University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2015 *Paper* 340

Computerizing Efficient Estimation of a Pathwise Differentiable Target Parameter

Mark J. van der Laan* Marco Carone†
Alexander R. Luedtke‡

*Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu
†Department of Biostatistics, University of Washington, mcarone@uw.edu
‡Division of Biostatistics, University of California, Berkeley, aluedtke@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper340

Computerizing Efficient Estimation of a Pathwise Differentiable Target Parameter

Mark J. van der Laan, Marco Carone, and Alexander R. Luedtke

Abstract

Frangakis *et al.* (2015) proposed a numerical method for computing the efficient influence function of a parameter in a nonparametric model at a specified distribution and observation (provided such an influence function exists). Their approach is based on the assumption that the efficient influence function is given by the directional derivative of the target parameter mapping in the direction of a perturbation of the data distribution defined as the convex line from the data distribution to a pointmass at the observation. In our discussion paper Luedtke *et al.* (2015) we propose a regularization of this procedure and establish the validity of this method in great generality. In this article we propose a generalization of the latter regularized numerical delta method for computing the efficient influence function for general statistical models, and formally establish its validity under appropriate regularity conditions. Our proposed method consists of applying the regularized numerical delta-method for nonparametrically-defined target parameters proposed in Luedtke *et al.* 2015 to the nonparametrically-defined maximum likelihood mapping that maps a data distribution (normally the empirical distribution) into its Kullback-Leibler projection onto the model. This method formalizes the notion that an algorithm for computing a maximum likelihood estimator also yields an algorithm for computing the efficient influence function at a user-supplied data distribution. We generalize this method to a minimum loss-based mapping. We also show how the method extends to compute the higher-order efficient influence function at an observation pair for higher-order pathwise differentiable target parameters. Finally, we propose a new method for computing the efficient influence function as a whole curve by applying the maximum likelihood mapping to a perturbation of the data distribution with score equal to an initial gradient of the pathwise derivative. We demonstrate each method with a variety of examples.

1 Introduction

Let $O_1, \ldots, O_n$ be independent and identically distributed copies of a random variable O with probability distribution P_0. Let $\mathcal{O}$ be a support of P_0. We will assume that $\mathcal{O} \subset \mathbb{R}^d$ is a Euclidean set of dimension d. It is assumed that P_0 is an element of a given set of probability distributions $\mathcal{M}$, which is called the statistical model. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a real valued statistical target parameter mapping that maps any probability distribution in the statistical model into a real number and the estimand of interest is given by $\Psi(P_0)$. The goal of this article is to construct a fully computerized efficient estimator of $\psi_0 = \Psi(P_0)$. The generalization to multidimensional target parameters Ψ is immediate, by simply applying our method to each component of Ψ.

It is assumed that Ψ is pathwise differentiable (Bickel et al., 1997) at any $P^0 \in \mathcal{M}$ for any parametric path $\{P^0_\delta : \delta\} \subset \mathcal{M}$ through P^0 at $\delta = 0$ in a user-supplied class of such paths:

$$\frac{d}{d\delta}\Psi(P^0_\delta)\Big|_{\delta=0} = P^0 D(P^0)S,$$

where S is the score of the path at $\delta = 0$ and $D(P^0)$ is a so-called gradient of the pathwise derivative at P^0. The scores and gradient are viewed as elements of the Hilbert space $L^2_0(P^0)$ of functions of O with mean zero and finite variance under P^0, endowed with inner product $\langle f, g\rangle_{P^0} = P^0 fg$, the covariance operator under P^0. Here we used the notation $Pf = \int f(o)dP(o)$. Let $\| f \|_{P^0} = \sqrt{\langle f, f\rangle_{P^0}}$ be the corresponding norm in this Hilbert space. Let $T(P^0)$ be the closure of the linear span of the set of scores generated by this class of parametric paths. This subspace of the Hilbert space $L^2_0(P^0)$ is referred to as the tangent space at P^0. The canonical gradient $D^*(P^0)$ is the unique gradient that is an element of this tangent space $T(P^0)$: $D^*(P^0) \in T(P^0)$. The canonical gradient is also called the efficient influence function. Thus $D^*(P^0)$ is either a score at P^0 itself of one of these paths $\{P^0_\delta : \delta\}$, or one can find a sequence of scores S_m such that $\| S_m - D^*(P^0) \|_{P^0} \to 0$ as $m \to \infty$. We assume that $D^*(P^0)$ is not only defined as an element of $L^2_0(P^0)$, but that $D^*(P^0)(o)$ is well defined for any possible realization of $O \sim P_0$, and

$$\| D^*(P^0) \|_\infty \equiv \sup_{o\in\mathcal{O}} | D^*(P^0)(o) | < \infty.$$

An estimator ψ_n of $\psi_0 = \Psi(P_0)$ is asymptotically efficient if and only if it is asymptotically linear at P_0 with influence function equal to the efficient influence function:

$$\psi_n - \psi_0 = P_n D^*(P_0) + o_P(1/\sqrt{n}),$$

where P_n is the empirical probability distribution so that $P_n f = \frac{1}{n}\sum_{i=1}^n f(O_i)$ (Bickel et al., 1997). For such an estimator we have $\sqrt{n}(\psi_n - \psi_0) \Rightarrow_d N(0, \sigma_0^2 = P_0 D^*(P_0)^2)$. Any regular asymptotically linear estimator has an asymptotic variance that is larger or equal than σ_0^2. One often refers to σ_0^2 as the generalized Cramer-Rao lower bound. In order for the efficient influence function to generate an achievable information bound σ_0^2 it is important that the class of submodels at P^0 in the above definition is chosen rich enough so that it generates a maximal tangent space $T(P^0)$.

A well known general method for construction of an efficient estimator is the one-step estimator (Bickel et al., 1997). Given an initial estimator $P_n^0 \in \mathcal{M}$ of P_0 one defines the one-step estimator as follows:

$$\psi_n^1 = \Psi(P_n^0) + P_n D^*(P_n^0).$$

Pathwise differentiable parameters allow a first-order Taylor expansion of the following form: for any pair $P, P^0 \in \mathcal{M}$

$$\Psi(P) - \Psi(P^0) = (P - P^0)D^*(P) + R_2(P, P^0) = -P^0 D^*(P) + R_2(P, P^0), \quad (1)$$

where $R_2(P, P^0)$ is a particular second-order term. Given $D^*(P)$, one can simply define $R_2(P, P^0)$ as $\Psi(P) - \Psi(P^0) + P^0 D^*(P)$ (Bickel et al., 1997). Given (1), and general empirical process results (van der Vaart and Wellner, 1996), it follows that ψ_n^1 is asymptotically efficient if $D^*(P_n^0)$ falls in a P_0-Donsker class with probability tending to 1, $P_0\{D^*(P_n^0) - D^*(P_0)\}^2 = o_P(1)$, and $R_2(P_n^0, P_0) = o_P(1/\sqrt{n})$.

Another general method for construction of an efficient estimator is the method of targeted maximum likelihood or, more generally, the method of targeted minimum loss-based estimation (van der Laan and Rubin, 2006; van der Laan, 2008; Rose and van der Laan, 2011). In this targeted maximum likelihood method one constructs a parametric path $\{P_n^0(\delta) : \delta\}$ dominated by P_n^0 so that

$$\frac{d}{d\delta} \log \frac{dP_n^0(\delta)}{dP_n^0}\bigg|_{\delta=0}$$

spans $D^*(P_n^0)$, and the first step TMLE-update of P_n^0 is then defined as $P_n^1 = P_n^0(\delta_n^0)$, where $\delta_n^0 = \arg\max_\delta P_n \log dP_n^0(\delta)/dP_n^0$. This process is then iterated till convergence defined by $\delta_n^k \approx 0$ at which point $P_n D(P_n^k) \approx 0$. If we denote the final update of P_n^0 with P_n^*, then the TMLE of ψ_0 is defined as the plug-in estimator $\Psi(P_n^*)$. In many examples the TMLE algorithm converges in one step or can be stopped at one or a few steps at which point $P_n D(P_n^k) = o_P(1/\sqrt{n})$ so that for practical purposes convergence has been achieved. In the more general targeted minimum loss-based estimation framework one utilizes

that $\Psi(P_0) = \Psi_1(Q_0)$ for some parameter $Q(P_0)$ and that the efficient influence function $D^*(P_0)$ can be represented as $D^*(Q_0, G_0)$ for some other nuisance parameter G_0. One now defines an initial estimator (Q_n^0, G_n^0) of (Q_0, G_0), and one replaces the $-\log dP^0(\delta)/dP^0$ loss by a user-supplied loss $L(Q_n^0(\delta))$ with submodel chosen so that $\frac{d}{d\delta} L(Q_n^0(\delta))\big|_{\delta=0}$ spans $D^*(Q_n^0, G_n^0)$. Again, given (1) and $P_n D^*(P_n^*) = 0$, it follows that

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^*(P_n^*) + R_2(P_n^*, P_0),$$

so that under the same conditions as mentioned above for the one-step estimator (with P_n^0 replaced by P_n^*), $\Psi(P_n^*)$ is asymptotically efficient.

Thus, construction of an efficient estimator of ψ_0 requires at minimal calculation of $D^*(P^0)(O_i)$, $i = 1, \ldots, n$, at an initial estimator P^0. Depending on the choice of least favorable submodel used in TMLE it may require calculation of $D^*(P^0)(o)$ at realizations o outside the sample, but for an appropriately chosen least favorable submodel it will only require $(D^*(P^m)(O_i) : i = 1, \ldots, n)$ at the m-th step.

Analytic computation of $D^*(P^0)$ can be challenging, especially in models $\mathcal{M}$ that have a tangent space at P^0 smaller than $L_0^2(P^0)$, involving calculation of a gradient of the pathwise derivative and subsequently projecting it on the tangent space $T(P^0)$ in $L_0^2(P^0)$. Even in nonparametric models where the (NP)MLE requires an implicit algorithm, typically $D^*(P^0)$ does not exists in closed form, but instead, its general definition requires inversion of an infinite dimensional linear Hilbert space operator (the so called information operator). Another minor complication is that the pathwise derivative requires coming up with an appropriate large enough class of parametric paths, although the only essential feature of this class is its corresponding tangent space.

Therefore, one can conclude that the analytic computation of the efficient influence function requires a skill that is not necessarily part of the toolkit of the applied computationally-savvy statistician that wants to compute an efficient estimator. This motivated the article (Frangakis et al., 2015) to develop a framework that allows for numerical computation of the efficient influence function given the statistical model and target parameter mapping. They proposed a numerical approximation for target parameter mappings on nonparametric statistical models based on the functional delta-method which defines the influence function of an estimator (and equivalently, a mapping defined on a nonparametric model) as a functional derivative in the direction of a convex line from the data distribution to a pointmass at an observation O_i. As pointed out in (Luedtke et al., 2015) this representation does not always apply and a regularization in which the pointmass is replaced by a kernel centered at the observation O_i was proposed and its validity established under weak regularity

conditions. The purpose of this article is to generalize this result to arbitrary statistical models. We present $D^*(P^0)(o)$ as a derivative of a function of ϵ at $\epsilon = 0$, where this function is determined by P^0. The evaluation of this function at ϵ is equivalent with computing an MLE (or, more generally, minimum loss estimator) based on an infinite sample from a simple ϵ-perturbation of P^0. This yields both an analytic representation as well as a numerical method for approximating $D^*(P^0)(o)$. In addition, this same MLE can be used for any pathwise differentiable target parameter, so that this potentially computationally intensive step does not need to be redone for each choice of target parameter. The minimization of the empirical risk (i.e., computing the MLE) or solving its score equations can be computationally challenging, but it is now a problem that can be solved computationally, without any need for expertise in efficiency theory and Hilbert space based analytic calculations, making efficient estimation much more accessible to the typical practitioner. We will suggest approximations to deal with the computational implementation of the MLE over the typically infinite dimensional model $\mathcal{M}$. Finally, we will pursue further generalizations to higher-order efficient influence functions and we show an alternative target parameter-specific perturbation of P^0 for which this method yields the whole efficient influence function as a function of O.

1.1 Organization of article

This article is organized as follows. In Section 2 we present the general method for evaluation of $D^*(P^0)(o)$, which sets up the function of ϵ determined by P^0, where evaluation of this function corresponds with maximizing a log-likelihood over a submodel of $\mathcal{M}$, and then defines $D^*(P^0)(o)$ as the derivative of this function at $\epsilon = 0$. This results in a numerical approximation by approximating this derivative with a difference at $\epsilon \approx 0$. There are a variety of steps that allows one to select simpler choices of such functions of ϵ, and they will be discussed. In Section 3 we present the main theorem and its proof. In this section, we also present sufficient conditions for the conditions of this main theorem based on a study of the log-likelihood as well as sufficient conditions based on a study of its corresponding score equations. The latter study yield an alternative analytic representation of the efficient influence function.

In Section 4 we present the general method for evaluation of $D^*(Q^0, G^0)(o)$, which sets up the function of ϵ determined by (Q^0, G^0), where evaluation of this function corresponds with minimizing an empirical risk over a subspace of the parameter space of Q, and then defines $D^*(Q^0, G^0)(o)$ as the derivative of this function at $\epsilon = 0$. The numerical approximation is obtained by approximating this derivavative with a difference at $\epsilon \approx 0$. In Section 5, analogue to Section

3, we present the theory and proof establishing the validity of this method. We also highlight an important special case of our theorem when it is known that the second-order term satisfies so called robustness w.r..t misspecification of G: $R_2((Q,G),(Q_1,G_1)) = 0$ if $G = G_1$. In that case, the numerical approximation will quickly converge to the derivative, and the conditions in our theorem are particularly weak.

Finally, in Sections 6, 7 and 8, we verify the conditions of our theorem for three examples, and discuss the practical implementation of the numerical approximation method. The examples cover parametric models, the bivariate right-censored data model, and a causal inference model. In Section 9 we provide a numerical method for computing the second-order efficient influence function in the case that the target parameter is second-order pathwise differentiable, thereby allowing for the construction of second-order one-step estimator or second-order TMLE based on such a numerical approximation. In Section 10 we present a similar method but applied to a targeted perturbation model $\{P^0_\epsilon : \epsilon\}$ with score at $\epsilon = 0$ equal to an initial gradient, which results in a numerical approximation of the efficient influence function $o \to D^*(P^0)(o)$ *at all* $o \in \mathcal{O}$. We show that the latter method is also tailored to computing a TMLE. We conclude with a discussion in Section 11.

2 General numerical method for calculation of efficient influence function, applied to data distribution P^0

In this article we will present two general methods for calculation of $D^*(P^0)(O_i)$ at an initial estimator P^0 of P_0 (we will suppress the n in P^0_n since n will be fixed throughout). In the first method we obtain a perturbation $\tilde{P}^0_\epsilon$ of P^0, while in the second more general method we obtain a perturbation $\tilde{Q}^0_\epsilon$ of the relevant part $Q^0 = Q(P^0)$. In this section we focus on the first method, and in a later Section 4 this method is generalized to the second method. The method involves a number of steps, where each step is covered by a subsection. In each subsection we will first present the step and subsequently the step will be discussed.

2.1 Step O: (Possibly) Reduce dimension of data based on P^0, without changing efficient influence function

This step can be skipped for an initial read and understanding, but becomes very practical once one starts implementing this method on high dimensional data structures O.

