(\cdot)).$$

Let's now solve the general case where we are concerned with proving asymptotic linearity of $\Phi(Q_n) \in \mathbb{R}$ as an estimator of $\Phi(Q) \in \mathbb{R}$, where $Q_n, Q \in (D_1, \|\cdot\|_1)$ for some normed linear space $(D_1, \|\cdot\|_1)$. In other words, we do not necessarily want to restrict ourselves to the case where $D_1 = \mathbb{R}^k$ or $D_1$ is a space containing cumulative distributions of $\mathbb{R}^d$-valued random variables as we did in the preceding theorem.

Supppose that you have been able to show that $Q_n$ is asymptotically linear in the following sense

$$(Q_n - Q)(\cdot) = \frac{1}{n}\sum_{i=1}^n f_P(X_i, \cdot) + R_n,$$

where $f_p(X_i, \cdot) \in D_1$, $i = 1, \ldots, n$ and $\|R_n\|_1 = o_P(1/\sqrt{n})$. In other words, $Q_n - Q$ can be approximated by an average of i.i.d. $D_1$-valued functions $f_P(X_i, \cdot)$.

Suppose that for any sequence of distributions $G_n$ for which $\|\sqrt{n}(G_n - G) - Z\|_1 \to 0$ for some function $Z \in (D_1, \|\cdot\|_1)$ with finite supnorm we have the following differentiability result for $\Phi$:

$$\Phi(G_n) - \Phi(G) = d\Phi(G)(G_n - G) + o(1/\sqrt{n}), \tag{1.32}$$

where $d\Phi(G)$ is a linear mapping.

Then

$$\Phi(Q_n) - \Phi(Q) = \frac{1}{n}\sum_{i=1}^{n} d\Phi(Q)(f_P(X_i, \cdot)) + o_P(1/\sqrt{n}).$$

In other words, then $\Phi(Q_n)$ is asymptotically linear with influence curve

$$IC_F(X_i) = d\Phi(Q)(f_P(X_i, \cdot)).$$

### 1.8.5 Exercise, computing the influence curve of Kaplan-Meier.

The following example functions as a guide and **exercise** for you to find the influence curve of the Kaplan-Meier estimator.

**Example 1.8.3** (Univariate right censoring) Recall that we represented the survival function hazard $\lambda(dt)$ as $P_1(dt)/\bar{P}(t)$, where

$$P_1(dt, 1) \equiv P(\widetilde{T} \in dt, \Delta = 1) \text{ and } \bar{P}(t) \equiv P(\widetilde{T} > t)$$

to itself. Consequently, we have that the integrated hazard $\Lambda(t) = \int_0^t \lambda(ds)$ can be represented as a

$$\Phi(P_1, \bar{P}) = \int_0^t \frac{P_1(ds)}{\bar{P}(s)}.$$

. We can estimate $\Lambda(t)$ by replacing $P_1$ and $\bar{P}$ by their empirical distributions: i.e. $\Lambda_n(t) = \Phi(P_1^n, \bar{P}_n)$, where

$$P_1^n(t, 1) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_i \leq t, \Delta_i = 1) \text{ and } \bar{P}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_i > t).$$

Now, the result above tells us that for showing asymptotic linearity of $\Lambda_n(t)$ we just need to verify differentiability condition (1.32) for $\Phi$, which basically has been done in the preceding example; just replace the role of $F_n$ by $P_1^n$ and the role of $S_n$ by $\bar{P}_n$. You should now be able to write down the influence curve of $\Lambda_n(t)$ by going through the same derivations as in the preceding example.

Recall that $S$ is a (product integral) mapping from $\Lambda$ to $S$. So $S(t) = \Phi(\Lambda)$ and $S_n(t) = \Phi(\Lambda_n)$ for a known $\Phi$. Hence once we have the linear approximation for $\Lambda_n - \Lambda$ we can obtain the linear approximation for $(S_n - S)(t)$ by finding the linear approximation (i.e. derivative) $d\Phi(\Lambda)(\Lambda_n - \Lambda)$. If we assume that $S$ is continuous, then the linear approximation is given by:

$$S_n(t) - S(t) \approx S(t)\int_0^t d(\Lambda_n - \Lambda)(s).$$

Substitute the linear approximation of $d\Lambda_n - d\Lambda$ and you will be able to write down the influence curve of $S_n$:

$$IC_{F,G}(\widetilde{T}, \Delta) = S(t) \left( \frac{I(\widetilde{T} \leq t, \Delta = 1)}{\bar{P}(\widetilde{T})} - \int_0^{t \wedge \widetilde{T}} \frac{P_1(ds)}{\bar{P}^2(s)} \right).$$

Given this influence curve how could you use this to construct an asymptotic confidence band for $S(t)$?

## 1.9 The role of the influence curve in robustness studies.

If an estimator $\theta_n$ of $\theta$ is asymptotically linear, then $(\theta_n - \theta \approx 1/n \sum_{i=1}^n IC_P(X_i)$. Suppose that $IC_P(x_0) = \infty$ (or very large), then by shifting one observation $X_i$ to $x_0$ we can completely blow up the difference $\theta_n - \theta$. If one observation has a large influence on the estimator one would say that the estimator is not robust. Preferably, one would like to have that

$$\| IC_P \|_\infty = \sup_x \mid IC_P(x) \mid < \infty.$$

In order to control the differences in the estimator for small shifts of the observations one would like to have a bound on the modulus of continuity of the influence curve:

$$W(\delta) = \sup_{|x-y|<\delta} \mid IC_P(x) - IC_P(y) \mid$$

or on a more global measure as (if $X \in \mathbb{R}$) the total variation of the influence curve as a function of the data $X$:

$$\| IC_P(x) \|_v \equiv \int \mid IC_P(dx) \mid .$$

This total variation norm could be generalized to $X \in \mathbb{R}^d$.

**Exercise.** Show that the Kaplan-Meier estimator $S_n(t)$ is robust if $\bar{G}(t) > 0$ in the sense that one can never blow up the estimator by changing one observation. This is due to the fact that $S_n(t)$ does not use the value of the observations which fall after $t$. Because of this property it can be shown that the Kaplan-Meier estimator is asymptotically normal under the minimal assumption that $\bar{G}(t) > 0$.

## 1.10 Smoothness of the influence curve in $P$ and estimation of the influence curve.

Above we discussed robustness of $\theta_n$ against a change of an observation $X_i$. It is also of interest to know how robust the estimator is towards deviations of the distribution of the data. Let $\theta_n^1$ be based on a sample from $P_1$ and $\theta_n$ be based on a sample from $P$. If $P_1$ is close to $P$ w.r.t. a certain distance measure, then we wonder if $\theta_n^1$ is close $\theta_n$. Since $\theta_n^1$ and $\theta_n$ are asymptoticaly normal with mean zero and variance $\mathrm{VAR}(IC_{P_1}(X))$ and $\mathrm{VAR}(IC_P(X))$, respectively, it is natural to take as distance between $\theta_n^1$ and $\theta_n$

$$\mid \mathrm{VAR}(IC_{P_1}(X)) - \mathrm{VAR}(IC_P(X)) \mid .$$

Therefore robustness towards deviations of the distribution of the data depends on the smoothness of the real valued functional

$$R(P) \equiv \mathrm{VAR}(IC_P(X)),$$

where we stress that the variance is taken under $P$. This smoothness is also exactly a measure of how well one can estimate $\mathrm{VAR}(IC_P(X))$, which is necessary for construction of confidence intervals. For example, if $IC_P$ depends on $P$ only through two moments of $P$, then we can estimate $R(P)$ with $R(P_n)$ and asymptotic linearity of $R(P_n)$ follows simply from the delta method for functions of two real valued variables. If $IC_P$ depends on $P$ through a whole cumulative distribution function $F$, then we can still estimate $R(P)$ with $R(P_n)$ and show asymptotic linearity of this estimator by the generalized delta-method. However, if $IC_P$ depends on a density $f$, then estimation of $R(P)$ would generally require estimation of the density $f$ and asymptotic linearity of such estimators is more delicate.

**Example 1.10.1** (Current status model). Consider the simple current status model: we observe $(C, \Delta = I(T \leq C))$ $C$ has density $g$, $T \sim F$, $C$ and $T$ independent. We are concerned with estimation of $\mu = \int r(t)(1 - F(t))dt$. As shown in the extended current status model section we have that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{r(c_i)(1-\Delta_i)}{g(C_i)}$$

has expectation $\mu$ and hence it is natural to estimate $\mu$ with

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \frac{r(c_i)(1-\Delta_i)}{g_n(C_i)},$$

where

$$g_n(c) = \frac{1}{nh}\sum_{i=1}^{n} K(c - C_i/h)$$

is a kernel density estimator with kernel $K$ and bandwidth $h$. It can been shown (if you want, then try it as a good exercise!), that $\mu_n$ is asymptotically linear:

$$\mu_n - \mu \approx \frac{1}{n}\sum_{i=1}^{n} \frac{r(C_i)}{g(C_i)}(F(C_i) - \Delta_i).$$

**Exercise.** Compute the variance $E(IC^2_{F,g}(C, \Delta))$ of the influence curve. Comment on the smoothness of the variance as a function of the distribution of the data. Propose an estimate of this variance and write down a 0.95-confidence interval for $\mu$.

