The International Journal of Biostatistics

Volume 2, *Issue* 1 2006 *Article* 7

Choice of Monitoring Mechanism for Optimal Nonparametric Functional Estimation for Binary Data

Nicholas P. Jewell, *Division of Biostatistics, School of Public Health, University of California, Berkeley*
Mark J. van der Laan, *Division of Biostatistics, School of Public Health, University of California, Berkeley*
Stephen Shiboski, *Department of Epidemiology and Biostatistics, University of California, San Francisco*

Recommended Citation:
Jewell, Nicholas P.; van der Laan, Mark J.; and Shiboski, Stephen (2006) "Choice of Monitoring Mechanism for Optimal Nonparametric Functional Estimation for Binary Data," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 7.
DOI: 10.2202/1557-4679.1031

Choice of Monitoring Mechanism for Optimal Nonparametric Functional Estimation for Binary Data

Nicholas P. Jewell, Mark J. van der Laan, and Stephen Shiboski

Abstract

Optimal designs of dose levels in order to estimate parameters from a model for binary response data have a long and rich history. These designs are based on parametric models. Here we consider fully nonparametric models with interest focused on estimation of smooth functionals using plug-in estimators based on the nonparametric maximum likelihood estimator. An important application of the results is the derivation of the optimal choice of the monitoring time distribution function for current status observation of a survival distribution. The optimal choice depends in a simple way on the dose-response function and the form of the functional. The results can be extended to allow dependence of the monitoring mechanism on covariates.

KEYWORDS: current status data, design, dose response, functional estimation, nonparametric maximum likelihood estimation

Author Notes: The first author thanks the Miller Institute of the University of California, Berkeley for support during the writing of this paper

1 Introduction

A common problem in dose-response experiments is estimation of the relationship between the level of a dose, C and the probability of a binary response, denoted by $F(C)$. Suppose the function $F = F_\theta$ is parametrically modeled by say a logistic or probit function, and that n_i observations are taken at a set of k dose levels $c_1, \ldots, c_k$. A natural design question relates to the optimal choice of k, $c_1, \ldots, c_k$ and $n_1, \ldots, n_k$, subject to a fixed total sample size $n = \sum n_i$, with regard to efficient estimation of all or some components of θ. See, for example, Sitter (1992) and the references therein. Such optimization often leads to one, two or three point designs that depend on the unknown value of θ and the optimality criterion used. For example, suppose F corresponds to a logistic distribution and interests focuses on estimation of the mean of F—in this case, also the so-called ED50—with optimality based on the asymptotic variance of the maximum likelihood estimator. Then, the optimal design is as a single-point design with all the observations taken at c, the mean of F (see, Wu, 1988).

A problem with such optimal designs is that the answer depends on the value of the unknown parameter(s). Usually, investigators assume that a preliminary estimate is available which is then used as the basis for a selected optimal design. Abdelbasit & Plackett (1983) note that three-point designs typically tend to be more robust to poor initial parameter estimates than two-point designs, supporting earlier intuitive ideas that the"less knowledge of the parameter values one has prior to the experiment, the more spread out the design should be and the more design points should be used" (Sitter, 1992, p. 1146). Sitter (1992) tackles the issue that the optimal design depends on unknown values of θ using a minimax approach over a region of possible values for θ, but does not consider that the parametric model for F_θ is also assumed to be known in advance. Here we formalize the lack of knowledge of both θ and, more generally, the form of F_θ, by considering the optimal choice of the dose levels where the form of F is unspecified and interest focuses on estimation of a single functional of F. This approach specifically illustrates how uncertainty regarding F leads to more "spread out" designs.

The ideas discussed here have immediate relevance to all dose-response experiments where there is little prior knowledge regarding the shape of the dose-response relationship. In addition, the results have immediate application to estimation of functionals of the distribution, F, of a survival random variable, T, where estimation is based on current status data; here, observation of T is restricted to knowledge of whether or not T exceeds a random independent monitoring time C. Nonparametric estimation of the survival

function, and semi-parametric techniques for related regression models, based on current status data, are reviewed in Jewell & van der Laan (2004). In detail, let T be the survival random variable of interest, with associated distribution function F. Assume that the monitoring time, C, is randomly selected from a distribution function G, independently of T. An independent and identically distributed sample of n individuals is therefore drawn from the joint distribution of (T, C); however, only $\{(\Delta_i, C_i) : i = 1, \ldots, n\}$ is observed where $\Delta = I(T \leq C)$. In this context, the design question relates to optimal choice of G for estimation of a given functional of F, based on such current status data. In some settings, choice of the monitoring times may not be under the control of the investigator; however, in many applications in carcinogenecity testing and cross-sectional disease incidence estimation, monitoring times may be pre-selected. We use current status notation in what follows below.

