Biometrics 71, 867–874
December 2015
DOI: 10.1111/biom.12362

Deductive Derivation and Turing-Computerization of Semiparametric Efficient Estimation

Constantine E. Frangakis,* Tianchen Qian, Zhenke Wu,*** and Ivan Diaz******

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
**email:* cfrangak@jhsph.edu
***email:* tqian@jhsph.edu
****email:* zhwu@jhu.edu
*****email:* idiaz@jhu.edu

Summary. Researchers often seek robust inference for a parameter through semiparametric estimation. Efficient semiparametric estimation currently requires theoretical derivation of the efficient influence function (EIF), which can be a challenging and time-consuming task. If this task can be computerized, it can save dramatic human effort, which can be transferred, for example, to the design of new studies. Although the EIF is, in principle, a derivative, simple numerical differentiation to calculate the EIF by a computer masks the EIF's functional dependence on the parameter of interest. For this reason, the standard approach to obtaining the EIF relies on the theoretical construction of the space of scores under all possible parametric submodels. This process currently depends on the correctness of conjectures about these spaces, and the correct verification of such conjectures. The correct guessing of such conjectures, though successful in some problems, is a nondeductive process, i.e., is not guaranteed to succeed (e.g., is not computerizable), and the verification of conjectures is generally susceptible to mistakes. We propose a method that can deductively produce semiparametric locally efficient estimators. The proposed method is computerizable, meaning that it does not need either conjecturing, or otherwise theoretically deriving the functional form of the EIF, and is guaranteed to produce the desired estimates even for complex parameters. The method is demonstrated through an example.

Key words: Compatibility; Deductive procedure; Gateaux derivative; Influence function; Semiparametric estimation; Turing machine.

1. Introduction

The desire for estimation that is robust to model assumptions has led to a growing literature on semiparametric estimation. Approximately efficient estimators can be obtained in general as the zeros of an approximation to the efficient influence function (EIF) (Tsiatis, 2007). Semiparametric estimation is useful, for example, for survival analysis (Cox, 1972), for estimating growth parameters in longitudinal studies (Liang and Zeger, 1986), and for estimating quantities under missing data (Robins et al., 1994), including treatment effects based on potential outcomes (Davidian et al., 2005; Crump et al., 2009). Here, we focus on problems in which the distribution of the observed data is, in principle, unrestricted, but where estimability requires use of lower dimensional working models.

Theoretical derivation of the EIF in such problems can be challenging. If this task can be computerized, it can save dramatic human effort, which can then be transferred, for example, to designing new studies. The EIF for the unrestricted problem can be written, in general, as a Gateaux derivative (Hampel, 1974). However, if simple numerical differentiation is used to calculate the EIF by a computer to avoid theoretical derivations, then the EIF's functional dependence on the parameter of interest is not revealed. For this reason, the derivative approach has not been generally used. Instead, the standard approach to obtaining the EIF is to construct theoretically the space of scores under all possible parametric submodels (Begun et al., 1983). This process currently depends on the correctness of conjectures about these spaces and the correctness of their verification. The correct guessing of such conjectures can succeed in some problems, but is a nondeductive process, i.e., is not guaranteed to succeed (e.g., is not computerizable) and, as with their verification, is generally susceptible to mistakes.

We propose a method that can deductively produce semiparametric locally efficient estimators even for complex parameters. In Section 2, we formulate the goal of a deductive method and show that it essentially requires numerical access to the functional dependence of the EIF on the parameter of interest. Section 3 shows how the concept of compatibility solves the functional dependence problem, and derives a deductive method. Throughout, we use the two-phase design as a test problem where the EIF is known theoretically, and we demonstrate our method with a study on asthma as an example. Section 4 discusses extensions, and Section 5 concludes with remarks.

2. The Problem of Deductive Computerization of Semiparametric Estimators

2.1. *The Goal of a Deductive Method*

Suppose, we conduct a study to measure data D_i, $i = 1, \ldots, n$, independent and identically distributed (iid) from an unknown distribution F, in order to estimate a root-n estimable

feature of the distribution

$$\tau(F). \tag{1}$$

Suppose τ has a nonparametric EIF denoted by $\phi(D_i, F - \tau, \tau)$, where $F - \tau$ denotes the remaining components of the distribution, other than τ. The goal is to find a deductive method that can derive ϕ and can compute estimators $\widehat{\tau}$ that solve

$$\sum_i \phi\{D_i, (F-\tau)_w, \tau\} = 0 \tag{2}$$

after substituting for $(F - \tau)$ estimates of a working model $(F - \tau)_w$. Under some regularity conditions, estimators solving (2) are consistent and locally efficient if the working estimators of $(F - \tau)_w$ are consistent with convergence rates larger than $n^{1/4}$ (van der Vaart, 2000). Our specific requirement that a method be "deductive and computerizable," means that the method should need neither conjecturing for, nor otherwise theoretically deriving the functional form of ϕ, and should be guaranteed to produce an estimate in the sense of Turing (1937) (i.e., use a discrete and finite set of instructions, and, for every input, finish in discrete finite steps).

