Bernoulli **1(4)**, 1995, 335–341

An identity for the nonparametric maximum likelihood estimator in missing data and biased sampling models

MARK J. VAN DER LAAN

Department of Statistics, Division of Biostatistics, University of California, Berkeley, CA 94703, USA

We derive an identity for the maximum likelihood estimator in nonparametric missing data models and biased sampling models, which almost says that this estimator 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.

Keywords: Asymptotic efficiency; efficient influence curve; empirical process

1. Introduction

Let $X_1, \ldots, X_n$ be n independently and identically distributed (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 a maximum likelihood estimator of (F, G) in the sense that for a dominating measure μ_m of P_{F_n,G_n} we have:

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

where P_n denotes the empirical distribution of $X_1, \ldots, X_n$. In particular, (1) will hold for a nonparametric maximum likelihood estimator (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 $\mu \, \mathrm{d}P_{F,G}/\mathrm{d}\mu = p_F p_G$ 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 into an F part and a G part so that F_n can be determined by just maximizing the relevant part of the log-likelihood. Also certain information calculations (below) do not depend on knowledge of G and hence we can do as if G is known.

In this paper we derive a crucial identity (see (9)) for missing data 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 the 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 *et al.* (1993). Then we will derive the identity and discuss its application in proving the 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.

2. 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 with respect to F_2 and that $\mathrm{d}F_1/\mathrm{d}F_2$ is bounded. For each $F_1 \ll_b F$ we define a line $f_\epsilon = (1+\epsilon h_1)f$ 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_ϵ 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 $g_\epsilon = (1+\epsilon h_2)g$ from G_1 to G, where $h_2 = (g_1 - g)/g \in L_0^2(G)$. These lines imply a one-dimensional submodel $P_{F_\epsilon, G_\epsilon}$ through $P_{F,G}$. We will assume that

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\log\left(\frac{\mathrm{d}P_{F_\epsilon,G_\epsilon}}{\mathrm{d}P_{F,G}}\right)\Bigg|_{\epsilon=0} = A_{F,G}(h_1) + B_{F,G}(h_2), \tag{2}$$

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

$$A_{F,G}(h_1) \equiv \frac{\mathrm{d}}{\mathrm{d}\epsilon}\log\left(\frac{\mathrm{d}P_{F_\epsilon,G}}{\mathrm{d}P_{F,G}}\right)\Bigg|_{\epsilon=0}$$

$$B_{F,G}(h_2) \equiv \frac{\mathrm{d}}{\mathrm{d}\epsilon}\log\left(\frac{\mathrm{d}P_{F,G_\epsilon}}{\mathrm{d}P_{F,G}}\right)\Bigg|_{\epsilon=0}.$$

The equalities in (2) and the limits in $\mathrm{d}/\mathrm{d}\epsilon$ are assumed to hold in $L^2(P_{F,G})$. In view of the linearity of Ψ the Cramér–Rao lower bound for the variance of $\sqrt{n}$-normed unbiased estimator of $\Psi(F) = \Psi(F_0)$ along the one-dimensional submodel $P_{F_\epsilon, G_\epsilon}$ with parameter ϵ is now given by:

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

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 $T_2(P_{F,G})$ to be the closure of $B_{F,G}(L_0^2(G))$, 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:

$$\sup_{h_1 \in L_0^2(F)} \left(\frac{\Psi\left(\int h_1 \, \mathrm{d}F\right)}{\|A^*_{F,G}(h_1)\|_{P_{F,G}}} \right)^2, \tag{4}$$

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

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

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

$$\Psi(F_1) - \Psi(F) = \Psi\left(\int h_1 \, \mathrm{d}F\right) = \langle \kappa(F, \Psi), h_1 \rangle_F.$$

Let $A^{*\mathrm{T}}_{F,G} : L_0^2(P_{F,G}) \to L_0^2(F)$ be the adjoint of $A^*_{F,G}$. If $\kappa(F, \Psi)$ lies in the range of the so-called *information operator* $I_{F,G} \equiv A^{*\mathrm{T}}_{F,G} A^*_{F,G}$, then

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

where

$$\ell^*(F, G, \Psi) = A^*_{F,G}(A^{*\mathrm{T}}_{F,G} A^*_{F,G})^-(\kappa(F, \Psi)). \tag{6}$$

If the latter holds, then, by the Cauchy–Schwarz inequality, the bound (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 (Bickel *et al.* 1993). 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

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

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:

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

Then it is easily verified that

$$A_{F,G}(h_1) \, \mathrm{d}P_{F,G} = -\frac{\alpha(F_1 - F, G)}{\alpha^2(F, G)} \mathrm{d}P'_{F,G} + \frac{\mathrm{d}P'_{F_1 - F, G}}{\alpha(F, G)}.$$

