Published in final edited form as:

J Am Stat Assoc. 2019 ; 114(527): 1174–1190. doi:10.1080/01621459.2018.1482752.

Toward computerized efficient estimation in infinite-dimensional models

Marco Carone¹, Alexander R. Luedtke², Mark J. van der Laan³

¹Department of Biostatistics, University of Washington

²Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

³Division of Biostatistics, University of California, Berkeley

Abstract

Despite the risk of misspecification they are tied to, parametric models continue to be used in statistical practice because they are simple and convenient to use. In particular, efficient estimation procedures in parametric models are easy to describe and implement. Unfortunately, the same cannot be said of semiparametric and nonparametric models. While the latter often reflect the level of available scientific knowledge more appropriately, performing efficient inference in these models is generally challenging. The efficient influence function is a key analytic object from which the construction of asymptotically efficient estimators can potentially be streamlined. However, the theoretical derivation of the efficient influence function requires specialized knowledge and is often a difficult task, even for experts. In this paper, we present a novel representation of the efficient influence function and describe a numerical procedure for approximating its evaluation. The approach generalizes the nonparametric procedures of Frangakis et al. (2015) and Luedtke et al. (2015) to arbitrary models. We present theoretical results to support our proposal, and illustrate the method in the context of several semiparametric problems. The proposed approach is an important step toward automating efficient estimation in general statistical models, thereby rendering more accessible the use of realistic models in statistical analyses.

Keywords

asymptotic efficiency; canonical gradient; efficient influence function; nonparametric and semiparametric models; pathwise differentiability

1 Introduction

1.1 Efficient estimation of smooth target parameters

Efficient estimation techniques are appealing because they maximally exploit available information and minimize the uncertainty of the resulting scientific findings. Efficiency is most often studied in an asymptotic sense due to its intractability in many finite-sample

problems. Characterizing asymptotic efficiency and constructing asymptotically efficient estimators has been an important focus of methodological and theoretical research in statistics. For convenience, from here and on, we ascribe without explicit mention an asymptotic sense to the terms *efficient* and *efficiency*

In the context of parametric models, a simple efficiency theory has been available for nearly a century, largely fueled by Fisher's work on maximum likelihood estimation. In such models, efficiency is characterized by the Crámer-Rao bounds and efficient estimators can generally be obtained via maximum likelihood (e.g., Hájek, 1970, 1972; Le Cam, 1972). When parametric models are adopted in practice, it is often because they are simple and convenient to use. However, the use of such models carries the risk of model misspecification, which may adversely affect the scientific process. In many scientific problems, the available background knowledge simply does not justify the use of such restrictive statistical models

Infinite-dimensional models – either nonparametric or semiparametric – offer a more flexible alternative. These richer models mitigate the potential for model misspecification and often more accurately reflect the level of available prior knowledge. Unfortunately, establishing efficiency bounds for target parameters in infinite-dimensional models can be a complex task. The development of a general efficiency theory, valid for arbitrary statistical models, is a more recent accomplishment: except for the early seminal contribution of Stein (1956), developments in this area began in the late 1970s and early 1980s with the works of Koshevnik and Levit (1977), Pfanzagl (1982) and Begun et al. (1983), among others, and continued throughout the 1990s (e.g., van der Vaart, 1991; Newey, 1994). Notably, it builds upon notions of differential geometry and functional analysis. In certain cases, a generalized notion of maximum likelihood, as described by Kiefer and Wolfowitz (1956), for example, can still be used to produce efficient estimators of certain parameters – see, for example, Wong and Severini (1991) for a general study under broad conditions. In other cases though, the statistical model or target parameter may be too complex for a maximum likelihood estimator to exist, let alone be well-behaved

Several approaches have been proposed for efficient estimation in the context of infinite-dimensional models. In problems in which the likelihood approach is problematic, some authors have successfully repaired the likelihood function (e.g., van der Laan, 1996; Maathuis and Wellner, 2008). Regularization techniques, including penalization and the method of sieves, have proven useful for devising efficient estimators (e.g., Shen, 1997; Newey, 1997; Chen, 2007). General approaches for efficient estimation based upon knowledge of the efficient influence function have been developed and extensively studied (e.g., Pfanzagl, 1982; Bickel et al., 1997; van der Laan and Robins, 2003; van der Laan and Rubin, 2006). These flexible approaches have been used in the past decade to perform valid inference incorporating a variety of machine learning tools. In particular, they have been employed to conduct robust and efficient inference via ensemble learning, wherein nuisance functions are estimated using a data-driven amalgamation of many learning strategies. The latter can range from simple parametric techniques to complex machine learning algorithms, and can include penalization and sieve-based learning methods. A variety of illustrations are provided in van der Laan and Rose (2011, 2018), for example

1.2 The efficient influence function and its role in efficient estimation

The efficient influence function, hereafter referred to as the EIF, is a key object in efficiency theory for infinite-dimensional models. It bears this name because it is the influence function of any efficient estimator of the parameter of interest relative to the statistical model with respect to which it was calculated. As such, the efficiency bound for this parameter, and thus the asymptotic variance of any efficient estimator, can be derived from the EIF. Knowledge of the EIF has important implications for practice. First, the relative efficiency of a candidate estimator can be assessed objectively by comparing its variance to the efficiency bound implied by the EIF. Second, using an EIF-based estimate of the efficiency bound, asymptotically valid Wald confidence intervals can be constructed using any available efficient estimator, thereby circumventing the need to rely on bootstrap schemes that may either be invalid or computationally infeasible in a given problem. Third, knowledge of the EIF can assist investigators in planning the design of their study by helping quantify the effect of different sampling strategies on the variability of any efficient estimator used. Concrete examples include the optimal selection of subgroups on which to process resource-intensive measurements in two-phase sampling designs (see, e.g., Gilbert et al., 2014) and the optimal design of group-sequential adaptive trials (e.g., Chambaz and van der Laan, 2014). Fourth, as it is the influence function of any efficient estimator, the EIF can be used to investigate the robustness of efficient estimators (e.g., Hampel et al., 2011). For example, the EIF can be used to assess the approximate (i.e., asymptotic) contribution of individual observations to the value of an efficient estimator. Finally, and perhaps most importantly, as suggested above, if the analytic form of the EIF is available, efficient estimators can themselves be constructed rather easily. To do so, several approaches may be used, including, for example, gradient-based estimating equations (e.g., van der Laan and Robins, 2003; Tsiatis, 2007), Newton-Raphson one-step corrections (e.g., Pfanzagl, 1982) and targeted minimum loss-based estimation (e.g., van der Laan and Rose, 2011). Thus, there is a strong motivation for deriving the EIF in a given statistical problem

Unfortunately, the analytic computation of the EIF is seldom straightforward. It generally involves finding a candidate influence function and projecting it onto the tangent space of the statistical model, which must itself be characterized – the effort can be mathematically intricate. Over the years, many techniques have been developed to facilitate this task in certain classes of problems – the discretization technique of Chamberlain (1987) is one such example. Nevertheless, this calculation remains a rather specialized skill, mastered mostly by a small collection of theoretically-inclined researchers. Worse still, in some problems, the EIF does not even have a closed form, rendering its evaluation and use particularly difficult (e.g., Geskus and Groeneboom, 1999; Quale et al., 2006). The theoretical derivation of EIFs is generally not in the skill set of practicing statisticians. Yet, in many problems, it is a skill needed to facilitate optimal inference in more realistic statistical models. The paucity of this skill may have impeded a broader appreciation and adoption of semiparametric and nonparametric techniques in applications

In view of this barrier, one naturally wonders whether a suitable numerical approximation could serve as substitute for the analytic form of the EIF, and further whether its calculation could be computerized. An affirmative answer to this question would render the

implementation of efficient inferential techniques in semiparametric and nonparametric models much more accessible to practitioners, and the impact on statistical practice could be significant. Recently, an important first step toward this goal was made by Frangakis et al. (2015): these authors proposed a simple numerical routine for calculating the EIF in the context of nonparametric models when the data are discrete-valued or when the parameter is a smooth functional of the distribution function. In our discussion of their article (see Luedtke et al., 2015), we suggested a regularization of their technique that is valid more broadly within the context of nonparametric models. Nevertheless, neither of these methods formally address the more difficult problem of numerically computing the EIF in semiparametric models. As opposed to nonparametric models, for which the tangent space is trivially described, semiparametric models generally have much more complex tangent spaces, projecting onto which may often require great experience. Identifying a numerical approach for computing the EIF in semiparametric models is therefore a more difficult but also more needed innovation

1.3 Contributions and organization of this article

In this article, we establish and study a novel representation of the EIF that directly generalizes the non-parametric results of Frangakis et al. (2015) and Luedtke et al. (2015) to arbitrary models. As we illustrate through several examples, this representation can be used to calculate either analytically or numerically the EIF of a given parameter in a given statistical model without the specialized knowledge traditionally required for this task. As such, it holds great promise in facilitating the computerization of efficient estimation not only in nonparametric models but also in semiparametric models, as we discuss below

This paper is organized as follows. In Section 2, we present our novel representation of the EIF for use in arbitrary statistical models and present high-level conditions under which it holds. In Section 3, we study each of these conditions and determine sufficient conditions under which the validity of this representation is guaranteed. We explore various practical issues regarding the implementation of our proposal and provide guidelines for use in Section 4. In Section 5, we illustrate the use of the approach in the context of four semiparametric problems. We provide concluding remarks in Section 6. While Theorem 1 is proved in the body of the paper, the proof of Theorems 2–5 are provided in an Appendix

2 Numerical calculation of the efficient influence function

2.1 Preliminaries

Suppose that we observe independent d -dimensional variates $X_1, X_2, \dots, X_n$ following a distribution P_0 known only to belong to the statistical model $\mathcal{M}$. We denote by $\mathcal{X}(P) \subseteq \mathbb{R}^d$ the sample space associated to $P \in \mathcal{M}$. We are interested in efficiently inferring about $\psi_0 := \Psi(P_0)$ using the available data, where $\Psi: \mathcal{M} \rightarrow \mathbb{R}^q$ represents a pathwise differentiable parameter mapping of interest. Pathwise differentiability ensures the parameter is a sufficiently smooth mapping so as to admit an efficiency theory (see, e.g., Pfanzagl, 1982; Bickel et al., 1997). We denote by $L_2^0(P)$ the Hilbert space of P -integrable functions from $\mathcal{X}(P)$ to $\mathbb{R}^q$ with mean zero and finite variance under P . The parameter Ψ is said to be

pathwise differentiable if there exists some $\chi_P \in L_2^0(P)$ such that, for each regular one-dimensional parametric submodel $\mathcal{M}_0 := \{P_\epsilon: \epsilon \in \mathcal{E}\} \subseteq \mathcal{M}$ with $\mathcal{E} \subset \mathbb{R}$ an interval containing zero and $P_{\epsilon=0} = P$, the pathwise derivative $\frac{d}{d\epsilon}\Psi(P_\epsilon)|_{\epsilon=0}$ can be represented as the inner product $\int \chi_P(u)s(u)dP(u)$, where s is the score for ϵ at $\epsilon=0$ in $\mathcal{M}_0$ (Pfanzagl, 1982). Any such element χ_P is said to be a gradient of Ψ at P relative to $\mathcal{M}$. The tangent space $T_{\mathcal{M}}(P) \subseteq L_2^0(P)$ of $\mathcal{M}$ at P is defined as the closure of the linear span of scores at P arising from regular one-dimensional P -dominated parametric submodels of $\mathcal{M}$ through P . The canonical gradient is the unique gradient contained in $T_{\mathcal{M}}(P)$ and corresponds to the EIF under sampling from P . Throughout, we will refer to the EIF at P as ϕ_P and write $\phi_P(x)$ for the evaluation of ϕ_P at the observation value x . The asymptotic variance of an efficient estimator of ψ_0 relative to model $\mathcal{M}$ is given by $\int \phi_{P_0}(u)\phi_{P_0}(u)^\top dP_0(u)$. Without loss of generality, we will assume $q=1$ as the general case can be dealt with using the developments herein applied to each component

If pathwise differentiability holds uniformly over paths in a neighborhood around P , for any $P_1 \in \mathcal{M}$ close enough to P , the parameter admits the linearization

$$\begin{aligned}\Psi(P_1) - \Psi(P) &= \int \phi_{P_1}(u)d(P_1 - P)(u) + R(P_1, P) \\ &= - \int \phi_{P_1}(u)dP(u) + R(P_1, P)\end{aligned}\tag{2.1}$$

where $R(P_1, P)$ is a second-order remainder term, and the second line follows since $\int \phi_{P_1}(u)dP_1(u) = 0$. This representation, which is no more than a first-order Taylor approximation over the model space, holds for most smooth parameters arising in practice. The precise form of R is generally established by hand on a case-by-case basis. This linearization is critical for motivating and studying the use of both Newton-Raphson one-step correction and targeted minimum loss-based estimation to construct efficient estimators. It is also at the heart of our current proposal for obtaining a numerical approximation to the EIF value $\phi_P(x)$ at a given distribution $P \in \mathcal{M}$ and observation value $x \in \mathcal{X}(P)$.