Consider a reduction $O_r = f(O, P^0)$ and denote the probability distribution of O_r under $O \sim P^0$ with P_r^0. Note that for each $P \in \mathcal{M}$, we can define the probability distribution P_r of $f(O, P)$ when $O \sim P$. That is, f defines a statistical model $\mathcal{M}_r$ for the reduced data structure O_r. Consider an analogue target parameter $\Psi_r : \mathcal{M}_r \to \mathbb{R}$. Suppose now that this choice of reduction f and parameter Ψ_r are chosen in such a way so that $\Psi_r(P_r^0) = \Psi(P^0)$, and that the canonical gradient of $\Psi_r : \mathcal{M}_r \to \mathbb{R}$ has a canonical gradient $D_r^*(P_r^0)(O_r)$ at P_r^0 equal to $D^*(P^0)(O)$:

$$D_r^*(P_r^0)(O_r) = D^*(P^0)(O).$$

In that case, we redefine our O_i as O_{ri}, $i = 1, \ldots, n$, P^0 as P_r^0, $\mathcal{M}$ as $\mathcal{M}_r$, Ψ as Ψ_r, and use our numerical algorithm to compute the efficient influence function of Ψ_r at P_r^0.

As we will see below, our method relies on setting a smoothing level $\lambda = \lambda(\epsilon)$. If O is very high dimensional, i.e. d is very large, then our proposed smoothing levels $\lambda(\epsilon)$ might become impractical computationally (i..e, ϵ will have to be selected closer to zero than computer precision can handle requiring special numerical methods). In these cases, such a reduction of the data from O to O_r can make the method more practical without any loss of its validity.

Let's consider an example. Let $O = (W, A, Y) \sim P^0$, $\mathcal{M}$ is the non-parametric model, and $\Psi(P) = E_P E_P(Y \mid A = 1, W)$. Suppose that the covariate vector is very high dimensional. Let $\bar{g}^0(W) = P^0(A = 1|W)$, $\bar{Q}^0(W) = E_{P^0}(Y \mid A = 1, W)$, and $Q_W^0 = P_W^0$ is the probability distribution of W under P^0. The efficient influence function at P^0 is given by (e.g., (Rose and van der Laan, 2011))

$$D^*(P^0)(o) = \frac{A}{\bar{g}^0(W)}(Y - \bar{Q}^0(W)) + \bar{Q}^0(W) - Q_W^0 \bar{Q}^0.$$

Suppose now that we define $O_r = (W_r \equiv (\bar{Q}^0(W), \bar{g}^0(W)), A, Y) \sim P_r^0$, and define $\Psi_r : \mathcal{M}_r \to \mathbb{R}$ as

$$\Psi_r(P_r) = E_{P_r} E_{P_r}(Y \mid A = 1, W_r).$$

Note that the dimension of O_r is only 4. Of course, by the same general formula, the efficient influence function of Ψ_r at P_r^0 is given by:

$$D^*(P_r^0)(O_r) = \frac{A}{g_r^0(W)}(Y - \bar{Q}_r^0(W)) + \bar{Q}_r^0(W) - Q_{W_r}^0\bar{Q}_r^0.$$

Now, note that $g_r^0(W) = P^0(A = 1 \mid \bar{Q}^0(W), \bar{g}^0(W)) = \bar{g}^0(W)$, $\bar{Q}_r^0(W) = E_{P^0}(Y \mid A = 1, \bar{Q}^0(W), \bar{g}^0(W)) = \bar{Q}^0(W)$, and $Q_{W_r}^0\bar{Q}_r^0 = Q_W^0\bar{Q}^0$. As a consequence, we have

$$D^*(P_r^0)(O_r) = \frac{A}{\bar{g}^0(W)}(Y - \bar{Q}^0(W)) + \bar{Q}^0(W) - Q_W^0\bar{Q}^0 = D^*(P^0)(O).$$

That is, the proposed data reduction O_r and target parameter Ψ_r yields the same efficient influence function $D^*(P^0)$. In general, if under P^0 certain conditional distributions depend on a subset of O only through a summary measure (depending on P^0), then one wants to include these summary measures in the reduced data O_r. This is formally demonstrated in our third example in Section 9. In the next steps, for notational convenience, we still use the notation O, Ψ and $\mathcal{M}$, assuming that these have already been redefined if indeed such a data reduction O_r was carried out.

2.2 Step I: Define perturbation of P^0 in direction of a single observation

Define a path $\{P_\epsilon^0 = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o} : \epsilon\}$, where $\Delta_{\lambda,o}$ is a probability distribution of O that 1) puts all its mass on a neighborhood $B(o:\lambda)$ around the observation o whose size is indexed by the smoothing parameter $\lambda \geq 0$ and 2) is absolute continuous w.r.t. P^0 if $\lambda > 0$, while $\Delta_{0,o}$ is the probability distribution that puts mass 1 on o. For example, $\Delta_{\lambda,o}$ could be defined as the probability distribution with density being the uniform distribution on the cube $\prod_{j=1}^d (o_j - \lambda, o_j + \lambda)$. It will be assumed that the value of λ is implied by ϵ, i.e. $\lambda = \lambda(\epsilon)$, so that P_ϵ^0 is fully defined by ϵ. Both ϵ and λ will be chosen to be close to zero, and appropriate rates for $\lambda(\epsilon)$ will be presented in our theorems.

Various remarks are in place here. Firstly, we note that for models $\mathcal{M}$ that are not nonparametric this path $\{P_\epsilon^0 : \epsilon\}$ is *not* a submodel of $\mathcal{M}$, even though its center $P^0 \in \mathcal{M}$ is in the model. Secondly, we also note that for each fixed $\lambda > 0$, the Radon-Nykodim derivative dP_ϵ^0/dP^0 exists, and for ϵ approximating 0 this Radon-Nykodim derivative approximates 1.

Thirdly, in various examples one can select $\lambda = 0$, but in examples in which $\Psi(P)$ depends on local features of P (such as density or conditional mean) and the model $\mathcal{M}$ is infinite dimensional, then it is often necessary to use a $\lambda(\epsilon) > 0$ at a rate in ϵ slower than ϵ. We suggest the following rule of thumb for deciding between $\lambda = 0$ and using an appropriate $\lambda(\epsilon) > 0$. Suppose Ψ is continuous at $P^0 \in \mathcal{M}$ w.r.t. uniform convergence of cumulative distribution functions in the following way: for a sequence $(P_m^0 : m) \subset \mathcal{M}$ for which the cumulative distribution functions F_m^0 of P_m^0 converge uniformly to the cumulative distribution function F^0 of P^0, we have $\Psi(P_m^0) \to \Psi(P^0)$ as $m \to \infty$. In that case, we can select $\lambda = 0$. If this is not the case then the continuity of Ψ at a P^0 relies on convergence of local features of the sequence P_m^0 such as convergence of the density of P_m^0 (w.r.t. P^0) to the density of P^0. Another rule of thumb one might apply is that if the MLE $\Psi(\tilde{P}_n^0)$ of $\Psi(P^0)$ under sampling from P^0 is consistent, where $\tilde{P}_n^0$ is an MLE over $\mathcal{M}$ of P^0 based on $O_1, \ldots, O_n \sim P^0$, possibly defined according to (Kiefer and Wolfowitz, 1956), then one sets $\lambda = 0$, but if the MLE requires regularization of some type, then we use a $\lambda(\epsilon) > 0$. The above rules of thumb are not meant to be theorems, but they are supposed to provide the practitioner with some practical guidance and understanding about why and when the smoothing λ may be essential. Either way, setting $\lambda = \lambda(\epsilon) > 0$ according to the rates suggested by our theorems will provide validity of the method, even if $\lambda = 0$ would have worked as well.

Finally, let's provide a formal mathematical motivation for the choice of $\lambda > 0$ which actually suggests concrete rates $\lambda(\epsilon)$ in ϵ. Suppose $\Psi(P)$ depends on P in an essential way through its density. Then a condition that allows easy verification of some of the key conditions in our theorems is that dP_ϵ^0/dP^0 converges to 1 as $\epsilon \to 0$ w.r.t. an appropriate norm. We have

$$\frac{dP_\epsilon^0}{dP^0} - 1 = (1-\epsilon) + \epsilon \frac{d\Delta_{\lambda,o}}{dP^0} - 1 = \epsilon \left(\frac{d\Delta_{\lambda,o}}{dP^0} - 1 \right).$$

This results in the following trivial but useful lemma.

Lemma 1 *Let* $\| f \|_\infty = \sup_{o \in \mathcal{O}} | f(o) |$ *be the supremum norm of a function* f. *If* $\sup_{o \in \mathcal{O}} d\Delta_{\lambda,o}/dP^0(o) < r(\lambda)$ *for some real valued function* $r(\lambda) > 0$, *then*

$$\| \frac{dP_\epsilon^0}{dP^0} - 1 \|_\infty = O(\epsilon r(\lambda)).$$

Thus, if $\lambda = \lambda(\epsilon)$ *is chosen so that* $\epsilon r(\lambda(\epsilon)) \to 0$, *then the density* dP_ϵ^0/dP^0 *converges uniformly to 1. For example, if* μ *is the Lebesgue measure,* $d\Delta_{\lambda,o}/d\mu$

is a d-variate uniform kernel density on the cube $\prod_{j=1}^{d}(o(j)-\lambda, o(j)+\lambda)$, $dP^0/d\mu > \delta$ *for some* $\delta > 0$ *on* $\prod_{j=1}^{d}(o(j)-\lambda, o(j)+\lambda)$, *then we have*

$$\frac{d\Delta_{\lambda,o}}{dP^0} \leq \frac{\lambda^{-d}}{\delta}.$$

In this case, we have supremum norm convergence of dP^0_ϵ/dP^0 *to* 1 *if* $\epsilon\lambda^{-d} \to 0$ *as* $\epsilon \to 0$, *so that we can select* $\lambda(\epsilon) >> \epsilon^{1/d}$.

In fact, to verify one of our key conditions, we might need this uniform convergence to occur at a rate such that quadratic differences involving integrals of $(dP^0_\epsilon/dP^0 - 1)^2$ will converge to zero at rate $o(\epsilon)$. Such second-order terms are naturally bounded by a constant times $\| dP^0_\epsilon/dP^0 - 1 \|^2_\infty$. In our uniform d-variate kernel example, this can thus be arranged to hold by

$$\epsilon\lambda^{-2d} = o(1), \text{ which holds if } \lambda(\epsilon) >> \epsilon^{1/2d}.$$

In essence, our theorem will demonstrate that we will be able to conclude that this choice of $\lambda(\epsilon) >> \epsilon^{1/(2d)}$, although potentially (much) too conservative, will provide guarantee that our numerical method works.

2.3 Step II: Define a submodel of our statistical model so that the efficient influence function at P^0 is still the same

In many cases one can define a smaller model $\mathcal{M}(P^0) \subset \mathcal{M}$ so that the canonical gradient of $\Psi : \mathcal{M}(P^0) \to \mathbb{R}$ at P^0 is identical to the canonical gradient $D^*(P^0)$ of $\Psi : \mathcal{M} \to \mathbb{R}$ at P^0. Specifically, the tangent space at P^0 of $\mathcal{M}(P^0)$ needs to equal the actual tangent space $T(P^0)$ at P^0 in model $\mathcal{M}$, or only exclude scores in $T(P^0)$ that are orthogonal to $D^*(P^0)$ in $L^2_0(P^0)$. By selecting $\mathcal{M}(P^0)$ smaller, our next step will be computationally less demanding. To start with one can define $\mathcal{M}(P^0) \subset \{P \in \mathcal{M} : dP/dP^0 < M\}$ so that all its probability distributions are absolutely continuous w.r.t. P^0 with a density uniformly bounded by a $M < \infty$. We assume that this step is carried out.

A common class of examples of models $\mathcal{M}$ are such that the density $dP/d\mu$ factorizes in $dP/d\mu = p_Q p_G$ for two variation independent parameters $P \to Q(P)$ and $P \to G(P)$ defined on $\mathcal{M}$ while $\Psi(P) = \Psi_1(Q(P))$ for some mapping Ψ_1. Since this factorization of the density of O implies that G is a so called orthogonal nuisance parameter of Q and thereby of $\Psi(P)$, it follows that the efficient influence function is not affected by knowledge on G. In this case, one should define $\mathcal{M}(P^0) = \{P \in \mathcal{M} : P \ll P^0, G(P) = G(P^0)\}$ as the

model in which G is treated as known and set equal to $G^0 = G(P^0)$. In greater generality, if each P in $\mathcal{M}$ is identified by two variation independent parameters $(Q(P), G(P))$, $\Psi(P) = \Psi_1(Q(P))$, and the tangent space of $G : \mathcal{M} \to \{G(P) : P \in \mathcal{M}\}$ at P^0 is orthogonal to the tangent space of Q at P^0, then one should define

$$\mathcal{M}(P^0) = \{P \in \mathcal{M} : P \ll P^0, G(P) = G(P^0)\}.$$

2.4 Step III: Define the MLE mapping at the perturbation P_ϵ^0 of P^0 in direction of single observation

We now define the MLE at the perturbation P_ϵ^0:

$$\tilde{P}_\epsilon^0 \equiv \arg \max_{P \in \mathcal{M}(P^0)} P_\epsilon^0 \log \frac{dP}{dP^0}. \tag{2}$$

It is assumed that this MLE $\tilde{P}_\epsilon^0$ exists and is an element of $\mathcal{M}(P^0)$, so that, in particular, $\tilde{P}_\epsilon^0 \ll P^0$.

Again, we have a few remarks. Firstly, because $P_\epsilon^0 \ll P^0$ for $\lambda > 0$, if $\lambda > 0$, this log-likelihood can be represented as $\int \log \frac{dP}{dP^0} \frac{dP_\epsilon^0}{dP^0} dP^0$ showing that it is a regularized likelihood in which all measures have densities w.r.t. P^0, allowing us to map a rate of convergence of dP_ϵ^0/dP^0 into a rate of convergence of $d\tilde{P}_\epsilon^0/dP^0$.

Secondly, note that if $\mathcal{M}(P^0)$ is the set of all possible probability distributions absolute continuous w.r.t. P^0, i..e. $\mathcal{M}$ is a nonparametric model, then $\tilde{P}_\epsilon^0 = P_\epsilon^0$.

Thirdly, very importantly, this MLE choice $\tilde{P}_\epsilon^0$ can be replaced by any other algorithm that maps P_ϵ^0 into a $\tilde{P}_\epsilon^0 \in \mathcal{M}$ as long as it satisfies $P_\epsilon^0 D^*(\tilde{P}_\epsilon^0) = 0$. In particular, as we will point out $\tilde{P}_\epsilon^0$ can be defined as a solution of a rich set of score equations of the above log-likelihood defining $\tilde{P}_\epsilon^0$.

In order to implement this MLE (2) one may define a finite partition $\cup_{j=1}^m \mathcal{O}_j$ of $\mathcal{O}$ (e.g., implied by a d-dimensional grid), and one approximates $\mathcal{M}(P^0)$ with $\mathcal{M}(P^0)_m$ defined by approximating each density $p = dP/dP^0$ with $P \in \mathcal{M}(P^0)$ by a histogram type-density $p^m = \sum_{j=1}^m I_{\mathcal{O}_j} p(m_j)$ that is constant within each $\mathcal{O}_j$ and equal to p at some midpoint $m_j \in \mathcal{O}_j$, $j = 1, \ldots, m$. In addition, one replaces P_ϵ^0 by such a histogram approximation $P_{\epsilon,m}^0 = (1 - \epsilon)P_m^0 + \epsilon\Delta_{\lambda,o}$ with $P_m^0 \in \mathcal{M}(P^0)_m$. One now defines the MLE $\tilde{P}_{\epsilon,m}^0 \in \mathcal{M}(P^0)_m$ as in (2). This discretized MLE corresponds now with maximizing a function that maps an m-dimensional vector into a real number (the corresponding log-likelihood value), so that one can use standard optimization routines for

finding the desired maximum. Here m does not need to be selected larger than is needed for $d_{K,L}(\tilde{P}^0_{\epsilon,m}, \tilde{P}^0_\epsilon) = O(1/n)$ under which condition the effect of m on the approximation of $D^*(P^0)(o)$ is negligible for the purpose of constructing an efficient estimator.