2.2. *Conjecturing and Functional Form as Barriers toward a Deductive Method*

2.2.1. A Test Problem: Estimating the Mean in a Two-Phase Design. To help make arguments concrete, we consider the following example where the EIF is well known. Suppose that in order to estimate the mean $\tau = E(Y)$ in a population, the researcher first obtains a simple random sample of individuals and records an easily measured covariate X_i. Then, the researcher is to measure the main outcome Y_i only for a subset denoted with $R_i = 1$, where the missing data mechanism is ignorable given X, i.e., $\text{pr}(R_i = 1 \mid Y_i, X_i) = \text{pr}(R_i = 1 \mid X_i)$ (Rubin, 1976). The final data D_i are $(X_i, R_i, Y_i R_i)$, $i = 1, \ldots, n$, iid from a distribution F, and, by ignorability, the parameter τ is identified from F as

$$\tau(F) = \int y(x)p(x)dx, \tag{3}$$

where $p(x)$ is the density of X_i; and $y(x)$ is the conditional expectation $E(Y_i \mid R_i = 1, X_i = x)$. For this problem, the EIF is known (e.g., Robins and Rotnitzky (1995) and Hahn (1998)) to be

$$\phi\{D_i, (F-\tau), \tau\} = \frac{R_i \cdot \{Y_i - y(X_i)\}}{e(X_i)} + y(X_i) - \tau, \tag{4}$$

where $e(x)$ is the propensity score of selection into the second phase, $\text{pr}(R_i = 1 \mid X_i = x)$. The derivation has, so far, been nondeductive because it is first based on conjectures on the score space over all submodels, which are then verified to be true (e.g., Hahn (1998)).

2.2.2. Current Estimation Methods Need the Functional Form of the EIF. Most existing approaches to using (2) first isolate a dependence of ϕ on τ, then replace the remaining dependence on F with a working model, and finally solve for τ. For example, in the test problem above, the most common approach to using (4) to estimate τ first obtains working functions $y_w(X_i)$ and $e_w(X_i)$, for example, using parametric MLEs, and estimates τ as the zero of the empirical sum of (4), to obtain the following:

$$\widehat{\tau}^{\text{ nonde-}}_{\text{ductive}} = \frac{1}{n}\sum_i \frac{R_i \cdot \{Y_i - y_w(X_i)\}}{e_w(X_i)} + y_w(X_i); \tag{5}$$

See, for example, Robins et al. (1994), Davidian et al. (2005), and Kang and Schafer (2007). While there also exist modified estimators like the targeted minimum loss estimator (TMLE) (van der Laan and Rubin, 2006), all methods that have been presented so far have advocated that it is critical to know the functional form dependence of ϕ on F, and so are nondeductive, hence, noncomputerizable without prior knowledge of the functional form.

2.2.3. The Gateaux Derivative Approach to EIF. For a general parameter τ, the EIF evaluated at an observation d' can be obtained as the Gateaux derivative

$$\phi(d', F) = \lim_{\epsilon \to 0} \frac{\tau(F_{d',\epsilon}) - \tau(F)}{\epsilon}, \text{ where} \tag{6}$$

$$F_{d',\epsilon} = (1-\epsilon)F + \epsilon \cdot 1 < d' >, \tag{7}$$

where $1 < d' >$ denotes a point mass at d' (Hampel, 1974). Calculating this derivative at a given d' and F is a deductive and computerizable operation. To demonstrate the ease of its derivation consider again the test problem with missing data.

Specifically, for a given observation $d' = (x', r', y'r')$ and a distribution F, it follows from (3), (7), and Bayes rule, that

$$\tau(F_{d',\epsilon}) = \int y_{d',\epsilon}(x)p_{d',\epsilon}(x)dx, \tag{8}$$

where $p_{d',\epsilon}(x) = (1-\epsilon)p(x) + \epsilon \cdot 1(x = x')$,

and

$$y_{d',\epsilon}(x) = \frac{\epsilon \cdot 1(x = x', r' = 1) \cdot y' + (1-\epsilon) \cdot p(x)e(x)y(x)}{\epsilon \cdot 1(x = x', r' = 1) + (1-\epsilon) \cdot p(x)e(x)},$$

where $1(\cdot)$ is 1 (or 0) if the logical statement $\cdot$ is true (or false). Then, (6) becomes

$$\phi(d', F) = \int \left[\frac{\partial y_{d',\epsilon}(x)}{\partial \epsilon} p_{d',\epsilon}(x)\right]_{\epsilon=0} dx + \int \left[y_{d',\epsilon}(x)\frac{\partial p_{d',\epsilon}(x)}{\partial \epsilon}\right]_{\epsilon=0} dx.$$

The first and second terms of the above are $\frac{r'\{y' - y(x')\}}{e(x')}$ and $y(x') - \tau$, respectively, which is the result (4) above.