2.2 Nonparametric models

Recently, Frangakis et al. (2015) presented one such proposal based on the representation of $\phi_P(x)$ as the Gateaux derivative of Ψ at P in the direction of $\delta_x - P$, where δ_x represents the degenerate distribution at x . Of course, this can also be seen as the pathwise derivative $\frac{d}{d\epsilon}\Psi(P_\epsilon)|_{\epsilon=0}$ of Ψ at P along the linear perturbation path $\{P_\epsilon := (1 - \epsilon)P + \epsilon\delta_x: 0 \leq \epsilon \leq 1\}$ between P and δ_x —this simple observation will be helpful when tackling arbitrary models. Here and throughout, any such derivative is of course interpreted as a right derivative. To computerize the process of calculating $\phi_P(x)$, these authors suggested approximating this derivative by the slope of the secant line through $(0, \Psi(P))$ and $(\epsilon, \Psi(P_\epsilon))$ for a small $\epsilon > 0$. In our discussion of Frangakis et al. (2015), we pointed out sufficient conditions that guarantee that this indeed approximates $\phi_P(x)$ (Luedtke et al., 2015). For example, this approach is valid whenever the model $\mathcal{M}$ is nonparametric and the sample space $\mathcal{X}(P)$ is

finite. However, if the parameter Ψ depends on local features of the distribution, this method may fail when $\mathcal{X}(P)$ is infinite, such as when any component of X is continuous under P . We proposed a slight modification of the procedure of Frangakis et al. (2015) to remedy this limitation. Specifically, we proposed replacing the degenerate distribution δ_x at x by a distribution $H_{x,\lambda}$ dominated by P and such that $\int g(u) dH_{x,\lambda}(u) \rightarrow g(x)$ as $\lambda \rightarrow 0$ for all g in a sufficiently large class of functions. This amounts to replacing the degenerate distribution by a nearly degenerate distribution with smoothing parameter $\lambda > 0$. In a technical report published contemporaneously, Ichimura and Newey (2015) also suggested this approach. As stated in Luedtke et al. (2015), under certain regularity conditions and provided M is nonparametric, it is generally the case that

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \left[\frac{d}{d\epsilon} \Psi(P_{\epsilon,\lambda}) \Big|_{\epsilon=0} \right] \quad (2.2)$$

where we have defined the linear perturbation path. $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$ Representation (2.2) is useful when the parameter is simple enough so that calculating the derivative of $\epsilon \mapsto \Psi(P_{\epsilon,\lambda})$ is analytically convenient. Otherwise, similarly as in Frangakis et al. (2015), the derivative in (2.2) can be approximated, for example, by the secant line slope

$$\frac{\Psi(P_{\epsilon,\lambda}) - \Psi(P)}{\epsilon}$$

for small ϵ and λ . This operation only requires the ability to evaluate Ψ on a given distribution. Generally, as we highlighted in Luedtke et al. (2015), ϵ must be chosen much smaller than λ to obtain an accurate approximation – this emphasizes that the limit and derivative operators in (2.2) cannot generally be interchanged. We discuss this point in detail later. Because representation (2.2) is a special case of the general result we present below, we defer a statement of regularity conditions and a formal proof until then

2.3 Arbitrary models

When the model M is not nonparametric, representation (2.2) generally does not hold. Except when $\epsilon = 0$, the linear perturbation path described by $P_{\epsilon,\lambda}$ is not necessarily contained in the model. As such, the parameter may not even be defined on this path. Even if it is, the approximation suggested by this representation generally only yields the EIF of Ψ relative to a nonparametric model rather than the actual model. While the nonparametric EIF is still an influence function in $\mathcal{M}$, it is not typically efficient relative to $\mathcal{M}$. This should not be surprising since the expression in (2.2) in no way acknowledges constraints implied by $\mathcal{M}$.

The linear perturbation approach of Frangakis et al. (2015) and Luedtke et al. (2015) does not generally lead to the EIF in semiparametric and parametric models. Nevertheless, a first-order analysis of the parameter mapping is helpful to determine conditions under which, in arbitrary models, the EIF can be expressed in terms of a pathwise derivative over some perturbation path. Below, we denote by $P_{\epsilon,\lambda}^* \in \mathcal{M}$ a candidate perturbation of P determined by a choice of $H_{x,\lambda}$ and ϵ . High-level conditions under which $P_{\epsilon,\lambda}^*$ serves as an adequate

perturbation are outlined in the theorem below. Here and throughout, we use the shorthand notation $\phi_{\epsilon, \lambda}^*$ to denote the evaluation $\phi_{P_{\epsilon, \lambda}^*}$ of the EIF at $P_{\epsilon, \lambda}^*$

Theorem 1.

The EIF of Ψ relative to $\mathcal{M}$ at $P \in \mathcal{M}$ evaluated at observation value x is given by

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \left[\frac{d}{d\epsilon} \Psi(P_{\epsilon, \lambda}^*) \right]_{\epsilon=0} \quad (2.3)$$

provided the following conditions hold:

- (A1) (near-solution of EIF estimating equation)
 $\lim_{\epsilon \rightarrow 0} \lim_{\lambda \rightarrow 0} \int \phi_{\epsilon, \lambda}^*(u) dP_{\epsilon, \lambda}(u) / \epsilon = 0;$
- (A2) (continuity of EIF) $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \int \phi_{\epsilon, \lambda}^*(u) d(H_{x, \lambda} - P)(u) = \phi_P(x)$
- (A3) (preservation of convergence rate) $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} R(P_{\epsilon, \lambda}^*, P) / \epsilon = 0.$

Proof. Setting $P_1 = P_{\epsilon, \lambda}^*$ in (2.1), we note that

$$\begin{aligned} \Psi(P_{\epsilon, \lambda}^*) - \Psi(P) &= - \int \phi_{\epsilon, \lambda}^*(u) dP(u) + R(P_{\epsilon, \lambda}^*, P) \\ &= \int \phi_{\epsilon, \lambda}^*(u) d(P_{\epsilon, \lambda} - P)(u) - \int \phi_{\epsilon, \lambda}^*(u) dP_{\epsilon, \lambda}(u) + R(P_{\epsilon, \lambda}^*, P). \end{aligned}$$

The result follows from (A1), (A2) and (A3) upon noting that, since, $P_{\epsilon, \lambda} - P = \epsilon(H_{x, \lambda} - P)$, we have that

$$\frac{\Psi(P_{\epsilon, \lambda}^*) - \Psi(P)}{\epsilon} = \int \phi_{\epsilon, \lambda}^*(u) d(H_{x, \lambda} - P)(u) - \frac{\int \phi_{\epsilon, \lambda}^*(u) dP_{\epsilon, \lambda}(u)}{\epsilon} + \frac{R(P_{\epsilon, \lambda}^*, P)}{\epsilon}.$$

□

We must now determine a strategy for constructing a perturbation $P_{\epsilon, \lambda}^*$, that can be expected to satisfy (A1)–(A3). Although the linear path $\{P_{\epsilon, \lambda} : 0 \leq \epsilon \leq 1\}$ is generally inadequate, it appears natural to consider the path determined by the projection $P_{\epsilon, \lambda}^*$ of $P_{\epsilon, \lambda}$ into $\mathcal{M}$, or a suitably regularized version thereof, according to an appropriate divergence. For any $P_1 \in \mathcal{M}$, we define $\mathcal{M}(P_1) := \{P_2 \in \mathcal{M} : P_2 \ll P_1\} \subseteq \mathcal{M}$ as the subset of all probability measures in $\mathcal{M}$ dominated by P_1 , and let $\mathcal{M}_{\text{NP}} \supseteq \mathcal{M}$ represent a nonparametric model containing $\mathcal{M}$. Then, we formally define a novel perturbation path in the model as

$$P_{\epsilon, \lambda}^* := \underset{P_1 \in \mathcal{M}(P)}{\operatorname{argmin}} \mathbb{D}(P_1, P_{\epsilon, \lambda}), \quad (2.4)$$

where the function $\mathbb{D} : \mathcal{M}_{\text{NP}} \times \mathcal{M}_{\text{NP}} \rightarrow [0, +\infty]$ satisfies the following conditions:

- (B1) for each $P_1, P_2 \in \mathcal{M}$, $\mathbb{D}(P_1, P_2) \geq 0$, and $\mathbb{D}(P_1, P_2) = 0$ if and only if $P_1 = P_2$

- (B1)** for any $P_2 \in \mathcal{M}$ and $P_1 \in \mathcal{M}(P_2)$ such that the Radon-Nikodym derivative of P_1 relative to P_2 is uniformly bounded in an interval $[m_0, m_1] \subset (0, +\infty)$, provided $P_1 \neq P_2$, the inequality

$$K_0(m_0, m_1) \leq \frac{\mathbb{D}(P_1, P_2)}{\int \left[\sqrt{\frac{dP_1}{dP_2}(u)} - 1 \right]^2 dP_2(u)} \leq K_1(m_0, m_1)$$

holds for constants $0 < K_0(m_0, m_1) < K_1(m_0, m_1) < +\infty$ depending on m_0 and m_1 alone

- (B3)** for each $P_2 \in \mathcal{M}_{NP}$ and each $K \in (0, +\infty)$, setting $P_2^* := \operatorname{argmin}_{P_1 \in \mathcal{M}(P_2)} \mathbb{D}(P_1, P_2)$, provided the Radon-Nikodym derivative of P_2^* relative to P_2 is uniformly bounded in an interval $[m_0, m_1] \subset (0, +\infty)$, there exists a map $B: [0, +\infty) \rightarrow [0, +\infty)$ possibly depending on (K, m_0, m_1) and with $\lim_{\delta \rightarrow 0} B(\delta) = 0$ for which

$$\left\| \frac{dP_2^*}{dP_2} - 1 \right\|_{2, P_2} \leq \delta \text{ implies that } \left| \int h(u) dP_2(u) \right| \leq \delta B(\delta)$$

for each $h \in T_{\mathcal{M}}(P_2^*)$ with $\|h\|_{\infty} < K$.

Condition (B1) simply requires $\mathbb{D}$ to be a divergence function. Condition (B2) allows us to relate values of this divergence and relevant L_2 norms. Condition (B3) is used to ensure condition (A1). Heuristically, it requires that, under a given distribution, elements of the tangent space at the projection of this distribution have mean nearly zero if the distribution is close enough to the model. Additional details are provided below.

At least two of the most important divergences fit into this general setup and satisfy all conditions: the Kullback-Leibler (KL) divergence and the Hellinger distance, corresponding respectively to

$$\mathbb{D}_{KL}: (P_1, P_2) \mapsto - \int \log \left[\frac{dP_1}{dP_2}(u) \right] dP_2(u) \text{ and } \mathbb{D}_H: (P_1, P_2) \mapsto \int \left[1 - \sqrt{\frac{dP_1}{dP_2}(u)} \right]^2 dP_2(u).$$

The fact that these particular divergences indeed satisfy (B1)–(B3) is established in the Supplementary Material. The KL divergence appears to be the most natural choice for several reasons. First, for this choice of divergence, for any given $P_2 \in \mathcal{M}_{NP}$ and any $h \in T_{\mathcal{M}}(P_2^*)$ with P_2^* the projection of P_2 into $\mathcal{M}(P_2)$, the equality $\int h(u) dP_2(u) = 0$ can be shown to hold exactly. To see this, suppose that h is in the interior of $T_{\mathcal{M}}(P_2^*)$ and construct the submodel $dP_{2,\eta}^* = (1 + \eta h) dP_2^*$ for η in a neighborhood of zero. Because P_2^* is a global minimizer of the KL divergence, we have that

$$0 = \frac{d}{d\eta} \int \log \left[\frac{dP_{2,\eta}^*}{dP_2}(u) \right] dP_2(u) \Big|_{\eta=0} = \int h(u) dP_2(u).$$

If h is instead on the boundary of $T_{\mathcal{M}}(P_2^*)$, the argument above can be repeated for a sequence of interior elements tending to h . Thus, for the KL divergence, an even stronger property than (B3) holds, and it implies in particular that $\int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) = 0$ for each ϵ and λ . As a consequence, a stronger version of (A1) holds – the EIF estimating equation is in fact exactly solved – which could improve the resulting approximation of the EIF in practice. Second, because of its central role in likelihood methods, the KL divergence has been studied extensively, and many strategies for likelihood maximization have been described in the statistical literature. Whether other appropriate divergences, such as the Hellinger distance, provide concrete benefits in some problems remains to be seen. In the illustrations of Section 5, we focus primarily on the KL divergence, although we also discuss the use of the Hellinger distance in one example. In the Supplementary Material, we also provide an example in which the use of a seemingly natural yet inappropriate divergence leads to a violation of (2.3). This highlights the importance of verifying conditions (B1)–(B3) for divergences other than $\mathbb{D}_{KL}$ and $\mathbb{D}_H$, both of which we have already verified to satisfy these conditions

