

Communication

Identity for NPMLE in Missing and Biased Sampling Models.

Mark J. van der Laan

Department of Statistics, Division of Biostatistics

University of California, Berkeley

CA 94703

laan@stat.berkeley.edu

May 15, 2001

Abstract

We derive an identity for NPMLE in non-parametric missing data models and biased sampling models, which almost says that the NPMLE is efficient. Application of empirical process theory to the identity provides us with a straightforward consistency and efficiency proof. The identity is illustrated with the random truncation model.

1 Introduction.

Let X_1, \dots, 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 μ_n of P_{F_n, G_n} we have:

$$P_{F_n, G_n} = \arg \max_{\{P_{F,G}: P_{F,G} \ll \mu_n\}} \int \log \left(\frac{dP_{F,G}}{d\mu_n} \right) dP_n(x), \quad (1)$$

where P_n denotes the empirical distribution of X_1, \dots, X_n . In particular, (1) 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 Ψ . 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 μ $dP_{F,G}/d\mu = p_F p_G$ 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 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 this paper we derive a crucial identity (see (9)) for missing and biased sampling models, by exploiting their specific structure (see (8)), and show how the combination of this identity with the efficient score equation (see (10)) often leads to a powerful identity (see (11)) 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 (1993a). 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 $\Psi(F_n)$. Finally we will illustrate our findings with the random truncation model. We remark here that the results are trivially extended to any parametrization $P_{\theta,\eta}$ having the same structure (8); just replace F by θ and G by η in the sequel.

2 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

$$\frac{d}{d\epsilon} \log \left(\frac{dP_{f_{h_1}(\epsilon), g_{h_2}(\epsilon)}}{dP_{F,G}} \right) \Big|_{\epsilon=0} = A_{F,G}(h_1) + B_{F,G}(h_2), \quad (2)$$

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

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

The equalities in (2) and the limits in $d/d\epsilon$ are assumed to hold in $L^2(P_{F,G})$. In view of linearity of Ψ 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 ϵ is now given by:

$$\left(\frac{\frac{d}{d\epsilon} \Psi(F_{h_1}(\epsilon)) \Big|_{\epsilon=0}}{\|A_{F,G}(h_1) + B_{F,G}(h_2)\|_{P_{F,G}}} \right)^2 = \left(\frac{\Psi(\int h_1 dF)}{\|A_{F,G}(h_1) + B_{F,G}(h_2)\|_{P_{F,G}}} \right)^2. \quad (3)$$

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 (3) 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 $B_{F,G}(h_2) = -\Pi(A_{F,G}(h_1) | T_2(P_{F,G}))$, where $\Pi(\cdot | T_2(P_{F,G}))$ 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:

$$\sup_{h_1 \in L_0^2(F)} \left(\frac{\Psi(\int h_1 dF)}{\|A_{F,G}^*(h_1)\|_{P_{F,G}}} \right)^2, \quad (4)$$

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

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

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

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

Let $A_{F,G}^{*\top} : L_0^2(P_{F,G}) \rightarrow L_0^2(F)$ be the *adjoint* of $A_{F,G}^*$. 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

$$\Psi(F_1) - \Psi(F) = \langle I_{F,G} I_{F,G}^-(\kappa(F, \Psi)), h_1 \rangle_F = \langle \ell^*(F, G, \Psi), A_{F,G}^*(h_1) \rangle_{P_{F,G}}, \quad (5)$$

where

$$\ell^*(F, G, \Psi) = A_{F,G}^* \left(A_{F,G}^{*\top} A_{F,G}^* \right)^- (\kappa(F, \Psi)). \quad (6)$$

If the latter holds, then by Cauchy-Schwarz (4) is given by the variance of $\ell^*(F, G, \Psi)$.

According to general theory this quantity (4), 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

$$\Psi(F_n) - \Psi(F) = \int \ell^*(F, G, \Psi)(x) d(P_n - P_{F,G})(x) + o_P(1/\sqrt{n}). \quad (7)$$

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

We will now show that (5) 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:

$$P_{F,G} = \frac{1}{\alpha(F, G)} P'_{F,G}, \text{ where } F \rightarrow \alpha(F, G) \text{ and } F \rightarrow P'_{F,G} \text{ are linear.} \quad (8)$$

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 (5) 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)

$$\Psi(F_1) - \Psi(F) = -\frac{\alpha(F, G)}{\alpha(F_1, G)} \int \ell^*(F_1, G, \Psi)(x)dP_{F,G}(x). \quad (9)$$

We want to apply this identity (9) 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 $\alpha \rightarrow 0$, then the identity (9) 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. This condition has been verified for a general class of missing data models which allow complete observations in van der Laan (1993b).