The problem with the derivative operation is that if simple numerical differentiation is used to calculate the EIF by a computer to avoid theoretical derivations, then the EIF's functional dependence on the parameter of interest τ and F is not revealed.

3. A Deductive Estimation Method

3.1. *Method*

A start to finding a deductive method is to appreciate from a new perspective a problem that nondeductive estimators such as (5) have. Specifically, nondeductive estimators are usually constructed from a dependence of the EIF ϕ on τ that is different from the variation-independent partition into $[(F-\tau), \tau]$ (this is probably because of the limitations of closed-form expressions). For example, the estimator $\widehat{\tau}^{\text{nonde-}}_{\text{ductive}}$ of (5) is a sample analogue of (i) the expression of the last appearance "τ" in the right hand side of (4), using (ii) a working expectation $y_w(x)$; and (iii) the empirical estimator for $p(x)$ to average over quantities of X_i. However, the parameters underlying (i), (ii), and (iii)—namely, τ, $y(x)$, and $p(x)$, respectively—are not variation independent, because τ is the average of $y(x)$ over $p(x)$. This creates an incompatibility: the value of the estimator $\widehat{\tau}^{\text{nonde-}}_{\text{ductive}}$ from this method differs (almost surely) from its defining expression $\tau(F)$ if for F we use the estimates in (ii) and (iii) that are used to produce $\widehat{\tau}^{\text{nonde-}}_{\text{ductive}}$.

The problem of incompatibility has been noted before as a nuisance (e.g., Newey (1998)) and has motivated compatible estimators like the TMLE (e.g., van der Laan and Rubin (2006)). Here, we show that, more fundamentally, the concept of incompatibility together with the Gateaux derivative creates a solution to the problem of deductive estimation. In particular, the previous section noted that evaluation of the Gateaux derivative at a working distribution F_w masks the dependence on τ. However, the same evaluation does contain evidence that parts of the working distribution F_w are misspecified, if the empirical sum of the Gateaux derivative is not zero. This evidence of misspecified F_w can be turned, by "εις άτοπον απαγωγή" ("reduction to the absurd"), into estimation for τ, where plausible values of τ are values $\tau(F)$ for distributions F for which the empirical sum of the Gateaux derivative is zero and therefore eliminates any evidence of misspecification.

Based on the above argument, we can construct the following method that solves the deductive computerization problem by addressing the above compatibility problem.

(step 1): Extend the working distribution F_w to a parametric model, say, $F_w(\delta)$, around F_w (i.e., so that $F_w(0) = F_w$), where δ is a finite dimensional vector. In this extension, we can keep unmodified the part of F_w that is known to be most reliably estimated (e.g., a propensity score elicited by physicians).

(step 2): Use the Gateaux numerical difference derivative

$$\begin{aligned}&\mathsf{Gateaux}\{\tau, F_w(\delta), D_i, \epsilon\}\\ &\quad := [\tau\{F_{w(D_i,\epsilon)}(\delta)\} - \tau\{F_w(\delta)\}]/\epsilon\end{aligned}$$

for a machine-small ϵ, to deduce the value of $\phi\{D_i, F_w(\delta)\}$ for arbitrary δ, and find

$$\widehat{\delta}^{opt} \text{ that minimizes the empirical variance of } \tau\{F_w(\widehat{\delta})\} \tag{9}$$

among all roots $\{\widehat{\delta}\}$ that solve the equation

$$\sum_i \left[\phi\{D_i, F_w(\widehat{\delta})\} \longleftarrow \mathsf{Gateaux}\{\tau, F_w(\widehat{\delta}), D_i, \epsilon\}\right] = 0, \tag{10}$$

where " $\longleftarrow$" means "computed as." Property (10) is the empirical analogue of the central, mean-zero property if the evaluated ϕ at $F_w(\widehat{\delta})$ is the true influence function of τ. An average of the EIF at a $F_w(\delta)$ that deviates from zero is evidence that the working distribution is incorrect. This step finds a distribution $F_w(\widehat{\delta})$ that eliminates such evidence. Technically, there may be no zeros, in which case $\widehat{\delta}$ can be defined as the minimizer of the absolute value of (10), although a better solution would be to make the model $F_w(\delta)$ more flexible (see below). More realistically, for a working model $F_w(\delta)$ there can be more than one zeros and so condition (9) selects the best one. Finally, although (9) is unambiguous if τ is a scalar, if τ is a vector then the researcher can minimize any one-dimensional criterion, such as, for example, the largest of the empirical variances of each of the components of $\tau\{F_w(\widehat{\delta})\}$.