2 Optimal Choice of G with F unspecified

Nonparametric maximum likelihood estimation of the distribution function, F (on $[0, \infty)$), of T from current status data is easily implemented using the pool-adjacent-violator algorithm (Ayer et al, 1955). Here we wish to select the distribution function, G, of C, in terms of minimizing the asymptotic variance of a specific functional estimate. We assume throughout that G is continuous with density function g.

The properties of the nonparametric maximum likelihood estimator F_n of F, based on current status data, were established by Groeneboom & Wellner (1992) who, in particular, considered the efficiency of smooth functionals of F_n as estimators of the corresponding functional of F. The estimator, F_n, is known to converge only at the rate $n^{-1/3}$. However, plug-in estimates of smooth functionals are asymptotically Gaussian, converging at the standard rate $n^{-1/2}$.

In detail, consider the parameter $\mu = \int (1 - F(u)) r(u) du$ for some function r, and the corresponding estimator $\mu_n = \int (1 - F_n(u)) r(u) du$. Suppose there is a constant $M < \infty$ so that (i) r is bounded on $[0, M]$, (ii) F is continuous with a density $f > 0$ on $[0, M]$ and zero elsewhere, and (iii) $g(c) = dG/dc > 0$ on $[0, M]$. Huang & Wellner (1995) proved that, for any pair (F, G) and function r that satisfy (i)–(iii), the estimator μ_n is regular and asymptotically linear with influence curve given by

$$IC = \frac{r(c)}{g(c)}(\Delta - F(c)), \tag{1}$$

DOI: 10.2202/1557-4679.1031

with variance

$$VAR(IC) = \int \frac{r^2(c)}{g(c)} F(c)(1 - F(c))dc. \tag{2}$$

The question we pose here is that, for a given r and F, what choice of the monitoring time distribution G minimizes the variance of the influence function for μ_n? That is we seek the G that minimizes the right hand side of (2).

To solve this optimization problem, we use a simple variational calculus analysis of (2) with respect to the density g corresponding to G. Specifically, let h be any bounded function in ${L_0}^2(G)$, the set of all square-integrable functions with respect to the measure dG that satisfy $\int h(c)dG(c) = 0$; then, for any g_0 and for a small enough number ϵ, $(1 + \epsilon h)g_0$ describes a one-dimensional family of densities that passes through g_0 at $\epsilon = 0$. If g_0 minimizes (2), it follows that the function

$$\epsilon \to \int \frac{r^2(c)}{(1 + \epsilon h(c))g_0(c)} F(c)(1 - F(c))dc \tag{3}$$

has a minimum at $\epsilon = 0$. That is,

$$\frac{d}{d\epsilon} \int \frac{r^2(c)}{(1 + \epsilon h(c))g_0(c)} F(c)(1 - F(c))dc \Big|_{\epsilon=0} = 0.$$

This yields $\int \frac{r^2(c)}{g_0(c)} F(c)(1 - F(c))h(c)dc = 0$. This is equivalent to saying that

$$\int \frac{r^2(c)}{[g_0(t)]^2} F(c)(1 - F(c))h(c)dG(c) = 0.$$

Since this is true for all h in ${L_0}^2(G)$, it follows that

$$\frac{r^2(c)}{[g_0(c)]^2} F(c)(1 - F(c)) = K,$$

for some constant K. Solving for K by normalizing then yields

$$g_0(c) = \frac{|r(c)|F(c)^{1/2}(1 - F(c))^{1/2}}{K^*}, \tag{4}$$

where the constant $K^* = \int |r(c)|F(c)^{1/2}(1 - F(c))^{1/2}dc$. To complete this analysis, we must show that this g_0 in fact describes a minimum of (3). This is seen by taking the second derivative of (3), and evaluating at $\epsilon = 0$; this yields

$$2 \int \frac{r^2(c)}{g_0(c)} F(c)(1-F(c))h(c)^2 dc = 2K^* \int |r(c)|F(c)^{1/2}(1-F(c))^{1/2}h(c)^2 dc > 0,$$

as desired.

Note that the conditions underlying the finite variance of (1), derived by Huang & Wellner (1995) and noted above, guarantee the finiteness of the normalizing constant K^*. It is interesting to note that, in the case of the mean of F, with $r(c) \equiv 1$, the resulting condition $\int F(c)^{1/2}(1-F(c))^{1/2}dc < \infty$ is equivalent to $\int (1-F(c))^{1/2}dc < \infty$ which is known as an $L_{2,1}$ condition, arising in the context of conditions for multiplier central limit theorems (see, for example, Ledoux & Talagrand, 1991, p.279). In this case, this is a slightly stronger condition than a second moment condition on F since $\int (1-F(c))^{1/2}dc < \infty$ implies that $\int c^2 dF(c) < \infty$. The condition is also implied by $\int c^j dF(c) < \infty$ for some $j > 2$.