It is useful to scrutinize the conditions of Theorem 1 in the context of the projected perturbation path. Condition (A1) drives in large part our generalization of the procedures of Frangakis et al. (2015) and Luedtke et al. (2015) to arbitrary models. Projecting the path $\{P_{\epsilon,\lambda} : 0 \leq \epsilon \leq 1\}$ into $\mathcal{M}$ to obtain $\{P_{\epsilon,\lambda}^* : 0 \leq \epsilon \leq 1\}$ is expected to ensure that the score-like equation described in (A1) is solved. Condition (A2) imposes mild continuity requirements on the EIF. Since R is a second-order term, $R(P_{\epsilon,\lambda}, P)$ is generally of order $O(\epsilon^2)$ for each fixed $\lambda > 0$. Condition (A3) requires that $R(P_{\epsilon,\lambda}^*, P)$ be of order $o(\epsilon)$ for λ small and ϵ sufficiently smaller. Determining how the projection step and the smoothing parameter $\lambda > 0$ affect the rate of this term is critical to establishing whether (A3) holds. This is studied in detail in the next section

Much of the effort required to use representation (2.3) to compute the EIF goes into identifying the projection $P_{\epsilon,\lambda}^*$ of $P_{\epsilon,\lambda}$ onto the model space. In some cases, this projection can be determined analytically, while in others, a numerical approach must be taken. Regardless, the definition of $P_{\epsilon,\lambda}^*$ does not involve the parameter of interest. Hence, the more challenging portion of the approach is exclusively model-specific, and once it has been successfully tackled, the resulting projection can be used for any parameter a practitioner may wish to study. This contrasts sharply with the traditional approach to deriving the EIF, wherein the statistician must first derive an influence function, characterize the tangent space of the model, and finally project the influence function onto this tangent space. In that approach, both the parameter-specific task – finding an influence function – and the model-specific task – studying the tangent space and how to project onto it – require specialized knowledge. Performing these tasks for a given parameter and model combination does not

necessarily provide an easy way of tackling other parameters, in contrast to the approach we propose

3 Verification of technical conditions

The validity of representation (2.3) is guaranteed to hold under the high-level technical conditions (A1), (A2) and (A3). We now identify lower-level sufficient conditions under which the projected perturbation path constructed in (2.4) satisfies conditions (A1), (A2) and (A3), and thus under which Theorem 1 holds

Here and throughout, given a function h defined on $\mathcal{X}(P)$, we write

$\|h\|_{k,P} := \left[\int |h(u)|^k dP(u) \right]^{\frac{1}{k}}$ and $\|h\|_{\infty,A} := \sup_{u \in A} |h(u)|$ for any set A . We denote by $\mathcal{S}_{x,\lambda}$ the support of $H_{x,\lambda}$ and we define

$$r_k(\lambda) := \left\| \frac{dH_{x,\lambda}}{dP} \right\|_{k,P}.$$

This Radon-Nikodym derivative of $H_{x,\lambda}$ relative to P is defined for each $\lambda > 0$ since $H_{x,\lambda}$ is dominated by P by construction. Unless P assigns positive mass to $\{x\}$ and $H_{x,\lambda}$ is the degenerate distribution at $\{x\}$, the value of $r_k(\lambda)$ usually tends to infinity as λ tends to zero. The rate at which this occurs will be critical in our study of the technical conditions listed in Theorem 1

3.1 Near-solution of the EIF estimating equation

By virtue of being a global optimizer, and in view of condition (B3), $P_{\epsilon,\lambda}^*$ is expected to solve, or at least nearly solve, a rich collection of equations indexed by elements in the tangent space of $\mathcal{M}$ at $P_{\epsilon,\lambda}^*$, including that exhibited in condition (A1). The following theorem establishes a formal regularity condition under which this is indeed the case

Theorem 2.—Suppose that there exists an interval $J_0 = [m_0, m_1] \subset (0, +\infty)$ such that for each small $\lambda > 0$ the Radon-Nikodym derivative of $P_{\epsilon,\lambda}^*$ relative to P is uniformly contained in J_0 over the support of P for sufficiently small ϵ . Then, condition (A1) holds.

The condition in the above theorem is expected to hold in many situations since for $\epsilon = 0$ and any $\lambda > 0$, $P_{\epsilon,\lambda} = P$ is in $\mathcal{M}$ and thus $P_{\epsilon,\lambda}^* = P$ trivially. As such, under continuity of the projection operator relative to the supremum norm, for ϵ small enough, the Radon-Nikodym derivative of $P_{\epsilon,\lambda}^*$ relative to P is expected to be close to one, and therefore bounded both away from zero and above. In practice, it is often possible to either verify this condition analytically or to assess its plausibility numerically

3.2 Continuity of the EIF

We relied on certain notions of continuity to establish the validity of representation (2.3). The theorem below highlights how the continuity requirement stated in condition (A2) can be more concretely verified

Theorem 3.—Suppose that $\lim_{\lambda \rightarrow 0} \int \phi_P(u) dH_{x,\lambda}(u) = \phi_P(x)$ and $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \int \phi_{\epsilon,\lambda}^*(u) dP(u) = 0$. Condition (A2) holds provided either

$$(a) \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \|\phi_{\epsilon,\lambda}^* - \phi_P\|_{\infty, s_{x,\lambda}} = 0 \quad \text{or} \quad (b) \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} r_2(\lambda) \|\phi_{\epsilon,\lambda}^* - \phi_P\|_{2,P} = 0.$$

The requirement that $\int \phi_P(u) dH_{x,\lambda}(u)$ approximates $\phi_P(x)$ as λ tends to zero simply stipulates that averaging ϕ_P with respect to a distribution eventually concentrating all its probability mass on $\{x\}$ should approximately yield $\phi_P(x)$. Furthermore, this theorem requires that $\int \phi_{\epsilon,\lambda}^*(u) dP(u)$ tends to zero, which is reasonable under some continuity since $\phi_{\epsilon,\lambda}^*$ tends to ϕ_P and $\int \phi_P(u) dP(u) = 0$. Beyond this, in order for condition (A2) to hold, it suffices either for $\phi_{\epsilon,\lambda}^*$ to approximate ϕ_P in supremum norm in a neighborhood of x or in $L_2(P)$ -norm at a rate faster than $r_2(\lambda)^{-1}$. These statements each hinge on a certain notion of continuity that appears needed whenever P is not finitely supported

3.3 Preservation of the rate of convergence

The proof of representation (2.3) hinges upon a linearization of the difference between $\Psi(P_{\epsilon,\lambda}^*)$ and $\Psi(P)$. To ignore the remainder term from this linearization, we require that $R(P_{\epsilon,\lambda}^*, P)/\epsilon$ be arbitrarily small for small enough λ and sufficiently smaller ϵ . The following theorem establishes that this is the case under certain regularity conditions

Theorem 4.—Suppose that there exists an interval $J_0 = [m_0, m_1] \subset (0, +\infty)$ such that for each small $\lambda > 0$ the Radon-Nikodym derivative of $P_{\epsilon,\lambda}^*$ relative to P is uniformly contained in J_0 over the support of P for sufficiently small ϵ . Suppose also that there exists some $0 < C < +\infty$ such that, for any $P_1 \in \mathcal{M}$ with Radon-Nikodym relative to P bounded above by m_1 over the support of P , we have that

$$|R(P_1, P)| \leq C \left\| \frac{dP_1}{dP} - 1 \right\|_{2,P}^2.$$

Then, condition (A3) holds

We have already discussed the first condition of this theorem since it is also used in Theorem 2. As for the second condition, it is often the case that the remainder term, being a second-order term arising from a linearization, can be bounded by the squared norm of the difference between the Radon-Nikodym derivative of $P_{\epsilon,\lambda}^*$ relative to P and its value at $\epsilon = 0$.

This inequality often follows quite easily from an application of the Cauchy-Schwartz inequality on the remainder term, and generally holds under rather mild conditions

4 Practical considerations

Representation (2.3) provides the theoretical foundations for our strategy for numerically approximating the EIF and thus constructing efficient estimators. The implementation of the approach suggested by this representation nevertheless presents specific challenges.

Practical guidelines, as provided below, may facilitate the successful implementation of our proposal in practice

4.1 Construction of the perturbation path

In constructing the linear perturbation path that defines $P_{\epsilon, \lambda}$ and thus $P_{\epsilon, \lambda}^*$, the nearly degenerate distribution at $\{x\}$ is used instead of its purely degenerate counterpart because it ensures that all distributions along the perturbation path are dominated by P . This is required to ensure the validity of the representation we have proposed. Clearly, there is no need for smoothing in the components of the data unit for which the corresponding marginal distribution implied by P is dominated by a counting measure. In fact, as we stress below, unnecessary smoothing can needlessly increase the computational burden of the approximation procedure. For components for which the corresponding marginal distribution is dominated by the Lebesgue measure, smoothing is generally needed. In practice, we suggest the use of product kernels for those components. Specifically, suppose that the data unit X is d -dimensional and can be partitioned into $X = (X_L, X_C)$, where $X_L := (X_{L1}, X_{L2}, \dots, X_{Ld1})$ and $X_C := (X_{C1}, X_{C2}, \dots, X_{Cd2})$ with $d_1 + d_2 = d$, and that the marginal distributions of X_L and X_C under P are respectively dominated by the Lebesgue measure and a discrete counting measure. In this case, we can typically use the product kernel with density function

$$u \mapsto h_{x, \lambda}(u) := \left[\prod_{j=1}^{d_1} K_{\lambda}(u_{Lj} - x_{Lj}) \right] \times \left[\prod_{j=1}^{d_2} I(u_{Cj} = x_{Cj}) \right] \quad (4.1)$$

relative to the sum of Lebesgue and counting measures, where $u := (u_L, u_C)$ with u_L and u_C possible realizations of X_L and X_C , respectively, and $K_{\lambda} : w \mapsto \lambda^{-1} K(\lambda^{-1} w)$ with K some symmetric and absolutely continuous univariate density function. The uniform kernel $K(w) := \mathbb{I}(-1 < w < +1)$ is particularly appealing due to its simplicity, which translates to greater practical feasibility of our numerical approximation procedure. If the uniform kernel is used, it is easy to verify that $r_K(\lambda) = \lambda^{-d(1-1/K)}$ provided, for example, the sectional evaluation $u_L \mapsto p(u_L, x_C)$ of the density p of P is continuous and bounded away from zero in a neighborhood of x_L . As suggested by Theorem 5 provided in the next subsection, for a secant line slope approximation to the derivative in (2.3) to be accurate, ϵ must generally be chosen such that $\epsilon \ll \lambda^{d_1}$. If d_1 is large, this requirement may be prohibitive, possibly even to the point of requiring a value of ϵ beyond the computer's standard level of precision and thus requiring special computational techniques. Of course, while this guideline is sufficient, it may be conservative in some applications. Later in this section, we provide a more practical means of selecting the value of ϵ and λ

As alluded to above, if we include smoothing over X_C as well in our choice of $H_{x, \lambda}$, we need $\epsilon \ll \lambda^{d_2}$. This can be much more prohibitive computationally than requiring that $\epsilon \ll \lambda^{d_1}$, particularly if d_2 is large. For this reason, smoothing in the construction of the linear perturbation path should be avoided for all components except those for which the corresponding marginal distribution under P is absolutely continuous. Additionally, for some parameters, smoothing can be avoided altogether for certain continuous components. As a

general guideline for which supporting theory remains to be developed, we expect no smoothing to be required at all if the parameter is sufficiently smooth at $P \in \mathcal{M}$, in the sense that $\Psi(P_m)$ tends to $\Psi(P)$ for any sequence $\{P_m \in \mathcal{M} : m = 1, 2, \dots\}$ for which the distribution function of P_m tends to that of P uniformly as m tends to infinity. Alternatively, if the MLE P_n^* of P based on observations $X_1, X_2, \dots, X_n$ from P is such that $\Psi(P_n^*)$ is a consistent estimator of $\Psi(P)$, no smoothing will generally be required. If, however, some regularization of the MLE is needed to ensure consistency, e.g., as in van der Laan (1996), smoothing will usually be critical

4.2 Finite-difference derivative approximations

When the projection of the linear perturbation path and the parameter of interest have an explicit form, it may be possible to compute analytically the pathwise derivative in representation (2.3). When analytic differentiation is possible, it is useful to do so since an approximation for $\phi_P(x)$ is then simply given by the evaluation of the pathwise derivative at $\epsilon = 0$ for a small $\lambda > 0$. The approximation thus involves only a single approximation parameter. Otherwise, numerical differentiation techniques must be used

Since the derivative in representation (2.3) can be expressed as the limit of the slope of the secant line between points $(0, \Psi(P))$ and $(\epsilon, \Psi(P_{\epsilon, \lambda}^*))$, the alternative representation