(step 3): Calculate the parameter at the EIF-fitted distribution $F_w(\widehat{\delta})$ as

$$\widehat{\tau}^{\text{deductive}} := \tau\{F_w(\widehat{\delta}^{opt})\}. \tag{11}$$

3.2. *Properties*

The above method is deductive because step 2 does not need the functional form of ϕ, but deduces it by the numerical Gateaux derivative (6). If δ is one-dimensional, then (10) is expected to have one root, and this can be found by numerical root-finding methods such as in Brent (1973) or quasi Newton-Raphson, by finding and using the numerical difference derivatives with respect to δ of the Gateaux derivative computation of ϕ. If δ has more dimensions, then $\widehat{\delta}^{\text{opt}}$ can be found by either iterative quasi Newton-Raphson or by numerical Lagrange multipliers, where (9) can be coded as the jackknife variance. Also, the above estimates for τ and the remaining model parameters are compatible, by construction.

The deductive estimator shares useful properties of so-far known, nondeductive estimators that take ϕ as given. Notably, suppose the actual expectation of $\phi(D_i, F_w)$ is zero for a working distribution when, say $\text{part}_1(F_w) = \text{part}_1(F)$, or,. . . ,or $\text{part}_K(F_w) = \text{part}_K(F)$. Then, the deductive estimator above is expected to be consistent as would be usual, nondeductive estimators (e.g., Scharfstein et al. (1999)). For example, for the two-phase design, suppose an original working function $y_w(x)$ has been obtained as the OLS fit $x'\widehat{\beta}^{ols}$ of a linear regression model $x'\beta$ for $E(Y \mid R = 1, X = x)$. Then, a simple model extension is to add to $x'\widehat{\beta}^{ols}$ a free parameter δ (this is the same as freeing-up (again) the intercept of $x'\widehat{\beta}^{ols}$ and let it be a parameter). The subsequent implementation steps for deriving the estimator for the mean estimand are

given in Appendix A. It is then easy to show (proof omitted) that this deductive estimator is doubly robust (Scharfstein et al., 1999): it is consistent either if the propensity score working model $e_w(X_i)$ (corresponding to $\text{part}_1(F_w)$ above) is correct, or if the regression working model $y_w(x)$ (corresponding to $\text{part}_2(F_w)$ above) is correct.

Also, the deductive estimator above shares with the TMLE the idea of extending the working model (Chaffee and van der Laan (2011)), and with other estimators the idea of empirical maximization (e.g., Rubin and van der Laan (2008)). The conditions for the deductive estimator to use the smallest empirical variance are similar to those used in (Rubin and van der Laan, 2008, Appendix 2) and are omitted here because of their technical nature. To our knowledge, all such existing work for local efficiency has considered it critical to have the theoretically derived form of the EIF based on the score theory. The contribution of the proposed method above is to show that this theory can be translated to estimation that can be computerized in general, by combining model extension with the Gateaux derivative.

The extension in step 1 can take different forms. For example, for the two-phase design, one can also compute an improved deductive estimator by extending δ to two dimensions (e.g., two coefficients) and minimizing the empirical variance as in step 2. If the space of distributions spanned by the one-dimensional-based extended model lies within the space spanned by the two-dimensional extended model, then the estimator based on the latter will have empirical variance at most that of the former estimator because of the larger space where minimization takes place.

3.3. *Feasibility Evaluations*

To evaluate the feasibility of our method, we applied it to the study analyzed by Huang et al. (2005), as an example of the two-phase design. The goal of that study was to compare rates of patient satisfaction for asthma care as the outcome Y (yes/no) among different physician groups (treatments). Physician groups differed in their distribution of patient covariates. So, in order to compare between, say, two physician groups, we set the goal to estimate the average (3) of patient satisfaction for each group, standardized by the distribution of patient covariates in the combined population of the two groups. This standardization of estimands to the covariate distribution on all patients is also used in the literature, for example, for point exposure studies (e.g., Rosenbaum and Rubin (1983)); and is more commonly now known as g-computation (based on Robins (1986)) also for longitudinal studies. The following covariates X were considered: age, gender, race, education, health insurance, drug insurance coverage, asthma severity, number of comorbidities, and SF-36 physical and mental scores.

We tested feasibility of the above method for the comparison within two pairs of groups, denoted in Table 1(i) as a_1 versus b_1 and a_2 versus b_2 (actual names omitted). We chose (a_1, b_1) as a pair for which the usual estimator $\widehat{\tau}^{\text{nonde-}}_{\text{ductive}}$ produces values diverging from the unadjusted rates for a_1 and b_1; and we chose (a_2, b_2) as a pair for which the usual estimator produces values shrinking from a_1 and b_1. The nondeductive estimator used as propensity score the quintiles of the logistic regression of group membership conditionally on X; and a working expectation y_w as the prediction from the logistic regression of patient satisfaction conditionally on X within each group. The deductive estimator uses the same propensity score, and, for step 1 of the method, extended the working expectation y_w by including back the intercept in the logistic regression for each group as a free parameter δ. The computation of ϕ for each δ in (10) was obtained by straightforward numerical differentiation for the Gateaux derivative; and the root $\widehat{\delta}$ was found by the method of Brent (1973) implemented by the function "uniroot" in R. See Appendix A for further details.