## 1.11 The role of the influence curve in efficiency studies.

Let $X_1, \ldots, X_n$ be $n$ i.i.d. observations from a probability measure $P_{F,G}$ on a measurable space, which is parametrized by two distributions $F$ and $G$ on two measurable spaces. We consider the model with $F$ and $G$ completely unspecified.

Let $(F_n, G_n)$ be an MLE of $(F, G)$ in the sense that for a dominating measure $\mu_n$ of $P_{F_n,G_n}$ we have:

$$P_{F_n,G_n} = \arg \max_{\{P_{F,G}: P_{F,G} \ll \mu_n\}} \int \log \left( \frac{dP_{F,G}}{d\mu_n} \right) dP_n(x), \tag{1.33}$$

where $P_n$ denotes the empirical distribution of $X_1, \ldots, X_n$. In particular, (1.33) will hold for an NPMLE as defined in Kiefer and Wolfowitz (1956). Suppose that we are interested in estimating $\Psi(F) \in \mathbb{R}$ for a certain *linear* real valued function $\Psi$. We will refer to $\Psi(F_n)$ as the NPMLE of $\Psi(F)$.

In nonparametric missing data models one will often encounter the situation where for any dominating measure $\mu$ $dP_{F,G}/d\mu = p_F p_G$ for certain functions $p_F$ and $p_G$, where $p_F$ does not depend on $G$ and $p_G$ does not depend on $F$; in particular, if $Y = \Phi(C, X)$ and the conditional distribution $G(\cdot \mid x)$ of $C$, given $X = x$ satisfies CAR. In such models the likelihood factorizes in a $F$ and $G$ part so that $F_n$ can be determined by just maximizing the relevant part of the loglikelihood. Also information calculations (below) do not depend on knowledge on $G$ and hence we can do as if $G$ is known.

In section we derive a crucial identity (see (1.41)) for missing and biased sampling models, by exploiting their specific structure (see (1.40)), and show how the combination of this identity with the efficient score equation (see (1.42)) often leads to a powerful identity (see (1.43)) for the NPMLE which forms an effective starting point for proving consistency and efficiency of $\Psi(F_n)$ (also in models where the NPMLE is highly implicit). The identity is an extension of the identity for missing data models derived in van der Laan (1995a). We will first review some efficiency theory as can be found in Bickel, Klaassen, Ritov and Wellner (1993) (we will abbreviate this reference with BKRW). Then we will derive the identity and discuss its application in proving efficiency of $\Psi(F_n)$.

We remark here that the results are trivially extended to any parametrization $P_{\theta,\eta}$ having the same structure (1.40); just replace $F$ by $\theta$ and $G$ by $\eta$ in the sequel.

### 1.11.1 Identity for NPMLE in biased sampling models.

Let $F \ll \mu_1$, $G \ll \mu_2$ and denote the corresponding densities with $f$ and $g$, respectively. If we write $F_1 \ll_b F_2$ for two measures $F_1, F_2$, then we mean that $F_1$ is absolutely continuous w.r.t. $F_2$ and that $dF_1/dF_2$ is bounded. For each $F_1 \ll_b F$ we define a line $f_{h_1}(\epsilon) = (1 + \epsilon h_1) f$ from $F_1$ to $F$, where $h_1 = (f_1 - f)/f \in L_0^2(F)$. Because $h_1$ is bounded it follows that $f_{h_1}(\epsilon)$ is also a well defined density for $\epsilon \in [-\delta, 1]$ for some $\delta > 0$.

Similarly, for each $G_1 \ll_b G$ we define a line $g_{h_2}(\epsilon) = (1 + \epsilon h_2) g$ from $G_1$ to $G$, where $h_2 = (g_1 - g)/g \in L_0^2(G)$. These lines imply a one-dimensional submodel $P_{f_{h_1}(\epsilon), g_{h_2}(\epsilon)}$ through $P_{f,g}$. We will assume that

$$\frac{d}{d\epsilon} \log \left( \frac{dP_{f_{h_1}(\epsilon), g_{h_2}(\epsilon)}}{dP_{F,G}} \right) \Bigg|_{\epsilon=0} = A_{F,G}(h_1) + B_{F,G}(h_2), \tag{1.34}$$

where the so called *score operators* $A_{F,G} : L_0^2(F) \to L_0^2(P_{F,G})$ and $B_{F,G} : L_0^2(G) \to L_0^2(P_{F,G})$ are

defined by

$$A_{F,G}(h_1) \equiv \frac{d}{d\epsilon} \log \left( \frac{dP_{f_{h_1}(\epsilon),g}}{dP_{F,G}} \right) \Big|_{\epsilon=0}$$
$$B_{F,G}(h_2) \equiv \frac{d}{d\epsilon} \log \left( \frac{dP_{f,g_{h_2}(\epsilon)}}{dP_{F,G}} \right) \Big|_{\epsilon=0}.$$

The equalities in (1.34) and the limits in $d/d\epsilon$ are assumed to hold in $L^2(P_{F,G})$. In view of linearity of $\Psi$ the Cramér-Rao lower bound for the variance of a $\sqrt{n}$-normed unbiased estimator of $\Psi(F) = \Psi(F_{h_1}(0))$ along the one-dimensional submodel $P_{F_{h_1}(\epsilon),G_{h_2}(\epsilon)}$ with parameter $\epsilon$ is now given by:

$$\left( \frac{\frac{d}{d\epsilon}\Psi(F_{h_1}(\epsilon))\big|_{\epsilon=0}}{\|A_{F,G}(h_1) + B_{F,G}(h_2)\|_{P_{F,G}}} \right)^2 = \left( \frac{\Psi(\int^{\cdot} h_1 dF)}{\|A_{F,G}(h_1) + B_{F,G}(h_2)\|_{P_{F,G}}} \right)^2. \tag{1.35}$$

Now, one obtains a Cramér-Rao lower bound for the whole model by taking the supremum of these one-dimensional lower bounds over $h_1 \in L_0^2(F)$ and $h_2 \in L_0^2(G)$. Because the numerator in (1.35) does not depend on $h_2$ we can maximize this bound by minimizing the denominator in $h_2$ for fixed $h_1$. For this purpose define $T_2(P_{F,G}) = \overline{B_{F,G}(L_0^2(G))}$, where the closure is taken in $L_0^2(P_{F,G})$. In order to minimize the denominator in $h_2$ one has to choose $h_2$ such that $B_{F,G}(h_2) = -\Pi(A_{F,G}(h_1) \mid T_2(P_{F,G}))$, where $\Pi(\cdot \mid T_2(P_{F,G}))$ denotes the projection operator in $L_0^2(P_{F,G})$ on the subspace $T_2(P_{F,G})$.

Hence the Cramér-Rao lower bound for the whole model is given by:

$$\sup_{h_1 \in L_0^2(F)} \left( \frac{\Psi(\int^{\cdot} h_1 dF)}{\|A_{F,G}^*(h_1)\|_{P_{F,G}}} \right)^2, \tag{1.36}$$

where $A_{F,G}^* : L_0^2(F) \to L_0^2(P_{F,G})$ is defined by:

$$A_{F,G}^*(h) = A_{F,G}(h) - \Pi\left(A_{F,G}(h) \mid T_2(P_{F,G})\right).$$

$A_{F,G}^*$ is called the *efficient score operator*. If $h_1 \to \Psi(\int^{\cdot} h_1 dF)$ as a mapping from $L_0^2(F)$ to $\mathbb{R}$ is continuous, then by the Riesz-representation theorem we have for a certain $\kappa(F, \Psi) \in L_0^2(F)$:

$$\Psi(F_1) - \Psi(F) = \Psi(\int^{\cdot} h_1 dF) = \langle \kappa(F, \Psi), h_1 \rangle_F.$$

Let $A_{F,G}^{*\top} : L_0^2(P_{F,G}) \to L_0^2(F)$ be the *adjoint* of $A_{F,G}^*$: for all $h \in L_0^2(F)$ and $v \in L_0^2(P_{F,G})$ we have:

$$\langle A_{F,G}^*(h), v \rangle_{P_{F,G}} = \langle h, A_{F,G}^{*\top}(v) \rangle_F.$$

If $\kappa(F, \Psi)$ lies in the range of the so called *information operator* $I_{F,G} \equiv A_{F,G}^{*\top} A_{F,G}^*$, then

$$\Psi(F_1) - \Psi(F) = \langle I_{F,G} I_{F,G}^-(\kappa(F, \Psi)), h_1 \rangle_F = \langle \ell^*(F, G, \Psi), A_{F,G}^*(h_1) \rangle_{P_{F,G}}, \tag{1.37}$$

where

$$\ell^*(F, G, \Psi) = A_{F,G}^* \left( A_{F,G}^{*\top} A_{F,G}^* \right)^- (\kappa(F, \Psi)). \tag{1.38}$$

If the latter holds, then by Cauchy-Schwarz (1.36) is given by the variance of $\ell^*(F,G,\Psi)$. $\ell^*$ can also be obtained by projecting an influence curve of an estimator for $\Psi(F)$ on the range of $A^*_{F,G}$; the projection of an influence curve on the tangent space equals the efficient influence curve.

