

# Identity for the NPMLE in Censored Data Models

Mark J. van der Laan  
Division of Biostatistics  
School of Public Health  
University of California, Berkeley

## Abstract

We derive an identity for nonparametric maximum likelihood estimators (NPMLE) and regularized MLEs in censored data models which expresses the standardized maximum likelihood estimator in terms of the standardized empirical process. This identity provides an effective starting point in proving both consistency and efficiency of NPMLE and regularized MLE. The identity and corresponding method for proving efficiency is illustrated for the NPMLE in the univariate right-censored data model, the regularized MLE in the current status data model and for an implicit NPMLE based on a mixture of right-censored and current status data. Furthermore, a general algorithm for estimation of the limiting variance of the NPMLE is provided.

**Keywords:** censored data, nonparametric maximum likelihood estimator, asymptotically efficient estimator, efficient influence curve

## 1 Introduction.

Let  $T$  be a random variable of interest with distribution function  $F$ . In many medical and biological applications the information on  $T$  is imprecise in the sense that we only observe  $X = \Phi(T, C)$  for a known many to one mapping  $\Phi$  and a censoring variable  $C$ . The probability distribution of  $X$  is indexed by the distribution  $F$  of  $X$  and the conditional distribution  $G$  of  $C$ , given  $X$ , and will be denoted with  $P_{F,G}$ . We let  $F$  be unspecified, but assume that the conditional distribution  $G(\cdot | T)$  of  $C$ , given  $T$ , satisfies *coarsening at random*. Coarsening at random was originally formulated by Heitjan and Rubin (1991), generalized by Jacobsen and Keiding (1994) and further generalized in Gill, van der Laan and Robins (1997).

A general definition of coarsening at random (CAR) in terms of the conditional distribution of  $X$ , given  $T$ , is given in Gill, van der Laan, Robins (1997): for each  $t, t'$

$$P_{X|T=t}(dx) = P_{X|T=t'}(dx) \text{ on } \{x : t \in C(x)\} \cap \{x : t' \in C(x)\}, \quad (1)$$

where  $C(x)$  represents the coarsening for  $T$  implied by  $x$ , i.e.  $C(x)$  consists of all  $t$  for which  $X = x$  is a possible observation. It is also possible to define coarsening at random

in terms of densities: for every  $t$  we have that for a dominating measure which satisfies (1) itself

$$p(x | t) = h(x) \text{ for some function } h.$$

In other words, for each  $t$  the density of  $x$ , given  $T = t$ , depends only on  $x = \Phi(t, c)$ . Coarsening at random is an attractive assumption since it implies factorization of the density  $p(x)$  (i.e. the likelihood of  $X = x$ ) into  $F$  and  $G$ -parts so that the maximum likelihood estimator of  $F$  does not depend on the chosen model for  $G$ .

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. copies of  $X$ . Let  $P_{F_n, G}$  be a nonparametric maximum likelihood estimator (NPMLE) in the sense of Kiefer and Wolfowitz (1956) in the model with  $G$  known, assuming it exists. Let  $\mu_n$  be a measure which dominates  $P_{F_n, G}$ . Then

$$P_{F_n, G} = \arg \max_{P_{F, G} \ll \mu_n} \int \log \left( \frac{dP_{F, G}}{d\mu_n} \right) dP_n(x), \quad (2)$$

where  $P_n \equiv \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  denotes the empirical distribution function of the  $X_i$ 's and  $G$  is fixed. Here  $\delta_x$  denotes the probability measure with mass 1 at  $x$ . Let  $\mu = F\kappa \equiv \int \kappa(t) dF(t) \in \mathbb{R}$  be a parameter of interest. We will refer to  $\mu_n = F_n\kappa = \int \kappa(t) dF_n(t)$  as the NPMLE of  $\mu$ .

This paper deals with proving efficiency of  $\mu_n$ . A concise summary of efficiency theory is that an estimator  $\mu_n$  of  $\mu$  is efficient if it can be asymptotically approximated by a sum of i.i.d. random variables, where the variance of these random variables equals the semiparametric information bound. This random variable is a uniquely defined function of  $X$ , depending on the true distribution of  $X$ , and it is called the *efficient influence function*. In the next section we will provide the most relevant summary of efficiency theory for the purposes of this paper.

Nonparametric maximum likelihood estimation has received much attention in the recent literature. Beyond many analyses of NPMLE in particular semiparametric models, there have been published some papers on general understanding of the asymptotic properties of the NPMLE. Gill and van der Vaart (1993) show that once one assumes that the NPMLE is  $\sqrt{n}$ -consistent and asymptotically normally distributed, it is also asymptotically efficient under weak regularity conditions. Van der Vaart (1992) provides general conditions under which the NPMLE is efficient, assuming uniform consistency of the NPMLE  $F_n$ . One of the key conditions is Fréchet-differentiability of an infinite dimensional estimating equation.

In this paper we consider an *identity* which represents the difference  $\mu_n - \mu$  as the empirical difference of the efficient influence function evaluated at  $(F_n, G)$ . We will provide conditions under which this identity is true and show how this identity provides an effective starting point in proving *both* consistency and efficiency of the NPMLE  $\mu_n$ . This identity approach exploits the convexity and linearity in  $F$  of the model for  $X$  and as a consequence does not require Fréchet-differentiability and does not require that consistency has been proved by other means.

The purpose of this paper is to show the effectiveness of this identity approach for proving consistency and efficiency of NPMLE (and regularized MLE) in a large class of censored data models. In order to illustrate the identity approach for proving consistency

and efficiency, we chose the *univariate censoring model*, the standard MLE in the *current status data model* and a regularized MLE in the same model with a useful AIDS application. Moreover, we analyze the implicit NPMLE for a mixture of current status and right-censored data in order to show how the approach can be applied to complicated models in which the NPMLE does not exist in closed form. The right-censored data example and the last example represent data structures  $\Phi(T, C)$  which allow a complete observation of  $T$ . The current status data structure only contains censored information on  $T$ .

In the first three examples the efficient influence function exists in closed form so that the identities can be verified explicitly. The examples show that if the efficient influence function exists in closed form, then one can directly verify the identity and it often has a surprising character because of its direct link between  $\mu_n - \mu$  and  $P_n - P_{F,G}$ .

The univariate censoring example shows that the identity for the NPMLE (which is the Kaplan-Meier estimator) in combination with empirical process theory yields a direct approach to consistency, efficiency, and bootstrap consistency. Though it is not the purpose of this example, this efficiency proof of the Kaplan-Meier estimator is new. The identity has been applied in Bitouzé (1995) to prove new exponential (non-asymptotic) bounds for the density of the smoothed Kaplan-Meier estimator. For an application of the identity to prove consistency and efficiency of NPMLE in a class of censored data models allowing complete observations on  $T$ , we refer to van der Laan (1996, chapter 3).

The NPMLE in the current status data model is analyzed in Groeneboom (1991) and Groeneboom and Wellner (1992). Here the identity does *not* exactly hold for the NPMLE and hence this example should be incorporated in this paper (our second example). However, as shown in the example, in such models the approximation of the identity forms a useful starting-point in an analysis. This is also shown by a paper of van der Geer (1994), where she proves efficiency of the NPMLE in a large class of mixture models and where her general proof uses the identity at  $F_n(\alpha_n) \equiv (1 - \alpha_n)F_n + \alpha_n F$  for a sequence  $\alpha_n \rightarrow 0$ .

The third example considers a regularized MLE  $F_n$  in the current status data model, which is obtained by replacing the empirical distribution  $P_n$  in the loglikelihood by a smoothed empirical distribution  $\tilde{P}_n$  and then maximizing over  $F$ , where  $\tilde{P}_n$  is an integrated kernel density estimator  $p_n$ . Here the conditions for the identity are trivially verified. This example illustrates that the identity approach is effective in proving efficiency of regularized MLE in general.

The extension of the presented approach for proving efficiency of the NPMLE to a general semiparametric model is given in van der Laan (1994). Moreover, it is shown in van der Laan (1995) that an extension of the identity holds in biased sampling models, like the random truncation model.

## 1.1 The organization of the paper.

In the next section we will provide a concise summary of efficiency theory which is relevant for understanding the derivation of the identity. In section 3 we prove a theorem stating the conditions under which the identity is true. Here it is also shown that this identity can be used to prove consistency and efficiency of the NPMLE  $F_n\kappa$ . Furthermore, we provide an algorithm for computing the actual semiparametric information bound for the

parameter  $F\kappa$  which is useful in construction of confidence intervals. Section 4 is devoted to the four examples.