An alternative method for approximating the solution $\tilde{P}^0_\epsilon$ is to take a very large Monte-Carlo sample from P^0_ϵ, and compute a regularized MLE (e.g., super-learner based on the log-likelihood loss) (or regular MLE if regularization is not needed). For example, one might use cross-validation to select among many candidate estimators of $\tilde{p}^0_\epsilon$ that are an element of $\mathcal{M}(P^0)$, just like one might compute the best density estimator of $\tilde{p}^0_\epsilon$ based on a super-learner. In this manner, one literally just applies a regularized MLE.

2.5 Step IV: Evaluate differential of target parameter at small value ϵ

Evaluate:

$$D^*_\epsilon(P^0)(o) \equiv \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon}. \tag{3}$$

Under our regularity conditions,

$$D^*(P^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon},$$

so that $D^*_\epsilon(P^0)(O) \approx D^*(P^0)(o)$ for $\epsilon \approx 0$.

We make the following remarks. Firstly, if one would fix $\lambda > 0$, then, under our regularity conditions, we have:

$$\lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon} = \Delta_{\lambda,o} D^*(P^0),$$

where the right-hand side is an average of values of $D^*(P^0)(o)$ in a λ-neighborhood of o. This makes clear that it is just a matter of letting λ converge to zero slowly enough relative to $\epsilon \to 0$.

We can also provide the following analogue analytic formula for calculation of $D^*(P^0)(o)$. Let $\tilde{P}^0_{\epsilon,\lambda}$ be the perturbation model in which λ and ϵ are two separate parameters (not linked through a rate $\lambda(\epsilon)$). Under our regularity conditions, we have

$$D^*(P^0)(o) = \lim_{\lambda \to 0} \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon}. \tag{4}$$

So in order to obtain this analytic formula for $D^*(P^0)(o)$ there is no need to worry about an appropriate rate of $\lambda(\epsilon)$: one fixes $\lambda > 0$, computes the derivative w.r.t ϵ of $\epsilon \to \Psi(\tilde{P}^0_{\epsilon,\lambda})$ at $\epsilon = 0$, and finally, one takes the limit of $\lambda \to 0$ of the resulting function of λ. Note that this analytic formula for $D^*(P^0)(o)$ does not require knowing about pathwise derivatives and projections in Hilbert spaces.

Under the weak differentiability condition on Ψ along the path $\{\tilde{P}^0_{\epsilon,\lambda} : \epsilon\}$ for a fixed $\lambda > 0$, a further simplification of this analytic formula (4) is given by:

$$D^*(P^0)(o) = \lim_{\lambda \to 0} \lim_{\epsilon \to 0} d\Psi(P^0) \left(\frac{\tilde{P}^0_{\epsilon,\lambda} - P^0}{\epsilon} \right), \tag{5}$$

where $d\Psi(P^0)(h) = \frac{d}{d\delta} \Psi(P^0 + \delta h)\big|_{\delta=0}$ is the Gateaux derivative of Ψ in the direction h.

Finally, we refer to our alternative analytic formula (9) based on score equations for $\tilde{P}^0_\epsilon$.

2.6 Applying the method to multiple observations simultaneously.

Suppose we replace in the above perturbation $\{P^0_\epsilon : \epsilon\}$ of P^0 $\Delta_{\lambda,o}$ by

$$\Delta_{\lambda,(o_i:i)} \equiv \frac{1}{n} \sum_{i=1}^{n} \Delta_{\lambda,o_i},$$

the average of the kernels Δ_{λ,o_i} across all observations o_i, $i = 1, \ldots, n$, and define again:

$$\bar{D}^*_{n,\epsilon}(P^0) \equiv \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon}.$$

Under the same regularity conditions as our Theorem, we obtain:

$$\frac{1}{n} \sum_{i=1}^{n} D^*(P^0)(o_i) \equiv \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon},$$

so that $\bar{D}^*_{n,\epsilon}(P^0) \approx \frac{1}{n} \sum_{i=1}^{n} D^*(P^0)(o_i)$. This can then be used to define the one-step estimator in one numerical step:

$$\psi^1_{n,\epsilon} = \Psi(P^0_n) + \bar{D}^*_{n,\epsilon}(P^0).$$

This proves that we can construct in one computational step an asymptotically efficient estimator.

However, statistical inference will also require to estimate the variance of this estimator. A common estimator of this variance is given by the empirical variance of the influence functions:

$$\sigma_n^2 = P_n D_\epsilon^*(P_n^0)^2.$$

If n is very large, one might approximate this variance estimator by

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{J}\left\{\sum_{i\in C_j} D_\epsilon^*(P_n^0)(o_i)\right\}^2,$$

where $\cup_j C_j$ is a partitioning of the total sample $\{1,\ldots,n\}$. One could now use the above numerical method with $\Delta_{\{O_i:i\in C_j\},\lambda} = \sum_{i\in C_j}\Delta_{O_i,\lambda}$ instead of $\Delta_{\lambda,o}$ to compute $\sum_{i\in C_j} D_\epsilon^*(P^0)(O_i)$ at once. In this manner, by selecting a size for the clusters C_j one can trade-off computation time and precision of the variance estimator.

One could also estimate the variance of the one-step estimator with the bootstrap, but that might require additional regularity conditions.

3 Main theorem establishing validity of the numerical method for computing efficient influence function

We will provide the proof of the method below, and then state the resulting main theorem.

By definition of the MLE $\tilde{P}_\epsilon^0$ as a maximizer of the log-likelihood over $\mathcal{M}(P^0)$, we have that $P_\epsilon^0 S = 0$ for any score in the tangent space $T(\tilde{P}_\epsilon^0)$ of one of our paths through and at $\tilde{P}_\epsilon^0$. If $D^*(\tilde{P}_\epsilon^0)$ (known to be an element of $T(\tilde{P}_\epsilon^0)$) is itself a score of one such path, then we obtain $P_\epsilon^0 D^*(\tilde{P}_\epsilon^0) = 0$. In some examples, the efficient influence function is not necessarily a score but can only be approximated by a sequence of scores. The following result provides now the desired equation.

Result 1 *Assume either* $\| \frac{dP_\epsilon^0}{d\tilde{P}_\epsilon^0} \|_{\tilde{P}_\epsilon^0} < \infty$ *or that* $D^*(\tilde{P}_\epsilon^0)$ *is an actual score in* $T(\tilde{P}_\epsilon^0)$. *Then,*

$$P_\epsilon^0 D^*(\tilde{P}_\epsilon^0) = 0.$$

The bounded norm assumption in this result is certainly expected to hold in great generality since a $\tilde{P}_\epsilon^0$ that assigns mass close to zero to an area where P_ϵ^0 has support would make the log-likelihood in (2) negative, while $\tilde{P}_\epsilon^0$ is supposed to maximize this log-likelihood. In particular, if P^0 is discrete, then $\tilde{P}_\epsilon^0$ puts positive mass on each of the support points of P^0 so that the assumption holds.
Proof of Result 1: Since $D^*(\tilde{P}_\epsilon^0) \in T(\tilde{P}_\epsilon^0)$, there exists a sequence of scores S_m at $\tilde{P}_\epsilon^0$ so that $\| D^*(\tilde{P}_{\epsilon,\lambda}^0) - S_m \|_{\tilde{P}_\epsilon^0} \to 0$ as $m \to \infty$. We have $P_\epsilon^0 S_m = 0$ for all m. Thus,

$$\begin{aligned} P_\epsilon^0 D^*(\tilde{P}_\epsilon^0) &= P_\epsilon^0\{D^*(\tilde{P}_\epsilon^0) - S_m\} \\ &= \int \frac{dP_\epsilon^0}{d\tilde{P}_\epsilon^0}\{D^*(\tilde{P}_\epsilon^0) - S_m\} d\tilde{P}_\epsilon^0 \\ &\leq \| \frac{dP_\epsilon^0}{d\tilde{P}_\epsilon^0} \|_{\tilde{P}_\epsilon^0} \| \{D^*(\tilde{P}_\epsilon^0) - S_m\} \|_{\tilde{P}_\epsilon^0} \\ &\to 0 \text{ as } m \to \infty. \end{aligned}$$

Here we used that $\frac{dP_\epsilon^0}{d\tilde{P}_\epsilon^0} < \infty$ has a bounded norm, which holds by assumption. This proves the result. □

Combining this with identity (1) at $P = \tilde{P}_\epsilon^0$ yields

$$\Psi(\tilde{P}_\epsilon^0) - \Psi(P^0) = (P_\epsilon^0 - P^0)D^*(\tilde{P}_\epsilon^0) + R_2(\tilde{P}_\epsilon^0, P^0).$$

If $d\tilde{P}_\epsilon^0/dP^0 - 1$ converges to zero w.r.t. an appropriate norm (such as $L^2(P^0)$-norm) at a rate $o(\sqrt{\epsilon})$, then one would have that the second-order term $R_2(\tilde{P}_\epsilon^0, P^0) = o(\epsilon)$. Note also that

$$(P_\epsilon^0 - P^0)D^*(\tilde{P}_\epsilon^0) = \epsilon(\Delta_{\lambda,o} - P^0)D^*(\tilde{P}_\epsilon^0),$$

where $\Delta_{\lambda,o} f$ is an average of $f(o')$ for values o' in a λ-neighborhood of o w.r.t. the probability distribution $\Delta_{\lambda,o}$. Thus, consistency of $\tilde{P}_\epsilon^0$, some continuity of $D^*(P)(O)$ at P^0 and o, and $\lambda(\epsilon) \to 0$ should imply

$$(\Delta_{\lambda,o} - P^0)D^*(\tilde{P}_\epsilon^0) = D^*(P^0)(o) + o(1).$$

This now establishes that

$$\lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}_\epsilon^0) - \Psi(P^0)}{\epsilon} = D^*(P^0)(o).$$

This proves the following general theorem in which we state the general conditions, without providing worked out sufficient conditions yet.

Theorem 1 *Assume*

Solving efficient influence function equation: *Given P^0, for any $\epsilon \in (-\delta, \delta)$ for some $\delta > 0$, we define a $\tilde{P}^0_\epsilon \in \mathcal{M}$ that satisfies $P^0_\epsilon D^*(\tilde{P}^0_\epsilon) = 0$. If $\tilde{P}^0_\epsilon$ is defined as the MLE (2), then it suffices to assume that either $D^*(\tilde{P}^0_\epsilon)$ is a score in $T(\tilde{P}^0_\epsilon)$ or we have $\| \frac{dP^0_\epsilon}{d\tilde{P}^0_\epsilon} \|_{\tilde{P}^0_\epsilon} < \infty$, for all $\epsilon \in (-\delta, \delta)$.*

Convergence rate of MLE as $\epsilon \to 0$: $R_2(\tilde{P}^0_\epsilon, P^0) = o(\epsilon)$;

Continuity of efficient influence function at P^0 and o:

$$\lim_{\epsilon \to 0} (\Delta_{\lambda,o} - P^0) D^*(\tilde{P}^0_\epsilon) = D^*(P^0)(o).$$

Then

$$D^*(P^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon}.$$

If one fixes $\lambda > 0$ as $\epsilon \to 0$, and the conditions above hold but now with the continuity of the efficient influence function condition replaced by

$$\lim_{\epsilon \to 0} (\Delta_{\lambda,o} - P^0) D^*(\tilde{P}^0_\epsilon) = \Delta_{\lambda,o} D^*(P^0),$$

then we have

$$\lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon} = \Delta_{\lambda,o} D^*(P^0).$$

Thus, if also $\lim_{\lambda \to 0} \Delta_{\lambda,o} D^*(P^0) = D^*(P^0)(o)$, an analytic formula of $D^*(P^0)(o)$ is given by:

$$D^*(P^0)(o) = \lim_{\lambda \to 0} \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon}.$$

In the next subsection we study the regularity condition named "continuity of the efficient influence function at P^0 and o". The most important condition of Theorem 1 is $R_2(\tilde{P}^0_\epsilon, P^0) = o(\epsilon)$, which may actually require a carefully selected rate $\lambda(\epsilon)$. To establish sufficient conditions for the latter, in the subsequent subsection we will study convergence of the MLE $\tilde{P}^0_\epsilon$ to P^0 with respect to the Kullback-Leibler divergence. We will then present the resulting corollary of our general Theorem 1, presenting sufficient and concrete conditions.

We then proceed with convergence rate results for the MLE based on using that $\tilde{P}^0_\epsilon$ solves the score equations defined by the log likelihood. The latter will be used to provide convergence in supremum norm of the density $d\tilde{P}^0_\epsilon / dP^0$, as well as establish a weaker convergence of the cumulative distribution function of $\tilde{P}^0_\epsilon$ to the cumulative distribution function of P^0 for the case that $\lambda = 0$ and $\Psi(P)$ is a smooth enough function of P. It will also provide us with an alternative analytic formula for $D^*(P^0)(o)$.

3.1 Continuity of efficient influence function at P^0 and o

The following theorem establishes a sufficient condition for the required continuity of efficient influence function in P and O.

Theorem 2 *Let $B(o:\lambda)$ be the support of $\Delta_{\lambda,o}$. Let $r(\lambda)$ be a constant so that $\sup_{o'\in\mathcal{O}} d\Delta_{\lambda,o}(o')/dP^0(o') < r(\lambda)$. Assume that $\| D^*(\tilde{P}^0_\epsilon) - D^*(P^0) \|_{P^0} \to 0$ as $\epsilon \to 0$, and that*

$$\lim_{\lambda\to 0} \int D^*(P^0)(o')d\Delta_{\lambda,o}(o') = D^*(P^0)(o).$$

In addition, suppose that one of the following two assumptions (A1), (A2) holds: As $\epsilon \to 0$, either

$$(A1): \sup_{o'\in B(o:\lambda(\epsilon))} | D^*(\tilde{P}^0_\epsilon) - D^*(P^0) | (o') \to 0,$$

or

$$(A2): r(\lambda(\epsilon))^{0.5} \| D^*(\tilde{P}^0_\epsilon) - D^*(P^0) \|_{P^0} \to 0.$$

Then, we have

$$\lim_{\epsilon\to 0} (\Delta_{\lambda(\epsilon),o} - P^0)D^*(\tilde{P}^0_\epsilon) = D^*(P^0)(o).$$

Proof: In this proof $\lambda = \lambda(\epsilon)$. To start with, by our first assumption, $P^0 D^*(\tilde{P}^0_\epsilon) \to P^0 D^*(P^0) = 0$ as $\epsilon \to 0$. In addition, we have

$$\begin{aligned}\int D^*(\tilde{P}^0_\epsilon)(o')d\Delta_{\lambda,o}(o') - D^*(P^0)(o) &= \int (D^*(\tilde{P}^0_\epsilon) - D^*(P^0))(o')d\Delta_{\lambda,o}(o') \\ &\quad + \int D^*(P^0)(o')d\Delta_{\lambda,o}(o') - D^*(P^0)(o').\end{aligned}$$

The second term is covered by assumption. Using Cauchy-Schwarz, the first term can be bounded by

$$\| d\Delta_{\lambda,o}/dP^0 \|_{P^0} \| D^*(\tilde{P}^0_\epsilon) - D^*(P^0) \|_{P^0} \leq r(\lambda)^{0.5} \| D^*(\tilde{P}^0_\epsilon) - D^*(P^0) \|_{P^0},$$

which converges to zero if (A2) holds. If we assume (A1), then the first term is bounded by

$$\sup_{o'\in B(o:\lambda)} | D^*(\tilde{P}^0_\epsilon) - D^*(P^0) | (o') | .$$

This proves the statement when assuming (A1). □

Under (A1), the above theorem would hold for $\lambda = 0$. Assumption (A2) is of interest as an alternative of (A1) since it provides the desired condition without relying on uniform convergence of the efficient influence function at the MLE $\tilde{P}^0_\epsilon$ to $D^*(P^0)$.