In many missing data models with independent censoring (van der Laan, 1993b), 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 (6)), 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) along this one dimensional submodel that the NPMLE (F_n, G_n) should solve this score:

$$\int \ell^*(F_n, G, \Psi)(x)dP_n(x) = 0. \quad (10)$$

Combining this so called *efficient score equation* with (9) for the pair $(F, F_1) = (F, F_n)$, we obtain:

$$\Psi(F_n) - \Psi(F) = \frac{\alpha(F, G)}{\alpha(F_n, G)} \int \ell^*(F_n, G, \Psi)(x)d(P_n - P_{F,G})(x). \quad (11)$$

Comparing (11) with (7) teaches us that identity (11) 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.

In the random truncation model, worked out below, the identity (11) can be explicitly written down. Here, we do not apply empirical process theory to the identity which would

provide us with an alternative efficiency proof for the well known and understood product limit-estimator; for the interested reader we refer to the completely worked out examples in van der Laan (1994).

Example 2.1 (Random truncation model). Let X_1, \dots, X_n be n i.i.d. copies of a real valued X with distribution function F on $[0, \infty)$, 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 on $[0, \infty)$, where G is completely unknown. X and C are independent. We observe (X_i, C_i) if $X_i < C_i$. So we are sampling from the conditional distribution instead of the distribution of (X, C) itself. Denote the observed (X_i, C_i) by (X'_i, C'_i) . We have $(X'_i, C'_i) \sim P_{F,G}$, where

$$dP_{F,G}(x, c) = \frac{1}{\alpha(F, G)} dF(x) dG(c) I(x < c)$$

and $\alpha(F, G) = P(X < C) = \int (1 - G)(x) dF(x)$. Let x_0 be fixed. We are concerned with estimating $\Psi(F) = F(x_0)$. We will assume that (here $\bar{G} = 1 - G$)

$$\int \frac{1}{F} dG < \infty \quad \text{and} \quad \int \frac{1}{\bar{G}} dF < \infty. \quad (12)$$

The score operator for F is given by: (we use the notation $Pf = \int f dP$)

$$A_{F,G} : L_0^2(F) \rightarrow L_0^2(P_{F,G}) : h_1 \rightarrow h_1(X') - P_{F,G}h_1$$

and the score operator for G is

$$B_{F,G} : L_0^2(G) \rightarrow L_0^2(P_{F,G}) : h_2 \rightarrow h_2(C') - P_{F,G}h_2.$$

It is easy to verify (see also BKRW, page 249) that the projection of $A_{F,G}(h_1)$ on the range of $B_{F,G}$ is given by:

$$\Pi(h_1 - P_{F,G}h_1 \mid T_2(P_{F,G})) = E(h_1(X') \mid C') - P_{F,G}h_1.$$

Hence the efficient score operator $A_{F,G}^*$ for F is given by:

$$A_{F,G}^*(h_1)(X', C') = h_1(X') - E(h_1(X') \mid C') = h_1(X') - \frac{\int_0^{C'} h_1 dF}{F(C')},$$

which indeed does not depend on G . Consequently, this is a model where one should expect that the efficient score equation (10) holds: $\int \ell^*(F_n, G, \Psi) dP_n = 0$, as will indeed appear to be the case.

Define

$$\begin{aligned} N_n(x) &\equiv \frac{1}{n} \sum_{i=1}^n I(X'_i \leq x) \\ Y_n(u) &\equiv \frac{1}{n} \sum_{i=1}^n I(X'_i \leq u, C'_i > u) \\ \Lambda(t) &\equiv \int_t^\infty \frac{dF(x)}{F(x-)}. \end{aligned}$$

Let N and Y be the expectations of N_n and Y_n : $N = P_{F,G}N_n$ and $Y = P_{F,G}Y_n$. If (12) holds, then the efficient influence function for estimating $F(x_0)$ is given by (see BKRW, page 244, formula (19)):

$$\ell^*(F, G, x_0)(x', c') = \alpha(F, G)F(x_0) \int_{x_0}^{\infty} \left(I(x' \leq u < c') \frac{F(du)}{\bar{G}(u)F^2(u-)} - \frac{I(x' \in du)}{\bar{G}(u)F(u-)} \right). \quad (13)$$