According to general theory this quantity (1.36), usually called the information bound, is also the optimal asymptotic variance of $\sqrt{n}(\Psi(F_n) - \Psi(F))$ if $\Psi(F_n)$ is a regular estimator (BKRW). For us the most relevant result from this theory is that $\Psi(F_n)$ is an asymptotically efficient estimator of $\Psi(F)$ if and only if

$$\Psi(F_n) - \Psi(F) = \int \ell^*(F,G,\Psi)(x) d(P_n - P_{F,G})(x) + o_P(1/\sqrt{n}). \tag{1.39}$$

Therefore $\ell^*(F,G,\Psi)$ is often called the *efficient influence curve* for estimating $\Psi(F)$.

We will now show that (1.37) has a convenient form in missing and biased sampling models. Because $\ell^*(F,G,\Psi) \perp T_2(P_{F,G})$ we have that

$$\langle \ell^*(F,G,\Psi), A^*_{F,G}(h_1)\rangle_{P_{F,G}} = \langle \ell^*(F,G,\Psi), A_{F,G}(h_1)\rangle_{P_{F,G}}.$$

Suppose now that $P_{F,G}$ satisfies the typical structure from missing and biased sampling models given by:

$$P_{F,G} = \frac{1}{\alpha(F,G)} P'_{F,G}, \text{ where } F \to \alpha(F,G) \text{ and } F \to P'_{F,G} \text{ are linear.} \tag{1.40}$$

Then it is easily verified that

$$A_{F,G}(h_1) dP_{F,G} = -\frac{\alpha(F_1 - F, G)}{\alpha^2(F,G)} dP'_{F,G} + \frac{dP'_{F_1-F,G}}{\alpha(F,G)}.$$

We also have:

$$\begin{aligned} dP_{F_1,G} - dP_{F,G} &= -\frac{\alpha(F_1 - F, G)}{\alpha(F_1,G)\alpha(F,G)} dP'_{F,G} + \frac{dP'_{F_1-F,G}}{\alpha(F_1,G)} \\ &= \frac{\alpha(F,G)}{\alpha(F_1,G)} A_{F,G}(h_1) dP_{F,G}. \end{aligned}$$

Consequently, we have that (1.37) reduces to the following identity for a pair $(F, F_1)$ with $F \ll_b F_1$: (we exchanged the roles of $F$ and $F_1$)

$$\Psi(F_1) - \Psi(F) = -\frac{\alpha(F,G)}{\alpha(F_1,G)} \int \ell^*(F_1,G,\Psi)(x) dP_{F,G}(x). \tag{1.41}$$

We want to apply this identity (1.41) to $F_1 = F_n$. Usually $F_n$ does not dominate $F$ so that this identity cannot be directly applied. However, notice that the identity holds in particular for $F_1 = F_n(\alpha) \equiv (1-\alpha)F_n + \alpha F$ for any $\alpha \in (0,1]$. Hence if $\ell^*(F_n(\alpha), G, \Psi)$ converges to $\ell^*(F_n, G, \Psi)$ in $L^1(P_{F,G})$ for $\alpha \to 0$, then the identity (1.41) holds also for $F_n$. Since $F_n(\alpha)$ converges to $F_n$ w.r.t. each norm this is a weak continuity condition on the efficient influence function.

In many missing data models with independent censoring (coarsening at random), in the random truncation model and line segments models (Laslett, 1982, Gill, van der Laan, Wijers, 1993) the efficient score operator $A^*_{F_n,G_n}$ at $(F_n, G_n)$ does not depend on $G_n$. Hence, $\ell^*(F_n, G, \Psi)$

lies in the closure of the range of $A^*_{F_n,G_n}$. If $\ell^*(F_n, G, \Psi)$ is actually lying in the range (so it is given by (1.38)), then it is a score corresponding with a one-dimensional submodel $P_{F_n(\epsilon),G_n(\epsilon)}$ and hence it follows by simply differentiating (1.33) along this one dimensional submodel that the NPMLE $(F_n, G_n)$ should solve this score:

$$\int \ell^*(F_n, G, \Psi)(x) dP_n(x) = 0. \tag{1.42}$$

Combining this so called *efficient score equation* with (1.41) for the pair $(F, F_1) = (F, F_n)$, we obtain:

$$\Psi(F_n) - \Psi(F) = \frac{\alpha(F, G)}{\alpha(F_n, G)} \int \ell^*(F_n, G, \Psi)(x) d(P_n - P_{F,G})(x). \tag{1.43}$$

Comparing (1.43) with (1.39) teaches us that identity (1.43) almost says that $\Psi(F_n)$ is efficient. We are now in the perfect setting to apply *empirical process theory* (see e.g. van der Vaart, Wellner, 1994): If $\ell^*(F_n, G, \Psi)\alpha(F, G)/\alpha(F_n, G)$ falls in a $P_{F,G}$-Donsker class with probability tending to 1, then this identity provides us with root-$n$ consistency of $\Psi(F_n)$. If also $\|\ell^*(F_n, G, \Psi)/\alpha(F_n, G) - \ell^*(F, G, \Psi)/\alpha(F, G)\|_{P_{F,G}}$ converges to zero in probability, then we have asymptotic efficiency.

## 1.12 Computation of the efficient influence curve for current status data.

Let $X_1, \ldots, X_n$ be $n$ i.i.d. copies of a real valued $X$ with distribution function $F$, where $F$ is completely unknown. Let $C_1, \ldots, C_n$ be $n$ i.i.d. copies of a real valued $C$ with distribution function $G$ which is completely unknown. $X$ and $C$ are independent. We observe

$$Y = (C, \Delta) \equiv (C, I(X \leq C)) \sim P_{F,G} = Fg\Delta + (1 - F)g(1 - \Delta),$$

where

$$dP_{F,G}(c, \Delta) = (1 - F(c))dG(c)\Delta + F(c)dG(c)(1 - \Delta).$$

We are concerned with estimating $\mu(F, R) = \int R(x)dF(x)$.

The score operator for $F$ is given by:

$$A_F : L^2_0(F) \to L^2_0(P_{F,G}) : A_F(h)(Y) = E_F(h(X) \mid Y).$$

It is orthogonal to the score operator for $G$. Hence $A_F$ is the efficient score operator. Its adjoint is given by $A^\top_G(v)(X) = E_G(v(Y) \mid X)$. So the information operator $I_{F,G} : L^2_0(F) \to L^2_0(F)$ is given by:

$$I_{F,G}(h)(x) = \int_x^\infty \frac{\int_0^c h(x)dF(x)}{F(c)} dG(c) + \int_0^x \frac{\int_c^\infty h(x)dF(x)}{1 - F(c)} dG(c). \tag{1.44}$$

The efficient influence curve for $\mu(F, R)$ is given by:

$$A_F I^{-1}_{F,G}(R - \mu(F, R)).$$

This requires showing that there exists a $h \in L^2_0(F)$ with $I_{F,G}(h) = R - \mu(F, R)$ in $L^2(F)$.

Taking derivatives w.r.t. $dG$ on both sides and using that $\int_x^\infty h dF = -\int_0^x h dF$ provides us with the equation

$$\int_0^x h dF = -\frac{dR(x)}{dG(x)} F(x)(1 - F(x)), \tag{1.45}$$

assuming that $R \ll G$. Define $r_G(x) \equiv dR/dG(x)$. If $r_G \ll F$, then (1.45) has a unique solution $h$ and the efficient influence curve is given by:

$$I^*(F, G)(c, \Delta) = -r_G(c)(1 - F(c))\Delta + r_G(c)F(c)(1 - \Delta).$$

# Chapter 2

# Semiparametric Multivariate Regression.

This chapter will explain existing methods for generalized additive models; these are regression models for which $E(Y \mid X_1, \ldots, X_k) = \phi_1(X_1) + \ldots + \phi(X_p)$ and where $\phi_i$ might be unknown, might be known to be monotone, might be known to be linear or might be known to be parametrized in another way, $i = 1, \ldots, p$. A good reference for these methods is Hastie and Tibshirani (1990). The generalized additive models method is implemented in Splus with the function "gam" and "ace". We will provide a self-contained understanding of these methods and introduce some new applications.

## 2.1 Estimating optimal transformations

Let $Y$, $X = (X_1, \ldots, X_p)$ be random variables with $Y$ the response and $X$ the predictors. Let $\Theta(Y)$ be a mean-zero function of $Y$ and let $\Phi(X) \equiv \sum_{i=1}^{p} \Phi_i(X_i)$ be an additive function of mean-zero functions of $X_1, \ldots, X_p$. The fraction of variance not explained ($\ell^2$) by a regression of $\Theta(Y)$ on $\Phi(X)$ is

$$\ell(\Theta, \Phi) = \frac{\mathrm{E}\left\{|\Theta(Y) - \Phi(X)|^2\right\}}{\mathrm{E}\Theta^2(Y)}. \tag{2.1}$$

Breiman and Friedman (1985) proposed an algorithm for minimizing $\ell(\Theta, \Phi)$ over $\Theta$ and any additive function $\Phi$. The algorithm is called "Alternating Conditional Expectations" (ACE) and it contains an outer and inner loop.