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_{\epsilon, \lambda}^*) - \Psi(P)}{\epsilon} \quad (4.2)$$

is available under the conditions of Theorem 1. This representation suggests that $\phi_P(x)$ may be approximated by $[\Psi(P_{\epsilon, \lambda}^*) - \Psi(P)]$ for appropriately small λ and ϵ . As indicated before, in general, the two limits in (4.2) cannot be interchanged. As a consequence, ϵ must be taken to be much smaller than λ . Practical guidelines for choosing ϵ and λ are provided in the next subsection. We note that, beyond the ability to compute the projection $P_{\epsilon, \lambda}^*$, use of (4.2) to approximate $\phi_P(x)$ requires no more than the ability to evaluate Ψ

Representation (4.2) is based on a two-point forward finite-difference approximation to the derivative in (2.3). A multi-point forward finite-difference scheme could be used to produce accurate approximations for possibly less prohibitive values of ϵ and λ . For $m \in \{1, 2, \dots\}$, the $(m + 1)$ -point forward finite-difference representation of the EIF value is

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \frac{\sum_{j=0}^m a_{j,m} \Psi(P_{j\epsilon, \lambda}^*)}{\epsilon} \quad (4.3)$$

for a specified vector $a_m := (a_{0,m}, a_{1,m}, \dots, a_{m,m})$. Fornberg (1988) gives a recursive formula for a_m and numerical values up to $m = 8$. For example, we have that $a_1 = (-1, +1)$, $a_2 = (-\frac{3}{2}, +2, -\frac{1}{2})$, $a_3 = (-\frac{11}{6}, +3, -\frac{3}{2}, +\frac{1}{3})$ and $a_4 = (-\frac{25}{12}, +4, -3, +\frac{4}{3}, -\frac{1}{4})$. Use of (4.3) as a basis for numerically approximating $\phi_P(x)$ indeed yields an improved approximation, as the following theorem indicates. Below, we denote by $\bar{\Psi}: \mathcal{M}_{NP} \rightarrow \mathbb{R}$ the mapping $P \mapsto \Psi(P^*)$

for P^* the projection of P into $\mathcal{M}(P)$, and by $\mathcal{M}_{0,\lambda}(P)$ the parametric submodel $\{P_{\epsilon,\lambda} : 0 < \epsilon < 1\}$.

Theorem 5.—*Suppose that conditions (A1)–(A3) hold, that $H_{x,\lambda}$ is a second-order kernel of the form (4.1), and that ϕ_P has bounded second derivatives in a neighborhood of x . Suppose also that, for small enough $\lambda > 0$, the parameter $\tilde{\Psi}$ is $(m_0 + 1)$ -times pathwise differentiable in $\mathcal{M}_{0,\lambda}(P)$, with gradient of order $m_0 + 1$ evaluated at $P_{\epsilon,\lambda}$ bounded uniformly for sufficiently small $\epsilon > 0$. Then, with $s := \min(m, m_0)$, we have that*

$$\frac{\sum_{j=0}^m a_{j,m} \Psi(P_{j\epsilon,\lambda}^*)}{\epsilon} = \phi_P(x) + O(\epsilon^s r_{s+1}(\lambda)^{s+1} + \lambda^2).$$

As indicated before, if $H_{x,\lambda}$ is taken to be the uniform kernel, we find that $r_{s+1}(\lambda) = \lambda^{-d_1 s/(s+1)}$, and so, the approximation error is of order $(\epsilon \lambda^{-d_1})^s + \lambda^2$. This result suggests that, at the cost of computing additional projections and evaluating the parameter at each of these projections, it may indeed be possible to increase approximation accuracy using a multi-point forward finite-difference schemes. Higher-order differentiability is required to ensure that such schemes yield valid results – in particular, this requires that the projected path be sufficiently smooth. More importantly, the theorem also indicates that there is no significant theoretical penalty incurred from using a multi-point scheme with number of points greater than the degree of differentiation, thereby supporting the use of multi-point schemes in practice. In Section 5, we show concrete benefits that may be derived from using a multi-point scheme in the context of Example 1

4.3 Selection of ϵ and λ values

When the pathwise derivative in (2.3) can be calculated analytically, the proposed approximation strategy only involves the smoothing parameter λ . The supporting theory suggests taking λ as small as possible. As we will illustrate in Section 5, in some cases there is little sensitivity to the choice of λ when (2.3) is used, and even a relatively large value of $\lambda > 0$ will yield stringent control of the approximation error

Whenever the involved projection is not available in closed form or differentiation with respect to ϵ is too cumbersome to perform analytically, a multi-point finite-difference scheme may be used to numerically approximate this analytic derivative, as described in the previous subsection. In such case, ϵ and λ must both be chosen, and more care is needed to ensure the reliability of the proposed procedure. The order of the limits in (4.2) and (4.3) suggests that we must select a small value of λ and even smaller value of ϵ . This was made more precise in the previous subsection, where in the case of a two-point scheme, for example, it is prescribed to choose ϵ to be much smaller than λ^{d_1} , with d_1 the number of components of P over which smoothing is performed. While this theoretical requirement may serve as a rough guide in practice, it does not provide a concrete means of selecting values for ϵ and λ . For this purpose, it is useful to produce a matrix representing the value of the approximation as a function of ϵ and λ , both ranging over an exponential scale – for example, we could consider both ϵ and λ in the set $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, \dots\}$. We refer

to the resulting display as an *epsilon-lambda plot*. As a convention, the y-axis is used to represent ϵ values while λ values are represented on the x-axis. Our theoretical findings suggest that the right balance between ϵ and λ will be achieved in a possibly curvilinear triangular region nested in the upper left portion of the epsilon-lambda plot. In this region, the finite-difference approximation of the EIF value should be essentially constant. One practical means of selecting ϵ and λ would then consist of identifying this region visually by determining the quasi-triangular region in the upper left portion of the matrix over which the approximated EIF value is fixed up to a certain level of precision. As an illustration, without yet providing details regarding the specific parameter and model under consideration, we may scrutinize the epsilon-lambda plot arising in Example 1 from Section 5 and based upon a two-point scheme. This plot, provided as Figure 1, suggests that, up to three decimal points, the EIF value of interest is -0.963 . This is indeed verified using theoretical calculations, as discussed in more detail in Section 5. The epsilon-lambda plot therefore may be a particularly useful tool for implementing the proposed approach for numerically approximating the EIF in practice

4.4 Numeric computation of the model space projection

In implementing our proposal, the main challenge consists of operationalizing the optimization problem that characterizes the projection of the linear perturbation path $\{P_{\epsilon,\lambda} : 0 \leq \epsilon \leq 1\}$ onto the model space $\mathcal{M}$. An analytic – or nearly analytic – form can be found for the projection in many problems, including the illustrations provided in Section 5. In other problems, the optimization problem is less analytically tractable and a numeric approach may be needed

A general strategy for numerically approximating the required projection is to instead consider the corresponding optimization problem over $\mathcal{M}_m$, where $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots \subseteq \mathcal{M}$ is a sequence of finite-dimensional submodels of $\mathcal{M}$ such that $\cup_{m=1}^{\infty} \mathcal{M}_m = \mathcal{M}$. We illustrate this in the context of families of tilted densities, though other parametrizations are possible. We note that any distribution Q dominated by P can be described as a tilted form $dQ(u) = \exp[h(u)]dP(u) / \int \exp[h(w)]dP(w)$ of P for some function h in a function class $\mathcal{H} = \mathcal{H}(\mathcal{M})$ determined by the model $\mathcal{M}$. Here, h characterizes the deviation of Q from P . It is often easier to determine suitable approximating finite-dimensional subspaces for $\mathcal{H}$ than for $\mathcal{M}$. Suppose that $\{h_1, h_2, \dots\} \subseteq \mathcal{H}$ forms a basis for $\mathcal{H}$, and let $\mathcal{H}_m$ denote the linear span of $\{h_1, h_2, \dots, h_m\}$. If P has density p relative to ν , the submodel $\mathcal{M}_m$ implied by $\mathcal{H}_m$ then consists of all distributions Q with density

$$\frac{dQ}{d\nu}(u) = \frac{\exp\left[\sum_{j=1}^m \beta_j h_j(u)\right] p(u)}{\int \exp\left[\sum_{j=1}^m \beta_j h_j(w)\right] p(w) \nu(dw)}$$

for some $\underline{\beta}_m := (\beta_1, \beta_2, \dots, \beta_m) \in \mathbb{R}^m$. The choice $\underline{\beta}_m = (0, 0, \dots, 0)$ leads to $Q = P$. Denote by $\underline{\beta}_{m,\epsilon,\lambda}^*$ the index of the projection $P_{\epsilon,\lambda}^*$ of $P_{\epsilon,\lambda}$ into $\mathcal{M}_m$. For small ϵ , it is expected that $P_{\epsilon,\lambda}^* \approx P$, which suggests that the corresponding index $\underline{\beta}_{m,\epsilon,\lambda}^*$ should be near zero. Thus,

the search for the optimizer can be focused in a neighborhood surrounding the origin in $\mathbb{R}^m$. In our experience, this simple observation can sometimes greatly accelerate the numerical optimization routine used. In practice, a sufficiently large m must be selected to ensure that the resulting approximation of the projection is accurate enough to ensure the validity of the numerical evaluation of the EIF based on (4.3). If the KL divergence is used as statistical distance, up to an additive constant, the resulting optimization problem is to maximize the objective function

$$\mathcal{L}(\underline{\beta}_m) := \sum_{j=1}^m \beta_j \int h_j(u) dP_{\epsilon, \lambda}(u) - \log \int \exp \left[\sum_{j=1}^m \beta_j h_j(u) \right] p(u) v(dw).$$

Since derivatives of $\mathcal{L}(\underline{\beta}_m)$ are easy to write down explicitly, many algorithms are available to solve this optimization problem efficiently, including the Newton-Raphson method. We note here that in this setup m should be taken as large as computationally feasible so as to minimize the bias induced by the approximation of $\mathcal{M}$ by $\mathcal{M}_m$. This is in contrast to traditional sieve estimation, where the choice of m generally requires a careful balance whenever the unsieved maximum likelihood estimator either does not exist or is inconsistent. This difference occurs because in our scheme all distributions in the regularized model are dominated by the distribution being projected into the model

It may sometimes be useful to consider a stochastic version of this deterministic optimization problem. For example, we could minimize the divergence between the approximating finite-dimensional submodel and the weighted empirical distribution based on a very large number of draws from a uniform mixture between P and $H_{x,\lambda}$, with weights $2(1 - \epsilon)$ and 2ϵ assigned to observations from P and $H_{x,\lambda}$, respectively. We are then faced with a standard parametric estimation problem, albeit one that may be high-dimensional. When a clever parametrization of the approximating submodel is used, it may be possible to use standard statistical learning techniques, including regularization methods from the machine learning literature, with computationally efficient and stable off-the-shelf implementations. When adopting this approach, it appears critical to ensure that the size of the dataset generated is very large compared to the richness of the approximating submodel, since otherwise the variability resulting from this parametric estimation problem could limit our ability to achieve the required level of accuracy. A detailed study of numerical strategies, including the approach described in this subsection, will be the focus of future work

4.5 Construction of an efficient estimator

As emphasized earlier, knowledge of the EIF facilitates the construction of efficient estimators in infinite-dimensional models. For example, if $\hat{P}_n$ is a consistent estimator of $P_0 \in \mathcal{M}$ based on independent draws $X_1, X_2, \dots, X_n$ from P_0 , the one-step Newton-Raphson estimator, defined as

$$\psi_n^+ := \Psi(\hat{P}_n) + \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(X_i),$$

is an efficient estimator of ψ_0 under certain regularity conditions. The one-step approach appears to be the constructive method most amenable to an implementation based on numerical approximations of the EIF. Indeed, if the analytic form of the EIF is not known, it suffices to numerically approximate the value of $\phi_{\hat{P}_n}(X_i)$ for each $i = 1, 2, \dots, n$, rather than the entire function $u \mapsto \phi_{\hat{P}_n}(u)$, to calculate ψ_n^+ . Thus, the procedure we propose can be used to approximate each of these n values. Nevertheless, when the projection step required to utilize the proposed representation of the EIF is computationally burdensome and the sample size n is large, computing each of these values may be challenging. One need not obtain an approximation of each $\phi_{\hat{P}_n}(X_i)$ if the objective is to simply compute the one-step estimator ψ_n^+ – in this case it suffices to obtain an approximation of the correction term