In all cases in Table 1(i), the deductive estimator gives answers very close to the nondeductive estimator. This suggests that, for this problem and data, the usual doubly robust estimator, although not derived compatibly, can be re-expressed compatibly by the set of parameter values derived by the deductive estimator. We have also studied computability of the deductive estimator for the estimand defined as the mean restricted to the patients with propensity scores in $(0.1, 0.9)$ (Table 1(ii)). For this estimand, for which the usual doubly robust estimator is very close to the plain average, the deductive estimator is, again, very close to the usual nondeductive estimator. What is most important is that, although both estimators produced their answers in less than a second for each group and estimand, the deductive estimator did not need knowledge of the closed form expression (4) for ϕ, whereas the usual estimator depended critically on that knowledge.

4. Extensions

Close observation of the method for the deductive estimator for the mean in the two-phase design, as detailed in Appendix A, actually reveals how to produce a locally semiparametric efficient estimator also for any other estimand in this design. To see this, suppose we denote by $y_w(t;x)$ the cumulative distribution function $\text{pr}_w(Y \le t \mid X = x, R = 1)$ of Y for the working model. Then, by Bayes rule, we have that the cumulative distribution, say $\text{pr}_{w(d',\epsilon)}(Y \le t \mid X = x; R = 1)$, of Y in the perturbed distribution $F_{d',\epsilon}$ of (7) at $d' = (x', r', y'r')$, is

$$1(y' \le t)\frac{\epsilon \cdot 1\,(x = x', r' = 1)}{\epsilon \cdot 1\,(x = x', r' = 1) + (1-\epsilon)\cdot p_w\,(x)\,e_w\,(x)} + y_w\,(t;x)\,\frac{(1-\epsilon)\cdot p_w\,(x)\,e_w\,(x)}{\epsilon \cdot 1\,(x = x', r' = 1) + (1-\epsilon)\cdot p_w\,(x)\,e_w\,(x)}. \tag{12}$$

Based on this measure, implementation of steps 1–3 of Section 3 is relatively easy and generalizable. We have implemented this method in order to derive a locally efficient semiparametric estimator also for the median estimand in the two-phase design. This deductive estimator for the median, for which we are aware of no other implemented estimator, is given in Appendix B. We have conducted several simulation experiments (omitted) in all of which the deductive estimator is consistent also for this estimand. A comprehensive report on the small sample properties of the deductive estimator for

Table 1

Feasibility of the deductive method for estimating the probability of patient satisfaction adjusted for covariates for two physician group pairs using data from the asthma study of Huang et al. (2005).

(i) All patients			Estimates of $\tau(F) = \int_{x\in A} y(x)p(x)dx$, A : {all x}			
Physician group (g)	n	Unadjusted % $pr(Y=1 \mid G=g)$	$\widehat{\tau}^{\text{nonde-ductive}}$ (%)	se	$\widehat{\tau}^{\text{deductive}}$ (%)	se
a_1	171	62.0	63.1	4.5	63.3	4.4
b_1	81	58.0	52.0	8.8	51.9	8.9
a_2	104	78.8	72.1	8.2	71.6	8.0
b_2	189	47.6	49.4	4.5	49.4	4.4
(ii) Patients with increased common support			A : patients with $\widehat{e}(x) \in (0.1, 0.9)^{(1)}$			
Physician group (g)	n	Unadjusted % $pr(Y=1 \mid G=g)$	$\widehat{\tau}^{\text{nonde-ductive}}$ (%)	se	$\widehat{\tau}^{\text{deductive}}$ (%)	se
a_1	107	65.4	65.3	5.2	65.4	5.2
b_1	76	59.2	59.3	6.9	59.1	6.8
a_2	95	77.9	75.6	6.2	75.3	6.2
b_2	154	46.8	46.2	5.1	46.3	5.0

(1) This estimand with increased "common support" (e.g., Crump et al. (2009)), excludes here 64, 5, 9, and 35 patients from a_1, b_1, a_2, b_2, respectively.

the median and for other more challenging estimands is of interest for future study.

In complex problems, it is possible that standard root-finding methods for (10) are unstable. In this section, we show that the Gateaux numerical derivative may still be used to construct a deductive estimation method that does not rely on solving an estimating equation.