The algorithm provides completely nonparametric estimators if no restrictions are enforced and it can be made as parametric or semiparametric as the user wants by simply enforcing the required constraints. For example, one might want to enforce linearity in $X_1$, monotonicity in $X_2$ and leave $X_3$ completely nonparametric.

Before we explain the algorithm we introduce some Hilbertspace definitions. Let $H_2(X_i)$ be the space of mean-zero functions of $X_i$, $i = 1, \ldots, p$, and let $H_2(Y)$ be the space of mean-zero functions of $Y$. Let $H_2(X) = H_2(X_1) + \ldots + H_2(X_p)$ be the sumspace; i.e. all sums $\phi_1(X_1) + \ldots + \phi_p(X_p)$. All these linear spaces are subspaces of the big Hilbertspace of all mean zero functions of $(Y, X)$ endowed with the variance norm, where we have as inner-product $\langle u, v \rangle \equiv E(u(X, Y)v(X, Y))$, the covariance. An inner product defines an orthogonality relation

between two functions by: $u \perp v$ if $\langle u, v\rangle = 0$. One can now also define the projection $P_X(V)$ of a function $V(Y)$ on $H_2(X)$ by

$$\langle V(Y) - P_X(V), \Phi(X)\rangle = 0 \text{ for all } \Phi \in H_2(X).$$

$P_X : H_2(Y) \to H_2(X)$ is a projection operator which computes for a given function $\Theta(Y)$ the additive function $\Phi(X)$ which is closest to $\Theta$ in the sense that $\|\Theta(Y) - \Phi(X)\|$ (variance norm) is minimal. Similarly, we can define the projection $P_Y(\Phi)$ of an additive function $\Phi(X) \in H_2(X)$ on $H_2(Y)$ by:

$$\langle \Phi(X) - P_Y(\Phi)(Y), V(Y)\rangle = 0 \text{ for all } V \in H_2(Y).$$

$P_Y : H_2(X) \to H_2(Y)$ is a projection operator which computes for a given additive function $\Phi(X)$ a function $\Theta$ of $Y$ which is closest to $\Phi$ in the sense that $\|\Phi(X) - \Theta(Y)\|$ is minimal. It is easily verified that $P_Y(\Phi) = E(\Phi(X) \mid Y)$, the conditional expectation operator:

$$E((\Phi(X) - E(\Phi(X) \mid Y))V(Y)) = EE\left((\Phi(X) - E(\Phi(X) \mid Y))V(Y) \mid Y\right) = 0.$$

If $X$ is univariate, then $P_X(V) = E(V(Y) \mid X)$ is also simply the conditional expectation operator, but if $X$ is multivariate, then $P_X V$ will only be equal to $E(V(Y) \mid X)$ if $E(V(Y) \mid X) \in H_2(X)$; i.e. if the conditional expection is described by an additive function. It will be relevant to notice that $P_X$ is the adjoint of $P_Y$. This is shown as follows:

$$\langle P_Y\Phi, \Theta\rangle = E(E(\Phi(X) \mid Y)\Theta(Y)) = E(\Phi(X)\Theta(Y)) = E(\Phi(X)(P_X\Theta)(Y)) = \langle \Phi, P_X\Theta\rangle.$$

Suppose the optimal $\Phi$ is given. Then $\ell$ is minimized over $\Theta$ under the restriction that $\|\Theta(Y)\| \equiv \sqrt{\int \Theta(Y)^2 dF(y)} = 1$ ($\|\cdot\|$ is the variance norm) by:

$$\Theta(Y) = \frac{E(\Phi(X) \mid Y)}{\|E(\Phi(X) \mid Y)\|} = \frac{P_Y(\Phi)}{\|P_Y(\Phi)\|}.$$

Moreover, if the optimal $\Theta(Y)$ is given, then $\ell$ is by definition of $P_X$ minimized over $\Phi \in H_2(X)$ by $P_X\Theta$. This tells us that the optimal solution $(\Theta, \Phi)$ is a solution of the following 2 equations:

$$\begin{aligned} \Theta(Y) &= \frac{P_Y(\Phi)}{\|P_Y(\Phi)\|} \\ \Phi(X) &= P_X\Theta. \end{aligned}$$

Moreover it suggests the following (outer loop) algorithm for solving these equations: Start with an initial guess $\Theta^0(Y)$ and iterate as follows till convergence

$$\Theta^{m+1}(Y) = \frac{P_Y(\Phi^{m+1})}{\|P_Y(\Phi^{m+1})\|} \text{ and } \Phi^{m+1}(X) = P_X\Theta^m.$$

Notice that this algorithm is equivalent with:

$$\Theta^{m+1}(Y) = P_Y P_X \Theta^m / \|P_Y P_X \Theta^m\|. \tag{2.2}$$

Computation of $P_X\Theta^m$ for a given $\Theta^m$ requires an algorithm itself (the inner loop) which we will explain after the outer loop.

### 2.1.1 Outer loop.

Before we discuss the inner loop we will deepen our understanding of the two equations the algorithm is solving for. By substitution of one in the other it follows that $\Theta$ and $\Phi$ solve the following uncoupled set of equations:

$$\begin{aligned} P_Y P_X(\Theta) &= \|P_Y(\Phi)\|\Theta \\ \Phi(X) &= P_X\Theta \end{aligned}$$

or

$$\begin{aligned} P_X P_Y \Phi &= \|P_Y(\Phi)\|\Phi \\ \Theta(Y) &= \frac{P_Y\Phi}{\|P_Y\Phi\|}. \end{aligned}$$

Restrict attention to the first set of equations; symmetric statements hold for the second set of equations. Since $P_Y$ is the adjoint of $P_X$ we have that $P_Y P_X : H_2(Y) \to H_2(Y)$ is self-adjoint: $(P_Y P_X)^\top = P_Y P_X$. This is an infinite dimensional equivalent of a matrix which is symmetric. For matrices it is well known that a symmetric matrix has an orthonormal basis of eigenvectors. A similar statement does not necessarily hold for self-adjoint Hilbertspace operators, but it holds if we assume that $P_Y P_X : H_2(Y) \to H_2(Y)$ is compact. Compactness means that $P_Y P_X$ maps the unit ball to a compact set (bounded, closed, and every sequence has a convergent subsequence). A sufficient (often stated) condition for the compactness of this operator is the following: let $f_{X,Y}$ be the joint density of $(X, Y)$ w.r.t. Lebesque measure and denote the marginals with $f_X$ and $f_Y$,

$$\int\int \frac{f_{X,Y}^2}{f_X f_Y} dxdy < \infty.$$

However, this assumption assumes unnecessarily smoothness of $f_{X,Y}$. For example, if $Y$ is discrete, then $P_Y P_X$ corresponds with a matrix and hence then $P_Y P_X$ is always compact; a bounded closed set in $\mathbb{R}^m$ is compact. The following more intuitive way to think about this compactness condition is the following: If $Y$ is non-discrete the condition will be satisfied when $X = x$ does not specify the support of $y$; so it excludes cases where $Y$ is a direct function of $X$ or where observing $X$ tells us that $Y$ lies in some interval which is a direct function of $X$.

Under the compactness assumption we know that $P_Y P_X$ (and $P_X P_Y$) have an orthonormal basis of eigenfunctions and it is directly verified that an eigenfunction $\Theta$ of $P_Y P_X$ implies an eigenfunction $P_X\Theta/\|P_X\theta\|$ for $P_X P_Y$ with the same eigenvalue and an eigenfunction $\Phi$ of $P_X P_Y$ implies an eigenfunction $P_Y\Phi/\|P_Y\Phi\|$ for $P_Y P_X$ with the same eigenvalue. In other words, the operators $P_Y P_X$ and $P_X P_Y$ have the same eigenvalues and 1-1 correspondence between the eigenspaces.

Any pair of such eigenfunctions $(\Theta, \Phi)$ with eigenvalue $\lambda = \|P_Y\Phi\|$ corresponds with a solution of the two equations $P_Y P_X\Theta = \lambda\Theta$, $\Phi(X) = P_X\Theta$. We are mainly interested in that pair of solutions with minimal $\ell^2$. For this purpose we express for each pair of eigenfunctions $(\Theta, \Phi)$ corresponding with eigenvalue $\lambda$

$$\ell^2(\Theta, \Phi) = 1 - 2E(\Theta\Phi) + E\Phi^2$$

in terms of $\lambda$. From $\lambda\Theta = P_Y\Phi$ it follows that $\lambda\Theta^2(Y) = E(\Theta(Y)\Phi(X))$. Since $E\Theta^2(Y) = 1$ it follows that $\lambda = E(\Theta(Y)\Phi(X))$. From the equation $\Phi(X) = P_X(\Theta)$ it follows that $\Phi^2(X) = \Phi(X)P_X(\Theta)$ and we have

$$E(\Phi(X)P_X(\Theta)) = E(\Phi(X)E(\Theta(Y) \mid X)) = E(\Phi(X)\Theta(Y)) = \lambda.$$

So $\lambda = E(\Phi^2(X))$. Consequently, we have for any pair of eigenfunctions $\Theta, \Phi$ corresponding with eigenvalue $\lambda$

$$\ell^2(\Theta, \Phi) = 1 - 2\lambda + \lambda = 1 - \lambda.$$

This shows that solving for the optimal transformations comes down to solving for the eigenfunction $\Theta$ of $P_Y P_X : H_2(Y) \to H_2(Y)$ with maximal eigenvalue.