$B_n := \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(X_i)$. This simple observation is useful because a slight modification to (2.3) yields a numerical procedure for approximating the required empirical average. Specifically, the proof of Theorem 1 can be adapted to show that, under similar regularity conditions, if we define the linear perturbation $\hat{P}_{n,\epsilon,\lambda} := (1-\epsilon)\hat{P}_n + \epsilon \frac{1}{n} \sum_{i=1}^n H_{X_i,\lambda}$ between $\hat{P}_n$ and a uniform mixture of nearly degenerate distributions with mass concentrated around $X_1, X_2, \dots, X_m$ then

$$B_n = \lim_{\lambda \rightarrow 0} \frac{d}{d\epsilon} \Psi(\hat{P}_{n,\epsilon,\lambda}^*)|_{\epsilon=0}$$

with $\hat{P}_{n,\epsilon,\lambda}^*$ denoting the projection of $\hat{P}_{n,\epsilon,\lambda}$ onto the model. As such, a numerical approximation of the one-step estimator can be computed in a single numerical step as

$$\Psi(\hat{P}_n) + \frac{d}{d\epsilon} \Psi(\hat{P}_{n,\epsilon,\lambda}^*)|_{\epsilon=0} \approx \Psi(\hat{P}_n) + \frac{\sum_{j=0}^m a_{j,m} \Psi(\hat{P}_{n,j\epsilon,\lambda}^*)}{\epsilon}$$

for appropriately selected ϵ and λ values. We note here that while one may wish to approximate B_n as accurately as feasible, it suffices that B_n be numerically approximated up to order $O_P(n^{-1/2})$ for the one-step estimator to preserve its desirable asymptotic properties. As such, the level of accuracy that must be enforced in practice very much depends on sample size

5 Illustration and numerical studies

To illustrate the use of (2.3), we study four examples in which the calculation of the EIF can be difficult for non-experts, whereas the suggested approach renders the problem straightforward. Below, we describe the steps involved in using representation (2.3) to approximate the EIF in the context of each of these examples, and provide numerical illustrations of its use. In the Supplementary Material, we show how representation (2.3) can instead be used to obtain the analytic form of the EIF without specialized knowledge

5.1 Example 1: Average value of density function with known mean

5.1.1 Background—Given a distribution P with univariate Lebesgue density p , the average density value parameter is given by

$$\Psi(P) := E_P[p(X)] = \int p(u)^2 du.$$

Estimation and inference for the average density value has been extensively studied in the semiparametric efficiency literature (e.g., Bickel and Ritov, 1988). We use this parameter as our first illustration because it is simple to describe yet requires specialized knowledge to study using conventional techniques. Suppose that $\mathcal{M}_{\text{NP}}$ denotes the nonparametric model consisting of all univariate absolutely continuous distributions with finite-valued density. Suppose that $\mu \in \mathbb{R}$ is fixed and known, and denote by $\mathcal{M} \subset \mathcal{M}_{\text{NP}}$ the semiparametric model consisting of all distributions in $\mathcal{M}_{\text{NP}}$ with mean μ . We wish to compute the EIF of Ψ relative to $\mathcal{M}$ at a $P \in \mathcal{M}$ evaluated at an observation value x .

The EIF $\phi_{\text{NP},P}$ of Ψ relative to the nonparametric model $\mathcal{M}_{\text{NP}}$ evaluated at $P \in \mathcal{M}$ is given by $u \mapsto \phi_{\text{NP},P}(u) := 2[p(u) - \Psi(P)]$. It is straightforward to derive this analytic form from first principles. Observing that $\mathcal{M} = \{P \in \mathcal{M}_{\text{NP}} : \Theta(P) = 0\}$, where $\Theta(P) := \int u dP(u) - \mu$ is a pathwise differentiable parameter with EIF relative to $\mathcal{M}_{\text{NP}}$ at $P \in \mathcal{M}$ given by $u \mapsto \phi_P(u) := u - \mu$, Example 1 of Section 6.2 of Bickel et al. (1997) | implies that the EIF of Ψ relative to $\mathcal{M}$ can be obtained as

$$u \mapsto \phi_P(u) := \phi_{\text{NP},P}(u) - \frac{\int \phi_{\text{NP},P}(w) \phi_P(w) dP(w)}{\int \phi_P(w)^2 dP(w)} \phi_P(u) = 2 \left[p(u) - \Psi(P) - \frac{\int (w - \mu) p(w) dP(w)}{\int (w - \mu)^2 dP(w)} (u - \mu) \right].$$

While the resulting analytic form of this EIF is relatively simple, its derivation hinges on specialized knowledge unlikely to be available to most practitioners. Use of our novel representation of the EIF provides an alternative approach that avoids the need for such knowledge, as highlighted below.

5.1.2 Implementation and results—To utilize our representation, we must understand how to project a given distribution $Q \in \mathcal{M}_{\text{NP}}$, say with Lebesgue density q , into $\mathcal{M}$, say relative to the KL divergence. Suppose that the support of Q has finite lower and upper limits a and b , respectively, satisfying that $a < \mu < b$. An application of the method of Lagrange multipliers yields that the maximizer in p of $\int \log p(u) dQ(u)$ over the class of all Lebesgue densities with mean μ is given by $q^*(u) := [1 - \xi_0(u - \mu)]^{-1} q(u)$, where $\xi_0 \in \mathbb{R}$ solves the equation $\mathbb{U}(\xi) = 0$ with

$$\mathbb{U}(\xi) := \int \left[\frac{u}{1 - \xi(u - \mu)} - \mu \right] dQ(u),$$

and lies strictly between $(a - \mu)^{-1}$ and $(b - \mu)^{-1}$

To compute $\phi_P(x)$ using the approach proposed in this paper, we first construct the linear perturbation $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$, where $H_{x,\lambda}$ is an absolutely continuous distribution that concentrates its mass on shrinking neighborhoods of the set $\{x\}$ as λ tends to zero. In our numerical implementation, we take $H_{x,\lambda}$ to be the uniform distribution on the interval $(x - \lambda, x + \lambda)$. The projection of $P_{\epsilon,\lambda}$ onto $\mathcal{M}$ is then obtained as described in the preceding paragraph with $Q = P_{\epsilon,\lambda}$. As such, it has a closed-form analytic expression up to the constant $\xi_0 = \xi_0(\epsilon, \lambda)$ that can be numerically solved. We may then approximate $\phi_P(x)$ by the secantline slope or using an improved multi-point finite-difference scheme, as suggested above

We evaluated this procedure numerically for a particular distribution P and observation value x . Specifically, we took P to be the Beta distribution with parameters $\alpha = 3$ and $\beta = 5$, and evaluated our numerical procedure for approximating the true value of $\phi_P(0.6) \approx -0.963$. Figure 2 provides the percent error of our numerical approximation based upon a secant line slope for various combination of values for ϵ and λ . This approximation is inaccurate if either ϵ is not small enough or if λ is too small relative to ϵ . For small λ and much smaller ϵ , the secant line slope approximates the true value of $\phi_P(x)$ with high accuracy, as depicted by the dark green squares in Figure 2. This plot confirms what theory suggests regarding the choice of ϵ and λ . It also reaffirms the usefulness of the epsilon-lambda plot for selecting appropriate values of ϵ and λ . Figure 3 compares the accuracy of multi-point finite-difference schemes using 2, 3, 4 or 5 points over the same range of values for ϵ and λ . As more points are used in approximation scheme, the high-accuracy region grows substantially, and a good approximation can be obtained with much less stringent choices of ϵ and λ . This plot confirms the significant practical benefits that can be derived from using multi-point schemes

5.2 Example 2: Center of a symmetric distribution

5.2.1 Background—In this example, the statistical problem is to efficiently estimate the location parameter indexing a univariate distribution in a semiparametric symmetry model. Specifically, we consider the model consisting of each distribution P with symmetric density p relative to a fixed dominating measure ν . Denoting by $P_{\mu,f}$ the distribution with density $u \mapsto p(u) = f(u - \mu)$ relative to ν , the model can be written as $\mathcal{M} = \{P_{\mu,f} : \mu \in \mathbb{R}, f \in \mathcal{F}_0(\nu)\}$ for $\mathcal{F}_0(\nu)$ the class of all densities relative to ν symmetric about zero. This semiparametric model is a subset of the nonparametric model $\mathcal{M}_{\text{NP}}$ of all distributions dominated by ν . If the center of symmetry is of scientific interest, the parameter considered is $\Psi(P_{\mu,f}) = \mu$. We wish to compute the EIF of Ψ relative to $\mathcal{M}$ at a distribution $P = P_{\mu,f}$ with $\mu \in \mathbb{R}$ and $f \in \mathcal{F}_0(\nu)$ evaluated at an observation value x

As one of the first problems considered in the development of efficiency theory for infinite-dimensional models – see Stein (1956) – this problem has a rich history in statistics. The EIF of Ψ relative to $\mathcal{M}$ at $P = P_{\mu,f}$ has the simple form

$$u \mapsto \phi_P(u) := -\frac{1}{I(f)} \frac{\dot{f}(u - \mu)}{f(u - \mu)},$$

where $\dot{f}$ is the first derivative of f and $I(f) := \int [f(w)\dot{f}(w)]^2 f(w)\nu(dw)$. This EIF can be derived in many ways, though in this case the efficient score approach for classical semiparametric models is particularly expedient. Tsiatis (2007) provides an accessible reference of this approach. We show below how the use of the representation in this paper negates the need for this specialized knowledge for EIF calculation

5.2.2 Implementation and results—As before, we must determine the projection Q^* of an arbitrary distribution $Q \in \mathcal{M}_{\text{NP}}$, say with density q , into $\mathcal{M}$. Because $Q^* \in \mathcal{M}$, there exist $\mu^* \in \mathbb{R}$ and $f^* \in \mathcal{F}_0(\nu)$ such that $Q^* = P_{\mu^*, f^*}$. If the KL divergence is used, we have that

$$\mu^* = \operatorname{argmax}_{v \in \mathbb{R}} \sup_{r \in \mathcal{F}_0(\nu)} \int \log r(u-v) q(u) \nu(du).$$

Using Jensen's inequality, it is not difficult to show that, for a fixed candidate center of symmetry ν , the maximizer $\hat{r}_\nu$ of $\int \log r(u-v) q(u) \nu(du)$ over $\mathcal{F}_0(\nu)$ is given explicitly as the symmetrization of q ,

$$u \mapsto \hat{r}_\nu(u) := \frac{q(v+u) + q(v-u)}{2}.$$

It follows then that μ^* is the maximizer of $v \mapsto A_{KL}(v) := \int \log \hat{r}_\nu(u-v) q(u) \nu(du)$ over $\mathbb{R}$, and f^* is given by $\hat{r}_{\mu^*}$. If instead the Hellinger distance is used, it is not difficult to show that

$$\mu^* = \operatorname{argmax}_{v \in \mathbb{R}} \sup_{r \in \mathcal{F}_0(\nu)} \int \sqrt{r(u-v)q(u)} \nu(du),$$

and for a fixed center of symmetry ν , the maximizer $\hat{r}_\nu$ of $\int \sqrt{r(u-v)q(u)} du$ over $\mathcal{F}_0(\nu)$ is given by

$$u \mapsto \hat{r}_\nu(u) := \frac{1}{B(v; q)} \left[\frac{\sqrt{q(v+u)} + \sqrt{q(v-u)}}{\sqrt{2}} \right]^2,$$

where $B(v; q) := 1 + \int \sqrt{q(u)q(2v-u)} du$. Then, μ^* is the maximizer of $v \mapsto \int \sqrt{\hat{r}_\nu(u-v)q(u)} du$, or equivalently, of $v \mapsto A_H(v) := \int \sqrt{q(u)q(2v-u)} du$, and as before, f^* is given by $\hat{r}_{\mu^*}$. Thus, once again, the projection Q^* has an explicit form up to an implicitly defined constant. Since $\Psi(Q^*) = \mu^*$, the parameter evaluation at the projection of Q simply equals the computed maximizer of A_{KL} or A_H

To compute $\phi_P(x)$ for $P = P_{\mu, f}$ we again begin by constructing the linear perturbation $P_{\epsilon, \lambda} := (1 - \epsilon)P + \epsilon H_{x, \lambda}$ where $H_{x, \lambda}$ denotes an absolutely continuous distribution that concentrates its mass on shrinking neighborhoods of the set $\{x\}$ as λ tends to zero. We then

compute the maximizer $\mu_{\epsilon, \lambda}^*$ of A_{KL} or A_H with $Q = P_{\epsilon, \lambda}$, and approximate $\phi_P(x)$ using, for example, the secant line slope $(\mu_{\epsilon, \lambda}^* - \mu)/\epsilon$.

