Biometrics , 1–12

# Discussion of "Deductive derivation and Turing-computerization of semiparametric efficient estimation" by Frangakis et al.

**Alexander R. Luedtke[1,*,†], Marco Carone[2,**,†], and Mark J. van der Laan[1,***]**

[†]ARL and MC contributed equally to this work.

[1]Division of Biostatistics, University of California, Berkeley

[2]Department of Biostatistics, University of Washington

*email: aluedtke@berkeley.edu

**email: mcarone@uw.edu

***email: laan@berkeley.edu

Key words: Deductive procedure; Gâteaux derivative; Influence function; Nonparametric estimation

This paper has been submitted for consideration for publication in *Biometrics*

## Overview

In their paper, the authors consider the construction of asymptotically efficient substitution estimators, for which they propose a computerized implementation. We read their work with great interest and much enthusiasm. We have been strongly advocating the use of substitution estimators and we infer from the authors' emphasis on 'compatibility' that they also value the merits of such estimators. The proposed approach incorporates an important and innovative departure from currently available methods in that the analytic computation of the efficient influence function (EIF), an ingredient generally required for efficient estimation in infinite-dimensional models, is replaced by a computerized approximation. We agree with the authors that the ability to computerize this task is both interesting and useful, and we applaud them for initiating work on this important problem. The computerization of calculations currently done by hand will surely play a key role in the future of statistics.

We highlight here certain caveats pertaining to the proposal as it currently stands and opportunities that lie ahead. In particular, we note that:

(1) the numerical approximation procedure suggested by the authors does not always yield the EIF, even in the 'unrestricted problem';
(2) a regularized extension of their procedure will exhibit much greater applicability in practice;
(3) the authors restrict their attention to efficient estimation in nonparametric models, thus avoiding models in which the EIF is most difficult to derive and computerization would be most useful;
(4) more precise guidelines may be needed to ensure the first computerization-based estimator the authors propose yields the desired large-sample properties.

We elaborate on these points below. While we support the authors' push for computerization,

we believe it is also important to highlight the risks it poses. In particular, we wish to stress that

(5) the black-box nature of a computerized procedure can mask underlying theoretical difficulties and provide unsuspecting practitioners seemingly sensible yet invalid results.

We provide an example from our own research where (5) can occur quite dramatically.

In the remainder, we adopt in large part the notation of the authors, except that we use $Z$ to denote the data unit $D$ to avoid expressions such as $\int f(d)dd$.

## 1. When does the proposed numerical approximation procedure yield the EIF?

Suppose that the model is nonparametric and that the parameter of interest $\tau$ is pathwise differentiable with EIF at $F$ denoted by $\phi_F^*$ . Denote by $L_2^0(F)$ the Hilbert space of mean-zero square-integrable real functions defined on the support of $F$. Suppose also that $\tau$ is Gâteaux differentiable at $F$ along all paths of the form $F_{\epsilon,H} := (1-\epsilon)F + \epsilon H$ indexed by $0 \leqslant \epsilon \leqslant 1$ and with $H$ any distribution function. Furthermore, suppose that there exists an element $\delta_F \in L_2^0(F)$ such that the linear representation

$$\frac{d}{d\epsilon}\tau(F_{\epsilon,H})\bigg|_{\epsilon=0} = \int \delta_F(z)d\{H-F\}(z)$$

holds. Then, it must follow that $\frac{d}{d\epsilon}\tau(F_{\epsilon,z})\big|_{\epsilon=0} = \phi_F^*(z)$, where $F_{\epsilon,z} := (1-\epsilon)F + \epsilon 1\langle z\rangle$ with $1\langle z\rangle$ the distribution function of the degenerate distribution at $z$. Thus, under the above conditions, the numerical approximation procedure proposed by the authors indeed yields the EIF.

Many parameters of interest depend on local features of the data-generating distribution and may fail to satisfy the conditions above. For such parameters, the proposed Gâteaux derivative does not necessarily yield the EIF. As the authors have verified, their approach is valid in the examples considered in their paper. Nevertheless, a seemingly innocuous

modification of their motivating example illustrates that the approach is not generally applicable, even in the setting of the unrestricted problem on which they focus. Let $G_0$ be a given marginal distribution of $X$ dominated by the marginal distribution of $X$ under $F$, and consider the parameter $\tau(F) := \int y(x) dG_0(x)$, interpretable as a reference population-adjusted mean outcome. This parameter is defined for each $F$ for which the authors' motivating parameter (see Equation 3 in their paper) is defined. Furthermore, it is pathwise differentiable in a nonparametric model and has EIF at $F$ evaluated at observation $z := (x, r, y)$ given by