## 2 Relevant efficiency theory.

Consider a given rich class of one-dimensional submodels with parameter  $\epsilon$  through  $P_{F,G}$  at  $\epsilon = 0$ . An estimator  $\mu_n$  is *regular* at  $P_{F,G}$  if  $\sqrt{n}(\mu_n - \mu)$  converges in distribution to a limit distribution which is undisturbed by each one-dimensional  $\epsilon = 1/\sqrt{n}$ -perturbation of the data-generating distribution  $P_{F,G}$  in this class of submodels at  $P_{F,G}$ . The class of regular estimators at  $P_{F,G}$  excludes estimators which are particular well adapted to the single element  $P_{F,G}$  like the estimator  $\mu_n = \mu$  of  $\mu$ . Therefore it can be viewed as a class of “well behaved” (in a neighborhood at  $P_{F,G}$ ) estimators. Efficiency theory is concerned with identifying a semiparametric information bound at  $P_{F,G}$  for the limit variance of estimators which are regular at  $P_{F,G}$ . This information bound is uniquely determined by the canonical gradient of the so called pathwise derivative of the parameter  $F\kappa$  defined on the Hilbert space generated by the scores of this given class of submodels. In this section we will derive this canonical gradient at  $P_{F,G}$  relative to a specified given class of submodels and define efficiency of an estimator in terms of this canonical gradient.

Since the likelihood factors into an  $F$  and a  $G$ -part the canonical gradient does not depend on the model for  $G$  (see Bickel, Klaassen, Ritov, Wellner, 1993, which will be abbreviated with BKRW). Therefore we can proceed as if  $G$  is known and thus fixed. To begin with we will need to define a rich class of submodels. Our class of submodels at  $P_{F,G}$  is induced by a class of submodels at  $F$ . Let  $\mu_1$  be a dominating measure of  $F$ , i.e.  $F \ll \mu_1$ , and denote the corresponding density with  $f$ . 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 uniformly bounded. For each  $F_1 \ll_b F$  we define a line from  $F_1$  to  $F$  by the densities  $f_h(\epsilon) = (1 + \epsilon h)f$ , where  $h = (f_1 - f)/f \in L_0^2(F)$ . By convexity of the parameter space the lines form submodels with parameter  $\epsilon \in [0, 1]$ . We endow  $L_0^2(F)$  with the usual inner product norm  $\|h\|_F^2 = \langle h, h \rangle_F = \int h^2 dF$ . The set of scores corresponding with these lines through  $F$  are given by:

$$S_1(F) \equiv \left\{ h \equiv \frac{dF_1 - dF}{dF} : F_1 \ll_b F, \|h\|_\infty < \infty \right\} \subset L_0^2(F).$$

It is trivially verified that the underlying tangent space  $T_1(F) \equiv \overline{\text{Lin}S_1(F)}$ , the  $L^2(F)$ -closure of the linear span of  $S_1(F)$ , is given by  $L_0^2(F)$ .

These *lines* yield one-dimensional submodels  $p_{f_h(\epsilon),G}$  through  $p_{f,G}$ . It is well known that (see BKRW, section 6.4) that for every  $h \in S_1(F)$

$$\left. \frac{d}{d\epsilon} \log \left( p_{f_h(\epsilon),G} \right) \right|_{\epsilon=0} (x) = A_F(h)(x), \quad (3)$$

where

$$A_F : L_0^2(F) \rightarrow L_0^2(P_{F,G}) : A_F(h) = E_F(h(T) | X). \quad (4)$$

$A_F$  is called the *score operator* of  $F$ . In other words, in censored data models the score operators are just conditional (given  $X$ ) expectation operators and are bounded linear operators.

Let  $h \in S_1(F)$ . The information lower bound for the variance of a regular estimator of  $F\kappa = F_h(0)\kappa$  along the one-dimensional submodel  $P_{F_h(\epsilon)}$  with parameter  $\epsilon$  is given by:

$$\left( \frac{\frac{d}{d\epsilon} F_h(\epsilon)\kappa \Big|_{\epsilon=0}}{\|A_F(h)\|_{P_{F,G}}} \right)^2 = \left( \frac{\int \kappa h dF}{\|A_F(h)\|_{P_{F,G}}} \right)^2. \quad (5)$$

One obtains an information lower bound for the whole model by taking the supremum of these one-dimensional lower bounds over  $h$  varying over the closure of the linear span of  $S_1(F)$ , which equals  $L_0^2(F)$  (see e.g. van der Vaart, 1988).

We will determine this supremum. Define

$$\kappa_F \equiv \kappa - \int \kappa(x) dF(x).$$

Let  $A_F^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F) : A_F^\top(v) = E_G(V(X) | T)$  be the *adjoint* of  $A_F$ . Suppose that  $\kappa_F$  lies in the range of  $A_F^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F)$ . Then there exists a  $\ell(\cdot | F, G, \kappa)$  such that

$$A_F^\top \ell(\cdot | F, G, \kappa) = \kappa_F. \quad (6)$$

This implies that

$$\langle \kappa, h \rangle_F = \langle \ell(\cdot | F, G, \kappa), A_F(h) \rangle_{P_{F,G}}. \quad (7)$$

Thus (6) implies that

$$\frac{1}{\epsilon} (F_h(\epsilon)\kappa - F\kappa) = \langle \kappa, h \rangle_F = \langle \ell(\cdot | F, G, \kappa), A_F(h) \rangle_{P_{F,G}},$$

which shows that (6) implies that  $V(P_{F,G})$  is pathwise differentiable with derivative given by  $\langle \ell(\cdot | F, G, \kappa), \cdot \rangle_{P_{F,G}}$ . Any solution  $\ell(\cdot | F, G, \kappa)$  of (6) is called a gradient of the pathwise derivative. If (6) has a solution, then there exists a unique gradient  $\tilde{\ell}(\cdot | F, G, \kappa) \in \overline{R(A_F)}$  which is called the *canonical gradient*. Thus (7) can be replaced by:

$$\langle \kappa, h \rangle_F = \langle \kappa_F, h \rangle_F = \langle \tilde{\ell}(\cdot | F, G, \kappa), A_F(h) \rangle_{P_{F,G}}. \quad (8)$$

Now, substitute (8) in (5). Then, by the Cauchy-Schwarz inequality, it follows that the information lower bound (5) is given by the variance (the  $L^2(P_{F,G})$ -norm) of  $\tilde{\ell}(\cdot | F, G, \kappa)$ .

We remark here that if  $\kappa_F$  lies in the range of the *information operator*  $I_F \equiv A_F^\top A_F$ , then

$$\tilde{\ell}(\cdot | F, G, \kappa) = A_F (A_F^\top A_F)^{-1} (\kappa_F), \quad (9)$$

where we denote  $I_F^{-1}(\kappa)$  for any element which is mapped to  $\kappa$ .

An estimator  $\mu_n$  of  $\mu$  is asymptotically linear with influence function  $IC(\cdot | F, \kappa) \in L_0^2(P_{F,G})$  if

$$\sqrt{n}(\mu_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC(X_i | F, G, \kappa) + o_P(1/\sqrt{n}).$$

Efficiency theory teaches us that the set of gradients, i.e. the solutions of (6), represents the set of potential influence functions of regular asymptotically linear estimators; in other words, any regular and asymptotically linear estimator has an influence function which is in this set of gradients. In addition, it teaches us that a regular estimator is *asymptotically efficient* at  $P_{F,G}$  if and only if it is asymptotically linear with influence function equal to the canonical gradient  $\tilde{\ell}(\cdot | F, G, \kappa)$ :

$$\sqrt{n}(\mu_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}(X_i | F, G, \kappa) + o_P(1/\sqrt{n}). \quad (10)$$

As a consequence, the canonical gradient  $\tilde{\ell}(\cdot | F, G, \kappa)$  is also called the efficient influence function.

### 3 The identity approach for proving efficiency of the nonparametric maximum likelihood estimator.

In this section we prove a theorem which provides conditions under which an identity, expressing  $\mu_n - \mu$  in terms of  $P_n - P_{F,G}$ , holds. The proof of this theorem is based on two lemmas. The first lemma states an identity in terms of the canonical gradient which follows from the fact that the pathwise derivative along lines of a linear parameter in a convex model has no remainder.