This variational calculus approach, of course, only derives a necessary condition for minimization of (2) assuming a minimum exists. However, with the form of (4) identified, it is straightforward to give a simple rigorous proof via the Cauchy-Schwarz inequality. In particular, given that the normalizing constant K^* is finite, we have, for any density function g,

$$\begin{aligned}\left[\int |r(c)|F(c)^{1/2}(1-F(c))^{1/2}dc\right]^2 &= \left[\int \frac{|r(c)|F(c)^{1/2}(1-F(c))^{1/2}}{g^{1/2}(c)} g^{1/2}(c)dc\right]^2 \\ &\leq \int \frac{r^2(c)F(c)(1-F(c))}{g(c)}dc \int g(c)dc \\ &= \int \frac{r^2(c)F(c)(1-F(c))}{g(c)}dc,\end{aligned}$$

with equality if and only if

$$g^{1/2}(c) = A\frac{|r(c)|F(c)^{1/2}(1-F(c))^{1/2}}{g^{1/2}(c)}$$

for some non-zero constant A. That is, $g(c) \equiv g_0(c)$ as given in (4). Thus,

$$\left[\int |r(c)|F(c)^{1/2}(1-F(c))^{1/2}dc\right]^2 \leq \int \frac{r^2(c)F(c)(1-F(c))}{g(c)}dc$$

for all densities g with equality if and only if g is given by (4).

We have thus shown that the optimal g_0 depends on the function r and F through (4). In the above we have assumed that the function r is fixed and known *a priori*. This has immediate application to estimation of any non-central moments of F, but does not directly apply to central moments where the relevant r depends of F itself. However, the result has immediate extension to such functionals as follows. Suppose a parameter of interest, ψ

is a (possibly non-linear) function of two parameters μ_1 and μ_2, of the kind considered above. That is, for $j = 1, 2$, $\mu_j = \int (1 - F(u)) r_j(u) du$ for two fixed functions r_1 and r_2. In particular, we assume that $\psi = h(\mu_1, \mu_2)$ where h is smooth. To make this concrete, consider σ_F^2, the variance of F with $\sigma_F^2 = \mu_2 - (\mu_1)^2$ where μ_1 and μ_2 are the first two non-central moments of F, respectively. In this case, $r_1 \equiv 1$ and $r_2(c) = 2c$.

Let the influence curves of μ_{n1} and μ_{n2}, the corresponding 'plug-in' estimators of μ_1 and μ_2, be denoted by IC_1 and IC_2, respectively. Then, the influence curve, IC_ψ, of $\psi_n = h(\mu_{1n}, \mu_{2n})$, is given by $\frac{\partial h}{\partial \mu_1}(\mu_1, \mu_2) IC_1 + \frac{\partial h}{\partial \mu_2}(\mu_1, \mu_2) IC_2$ by the delta method. Specifically, $IC_\psi = (\frac{\partial h}{\partial \mu_1} r_1 + \frac{\partial h}{\partial \mu_2} r_2) \frac{(\Delta - F(c))}{g(c)}$ using (1). For example, the influence curve for estimation of σ_F^2 is just $\frac{(2c - 2\mu_1)}{g(c)}(\Delta - F(c))$.

The variance of IC_ψ is then given by (2) with r replaced by $(\frac{\partial h}{\partial \mu_1} r_1 + \frac{\partial h}{\partial \mu_2} r_2)$. Miminimization of this variance over g follows exactly as before (even though the replacement for the function r is not known *a priori* and depends on F) yielding an optimal g_0 given by (4) with r replaced in the exact same manner. Estimation of the variance of F thus yields an optimal choice of g given by