$$\phi_F^*(z) := \frac{r\{y - y(x)\}}{e(x)} \cdot \frac{p_0(x)}{p(x)} \ ,$$

where $p_0$ and $p$ are the density functions associated to the marginal distribution of $X$ under $G_0$ and $F$, respectively, relative to a common dominating measure $\mu$. We can verify that, whenever $\mu$ is the Lebesgue measure, $\frac{d}{d\epsilon}\tau(F_{\epsilon,z})\big|_{\epsilon=0} = 0 \neq \phi_F^*(z)$, indicating that the proposed Gâteaux derivative approach fails to yield the EIF. This is truly worrisome since there is no prima facie indication to suggest failure for this simple parameter and model. Clearly, additional conditions, such as those proposed above, must be imposed on a given parameter $\tau$ to ensure the validity of the proposed approach. In any given problem, the analytic verification of this condition may be (nearly) as difficult as simply computing the EIF by hand.

We note additionally that the proposed Gâteaux derivative is not defined for many parameters, even within a nonparametric model. As a simple illustration, consider the (nonparametric) model consisting of all absolutely continuous distributions, and suppose we wish to estimate the average density value

$$\psi(F) := \int f(z) dF(z) = \int f^2(z) dz \ ,$$

where $f$ is the Lebesgue density associated to the distribution function $F$. The parameter $\psi$ is pathwise differentiable and has EIF at $F$ given by $\phi_F^* := 2\,[f - \psi(F)]$. Furthermore, under

smoothness conditions on $f$, regular and asymptotically efficient estimators of $\psi(F)$ exist – see, e.g., Ibragimov and Hasminskii (1978), and Bickel and Ritov (1988). Nonetheless, the 'submodel' considered by the authors is not truly a submodel since the distributions that comprise it are not absolutely continuous. The parameter is thus ill-defined at each $F_{\epsilon,z}$ with $\epsilon \neq 0$. The computerized approach described therefore cannot be implemented even though the model is nonparametric. Although not sufficiently stringent to constrain the tangent space and yield a proper semiparametric model, the model constraints are strong enough to forbid the type of submodels used by the authors.

These examples highlight that the proposed Gâteaux derivative only provides a representation of the EIF under possibly strict conditions on the parameter and model considered. This is why the EIF is generally obtained via the more general concept of pathwise derivative: if the data-generating distribution $F$ is known only to lie in some model $\mathcal{M}$, the EIF $\phi_F^*$ is the unique element of the tangent space $T_{\mathcal{M}}(F)$ of $\mathcal{M}$ at $F$ such that

$$\left.\frac{d}{d\epsilon}\tau(F_\epsilon)\right|_{\epsilon=0} = \int \phi_F^*(z)h(z)dF(z)$$

for every score $h \in T_{\mathcal{M}}(F)$ and every regular one-dimensional regular parametric submodel $F_\epsilon$ through $F$ at $\epsilon = 0$ and with score $h$ for $\epsilon$ at $\epsilon = 0$. When the model is semiparametric, this implicit representation does not seem nearly as amenable to computerization as the explicit (but more narrowly applicable) Gâteaux representation the authors have used. However, as we argue below, it may be used to resolve the difficulties discussed so far in some nonparametric problems.

## 2. Broadening the applicability of the proposed approach via regularization

It is often possible to regularize the authors' proposal to yield the EIF when the model is nonparametric. Rather than taking a point mass-contaminated submodel, we will consider any regular one-dimensional parametric submodel $F_{\epsilon,z,\lambda}$ with score

$$\frac{dH_{z,\lambda}}{dF} - 1 \in L_2^0(F)$$

at $\epsilon = 0$ for some sequence of probability distributions $H_{z,\lambda}$ dominated by $F$ and symmetric about $z$ such that $\int g(u)dH_{z,\lambda}(u) \to g(z)$ as $\lambda \to 0$ for all $g$ in a large class $\mathcal{G}_z$ of functions (e.g., all functions continuous in a neighborhood of $z$). Often, we may simply choose the submodel $F_{\epsilon,z,\lambda} = (1-\epsilon)F + \epsilon H_{z,\lambda}$. We note that if $F$ places positive mass at $z$, the authors' original proposal of setting $H_{z,\lambda} := 1\langle z\rangle$ for all $\lambda$ falls in this framework. The pathwise differentiability of $\tau$ yields

$$\frac{d}{d\epsilon}\tau(F_{\epsilon,z,\lambda})\bigg|_{\epsilon=0} = \int \phi_F^*(z_0)\left[\frac{dH_{z,\lambda}}{dF}(z_0) - 1\right]dF(z_0) = \int \phi_F^*(z_0)dH_{z,\lambda}(z_0) \longrightarrow \phi_F^*(z)\ ,$$

as $\lambda$ tends to zero, where the limit holds provided $\phi_F^* \in \mathcal{G}_z$.