**Lemma 3.1** *Let  $F_n$  be an estimator of  $F$ . Define for  $\alpha \in [0, 1]$ ,  $F_n(\alpha) = (1 - \alpha)F_n + \alpha F$ . For a given  $F$ ,  $\tilde{\ell}(\cdot | F, G, \kappa)$  denotes the efficient influence function for  $F\kappa$  at  $P_{F,G}$ . Suppose that*

- $F\kappa$  is pathwise differentiable at every  $F \in \{F_n(\alpha) : \alpha \in [0, 1]\}$ .
- $F \ll_b F_n$  or  $\tilde{\ell}(\cdot | F_n(\alpha), G, \kappa) \rightarrow \tilde{\ell}(\cdot | F_n, G, \kappa)$  for  $\alpha \rightarrow 0$  w.r.t. the  $L^1(P_{F,G})$ -norm.

Then

$$F_n\kappa - F\kappa = - \int \tilde{\ell}(x | F_n, G, \kappa) dP_{F,G}(x). \quad (11)$$

**Proof.** The left-hand side in (7) equals  $1/\epsilon (F_{h_1}(\epsilon)\kappa - F\kappa)$ . If  $h_1 = (f_1 - f)/f$ , then by the linearity of  $\Psi$  this equals  $F_1\kappa - F\kappa$ . Suppose that  $F \rightarrow P_{F,G}$  is linear. Then

$$A_F(h_1) = \frac{dP_{F_1} - dP_{F,G}}{dP_{F,G}}.$$

Hence, then the right-hand side of (7) equals  $\int \tilde{\ell}(\cdot | F, G, \kappa)(x) dP_{F_1}(x)$ , which, by interchanging the roles of  $F$  and  $F_1$ , yields the following identity: for all  $F \ll_b F_1$

$$F_1\kappa - F\kappa = - \int \tilde{\ell}(x | F_1, G, \kappa)(x) dP_{F,G}(x). \quad (12)$$

We want to apply this result to  $F_1 = F_n$ . Usually  $F_n$  does not dominate  $F$  so that this identity cannot be directly applied (i.e. the condition  $F \ll_b F_n$  fails). However,

because the identity holds for  $F_n(\alpha) \equiv (1 - \alpha)F_n + \alpha F \gg_b F$ ,  $\alpha \in (0, 1]$ , we have that if  $\tilde{\ell}(\cdot | F_n(\alpha), G, \kappa) \rightarrow \tilde{\ell}(\cdot | F_n, G, \kappa)$  for  $\alpha \rightarrow 0$  w.r.t. the  $L^1(P_{F,G})$ -norm (or  $L^2$ ), then this identity holds also at  $F_1 = F_n$ . This proves the lemma.  $\square$

It is important to note that  $\tilde{\ell}(\cdot | F_1, G, \kappa)$  is not necessarily the same as  $\tilde{\ell}(\cdot | F, G, \kappa)$  with  $F$  replaced by  $F_1$  since the efficient influence function at  $F_1$  might be a different functional of  $F_1$  than the efficient influence function at  $F$  is of  $F$ . For example, in the current status data example it appears that the efficient influence function is a different functional of discrete measures (such as the NPMLE  $F_n$ ) than it is of continuous measures (such as the population d.f.  $F$ ) and as a consequence the continuity condition fails at the NPMLE  $F_n$ . If  $F_n$  is a regularized MLE of a smooth  $F$ , then one will have that  $F \ll_b F_n$  so that identity (11) holds without any need to verify the *continuity condition*. The identity (11) can be explicitly verified for the regularized MLE in the current status data model; see our third example. If  $\tilde{\ell}(\cdot | F_n(\alpha), G, \kappa)$  is the same functional in  $F_n(\alpha)$  as  $\tilde{\ell}(\cdot | F_n, G, \kappa)$  is in  $F_n$ , then the continuity condition of lemma 3.1 can be easily verified, and the desired identity (11) follows. This follows from the fact that  $F_n(\alpha) - F_n = \alpha(F - F_n)$  and hence  $F_n(\alpha) \rightarrow F_n$  for  $\alpha \rightarrow 0$  with respect to all norms. In van der Laan (1996, chapter 3) the continuity condition has been verified for almost all data structures  $\Phi(T, C)$  which allow complete observations on  $T$ .

The following lemma provides a sufficient condition for the NPMLE  $F_n$  to solve the *efficient score equation* (13). This condition holds in all examples we encountered.

**Lemma 3.2** *Let  $F_n$  be an estimator of  $F$  satisfying (2) for a dominating measure  $\mu_n$  of  $P_{F_n, G}$ . Assume that  $\kappa_{F_n}$  lies in the range of the information operator  $I_{F_n} : L_0^2(F_n) \rightarrow L_0^2(F_n)$ ,  $h(x) \equiv I_{F_n}^-(\kappa_{F_n})(x)$  is a well defined function for  $F$ -almost all  $x$  with finite supremum norm. Then*

$$0 = \int \tilde{\ell}(x | F_n, G, \kappa)(x) dP_n(x). \quad (13)$$

**Proof.** If  $\kappa_{F_n}$  lies in the range of the *information operator*, then

$$\tilde{\ell}(\cdot | F_n, G, \kappa) = A_{F_n} I_{F_n}^-(\kappa_{F_n})$$

and thus  $\tilde{\ell}(\cdot | F_n, G, \kappa)$  lies in the range of  $A_{F_n}$ . Assume now that formula (9) holds (can be extended to hold) pointwise so that for all  $x$ ,  $\tilde{\ell}(x | F_n, G, \kappa) = A_{F_n}(h)(x)$ , where  $h(x) = I_{F_n}^-(\kappa_{F_n})(x)$  has finite supremum-norm. Then  $\tilde{\ell}(x | F_n, G, \kappa)$  is actually a score corresponding to the one-dimensional submodel  $P_{F_n, h(\epsilon), G}$ , where  $dF_{n, h(\epsilon)}(t) = (1 + \epsilon h(t)) dF_n(t)$ :

$$\tilde{\ell}(X_i | F_n, G, \kappa) = A_{F_n}(h)(X_i) = \frac{d}{d\epsilon} \log \left( \frac{dP_{F_n, h(\epsilon), G}}{d\mu_n} \right) (X_i) \Big|_{\epsilon=0}, \quad i = 1, \dots, n,$$

where  $\mu_n$  can be any dominating measure of  $P_{F_n, G}$ . Since the supremum norm of  $h$  is finite we have that  $F_{n, h(\epsilon)}$  is a submodel for  $\epsilon \in (-\delta, \delta)$  for some  $\delta > 0$ . Moreover, since  $F_n$  maximizes the likelihood (2) it follows that the derivative of

$$\epsilon \rightarrow \int \log \left( \frac{dP_{F_n, h(\epsilon)}}{d\mu_n} \right) (x) dP_n(x)$$

at  $\epsilon = 0$  equals zero. Thus

$$0 = \frac{d}{d\epsilon} \int \log \left( \frac{dP_{F_n, h(\epsilon)}}{d\mu_n} \right) (x) dP_n(x) \Big|_{\epsilon=0} = \int \tilde{\ell}(x | F_n, G, \kappa) dP_n(x). \quad (14)$$

Because the integral is a finite sum, the exchange of integration and differentiation always holds.  $\square$

Combining the identity (11) with the efficient score equation (13) yields the following important identity (15) which expresses  $\mu_n - \mu$  directly in terms of the empirical difference  $(P_n - P_{F,G})\tilde{\ell}(\cdot | F_n, G, \kappa)$ , where we used the notation  $Pf \equiv \int f(x)dP(x)$  for a real valued function  $f$  of  $X$ .

**Theorem 3.1** *Let  $F_n$  be an estimator of  $F$  satisfying (2) for a dominating measure  $\mu_n$  of  $P_{F_n, G}$ . Assume that  $\kappa_{F_n}$  lies in the range of the information operator  $I_{F_n} : L_0^2(F_n) \rightarrow L_0^2(F_n)$ ,  $h(x) \equiv I_{F_n}^-(\kappa_{F_n})(x)$  is a well defined function for  $F$ -almost all  $x$  with finite supremum norm.*

*Suppose that*

- $F\kappa$  is pathwise differentiable at every  $F \in \{F_n(\alpha) : \alpha \in [0, 1]\}$  with canonical gradient  $\tilde{\ell}(\cdot | F, G, \kappa)$ , where  $F_n(\alpha) \equiv (1 - \alpha)F_n + \alpha F$ .
- $F \ll_b F_n$  or  $\tilde{\ell}(\cdot | F_n(\alpha), G, \kappa) \rightarrow \tilde{\ell}(\cdot | F_n, G, \kappa)$  for  $\alpha \rightarrow 0$  w.r.t. the  $L^1(P_{F,G})$ -norm,

*Then*

$$F_n\kappa - F\kappa = \int \tilde{\ell}(x | F_n, G, \kappa)(x) d(P_n - P_{F,G})(x). \quad (15)$$

The identity (15) provides an effective starting point to prove efficiency of the NPMLE  $\mu_n$  (compare (15) with (10)). The following theorem presents the identity approach for proving efficiency of  $\mu_n$  based on the identity (15). It is a direct corollary of the equicontinuity of an empirical process indexed by a Donsker class (see e.g. van der Vaart, Wellner, 1996, page 89). For the definition of a Donsker class we refer to van der Vaart, Wellner (1996, page 80); roughly speaking, a  $P$ -Donsker class is a class  $\mathcal{F}$  of real valued functions of  $X \sim P$  for which the Central Limit Theorem holds for  $\int f(x)(dP_n - dP)(x)$  uniformly in  $f \in \mathcal{F}$ .