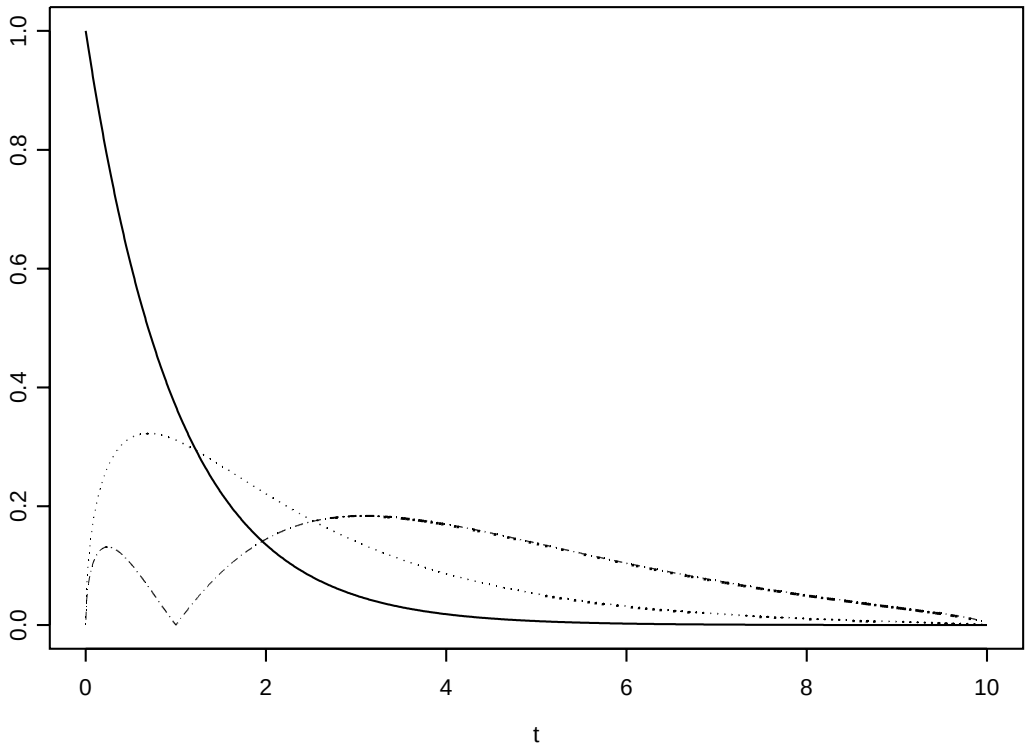
$$g_0(c) = \frac{|2c - 2\mu_1| F(c)^{1/2} (1 - F(c))^{1/2}}{K^*}, \tag{5}$$

where the normalizing constant $K^* = \int |2c - 2\mu_1| F(c)^{1/2} (1 - F(c))^{1/2} dc$ and μ_1 is just the mean of F. We note that this argument is quite general and can be immediately extended to estimation of parameters that are (potentially non-linear) functions of $k > 2$ functionals of the kind considered by Huang & Wellner (1995).

We briefly consider two simple examples where interest focuses on (i) the mean, and (ii) the variance of F. For the mean, take $r(c) \equiv 1$ as noted. Here, the optimal choice is $g_0 \propto F^{1/2}(1 - F)^{1/2}$; thus monitoring times (or doses) should be concentrated around the median of F. Note that, with the mean, the optimal design is symmetric around the mean when F itself is symmetric as should be expected. Alternatively, for the variance, take $r(c) = 2c - 2E(F)$, with the subsequent optimal choice given by $g_0(c) \propto |2c - 2E(F)| F^{1/2}(1 - F)^{1/2}$. In this case, monitoring times (doses) will be much less concentrated around the median of F with more weight given to values in the tails of F, again as suggested by intuition. These simple examples confirm the obvious notion that optimal choice of monitoring times (or doses) depend crucially on what functional is being targeted. We comment further on this issue in §5.

For illustration, suppose the unknown F is described by an exponential distribution with mean 1, conditional on being less than 10. Figure 1 illustrates the optimum choice of g for estimation of the mean and variance of F, based

Figure 1: Optimal Choice of Monitoring Time Density, g_0, for Nonparametric Estimates of the Mean (dotted line), and Variance (dash-dotted line) of the Distribution Function F (with density given by the solid line)



on the nonparametric maximum likelihood estimator.

It is important to note whether the gain in efficiency from use of the optimal g is of practical importance. This can now be investigated for any combination of F and r since we explicitly know the optimal variance (2) using (4). For example, suppose F is Uniform on $[0,1]$, $r(c) \equiv 1$, and the monitoring density g is also selected to be Uniform on $[0,1]$. Then $VAR(IC) = \int_0^1 c(1-c)dc = 0.1666$, whereas the optimal variance using $g_0(c) = c^{1/2}(1-c)^{1/2}(8/\pi)I_{(0<c<1)}$ is simply $(\pi/8)^2 = 0.1542$. In this case, the relative efficiency of the simple sub-optimal design for estimating the mean of F is 93%, so that the gain from the use of the optimal g is not great. Poorer choices of a monitoring

Figure 2: Monitoring Time Densities, g, for Nonparametric Estimates of the Mean and Variance of F, the Uniform Distribution on $[0, 1]$; the Optimal g for Estimation of the Mean and Variance of F are the Dash-dotted and Solid Lines, Respectively; g Uniform is Long-dash Line; g_5 is the Dotted Line