For the remainder of this section we assume that the component of $z$ that requires smoothing is univariate and absolutely continuous with respect to Lebesgue measure. For any real $w$ and $\lambda > 0$, we denote by $U_{w,\lambda}$ the distribution function of the uniform distribution over $(w-\lambda, w+\lambda)$. Below, we revisit the two examples discussed in the previous section.

In the average density value example, suppose that $F$ is absolutely continuous and let $H_{z,\lambda} = U_{z,\lambda}$. In this setting $\psi(F_{\epsilon,z,\lambda})$ is well-defined for each $\lambda > 0$ since then $F_{\epsilon,z,\lambda}$ is absolutely continuous. The Gâteaux derivative of $\psi$ at $F$ along this submodel can be computed using that

$$\frac{\psi(F_{\epsilon,z,\lambda}) - \psi(F)}{\epsilon} = \phi_F^*(z) + 2(1-\epsilon)\left[\frac{F(z+\lambda) - F(z-\lambda)}{2\lambda} - f(z)\right] + \epsilon\left[\frac{1}{2\lambda} + \tau(F) - 2f(z)\right]$$

with $\phi_F^*(z)$ the EIF of $\psi$ at $F$ and evaluated at $z$. For $\epsilon$ and $\lambda$ small,

$$\frac{\psi(F_{\epsilon,z,\lambda}) - \psi(F)}{\epsilon} \doteq \phi_F^*(z) + \frac{\epsilon}{2\lambda}\ .$$

It follows then that, for small $\lambda$ and much smaller $\epsilon$, this secant slope approximates $\phi_F^*(z)$, as desired. Here, the regularization allows us to circumvent the fact that the parameter is not defined along the point mass-contaminated submodels.

For the reference population-adjusted mean outcome, suppose that the covariate $X$ is a univariate continuous random variable and $Y$ is a binary outcome. If $H_{z,\lambda}$ is the distribution

where $(R, Y)$ equals $(r, y)$ with probability one and $X$ is drawn from $U_{x,\lambda}$, we have that

$$\frac{\tau(F_{\epsilon,z,\lambda}) - \tau(F)}{\epsilon} = \phi_F^*(z) + \left\{ \int \phi_F^*(z_1) dH_{z,\lambda}(z_1) - \phi_F^*(z) \right\} + \frac{\epsilon}{\lambda} \int \phi_F^*(x_1, r, y) \left[ \frac{\lambda \{2\lambda p(x_1) e(x_1) - 1\}}{2\lambda(1-\epsilon) p(x_1) e(x_1) + \epsilon} \right] dU_{x,\lambda}(x_1)$$

with $\phi_F^*(z)$ the EIF of $\tau$ at $F$ and evaluated at $z$. Under appropriate smoothness conditions and provided $e(x)p(x) > 0$, the secant slope approximates $\phi_F^*(z)$ once again for small $\lambda$ and much smaller $\epsilon$. The proposed regularization therefore allows recovery of the EIF in an example where, although the parameter is well-defined along the point mass-contaminated submodels, the Gâteaux derivative along these submodels does not yield the EIF. This result holds when the outcome $Y$ is continuous as well provided $H_{z,\lambda}$ also includes regularization in the $Y$ component.

This regularization strategy is somewhat subtle since its result depends on the relative rate at which the approximation parameters vanish. To ensure recovery of the EIF, this rate must satisfy a certain condition whose derivation required analytic work. We conjecture that, in a large class of problems, the condition $\epsilon = o(\lambda)$ will suffice when smoothing occurs in a single dimension. In any case, this regularization extends the applicability of the authors' proposal to more settings.

## 3. Restriction to nonparametric models

As they explicitly recognize in their paper, the authors have focused exclusively on efficient inference in an unrestricted model. It is important to emphasize this since applicability in the context of *semiparametric* models remains at this time unrealized.

In nonparametric models, the tangent space is by definition fully understood, and as such, there is no need to conjecture about its structure. The EIF is also the only possible gradient, and in our experience, its analytic computation is often – though not always, as we mention in the conclusion – rather straightforward. Intricacies generally arise in semiparametric models (i.e., models with restricted tangent spaces). As the authors suggest, efficient inference within

semiparametric models is often complicated because of the need to characterize the tangent space and project onto it. Indeed, analytic computation of the EIF is generally accomplished by identifying a gradient and projecting it orthogonally onto the tangent space, an often difficult endeavor. Computerization would be particularly appealing in such cases.