**Theorem 3.2** *Let  $\mathcal{G}$  be a given class of real valued functions of  $T$ . Suppose that the conditions of the preceding theorem hold for every  $\kappa \in \mathcal{G}$  so that:*

$$F_n\kappa - F\kappa = \int \tilde{\ell}(\cdot | F_n, G, \kappa) d(P_n - P_{F,G}) \text{ for every } \kappa \in \mathcal{G}.$$

*Assume that the probability that for all  $\kappa \in \mathcal{G}$ ,  $\tilde{\ell}(\cdot | F_n, G, \kappa)$  falls in a  $P_{F,G}$ -Donsker class converges to 1 when the sample size  $n$  converges to infinity. Then*

$$\sup_{\kappa \in \mathcal{G}} |F_n\kappa - F\kappa| = O_P(1/\sqrt{n}). \quad (16)$$

*If also  $\|\tilde{\ell}(\cdot | F_n, G, \kappa) - \tilde{\ell}(\cdot | F, G, \kappa)\|_{P_{F,G}}$  converges to zero with probability tending to 1, then  $F_n\kappa$  is an asymptotically efficient estimator of  $F\kappa$ .*

**Remark about proving the Donsker class condition of theorem 3.2.** Proving the Donsker class condition usually requires the most work, but it is also the work which cannot be avoided (see also van der Vaart, 1992). In van der Vaart, Wellner (1996) several Donsker classes of interest have been identified. In particular, we have that the class of univariate real valued functions with a uniform bound on the variation is a Donsker class and this can be generalized to multivariate functions if we replace the variation norm by a uniform sectional variation norm (Gill, van der Laan, Wellner, 1995, van der Laan, 1996, chapter 1). In our examples we prove the Donsker class conditions by bounding the variation of the terms appearing in  $\tilde{\ell}(\cdot | F_n, G, \kappa)$ . In our fourth example these terms are implicitly defined in terms of the inverse of the information operator. In this case our approach is to prove that  $I_{F_n}^{-1}(\kappa_{F_n})$  is of bounded variation uniformly in  $F_n$  so that formula (9) can be used to show that  $\tilde{\ell}(\cdot | F_n, G, \kappa)$  is Donsker.

### 3.1 An algorithm for construction of asymptotic confidence intervals based on the nonparametric maximum likelihood estimator.

Suppose that the information operator  $I_F : L_0^2(F) \rightarrow L_0^2(F)$  is invertible and onto. Then  $F\kappa$  is pathwise differentiable with efficient influence function  $A_F I_F^{-1}(\kappa_F) \in L_0^2(P_{F,G})$ . If the conditions of theorem 3.2 are satisfied, then the standardized NPMLLE  $\sqrt{n}(F_n\kappa - F\kappa)$  is asymptotically normally distributed with mean zero and variance equal to the variance of the efficient influence function. Suppose that we can compute  $\tilde{\ell}(X_i | F_n, G, \kappa)$  at every observation  $X_i, i = 1, \dots, n$ . The variance of the efficient influence function can then be estimated with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \tilde{\ell}(X_i | F_n, G_n, \kappa)^2,$$

with  $G_n$  being an estimator of  $G$ . Computation of  $\tilde{\ell}(\cdot | F_n, G, \kappa)$  requires the computation of the inverse of the information operator. Therefore it is important to have an algorithm for inverting this information operator and a condition which guarantees that the inverse is a bounded linear operator so that one can actually trust the value  $\hat{\sigma}^2$ .

**Lemma 3.3** *Let  $I_F = A_F^\top A_F : (L_0^2(F), \langle \cdot, \cdot \rangle_F) \rightarrow (L_0^2(F), \langle \cdot, \cdot \rangle_F)$  be the information operator. Assume that for all  $h \in L_0^2(F)$  with  $\|h\|_F > 0$  we have  $\|A_F(h)\|_{P_{F,G}} > 0$ . Then  $I_F$  is 1-1.*

*Assume that there exists a  $\delta > 0$  so that for all  $h \in L_0^2(F)$  we have  $\|A_F(h)\|_{P_{F,G}} \geq \delta \|h\|_F$  for some  $\delta > 0$ . Then  $I_F$  is onto and has bounded inverse with operator norm smaller than or equal to  $1/\delta^2$  and its inverse is given by*

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

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

$$\|A_F^\top A_F(h)\|_F = \|A_F^\top A_F(h)\|_F \|h\|_F$$

$$\begin{aligned}
&\geq \langle A_F^\top A_F(h), h \rangle_F \\
&= \langle A_F(h), A_F(h) \rangle_{P_{F,G}}.
\end{aligned}$$

If  $\|A_F(h)\|_{P_{F,G}} > 0$  for all  $\|h\|_F = 1$ , then  $I_F$  is 1-1 and hence invertible. Moreover, if we have  $\|A_F(h)\|_{P_{F,G}} > \delta\|h\|_F$ , then the inverse is bounded. By using Cauchy sequences and the completeness of a Hilbert space it is easily checked that  $\|A_F(h)\|_{P_{F,G}} > \delta\|h\|_F$  implies that the range of  $A_F$  is closed. This implies that  $I_F$  is onto (BKRW, 1993). It remains to prove "explicit" form of the inverse.

We have that  $I_F = I - (I - I_F)$ . If  $\|A_F(h)\|_{P_{F,G}} > \delta$ , then we have that  $\|I_F(h)\|_F > \delta^2$ . Because  $I_F$  is self-adjoint we have that

$$\sup_{\|h\|_F=1} \|I_F(h)\|_F = \sup_{\|h\|_F=1} |\langle h, I_F(h) \rangle_F| = \sup_{\|h\|_F=1} \langle A_F(h), A_F(h) \rangle_{P_{F,G}} \leq 1.$$

So the self-adjoint operator  $I - I_F$  is also positive and hence its norm is given by:

$$\sup_{\|h\|_F=1} \langle h, (I - I_F)(h) \rangle_F.$$

Because  $\langle h, I_{F,G}(h) \rangle = \langle A_F(h), A_F(h) \rangle \geq \delta^2$  it follows that this norm is smaller than  $1 - \delta^2$ . Consequently, the inverse of  $I_F$  is given by the Neumann series of  $I - I_F$  which converges for all  $h \in L_0^2(F)$ .  $\square$

As a consequence of this invertibility result for the information operator we can determine  $h = I_F^{-1}(\kappa)$  with the following iterative algorithm:

$$h^{k+1} = \kappa - (I - I_F)(h^k). \tag{17}$$

Notice that  $\|A_F(h)\|_{P_{F,G}} > \delta\|h\|_F$  is also a necessary condition for bounded invertibility because if this does not hold then there exist submodels  $P_{F_h(\epsilon),G}$  with arbitrarily small information for estimating  $\epsilon$ .

## 4 Examples

**Example 4.1 (Univariate Censoring Model).** Let  $T_1, \dots, T_n$  be  $n$  i.i.d. copies of a real valued  $T$  with distribution function  $F$ , where  $F$  is completely unknown. Let  $C_1, \dots, C_n$  be  $n$  i.i.d. copies of a real valued  $C$  with distribution function  $G$ , where  $G$  is completely unknown. We will assume that  $T$  and  $C$  are independent. We observe

$$X_i = (Z_i, D_i) = \Phi(T_i, C_i) \equiv (T_i \wedge C_i, I(T_i \leq C_i)) \sim P_{F,G} \equiv (F \times G)\Phi^{-1}.$$

The independence between  $T$  and  $C$  implies that the censoring mechanism satisfies coarsening at random. We are interested in estimating the *survival function*  $\Psi(F) = S(t) \equiv 1 - F(t)$ . It is well known that the nonparametric maximum likelihood estimator is the Kaplan-Meier estimator. This estimator has been extensively analyzed. For an overview of work done in this field we refer to Andersen, Borgan, Gill and Keiding (1993).