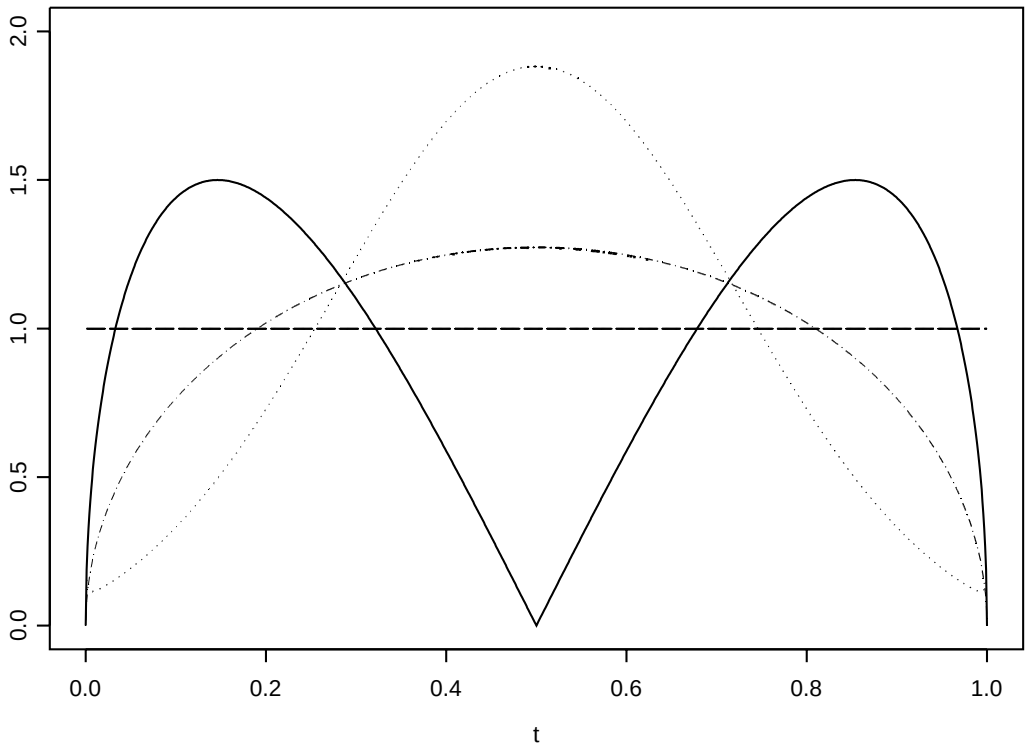


Table 1: Asymptotic Variances of the Influence Curves for Nonparametric Estimators of the Mean and Variance of the Distribution Function F Based on Various Choices of Monitoring Time Densities g: g_0 is the optimal density for the particular choice of F and functional; g_5, supported on $[0,1]$ is defined in the text; ${g_5}^*$ is g_5 rescaled to be supported on $[0,10]$

F	g	Estimating Mean of F		Estimating Variance of F	
		$Var(IC)$	Rel. Effic.	$Var(IC)$	Rel. Effic.
Uniform	g_0	0.1542	100%	0.0278	100%
	Uniform	0.1666	93%	0.0333	83%
	g_5	0.1888	82%	0.0710	39%
Truncated Exponential	g_0	2.4018	100%	22.44	100%
	Uniform	4.9959	48%	29.39	76%
	${g_5}^*$	14.8670	16%	85.01	26%

density for the mean here might correspond to densities of the form $g_p(c) = c^{-1}(1-c)^{-1}\left(\log(1/c(1-c))\right)^{-p}/K_p$ for $p > 1$. See Figure 2 for a graph of g_5 for which $K_5 = 0.4150$; for this choice of g, $VAR(IC) = 0.1888$ for estimation of the mean of F. Now, the relative efficiency has dropped to 82%, that, although non-trivial, is still not catastrophic. Table 1 provides similar calculations for a similar set of monitoring densities where F is the truncated unit exponential, supported on $[0,10]$, discussed above.

As noted, for estimation of the variance of F, we take $r(t) = 2t - E(F)$. In this case, with both F and g given by the Uniform distribution and density on $[0,1]$, respectively, $VAR(IC) = \int_0^1 (2c - \frac{1}{2})^2 c(1-c)dc = 0.0333 = 1/30$. The optimal variance is $\left[\int_0^1 |2c-1| c^{1/2}(1-c)^{1/2} dc\right]^2 = 0.0278$. Now, for this functional, the relative efficiency of the sub-optimal design is 83%, so that the gain from the use of the optimal g is more substantial. Similar calculations for other choices of g, and when F is the truncated unit exponential on $[0,10]$, are also given in Table 1. When F is Uniform on $[0,1]$, Figure 2 graphs the various monitoring densities g considered for both estimation of the mean and variance of F.

When F is Uniform, the density g_5 places more mass in the center of the interval and less on the extremes than the optimal density for estimating the mean of F nonparametrically. On the other hand, the Uniform monitoring density is 'closer' to the optimal density resulting in slightly better efficiency.