We evaluated this EIF approximation scheme by implementing the numerical algorithm above to obtain the value of $\phi_P(x)$ for $x = 1$ and P taken to be the normal distribution with mean zero and unit variance. Using the known analytic form of ϕ_P provided above, we find that $\phi_P(x) = x$ in this case – this is expected since this is the influence function of the sample mean, and the latter is an efficient estimator of the center of symmetry in a normal location family. The value we are seeking to approximate is $\phi_P(1) = 1$. The relative error of the secant line slope approximation based on the KL divergence is displayed in Figure 4 for different combinations of ϵ and λ . The pattern observed is similar as in Example 1, with high accuracy obtained for small λ and even smaller ϵ . The implementation using the Hellinger distance led to very similar results – in fact, the resulting relative error plot was indistinguishable from the plot in Figure 4

5.3 Example 3: Coefficient in semiparametric Poisson regression model

5.3.1 Background—We now consider the problem of efficient estimation of the regression coefficient in a partially linear multiplicative Poisson regression model. Suppose that the data unit is $X := (Y, Z, W) \sim P \in \mathcal{M}$, where Y is a count outcome, Z is a univariate exposure, and W is a vector of possible confounders. Let G be the space of univariate real-valued functions, and let $\mathcal{F}$ be the space of all bivariate distribution functions. We consider the model $\mathcal{M} := \{P_{\beta, g, F} : \beta \in \mathbb{R}, g \in \mathcal{G}, F \in \mathcal{F}\}$, where $P_{\beta, g, F}$ denotes the joint distribution implying

$$P_{\beta, g, F}(Y = y | Z = z, W = w) = e^{-\mu_{\beta, g}(z, w)} \cdot \frac{\mu_{\beta, g}(z, w)^y}{y!}, \quad y = 0, 1, 2, \dots$$

with $\log \mu_{\beta, g}(z, w) = \beta z + g(w)$, and $P_{\beta, g, F}(Z \leq z, W \leq w) = F(z, w)$, for each z and w . If Z is the exposure of interest, the coefficient β serves as a parsimonious summary of the relationship between Y and Z adjusting for W . As such, we consider the parameter $\Psi(P_{\beta, g, F}) = \beta$. We wish to compute the EIF of Ψ relative to $\mathcal{M}$ at distribution $P = P_{\beta, g, F}$ evaluated at observation value $x := (y, z, w)$

Again, efficiency calculations in this case can most expediently be performed using techniques for classical semiparametric models. The EIF can be expressed as an appropriate renormalization of the efficient score for β , defined as the projection of the regular score for β into the orthogonal complement of the nuisance tangent space. Using this technique, the EIF of Ψ relative to $\mathcal{M}$ at $P = P_{\beta, g, F}$ is found to be

$$x \mapsto \phi_P(x) := \frac{[z - a_P(w)][y - \mu_{\beta, g}(z, w)]}{\text{var}_P\{[Z - a_P(W)][Y - \mu_{\beta, g}(Z, W)]\}},$$

where $a_P(w) := E_P(ZY | W = w) / E_P(Y | W = w)$ for each w . Below, we discuss the use of our proposed numerical approach for computing this EIF without using any knowledge of semiparametric efficiency theory.

5.3.2 Implementation and results—Let $Q \in \mathcal{M}_{NP}$ be an arbitrary distribution. The projection Q^* of Q into $\mathcal{M}$ can be written as $Q^* = P_{\beta^*, g^*, F^*}$ for some $\beta^* \in \mathbb{R}$, $g^* \in G$ and $F^* \in \mathcal{F}$. Here, we focus on the KL divergence. Writing $\mathbb{L}(\beta, g) := \iint_{z, w} \sum_y \log p_{\beta, g}(y | z, w) q(y | z, w) F_Q(dz, dw)$ with $p_{\beta, g}(y | z, w)$ denoting the conditional probability mass function of Y given $Z = z$ and $W = w$ evaluated at y and implied by parameter choices (β, g) , and with F_Q denoting the distribution of (Z, W) implied by Q , we have that $\beta^* = \operatorname{argmax}_{\beta \in \mathbb{R}} \sup_{g \in G} \mathbb{L}(\beta, g)$. For fixed β , the maximizer $\hat{g}_\beta$ of $\mathbb{L}(\beta, g)$ is given by $w \mapsto \hat{g}_\beta(w) := \log E_Q(Y | W = w) - \log E_Q(e^{\beta Z} | W = w)$. Under conditions allowing the interchange of integrals and derivatives, the maximizer β^* of $\mathbb{L}_0(\beta) := \mathbb{L}(\beta, \hat{g}_\beta)$ can be shown to also be the solution of the profile score equation $\mathbb{U}_0(\beta) = 0$, where

$$U_0(\beta) := E_Q \left[E_Q(ZY | W) - \frac{E_Q(Y | W) E_Q(Z e^{\beta Z} | W)}{E_Q(e^{\beta Z} | W)} \right].$$

It then follows that $g^* = \hat{g}_{\beta^*}$. Since $\mathcal{M}$ does not restrict the distribution of (Z, W) , we have that $F^* = F_Q$.

To compute $\phi_P(x)$ for $P = P_{\beta, g, F}$, we begin by constructing the linear perturbation $P_{\epsilon, \lambda} := (1 - \epsilon)P_\epsilon H_{x, \lambda}$. If Z is discrete-valued under P , $H_{x, \lambda}$ is a joint distribution for X that concentrates its mass on shrinking neighborhoods of the set $\{x\}$ as λ tends to zero and such that the event $\{Y = y, Z = z\}$ has probability one under $H_{x, \lambda}$ for every $\lambda > 0$. If Z is continuous-valued, the latter requirement can instead be taken to be that $\{Y = y\}$ has probability one under $H_{x, \lambda}$. We then compute $\beta_{\epsilon, \lambda}^*$ as the solution of $\mathbb{U}_0(\beta) = 0$ with $Q = P_{\epsilon, \lambda}$, and approximate the EIF value $\phi_P(x)$ by the secant line slope $(\beta_{\epsilon, \lambda}^* - \beta)/\epsilon$. We evaluated our proposed numerical scheme for approximating the value of $\phi_P(x)$ for $x = (1, 0, 0.5)$ and a distribution P under which: conditional on $(Z, W) = (z, w)$, Y has a Poisson distribution with mean $\exp\{\log(2)z + w\}$; conditional on $W = w$, Z is a Bernoulli random variable with success probability $\expit(1 - 2w)$; and W has a uniform distribution on the interval $(0, 1)$. The true value of the EIF can be calculated to be $\phi_P(1, 0, 0.5) \approx 0.793$. Figure 5 provides the relative error of our proposed secant line approximation for different combinations of ϵ and λ , and confirms patterns observed in previous examples.

5.4 Example 4: G-computation parameter under Markov structure

5.4.1 Background—We finally consider efficient estimation of a more complex parameter arising in the causal inference literature. Suppose that the data unit consists of the longitudinal observation $X := (L_0, A_0, \dots, L_K, A_K, L_{K+1}) \sim P$, where $L_0, L_1, \dots, L_K$ is a sequence of measurements collected at $K+1$ distinct instances through time, L_{K+1} is the outcome of interest, and $A_0, A_1, \dots, A_K$ are intervention indicators corresponding to each pre-outcome

timepoint. For simplicity, we consider all treatment indicators to be binary. Let $\mathcal{M}_{\text{NP}}$ be a nonparametric model for P . We may be interested in the covariate-adjusted, treatment-specific mean $\Psi(P)$ corresponding to the intervention $(A_0, A_1, \dots, A_K) = (1, 1, \dots, 1)$. Mathematically, $\Psi(P)$ is given explicitly by $EP[m_{0,P}(L_0)]$ via the G-computation recursion

$$m_j, P(\bar{\ell}_j) := EP[m_{j+1}, P(\bar{\ell}_{j+1}) | \bar{L}_j = \bar{\ell}_j, A_j = A_{j-1} = \dots = A_0 = 1]$$

for $j = K, K-1, \dots, 0$, where we have set $m_{K+1}, P(\bar{\ell}_{K+1}) := L_{K+1}$. Here, for any vector $u := (u_0, u_1, \dots)$ we write $\bar{u}_k := (u_0, u_1, \dots, u_k)$. This parameter only depends on P through the conditional distribution $P_{j+1,1}$ of $\bar{L}_{j+1}$ given $\bar{L}_j$ and $A_0 = A_1 = \dots = A_j = 1$ for $j = 0, 1, \dots, K$, and the marginal distribution $P_{0,1}$ of L_0 . Under certain untestable causal assumptions, $\Psi(P)$ corresponds to the mean of the counterfactual outcome Y defined by an intervention setting all treatment nodes to one (Robins, 1986). With respect to $\mathcal{M}_{\text{NP}}$, or any model with restrictions only on the conditional distribution of A_j given $\bar{A}_{j-1}$ and $\bar{L}_j$ possibly for any $j \in \{0, 1, \dots, K\}$, the EIF of Ψ at P is known to be given by $\phi_{\text{NP}, P} := \sum_{j=0}^{K+1} \phi_{j, \text{NP}, P}$, where $\phi_{0, \text{NP}, P}(x) := m_{0, P}(x) - \Psi(P)$ and

$$\phi_{j, \text{NP}, P}(x) := \frac{a_0 a_1 \dots a_{j-1}}{\prod_{r=0}^{j-1} P(A_r = 1 | \bar{L}_r = \bar{\ell}_r, A_0 = A_1 = \dots = A_{r-1} = 1)} \{m_j, P(\bar{\ell}_j) - m_{j-1}, P(\bar{\ell}_{j-1})\}$$

for $j = 1, 2, \dots, K+1$.

We consider instead the model $\mathcal{M}$ consisting of the subset of distributions in $\mathcal{M}_{\text{NP}}$ under which, for each $j = 2, 3, \dots, K+1$, L_j and $\bar{L}_{j-2}$ are independent given L_{j-1} and $A_{j-1} = A_{j-2} = \dots = A_0 = 1$. If $P \in \mathcal{M}$, we note that $m_j, P(\bar{\ell}_j) = m_j, P(\ell_j)$ for each j . The EIF of Ψ at P relative to $\mathcal{M}$ is given by, where $\phi_P := \sum_{j=0}^{K+1} \phi_{j, P}$ and $\phi_{j, P}$ is defined pointwise as

$$\begin{aligned} x \mapsto \phi_{j, P}(x) &:= EP[\phi_{j, \text{NP}, P}(X) | L_j = \ell_j, L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1}] \\ &\quad - EP[\phi_{j, \text{NP}, P}(X) | L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1}] \\ &= a_0 a_1 \dots a_{j-1} \cdot T_j(P)(x) \cdot \{m_j, P(\ell_j) - m_{j-1}, P(\ell_{j-1})\} \end{aligned}$$

for $j = 1, 2, \dots, K+1$, and we use $T_j(P)(x)$ to denote

$$EP \left[\frac{1}{\prod_{r=0}^{j-1} P(A_r = 1 | \bar{L}_r, A_0 = A_1 = \dots = A_{r-1} = 1)} \middle| L_j = \ell_j, L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1} \right].$$

Deriving this expression requires specialized knowledge and familiarity with efficiency theory for longitudinal structures. Furthermore, even given this analytic expression, the EIF may often be difficult to compute since it involves rather elaborate conditional expectations

5.4.2 Implementation and results—Again, we must understand how to project a given distribution $Q \in \mathcal{M}_{\text{NP}}$ into $\mathcal{M}$, say according to the KL divergence. Given a dominating

measure ν , we denote the density function of Q with respect to ν as q . Furthermore, we denote by q_{L_j} the density of the conditional distribution of L_j given $\bar{L}_{j-1}$ and $\bar{A}_{j-1}$, and by q_{A_j} the density of the conditional distribution of A_j given $\bar{L}_j$ and $\bar{A}_{j-1}$. We also denote by $q_{L_{j,1}}$ the density q_{L_j} with $\bar{a}_{j-1} = (1, 1, \dots, 1)$. We use the same notational convention for any other candidate density r . Because for any candidate r we can write

$$\int \log r(u) dQ(u) = \sum_{j=0}^{K+1} \int \log r_{L_j}(\ell_j | \bar{\ell}_{j-1}, \bar{a}_{j-1}) dQ(u) + \sum_{j=0}^K \int \log r_{A_j}(a_j | \bar{\ell}_j, \bar{a}_{j-1}) dQ(u)$$

and $\mathcal{M}$ can be written as a product model for the set of conditional distributions implied by the joint distribution, the required optimization problem can be performed separately for each conditional density. Because computing $\Psi(Q^*)$ does not require any component of Q^* beyond $q_{L_{j,1}}^*$ for $j = 0, 1, \dots, K+1$, we focus our attention on the corresponding optimization problems alone. Below, we denote by $\tilde{q}(\ell_j, \ell_{j-1})$ the marginalized density $\iint \dots \int q(\ell_0, 1, \ell_1, 1, \dots, \ell_{j-1}, 1, \ell_j) \nu(d\ell_0, d\ell_1, \dots, d\ell_{j-2})$. To find $q_{L_{j,1}}^*$ for $j = 2, 3, \dots, K+1$, we must maximize the criterion

$$\begin{aligned} \mathbb{L}_j(r_{L_{j,1}}) &:= \iint \dots \int \log r_{L_{j,1}}(\ell_j | \ell_{j-1}) q(\ell_0, 1, \ell_1, 1, \dots, \ell_{j-1}, 1, \ell_j) \nu(d\ell_0, d\ell_1, \dots, d\ell_j) \\ &= \iint \log r_{L_{j,1}}(\ell_j | \ell_{j-1}) \tilde{q}(\ell_{j-1}, \ell_j) \nu(d\ell_{j-1}, d\ell_j) \\ &= \iint \log r_{L_{j,1}}(\ell_j | \ell_{j-1}) \frac{\tilde{q}(\ell_{j-1}, \ell_j)}{\int \tilde{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)} \nu(d\ell_j) \int \tilde{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j) \nu(d\ell_{j-1}) \end{aligned}$$

over the class of candidate conditional densities that do not depend on $\bar{\ell}_{j-2}$, here represented by $r_{L_{j,1}}$. Since for each fixed ℓ_{j-1} the mapping