In this model  $\tilde{\ell}(X | F, G, t)$  can be explicitly written down so that we can explicitly verify identity (15), which we will do now. Define

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

Let  $H = 1 - G$ ,  $N = E_{F,G}N_n$  and  $Y = E_{F,G}Y_n$ . It is well known (e.g. Wellner, 1982, Gill, 1993) that if  $H(t) > 0$ , then the efficient influence function for estimating  $S(t)$  is given by:

$$\tilde{\ell}(X | F, G, t)(z, d) = -S(t) \int_0^t \frac{I(z \in dv, d = 1) - I(z \geq v)d\Lambda(v)}{S(v)H(v-)}.$$

Consequently, the efficient score-equation for the NPMLE  $S_n$  is given by:

$$P_n \tilde{\ell}(\cdot | F_n, G, t) = S_n(t) \int_0^t \frac{dN_n(v) - Y_n d\Lambda_n(v)}{S_n(v)H(v-)}.$$

This is zero if and only if  $d\Lambda_n(v) = dN_n(v)/Y_n(v)$  which implies that  $S_n(t) = \prod_{(0,t]}(1 - dN_n/Y_n)$ , where  $\prod_{(0,t]}$  is a *product integral* and stands for a limit of approximating finite products over partitions of  $(0, t]$  as the partitions become finer. So the efficient-score equation is uniquely solved by the Kaplan-Meier estimator. This verifies the efficient score-equation (13).

It remains to verify the identity (11), i.e.  $S_n(t) - S(t) = -P_{F,G} \tilde{\ell}(\cdot | F_n, G, t)$ , which is here given by:

$$S_n(t) - S(t) = S_n(t) \int_0^t \frac{dN(v) - Y d\Lambda_n(v)}{S_n(v)H(v-)}. \quad (18)$$

We know that  $dN = Y d\Lambda$ ,  $Y = S_- H_-$ . So  $dN - Y d\Lambda_n = S_- H_- (d\Lambda - d\Lambda_n)$ , where  $H_-$  cancels with the denominator. Therefore (18) is equivalent to:

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

where we used  $S_n(t)/S_n(v) = \prod_{(v,t]} (1 - dN_n/Y_n)$ . This is the well known *Duhamel equation* for the univariate product integral (Gill and Johansen, 1990). This proves the identity (15) for the NPMLE in the univariate censoring model:

$$S_n(t) - S(t) = (P_n - P_{F,G}) \tilde{\ell}(\cdot | F_n, G, t). \quad (19)$$

To finish the efficiency proof it remains to verify the  $P$ -Donsker class and  $P$ -consistency condition of theorem 3.2.

Notice that  $\tilde{\ell}(X | F_n, G, t)$  is a sum of two monotone functions and both parts are bounded by  $c/H(t)$  for a constant  $c$ . The class of bounded monotone functions is a uniform-Donsker class (see van der Vaart and Wellner, 1996). Thus theorem 3.2 applied to the parameter  $H(t)S(t)$  provides us with the following whole line result:

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

It follows trivially that  $\sup_{t \in [0, \infty]} \|H(\tilde{\ell}(\cdot | F_n, G, t) - \tilde{\ell}(\cdot | F, G, t))\|_{P_{F,G}} \rightarrow 0$  in probability. This provides us with supremum norm (on the whole line) efficiency of  $HS_n$  as an estimator of  $HS$  under no assumptions at all. In particular, this implies efficiency of  $S_n$  on any rectangle  $[0, \tau]$  with  $H(\tau) > 0$ . Some deeper whole-line and bootstrap results based on the identity (19) can be found in Gill (1993).

Now, we will give an example of a missing data model with no complete observations. Here the identity (1) holds at  $F_n(\alpha)$  for any  $\alpha$ , but not at the discrete  $F_n$  itself.

**Example 4.2 (Current Status Data).** Let  $T_1, \dots, T_n$  be  $n$  i.i.d. copies of a real valued  $T$  with distribution function  $F$ , where  $F$  is completely unknown. Let  $C_1, \dots, C_n$  be  $n$  i.i.d. copies of a real valued  $C$  with distribution function  $G$  which is completely unknown.  $T$  and  $C$  are independent. We observe

$$X = (C, \Delta) \equiv (C, I(T \leq C)) \sim P_{F,G},$$

where

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

We are concerned with estimating  $\mu = FR = \int R(t)dF(t)$ . So if we set  $R(t) = t^k$ , then we are estimating the  $k$ 'th moment of  $F$ .

The NPMLE for  $F$  has been analyzed in Groeneboom (1991) and in Groeneboom and Wellner (1992). Alternative shorter proofs for asymptotic normality are given by van der Geer (1994), which uses the identity (11) for the sequence  $F_n(\alpha)$ , and Huang and Wellner (1994).

The score operator for  $F$  is given by:

$$A_F : L_0^2(F) \rightarrow L_0^2(P_{F,G}) : A_F(h)(Y) = E_F(h(T) | X),$$

where  $X = (C, \Delta)$ . Its adjoint is given by  $A_F^\top(v)(T) = E_G(v(X) | T)$ . So the information operator  $I_F : L_0^2(F) \rightarrow L_0^2(F)$  is given by:

$$I_F(h)(t) = \int_t^\infty \frac{\int_0^c h(t)dF(t)}{F(c)} dG(c) + \int_0^t \frac{\int_c^\infty h(t)dF(t)}{1 - F(c)} dG(c). \quad (20)$$

Firstly, we will find the canonical gradient of the pathwise derivative of  $\mu$  at  $F$  and the NPMLE  $F_n$ . Consider the equation  $I_F(h) = R - \mu$  in  $h \in L^2(F)$ . Taking derivatives with respect to  $G$  on both sides and using  $\int_t^\infty h dF = -\int_0^t h dF$  yields the equation

$$\int_0^t h dF = -\frac{dR(t)}{dG(t)} F(t)(1 - F(t)), \quad (21)$$

assuming that  $R \ll G$ . Define  $r(x) \equiv dR/dG(x)$ . If  $r \ll F$ , then (21) has a unique solution  $h$  and the efficient influence function is given by:

$$\tilde{\ell}(X | F, r) = A_F(h)(c, \Delta) = -r(c)(1 - F(c))\Delta + r(c)F(c)(1 - \Delta). \quad (22)$$

Let now  $F = F_n$  be an NPMLE of  $F$ . There is no uniquely defined NPMLE in this problem, but following the convention in Groeneboom and Wellner (1992) we can choose a discrete NPMLE  $F_n$ . Denote its support with  $\{x_1, x_2, \dots, x_k\}$ . Solving (21) in  $L^2(F_n)$  means that we only need equality at the support points  $x_i, i = 1, \dots, k$ . So we need:

$$\int_{x_i}^{\infty} \frac{\int_0^c h(t) dF_n(t)}{F_n(c)} dG(c) + \int_0^{x_i} \frac{\int_c^{\infty} h(t) dF_n(t)}{1 - F_n(c)} dG(c) = R(x_i) - \mu(F, R).$$

Use the fact that  $\int_c^{\infty} h(t) dF_n(t) = -\int_0^c h(t) dF_n(t)$ . Now, it follows that

$$R(x_{i+1}) - R(x_i) = - \int_{(x_i, x_{i+1}]} \left( \frac{\int_0^c h(t) dF_n(t)}{F_n(c)} + \frac{\int_0^c h(t) dF_n(t)}{1 - F_n(c)} \right) dG(c).$$

However,  $\int_0^c h dF_n$  and  $F_n(c)$  are equal to  $\int_0^{x_i} h dF_n$  and  $F_n(x_i)$ , respectively, for  $c \in (x_i, x_{i+1})$ . Thus

$$\int_0^{x_i} h dF_n = - \frac{R(x_{i+1}) - R(x_i)}{G(x_{i+1}) - G(x_i)} F_n(x_i)(1 - F_n(x_i)), \quad i = 1, \dots, k$$

which has a solution  $h \in L_0^2(F_n)$ . The efficient influence function is given by  $A_{F_n}(h)$ .