DOI: 10.2202/1557-4679.1031

When estimating the variance of F nonparametrically, note that the optimal g places no mass at the center of the interval and much more pass towards the two extremes of the interval (see Figure 2). In this case, g_5 places substantially more mass at the center than is optimal resulting in a considerable loss of efficiency to 39%. When F is the truncated unit exponential on $[0, 10]$ however, the monitoring density ${g_5}^*$ places its mass towards the center of the interval $[0, 10]$, far from the mean of F, resulting in a disastrous loss of efficiency to 16%; that the Uniform monitoring density does slighlty better—with an efficiency of 48%—results from it placing much more mass at the left-hand side of the interval as the optimal g does (see Figure 1). These efficiences improve slightly when considering estimation of the variance of F since more mass towards both tails of F is preferable.

It is possible to consider comparisons of the nonparametric optimal design with various parametric counterparts although we do not provide formal numerical results here. Of course, the parametric optimal designs do not provide data that allow nonparametric identifiability of functionals of F so that in this sense the parametric optimal designs have zero efficiency in the nonparametric setting. Further, the two approaches essentially address different questions. If one uses a parametric model incorrectly, the primary concern will be bias rather than efficiency. (For a genral discussion of paremtric and nonparametric efficincy, we refer to Bickel et al. (1993). Note, however, qualitatively that the nonparametric optimal design does indeed 'spread out' the selection of dose levels or monitoring times.

These simple calculations reinforce the nessity for an investigator to think carefully about which functionals of F are of primary interest before selecting a design choice of G. We note that it is straightforward to sample from a pre-selected G in choosing a finite set of doses or monitoring times.

3 Allowing the Optimal Choice of G to Depend on Covariates

We extend the result of §2 to allow for monitoring designs that are allowed to depend on a k dimensional fixed covariate Z. The assumption that C and T are independent is now relaxed to C being independent of T, given Z. In nonparametrically estimating the functional μ, based on observed data

$\{(\Delta_i, C_i, Z_i) : i = 1, \ldots, n\}$, the efficient influence curve is given by

$$\begin{aligned} IC_{eff}(c) &= \frac{r(c)\{\Delta - F(c|Z)\}}{g(c|Z)} + \int_0^\infty r(u)\{1 - F(u|Z)\}du - \mu \\ &= \frac{r(c)\{\Delta - F(c|Z)\}}{g(c|Z)} + \int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du, \end{aligned}$$

with $\bar{F} = 1 - F$, a special case of (4.12) in van der Laan & Robins (2003, p. 242). The variance of this influence curve is then

$$\begin{aligned} E\left(IC_{eff}{}^2\right) &= E\left[\frac{r^2(C)\left\{\Delta - F(c|Z)\right\}^2}{g^2(C|Z)}\right] \\ &\quad + E\left[\left(\int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du\right)^2\right] \\ &\quad + 2E\left[\frac{r(C)\{\Delta - F(c|Z)\}}{g(C|Z)}\int_0^\infty r(u)\{\bar{F}(u|Z) - \bar{F}(u)\}du\right] \\ &= E\left[\frac{r^2(C)\left\{\Delta - F(c|Z)\right\}^2}{g^2(C|Z)}\right] + \phi(F_Z), \end{aligned}$$

where $E(\cdot)$ is the expectation with respect to the data generating distribution, and F_Z is the marginal distribution of Z, which does not depend on g. The second step in this derivation follows from taking conditional expectations, first with respect to C and then with respect to Z. We now seek the optimal set of conditional densities $g(c|Z)$ that minimizes the expectation

$$E\left[\frac{r^2(c)\left\{\Delta - F(c|Z)\right\}^2}{g^2(c|Z)}\right] = E_{F_Z}\left[\int_0^\infty \frac{r^2(c)F(c|Z)\left\{1 - F(c|Z)\right\}}{g(c|Z)}dc\right].$$

For a fixed Z, an identical argument to §2 shows that the density that optimizes $\left[\int_0^\infty \frac{r^2(c)F(c|Z)\{1-F(c|Z)\}}{g(c|Z)}dc\right]$ is given by

$$g_0(c|Z) = \frac{|r(c)|F(c|Z)^{1/2}(1 - F(c|Z))^{1/2}}{K^*(Z)}, \tag{6}$$

with the normalizing constant $K^*(Z) = \int |r(c)|F(c|Z)^{1/2}(1 - F(c|Z))^{1/2}dc$, as before, again assumed finite for all relevant Z. It immediately follows that the densities (6), for all Z, provide the optimal conditional monitoring densities. Practical implementation of this result, of course, requires knowledge of the distribution functions $F(t|Z)$ for appropriate Z which limits applicability with high-dimensional Z.