Does the outer-loop algorithm converge to the maximal eigenvalue eigenfunction? Firstly, it is clear that $P_Y P_X$ maps an eigenspace corresponding with a certain eigenvalue to itself and it maps the orthonormal complement of a certain eigenspace to itself. Hence if we start the algorithm (2.2) with a $\Theta^0$ in the orthocomplement of the maximal eigenvalue eigenspace, then the algorithm (which just applies $P_Y P_X$ iteratively) will never leave this orthocomplement in the subsequent steps. So if we start with such a $\Theta^0$ we will not converge to the optimal eigenfunction. However, it can be shown that if we start with a $\Theta^0$ for which its projection on the maximal eigenspace is not zero, then the algorithm will converge to the maximal eigenvalue solution; for the proof we refer to Breiman and Friedman (1985).

If one is interested in the eigenfunctions corresponding with the second eigenvalue one simply starts with a $\Theta^0$ for which $\langle \Theta^0, \Theta^* \rangle = 0$, where $\Theta^*$ is the optimal eigenfunction. We should be a little more precise here. If the eigenspace corresponding with the maximal eigenvalue is more than 1-dimensional, then starting with a $\Theta^0$ which is orthogonal to the maximal $\Theta$ one just found will give us another maximal eigenfunction. So by starting with an initial $\Theta^0$ which is orthogonal to all (already) found eigenfunctions we can work our way down and find all eigenfunctions.

If the first and second eigenvalue are close to each other, then the corresponding eigenfunctions provide (almost) the same measure of fit. Hence, in this case, it might be of practical interest to compute also the second-optimal-eigenfunction; for example, the transformations of $X, Y$ described by the second-optimal pair of eigenfunctions might be easier interpret in practice and hence preferable. Similarly, if the eigenspace corresponding with the first eigenvalue is two dimensional, then one want to consider both eigenfunction transformations.

### 2.1.2 Applications of outer loop.

In practice, we replace the conditional expectations by data smooths and $\|V\|$ is replaced by its estimate $1/n \sum_{i=1}^n V(Y_i)^2$. Then the outerloop will provide us with a $\Theta_n$ and $\Phi_n$ for which the estimated variance of $\Theta_n(Y) - \Phi_n(X)$ is minimal. ACE is implemented in SPLUS with the function "ace". One should realize that a model $Y = \beta X + e$, $e \sim N(o, \sigma^2)$, $X$ some distribution, does not necessarily imply an outcome of ACE given by $\Theta(Y) = Y$; ACE just finds the model $\Theta(Y) = \Phi(X) + e$ with $Ee^2$ minimal and that one might, for example, correspond with an asymmetric error distribution. A model $\Theta(Y) = \Phi(X) + e$ can only be directly computed if $(\Theta, \Phi)$ is a pair of eigenfunctions which is satisfied if and only if $E(e \mid X) = 0$, $E(\Phi(X) \mid Y) = c\Theta(Y)$ for some constant $c$. These conditions are not very strong at all in the sense that it is essentially only one-dimensional restriction on the conditional distribution of $X$, given $Y$. However, for a given popular model it will usually fail to be true.

The transformation $\Theta_n$ tells us that $\Theta_n(Y)$ would have been a better variable to regress on an additive model than $Y$. If one does not want a transformation of $Y$, then one simply enforces $\Theta(Y)$ to be linear which means that in the algorithm $E(\Phi(X) \mid Y)$ is estimated with linear (weighted) regression. This will lead to better regression estimates of $E(Y \mid X)$, then applying the inverse of $\Theta_n$ to the estimate $\Theta_n(Y) = \Phi(X) + e$.

### 2.1.3 Applying ACE in determining scores empirically.

Jean Norris is working on a (huge) data-set which is used to establish relations between the health-status of the American citizen and its nutrition intake (vitamins, minerals etc.) hereby correcting for other factors which might explain the observed difference in health status. For each person a number of variables (say five) $\vec{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)$ were asked for determining his health status. It is of interest to determine a summary of these 5 variables which can be very well predicted from the $X$-variables related to Nutrition:

$$Y_s = \alpha^\top \vec{Y} = \alpha_1 Y_1 + \ldots + \alpha_5 Y_5,$$

where $\alpha_i \geq 0$, $i = 1, \ldots, 5$. The following method could be used to determine empirically the weights $\alpha$. We look at the models

$$\alpha^\top \vec{Y} = \Phi(X) + \epsilon, \text{ where } E(\epsilon \mid X) = 0.$$

We are interested in finding $(\alpha, \Phi)$, $\|\alpha^\top \vec{Y}\| = 1$, for which $E(\epsilon^2)$ is minimal. Imitating the ACE-idea, the following algorithm will converge to this optimal solution: Start with a $\alpha_0^\top \vec{Y}$ with norm 1. Compute $\Phi_1(X) = E(\alpha_0^\top \vec{Y} \mid X)$ assuming an additive model. Compute $\alpha_1^\top \vec{Y} = E(\Phi_1(X) \mid Y)/\|E(\Phi_1(X) \mid Y)\|$ assuming a linear model and proceed. So it is the outer loop of ACE, but $\alpha^\top \vec{Y}$ is here playing the role of $\Theta(Y)$. In the linear regression steps one needs to enforce $\alpha_i > 0$ (just set them zero if they happen to be negative), but if one starts with a $\alpha > 0$, then we expect that no negative $\alpha_i$ will appear. It is clear that the same algorithm works for more nonparametric regressions than $\alpha^\top \vec{Y}$.

## 2.2 The inner loop.

How do we compute $P_X h$, i.e. the projection of $h(Y)$ on the sumspace $H_2(X)$, for a given $h(Y)$. Interestingly enough this can be carried out only using the marginal projection operators $P_j$ which project a function of $X$ or $Y$ on $H_2(X_j)$, $j = 1, \ldots, p$. Recall that $P_j(v)(X_j) = E(v \mid X_j)$. For notational convenience we will set $p = 3$. Let $P_X h = h_1(X_1) + h_2(X_2) + h_3(X_3)$. The following algorithm computes $P_X h$. So if $\|f_1(X_1) + f_2(X_2) + f_3(X_3)\| = 0$ implies $f_1 = f_2 = f_3 = 0$, then the algorithm finds the unique $h_1, h_2, h_3$.

$$\begin{aligned}
h_3^1 &= P_3 h \\
h_2^1 &= P_2(h - h_3^1) \\
h_1^m &= P_1(h - h_2^m - h_3^m) && (2.3) \\
h_2^{m+1} &= P_2(h - h_1^m - h_3^{m+1}) && (2.4) \\
h_3^{m+1} &= P_3(h - h_1^m - h_2^m). && (2.5)
\end{aligned}$$

For showing the convergence of the algorithm (as done in the ACE-appendix in BKRW, 1993) one simply uses that $(I - P_j)P_j = 0$ (in other words, $P_j^2 = P_j$, projecting a function $v(X_j)$ on $H_2(X_j)$ gives $v(X_j)$ itself, of course) and (2.3, 2.4,2.5) to show

$$h - h_1^{m+1} - h_2^{m+1} - h_3^{m+1} = (I - P_1)(I - P_2)I - P_3)(h - h_1^m - h_2^m - h_3^m). \tag{2.6}$$

This works as follows:

$$(I - P_3)(h - h_1^m - h_2^m - h_3^m) = (I - P_3)(h - h_1^m - h_2^m) = h - h_1^m - h_2^m - h_3^{m+1},$$

where we used $(I - P_3)h_3^m = 0$ at the first equality and (2.5) at the second equality. Now, we have to apply $I - P_2$ to this result. The same two steps, $(I - P_2)h_2^m = 0$ and (2.4) tells us that

$$(I - P_2)(h - h_1^m - h_2^m - h_3^{m+1}) = h - h_1^m - h_2^{m+1} - h_3^{m+1}.$$

Now, applying $I - P_1$ in the same way gives us the result. Result (2.6) implies

$$h - h_1^{m+1} - h_2^{m+1} - h_3^{m+1} = (I - P_1)^m(I - P_2)^m(I - P_3)^m(h - h_3^1 - h_2^1 - h_1^1).$$

We define the operator $Q \equiv (I - P_1)(I - P_2)(I - P_3)$. We refer to a nice result of Von Neumann (see BKRW) which shows that for $m \to \infty$

$$Q^m(v) \to v - \Pi(v \mid H_2(X)).$$

This implies for any $h_i^1 \in H_2(X_i)$, $i = 1, 2, 3$

$$h - h_1^{m+1} - h_2^{m+1} - h_3^{m+1} \to h - h_3^1 - h_2^1 - h_1^1 - \Pi(h - h_3^1 - h_2^1 - h_1^1 \mid H_2(X)) = h - \Pi(h \mid H_2(x)).$$

This shows that

$$h_1^{m+1} + h_2^{m+1} + h_3^{m+1} \to \Pi(h \mid H_2(X)),$$

which completes the proof.