3.2 Convergence rate of the MLE w.r.t. log-likelihood dissimilarity

The following theorem provides a rate of convergence result for the MLE $\tilde{P}_\epsilon^0$ as $\epsilon \to 0$.

Theorem 3 *Recall that, by assumption, $\mathcal{M}(P^0)$ only includes distributions P with $dP/dP^0 < M$, so that $d\tilde{P}_\epsilon^0/dP^0 < M$ for all $\epsilon \in [0, \delta)$ for some $\delta > 0$. Let $L(P, P^0) = -\log dP/dP^0$. Let $r(\lambda)$ be a rate in λ so that*

$$\| \frac{d\Delta_{\lambda,o}}{dP^0} \|_{P^0} < r(\lambda).$$

By Lemma 1, if the dimension of O is d, $\Delta_{\lambda,o}$ is absolute continuous w.r.t. Lebesgue measure μ with a multivariate uniform kernel density centered at o with bandwidth λ, $dP^0/d\mu > \delta > 0$ for some $\delta > 0$ on the support of this kernel, then this rate is given by $r(\lambda) = O(\lambda^{-d})$.

Then,

$$P^0 L(\tilde{P}_\epsilon^0, P^0) = O(\epsilon^2 r(\lambda)^2).$$

Proof: We have

$$\begin{aligned}
0 &\leq P^0 L(\tilde{P}_\epsilon^0, P^0) \\
&= (P^0 - P_\epsilon^0) L(\tilde{P}_\epsilon^0, P^0) + P_\epsilon^0 L(\tilde{P}_\epsilon^0, P^0) \\
&\leq (P^0 - P_\epsilon^0) L(\tilde{P}_\epsilon^0, P^0) \\
&= -\epsilon(\Delta_{\lambda,o} - P^0) L(\tilde{P}_\epsilon^0, P^0) \\
&= -\epsilon \int \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} L(\tilde{P}_\epsilon^0, P^0) dP^0 \\
&\leq \epsilon \| \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \|_{P^0} \| L(\tilde{P}_\epsilon^0, P^0) \|_{P^0} .
\end{aligned}$$

For the log-likelihood loss we have the property that

$$P^0 \{L(P_1, P^0)\}^2 \leq M P^0 L(P_1, P^0)$$

for some $M < \infty$ defined in terms of $\sup_o d\tilde{P}_\epsilon^0/dP^0 < \infty$ (Lemma 2 in (van der Laan et al., 2004)). This is based on Lemma in (van der Vaart, 1998) (Lemma 5.35, asymptotic statistics) stating that for any pair (P^0, P) $P^0 \log dP^0/dP \geq \int (\sqrt{dP/dP^0} - 1)^2 dP^0$. Thus, we have shown that for ϵ small enough

$$P^0 L(\tilde{P}_\epsilon^0, P^0) \leq M\epsilon \| \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \|_{P^0} \sqrt{P^0 L(\tilde{P}_\epsilon^0, P^0)},$$

which proves

$$P^0 L(\tilde{P}_\epsilon^0, P^0) = O(\epsilon^2) \parallel\parallel \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \parallel_{P^0}^2 .$$

This completes the proof. □

3.3 Corollary of Theorem 1.

Given this rate of convergence result of the MLE, and the fact that the square-root of Kullback-Leibler divergence is equivalent with the $L^2(P^0)$-norm when assuming $\limsup_{\epsilon\to 0} \parallel d\tilde{P}_\epsilon^0/dP^0 \parallel_\infty < \infty$ (van der Vaart, 1998), one will typically be able to show that $R_2(\tilde{P}_\epsilon^0, P^0) = O(\epsilon^2 r^2(\lambda))$ (by using Cauchy-Schwarz inequality). This results in the following corollary of Theorem 1.

Theorem 4 *Let $r(\lambda)$ be a rate in λ so that*

$$\parallel \frac{d\Delta_{\lambda,o}}{dP^0} \parallel_{P^0} < r(\lambda).$$

We make the following assumptions.

Solving efficient influence function equation: *Given P^0, for any $\epsilon \in (-\delta, \delta)$ for some $\delta > 0$, $\tilde{P}_\epsilon^0 \in \mathcal{M}$ is defined and it satisfies $d\tilde{P}_\epsilon^0/dP^0 < M < \infty$ for some $M < \infty$, and $P_\epsilon^0 D^*(\tilde{P}_\epsilon^0) = 0$. If $\tilde{P}_\epsilon^0$ is defined as the MLE (2), then it suffices to assume that either $D^*(\tilde{P}_\epsilon^0)$ is a score in $T(\tilde{P}_\epsilon^0)$ or we have $\parallel \frac{dP_\epsilon^0}{d\tilde{P}_\epsilon^0} \parallel_{\tilde{P}_\epsilon^0} < \infty$, for all $\epsilon \in (-\delta, \delta)$.*

Convergence rate of MLE as $\epsilon \to 0$: *Assume that*

$$R_2(\tilde{P}_\epsilon^0, P^0) < C \parallel d\tilde{P}_\epsilon^0/dP^0 - 1 \parallel_{P^0}^2 \text{ for some } C < \infty.$$

Assume that $\lambda(\epsilon)$ is chosen so that $\epsilon^2 r^2(\lambda(\epsilon)) = o(\epsilon)$;

Continuity of efficient influence function at P and o: *Assume that*

$$\parallel D^*(\tilde{P}_\epsilon^0) - D^*(P^0) \parallel_{P^0} < C \parallel d\tilde{P}_\epsilon^0/dP^0 - 1 \parallel_{P^0}$$

for some (universal in $\epsilon \in (-\delta, \delta)$) $C < \infty$, and that

$$\lim_{\lambda\to 0} \int D^*(P^0)(o') d\Delta_{\lambda,o}(o') = D^*(P^0)(o).$$

Then,

$$\| d\tilde{P}_\epsilon^0/dP^0 - 1 \|_{P^0}^2 = O(\epsilon^2 r^2(\lambda(\epsilon))).$$

In addition,

$$D^*(P^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}_\epsilon^0) - \Psi(P^0)}{\epsilon}.$$

If one fixes $\lambda > 0$ as $\epsilon \to 0$, then we have

$$\lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}_\epsilon^0) - \Psi(P^0)}{\epsilon} = \Delta_{\lambda,o} D^*(P^0)(o).$$

3.4 Convergence of MLE based on score equation of MLE

The above Theorem 4 provides a concrete rate $\lambda(\epsilon)$ that yields the validity of our numerical method. This Theorem 4 selects $\lambda(\epsilon)$ in such a way so that the density dP_ϵ^0/dP^0 converges to 1 at a specified rate in ϵ and, as a consequence, the corresponding MLE $d\tilde{P}_\epsilon^0/dP^0$ converges to 1 at the same rate.

However, there are some issues we have not addressed yet. Suppose that our target parameter Ψ allows the choice $\lambda = 0$. Then, the previous theorem is not applicable, but Theorem 1 still is. In this case, dP_ϵ^0/dP^0 is not even defined. However, the cumulative distribution function F_ϵ^0 of P_ϵ^0 converges at rate ϵ to the cumulative distribution function F^0 of P^0. In this subsection we provide a theorem that allows us to prove that this uniform convergence of cumulative distribution function yields the uniform convergence of $\tilde{F}_\epsilon^0$ to F^0 at the same rate ϵ.

The method below will be formulated to be general enough so that one can also use it to establish uniform convergence of $d\tilde{P}_\epsilon^0/dP^0$ to 1 at a rate for the case that we use a $\lambda(\epsilon) > 0$. Such a result might not be necessary for applying Theorem 4, but nonetheless it tells us about how well our numerical algorithm will converge. In addition, one might be defining $\tilde{P}_\epsilon^0$ as a solution of the score equations, instead of as an MLE (even though these might agree to be the same), so that one should also be able to establish the desired convergence of its density $d\tilde{P}_\epsilon^0/dP^0$ based on the score equations.

Let $\tilde{p}_\epsilon^0 = d\tilde{P}_\epsilon^0/dP^0$ be the density w.r.t. P^0. Let $P_\epsilon^0 S_h(\tilde{p}_\epsilon^0) = 0$ be a score equation for the MLE $\tilde{P}_\epsilon^0$ obtained by differentiating the log likelihood in (2) along a path $\{\tilde{P}_{\epsilon,h,\delta}^0 : \delta\}$ (dominated by P^0) through $\tilde{P}_\epsilon^0$ at $\delta = 0$ and score $S_h(\tilde{p}_\epsilon^0)$ at $\delta = 0$. One can select a whole collection of such paths and we will denote the index set containing all possible choices h with $\mathcal{H}$. We assume that this set of scores is chosen rich enough so that the equation $P_\epsilon^0 S_h(\tilde{p}_\epsilon^0) = 0$ for all $h \in \mathcal{H}$ uniquely identifies $\tilde{p}_\epsilon^0$ among all densities in $\mathcal{M}(P^0)$ (w.r.t. P^0).

This is crucial in order to be able to establish the invertibility condition on U_1 in our proof below.

Recall that $O \in \mathbb{R}^d$ and that $\mathcal{O}$ is a support of $O \sim P_0$. Let $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$. Let $(D(\mathcal{O}), \| \cdot \|)$ be a Banach space of multivariate real valued functions $f : \mathcal{O} \subset \mathbb{R}^d \to \mathbb{R}$ containing all densities of $\mathcal{M}$ w.r.t. P^0, endowed with norm $\| \cdot \|$. The latter norm could be the supremum norm $\| f \| = \sup_{o \in \mathcal{O}} \mid f(o) \mid$, the $L^2(P^0)$-norm $\| f \| = \sqrt{\int f^2(o) dP^0(o)}$, or the supremum norm $\| \int_0^{\cdot} f dP^0 \|_\infty$ of the corresponding cumulative distribution functions.

One should select the norm $\| \cdot \|$ so that $\| p^0_\epsilon - p^0 \|$ converges to zero at rate ϵ, and the result of the theorem will then establish this same or similar rate of convergence for $\| \tilde{p}^0_\epsilon - p^0 \|$. So if one selects $\lambda = 0$, then one might select the supremum norm of the cumulative distribution function, which is the weakest norm among the three examples above. However, if Ψ depends on local features and one thus selects a $\lambda > 0$, then one needs to select one of the density norms so that we get the desired convergence of the density of $\tilde{P}^0_\epsilon$. Another important point is that the stronger the available convergence result for $P^0_\epsilon - P^0$ and thus the stronger the norm $\| \cdot \|$ one selects, the weaker the required Frechet differentiability condition (6) below is which is defined in terms of this norm $\| \cdot \|$.

Given P^0, let $U(\tilde{p}, \epsilon) \equiv (P^0_\epsilon S_h(\tilde{p}) : h \in \mathcal{H})$. Let $\ell^\infty(\mathcal{H})$ be the class of real valued function $f : \mathcal{H} \to \mathbb{R}$ endowed with the supremum norm. We assume that $U : (D(\mathcal{O}), \| \cdot \|_1) \times [-\delta^*, \delta^*] \to \ell^\infty(\mathcal{H})$. That is, U is a mapping that maps any function in $D(\mathcal{O})$ and real number in $(-\delta^*, \delta^*)$ into a function in $\ell^\infty(\mathcal{H})$.

We have $U(\tilde{p}^0_\epsilon, \epsilon) = 0$ and $U(p^0, 0) = 0$, where $p^0 = dP^0/dP^0 = 1$, but for notational ease we denote it with p^0. This yields the starting identity:

$$\begin{aligned} U(\tilde{p}^0_\epsilon, 0) - U(p^0, 0) &= -\{U(\tilde{p}^0_\epsilon, \epsilon) - U(\tilde{p}^0_\epsilon, 0)\} \\ &= -\epsilon(\Delta_{\lambda,o} - P^0) S(\tilde{p}^0_\epsilon), \end{aligned}$$

where we define $S(\tilde{p}) = (S_h(\tilde{p}) : h \in \mathcal{H})$. Let $U_1(p^0, 0)(f) = \frac{d}{d\delta} U(p^0 + \delta f, 0)\big|_{\delta=0}$ be the Gateaux derivative of $p \to U(p, 0)$ at p^0 in the direction $f \in D(\mathcal{O})$. We note that $U_1(p^0, 0) : (D(\mathcal{O}), \| \cdot \|_1) \to \ell^\infty(\mathcal{H})$. We assume that U is Frechet differentiable in its first coordinate at p^0 in the sense that $U_1(p^0, 0)$ yields the desired linear approximation in the following uniform sense:

$$\lim_{\epsilon \to 0} \frac{\| U(\tilde{p}^0_\epsilon, 0) - U(p^0, 0) - U_1(p^0, 0)(\tilde{p}^0_\epsilon - p^0) \|_\infty}{\| \tilde{p}^0_\epsilon - p^0 \|_1} = 0. \tag{6}$$

So we obtained the following:

$$U_1(p^0,0)(\tilde{p}^0_\epsilon - p^0) = -\epsilon(\Delta_{\lambda,o} - P^0)S(\tilde{p}^0_\epsilon) + o(\parallel \tilde{p}^0_\epsilon - p^0 \parallel_1).$$

We also assume that the linear mapping $U_1(p^0,0) : (D(\mathcal{O}), \parallel \cdot \parallel_1) \to \ell^\infty(\mathcal{H})$ has a bounded inverse $U_1(p^0,0)^{-1} : \ell^\infty(\mathcal{H}) \to (D(\mathcal{O}), \parallel \cdot \parallel_1)$. Then, we obtain:

$$\tilde{p}^0_\epsilon - p^0 = -\epsilon U_1(p^0,0)^{-1}(\Delta_{\lambda,o} - P^0)S(\tilde{p}^0_\epsilon) + o(\parallel \tilde{p}^0_\epsilon - p^0 \parallel_1). \tag{7}$$

We also assume that for a specified bound $r_1(\lambda)$

$$\sup_{\epsilon \in (-\delta^*, \delta^*)} \sup_{h \in \mathcal{H}} \mid (\Delta_{\lambda,o} - P^0)S_h(\tilde{p}^0_\epsilon) \mid = O(r_1(\lambda)). \tag{8}$$

Note that if $\lambda > 0$, and one of the density norms $\parallel \cdot \parallel$ is chosen, using the Cauchy-Schwarz inequality, we can bound the left-hand side in (8) with:

$$\parallel \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \parallel_{P^0} \sup_{h \in \mathcal{H}, \epsilon \in (-\delta^*, \delta^*)} \parallel S_h(\tilde{p}^0_\epsilon) \parallel_{P^0} .$$

The first factor can be bounded by $O(r(\lambda)^{0.5})$ (see Theorem 4), while convergence of $\tilde{p}^0_\epsilon - p^0$ should provide a bound $O(1)$ on the second factor, so that for the case that $\lambda > 0$ and a density norm is selected $r_1(\lambda) = r(\lambda)^{0.5}$.

Under the assumption (8) we have

$$\parallel U_1(p^0,0)^{-1}(\Delta_{\lambda,o} - P^0)S(\tilde{p}^0_\epsilon) \parallel = O(r_1(\lambda)).$$

Taking the $\parallel \cdot \parallel$-norm on both sides of (7) yields then:

$$\parallel \tilde{p}^0_\epsilon - p^0 \parallel_1 = O(\epsilon r_1(\lambda)) + o(\parallel \tilde{p}^0_\epsilon - p^0 \parallel_1).$$

This implies

$$\parallel \tilde{p}^0_\epsilon - p^0 \parallel_1 = O(\epsilon r_1(\lambda)).$$

We state this convergence of the density $\tilde{p}^0_\epsilon$ to p^0 w.r.t $\parallel \cdot \parallel$-norm in the theorem below, but we first proceed with the derivation of an analytic formula for $D^*(P^0)(o)$.