Suppose that the parameter $\tau(F)$ depends on F only through a set of variation-independent parameters $q_j(F) : j = 1, \ldots, J$. Such is the case of parameter (3) in our example, with $q_1(F; x) = y(x)$ and $q_2(F; x) = p(x)$. In an abuse of notation, let $\tau(q_1(F), \ldots, q_J(F)) := \tau(F)$. Since the parameters q_j are variation independent, the Gateaux derivative expression of ϕ in (6) reduces to

$$\phi(d', F) = \sum_{j=1}^{J} \lim_{\epsilon \to 0} \frac{\tau(q_1(F), \ldots, q_j(F_{d',\epsilon}), \ldots, q_J(F)) - \tau(F)}{\epsilon}.$$

This expression provides the decomposition $\phi(d', F) = \sum_{j=1}^{J} \phi_j(d', F)$, where ϕ_j is the nonparametric efficient score associated to q_j. Once the Gateaux numerical derivatives ϕ_j have been computed, it is possible to implement a standard TMLE without knowledge of the functional form of ϕ. We only provide a brief recap of the TMLE template, since extensive discussions are presented elsewhere (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). For each q_j, consider a loss function $L_j(q_j; D)$ whose expectation is minimized at the true value of q_j. Consider also a working model q_{jw} and a parametric extension $q_{jw}(\delta)$ satisfying

$$\frac{d}{d\delta} L(q_{jw}(\delta); d') \bigg|_{\delta=0} = \phi_j(d').$$

In our example, since q_j are components of the likelihood, the negative log-likelihood loss function and the exponential family may be used in this step:

$$L(q_j; d) = -\log q_j(d),$$
$$q_{jw}(\delta; d) \propto \exp(\delta \phi_j(d)) q_{jw}(d). \tag{13}$$

The TMLE is then defined by an iterative procedure that, at each step, estimates δ by minimizing the expected sum of the loss functions $L_j(q_{jw}(\delta); \cdot)$. An update of the working model is then computed as $q_{jw} \leftarrow q_{jw}(\widehat{\delta})$, and the process is repeated until convergence. The TMLE is defined by $\widehat{\tau} = \tau(q^{\star}_{1w}, \ldots, q^{\star}_{Jw})$, where $q^{\star}_{jw}$ denotes the estimate obtained in the last step of the iteration. Like the estimator presented in Section 3, the TMLE is a compatible estimator, and solves the EIF estimating equation. Unlike the estimator of Section 3, the TMLE does not require direct solution of that equation. However, the TMLE may be computationally more intensive, as it is iterative and may require numerical integration for computation of the proportionality constant in (13).

5. Remarks

We proposed a deductive method to produce semiparametric estimators that are locally efficient. The method does not rely on conjectures of tangent spaces and is not susceptible to possible errors in the verification of such conjectures. Instead, the new method relies on computability of the estimand τ for specified working distributions of the observed data F, and on numerical methods for differentiation and for root finding.

Although we have focused on local efficiency of originally unrestricted problems, one can see a path toward finding a deductive method also for problems with restrictions set a priori. Such a path can explore, first, nesting the restricted

problem within an unrestricted one, and then, making use of the proposed deductive method for the unrestricted problem, modified to impose numerically the nested restrictions. Such deductive methods can save dramatic amounts of human effort on essentially computerizable processes, and allow the transfer of that effort to other statistically demanding parts of the scientific process such as the efficient design of new studies.

6. Supplementary Materials

In Appendix C, which can be accessed at the *Biometrics* website on Wiley Online Library, we discuss a template for establishing large sample normality of the deductive estimator. Computer code and a run example data are available with this paper at the Biometrics website on Wiley Online Library. Instructions to use the methods of the article on deductive estimation can be found at: http://www.biostat.jhsph.edu/~cfrangak/papers/deduction

Acknowledgements

We thank the Editor, an Associate Editor, and two referees for helpful comments, and the NIH for partial financial support. The article has its seeds in part in critical discussions with Dr. Spyridon Kotsovilis on the scientific meaning of computability and insight, and has benefited by helpful discussions with Mark van der Laan, Michael Rosenblum, Daniel Scharfstein, Stijn Vansteelandt, and Kyrana Tsapkini.

References

Begun, J. M., Hall, W., Huang, W.-M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *The Annals of Statistics* **11**, 432–452.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Chaffee, P. and van der Laan, M. J. (2011). Targeted minimum loss based estimation based on directly solving the efficient influence curve equation. *UC Berkeley Division of Biostatistics Working Paper Series, Working Paper 287* .

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science* **20**, 261–301.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Huang, I., Frangakis, C., Dominici, F., Diette, G., and Wu, A. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* **40**, 253–278.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Newey, W. (1998). Undersmoothing and bias corrected functional estimation. *Massachusetts Institute of Technology, Department of Economics Working Paper, No. 98-17*.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modeling* **7**, 1393–1512.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. and van der Laan, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics* **4**, Article 5, DOI: 10.2202/1557–4679.1084.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Tsiatis, A. A. (2007). *Semiparametric Inference and Missing Data*. New York, NY: Springer.