Addressing models which are semiparametric is an important outstanding problem in this realm. We are very much interested in such an extension, and certainly encourage further research along this line. We consider such an extension crucial for fulfilling the promise of computerized semiparametric efficient estimation.

## 4. Practical implementation of the proposed estimation approach

The authors propose two different deductive estimation strategies, both of which involve the construction of parametric submodels through an initial estimate of $F$ (or whichever portion of it may be relevant). In their first proposal, the authors suggest selecting among all distributions contained in a chosen submodel the one at which the approximated EIF has empirical mean closest to zero. However, no general guideline is provided on how a submodel should be chosen. There may be infinitely many distributions that solve the EIF estimating equation, most of which are inadequate estimates of $F$. If the construction of a revised estimate of $F$ only focuses on solving the EIF estimating equation, this estimate may have poor statistical properties. This matters because, as the authors indicate in their Supplementary Material, solving the EIF estimating equation does not suffice for asymptotic efficiency of the resulting substitution estimator: the revised estimator must still be a good estimator of $F$. As discussed in van der Laan and Rubin (2006) and implemented in Chaffee and van der Laan (2011), it is advantageous to adopt the least-favorable submodel when constructing the estimators described by the authors, though this may be more difficult to implement in practice using computerization.

The authors' second proposal is the computerization of a standard TMLE implementa-

tion: successive estimates are obtained by finding the minimizer of the empirical risk over an approximate least-favorable parametric submodel through the current estimate of $F$. Heuristically, TMLE preserves the statistical properties of the initial estimator of $F$ by ensuring that the empirical risk of successive estimates of $F$ never increases. As such, the TMLE algorithm is devised precisely to produce an estimate of $F$ that not only solves the EIF estimating equation but is also more likely to satisfy the regularity conditions required for asymptotic efficiency.

In view of this, it appears to us that (a regularization of) the authors' second proposal may be more likely to exhibit good statistical properties and yield desirable results in practice than the first proposal.

## 5. Potential for abuse inherent to computerized approaches

This last point is not a limitation of the authors' proposal per se but rather a call for caution applying to any procedure that automates the computation of the EIF. While the task of analytically deriving the EIF can be quite onerous, the exercise can be informative. For example, the process will often clarify when the parameter is pathwise differentiable, when the EIF has desirable properties (e.g., boundedness, robustness to misspecification), and under which regularity conditions the resulting estimator will be asymptotically linear and efficient. Invariably, the theoretical work involved translates directly into concrete guidelines for practice. Computerization would necessarily veil any such insights. At worst, this could result in the blind use of a supposedly efficient estimator even when it can be shown that no regular root-$n$ consistent estimator even exists. In such a case, the user may be completely oblivious to the invalidity of the inferences drawn. We describe an example arising in our research.

Suppose that $(X, R, Y) \sim F$, where $X$ denotes a patient's baseline covariate vector, $R$ is a binary treatment indicator and $Y$ an outcome of interest. The problem of estimating

the adjusted mean outcome $\tau(F) := E_F\left[E_F\left(Y|R = r_*(X), X\right)\right]$ under the optimal treatment rule $x \mapsto r_*(x)$, defined as

$$r_*(x) := I\left(E_F\left(Y|R = 1, X = x\right) > E_F\left(Y|R = 0, X = x\right)\right) \ ,$$

has been of much interest to investigators (see, e.g., Chakraborty and Moodie, 2013). It is known that if $P_F\left(E_F\left(Y|R = 1, X\right) = E_F\left(Y|R = 0, X\right)\right) > 0$, in which case $F$ is referred to as an *exceptional law*, this parameter is generally not pathwise differentiable (Robins and Rotnitzky, 2014). Nevertheless, the left- and right-sided pathwise derivatives may still exist along every smooth one-dimensional parametric submodel through $F$ (Hirano and Porter, 2012). This suggests that, even though it does not and clearly cannot yield the EIF since the latter does not exist, the Gâteaux derivative described by the authors exists. Interested readers can verify these facts in the simpler no-covariate scenario where $r_*$ is taken to be the marginally best treatment. Practitioners may therefore implement this approach and expect the resulting inference, known to be incorrect, to instead be valid. In the same spirit, the bootstrap is an example of how the simplicity of ready-to-use techniques can be both a curse and a blessing, empowering practitioners immensely all the while leading to widespread (and generally involuntary) abuse.