Suppose now that λ is fixed (not a function of ϵ). Then, we have obtained:

$$\tilde{p}^0_{\epsilon,\lambda} - p^0 = -\epsilon U_1(p^0,0)^{-1}(\Delta_{\lambda,o} - P^0)S(\tilde{p}^0_\epsilon) + o(\epsilon).$$

To start with, assume the following for fixed $\lambda > 0$: if $\parallel \tilde{p}^0_{\epsilon,\lambda} - p^0 \parallel = O(\epsilon)$, then $\sup_{h \in \mathcal{H},o} \mid S_h(\tilde{p}^0_\epsilon) - S_h(p^0) \mid (o) = o(1)$. Then, the right-hand side of the above displayed equation is approximated by $-\epsilon U_1(p^0,0)^{-1}(\Delta_{\lambda,o}S(p^0)) + o(\epsilon)$.

In addition, suppose Ψ is (Hadamard) differentiable at P^0 in the following sense: if $(\tilde{p}^0_{\epsilon,\lambda} - p^0)/\epsilon$ converges w.r.t. $\|\cdot\|$ to f as $\epsilon \to 0$, then

$$\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0) = d\Psi(P^0)(\tilde{p}^0_{\epsilon,\lambda} - p^0) + o(\epsilon).$$

Then, it follows

$$\frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon} = -d\Psi(P^0)U_1(p^0,0)^{-1}(f_{\lambda,o}) + o(1),$$

where $f_{\lambda,o} \equiv \Delta_{\lambda,o}S(p^0)$. Finally, assume that

$$\lim_{\lambda\to 0}\sup_{h\in\mathcal{H}} \mid \Delta_{\lambda,o}S_h(p^0) - S_h(p^0)(o) \mid = 0.$$

Then, we have shown

$$\lim_{\lambda\to 0}\lim_{\epsilon\to 0} \frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon} = -d\Psi(P^0)U_1(p^0,0)^{-1}(S(p^0)(o)).$$

Notice that for a fixed o, $h \to S_h(p^0)(o) \in \ell^\infty(\mathcal{H})$, so that indeed this inverse $U_1(p^0,0)^{-1}(S(p^0)(o))$ is well defined. Since we also showed that the left-hand side equals $D^*(P^0)(o)$ this proves the following analytic formula for the efficient influence function

$$D^*(P^0)(o) = -d\Psi(P^0)U_1(p^0,0)^{-1}(S(p^0)(o)).$$

In some statements below we use the notation $\tilde{p}^0_{\epsilon,\lambda}$ to stress the dependence on λ.

Theorem 5 *Given P^0, let $U(\tilde{p}^0_\epsilon, \epsilon) \equiv (P^0_\epsilon S_h(\tilde{p}^0_\epsilon) : h \in \mathcal{H})$ be a collection of score equations solved by $\tilde{p}^0_\epsilon = d\tilde{P}^0_\epsilon/dP^0$ for a given ϵ.*

Let $O \in \mathbb{R}^d$ and let $\mathcal{O}$ be a support of $O \sim P_0$. Let $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$. Let $(D(\mathcal{O}), \|\cdot\|_1)$ be a Banach space of multivariate real valued functions $f : \mathcal{O} \subset \mathbb{R}^d \to \mathbb{R}$ containing all densities of $\mathcal{M}$ w.r.t. P^0, endowed with a norm $\|\cdot\|$. Let $\ell^\infty(\mathcal{H})$ be the class of real valued function $f : \mathcal{H} \to \mathbb{R}$ endowed with the supremum norm. We assume that $U : (D, \|\cdot\|_1) \times [-\delta^*, \delta^*] \to \ell^\infty(\mathcal{H})$.

We make the following assumptions:

Solves score equation: *for each $\epsilon \in (-\delta^*, \delta^*)$, $\tilde{p}^0_\epsilon$ is well defined and solves $U(\tilde{p}^0_\epsilon, \epsilon) = 0$ and, in particular, $U(p^0, 0) = 0$;*

Frechet differentiability at P^0: *Let $U_1(p^0,0)(f) = \frac{d}{d\delta}U(p^0+\delta f,0)\big|_{\delta=0}$ be the Gateaux derivative in direction $f \in D(\mathcal{O})$. We note that $U_1(p^0,0) : (D(\mathcal{O}), \| \cdot \|_1) \to \ell^\infty(\mathcal{H})$. We assume that U is Frechet differentiable in its first coordinate at p^0 in the sense that $U_1(p^0,0)$ yields the desired linear approximation in the following uniform sense:*

$$\lim_{\epsilon\to 0} \frac{\| U(\tilde{p}^0_\epsilon,0) - U(p^0,0) - U_1(p^0,0)(\tilde{p}^0_\epsilon - p^0) \|_\infty}{\| \tilde{p}^0_\epsilon - p^0 \|_1} = 0.$$

Bounded inverse of the derivative at P^0: *Assume that the linear mapping $U_1(p^0,0) : (D(\mathcal{O}), \| \cdot \|_1) \to \ell^\infty(\mathcal{H})$ has a bounded inverse $U_1(p^0,0)^{-1} : \ell^\infty(\mathcal{H}) \to D(\mathcal{O})$.*

Consistency condition: *Assume that for some $\lambda \to r_1(\lambda)$, we have*

$$\sup_{\epsilon\in(-\delta^*,\delta^*)} \sup_{h\in\mathcal{H}} | (\Delta_{\lambda,o} - P^0)S_h(\tilde{p}^0_{\epsilon,\lambda}) | = O(r_1(\lambda)).$$

Then,

$$\| \tilde{p}^0_{\epsilon,\lambda} - p^0 \| = O(\epsilon r_1(\lambda)).$$

Suppose now that the above conditions and the next two conditions hold for a fixed $\lambda > 0$. Firstly, if $\| \tilde{p}^0_\epsilon - p^0 \| = O(\epsilon)$, then

$$\sup_{h\in\mathcal{H},o} | S_h(\tilde{p}^0_{\epsilon,\lambda}) - S_h(p^0) | (o) \to 0,$$

as $\epsilon \to 0$. Secondly, Ψ is (Hadamard) differentiable at P^0 in the following sense: if $(\tilde{p}^0_{\epsilon,\lambda} - p^0)/\epsilon$ converges w.r.t. $\| \cdot \|$ to f as $\epsilon \to 0$, then

$$\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0) = d\Psi(P^0)(\tilde{p}^0_{\epsilon,\lambda} - p^0) + o(\epsilon).$$

Finally, assume that

$$\lim_{\lambda\to 0} \sup_{h\in\mathcal{H}} | \Delta_{\lambda,o}S_h(p^0) - S_h(p^0)(o) | = 0.$$

Then,

$$\begin{aligned} D^*(P^0)(o) &= \lim_{\lambda\to 0}\lim_{\epsilon\to 0} \frac{\Psi(\tilde{P}^0_{\epsilon,\lambda}) - \Psi(P^0)}{\epsilon} \\ &= -d\Psi(P^0)U_1(p^0,0)^{-1}(S(p^0)(o)). \end{aligned} \tag{9}$$

The analytic formula (9) for $D^*(P^0)(o)$ is the generalization of the formula $df(\beta^0)\{-P^0\frac{d}{d\beta^0}S(\beta^0)\}^{-1}S(\beta^0)(o)$ for the efficient influence function of $f(\beta)$ for a parametric model $\{p_\beta : \beta\}$ at β^0 in terms of the gradient $df(\beta^0)$ of f and the inverse of the information matrix applied to the score vector $S(\beta^0))(o)$.

As with the Kullback-Leibler divergence result for the MLE, the convergence of $\tilde{p}^0_\epsilon - p^0$ to zero at rate $\epsilon r_1(\lambda)$ will typically immediately imply that $R_2(\tilde{P}^0_\epsilon, P^0) = O(\epsilon^2 r_1(\lambda)^2)$, thereby providing a rate $\lambda(\epsilon)$ for which $\epsilon^2 r_1(\lambda)^2 = o(\epsilon)$, as required in our general Theorem 1.

3.5 Existence of desired MLE based on implicit function theorem

In the previous subsections we established a rate of convergence of the MLE $\tilde{p}^0_\epsilon$ to p^0 w.r.t. Kullback-Leibler divergence (and thereby $L^2(P^0)$-norm) and a user-supplied norm $\| \cdot \|$, respectively, where this rate is expressed in terms of ϵ and $\lambda(\epsilon)$. This allowed us to provide a slow enough rate $\lambda(\epsilon)$ for which $\| \tilde{p}^0_\epsilon - p^0 \|$ converges to zero fast enough so that $R_2(\tilde{P}^0_\epsilon, P^0) = o(\epsilon)$, a crucial condition of Theorem 1.

Another key condition of our general Theorem 1 is the actual existence of the MLE (or solution of score equation) $\tilde{P}^0_\epsilon$ solving $P^0_\epsilon D^*(\tilde{P}^0_\epsilon) = 0$, given P^0 and ϵ. Of course, given a particular $\mathcal{M}(P^0)$, one might be able to explicitly establish this result.

The goal of this subsection is to utilize the so called implicit function theorem in order to establish this existence. This will now rely on stronger differentiability conditions on the score equation, but these conditions are still very reasonable if we use a supremum norm $\| dP/dP^0 \|_\infty = \sup_{o\in\mathcal{O}} | dP/dP^0 | (o)$ on the density, and a fixed $\lambda > 0$.

The following theorem is an immediate consequence of the implicit function theorem (see e.g. Chapter 6 in (van der Laan, 1996b)).

Theorem 6 *In this theorem we will let $\lambda > 0$ be fixed and thus be disentangled from ϵ. Given P^0, consider the following function in $(\epsilon, \tilde{p})$: $U(\tilde{p}, \epsilon) \equiv (P^0_\epsilon S_h(\tilde{p}) : h \in \mathcal{H})$, where $P^0_\epsilon = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o}$. Let $O \in \mathbb{R}^d$ and let $\mathcal{O}$ be a support of $O \sim P_0$. Let $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$. Let $(D(\mathcal{O}), \| \cdot \|_\infty)$ be a Banach space of multivariate real valued functions $f : \mathcal{O} \subset \mathbb{R}^d \to \mathbb{R}$ containing all densities of $\mathcal{M}$ w.r.t. P^0, endowed with the supremum norm over $\mathcal{O}$. Let $(D(\mathcal{O}), \| \cdot \|_\infty) \times [-\delta^*, \delta^*])$ be the product Banach space endowed with the max norm: $\| (f, \epsilon) \| = \max(\| f \|_\infty, | \epsilon |)$. Let $\ell^\infty(\mathcal{H})$ be the class of real valued function $f : \mathcal{H} \to \mathbb{R}$ endowed with the supremum norm. Note that $U : (D, \| \cdot \|_\infty) \times [-\delta^*, \delta^*] \to \ell^\infty(\mathcal{H})$.*

Beyond this general definition of U, we make the following assumptions:

efficient influence function spanned by score equations: *Assume that a solution $\tilde{p}^0_\epsilon$ of $U(\tilde{p}, \epsilon) = 0$ also satisfies $P^0_\epsilon D^*(\tilde{p}^0_\epsilon) = 0$, and that $U(p^0, 0) = 0$.*

Continuous Frechet differentiability of U at $(p^0, 0)$: *We assume that U is Frechet differentiable at any $(\tilde{p}, \epsilon)$ in neighborhood of $(p^0, 0)$:*

$$\lim_{\delta \to 0} \sup_{\|f,e\| \leq 1} \frac{\| U(\tilde{p} + \delta f, \epsilon + \delta e) - U(\tilde{p}, \epsilon) - \delta dU(\tilde{p}, \epsilon)(f, e) \|_\infty}{\delta} = 0.$$

In addition, we assume that the derivative $dU(p^0, 0)$ is continuous at $(p^0, 0)$: if $(\tilde{p}_n, \epsilon_n)$ converges to $(\tilde{p}, \epsilon)$ w.r.t $\| \cdot \|$, then

$$\sup_{\|(f,e)\|<1} \| dU(\tilde{p}_n, \epsilon_n)(f, e) - dU(\tilde{p}, \epsilon)(f, e) \|_\infty \to 0.$$

Let $U_1(p^0, 0)(f) = \frac{d}{d\delta} U(p^0 + \delta f, 0)\big|_{\delta=0}$ be the Gateaux derivative in the direction $f \in D(\mathcal{O})$ and let $U_2(p^0, 0)(e) = \frac{d}{d\epsilon} U(p^0, \epsilon)\big|_{\epsilon=0}$. We note that the derivative $dU(p^0, 0) : (D\mathcal{O}), \| \cdot \|_\infty) \times [-\delta^*, \delta^*]$ is a linear operator given by

$$dU(p^0, 0)(f, e) = U_1(p^0, 0)(f) + U_2(p^0, 0)(e).$$

We also note that $U_1(p^0, 0) : (D(\mathcal{O}), \| \cdot \|_\infty) \to \ell^\infty(\mathcal{H})$.

Bounded invertibility of derivative: *Assume that the linear mapping $U_1(p^0, 0) : (D(\mathcal{O}), \| \cdot \|_\infty) \to \ell^\infty(\mathcal{H})$ has a bounded inverse $U_1(p^0, 0)^{-1} : \ell^\infty(\mathcal{H}) \to (D(\mathcal{O}), \| \cdot \|_\infty)$.*

Then there are open neighborhoods $\mathcal{A}_0 \subset [-\delta^*, \delta^*]$ of 0 and $\mathcal{B}_0 \subset (D(\mathcal{O}), \| \cdot \|_\infty)$ of p^0 such that for each $\epsilon \in \mathcal{A}_0$ there is a unique $\tilde{p}^0_\epsilon \in \mathcal{B}_0$ such that $U(\tilde{p}^0_\epsilon, \epsilon) = 0$. Moreover, if we define $\Theta : \mathcal{A}_0 \to \mathcal{B}_0$ so that $\tilde{p}^0_\epsilon = \Theta(\epsilon)$ for $\epsilon \in \mathcal{A}_0$, then for $\mathcal{A}$ and $\mathcal{B}$ small enough, Θ is continuously differentiable mapping from $\mathcal{A}$ into $\mathcal{B}$. Its derivative is given by:

$$\Theta'(\epsilon) = -(U_1(\Theta(\epsilon), \epsilon))^{-1} U_2(\Theta(\epsilon), \epsilon).$$

The important implication of this theorem is that under the stated differentiability and invertibility condition on the score equation $U()$ at a fixed $\lambda > 0$, one now knows that for a given $\epsilon \approx 0$, there exists a unique $\tilde{p}^0_\epsilon$ in a neighborhood of $\tilde{p}^0$ and the mapping $\epsilon \to \tilde{p}^0_\epsilon$ is continuously differentiable as a mapping from a neighborhood of 0 into the Banach space endowed with

the supremum norm. If for a given $\epsilon \approx 0$, one finds a solution of the score equation, one will have to verify that it is close to p^0, since only then we have the guarantee that this will be this unique solution in the neighborhood of p^0 that is very smooth in ϵ. So due to this theorem, we now know that if we select ϵ small enough, we will be able to find a solution $\tilde{p}^0_\epsilon$ close to p^0, thereby establishing the existence. Moreover, the continuous differentiability of this local solution teaches us immediately that for any fixed $\lambda > 0$, we have

$$\| \tilde{p}^0_\epsilon - p^0 \|_\infty = O(\epsilon).$$

So it establishes that for each fixed $\lambda > 0$ the solution $\tilde{p}^0_\epsilon$ converges in supremum norm to p^0 at rate $O(\epsilon)$. Under the conditions stated in this theorem for fixed λ, and $\lim_{\lambda \to 0} \sup_h \mid \Delta_{\lambda,o} S_h(p^0) - S_h(p^0)(o) \mid = 0$, the analytic formula (9) follows as well.