We also have:

$$\mathrm{d}P_{F_1,G} - \mathrm{d}P_{F,G} = -\frac{\alpha(F_1 - F, G)}{\alpha(F_1, G)\alpha(F, G)}\mathrm{d}P'_{F,G} + \frac{\mathrm{d}P'_{F_1-F,G}}{\alpha(F_1, G)}$$
$$= \frac{\alpha(F, G)}{\alpha(F_1, G)} A_{F,G}(h_1)\,\mathrm{d}P_{F,G}.$$

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):

$$\Psi(F_1) - \Psi(F) = -\frac{\alpha(F, G)}{\alpha(F_1, G)}\int \ell^*(F_1, G, \Psi)(x)\,\mathrm{d}P_{F,G}(x). \tag{9}$$

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 $\alpha \to 0$, then (9) holds also for F_n. Since $F_n(\alpha)$ converges to F_n with respect to 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 *et al.* 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 to 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:

$$\int \ell^*(F_n, G, \Psi)(x)\,\mathrm{d}P_n(x) = 0. \tag{10}$$

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

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

Comparing (11) with (7) teaches us that (11) almost says that $\Psi(F_n)$ is efficient. We are now in the perfect setting to apply empirical process theory (see, for example, van der Vaart and 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 (*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

$$\mathrm{d}P_{F,G}(x,c) = \frac{1}{\alpha(F,G)} \mathrm{d}F(x)\,\mathrm{d}G(c) I(x < c)$$

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

$$\int \frac{1}{F}\mathrm{d}G < \infty \qquad \int \frac{1}{\bar{G}}\mathrm{d}F < \infty. \tag{12}$$

Using the notation $Pf = \int f\,\mathrm{d}P$, the score operator for F is given by

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

and the score operator for G is

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

It is easy to verify (see also Bickel *et al.* 1993, p. 249) that the projection of $A_{F,G}(h_1)$ on the range of $B_{F,G}$ is given by:

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

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

$$A^*_{F,G}(h_1)(X', C') = h_1(X') - \mathrm{E}\{h_1(X')|C'\} = h_1(X') - \frac{\int_0^{C'} h_1\,\mathrm{d}F}{F(C')},$$

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

Define

$$N_n(x) \equiv \frac{1}{n}\sum_{i=1}^n I(X'_i \le x)$$

$$Y_n(u) \equiv \frac{1}{n}\sum_{i=1}^n I(X'_i \le u,\ C'_i > u)$$

$$\Lambda(t) \equiv \int_t^\infty \frac{\mathrm{d}F(x)}{F(x-)}.$$

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 Bickel *et al.* 1993, p. 244, formula (19)):

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

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

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

This holds if and only if $\mathrm{d}\Lambda_n(u) \equiv \mathrm{d}F_n(u)/F_n(u-) = \mathrm{d}N_n(u)/Y_n(u)$, which implies that $F_n(t) = \prod_{(t,\infty)}(1 - \mathrm{d}N_n/Y_n)$. 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 Woodroofe (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 (9), i.e. $F_n(x_0) - F(x_0) = -(\alpha/\alpha_n)P_{F,G}\ell^*(F_n,G,x_0)$. Substitution of (13) and taking the expectation with respect to $P_{F,G}$ within the integral tells us that (9) is here given by:

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

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

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

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

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

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and adaptive inference in semiparametric models*. Baltimore, MD: Johns Hopkins University Press.

Gill, R.D. and Johansen, S. (1990) A survey of product integration with a view towards application in survival analysis. *Ann. Statist.*, **18**, 1501–1555.

Gill, R.D., van der Laan, M.J. and Wijers, B.J. (1993) The line-segment problem. Submitted to *Bernoulli*.

Keifer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Statistic.*, **27**, 887–906.

Keiding, N. and Gill, R.D. (1990) Random truncation models and Markov processes. *Ann. Statist.*, **18**, 582–602.

Laslett, G.M. (1982) The survival curve under monotone density constraints with applications to two-dimensional line segment processes, *Biometrika*, **69**, 153–160.

van der Laan, M.J. (1993a) General identity for linear parameters in convex models with application to efficiency of NPMLE. Submitted to *Ann. Statist.*

van der Laan, M.J. (1993b) *Efficient and Inefficient Estimation in Semiparametric Models*. Doctoral thesis, Department of Mathematics, Utrecht, the Netherlands. Will appear as CWI tract 44, Centre for Mathematics and Computer Science, Amsterdam.

van der Laan, M.J. (1994) Proving efficiency of NPMLE and identities. Technical report 44, Group in Biostatistics, Berkeley.

van der Vaart, A.W. (1991) On differentiable functionals. *Ann. Statist.*, **19**, 178–204.

van der Vaart, A.W. and Wellner, J.A. (1995) *Weak Convergence and Empirical Processes*, Springer Verlag, New York, (to appear).

Wang, M.C., Jewell, N.P. and Tsai, W.Y. (1986) Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.*, **14**, 1597–1605.

Woodroofe, M. (1985) Estimating a distribution function with truncated data. *Ann. Statist.*, **13**, 163–177. Correction: *Ann. Statist.*, **15** (1987), 883.

Received May 1994 and revised March 1995