DOI: 10.2202/1557-4679.1031

4 Further Extensions: Regression Parameters

In many examples, particularly in the presence of covariates, interest focuses on functionals that are not merely based on the marginal distribution F. For example, if we assume a regression model linking T with Z of the form

$$E(T|Z) = \alpha + \beta Z, \tag{7}$$

we may wish to select a monitoring distribution to optimize estimation of β. A simple example of this occurs in a two group comparison of the mean of T. As before, van der Laan & Robins (2003, p.242) provides the relevant efficient influence curve for estimation of a smooth functional $\mu(F_{T,Z})$ of the joint distribution $F_{T,Z}$ of (T, Z). In particular, suppose $D(T, Z)$ is the efficient influence curve for $\mu(F_{T,Z})$ in the full data world where $\{(T_i, Z_i) : i = 1, \ldots, n\}$ is observed. Let $a_{g|Z}$ be the left end point of the support of the density $g(\cdot|Z)$. Then, the analogous efficient influence curve based on $\{(\Delta_i, C_i, Z_i) : i = 1, \ldots, n\}$ is given by

$$\begin{aligned} IC_{eff} &= \frac{D'(c, Z)\{F(c|Z) - \Delta\}}{g(c|Z)} + \int_0^\infty D'(u, Z)\{1 - F(u|Z)\}du \\ &\quad + D(a_{g|Z}, Z) \qquad (8) \\ &\equiv \frac{D'(c, Z)\{F(c|Z) - \Delta\}}{g(c|Z)} + H(Z, F_{T|Z}, a_{g|Z}), \end{aligned}$$

where $D'(t, Z) = \frac{\partial D(t,Z)}{\partial t}$. The variance of this influence curve is then $E(IC_{eff}^2)$ which is comprised of three terms: (i) $E\left[\frac{D'(c,Z)\{F(c|Z)-\Delta\}}{g(c|Z)}\right]^2$, (ii) $2E\left(\left[\frac{D'(c,Z)\{F(c|Z)-\Delta\}}{g(c|Z)}\right] H(Z, F_{T|Z}, a_{g|Z})\right)$, and (iii) $E\left[H(Z, F_{T|Z}, a_{g|Z})\right]^2$. The second of these terms is zero, as can be seen by first calculating the expectation conditional on C and Z so that only the term $\{F(c|Z) - \Delta\}$ is random. The third term only depends on the conditional density $g(c|Z)$ through its left end support point $a_{g|Z}$. Suppose we now consider a quasi-minimization of $E(IC_{eff}^2)$ where we consider only the set of monitoring densities $g(c|Z)$ where $a_{g|Z}$ coincides with the left end point of the support of $F_{T|Z}$. Then, term (iii) does not depend on the shape of $g(c|Z)$ so that the minimization needs only consider term (i). The same approach as in §3 shows that the quasi-optimal conditional monitoring densities given by

$$g_0(c|Z) = \frac{|D'(c, Z)|F(c|Z)^{1/2}(1 - F(c|Z))^{1/2}}{K^*(Z)},$$

with normalizing constant $K^* = \int |D'(c,Z)|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}dc$, minimizes the expectation, conditional on Z, and thus minimizes the full expectation. Note that the quasi-optimal densities share the same left end support points as $F_{T|Z}$ as proscribed.

We note the use of the term 'quasi-optimal' here since, in principle, there may be optimal choices of $g(c|Z)$ that do not possess the same left end support points as $F_{T|Z}$ that outperform these quasi-optimal choices, although $a_{g|Z}$ cannot be smaller than the left end support point of $F_{T|Z}$. However, describing these monitoring densities is more complex since we must offset the role of $g(c|Z)$ in both terms (i) and (iii) above. Further, our quasi-optimal choices make practical sense in that the support of $g_0(c|Z)$ is always selected to agree with that of $F_{T|Z}$.

To illustrate the calculations, we consider simple linear regression as in (7) although the ideas extend readily to more complex regression models. To be specific, consider estimation of β in (7), noting that the efficient influence curve for estimation of α and β is given by

$$D(T,Z) = V^{-1}\begin{pmatrix} 1 \\ Z \end{pmatrix}(T-\alpha-\beta Z),$$

where $V = E\begin{pmatrix} 1 & Z \\ Z & Z^2 \end{pmatrix}$ is a 2×2 matrix that depends solely on the distribution of the covariate Z. Thus

$$D'(t,Z) = V^{-1}\begin{pmatrix} 1 \\ Z \end{pmatrix}.$$

Note that since $D(T,Z)$ is linear in T, and since we assume that each $F(T|Z)$ has finite support, it follows that $\int_0^\infty D'(u,Z)\{1-F(u|Z)\}du = \alpha+\beta Z - a_{F|Z}$. Thus, our restriction to monitoring densities g with $a_{g|Z} = a_{F|Z}$ in this case makes the second and third terms of (8) sum to zero, simplifying both the quasi-optimality argument above and the calculation of the variance of the efficient influence curve. In summary, for quasi-optimal estimation of β we have

$$g_0(c|Z) = \frac{|v_{12}+v_{22}Z|F(c|Z)^{1/2}(1-F(c|Z))^{1/2}}{K^*(Z)},$$

with v_{ij} being the ijth element of V^{-1}, and $K^*(Z)$ a normalizing constant, as above.