$\ell_j \mapsto \tilde{q}(\ell_{j-1}, \ell_j) / \int \tilde{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)$ defines a proper conditional density, by Jensen's inequality, $\mathbb{L}_j(r_{L_{j,1}})$ is maximized by

$$q_{L_{j,1}}^*(\ell_j | \ell_{j-1}) = \frac{\tilde{q}(\ell_{j-1}, \ell_j)}{\int \tilde{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)}.$$

It is easy to see that $\mathcal{M}$ constrains neither $r_{L_1,1}$ nor r_{L_0} and therefore $q_{L_1,1}^* = q_{L_1,1}$ and $q_{L_0}^* = q_{L_0}$. Thus, in the context of a longitudinal data structure, the projection of any given distribution Q into a model only constrained by a Markov structure has an explicit analytic form

To compute $\phi_P(x)$ numerically, we first construct the linear perturbation $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$, where $H_{x,\lambda}$ is a distribution dominated by P and concentrating its mass in shrinking neighborhoods of the set $\{x\}$ as λ tends to zero. The projection $P_{\epsilon,\lambda}^*$ onto $\mathcal{M}$ is obtained by setting $Q = P_{\epsilon,\lambda}$ in the preceding paragraph. As in previous examples, we may approximate

$\phi_{\mu}(x)$ by the secant line slope for small λ and even smaller ϵ . Because in this example $P_{\epsilon, \lambda}^*$ is available in closed form, $\phi_{\mu}(x)$ can alternatively be approximated by the evaluation of $\frac{d}{d\epsilon}\Psi P_{\epsilon, \lambda}^*$ for small λ

For simplicity, in our numerical evaluation of the EIF, we restricted our attention to a setting with $K = 2$ post-baseline time-points. We considered the joint distribution P of X defined in terms of the following conditional distributions. The baseline covariate L_0 has a discrete uniform distribution on the set $\{0, 1, 2, 3, 4\}$. Given $L_0 = l_0$, A_0 has a Bernoulli distribution with success probability $\text{expit}(-1 + 0.5l_0)$. Given $A_0 = a_0$ and $L_0 = l_0$, L_1 has a normal distribution with mean $3l_0 - 3a_0$ and variance 4. Given $L_1 = l_1$, $A_0 = a_0$ and $L_0 = l_0$, A_1 has a Bernoulli distribution with success probability $\text{expit}\{-5 + c_{10}(l_1) + a_0 + 0.5l_0\}$, where we define c_{10} to be the trimming function $u \mapsto 10 \cdot I_{(-\infty, -10)}(u) + u \cdot I_{[-10, +10]}(u) + 10 \cdot I_{(+10, +\infty)}(u)$. Given $A_1 = a_1$, $L_1 = l_1$, $A_0 = a_0$ and $L_0 = l_0$, Y has a Bernoulli distribution with success probability $\text{expit}\{-1 + 0.5c_{10}(l_1) - 0.5a_1 - a_0\}$. We evaluated approximations of $\phi_{\mu}(x)$ for $x = (0, 1, 2, 1, 1)$ based on either the secant line slope or the analytic pathwise derivative. We report the relative error of the approximation obtained using the secant line slope in Figure 6 and using the analytic derivative in Figure 7. Once again, the pattern observed in Figure 6 is similar to the pattern seen in other examples. From Figure 7, we note that a high level of accuracy is achieved with a relatively large $\lambda > 0$. Thus, use of the analytic derivative essentially eliminates the need for careful selection of approximation parameters otherwise needed. These implementations were repeated at additional values of x . Because very similar patterns emerged, results are not reported here

6 Concluding remarks

The novel representation of the EIF we have presented in this paper provides a novel approach for analytically deriving the EIF and suggests natural strategies for numerically approximating the EIF. This representation holds in arbitrary models under mild regularity conditions. Use of this representation requires the ability to project a given distribution into the statistical model – if the KL divergence is used, this is essentially no more than a maximum likelihood step that can be tackled by most practitioners. Most importantly, the involved work requires neither knowledge of efficiency theory nor familiarity with concepts from functional analysis or differential geometry. As such, this representation has the potential of democratizing the calculation of the EIF and thus the construction of efficient estimators in nonparametric and semiparametric models. Even for seasoned researchers in semiparametric and nonparametric theory, it provides an alternate means of tackling difficult problems, including those for which the EIF is either difficult or impossible to derive analytically

In most problems, we anticipate the analytic work required to project the linear perturbation path onto the model space to be much simpler than that needed for the conventional tangent space approach. Nevertheless, this may still constitute a barrier in practice. However, because the task of projecting onto the model space represents no more than an optimization problem, albeit an infinite-dimensional one, computational tools may readily be used to circumvent most, if not all, analytic work otherwise required. This is encouraging since

strong computational skills are commonplace in statistics and data science. Furthermore, we expect the numerical challenge to become increasingly surmountable as the capability of our computational devices continues to grow over time. While in this paper we established theoretical results that may facilitate computerization, the practical challenges of numerical computerization must still be carefully studied and addressed to achieve the goal of automated efficient estimation. For example, robust protocols must be devised for automatically selecting suitable values of approximation parameters involved in the numerical computerization of efficient estimators. These important questions are the focus of ongoing research

As with all methods that incorporate some level of automation and more readily lend themselves to use by non-specialists, there is a clear potential for misuse of the results we have presented. This appears to be an inevitable risk inherent to this type of proposal, and it equally applies to some of the most celebrated tools in current statistical practice, including the bootstrap. Deriving the EIF analytically undoubtedly remains the gold-standard approach and it should be preferred whenever possible since much information can be learned about the problem at hand from the analytic form of the EIF. In particular, verification of the regularity conditions invoked in this paper can be difficult without prior analytic knowledge of the EIF. Nevertheless, the representation introduced in this paper has the potential of serving as an important new tool in the arsenal of statistical researchers and practitioners alike for performing semiparametric and nonparametric analyses. Devising algorithms for verifying the required regularity conditions in any given problem is an important avenue for future research

We have noted that a distinct advantage of the representation we have provided is that once it has been used to compute the EIF of a certain parameter in a given statistical model, the EIF of any other parameter can be obtained without any additional work since the bulk of the work required is exclusively model-specific. Nevertheless, the involved computational work must be repeated for each observation value at which we wish to evaluate the EIF. In particular, this makes it difficult to approximate the entire EIF as a function, particularly in the case of continuous or longitudinal data units. While the one-step approach only requires the EIF at the observed data points, the implementation of other efficient estimators with potentially better properties, such as targeted minimum loss-based estimators (TMLE), generally requires the entire EIF. The representation presented in this paper is therefore not conducive to a computerized implementation of TMLE. There is promise that alternative representations may be better suited for this purpose – this is also an area of active research

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge the support of the Career Development Fund of the Department of Biostatistics at the University of Washington (MC), the New Development Fund of the Fred Hutchinson Cancer Research Center (ARL), and of NIH/NIAID grants 5UM1AI068635 (MC,ARL) and 5R01AI074345 (MJvdL)

Appendix

Throughout this Appendix, we will use the acronym RND to denote the Radon-Nikodym derivative.

Proof of Theorem 2. Denote by $r_{\infty}(\lambda)$ the supremum norm over $\mathcal{X}(P)$ of the RND of $H_{x,\lambda}$ relative to P . It is easy to verify that, for any fixed $\lambda > 0$, the RND of P relative to $P_{\epsilon,\lambda}$ is uniformly bounded in $(0.5, 1.5)$ over $\mathcal{X}(P)$ provided $\epsilon < 1/\max[r_{\infty}(\lambda) - 1, 3]$. Using the fact that $P_{\epsilon,\lambda}^*$ minimizes $\mathbb{D}(P_1, P_{\epsilon,\lambda})$ over all $P_1 \in \mathcal{M}(P_{\epsilon,\lambda})$, and that $P_1 \in \mathcal{M}(P_{\epsilon,\lambda})$ since $P_{\epsilon,\lambda}$ dominates P by construction, we then have that

$$\begin{aligned} 0 \leq \mathbb{D}(P_{\epsilon,\lambda}^*, P_{\epsilon,\lambda}) &\leq \mathbb{D}(P, P_{\epsilon,\lambda}) \leq \tilde{K}_1 \int \left[\sqrt{\frac{dP}{dP_{\epsilon,\lambda}}}(u) - 1 \right]^2 dP_{\epsilon,\lambda}(u) = \tilde{K}_1 \int \left[\sqrt{\frac{dP_{\epsilon,\lambda}}{dP}}(u) - 1 \right]^2 dP(u) \\ &\leq \tilde{K}_1 \int \left[\frac{dP_{\epsilon,\lambda}}{dP}(u) - 1 \right]^2 dP(u) \leq \tilde{K}_1 \epsilon^2 r_2(\lambda)^2, \end{aligned}$$

where we have set $\tilde{K}_1 := K_1(0.5, 1.5)$ and used conditions (B1) and (B2). It is straightforward to show that the RND of $P_{\epsilon,\lambda}^*$ relative to $P_{\epsilon,\lambda}$ is uniformly bounded between $0.5m_0$ and $1.5m_1$ for $\epsilon < 1/\max[r_{\infty}(\lambda) - 1, 3]$. Using condition (B2), this allows us to write that

$$\begin{aligned} 0 \leq \left\| \frac{dP_{\epsilon,\lambda}^*}{dP_{\epsilon,\lambda}} - 1 \right\|_{2, P_{\epsilon,\lambda}}^2 &= \int \left[\frac{dP_{\epsilon,\lambda}^*}{dP_{\epsilon,\lambda}}(u) - 1 \right]^2 dP_{\epsilon,\lambda}(u) \\ &= \int \left[\sqrt{\frac{dP_{\epsilon,\lambda}^*}{dP_{\epsilon,\lambda}}}(u) - 1 \right]^2 \left[\sqrt{\frac{dP_{\epsilon,\lambda}^*}{dP_{\epsilon,\lambda}}}(u) + 1 \right]^2 dP_{\epsilon,\lambda}(u) \\ &\leq \tilde{K}_2 \int \left[\sqrt{\frac{dP_{\epsilon,\lambda}^*}{dP_{\epsilon,\lambda}}}(u) - 1 \right]^2 dP_{\epsilon,\lambda}(u) \leq \tilde{K}_0 \tilde{K}_2 \mathbb{D}(P_{\epsilon,\lambda}^*, P_{\epsilon,\lambda}) \leq \tilde{K}_0 \tilde{K}_1 \tilde{K}_2 \epsilon^2 r_2(\lambda)^2 \end{aligned}$$

with $\tilde{K}_0 := K_0(0.5m_0, 1.5m_1)$ and $\tilde{K}_2 := 1.5m_1 + 2\sqrt{1.5m_1} + 1$. Writing $\tilde{K} := \sqrt{\tilde{K}_0 \tilde{K}_1 \tilde{K}_2}$, in view of the fact that $\phi_{\epsilon,\lambda}^* \in T_{\mathcal{M}}(P_{\epsilon,\lambda}^*)$, the latter inequality implies, in view of condition (B3), that

$$\frac{\left| \int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) \right|}{\epsilon} \leq \frac{\tilde{K} \epsilon r_2(\lambda) B(\tilde{K} \epsilon r_2(\lambda))}{\epsilon} = \tilde{K} r_2(\lambda) B(\tilde{K} \epsilon r_2(\lambda)) \rightarrow 0$$

as ϵ tends to zero. This establishes that $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) / \epsilon = 0$, as required. $\square$

Proof of Theorem 3. We first note that

$$\begin{aligned} &\left| \int \phi_{\epsilon,\lambda}^*(u) d(H_{x,\lambda} - P)(u) - \phi_P(x) \right| \\ &= \left| \int \{ \phi_{\epsilon,\lambda}^*(u) - \phi_P(u) \} dH_{x,\lambda}(u) + \int \phi_P(u) dH_{x,\lambda}(u) - \phi_P(x) - \int \phi_{\epsilon,\lambda}^*(u) dP(u) \right| \\ &\leq \left| \int \{ \phi_{\epsilon,\lambda}^*(u) - \phi_P(u) \} dH_{x,\lambda}(u) \right| + \left| \int \phi_P(u) dH_{x,\lambda}(u) - \phi_P(x) \right| + \left| \int \phi_{\epsilon,\lambda}^*(u) dP(u) \right| \end{aligned}$$

and because by assumption the second and third summands on the second line tend to zero as λ tends to zero, it suffices to study the first summand. We can bound this term by $\|\phi_{\epsilon, \lambda}^* - \phi_P\|_{\infty, S_{X, \lambda}}$ and so, if condition (a) holds, the result follows immediately.