Define the step function  $r_n(x) \equiv (R(x_{i+1}) - R(x_i)) / (G(x_{i+1}) - G(x_i))$  if  $x \in (x_i, x_{i+1}]$ . We have shown that the efficient influence function at  $F_n$ , i.e.  $A_{F_n}(h)$ , is given by:

$$\tilde{\ell}(X | F_n, r) = -r_n(c)(1 - F_n(c))\Delta + r_n(c)F_n(c)(1 - \Delta).$$

Recall that  $\tilde{\ell}(X | F_n, r)$  is defined as the efficient influence function at  $P_{F_n, G}$  for the parameter  $\mu = \int R dF$ , which happens here to involve  $r_n$  which is a function of  $F_n, R$  and  $G$ . So the dependence of  $\tilde{\ell}(X | F_n, r)$  on  $F_n$  is not only through  $(1 - F_n)$  and  $F_n$ , but it also enters in  $r_n$  while this is not the case for the efficient influence function at a continuous  $F$ . So here we have an example where the continuity condition in lemma 3.1 for the efficient influence function for sequences  $F_n(\alpha)$  fails. In fact, we have

$$\begin{aligned} P_{F, G} \tilde{\ell}(\cdot | F_n, r) &= \int r_n(F - F_n) dG = \int r(F - F_n) dG + \int (r_n - r)(F - F_n) dG \\ &= \mu - \mu_n + \int (r_n - r)(F - F_n) dG. \end{aligned} \quad (23)$$

So indeed the exact identity (11) does not hold, due to the fact that  $\tilde{\ell}(X | F_n, r)$  involves  $r_n$  instead of  $r$  itself.

The conditions of lemma 3.2 follow directly from our representation of the efficient influence function  $\tilde{\ell}(X | F_n, r)$  as  $A_{F_n}(h)$ . We will write down the efficient score equation and determine what it teaches us about the NPMLE  $F_n$ . If  $P_n$  is the empirical distribution of the data  $(C_i, \Delta_i)$  and  $G_n(\cdot) = P_n(\cdot, 0) + P_n(\cdot, 1)$  is the empirical distribution of the  $C_i$ 's,

then the histogram versions of  $P_n(\cdot, 1)$  and  $G_n$  on the support points  $x_i$  of  $F_n$  are defined by the jumps:

$$\begin{aligned} N_n(\Delta x_i) &= P_n((x_i, x_{i+1}], 1) = \frac{1}{n} \#\{j : c_j \in (x_i, x_{i+1}], \Delta_j = 1\} \\ Y_n(\Delta x_i) &= G_n((x_i, x_{i+1}]) = \frac{1}{n} \#\{j : c_j \in (x_i, x_{i+1}], \Delta_j = 0\}. \end{aligned}$$

The efficient score equation is given by

$$0 = P_n \tilde{\ell}(\cdot | F_n, r) = \int -r_n(1 - F_n) dN_n + \int r_n F_n dY_n.$$

This is solved by:

$$F_n(x_i) = \frac{N_n(\Delta x_i)}{(N_n + Y_n)(\Delta x_i)} = \frac{P_n((x_i, x_{i+1}], 1)}{G_n((x_i, x_{i+1}])}, \quad i = 1, \dots, k. \quad (24)$$

So  $F_n$  is determined up to its support points, where we have some knowledge about the support because we know that the support has to be chosen such that  $F_n$  is monotone. The support points are chosen such that the likelihood is maximized and that information is not contained in the score equations. However, we note that  $F_n$  is just a fraction of two histogram density estimators and that suggests that we could replace these by kernel density estimators. We will do this in the next example.

Combining the efficient score equation with the identity (23) yields

$$\mu_n - \mu = \int \tilde{\ell}(x | F_n, r) d(P_n - P_{F,G})(x) + \int (r_n - r)(F_n - F) dG.$$

An efficiency proof would now involve an additional big step, namely we first need to show that  $\int (r_n - r)(F_n - F) dG = o_P(1/\sqrt{n})$ . This involves knowledge about the support points of  $F_n$  (to control  $r_n - r$ ) and it requires a rate result for  $F_n$ . A nice and elegant efficiency proof is given in Huang and Wellner (1994).

Suppose that the MLE  $F_n$  is smooth so that  $F_n \gg F$ . Then the identity (11) holds and it is thus not necessary to verify the continuity condition of lemma 3.1. This is the case if  $F_n$  is a regularized MLE which is obtained by maximizing  $\int \log(p_{F,G}) d\tilde{P}_n$ , where  $\tilde{P}_n$  is a smoothed empirical distribution function. A regularized MLE  $F_n$  will solve the smoothed efficient score equation  $\tilde{P}_n \tilde{\ell}(\cdot | F_n, G, \kappa) = 0$  which leads to the identity:

$$F_n \kappa - F \kappa = (\tilde{P}_n - P_{F,G}) \tilde{\ell}(\cdot | F_n, G, \kappa).$$

The regularized MLE in the current status data model has a strong application in the following model:

**Example 4.3 (doubly-censored current status data in an aids model).** In HIV-partner studies one encounters the following kind of problem as modeled in Jewell, Malani and Vittinghoff (1994). Consider a partnership. Let  $I$  be the chronological time of HIV-infection of the initially infected partner (the index case) and let  $J$  denote the infection time of the partner who is infected second. We are concerned with nonparametric-estimation of the distribution  $G$  of  $T = J - I$ , the time between infection dates of the two partners.

In this AIDS-model we observe two times  $A$  and  $B$  of which we know that  $A < I < B$ . Given  $A, B$  we assume that  $I$  follows a uniform distribution on  $(A, B)$ . Moreover, we still observe  $\Delta \equiv I(J \leq B)$ , where  $\Delta = 1$  if both partners are infected at time  $B$  and  $\Delta = 0$  if only one partner is infected at time  $B$ . Let  $X = (A, B, \Delta)$  represent the data on the partners.

If we define  $C = B - A$ , then by using that  $T$  is independent of  $I$ , given  $A, B$ , and  $T$  is independent of  $A, B$  we obtain:

$$P(\Delta = 1 | A, B) = P(\Delta = 1 | C) = F(C) \equiv \int_0^C \frac{C-t}{C} dG(t) = G(C) - \frac{1}{C} \int_0^C tdG(t). \quad (25)$$

Therefore we can reduce the data  $X$  to  $(C, \Delta)$  without loss of information for estimation of  $G$ . Let  $h$  be the unknown density of  $C = B - A$  on  $[0, \tau]$ . Then the density of the data  $(C = B - A, \Delta)$  is

$$p(c, \Delta) = F_G(c)h(c)^\Delta ((1 - F_G(c))h(c))^{1-\Delta}$$

and hence this is just a submodel of the current status data model as described in the second example ( $F_G$  is playing the role of  $F$  and  $h$  is playing the role of the censoring density  $g$ ). Here one has to notice that  $F_G$  is a distribution function.

In van der Laan, Bickel, Jewell (1997) it is shown that (in spite of the fact that the ranges of the score operators for the two models are different) the tangent space in the AIDS-model, where  $F_G$  is restricted, equals the tangent space of the current status data model assuming nothing about  $F$ . This implies that the NPMLE  $F_n$  for the nonparametric current status data model provides us with efficient estimators of smooth functionals of  $F_G$ . In other words, the particular knowledge that  $F$  can be written as  $F_G$  for some  $G$ , does not help in estimating smooth functionals of  $F$ . Therefore it is appropriate to first consider estimation of  $F_G$  as if we are in the nonparametric current status data model and then recover information about  $G$  from this estimate. In order to convert an estimate of  $F_G$  to an estimate of  $G$  we can use the relation:

$$G(c) = F_G(c) + cf_G(c), \quad (26)$$

where  $f_G$  is the derivative of  $F_G$ . Here it should be noted that  $F_G$  is differentiable for all  $G$  (so also if  $G$  is discrete).

As shown in the second example, given its support  $x_1, \dots, x_n$ , the NPMLE  $F_n$  of  $F_G$  in the nonparametric current status data model is given by

$$F_n(x) = \frac{P_n((x_i, x_{i+1}], 1)}{P_n((x_i, x_{i+1}])}, \text{ if } x \in (x_i, x_{i+1}], \quad (27)$$

where  $P_n$  is the empirical distribution of  $P_{F,G}$ . So  $F_n$  is just a fraction of histogram estimates, one based on the  $C_i$ 's with  $\Delta_i = 1$  and one based on all the  $C_i$ 's. By the relation (26) we need a smoothed version of  $F_n$  so that the derivative  $f_n$  of  $F_n$  estimates  $f$  consistently. Therefore we replace the histogram estimates in numerator and denominator in (27) by kernel density estimators  $p_n(\cdot, 1)$  of  $p_F(\cdot, 1) = F(\cdot)h(\cdot)$  and  $h_n(\cdot) = p_n(\cdot, 1) + p_n(\cdot, 0)$  of  $h$ , respectively, both with appropriate bandwidth so that the derivatives of these kernel density estimators are consistent. Then we obtain

$$\tilde{F}_n(c) \equiv \frac{p_n(c, 1)}{h_n(c)}. \quad (28)$$

In van der Laan, Bickel, Jewell (1997) it is shown that (28) is the regularized MLE corresponding with the density estimator  $p_n$ . The estimator (28) provides us via (26) with an estimator of  $G$  itself.