We note, additionally, that there may be no sensible approach for algorithmically detecting failures of pathwise differentiability in practice. In the example above, such attempts would necessarily fail. Because the investigator does not know a priori whether or not $F$ is an exceptional law, $F_w(\delta)$ may not be an exceptional law for any particular choice of $\delta$ – in fact, it may be non-exceptional for all $\delta$. The parameter $\tau$ may therefore be pathwise differentiable and admit an EIF at each $F_w(\delta)$. Under reasonable regularity conditions, any estimator solving the EIF estimating equation, including the proposed deductive approach, converges at root-$n$ rate but will suffer from non-negligible asymptotic bias (Luedtke and van der Laan, 2014). Most importantly, the proposed method will give no indication that the estimator is

asymptotically biased, and the user will have no apparent reason to distrust the resulting inference.

This example shows that the underlying assumption of pathwise differentiability is essential to the proposed computerized method and the regularization thereof. The verification of such an assumption can be challenging even analytically. In fact, we have had personal experience with an error in the elicitation of regularity conditions required for the pathwise differentiability of this parameter. There is certainly an appeal for an automated method for establishing regularity conditions under which the parameter is pathwise differentiable at the data-generating distribution. Though beyond the scope of the current work, this is an important area for future research.

## Concluding remarks

Again, we commend the authors for initiating what is sure to be an exciting line of research on the role of computerization in statistics. We have noted that the authors' proposed numerical approximation does not always yield the EIF, even in the unrestricted problem. Nonetheless, we have argued that a regularization of their approach appears to mostly retain the deductive nature of the original proposal, despite requiring some additional regularity conditions to work in practice.

Even though analytic computation of the EIF is straightforward in the examples the authors have considered and computerization is therefore not needed, we appreciate the immense pedagogical value of their simple illustrations. Even in the context of the unrestricted problem the authors have focused on, there are important problems in which deductive estimation could be quite valuable in practice. One such problem concerns efficient estimation of a bivariate survival function when the underlying pair of failure times is subject to bivariate censoring known only to satisfy coarsening at random (Gill et al., 1997). In this problem, the NPMLE exists but is inconsistent (Tsai et al., 1986). Although the induced model for

the observed data is nonparametric, the EIF does not exist in closed form, rendering the implementation and analysis of an efficient estimator challenging (van der Laan, 1996). As such, this would seem to be a particularly interesting test case for the proposed method (or any regularization thereof). In our opinion, validation of the deductive method in the context of this problem could serve as truly compelling advertisement for the approach.

We believe it is neither yet within reach nor desirable to eliminate the analytic calculations involved in the construction of efficient estimators. For example, without these calculations, the verification of general regularity conditions (e.g., as listed by the authors in their Supplementary Material) seems challenging for any deductive method, and so, such a method could be prone to misuse. Further, the authors' underlying assumption of pathwise differentiability could also be violated in problems of interest. Nevertheless, we remain excited by the stimulating conversation the authors have initiated with their work. We see in practice much value to computerized efficient estimation, particularly in tandem with suitable theoretical work, and we look forward to further research on this topic. Such research could, for example, elucidate how computerized methods may be used to verify analytic computations of the EIF, or how they may flag scenarios in which further analytic consideration would be especially warranted, particularly when computation of the EIF is intractable.

**Acknowledgements**

ARL was supported by the NDSEG Fellowship Program of the U.S. Department of Defense. MC was supported by a Genentech Endowed Professorship at the University of Washington. MJvdL was supported by NIH grant R01 AI074345-06.

**References**

Bickel, P. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā* **50,** 381–393.

Chaffee, P. and van der Laan, M. (2011). Targeted minimum loss based estimation based

on directly solving the efficient influence curve equation. Technical report, UC Berkeley Department of Biostatistics Working Paper Series.

Chakraborty, B. and Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Gill, R., van der Laan, M., and Robins, J. (1997). Coarsening at random: characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer.

Hirano, K. and Porter, J. (2012). Impossibility results for nondifferentiable functionals. *Econometrica* **80,** 1769–1790.

Ibragimov, R. and Hasminskii, I. (1978). On the non-parametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, pages 41–52.

Luedtke, A. and van der Laan, M. (2014). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Technical report, UC Berkeley Division of Biostatistics Working Paper Series.

Robins, J. and Rotnitzky, A. (2014). Discussion of "Dynamic treatment regimes: Technical challenges and applications". *Electronic Journal of Statistics* **8,** 1273–1289.

Tsai, W., Leurgans, S., and Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *The Annals of Statistics* **14,** 1351–1365.

van der Laan, M. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics* **24,** 596–627.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. Technical report, UC Berkeley Department of Biostatistics Working Paper Series.