4 General numerical method for calculation of efficient influence function, applied to relevant part of P^0

As before, the method involves a number of steps, and below we provide these steps. Step 0, and the discussions and remarks in Section 2 are equally relevant for this method, but are not repeated here.

4.1 Step I: Define relevant part of data generating distribution and loss function

Let $\Psi : \mathcal{M} \to \mathbb{R}$ be such that $\Psi(P) = \Psi_1(Q(P))$ for some Ψ_1 and $Q : \mathcal{M} \to \mathcal{F} \equiv \{Q(P) : P \in \mathcal{M}\}$. Let $L(Q)(O)$ be a loss function so that

$$Q(P_0) = \arg\min_{Q \in \mathcal{F}} P_0 L(Q).$$

Let $Q^0 = Q(P^0)$. We assume that there exists a collection of paths $\{Q^0_\delta : \delta\} \subset \mathcal{F}$ in $\mathcal{F}$ through Q^0 at $\delta = 0$ so that the closure of the linear span of its generalized scores $\frac{d}{d\delta} L(Q^0_\delta)\big|_{\delta=0}$ in $L^2_0(P^0)$ contains $D^*(P^0)$.

The efficient influence function will depend on P^0 through Q^0 and a nuisance parameter $G^0 = G(P^0)$, so that we can also denote it with $D^*(Q^0, G^0)$. For most loss functions $L(Q)$, the above condition that $D^*(P^0)$ needs to be a score of this loss L will typically require that G is a nuisance parameter whose tangent space at P^0 is orthogonal to the tangent space of Q at P^0. However,

we suspect that there are cases in which the loss $L(Q) = L_{G^0}(Q)$ depends on G^0 and where one can still generate the $D^*(P^0)$ as score of $L_{G^0}(Q^0_\delta)$ at $\delta = 0$ even when G is not orthogonal to Q.

4.2 Step I: Define a perturbation of P^0 in the direction of a single observation

We define the same path $P^0_\epsilon = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o}$ as before, where $\lambda = \lambda(\epsilon)$. Given (Q^0, G^0) defined above, we can select P^0 as any probability distribution that is compatible with these two initial estimators (Q^0, G^0).

4.3 Define a submodel of our statistical model for relevant part so that the efficient influence function at P^0 is still the same

As discussed earlier, in many cases one can define a smaller model $\mathcal{M}(P^0) \subset \mathcal{M}$ so that the canonical gradient of $\Psi : \mathcal{M}(P^0) \to \mathbb{R}$ at P^0 is identical to the canonical gradient $D^*(P^0)$. Let $\mathcal{F}(P^0) = \{Q(P) : P \in \mathcal{M}(P^0)\} \subset \mathcal{F}$ be the corresponding model for Q_0.

4.4 Define the Minimum Loss Estimator (MLE) mapping at the perturbation of P^0

We now define the minimum loss estimator (MLE) mapping applied to the perturbation P^0_ϵ:

$$\tilde{Q}^0_\epsilon \equiv \arg \min_{Q \in \mathcal{F}(P^0)} P^0_\epsilon L(Q). \tag{10}$$

It is assumed that this MLE $\tilde{Q}^0_\epsilon$ exists and is an element of $\mathcal{F}(P^0)$.

This choice $\tilde{Q}^0_\epsilon$ can be replaced by any $\epsilon \to \tilde{Q}^0_\epsilon$ that satisfies that for each $\epsilon \in (-\delta^*, \delta^*)$ (for some $\delta^* > 0$)

$$P^0_\epsilon D^*(\tilde{Q}^0_\epsilon, G^0) = 0.$$

4.5 Evaluate differential of target parameter at small value ϵ

Evaluate

$$D^*_\epsilon(P^0)(o) \approx \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon} \text{ for } \epsilon \approx 0,$$

where we remind the reader that $\lambda = \lambda(\epsilon)$ is a function of ϵ so that $\epsilon \approx 0$ implies $\lambda(\epsilon) \approx 0$.

Under our regularity conditions,

$$D^*(P^0)(o) \equiv \lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon},$$

so that $D^*_\epsilon(P^0)(o)$ indeed approximates $D^*(P^0))(o)$.

If one would fix $\lambda > 0$, then we have:

$$\lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon} = \Delta_{\lambda,o} D^*(P^0).$$

5 Main Theorem for establishing validity of method applied to relevant part of P^0

The validity of the method is proven as follows. Firstly, we have $P^0_\epsilon D^*(\tilde{Q}^\epsilon_0, G^0) = 0$. The analogue of identity of (1) is given by: for any pair $(\tilde{Q}^0, Q^0 = Q(P^0))$ and $G^0 = G(P^0)$, we have

$$\Psi_1(\tilde{Q}^0) - \Psi_1(Q^0) = -P^0 D^*(\tilde{Q}^0, G^0) + R_2((\tilde{Q}^0, G^0), (Q^0, G^0)), \tag{11}$$

where $R_2((Q_1, G_1), (Q_2, G_2))$ is a second-order term in differences $(Q_1 - Q_2)$ and $(G_1 - G_2)$. This yields:

$$\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0) = (P^0_\epsilon - P^0) D^*(\tilde{Q}^0_\epsilon, G^0) + R_2((\tilde{Q}^0_\epsilon, G^0), (Q^0, G^0)). \tag{12}$$

As before we assume that $\lambda = \lambda(\epsilon)$ is chosen so that the strong consistency condition $R_2((\tilde{Q}^0_\epsilon, G^0), (Q^0, G^0)) = o(\epsilon)$ holds. This gives then

$$\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0) = \epsilon(\Delta_{\lambda,o} - P^0) D^*(\tilde{Q}^0_\epsilon, G^0) + o(\epsilon).$$

Finally, we need a continuity condition on the efficient influence function so that

$$(\Delta_{\lambda,o} - P^0) D^*(\tilde{Q}^0_\epsilon, G^0) = D^*(Q^0, G^0)(o) + o(1). \tag{13}$$

This proves the following analogue of Theorem 1.

Theorem 7 *Assume*

Solving efficient influence function equation: *Given P^0, for any $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$, we define a $\tilde{Q}^0_\epsilon \in \mathcal{F}(P^0)$ that satisfies $P^0_\epsilon D^*(\tilde{Q}^0_\epsilon, G^0) = 0$.*

Convergence rate of MLE as $\epsilon \to 0$: $R_2((\tilde{Q}^0_\epsilon, G^0), (Q^0, G^0)) = o(\epsilon)$;

Continuity of efficient influence function at P^0 and O:

$$\lim_{\epsilon \to 0}(\Delta_{\lambda,o} - P^0)D^*(\tilde{Q}^0_\epsilon, G^0) = D^*(Q^0, G^0)(o).$$

Then

$$D^*(Q^0, G^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon}.$$

If one fixes $\lambda > 0$ as $\epsilon \to 0$, and the conditions above hold but now with the continuity of the efficient influence function condition replaced by

$$\lim_{\epsilon \to 0}(\Delta_{\lambda,o} - P^0)D^*(\tilde{Q}^0_\epsilon, G^0) = \Delta_{\lambda,o}D^*(Q^0, G^0),$$

then we have

$$\lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi(Q^0)}{\epsilon} = \Delta_{\lambda,o}D^*(Q^0, G^0)(o).$$

5.1 Continuity of efficient influence function condition

The generalization of Theorem 2 is immediate and is given by the following.

Theorem 8 *Let $B(o : \lambda)$ be the support of $\Delta_{\lambda,o}$. Let $r(\lambda)$ be a constant so that $\sup_{o' \in \mathcal{O}} d\Delta_{\lambda,o}/dP^0(o') < r(\lambda)$. Assume that $\| D^*(\tilde{Q}^0_\epsilon, G^0) - D^*(Q^0, G^0) \|_{P^0} \to 0$ as $\epsilon \to 0$, and that*

$$\lim_{\lambda \to 0} \int D^*(Q^0, G^0)(o')d\Delta_{\lambda,o}(o') = D^*(Q^0, G^0)(o).$$

In addition, suppose that one of the following two assumptions (A1), (A2) holds: as $\epsilon \to 0$, either

$$(A1) : \sup_{o' \in B(o:\lambda(\epsilon))} | D^*(\tilde{Q}^0_\epsilon, G^0) - D^*(Q^0, G^0) | (o') \to 0,$$

or

$$(A2) : r(\lambda(\epsilon))^{0.5} \| D^*(\tilde{Q}^0_\epsilon, G^0) - D^*(Q^0, G^0) \|_{P^0} \to 0.$$

Then,

$$\lim_{\epsilon \to 0}(\Delta_{\lambda(\epsilon),o} - P^0)D^*(\tilde{Q}^0_\epsilon, G^0) = D^*(Q^0, G^0)(o).$$

5.2 Convergence rate of MLE condition.

The following theorem generalizes Theorem 3.

Theorem 9 *Let $L(Q, Q^0) = L(Q) - L(Q^0)$. We assume the following property of the loss-function $L(Q)$ (see (van der Laan et al., 2006; van der Vaart et al., 2006)) wander:*

$$\sup_{\epsilon \in (-\delta^*, \delta^*)} \frac{P^0\{L(\tilde{Q}^0_\epsilon, Q^0)\}^2}{P^0 L(\tilde{Q}^0_\epsilon, Q^0)} < M \text{ for some } M < \infty.$$

Let $r(\lambda)$ be a rate in λ so that

$$\| \frac{d\Delta_{\lambda,o}}{dP^0} \|_{P^0} < r(\lambda).$$

Lemma 1 provides conditions under which $r(\lambda) = O(\lambda^{-d})$.

Then,

$$P^0 L(\tilde{Q}^0_\epsilon, Q^0) = O(\epsilon^2 r^2(\lambda)).$$

Proof: We have

$$\begin{aligned}
0 &\leq P^0 L(\tilde{Q}^0_\epsilon, Q^0) \\
&= (P^0 - P^0_\epsilon) L(\tilde{Q}^0_\epsilon, Q^0) + P^0_\epsilon L(\tilde{Q}^0_\epsilon, Q^0) \\
&\leq (P^0 - P^0_\epsilon) L(\tilde{Q}^0_\epsilon, Q^0) \\
&= -\epsilon(\Delta_{\lambda,o} - P^0) L(\tilde{Q}^0_\epsilon, Q^0) \\
&= -\epsilon \int \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} L(\tilde{Q}^0_\epsilon, Q^0) dP^0 \\
&\leq \epsilon \| \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \|_{P^0} \| L(\tilde{Q}^0_\epsilon, Q^0) \|_{P^0} .
\end{aligned}$$

By assumption

$$\sup_{\epsilon \in (-\delta^*, \delta^*)} \frac{P^0\{L(\tilde{Q}^0_\epsilon, Q^0)\}^2}{P^0 L(\tilde{Q}^0_\epsilon, Q^0)} < M$$

for some $M < \infty$.

Thus, we have shown that

$$P^0 L(\tilde{Q}^0_\epsilon, Q^0) \leq M\epsilon \| \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \|_{P^0} \sqrt{P^0 L(\tilde{Q}^0_\epsilon, Q^0)},$$

which proves

$$P^0 L(\tilde{Q}^0_\epsilon, Q^0) = O(\epsilon^2) \| \frac{d\Delta_{\lambda,o} - dP^0}{dP^0} \|^2_{P^0} .$$

This completes the proof. □

5.3 Corollary of Theorem 7.

The previous two subsections provide the following corollary of Theorem 7.

Theorem 10 *Let $r(\lambda)$ be a rate in λ so that*

$$\| \frac{d\Delta_{\lambda,o}}{dP^0} \|_{P^0} < r(\lambda).$$

Lemma 1 provides conditions under which $r(\lambda) = O(\lambda^{-d})$. Let

$$d_{P^0}(Q, Q^0) = P^0\{L(Q) - L(Q^0)\}.$$

We assume the following property of the loss-function $L(Q)$:

$$\sup_{\epsilon \in (-\delta^*, \delta^*)} \frac{P^0\{L(\tilde{Q}^0_\epsilon, Q^0)\}^2}{P^0 L(\tilde{Q}^0_\epsilon, Q^0)} < M \text{ for some } M < \infty.$$

Assume

Solving efficient influence function equation: *Given P^0, for any $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$, we define a $\tilde{Q}^0_\epsilon \in \mathcal{F}(P^0)$ that satisfies $P^0_\epsilon D^*(\tilde{Q}^0_\epsilon, G^0) = 0$;*

Convergence rate of MLE as $\epsilon \to 0$: *If $R_2((\tilde{Q}^0_\epsilon, G^0), Q^0, G^0)) = 0$, then skip this condition. Otherwise, assume that $R_2((\tilde{Q}^0_\epsilon, G^0), (Q^0, G^0)) < C d_{P^0}(\tilde{Q}^0_\epsilon, Q^0)$ for some $C < \infty$, and that $\lambda(\epsilon)$ is chosen so that $\epsilon^2 r^2(\lambda(\epsilon)) = o(\epsilon)$;*

Continuity of efficient influence function at P and o: *Assume that $\| D^*(\tilde{Q}^0_\epsilon, G^0) - D^*(Q^0, G^0) \|_{P^0} < C\sqrt{d_{P^0}(\tilde{Q}^0_\epsilon, Q^0)}$ for some universal $C < \infty$, and that*

$$\lim_{\lambda \to 0} \int D^*(Q^0, G^0)(o) d\Delta_{\lambda,o}(o) = D^*(Q^0, G^0)(o).$$

Then,

$$d_{P^0}(\tilde{Q}^0_\epsilon, Q^0) = O(\epsilon^2 r^2(\lambda(\epsilon))).$$

In addition,

$$D^*(Q^0, G^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon}.$$

If one fixes $\lambda > 0$ as $\epsilon \to 0$, then we have

$$\lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0)}{\epsilon} = \Delta_{\lambda,o} D^*(Q^0, G^0).$$

5.4 Theorem relying on second-order term being zero.

In many examples, so called problems in which the efficient influence function is double robust (see e.g., (Robins et al., 2000; Rotnitzky et al., 2012; van der Laan and Robins, 2003; Rose and van der Laan, 2011), we have $R_2((\tilde{Q}^0, G^0), (Q^0, G^0)) = 0$ for any $(\tilde{Q}^0, Q^0, G^0)$. In that case, the identity (12) becomes:

$$\Psi_1(\tilde{Q}^0_\epsilon) - \Psi_1(Q^0) = \epsilon(\Delta_{\lambda,o} - P^0)D^*(\tilde{Q}^0_\epsilon, G^0). \tag{14}$$

So now we only have to assume (13) to obtain the desired result. We will state this remarkable powerful theorem for this important special case.

Theorem 11 *Let $r(\lambda)$ be a rate in λ so that*

$$\| \frac{d\Delta_{\lambda,o}}{dP^0} \|_{P^0} < r(\lambda).$$

Lemma 1 provides conditions under which $r(\lambda) = O(\lambda^{-d})$. Let

$$d_{P^0}(Q, Q^0) = P^0\{L(Q) - L(Q^0)\}.$$

We assume the following property of the loss-function $L(Q)$:

$$\sup_{\epsilon \in (-\delta^*, \delta^*)} \frac{P^0\{L(\tilde{Q}^0_\epsilon, Q^0)\}^2}{P^0 L(\tilde{Q}^0_\epsilon, Q^0)} < M \tag{15}$$

for some $M < \infty$.