### 2.2.1 Applications of the inner loop ACE-algorithm.

We can use the inner loop algorithm for estimation of $E(Y \mid X)$ assuming an additive model

$$E(Y \mid X) = \Phi(X) = \Phi_1(X_1) + \ldots + \Phi(X_p).$$

The algorithm tells us that we only need to estimate the marginal conditional expectations $P_j = E(\cdot \mid X_j)$. If $X_j$ is known to be discrete, then one simply estimates $P_j(V)(x_j)$ by the average of the $Y$'s corresponding with $X_j = x_j$. If $X_j$ is continuous, then one uses a smoother. The algorithm is very easily adjusted for fitting semiparametric additive models. For example, if $\Phi_1(X_1)$ is known to be linear, then one estimates $P_1(V)$ with the least squares estimate. So for each single covariate $X_i$ one can decide to fit a particular form of $\Phi_i$ ranging from assuming a parametric form or only assuming monotonicity or whatever. If the enforced conditions on $\Phi_i$ are actually true, then that would imply a real gain; for example, as a shown later, one obtains root-$n$ estimates for a parametric $\Phi_i$ in this way.

If one assumes a model

$$Y = \beta X + e,$$

where $e \mid X$ has a mean zero density belonging to the exponential model, then it is well known that the conditional (on $X$) maximum likelihood estimator minimizes

$$\sum_{i=1}^{n} w_i^2 (Y_i - \beta X_i)^2,$$

where $w_i^2 = 1/E(e^2 \mid X_i)$. Hence in estimating $P_j$ for the parametric components one should use weighted least squares. The weighted least squares has been implemented for the linear parts of the regression in gam. Weighted least squares is equivalent with normal least squares on $Y_i/w_i$, $X_i/w_i$, for estimates $w_i$, which corresponds with $e_i/w_i$ which has constant variance. If one does not trust the estimates of $w_i$, then one can as well just apply standard least squares for the parametric $P_j$ and standard (i.e. not transforming $Y$ to have equal variance) smoothing for the nonparametric $P_j$.

## 2.3 Inner loop in generalized additive models.

The inner loop algorithm has also a nice application in non-linear additive models:

$$Y = f_0(\Phi(X)) + e,\ E(e \mid X = x) = 0 \text{ for some known real valued } f_0.$$

For example, the logistic regression model can be formulated as: $Y \in 0, 1$, $X$ is vector of covariates,

$$Y = \frac{1}{1 + \exp(-\Phi(X))} + e,\ E(e \mid x) = 0.$$

The logistic linear regression model is a popular model since the log of the odds-ratio (being close to relative risk if $P(Y = 1)$ is small) for $Y$ being 1 or 0 and $X_1 = x_1$ or $X_1 = x_1 + 1$, fixing $X_2, \ldots, X_p$, is given by $\beta_1$. In the additive logistic regression model this same log-odds-ratio is given by

$$\Phi_1(x_1 + 1) - \Phi_1(x_1).$$

So this nonparametric additive model allows for arbitrary curves for the log-odds ratio and does not enforce a constant odds-ratio in $x_1$.

How do we apply the inner loop to compute an estimate of $\Phi$? The algorithm is a simple combination of Newton-Raphson and ACE. Let $\Phi^0$ be an initial quess of $\Phi$ and set $m = 0$. We have

$$\begin{aligned} Y &= f_0(\Phi(X)) + e \\ &= f_0(\Phi^m(X)) + f_0'(\Phi^m(X))(\Phi - \Phi^m)(X) + R_m(X) + e \\ &= f_0(\Phi^m(X)) - f_0'(\Phi^m(X)) + f_0'(\Phi^m(X))\Phi(X) + R_m + e, \end{aligned}$$

where $\mid R_m(X) \mid \leq \|f_0^{(2)}\|_\infty (\Phi^m(X) - \Phi(X))^2$. The first two terms on the right-hand side are now known and will be denoted with $g_m(X)$:

$$g_m(X) \equiv f_0(\Phi^m(X)) - f_0'(\Phi^m(X)).$$

Define

$$Y_m \equiv \frac{Y - g_m(X)}{f_0'(\Phi^m(X))}, \qquad e_m = \frac{e}{f_0'(\Phi^m(X))}$$

and

$$R'_m(X) = R_m(X)/f'_0(\Phi^m(X)).$$

Then we have:

$$Y_m = \Phi(X) + R'_m(X) + e_m.$$

Notice that $E(e \mid X) = 0$ implies $E(e_m \mid X) = 0$. Hence if we ignore $R'_m(X)$, then $\Phi(X) = E(Y_m \mid X)$ and hence can be obtained with the inner loop of ACE. Now, we set $\Phi^{m+1}$ equal to this estimate and we iterate. So the algorithm computes $Y_m$, applies ACE to find $\Phi^{m+1}$, computes $Y_{m+1}$, applies ACE to find $\Phi^{m+2}$ etc, and hence it is just matter of applying a number of times ACE with the same set of covariates but each time an adjusted $Y$. Of course, weighted least squares will be part of the ACE-procedure, again.

### 2.3.1 Proof algorithm.

Let $\Pi(Y_m \mid X)$ be the projection of $Y_m$ on the sumspace $H_1(X_1) + \ldots + H_p(X_p)$. We have

$$E(Y_m \mid X) = \Phi(X) + R'_m(X) \tag{2.7}$$

If $\|f_0^{(2)}\|_\infty < \infty$, then we have

$$\|R_m(X)\| \leq c\|(\Phi^m(X) - \Phi(X))^2\|. \tag{2.8}$$

we have that

$$\|E(Y_m \mid X) - \Pi(Y_m \mid X)\| \leq \|E(Y_m \mid X) - \Phi(X)\| = \|R_m(X)\|. \tag{2.9}$$

However, by definition we have $\Pi(Y_m \mid X) = \Phi^{m+1}(X)$. Using this and (2.7) tells us that (2.9) is equivalent with:

$$\|\Phi(X) + R'_m(X) - \Phi^{m+1}(X)\| \leq R'_m(X)\|.$$

This implies (also using (2.8) that

$$\|\Phi^{m+1}(X) - \Phi(X)\| \leq 2\|R'_m(X)\| \leq 2c\|(\Phi^m(X) - \Phi(X))^2\|.$$

From this equation it follows that if we start close enough to $\Phi$, then we have quadratic convergence.

This proves convergence of the algorithm. What rate can we expect for our estimated $\Phi$? If in this algorithm $\Pi(Y_m \mid X)$ is estimated with the empirical inner-loop of ACE, then we obtain in the same way:

$$\|\hat{\Phi}^{m+1}(X) - \Phi(X)\| \leq 2c\|(\hat{\Phi}^m(X) - \Phi(X))^2\| + r(n, m),$$

where $r(n, m) = \|\hat{\Pi}(Y_m \mid X) - \Pi(Y_m \mid X)\|$. So for our final estimator (the result of the algorithm) we have $m = \infty$ and hence we have:

$$\|\hat{\Phi}(X) - \Phi(X)\| \leq 2c\|(\hat{\Phi}(X) - \Phi(X))^2\| + r(n).$$

This equation implies that

$$\|\hat{\Phi}(X) - \Phi(X)\| = O(r(n)),$$

which showes that we will have the same consistency results for $\Phi$ as we have for standard ACE-estimation of $\Pi(Y \mid X)$ for some $Y$.

## 2.4 Inner loop in Cox-proportional hazards.

The Cox-proportional hazards model models the hazard of $Y$, given a set of covariates:

$$\lambda(y \mid X) = \lambda_0(y) \exp(\beta^\top X).$$

Let $\Lambda_0(y)$ be the cumulative hazard function. This model is equivalent with

$$\log(\Lambda_0(y)) = -\beta^\top X + \epsilon, \ \exp(\epsilon) \sim \text{Exponential}(1).$$

This is proved as follows:

$$\begin{aligned} S(y \mid x) &\equiv P(Y > y \mid x) \\ &= P(\Lambda_0(Y) > \Lambda_0(y) \mid x) \\ &= P(\exp(\epsilon) > \Lambda_0(y) \exp(\beta^\top x) \mid x) \\ &= \exp(-\Lambda_0(y))^{\exp(\beta^\top x)} \\ &= S_0(y)^{\exp(\beta^\top x)}. \end{aligned}$$

So

$$P(Y \in dy \mid x) = S(y \mid x) \exp(\beta^\top x) \frac{dF_0(y)}{S_0(y)}$$

which proves that $P(Y \in dy \mid x)/P(Y > y \mid x) = \lambda_0(y) \exp(\beta^\top X)$.