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

$$0 = P_n \ell^*(F_n, G, t) = \alpha(F_n, G)F_n(x_0) \int_{x_0}^{\infty} Y_n(u) \frac{F_n(du)}{\bar{G}(u)F_n^2(u-)} - \frac{N_n(du)}{\bar{G}(u)F_n(u-)}.$$

This holds if and only if $d\Lambda_n(u) \equiv dF_n(u)/F_n(u-) = dN_n(u)/Y_n(u)$ which implies that $F_n(t) = \prod_{(t, \infty)} (1 - dN_n/Y_n)$. This verifies the efficient score equation for F_n . $F_n(t)$ is the well known product limit estimator for the random truncation model. Asymptotic results of this estimator have been obtained by Woodroffe (1985), Wang, Jewell and Tsai (1986), Keiding and Gill (1990) and van der Vaart (1991); under assumption (12) F_n is asymptotically efficient.

It remains to verify the identity (9), i.e. $F_n(x_0) - F(x_0) = -\frac{\alpha}{\alpha_n} P_{F,G} \ell^*(F_n, G, x_0)$. Substitution of (13) and taking the expectation w.r.t. $P_{F,G}$ within the integral tells us that (9) is here given by:

$$F_n(x_0) - F(x_0) = -\frac{\alpha}{\alpha_n} \alpha_n F_n(x_0) \int_{x_0}^{\infty} Y(u) \frac{F_n(du)}{\bar{G}(u)F_n^2(u-)} - \frac{N(du)}{\bar{G}(u)F_n(u-)}. \quad (14)$$

We have that $Y = F\bar{G}/\alpha$ and $N(du) = F(du)\bar{G}(u)/\alpha$. Hence (14) can be rewritten as follows:

$$\begin{aligned} F_n(x_0) - F(x_0) &= -F_n(x_0) \int_{x_0}^{\infty} \frac{F(u)}{F_n(u-)} (\Lambda_n - \Lambda)(du) \\ &= -\int_{x_0}^{\infty} \frac{F_n(x_0)}{F_n(u-)} (\Lambda_n - \Lambda)(du) F(u) \\ &= -\int_{x_0}^{\infty} \prod_{(x_0, u)} (1 - \Lambda_n(dv)) (\Lambda_n - \Lambda)(du) \prod_{(u, \infty)} (1 - \Lambda(dv)), \end{aligned}$$

which is just the well known *Duhamel equation* for the product integral (see Gill and Johansen, 1990, page 1519, Theorem 6).

This proves the identity (11) for the NPMLE, i.e. the product limit estimator, in the random truncation model.

References

- P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and adaptive inference in semiparametric models*, Johns Hopkins University Press, Baltimore.
- 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 B.J. Wijers (1993), *The line-segment problem*, submitted to Bernoulli.

- 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.
- N. Keiding and R.D. Gill (1990), Random truncation models and Markov processes, *Ann. Statist.* **18**, 582–602.
- G.M. Laslett (1982), The survival curve under monotone density constraints with applications to two-dimensional line segment processes, *Biometrika* **69** 153–160.
- M.J. van der Laan (1993a), *General identity for linear parameters in convex models with application to efficiency of NPML*, Submitted to the Annals of Statistics.
- M.J. van der Laan (1993b), *Efficient and inefficient estimation in semiparametric models*, thesis, ISBN nr. 90-393-0339-8, Department of Mathematics, Utrecht, the Netherlands. Will appear as CWI tract **44**, Centre for Mathematics and Computer Science, Amsterdam.
- M.J. van der Laan (1994), *Proving efficiency of NPML and identities*, technical report **44**, Group in Biostatistics, Berkeley.
- A.W. van der Vaart (1991), On differentiable functionals, *Ann. Statist.* **19**, 178–204.
- A.W. van der Vaart and J.A. Wellner (1994), *Weak convergence and empirical processes*. IMS Lecture Notes-Monograph Series (to appear).
- M.C. Wang, N.P. Jewell and W.Y. Tsai (1986), Asymptotic properties of the product limit estimate under random truncation, *Ann. Statist.* **14**, 1597–1605.
- M. Woodroffe (1985), Estimating a distribution function with truncated data, *Ann. Statist.* **13**, 163–177. Correction: *Ann. Statist.* **15** (1987), 883.