Turing, A. (1937). On computable numbers, with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society* **42**, 230–265.

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer.

van der Laan, M. J. and Rubin, D. B. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**, Article 11, DOI:10.2202/1557–4679.1043.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Received June 2014. *Revised December* 2014. *Accepted January* 2015.

Appendix

Appendix A: Details for Deductive Estimation of the Mean with Working Model as in the Example

This section provides the details for how steps 1–3 of the general method of Section 3 are implemented in the data example given in that section.

(Preliminaries) : Coding of functions for the estimands at working and perturbed distributions.

First, a working distribution $F_w(\widehat{\beta})$ was specified as follows:

(i) the working distribution, $p_w(\cdot)$, of X, was taken to be the empirical distribution with point-mass $1/n$ at each

observed X_i (one can also assign weights other than $1/n$ for standardizing to different population);

(ii) the working propensity score, $e_w(\cdot)$, was taken to be the fit from a logistic regression;

(iii) the working outcome regression, $y_w(\cdot)$ for $E(Y \mid X = \cdot, R = 1)$, was taken to be the fit from the logistic regression:

$$y_w(x, \widehat{\beta}) = \text{expit}\{\widehat{\beta}_0 + \widehat{\beta}_1 x^{(1)} + \cdots + \widehat{\beta}_p x^{(p)}\}, \tag{A.1}$$

where $x = \left(x^{(1)}, \ldots, x^{(p)}\right)$ is p-dimensional covariate vector and expit is the inverse logit.

Then, functions were coded for the estimands $\tau\left\{F_w(\beta)\right\}$ and $\tau\left\{F_{w(D_i,\epsilon)}(\beta)\right\}$, i.e., the perturbation at the data point $D_i = (X_i, R_i, Y_i R_i)$ and arbitrary β, $0 < \epsilon < 1$. Based on the general formula (8) and the above working distributions, these functions are

$$\tau\left\{F_w(\beta)\right\} = \sum_{j=1}^{n} y_w(X_j, \beta)\, p_w(X_j), \tag{A.2}$$

$$\tau\left\{F_{w(D_i,\epsilon)}(\beta)\right\} = \sum_{j=1}^{n} y_{w(D_i,\epsilon)}(X_j, \beta)\, p_{w(D_i,\epsilon)}(X_j), \tag{A.3}$$

where the components of $F_{w(D_i,\epsilon)}(\beta)$ are derived using Bayes rule:

$$p_{w(D_i,\epsilon)}(x) = (1-\epsilon)\, p_w(x) + \epsilon \cdot 1(x = X_i),$$

$$y_{w(D_i,\epsilon)}(x, \beta) = \frac{\epsilon \cdot 1(x = X_i, R_i = 1)\, Y_i + (1-\epsilon) \cdot p_w(x)\, e_w(x)\, y_w(x, \beta)}{\epsilon \cdot 1(x = X_i, R_i = 1) + (1-\epsilon) \cdot p_w(x)\, e_w(x)}.$$

Then, steps 1–3 of Section 3 were implemented as follows.

(step 1): The extended working model $F_w(\delta)$ of Section 3 was defined by adding to $x'\widehat{\beta}$ a free parameter δ. Specifically, for a given δ, the extended working distribution, denoted here more precisely by $F_w(\widehat{\beta}_\delta)$, takes the working distributions for the covariate and for the propensity score as in the working models (i) and (iii), but takes the working regression $E(Y \mid X = x, R = 1)$ to be $y_w(x, \widehat{\beta}_\delta)$ (see (A.1)) where $\widehat{\beta}_\delta = \left(\delta + \widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)$. Note that here δ is 1-dim, and we have $F_w(\widehat{\beta}_\delta)|_{\delta=0} = F_w(\widehat{\beta})$.

(step 2): The empirical influence function is numerically computed and solved for its zero. To do this, this step starts with a candidate δ (say 0). Then,

(i) for a small ϵ, this step computes $\tau\{F_w(\widehat{\beta}_\delta)\}$ and $\tau\{F_{w(D_i,\epsilon)}(\widehat{\beta}_\delta)\}$ using the functions defined in (A.2)-(A.3), and hence computes the numerical derivative

$$\phi\{D_i, F_w(\widehat{\beta}_\delta)\} := \frac{\tau\{F_{w(D_i,\epsilon)}(\widehat{\beta}_\delta)\} - \tau\{F_w(\widehat{\beta}_\delta)\}}{\epsilon};$$

(ii) the sum $\sum_{i=1}^{n} \phi\{D_i, F_w(\widehat{\beta}_\delta)\}$ is computed for the candidate δ;

(iii) substeps (i)–(ii) above are repeated using the bisection method to find a $\widehat{\delta}$ such that the sum $\sum_{i=1}^{n} \phi\{D_i, F_w(\widehat{\beta}_{\widehat{\delta}})\}$ is 0 (note that because δ has dimension 1, there is no search to optimize the empirical variance).