This regression setup suggests to consider the much less parametric model:

$$\log(\Lambda_0(y)) = f(X) + \epsilon, \ \exp(\epsilon) \sim \text{Exponential}(1), \tag{2.10}$$

where $f(X)$ is an additive model. The following algorithm can be used to find $f$. By replacing the distribution of $(X, Y)$ by the empirical distribution we obtain an algoritm for estimation of $f$. This model has been considered in the literature; we refer to Andersen, Borgan, Gill, Keiding (1993).

Given $f$ the likelihood over $\Lambda$ is maximized by

$$\Lambda_0(y) = \int_0^y \frac{dF_0(u)}{E(e^{f(X)} I(Y \geq u))} \equiv \Phi_1(f)(y). \tag{2.11}$$

This statement is just the analogue of the well known linear case where $f(X) = \beta^\top X$. Let $\mu_1$ be the expectation of $\epsilon$. Let $\Phi(f) \equiv \log(\Phi_1(f)) - \mu_1$ which equals $\Psi_0(Y) \equiv \log(\Lambda_0(Y))$. Given $\Lambda_0$ we have that

$$f(X) = E(\log(\Lambda_0)(Y) \mid X) = E(\Phi(f)(Y) \mid X). \tag{2.12}$$

Notice that $\log(\Phi_1(f + c)) = \log(\Phi_1(f)) + c$. This shows that (2.12) is solved by any $f + c$ for a constant $c$. This is due to the fact that the constant term in $f(X)$, i.e. $f(0)$, can aswell be put in $\Lambda_0$. We can solve this by enforcing $f(0) = 0$; in the class of all functions $f$ with $f(0) = 0$ (2.12) has a unique solution.

This suggests the following algorithm for solving $f$: Start with a $f^0$ with $f^0(0) = 0$ and iterate:

$$f^{m+1}(X) = E(\Phi(f^m)(Y) \mid X).$$

Below we will show that the algorithm will converge if our initial value is close enough to the solution $f$ of (2.12). One could probably get close enough to $f$ by replacing in the first steps $f^m(x)$ by $\beta^m f^m(x)$, where $\beta^m$ is the maximizer of the partial likelihood for the linear Cox-proportional hazards model with $f^m(X)$ as covariates:

$$\lambda(y \mid X) = \lambda_0(y) \exp(\beta^\top f^m(X)).$$

This will speed up the convergence since the initial values $\beta^{m,\top} f^m(X)$ will be closer to $f$.

It is clear how this algorithm is implemented with data. $\beta^m$ will now be a function of the empirical distribution, the maximizer of the partial likelihood given $f^m(X)$ as covariates, and the expectation in $\Phi_1(f)$ is replaced by the empirical mean, and $E(\Phi(f^m)(Y) \mid X)$ is estimated with the inner loop of ACE. The algorithm will provide us with an estimate $\hat{f}(X)$ of the functional covariate $f(X)$ which implies an estimate of $\Lambda_0$ via (2.11). Moreover, as already carried out in the Splus function "gam" we can implement a stepwise-selection method for selecting the most relevant covariates and/or the number of degrees of freedom used in estimating $f_i$ at each step where we compute $E(\Phi(f^m)(Y) \mid X)$ for a given $f^m$.

After having estimated $f$ and $\Lambda_0$ using this algorithm one can compute the errors:

$$\hat{\epsilon}_i = \log(\hat{\Lambda}_0)(Y_i) - \hat{f}(X_i) - \mu_1.$$

Since (2.12) was an estimating equation following from the Cox-model our $\hat{f}$ will be a consistent estimator of $f$, assuming that the Cox-model was correct. One can use these estimated errors to fit an error distribution and in particular to test if $\epsilon = \log(\exp(1))$ which tests the validity of the nonparametric Cox-model. These estimated errors have been used for that purpose in the literature. A plot of these errors or its estimated density form nice diagnostic pictures for the validity of Cox; for example, by plotting $\hat{\epsilon}_i$ against $Y_i$ one can test if some of the used covariates are time(i.e. $Y$)-dependent.

One should realize that so far all we have assumed for the model of $Y, X$ is that for some $f$

$$\Phi_{P_{X,Y}}(f)(Y) = f(X) + \epsilon, \text{ where } E(\epsilon \mid x) = 0$$

and where we stressed the dependence of $\Phi$ on the distribution of $X, Y$. In particular, we know that if $P_{X,Y}$ satisfies Cox, then $\epsilon$ has a log-exponential(1) distribution and $f$ is the functional covariate. However, since the error distribution is completely unspecified it is clear that this model is much larger than Cox-proportional hazards. Therefore this method is much more nonparametric then another approach based on the partial likelihood, due to the fact that inner loop of ACE works with any error distribution. As a consequence this method will not be fully efficient for estimation of $f$ if one assumes Cox is valid (a fully efficient method would use that the error distribution is known and hence should be fully based on the partial likelihood, such a method is suggested by Hastie and Tibshirani in their book "Generalized Additive Models"). The nice property of the method here is that it immediately checks the validity of nonparametric Cox. If one concludes that the error distribution is log(exp(1)), then we could now slightly improve the estimator by doing linear Cox-proportional hazards with covariate $\hat{f}(X)$. In other words, we use $\hat{\beta}^\top \hat{f}(X)$ as functional covariate, where $\hat{\beta}$ is the partial likelihood estimator. If $\hat{f}$ is close to the actual $f$, then we will be close to being fully efficient, because the partial likelihood exploits the specific error distribution fully.

If it appears that the error distribution has a variance depending on $x$ (which implies that Cox is not valid), then ACE would have used the weighted least squares approach in carrying

out the ACE-inner loop for the linear terms. This would provide us with optimal estimate $\hat{f}$ assuming the large model. Suppose that there is a significant deviation from Cox. Then a question of interest is if we can recover from the estimate of $f$ an estimate of the survival function of $Y$, given $X$. It is also of interest to know how to use the estimated errors to test against particular alternatives.

In the case that Cox was not appropriate, then $(\hat{f}, \hat{\Lambda}_0)$ does not correspond with the hazard of $Y$, given $X$, as suggested by the Cox-model. However, this estimate, together with an estimate of the error distribution, implies a nonparametric estimator of $S(y \mid x)$ as follows: here $\Psi(Y) = \Phi(f)(Y)$.

$$\begin{aligned} S(y \mid x) &= P(\Psi(Y) > \Psi(y) \mid x) \\ &= P(\epsilon > \Psi(y) - f(x) \mid x) \\ &= S_\epsilon(\Psi(y) - f(x)), \end{aligned}$$

where $S_\epsilon$ is the survival function of $\epsilon$. So we can estimate $S(Y \mid x)$ by replacing $f$ by $\hat{f}$, $\Psi$ by $\log(\hat{\Lambda}_0)$ and $S_\epsilon$ by the estimate based on $\hat{\epsilon}_i$, $i = 1, \ldots, n$. If we would have observed the $\epsilon_i$, then $\hat{S}_\epsilon$ can be perfectly well estimated using these i.i.d. observations. It remains to be investigated how the replacement of $\epsilon_i$ by $\hat{\epsilon}_i$ influences the bahavior of the estimator.

### 2.4.1 Proof algorithm.

We will now prove that the algorithm solves for the actual $f$. Since $f$ solves (2.12) we have:

$$f^{m+1}(X) - f(X) = E(\Phi(f^m)(Y) - \Phi(f)(Y) \mid X).$$

By Tayloring $\log(x)$ and $e^x$ we have for a function $h$ with $\|h\|_\infty$ small:

$$\begin{aligned} \Phi(f+h)(Y) - \Phi(f)(Y) &= \log(\Phi_1(f+h))(Y) - \log(\Phi_1(f))(Y) \\ &\approx \frac{1}{\Lambda_0(Y)} (\Phi_1(f+h))(Y) - \Phi_1(f))(Y)) \\ &= \frac{1}{\Lambda_0(Y)} \left( \int_0^Y \frac{E((e^{f+h} - e^f)(X) I(Y > u))}{E(e^f(X) I(Y > u)) E(e^{f+h}(X) I(Y > u)} dF_0(u) \right) \\ &\approx \frac{1}{\Lambda_0(Y)} \left( \int_0^Y \frac{E((e^{f(X)} h(X) I(Y > u))}{E^2(e^f(X) I(Y > u))} dF_0(u) \right). \end{aligned}$$

Notice that $E(e^{f(X)} h(X) I(Y > u)) = E(e^{f(X)} h(X) S(u \mid X))$. Suppose now that

$$E(e^{f(X)} h(X) S(u \mid X)) \leq \delta E(e^{f(X)} S(u \mid X)) \|h\|_\infty \tag{2.13}$$

for some $\delta$. Then it follows that

$$\begin{aligned} \Phi(f+h)(Y) - \Phi(f)(Y) \leq \frac{\delta \|h\|_\infty}{\Lambda_0(Y)} \left( \int_0^Y \frac{E((e^{f(X)} S(u \mid X))}{E^2(e^f(X) S(u \mid X))} dF_0(u) \right) \\ = \delta \|h\|_\infty. \end{aligned}$$

The assumption (2.13) is satisfied for a $\delta < 1$ if $h \neq \|h\|_\infty$. Since $h$ is here playing the role of $f^m - f$ this shows that if $f^0$ is close enough to $f$, then $\|f^{m+1} - f\|_\infty \leq \delta_m \|f^m - f\|_\infty$, where

$\delta_m < 1$. Since $h^m = f^m - f$ will typically be partly positive and partly negative valued one can expect that (2.13) will be true for a $\delta$ essentially smaller than 1, also when $\|h^m\|_\infty$ approximates 0.