Let  $\mu = \int R dG$  be the parameter of interest for some function  $R$ . If  $R_1(x) = R(x) - xr(x) \in L^2(F_G)$  and  $\lim_{x \rightarrow \infty} R(x)/x = 0$ . Then we have:

$$\begin{aligned}
\mu &\equiv \int R dG = \int \frac{R(t)}{t} t dG(t) \\
&= \int_0^\infty \left( \int_t^\infty \frac{R(x) - xr(x)}{x^2} dx \right) t dG(t) \\
&= \int R_1(x) \left( \frac{1}{x^2} \int_0^x t dG(t) \right) dx \\
&= \int R_1(x) dF_G(x).
\end{aligned} \tag{29}$$

At the second line we applied Fubini's theorem using  $\lim_{x \rightarrow \infty} R(x)/x = 0$  so that  $\int_t^\infty R_1(x)/x^2 dx = R(t)/t$ . Thus an estimator of  $\int R_1 dF$  yields an estimator of  $\int R dG$ . The efficient influence function  $\tilde{\ell}(X | \tilde{F}_n, h, R_1)$  at  $(\tilde{F}_n, h)$  is given by (22).

The identity (11) tells us now that

$$\tilde{F}_n R_1 - F R_1 = -P_{F,G} \tilde{\ell}(\cdot | \tilde{F}_n, h, R_1). \tag{30}$$

This identity can be trivially *explicitly* verified by substitution of (28).

Using  $h_n(c) = p_n(c, 1) + p_n(c, 0)$  and (28) it also follows that the smoothed efficient score equation holds:

$$\tilde{P}_n \tilde{\ell}(\cdot | \tilde{F}_n, h, R_1) = 0 \text{ for all differentiable } R_1 \text{ on } [0, \tau].$$

This provides us with the identity:

$$\tilde{F}_n R - F R = \left( \tilde{P}_n - P_{F,G} \right) \tilde{\ell}(\cdot | \tilde{F}_n, h, R_1). \tag{31}$$

The efficiency proof for  $\mu(\tilde{F}_n)$  can now be straightforwardly completed by using empirical process theory and choosing a bandwidth such that  $\tilde{P}_n - P_n = o_P(1/\sqrt{n})$ . This and a complete analysis of the estimation problem is carried out in van der Laan, Bickel and Jewell (1997).

Finally, we provide an example in which the NPMLLE is implicit.

**Example 4.4 (A mixture of right-censored and current status data).** It is quite common in studies that, by lack of resources, one selects a subsample whose subjects are followed up while the others are only monitored at time of entrance in the study. In such studies one might have subjects for which one has current status data and the subjects of the selected sample are followed up till the survival time of interest is observed or till the end of the experiment. Let  $T \sim F$  be the survival time of interest. Let  $C$  be the monitoring time or right-censoring time which is always observed and let  $\xi \in \{0, 1\}$  be the indicator for the selected sample. We assume that  $(C, \xi)$  is independent of  $T$  and we denote the subdistributions of  $(C, 0)$  and  $(C, 1)$  with  $G_0$  and  $G_1$ , respectively. We observe

$(C, \xi)$ , and if  $\xi = 1$  (i.e. the subjects belong to the selected sample), we observe  $(T \wedge C, \Delta)$  and if  $\xi = 0$ , we observe  $\Delta$ . The distribution of the data  $X$  is given by:

$$\begin{aligned} P(C \in dc, \xi = 1, T \in dt, \Delta = 1) &= G_1(dc)F(dt) \\ P(C \in dc, \xi = 1, \Delta = 0) &= G_1(dc)S(c) \\ P(C \in dc, \xi = 0, \Delta = 0) &= G_0(dc)S(c) \\ P(C \in dc, \xi = 0, \Delta = 1) &= G_0(dc)F(c). \end{aligned}$$

Without any loss of information for  $F$  we can pool together the 2 types of observations corresponding with  $\Delta = 0$ . We define a new  $\delta$  to indicate the remaining three types of observations: let  $\delta = 1$  if  $\xi = 1, \Delta = 1$ , let  $\delta = 2$  if  $\Delta = 0$  and let  $\delta = 3$  if  $\xi = 0, \Delta = 1$ . Let  $G(c) = G_0(c) + G_1(c)$ . Then the distribution of the data is given by:

$$I(\delta = 1)F(dt)G_1(dc) + I(\delta = 2)S(c)G(dc) + I(\delta = 3)F(c)G_0(dc).$$

We are concerned with estimation of  $F(t)$ .

Each observation implies a region for  $T$ . A NPMLE of  $F$  should put mass on at least one point in each  $T$ -region. Let  $\{x_1, \dots, x_k\}$  be the set obtained as follows: start with all uncensored  $T_i$ , now for each  $T$ -region corresponding with a censored observation which does not contain an uncensored observation add one chosen point in this region. Let  $F_n$  be the MLE over all  $F$  with support given by  $\{x_1, \dots, x_k\}$ . This NPMLE of  $F$  can be determined with the EM-algorithm by iterating the self-consistency equation

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n P_{F_n}(T \leq t \mid Y_i)$$

with an initial estimator with support  $\{x_1, \dots, x_k\}$  (see van der Laan, 1996, chapter 3, for existence). In this example we will show how one can prove uniform consistency and efficiency of  $F_n$  with the identity approach under the following assumptions. Firstly, let  $F$  have compact support  $[0, \tau]$ ,  $S(\tau-) > 0$ , and  $\bar{G}_1 > \delta > 0$  on the support of  $F$ ; this can be arranged by artificially right-censoring all uncensored observations larger than  $\tau$  at  $\tau$ , where  $\bar{G}_1(\tau) > 0$ . Secondly, we will assume that  $G_0 = 0$  on  $[0, \delta)$  for some  $\delta > 0$  and  $F(\delta) > 0$ .

The score operator  $A_F : L_0^2(F) \rightarrow L_0^2(P_{F,G})$  is given by:

$$A_F(h) = h(T)I(\Delta = 1) + \frac{\int_C^\infty hdF}{S(C)}I(\delta = 2) + \frac{\int_0^C hdF}{F(C)}I(\delta = 3).$$

It is straightforward to determine the adjoint  $A_F^\top : L_0^2(P_{F,G}) \rightarrow L_0^2(F)$  of  $A_F$  which is given by:

$$A_F^\top(V)(T) = \int_T^\infty V(c, T, 1)G_1(dc) + \int_0^T V(c, 2)G(dc) + \int_T^\infty V(c, 3)G_0(dc).$$

So the information operator  $I_F = A_F^\top A_F$  is given by:

$$I_F(h)(T) = h(T)\bar{G}_1(T) + \int_0^T \frac{\int_c^\infty hdF}{S(c)}G(dc) + \int_T^\infty \frac{\int_0^c hdF}{F(c)}G_0(dc).$$

Since  $\bar{G}_1 > \delta > 0$  on the support of  $F$  we have  $\|A_F(h)\|_{P_{F,G}}^2 \geq \delta \|h\|_F$ . By lemma (3.3) this implies that  $I_F : L_0^2(F) \rightarrow L_0^2(F)$  is one-to-one, onto, with bounded inverse. Consider the equation

$$I_{F_n}(h)(T) = \kappa(T) - F_n \kappa.$$

We know that this equation has a unique  $L^2(F_n)$ -solution. For proving that the efficient score equation holds we need that this equation has a pointwise well defined solution  $h$  such that  $I_{F_n}(h) = \kappa - F_n \kappa$  pointwise and that  $h$  is uniformly bounded. We can show this as follows.