Assume

Solving efficient influence function equation: *Given P^0, for any $\epsilon \in (-\delta^*, \delta^*)$ for some $\delta^* > 0$, we define a $\tilde{Q}^0_\epsilon \in \mathcal{F}(P^0)$ that satisfies $P^0_\epsilon D^*(\tilde{Q}^0_\epsilon, G^0) = 0$.*

Convergence of MLE as $\epsilon \to 0$: *Assume that $\lambda(\epsilon)$ is chosen so that $\epsilon^2 r^2(\lambda(\epsilon)) = o(1)$;*

Continuity of efficient influence function at P and O: *Assume that $d_{P^0}(\tilde{Q}^0_\epsilon, Q^0) \to 0$ as $\epsilon \to 0$ implies $\| D^*(\tilde{Q}^0_\epsilon, G^0) - D^*(Q^0, G^0) \|_{P^0} \to 0$ as $\epsilon \to 0$. Assume also that*

$$\lim_{\lambda \to 0} \int D^*(Q^0, G^0)(o')d\Delta_{\lambda,o}(o') = D^*(Q^0, G^0)(o).$$

Then,

$$d_{P^0}(\tilde{Q}_\epsilon^0, Q^0) = o(1).$$

In addition,

$$D^*(Q^0, G^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}_\epsilon^0) - \Psi_1(Q^0)}{\epsilon}.$$

If one fixes $\lambda > 0$ as $\epsilon \to 0$, then we have

$$\lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}_\epsilon^0) - \Psi_1(Q^0)}{\epsilon} = \Delta_{\lambda,o} D^*(Q^0, G^0)(o).$$

6 Example I: Parametric model

6.1 Parametric model: Assuming invertibility of information matrix

Let $O \sim P_{\theta_0} \in \mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for some Euclidean subset $\Theta \in \mathbb{R}^L$ for some L. This model could still be a nonparametric model if O is discrete valued. Suppose that our statistical target parameter is $\Psi(P_\theta) = \Psi_1(\theta) \in \mathbb{R}$ for some specified function Ψ_1 that is continuously differentiable, where it is assumed that θ is strongly identifiable from P_θ in the sense that $d_{KL}(p_{\theta_m}, p_\theta) \to 0$ as $m \to \infty$ implies $\| \theta_m - \theta \| \to 0$.

Let S_θ be the score vector, $I(\theta) = -P_\theta \frac{d}{d\theta} S_\theta$ the information operator, assumed to be invertible, and $f(\theta) = \frac{d}{d\theta}\Psi_1(\theta)$ is the gradient of Ψ_1. Then the efficient influence function at P_θ is given by $D^*(\theta)(o) = f(\theta)^\top I(\theta)^{-1} S_\theta(o)$. Let's assume that either all probability distributions in $\mathcal{M}$ are absolutely continuous w.r.t. Lebesgue measure (O is continuous) or counting measure (O is discrete). Let's denote this dominating measure with μ and the density of P_θ is denoted with $p_\theta = dP_\theta/d\mu$.

Let $P^0 = P_{\theta^0} \in \mathcal{M}$ be given, and let $P_\epsilon^0 = (1-\epsilon)P^0 + \epsilon\Delta_o$, where $\Delta_o(A) = I(o \in A)$ is the probability distribution that puts mass 1 on o. Even if O is continuous we will use a $\lambda = 0$, in which case dP_ϵ^0/dP^0 does not exist. The MLE applied to P_ϵ^0 is defined as

$$\tilde{P}_\epsilon^0 = \arg\max_{P \in \mathcal{M}} P_\epsilon^0 \log p.$$

Equivalently, we could define this MLE in terms of θ:

$$\tilde{\theta}_\epsilon^0 = \arg\max_{\theta \in \Theta} P_\epsilon^0 \log p_\theta.$$

This corresponds with determining a maximum of a multivariate real valued function. The only practical complication occurs if O is continuous, in which

case evaluation of this function involves an integral $\int \log p_\theta(o) dP^0_\epsilon(o)$, which might then be either approximated through histogram approximations of p_θ and p^0_ϵ, or by a sample mean over a large Monte Carlo sample from P^0_ϵ. Note that a proportion ϵ of the observations in the Monte-Carlo sample are equal to o. The latter Monte-Carlo method seems to be the most appealing, since it just requires computing a standard MLE (as one would compute on a data set). The numerical approximation of the efficient influence function at o is now given by

$$\frac{\Psi_1(\tilde{\theta}^0_\epsilon) - \Psi_1(\theta^0)}{\epsilon}.$$

Let's now verify the validity of the method by verifying the conditions of Theorem 1. Firstly, the MLE solves the L-dimensional equation $P^0_\epsilon S(\tilde{\theta}^0_\epsilon)$, and since the efficient influence function is a linear combination of the L scores, this immediately implies $P^0_\epsilon D^*(\tilde{\theta}^0_\epsilon) = 0$. Secondly, if $p^0 > \delta > 0$ on $\mathcal{O}$, then a standard consistency proof based on the log-likelihood (simplified version of proof of Theorem 9), establishes that $d_{KL}(p_{\tilde{\theta}^0_\epsilon}, p^0)$ converges to zero at rate $O(\epsilon)$ (since dP^0_ϵ/dP^0 does not exist when O is continuous, we do not immediately obtain the desired ϵ^2-rate). By the strong identifiability of θ from p_θ this now yields $\| \tilde{\theta}^0_\epsilon - \theta^0 \| \to 0$ as $\epsilon \to 0$. We will now proceed with applying Theorem 5 (i.e., a standard M-estimator analysis in this finite dimensional case) based on the score equations $P^0_\epsilon S(\tilde{\theta}^0_\epsilon) = 0$ and $P^0 S(\theta^0) = 0$. It follows that if $\theta \to I(\theta)$, $\theta \to S_\theta \in L^2(P_{\theta^0})$ is continuous at θ^0, and $I(\theta^0)$ is invertible (thereby also $\theta \to I(\theta)^{-1}$ is continuous at θ^0), then

$$(\tilde{\theta}^0_\epsilon - \theta^0)/\epsilon = I(\theta^0)^{-1} S_{\theta^0}(o) + o(1).$$

Finally, since Ψ_1 is differentiable at θ^0 the latter implies

$$\frac{\Psi_1(\tilde{\theta}^0_\epsilon) - \Psi_1(\theta^0)}{\epsilon} = f(\theta^0)^\top I^{-1}_{\theta^0} S_{\theta^0}(o) + o(1),$$

so that its limit for $\epsilon \to 0$ equals $D^*(P_{\theta^0})(o)$.

6.2 Parametric model, only assuming identifiability of target parameter.

As above, let $O \sim P_{\theta_0} \in \mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for some Euclidean subset $\Theta \in \mathbb{R}^L$ for some L, and that the statistical target parameter is given by $\Psi(P_\theta) = \Psi_1(\theta) \in \mathbb{R}$ for some specified function Ψ_1 that is continuously differentiable. It is assumed that $\Psi : \mathcal{M} \to \mathbb{R}$ is pathwise differentiable at P with canonical gradient $D^*(P)$, for all $P \in \mathcal{M}$.

Let S_θ be the score vector. The tangent space is the finite dimensional linear space spanned by the components of this score vector, possibly smaller than L if the model is over-parameterized. Since $D^*(P)$ is an element of the tangent space, it is a linear combination of scores. Let $P^0 = P_{\theta^0} \in \mathcal{M}$ be given, and let $P^0_\epsilon = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o}$, where $\delta_{\lambda,o}(x) = \frac{1}{\lambda^d}K\left(\frac{x-0}{\lambda}\right)$ is a multivariate density kernel centered at o with bandwidth λ.

As above let's assume that either all probability distributions in $\mathcal{M}$ are absolutely continuous w.r.t. Lebesgue measure (O is continuous) or counting measure (O is discrete). Let's denote this dominating measure with μ and the density of P_θ is denoted with $p_\theta = dP_\theta/d\mu$. Contrary to the case in which we assume identifiability of θ from P_θ above, if O is continuous we will use a $\lambda > 0$, so that we can apply our rate of convergence $d_{KL}(\tilde{P}^0_\epsilon, P^0) = O(\epsilon^2 r^2(\lambda))$.

The MLE applied to P^0_ϵ is defined as

$$\tilde{P}^0_\epsilon = \arg\max_{P\in\mathcal{M}} P^0_\epsilon \log p.$$

Since we are not assuming identifiability of θ, this MLE can be compatible with a set of values in the parameter space Θ.

To implement the MLE we apply a standard MLE to a large Monte-Carlo sample. The numerical approximation of the efficient influence function at o is now given by

$$\frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon}.$$

Let's now verify the validity of the method by verifying Theorem 1. Firstly, the MLE solves the L-dimensional equation $P^0_\epsilon S(\tilde{\theta}^0_\epsilon)$, and since the efficient influence function is a linear combination of the L scores, this immediately implies $P^0_\epsilon D^*(\tilde{P}^0_\epsilon) = 0$. If $p^0 > \delta > 0$, then Theorem 3 yields that $d_{KL}(\tilde{P}^0_\epsilon, P^0) = O(\epsilon^2 r^2(\lambda))$. Thus, we select $\lambda = \lambda(\epsilon)$ so that $\epsilon\lambda^{-d} = o(1)$. To verify the continuity condition of the efficient influence function we can apply Theorem 2. So the above consistency need to be used to establish $\| D^*(\tilde{P}^0_\epsilon) - D^*(P^0) \|_{P^0} \to 0$ as $\epsilon \to 0$, and that

$$\lim_{\lambda\to 0} \int D^*(P^0)(o')d\Delta_{\lambda,o}(o') = D^*(P^0)(o).$$

In addition, we might assume condition (A1) of Theorem 2: $\sup_{o'\in B(o:\lambda(\epsilon))} | D^*(\tilde{P}^0_\epsilon) - D^*(P^0) | (o') \to 0 = o(1)$. These are very weak regularity conditions. This verifies all conditions of Theorem 1 and thus establishes the validity of the numerical method for calculating $D^*(P^0)(o)$, under weak regularity conditions.

7 Example II: Estimation of bivariate survival function based on bivariate right-censored data.

In many censored data models in which the full-data model is nonparametric (and certainly when it is not), the efficient influence function does not exist in closed form. For example, interval censoring with more than two monitoring times, double censored data in which the survival time is subject to both left and right-censoring, bivariate right-censored data, and so on (van der Laan, 1996b,a; Chang and Yang, 1987; Chang, 1990; Groeneboom and Wellner, 1992). Essentially, whenever the censoring is defined by multiple censoring variables, typically the efficient influence function does not exist in closed form. Here we select one of such type of censored data structures to demonstrate the applicability of our numerical method.

Let (T_1, T_2) be a bivariate survival time and denote its bivariate cumulative distribution function with Q_0. Let (C_1, C_2) be a bivariate censoring time, and assume that (C_1, C_2) is independent of (T_1, T_2). Let G_0 be the cumulative distribution function of (C_1, C_2). Let the observed data on a unit be given by $O = (\tilde{T}_1 = \min(T_1, C_1), \Delta_1 = I(T_1 \leq C_1), \tilde{T}_2 = \min(T_2, C_2), \Delta_2 = I(T_2 \leq C_2))$. The probability distribution P_0 of O is determined by (Q_0, G_0), and can thus be denoted with P_{Q_0,G_0}, where the statistical model is defined as $\mathcal{M} = \{P_{Q,G} : Q \in \mathcal{F}, G \in \mathcal{G}\}$, where $\mathcal{F}$ and $\mathcal{G}$ consists of all bivariate cumulative distribution functions. This is not a fully nonparametric statistical model since the assumption of independent censoring is stronger than coarsening at random.

Let's assume that (T_1, T_2) and (C_1, C_2) are continuous, so that any Q and G are absolutely continuous w.r.t. Lebesgue measure with densities q and g, respectively. Each subdistribution of O has now a Lebesque density:

$$\begin{aligned}
p_{Q,G}(\tilde{t}_1, \tilde{t}_2, 1, 1) &= \bar{G}(\tilde{t}_1, \tilde{t}_2) q(\tilde{t}_1, \tilde{t}_2) \\
p_{Q,G}(\tilde{t}_1, \tilde{t}_2, 1, 0) &= \int_{\tilde{t}_1}^{\infty} g(t_1, \tilde{t}_2) dt_1 \int_{\tilde{t}_2}^{\infty} q(\tilde{t}_1, t_2) dt_2 \\
p_{Q,G}(\tilde{t}_1, \tilde{t}_2, 0, 1) &= \int_{\tilde{t}_2}^{\infty} g(\tilde{t}_1, t_2) dt_2 \int_{\tilde{t}_1}^{\infty} q(t_1, \tilde{t}_2) dt_1 \\
p_{Q,G}(\tilde{t}_1, \tilde{t}_2, 0, 0) &= g(\tilde{t}_1, \tilde{t}_2) \bar{Q}(\tilde{t}_1, \tilde{t}_2),
\end{aligned}$$

where $\bar{G}(t_1, t_2) = P(C_1 > t_1, C_2 > t_2)$ and $\bar{Q}(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Let p_Q and h_G denote the factors of the density $p_{Q,G}$: $p_{Q,G} = p_Q h_G$. The statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by $\Psi(P) = \bar{Q}(t_{10}, t_{20}) =$

$P(T_1 > t_{10}, T_2 > t_{20})$ for some given point (t_{10}, t_{20}) in the plane.

Let G^0, Q^0 be given, and $P^0 = P_{Q^0,G^0}$. Let $\psi^0 = \Psi(P^0)$. We will assume that Q^0 has compact support $[0, \tau_1] \times [0, \tau_2] \subset \mathbb{R}^2_{\geq 0}$, $\bar{G}^0(\tau_1, \tau_2) > 0$, and $q^0 > \delta > 0$ on $[0, \tau_1] \times [0, \tau_2]$ for some $\delta > 0$. Under this condition, Ψ is pathwise differentiable at $P^0 = P_{Q^0,G^0}$ and the efficient influence function $D^*(Q^0, G^0)$ can be represented in the following manner. Define the nonparametric score operator $A_{Q^0} : L_0^2(Q^0) \to L_0^2(P^0)$ as $A_{Q^0}(S)(O) = E_{Q^0}(S(T_1, T_2) \mid O)$, and its adjoint $A^\top_{G^0} : L_0^2(P^0) \to L_0^2(Q^0)$ as $A^\top_{G^0}(V)(T) = E_{G^0}(V(O) \mid T_1, T_2)$. Then, one can define the so called nonparametric information operator $I_{P^0} : L_0^2(Q^0) \to L_0^2(Q^0)$ as $I_{P^0} = A_{Q^0} A^\top_{G^0}$. Under the above conditions on Q^0, G^0, we have that $I_{P^0} : L_0^2(Q^0) \to L_0^2(Q^0)$ is invertible and has a bounded inverse $I^{-1}_{P^0}$, and

$$D^*(P^0) = A_{Q^0} I^{-1}_{P^0}(\kappa_{\psi^0}),$$

where $\kappa_{Q^0}(T_1, T_2) = I(T_1 > t_{10}, T_2 > t_{20}) - \Psi_1(Q^0)$, where Ψ_1 is defined by $\Psi(P) = \Psi_1(Q(P))$ (van der Laan, 1996a,b).

In general, there is no closed form solution for $D^*(P^0)$, but the inverse of the information operator can be represented with the Neumann series $I^{-1}_{P^0} = \sum_{k=0}^{\infty}(I - I_{P^0})^k$, where I denotes the identity operator. Let $D^*(P^0)$ be a pointwise defined version of the above defined $D^*(P^0)$ defined in $L_0^2(P^0)$ that is continuous at the given observation o at which we want to compute $D^*(P^0)(o)$.