This proves that if we start close to a solution of (2.12), then we will converge to this solution. The fact that (2.12) has only one solution $f$ follows from identifiability of $(f, \Lambda_0)$ from $P_{f,\Lambda_0}$ which follows from the known identifiability result for the standard Cox-model, i.e. of $(\beta^\top X, \Lambda_0)$ from $P_{\beta^\top X,\Lambda_0}$.

## 2.5 Consistency for inner loop of ACE.

Let $\Pi(\Psi(Y) \mid H_2(X_1) + H_2(X_2)) = h_1(X_1) + h_2(X_2)$ the projection of a given function $\Psi(Y)$ on $H_2(X) = H_2(X_1) + H_2(X_2)$, where $H_2(X_i)$ is some sub-Hilbertspace of the space of all functions of $X_i$ with finite variance. Furthermore, as defined earlier, $P_i$ is the projection on $H_2(X_i)$, and $Q_i$ is the projection on the orthonormal complement of $H_2(X_i)$. If one knows that $h_1$ is linear in $X_1$, then $H_2(X_1)$ is the set of all functions linear in $X_1$ which is clearly a Hilbertspace. Similarly, one can represent $h_1$ as some linear combination of basis functions. However, if $h_1(X_1)$ is only known to be monotone, then one cannot choose $H_2(X_1)$ to be the space of all monotone functions because the difference of two monotone functions is not monotone and hence this is not a linear space. Hence $H_2(X_1)$ will always equal to closure of the linear extension of the set of functions we know $h_1$ belongs to; in the latter case, that is simply every function of $X_1$.

When estimating $P_1$ one should use specific knowledge about $h_1$. So if $h_1$ is known to be monotone, then one estimates $E(f(Y, X_2)) \mid X_1)$ in every step of the inner loop of ACE with monotonic regression (Pool Adjacent Violators Algorithm) methods and if $h_1$ is known to be linear we estimate the conditional expectation with linear regression. We will assume that $P_1$ can be estimated with a better rate than $P_2$. The conclusion of the derivations below will be that we can estimate $h_1$ just as well as we can estimate $P_1 h_2$, $h_2$ unknown, which teaches us that if $H_2(X_1)$ is finite dimensional, then we can estimate $h_1$ at root-$n$ rate, even when we estimate $h_2$ nonparametrically! It also tells us the results for the general multivariate case: Say we are projecting on a sumspace of four Hilbert-spaces $H_2(X_i)$, $i = 1, 2, 3, 4$. This sumspace can also be represented as a sumspace of two Hilbert-spaces: $H_2(X_1, X_2) = H_2(X_1) + H_2(X_2)$ and $H_2(X_3, X_4) = H_2(X_3) + H_2(X_4)$. Now, our result says that we can estimate $h_1 + h_2$ just as well as we can estimate $\Pi(h_3 + h_4 \mid H_2(X_1, X_2))$ and $h_3 + h_4$ just as well as we can estimate $\Pi(h_1 + h_2 \mid H_2(X_3, X_4))$. Now, apply our result to $\Pi(h_3 + h_4 \mid H_2(X_1, X_2))$ and $\Pi(h_1 + h_2 \mid H_2(X_3, X_4))$ to obtain that we can estimate $h_i$ just as well as $P_i f$ for some $f$, $i = 1, 2, 3, 4$.

The inner-loop of ACE for computing $h_1 + h_2$ tells us that $h_1, h_2$ solves:

$$\begin{aligned} h_1(X_1) &= P_1(\Psi(Y) - h_2(X_2)) \\ h_2(X_2) &= P_2(\Psi(Y) - h_1(X_1)). \end{aligned}$$

Subtstitution of the first equation in the second gives us the uncoupled set of equations:

$$h_2(X_2) - P_2 P_1(h_2) = P_2 Q_1 \Psi \tag{2.14}$$

$$h_1(X_1) = P_1(\Psi(Y) - h_2(X_2)). \tag{2.15}$$

Let now $P_2^n, P_1^n$ be estimates of $P_1, P_2$. Then the estimated $h_i^n$ as computed with the ACE-inner loop based on the data solve:

$$\begin{aligned} h_2^n(X_2) - P_2^n P_1^n(h_2^n) &= P_2^n Q_1^n \Psi & (2.16)\\ h_1^n(X_1) &= P_1^n(\Psi(Y) - h_2^n(X_2)). & (2.17)\end{aligned}$$

By subtracting (2.16) from (2.14) and doing the telescoping we obtain:

$$\begin{aligned} h_2^n - h_2 &= (P_2^n Q_1^n - P_2 Q_1)(\Psi) + (P_2 - P_2^n)P_1 h_2 + P_2^n(P_1 - P_1^n)h_2 - P_2^n P_1^n(h_2^n - h_2).\\ &\equiv \epsilon_n + \epsilon_2^n + \epsilon_1^n - P_2^n P_1^n(h_2^n - h_2).\end{aligned}$$

If $\rho(H_2(X_1), H_2(X_2)) < 1$, then $\sup_{\|h\|=1} \|P_2 P_1(h)\| < 1$. Hence, if $P_i^n$ are consistent in the sense that

$$\|P_2^n P_1^n\| \to \|P_2 P_1\| \text{ in probability,}$$

then we also have that $\|P_2^n P_1^n\| < 1$ for $n$ large enough.

This implies that the linear operator $(I + P_2^n P_1^n) : H_2(X,Y) \to H_2(X,Y)$ has an inverse given by

$$\sum_{k=0}^{\infty} (-1)^k (P_2^n P_1^n)^k$$

with norm bounded by $1/1 - \|P_2^n P_1^n\|$. Consequently, we have

$$\|(h_2^n - h_2)\| = \|\sum_{k=0}^{\infty} (-1)^k (P_2^n P_1^n)^k \left(\epsilon_n + \epsilon_2^n + \epsilon_1^n\right)\| \leq \frac{1}{1 - \|P_2^n P_1^n\|} \|\epsilon_n + \epsilon_2^n + \epsilon_1^n\|.$$

This implies that $h_2^n - h_2$ can be estimated with a rate given by (recall that $\epsilon_2^n$ is slower than $\epsilon_1^n$)

$$\|\epsilon_2^n\| = \|P_2^n P_1 h_2 - P_2 P_1 h_2\|.$$

Let's now see what this means for estimation of $h_1$: taking the difference between the $h_1$ and $h_1^n$ equations provides us with:

$$h_1^n - h_1 = (P_1^n - P_1)(h_2) - P_1^n(h_2^n - h_2).$$

So if the second term converges slower to zero than the first term, then the fact that $P_2$ was harder to estimate than $P_1$ lowered the rate for estimating $h_1$. However, if $H_2(X_1)$ is finite dimensional, then $P_1 h_2$, which is some smooth function of $h_2$ (it depends on $h_2$ only through a finite dimensional vector), is a parameter ($h_2$ unknown), can be estimated at root-$n$ rate. Hence one should expect that $P_1 h_2^n$ converges at root-$n$ rate to $P_1 h_2$.

This is easily shown by going back to the equation we derived for $h_2^n - h_2$, but apply now $P_1$ to both sides:

$$P_1^n h_2^n - P_1^n h_2 = P_1^n \epsilon_n + P_1^n \epsilon_2^n + P_1^n \epsilon_1^n - P_1^n P_2^n P_1^n (h_2^n - h_2).$$

So if we define $\mu_2^n \equiv P_1^n h_2^n - P_1^n h_2$ and use again that $(I - P_1^n P_2^n)^{-1}$ is given by the Neumann series of $P_1^n P_2^n$, then we obtain:

$$\begin{aligned} \|\mu_2^n\| &= \|\sum_{k=0}^{\infty} (-1)^k (P_1^n P_2^n)^k \left(P_1^n \epsilon_n + P_1^n \epsilon_2^n + P_1^n \epsilon_1^n\right)\|\\ &\leq \frac{1}{1 - \|P_2^n P_1^n\|} \|P_1^n(\epsilon_n + \epsilon_2^n + \epsilon_1^n)\|.\end{aligned}$$

Standard analysis (the $\epsilon$ terms only involve smoothers based on the data) can now be used to show that $\|P_1^n(\epsilon_n + \epsilon_2^n + \epsilon_1^n)\|$ converges to zero at root-$n$ rate. This kind of analysis is based on the idea that the smoother itself does not converge at root-$n$ rate, but if we integrate out the smoother we obtain a smoothed empirical cumulative distribution function which can be shown to converge at root-$n$-rate under some conditions on kernel and bandwidth.