The equation  $I_{F_n}(h)(t) = \kappa_n(t) \equiv \kappa(t) - F_n \kappa$  is equivalent to the following equation:

$$h(t) = \frac{1}{\bar{G}_1(t)} \left( \kappa_n(t) + \int_0^t \frac{\int_c^\infty h dF_n}{S(c)} G(dc) + \int_t^\infty \frac{\int_0^c h dF_n}{F_n(c)} G_0(dc) \right). \quad (32)$$

For the moment denote the right-hand side by  $C_{F_n}(h, \kappa_n)(t)$ : i.e. we consider the equation  $h(t) = C_{F_n}(h, \kappa_n)(t)$ . We know that there exists an  $h' \in L^2(F_n)$ , which is unique in  $L^2(F_n)$ , with  $\|I_{F_n}(h') - \kappa_n\|_{F_n} = 0$ : i.e.  $\|h' - C_{F_n}(h', \kappa_n)\|_{F_n} = 0$ . Notice that if  $\|h - g\|_{F_n} = 0$ , then for each  $t$ ,  $C_{F_n}(h - g, \kappa_n)(t) = 0$ . So even if  $h'$  is only uniquely determined in  $L^2(F_n)$ , then  $C_{F_n}(h', \kappa_n)(t)$  is uniquely determined for each  $t$ . Now, we can define  $h(t) \equiv C_{F_n}(h', \kappa)(t)$ . Then  $\|h - h'\|_{F_n} = \|C_{F_n}(h', \kappa_n) - h'\|_{F_n} = 0$ . So in this way we have found a solution  $h$  (a version of  $h'$ ) of (32) which holds for each  $t$  instead of only in the  $L^2(F_n)$  sense.

Now, we want to show that  $h$  has a finite supremum norm. We have by  $\bar{G}_1 > \delta > 0$  and the Cauchy-Schwarz-inequality that

$$|h(t)| \leq \frac{1}{\delta} \left( |\kappa_n(t)| + \|h\|_F \int_0^t \frac{G(dc)}{\sqrt{S_n(c)}} + \|h\|_F \int_t^\infty \frac{G_0(dc)}{\sqrt{F_n(c)}} \right).$$

By the fact that  $F_n$  solves the self-consistency equation, we have that  $F_n(A)$  is larger than the fraction of uncensored observations in  $A$ . This empirical fraction converges uniformly to  $\int_A \bar{G}_1(t) dF(t) \geq \delta F(A)$ , by our assumption that  $\bar{G}_1 > \delta > 0$ . By our assumptions this proves that  $F_n(c)$  and  $S_n(c)$  are uniformly bounded away from zero on the support of  $G_0$  and  $G$ , respectively. This proves that  $\|h\|_\infty \leq M \|\kappa_n\|_\infty$  for some  $M < \infty$ . Similarly, we prove that  $\|h\|_v \leq M \|\kappa_n\|_v$ , where  $\|\cdot\|_v$  denotes the variation norm.

We are now ready to apply the identity approach for proving consistency and efficiency. Firstly, we have that  $\tilde{\ell}(X | F_n, G, t) = A_{F_n}(h_n)$  with  $h_n = I_{F_n}^{-1}(\kappa_n)$ , where we showed that  $h_n$  has finite supremum-norm. Consequently,  $\tilde{\ell}(X | F_n, G, t)$  is the score of a one dimensional submodel  $P_{F_n, \epsilon}$ , with  $dF_{n, \epsilon} = (1 + \epsilon h_n) dF_n$ , and thus

$$P_n \tilde{\ell}(\cdot | F_n, G, t) = 0.$$

Above we showed that the information operator  $I_F$  has a bounded inverse w.r.t. the supremum-norm, where the norm of  $I_F^{-1}$  does not depend on  $F$ , as long as  $F$  satisfies our assumptions. Therefore the continuity condition in lemma 3.1 can now be straightforwardly verified. This proves that

$$F_n(t) - F(t) = (P_n - P_{F,G}) \tilde{\ell}(\cdot | F_n, G, t).$$

Since the score operator (of any censored data model) maps monotone functions to monotone functions and every function of bounded variation is the difference of two monotone functions, it follows that  $\|A_F(h)I(\delta = i)\|_v \leq M\|h\|_v$  for some  $M < \infty$ . Consequently,  $\tilde{\ell}(X | F_n, G, t)$  falls with probability tending to 1 in the class of functions with variation smaller than  $M$ , for some fixed  $M < \infty$ . The class of functions with a uniform bound on the variation is a Donsker class. This proves that

$$\sup_{t \in [0, \tau]} |F_n(t) - F(t)| = O_P(1/\sqrt{n}).$$

Using this uniform-consistency result and the supremum-norm invertibility of the information operator it is easily shown that

$$\|\tilde{\ell}(\cdot | F_n, G, t) - \tilde{\ell}(\cdot | F, G, t)\|_{P_{F,G}} \rightarrow 0 \text{ in probability.}$$

This proves that  $F_n(t)$  is asymptotically efficient.

## Acknowledgement.

I thank the referee for his/her constructive and helpful comments.

## References

- P.K. Andersen, Ø. Borgan, R.D. Gill and N. Keiding (1993), *Statistical models based on counting processes*, Springer, New York.
- P.J. Bickel, A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.
- D. Bitouzé (1995), *Estimation de fonctionnelles d'une densité à partir d'observations directes ou censurées*, Université de Paris-Sud, Janvier 1995.
- S. van der Geer (1994), *Asymptotic normality in mixture models*, preprint University of Leiden, the Netherlands.
- R.D. Gill (1983), Large sample behaviour of the product-limit estimator on the whole line, *Ann. Statist.* **11** 49–58.
- R.D. Gill and A.W. van der Vaart (1993), Non- and semi-parametric maximum likelihood estimators and the von Mises method - II, *Scand. J. Statist.* **20**, 271–288.
- R.D. Gill (1993), *Lectures on survival analysis*, In: D. Bakry, R.D. Gill and S. Molchanov, École d'Été de Probabilités de Saint Flour XXII–1992, ed. P. Bernard; Springer Lecture Notes in Mathematics.
- R.D. Gill and S. Johansen (1990), A survey of product integration with a view towards application in survival analysis, *Ann. Statist.* **18**, 1501–1555.
- R.D. Gill, M.J. van der Laan and J.M. Robins (1997), *Coarsening at Random*, to appear in the *Proceedings of the Seattle Symposium in Biostatistics*, 1995.
- R.D. Gill, M.J. van der Laan and J.A. Wellner (1995), Inefficient estimators of the bivariate survival function for three models, *Ann. Inst. Henri Poincaré* **31**, 545–597.
- P. Groeneboom (1991), *Nonparametric maximum likelihood estimators for interval censoring and deconvolution*, Technical report nr. **378**, Department of Statistics Stanford University, California.

- P. Groeneboom and J.A. Wellner (1992), *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser verlag.
- D.F. Heitjan and D.B. Rubin (1991), Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.
- M. Jacobsen and N. Keiding (1994), *Coarsening at random in general sample spaces and random censoring in continuous time*, preprint, Institute of Math. Stat. and Dept. of Biostatistics, University of Copenhagen.
- N.P. Jewell, H.M. Malani and E. Vittinghoff (1994), Nonparametric estimation for a form of doubly censored data with application to two problems in aids, *JASA* **89**, 7–18.
- J. Kiefer and J. Wolfowitz (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Statist.* **27**, 887–906.
- M.J. van der Laan (1995), An identity for the nonparametric maximum likelihood estimator in missing data and biased sampling models, *Bernoulli* **1(4)**, 335–341.
- M.J. van der Laan (1996), *Efficient and inefficient estimation in semiparametric models*, CWI-tract # 114, Centre for Mathematics and Computer Science, Amsterdam.
- M.J. van der Laan (1994), *Proving efficiency of NPML and identities*, technical report #44, Group of Biostatistics, Berkeley.
- M.J. van der Laan, P.J. Bickel and N.P. Jewell (1997), *Singly and doubly censored current status data: Estimation, Asymptotics and Regression*, to appear in *Scand. J. Stat.*.
- A.W. van der Vaart (1988), *Statistical estimation in large parameter spaces*, CWI Tract, Centre for Mathematics and Computer Science, Amsterdam.
- A.W. van der Vaart and J.A. Wellner (1996), *Weak convergence and empirical processes*. Springer Verlag.
- A.W. van der Vaart (1992), Efficiency of infinite dimensional M-estimators, Technical Report, Free University Amsterdam, the Netherlands.
- J.A. Wellner (1982), Asymptotic optimality of the product limit estimator, *Ann. Statist.* **10**, 595–602.
- J. Huang and J.A. Wellner (1994), *Asymptotic normality of the NPML of the mean for interval censored data, case I*, submitted to *Statistica Neerlandica*.