Numerical comparisons of the efficiency of various design choices for the monitoring densities at all values of Z can now be carried out directly. In particular, if we assume that each $F(T|Z)$ is Uniform with mean $\alpha+\beta Z$

DOI: 10.2202/1557-4679.1031

and with the same size of support interval for all Z, the calculations for the mean functional of §2 apply immediately to this regression setting, providing identical relative efficiencies. Similarly, the assumption of a truncated unit exponential distribution, as used in §2, also applies to the regression model (7), presumably on the $\log T$ scale, with the same caveat about support intervals at different Z.

For the results in §2–3, it is desirable to allow that some of the components of Z be time-dependent. In this case, the efficient influence curve is the implicit solution to an integral equation, and so it is not easy to see how optimization can proceed straightforwardly. In practice, discrete sequential choice of future monitoring times might be based on current values of the time dependent covariates using the results of §3.

5 Discussion

The simple optimal result developed here suggests that reasonable non-optimal selections of a dose, or monitoring, density are unlikely to introduce substantial additional variability for estimation of the mean, as compared to an optimal choice. Nevertheless, the work here is of value in that it allows such quantitative calculations for any given scenario. In addition, the value of the optimal design may be substantially greater in estimation of other functionals.

In practice, of course, F is no more known a priori than θ in the parametric setting. Thus, an optimum design based on a presumed F may be somewhat different than the true F in the experimental setting. We suggest therefore that a series of plausible Fs be considered along with the associated optimum design. Then, for each such F, the relevant variance of the desired functional can be calculated from (2) over the range of possible optimal designs under consideration. As in Sitter (1992) a minimax criterion could then be used to select a particular design that is robust to some misspecification of F. At the very least, optimum nonparametric and parametric designs can be compared to illuminate how much the design depends on a particular parametric model choice. Similarly, to exploit the role of covariates a plausible regression model for $F(c|Z)$ must be invoked to derive the optimal monitoring densities $g(c|Z)$.

We have focused here on estimation of a single functional. In many examples, investigators may wish to estimate several functionals efficiently and simultaneously. In principle, the joint influence curve can be calculated as in (1) although now we have several possible optimality criteria, including D-, A-, and E-optimality (see Sitter, 1992). Any of these approaches can serve as the basis of optimal choice of g.

In the context of estimation of functionals of a survival distribution, the methods presented here naturally raise similar questions for situations where individuals may be monitored at multiple times. It is likely that substantial information is gained by increasing the number of observations as compared to simple choice of the timing of a single monitoring. It would thus be of interest to consider the trade-off between increasing the number of sampled individuals monitored at a single time as compared to monitoring a fixed number of individuals but at more than one time using a similar variational approach and taking advantage of results on the efficient influence function for more general interval censoring (see, for example, Geskus & Groeneboom, 1996, 1997, 1999)

References

Abdelbasit, K.M., Plackett, R.L. (1983). Experimental design for binary data. *J. Amer. Statist. Assn.* **78**, 90-8.

Ayer, M, Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641-7.

Bickel, P., Klaassen, C.A.J., Ritov, Y., Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Geskus, R.B., Groeneboom, P. (1996). Asymptotically optimal estimation of smooth functionals for interval censoring, part 1. *Statist. Neer.* **50**, 69-88.

Geskus, R.B., Groeneboom, P. (1997). Asymptotically optimal estimation of smooth functionals for interval censoring, part 2. *Statist. Neer.* **51**, 201-19.

Geskus, R.B., Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *Ann. Statist.* **27**, 627-74.

Groeneboom, P., Wellner, J.A. (1992) *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution*. Boston: Birkhäuser.

DOI: 10.2202/1557-4679.1031

Huang, J., Wellner, J.A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statist. Neer.* **49**, 153-63.

Jewell, N.P., van der Laan, M.J. (2004). Current status data: Review, recent developments and open problems. In *Advances in Survival Analysis* , Handbook in Statistics #23, 625-42, Amsterdam: Elsevier.

Ledoux, M., Talagrand, M. (1991) *Probability in Banach Spaces*. New York: Springer.

van der Laan, M.J., Robins, J.M. (2003) *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Sitter, R.R. (1992). Robust designs for binary data. *Biometrics* **48**, 1145-55.

Wu, C.F.J. (1988). Optimal design for percentile estimation of a quantal response curve. In *Optimal Design and Analysis of Experiments*, Amsterdam: Elsevier.