(step 2): The estimate $\widehat{\tau}^{\text{deductive}}$ is computed using the function (A.2), giving

$$\widehat{\tau}^{\text{deductive}} := \tau\{F_w(\widehat{\beta}_{\widehat{\delta}})\}. \tag{A.4}$$

Appendix B: Deductive Estimation of the Median in the Two-Phase Design

This section describes how steps 1–3 of the general method of Section 3 are implemented to estimate the median outcome in the two-phase design, that is,

$$\tau := \text{median}(F) = \inf_t \left\{ t : \int \text{pr}(Y \le t \mid X = x, R = 1) p(x) dx \ge 0.5 \right\}, \tag{B.1}$$

where the last equality follows by ignorability in the two-phase design.

(Preliminaries) : Coding of functions for the estimands at working and perturbed distributions.

First, consider a working distribution $F_w(\widehat{\theta})$, with the working distribution, $p_w(\cdot)$, of X, and the working propensity score, $e_w(\cdot)$, as (i) and (ii) in Appendix A; and with

(iii') the working conditional distribution for the outcome given X to be the MLE fit from a normal regression $N(\widehat{\beta}_0 + \widehat{\beta}_1 x^{(1)} + \cdots + \widehat{\beta}_p x^{(p)}, \widehat{\sigma}^2)$, and denote the cumulative distribution by

$$y_w(t; x, \widehat{\theta}) := \text{pr}\left(Y \le t \mid X = x\cdot, R = 1, \widehat{\theta}\right), \tag{B.2}$$

where $\theta = (\beta, \sigma^2)$.

Then, the median $\tau\left\{F_w(\theta)\right\}$ and $\tau\left\{F_{w(D_i,\epsilon)}(\theta)\right\}$, i.e., the perturbation at the data point $D_i = (X_i, R_i, Y_i R_i)$ and arbitrary θ, $0 < \epsilon < 1$, can be easily derived based on the general formula (8) and the above working distributions, as

$$\tau\left\{F_w(\theta)\right\} = \inf_t \left\{\sum_{j=1}^{n} y_w(t; X_j, \theta)\, p_w(X_j) \ge 0.5\right\}, \tag{B.3}$$

$$\tau\left\{F_{w(D_i,\epsilon)}(\theta)\right\} = \inf_t\left\{\sum_{j=1}^{n} y_{w(D_i,\epsilon)}(t; X_j, \theta)\, p_{w(D_i,\epsilon)}(X_j) \geq 0.5\right\},$$

where the components of $F_{w(D_i,\epsilon)}(\theta)$ are derived using Bayes rule (similar argument to Appendix A)

$$p_{w(D_i,\epsilon)}(x) = (1-\epsilon)\, p_w(x) + \epsilon \cdot 1(x = X_i),$$

$$y_{w(D_i,\epsilon)}(t; x, \theta) = 1(Y_i \leq t)\frac{\epsilon \cdot 1(x = X_i, R_i = 1)}{\epsilon \cdot 1(x = X_i, R_i = 1) + (1-\epsilon)\cdot p_w(x)\, e_w(x)} + y_w(t; x, \theta)\frac{(1-\epsilon)\cdot p_w(x)\, e_w(x)}{\epsilon \cdot 1(x = X_i, R_i = 1) + (1-\epsilon)\cdot p_w(x)\, e_w(x)}. \quad \text{(B.4)}$$

Then, steps 1–3 of Section 3 were implemented as follows.

(step 1): The extended working model $F_w(\delta)$ of Section 3 was defined by freeing-up the intercept of $\widehat{\beta}$. Specifically, for a given δ, the extended working distribution, denoted here more precisely by $F_w(\widehat{\theta}_\delta)$, takes the working distributions for the covariate and for the propensity score as in the working model (i)-(ii), but takes the cumulative distribution $\text{pr}(Y \leq t \mid X = x, R = 1)$ to be $y_w(t; x, \widehat{\theta}_\delta)$ (see (B.2)) where $\widehat{\theta}_\delta = \left(\delta + \widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p, \widehat{\sigma}^2\right)$.

(step 2): The empirical influence function is numerically computed and solved for its zero in exactly the same way as in step 2 of Appendix A.

(step 3): The estimate $\widehat{\tau}^{\text{deductive}}$ is computed using the function (B.3), giving

$$\widehat{\tau}^{\text{deductive}} := \tau\{F_w(\widehat{\theta}_\delta)\}. \quad \text{(B.5)}$$

Note: Because (B.4) actually shows the full measure for Y under the extended working models, it can be used to compute, under these models, any estimand that can be computed based on the original working models. The above discussion, then, also serves to produce locally semiparametric efficient estimators for any other such estimand in this design.