Alternatively, we can write this term as

$$\begin{aligned} \left| \int \{ \phi_{\epsilon, \lambda}^*(u) - \phi_P(u) \} dH_{X, \lambda}(u) \right| &= \left| \int \frac{dH_{X, \lambda}}{dP}(u) \{ \phi_{\epsilon, \lambda}^*(u) - \phi_P(u) \} dP(u) \right| \\ &\leq \|\phi_{\epsilon, \lambda}^* - \phi_P\|_{2, P} \left\| \frac{dH_{X, \lambda}}{dP} \right\|_{2, P} \end{aligned}$$

and thus, if condition (b) holds, the result is also guaranteed to hold $\square$

Proof of Theorem 4. We will build upon notation and results established in the proof of Theorem 2. Below, we take a small but otherwise arbitrary $\lambda > 0$ and consider only $\epsilon < 1/\max[r_{\infty}(\lambda) - 1, 3]$. We first note that

$$\frac{dP_{\epsilon, \lambda}^*}{dP} - 1 = \frac{dP_{\epsilon, \lambda}}{dP} \left(\frac{dP_{\epsilon, \lambda}^*}{dP_{\epsilon, \lambda}} - 1 \right) + \left(\frac{dP_{\epsilon, \lambda}}{dP} - 1 \right).$$

We can write that

$$\begin{aligned} \left\| \frac{dP_{\epsilon, \lambda}}{dP} \left(\frac{dP_{\epsilon, \lambda}^*}{dP_{\epsilon, \lambda}} - 1 \right) \right\|_{2, P}^2 &= \int \left[\frac{dP_{\epsilon, \lambda}}{dP}(u) \right]^2 \left[\frac{dP_{\epsilon, \lambda}^*}{dP_{\epsilon, \lambda}}(u) - 1 \right]^2 dP(u) \\ &\leq \left\| \frac{dP_{\epsilon, \lambda}}{dP} \right\|_{\infty, \mathcal{X}(P)} \left\| \frac{dP_{\epsilon, \lambda}^*}{dP_{\epsilon, \lambda}} - 1 \right\|_{2, P_{\epsilon, \lambda}}^2 \leq [1 - \epsilon + \epsilon r_{\infty}(\lambda)] [\tilde{K} \epsilon r_2(\lambda)]^2 \end{aligned}$$

and additionally, that

$$\left\| \frac{dP_{\epsilon, \lambda}}{dP} - 1 \right\|_{2, P}^2 = \epsilon^2 \int \left[1 - \frac{dH_{X, \lambda}}{dP}(u) \right]^2 dP(u) \leq \epsilon^2 r_2(\lambda)^2.$$

Thus, using the above inequalities and the triangle inequality, we find that

$$0 \leq |R(P_{\epsilon, \lambda}^*, P)| \leq C \left\| \frac{dP_{\epsilon, \lambda}^*}{dP} - 1 \right\|_{2, P}^2 \leq \epsilon^2 r_2(\lambda)^2 [\sqrt{1 - \epsilon + \epsilon r_{\infty}(\lambda)} \tilde{K} + 1]^2,$$

which then implies that $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} R(P_{\epsilon, \lambda}^*, P)/\epsilon = 0$, as required. $\square$

Proof of Theorem 5. We will denote by $f_{\lambda}(\epsilon, \lambda)$ the derivative $\frac{d^r}{d\epsilon^r} \Psi(P_{\epsilon, \lambda}^*)$ and by $\phi_{\epsilon, \lambda, m}^{\circ}$ the $(m+1)$ -point forward finite-difference approximation to $\phi_P(x)$. The approximation error can be written as

$$\phi_{\epsilon, \lambda, m}^{\circ} - \phi_P(x) = [\phi_{\epsilon, \lambda, m}^{\circ} - f_1(0, \lambda)] + [f_1(0, \lambda) - \phi_P(x)],$$

where we will see that the first bracketed summand benefits from use of the multi-point scheme. Define the Taylor remainder $R_{j\epsilon, \lambda, s} := f_0(j\epsilon, \lambda) - \sum_{r=1}^s \frac{(j\epsilon)^r}{r!} f_r(0, \lambda)$. Setting $A(r, m) := \sum_{j=1}^m j^r a_{j, m}$, in view of the fact that $A(0, m) = 0$, $A(1, m) = 1$ and $A(r, m) = 0$ for $r = 2, 3, \dots, m$ (Fornberg, 1988), we note that

$$\begin{aligned} \phi_{\epsilon, \lambda, m}^{\circ} &= \frac{1}{\epsilon} \sum_{j=1}^m a_{j, m} \Psi(P_{j\epsilon, \lambda}^*) = \sum_{j=1}^m a_{j, m} \left[\frac{\Psi(P_{j\epsilon, \lambda}^*) - \Psi(P)}{\epsilon} \right] = \sum_{j=1}^m j a_{j, m} \left[\frac{\Psi(P_{j\epsilon, \lambda}^*) - \Psi(P)}{j\epsilon} \right] \\ &= \sum_{j=1}^m j a_{j, m} \left[f_1(0, \lambda) + \sum_{r=2}^s \frac{(j\epsilon)^{r-1}}{r!} f_r(0, \lambda) + \frac{R_{j\epsilon, \lambda, s}}{j\epsilon} \right] \\ &= f_1(0, \lambda) A(1, m) + \sum_{r=2}^s \frac{\epsilon^{r-1}}{r!} f_r(0, \lambda) A(r, m) + \frac{1}{\epsilon} \sum_{j=1}^m a_{j, m} R_{j\epsilon, \lambda, s} = f_1(0, \lambda) + \frac{1}{\epsilon} \sum_{j=1}^m a_{j, m} R_{j\epsilon, \lambda, s}, \end{aligned}$$

where the second line follows from the first by a Taylor expansion of $\eta \rightarrow \Psi(P_{\eta, \lambda}^*)$ around $\eta = 0$ evaluated at $\eta = j\epsilon$. By Taylor's theorem, the remainder term has the form

$$R_{j\epsilon, \lambda, s} = \frac{(j\epsilon)^{s+1}}{(s+1)!} f_{s+1}(\xi_{j, \epsilon, \lambda}, \lambda)$$

for some $\xi_{j, \epsilon, \lambda}$. By the differentiability condition, for small $\omega > 0$, we can write

$$f_{s+1}(\omega, \lambda) = \int \cdots \int Y_{s+1, \lambda}(P_{\omega, \lambda})(\underline{u}_{s+1}) \prod_{k=1}^{s+1} \left[\frac{dH_{x, \lambda}}{dP}(u_k) - 1 \right] dP(u_1) dP(u_2) \cdots dP(u_{s+1}),$$

where $\underline{u}_{s+1} := (u_1, u_2, \dots, u_{s+1})$ and $Y_{s+1, \lambda}(P_{\omega, \lambda})$ denotes an $(s+1)^{th}$ order gradient of $\tilde{\Psi}$ relative to model $\mathcal{M}_{0, \lambda}$ evaluated at $P_{\omega, \lambda}$ (see, e.g., van der Vaart, 2014). We thus find that

$$|f_{s+1}(\xi_{j, \epsilon, \lambda}, \lambda)| \leq \sup_{0 < \omega < m\epsilon} |f_{s+1}(\omega, \lambda)| \leq \sup_{0 < \omega < m\epsilon} \|Y_{s+1, \lambda}(P_{\omega, \lambda})\|_2 P_{s+1} \cdot [1 + r_{s+1}(\lambda)]^{s+1}$$

with $P^{s+1} := P \times \cdots \times P$ is a product measure of dimension $s+1$. For and λ small enough, this gives that

$$\begin{aligned} |\phi_{\epsilon, \lambda, m}^{\circ} - f_1(0, \lambda)| &\leq \left| \frac{1}{\epsilon} \sum_{j=1}^m a_{j, m} \frac{(j\epsilon)^{s+1}}{(s+1)!} f_{s+1}(\xi_{j, \epsilon, \lambda}, \lambda) \right| \leq \epsilon^s \sum_{j=1}^m |a_{j, m}| \frac{j^{s+1}}{(s+1)!} |f_{s+1}(\xi_{j, \epsilon, \lambda}, \lambda)| \\ &\leq \epsilon^s [1 + r_{s+1}(\lambda)]^{s+1} \frac{\sup_{0 < \omega < m\epsilon} \|Y_{s+1, \lambda}(P_{\omega, \lambda})\|_2 P_{s+1}}{(s+1)!} \sum_{j=1}^m j^{s+1} |a_{j, m}| \\ &= O(\epsilon^s r_{s+1}(\lambda)^{s+1}). \end{aligned}$$

Finally, we note that $f_1(0, \lambda) = \int \phi_P(u) H_{x, \lambda}(du)$ in view of the proof of Theorem 1. If $u \mapsto \phi_P(u)$ has bounded second derivatives near $u = x$, and $H_{x, \lambda}$ is a second-order kernel function, it follows from standard results in the kernel smoothing literature that the systematic deviation $f_1(0, \lambda) - \phi_P(x)$ has order $O(\lambda^{-2})$. $\square$

References

- Begun J, Hall W, Huang W, and Wellner J (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, pages 432–452.
- Bickel P, Klaassen C, Ritov Y, and Wellner J (1997). Efficient and adaptive estimation for semiparametric models. Springer.
- Bickel P and Ritov Y (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates *Sankhyā*: The Indian Journal of Statistics, Series A, pages 381–393.
- Chambaz A and van der Laan MJ (2014). Inference in targeted group-sequential covariate-adjusted randomized clinical trials. *Scandinavian Journal of Statistics*, 41(1):104–140.
- Chamberlain G (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.
- Chen X (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.
- Fornberg B (1988). Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51(184):699–706.
- Frangakis C, Qian T, Wu Z, and Díaz I (2015). Deductive derivation and Turing-computerization of semiparametric efficient estimation (with discussion). *Biometrics*.
- Geskus R and Groeneboom P (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *The Annals of Statistics*, 27(2):627–674. et al.
- Gilbert PB, Yu X, and Rotnitzky A (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in medicine*, 33(6):901–917. [PubMed: 24123289]
- Hájek J (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330.
- Hájek J (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194
- Hampel FR, Ronchetti EM, Rousseeuw PJ, and Stahel WA (2011). *Robust statistics: the approach based on influence functions*, volume 196 John Wiley & Sons
- Ichimura H and Newey W (2015). The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*
- Kiefer J and Wolfowitz J (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906
- Koshevnik Y and Levit B (1977). On a non-parametric analogue of the information matrix. *Theory of Probability & Its Applications*, 21(4):738–753
- Le Cam L (1972). Limits of experiments. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 245–261
- Luedtke A, Carone M, and van der Laan M (2015). A discussion of “Deductive derivation and Turing-computerization of semiparametric efficient estimation” by Frangakis et al. *Biometrics*
- Maathuis MH and Wellner JA (2008). Inconsistency of the mle for the joint distribution of interval-censored survival times and continuous marks. *Scandinavian Journal of Statistics*, 35(1):83–103
- Newey W (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382
- Newey WK (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168
- Pfanzagl J (1982). *Contributions to a general asymptotic statistical theory*. Springer
- Quale CM, van der Laan MJ, and Robins JR (2006). Locally efficient estimation with bivariate right-censored data. *Journal of the American Statistical Association*, 101(475):1076–1084
- Robins JM (1986). A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512
- Shen X (1997). On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591

- Stein C (1956). Efficient nonparametric testing and estimation. In Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1, pages 187–195
- Tsiatis A (2007). Semiparametric theory and missing data. Springer Science & Business Media
- van der Laan M (1996). Efficient estimation in the bivariate censoring model and repairing npml. The Annals of Statistics, 24(2):596–627
- van der Laan M and Robins J (2003). Unified methods for censored longitudinal data and causality. Springer.
- van der Laan M and Rose S (2011). Targeted learning: causal inference for observational and experimental data. Springer
- van der Laan M and Rubin D (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1)
- van der Laan MJ and Rose S (2018). Targeted learning in data science: causal inference for complex longitudinal studies. Springer
- van der Vaart A (1991). On differentiable functionals. The Annals of Statistics, pages 178–204.
- van der Vaart A (2014). Higher order tangent spaces and influence functions. Statistical Science, 29(4):679–686
- Wong W and Severini T (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. The Annals of Statistics, pages 603–632

Figure 1:
Epsilon-lambda plot of approximated values of the EIF using a secant line slope as a function of ϵ and λ in Example 1

Figure 2:

Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 1

Figure 3:

Absolute % error in the approximation of the EIF value using forward multi-point schemes with different number of points (top left: 2, top right: 3, bottom left: 4, bottom right: 5) as a function of ϵ and λ in Example 1. Color coding of entries is identical as in Figure 2

Figure 4:

Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 2 based on either the Kullback-Leibler divergence. Corresponding plot based on Hellinger distance is identical and not shown

Figure 5:
Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 3

Figure 6:

Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 4

Figure 7:
Absolute % error in the approximation of the EIF value using analytic differentiation as a function of λ in Example 4