Instead of using this highly involved algorithm in terms of the inverse of the nonparametric information operator (as carried out in (Quale et al., 2006)) for computing $D^*(P^0)(o)$, we now want to use a numerical approximation method proposed in this article. Let $o = (\tilde{t}_1, \tilde{t}_2, \delta_1, \delta_2)$. Let $P^0_\epsilon = (1 - \epsilon)P^0 + \epsilon\Delta_{\lambda,o}$ be the ϵ-perturbation of P^0 based on a bivariate kernel K with bandwidth λ:

$$\Delta_{\lambda,o}(x_1, x_2, b_1, b_2) = I(b_1 = \delta_1, b_2 = \delta_2)\frac{1}{\lambda^2}K\left((\tilde{t}_1 - x_1)/\lambda, \tilde{t}_2 - x_2)/\lambda)\right).$$

We define the corresponding perturbation $\tilde{Q}^0_\epsilon$ of Q^0 by the MLE mapping:

$$\tilde{Q}^0_\epsilon = \arg\max_{Q \in \mathcal{F}(Q^0)} P^0_\epsilon \log p_Q,$$

where $\mathcal{F}(Q^0)$ is the set of all bivariate distributions absolute continuous w.r.t. Q^0. This MLE can be computed with the EM-algorithm: start with a $F^0 = Q^0$, and for $m = 1, \ldots$, compute

$$F^m(t_1, t_2) = E_{P^0_\epsilon} E_{F^{m-1}}(I(T_1 > t_1, T_2 > t_2) \mid O), \text{ for all } (t_1, t_2) \in [0, \tau].$$

The limit of this algorithm as $m \to \infty$ is $\tilde{Q}^0_\epsilon$.

In practice, we may approximate and implement this MLE $\tilde{Q}_\epsilon^0$ as follows resulting in an approximation $\tilde{Q}_{\epsilon,d}^0$ with a histogram density. Firstly, one defines a discrete grid approximation τ^d of the support $[0,\tau]$ of Q^0, and let q_d^0 be the corresponding histogram density approximation of the density q^0 of Q^0. For each $(t_1,t_2) \in \tau_d$, let $R(t_1,t_2)$ be a rectangle, so that $[0,\tau] = \cup_{(t_1,t_2)\in\tau^d} R(t_1,t_2)$. This then also defines a histogram based approximation Q_d^0 of the cumulative bivariate distribution function Q^0. Similarly, let G_d^0 be an histogram based approximation of G^0 based on the same grid. For each little rectangle $R(t_1,t_2)$ in the partitioning defined by τ^d, identified by $(t_1,t_2) \in \tau^d$, let $p^0(t_1,t_2)$ be the probability under q_d^0 to fall in this rectangle (i.e., this is just the histogram density q_d^0 integrated over this rectangle). Now, we start with p_d^0, set $m = 0$, and run the EM-algorithm for a multinomial probability distribution: for $m = 1, \ldots$, define

$$p_d^m(t_1,t_2) = E_{P_\epsilon^0} E_{p_d^{m-1}}(I((T_1,T_2) \in R(t_1,t_2)) \mid O), \text{ for all } (t_1,t_2) \in \tau^d.$$

Here one can approximate the expectation w.r.t. P_ϵ^0 with a large Monte Carlo sample from P_ϵ^0. The limit p_d^∞ as $m \to \infty$ of this algorithm is a discrete distribution representing the mass the MLE $\tilde{Q}_\epsilon^0$ gives to each rectangle, so that it defines a histogram density estimator q_d^∞ whose bivariate cumulative distribution function is our desired approximation $\tilde{Q}_{\epsilon,d}^0$ of $\tilde{Q}_\epsilon^0$.

An alternative strategy for approximating $\tilde{Q}_\epsilon^0$ is the one mentioned earlier: replace each p_Q with $Q \in \mathcal{F}(Q^0)$ by a histogram approximation p_Q^d, replace p_ϵ^0 by a histogram approximation $p_{\epsilon,d}^0$, and define the MLE as

$$\tilde{p}_{\epsilon,d}^0 = \arg \max_{p_Q^d, Q\in\mathcal{F}(Q^0)} P_{\epsilon,d}^0 \log p_Q^d.$$

Finally, one could also take a very large Monte-Carlo sample from P_ϵ^0 and compute a regularized MLE of $\tilde{Q}_\epsilon^0$ such as the regularized MLE in van der Laan (1996a,b).

The numerical approximation of $D^*(Q_d^0, G_d^0)(o)$ is now given by:

$$\frac{\Psi(\tilde{Q}_{\epsilon,d}^0) - \Psi(Q_d^0)}{\epsilon},$$

where ϵ and λ will need to be chosen small enough, as discussed in detail below. Since the perturbation $\tilde{Q}_{\epsilon,d}^0$ does not depend on the target parameter, one can use this same $\tilde{Q}_{\epsilon,d}^0$ to compute the efficient influence function for the survival probability at (t_1,t_2) for any $(t_1,t_2) \in \tau^d$. So even though running this EM algorithm might take some serious computer time, in the end we obtain the efficient influence function for a large class of target parameters.

In order to establish the validity of the numerical method for approximating $D^*(P_d^0)$ at a fixed discretization d, and recommended values for λ, ϵ, we apply Theorem 11. To start with, we note that Lemma 1 we have $r(\lambda) = \lambda^{-2}$. Since this is a CAR-censored data model, we have $R_2((\tilde{Q}^0, G^0), (Q^0, G^0)) = 0$ for any $\tilde{Q}^0$, given (Q^0, G^0). We apply this theorem with the log-likelihood loss: $L(Q) = -\log p_Q$. Property (15) is known to hold if $p_{\tilde{Q}^0_{\epsilon,d}}$ and $p_{Q^0_d}$ are bounded away from zero on the support $[0, \tau]$. By assumption on q_d^0 and $\bar{G}_d^0(\tau) > 0$, $p_{Q_d^0}$ is bounded away from zero. In addition, by a standard argument based on the EM algorithm as a redistribution algorithm it follows that $\tilde{q}^0_{\epsilon,d} \geq p^0_{\epsilon,d}(\cdot, (1,1))$, and the latter is bounded away from zero since $\bar{G}_d^0(\tau) > 0$ and q_d^0 bounded away from zero on $[0, \tau]$. So this proves (15).

Regarding the first main condition of this theorem we need to prove that $P^0_{\epsilon,d} D^*(\tilde{Q}^0_{\epsilon,d}, G^0_d) = 0$. Since $D^*(\tilde{Q}^0_{\epsilon,d}, G^0_d) = A_{\tilde{Q}^0_{\epsilon,d}} I^{-1}_{\tilde{Q}^0_{\epsilon,d}, G^0_d}(\kappa_{\tilde{Q}^0_{\epsilon,d}})$ it follows that it is an actual score at $P_{\tilde{Q}^0_{\epsilon,d}, G^0_d}$. Here we also use that, due to the discretization, the bounded invertibility of the information operator in $L_0^2(Q_d^0)$ norm also implies bounded invertibility w.r.t. supremum/max norm, since for finite dimensional spaces all norms are equivalent. As a consequence, it is not only a score in $L_0^2(P^0)$, it also has a bounded max norm so that we can construct a submodel with this score. This establishes the first main condition.

The theorem now teaches us that we need to select $\lambda(\epsilon)$ so that $\epsilon/\lambda \to 0$. Under this condition, we have $d_{P^0}(\tilde{Q}^0_{\epsilon,d}, Q^0_d) \to 0$ as $\epsilon \to 0$, which implies $\tilde{q}^0_{\epsilon,d} - q^0_d$ converges to zero in $L^2(Q^0_d)$-norm. We now need to verify the continuity of the efficient influence function conditions. For notational convenience, in the following formulas, let $\tilde{Q}^0_d$ denote $\tilde{Q}^0_{\epsilon,d}$. We have

$$\begin{aligned} D^*(\tilde{Q}^0_d, G^0_d) - D^*(Q^0_d, G^0_d) &= A_{\tilde{Q}^0_d} I^{-1}_{\tilde{Q}^0_d, G^0_d}(\kappa_{\tilde{Q}^0_d}) - A_{Q^0_d} I^{-1}_{Q^0_d, G^0_d}(\kappa_{Q^0_d}) \\ &= (A_{\tilde{Q}^0_d} - A_{Q^0_d}) I^{-1}_{Q^0_d, G^0_d}(\kappa + Q^0_d)) \\ &+ A_{\tilde{Q}^0_d}\{I^{-1}_{\tilde{Q}^0_d, G^0_d} - I^{-1}_{Q^0_d, G^0_d}\}(\kappa_{Q^0_d}) + A_{\tilde{Q}^0_d} I^{-1}_{\tilde{Q}^0_d, G^0_d}(\kappa_{\tilde{Q}^0_d}) - \kappa_{Q^0_d}). \end{aligned}$$

Regarding the second term on the right-hand side we note that

$$I^{-1}_{\tilde{Q}^0_d, G^0_d} - I^{-1}_{Q^0_d, G^0_d} = -I^{-1}_{\tilde{Q}^0_d, G^0_d}(I_{\tilde{Q}^0_d, G^0_d} - I_{Q^0_d, G^0_d}) I^{-1}_{Q^0_d, G^0_d}.$$

At a fixed discretization, for any Q, I^{-1}_{Q, G^0_d} has a bounded inverse w.r.t. supremum norm (it is a finite dimensional problem in which case all norms are equivalent). In addition, we have that $\tilde{q}^0_{\epsilon,d} - \tilde{q}^0_d$ converges in supremum norm to zero as $\epsilon \to 0$, and all denominators in $A_{\tilde{Q}_d}$ and $A_{Q^0_d}$ and thereby $I_{\tilde{Q}^0_d}$ are bounded away from zero. In this manner, it is straightforward to show that

$\| D^*(\tilde{Q}^0_{\epsilon,d}, G^0_d) - D^*(\tilde{Q}^0_{\epsilon,d}, G^0_d) \|_{\tilde{P}^0_\epsilon}$ converges to zero as $\epsilon \to 0$, which confirms the first condition in this continuity condition of the theorem.

Finally, we need to show that $D^*(P^0_d)(o) = \lim_{\lambda \to 0} \Delta_{\lambda,o} D^*(P^0_d)$. This relies on $O \to D^*(P^0_d)(O)$ being continuous at o, which holds trivially at this discrete P^0_d. This verifies all the conditions of Theorem 11, and thus proves the validity of the numerical approximation:

$$D^*(Q^0_d, G^0_d)(o) = \lim_{\epsilon \to 0} \frac{\Psi_1(\tilde{Q}^0_{\epsilon,d}) - \Psi_1(Q^0_d)}{\epsilon}.$$

We conclude with some remarks regarding the role of discretization in this proof. Suppose that one would be able to establish that the nonparametric information operator $I_{Q,G^0} : L^2_0(Q) \to L^2_0(Q)$ has a bounded inverse w.r.t. the supremum norm, thus not relying on the discretization. Then, by the same proof as above, we would establish our desired consistency condition for the efficient influence function, *uniformly in any grid approximation*. Unfortunately, bounded invertibility of the information operator w.r.t supremum norm at a continuous Q^0 is an unknown result in this complex bivariate right-censored data model. Nonetheless, one expects it to hold: the structure of the information operator that causes the analytic challenges is due to the singly-censored observations, while we have shown that the nonparametric information operator for general censored data structures (with positive probability on uncensored observations) has a nice bounded inverse w.r.t supremum norm (and even variation norm) once the regions induced by the censored observations have full dimension (instead of lines in a plane that have probability zero). There is no sensible reason to believe that extra censoring makes the information operator more invertible, on the contrary, so that it is a pure technical issue. In fact, a practical study of the Neuman series inverse in Quale et al. (2006) showed a stable inverse in terms of the grid selected, practically confirming our conjecture that the information operator is indeed invertible w.r.t. supremum norm .

If in truth the information operator has not a bounded inverse w.r.t. supremum norm, then it might be possible that the norm of the inverse of the information operator increases to larger and larger values when one lets the mesh of the grid converge to zero. In particular, that might then mean that in practice, when using a reasonably fine partitioning, it takes a very small ϵ and λ to obtain the desired approximation. By the lack of this supremum norm invertibility result, it is also unclear how to prove that $D^*(P^0_d)$ converges to $D^*(P^0)$ as the grid gets finer and finer, even though, again, one certainly expects this to hold. At a discretized P^0_d, one could even have selected $\lambda = 0$, but in that case one might start relying on the grid approximation playing the role

of λ, in which case it might become important that the grid converges slowly relative to ϵ as $\epsilon \to 0$, while if we use our λ, given the conjectured supremum norm invertibility, the numerical approximation works for the above choice of (ϵ, λ) satisfying $\epsilon/\lambda^2 \to 0$, whatever grid we select. We prefer that λ plays the role of regularizing the MLE, so that the convergence of our numerical algorithm is only driven by the single parameter λ, while the choice of discretization is purely a computational consideration (chosen so that the error due to discretization is negligible for practical purposes).

8 Example III: Counterfactual mean for two time point intervention

We refer to (van der Laan and Gruber, 2012; Petersen et al., 2014; Bang and Robins, 2005). Let $O = (L(0), A(0), L(1), A(1), Y = L(2)) \sim P_0$ and assume a nonparametric statistical model $\mathcal{M}$. Suppose that $Y \in \{0, 1\}$ is binary. Let $d = (d_0, d_1)$ be a dynamic treatment regimen for assigning treatment $A(0)$ and $A(1)$. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be defined by:

$$\Psi(P) = E_P E_P(E_P(Y \mid \bar{L}(1), \bar{A}(1) = d(\bar{L}(1))) \mid L(0), A(0) = d_0(L(0))).$$

Under a causal model, the sequential randomization assumption, and a positivity assumption, $\Psi(P_0)$ equals the mean counterfactual outcome $E_0(Y_d)$ under dynamic treatment rule d. Let $P^0 \in \mathcal{M}$ be given, and let $P^0_\epsilon = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o}$ be the perturbation of P^0 in the direction of a smoothed pointmass at o, where

$$\delta_{\lambda,o}(\bar{l}'(2), \bar{a}'(1)) = I(\bar{a}'(1) = \bar{a}(1), l(0) = l'(0), y' = y)K_\lambda(l'(1) - l(1)).$$

Let $L(1) \in \mathbb{R}^d$. For an $x \in \mathbb{R}^d$ and d-variate product density kernel K, we define $K_\lambda(x) = \frac{1}{\lambda^d}K(x/\lambda)$, and $K(x) = \prod_{j=1}^d K_j(x_j)$. In this special target parameter $\Psi(P)$ involving an expectation over $L(0)$ w.r..t its true marginal distribution (instead of some other distribution), it is not necessary to also smooth in $L(0)$ (Luedtke et al., 2015).

We have that any P is determined by $Q(P)$ and $G(P)$, where $Q(P)$ are the conditional distributions of the L-nodes and $G(P)$ the conditional distributions of the A-nodes. The parameter $\Psi(P)$ only depends on P through $Q(P)$. Let $P^0 = (Q^0, G^0)$ and $P^0_\epsilon = (Q^0_\epsilon, G^0_\epsilon)$. In this nonparametric model we have that the MLE mapping $\tilde{Q}^0_\epsilon$ applied to P^0_ϵ equals Q^0_ϵ. Thus,

$$\Psi(\tilde{P}^0_\epsilon) = E_{P^0_\epsilon} E_{P^0_\epsilon}(E_{P^0_\epsilon}(Y \mid \bar{L}(1), \bar{A}(1) = d(\bar{L}(1))) \mid L(0), A(0) = d(L(0))),$$

where each conditional expectation only depends on P_ϵ^0 through Q_ϵ^0. We approximate $D^*(P^0)(o)$ with $\{\Psi(P_\epsilon^0) - \Psi(P^0)\}/\epsilon$ for an $\epsilon \approx 0$.

Let's first establish the validity of this method. We will apply Theorem 11, using that $R_2(\tilde{Q}_\epsilon^0, G^0, Q^0, G^0) = 0$. We have $P_\epsilon^0 D^*(\tilde{Q}_\epsilon^0, G^0) = 0$ since $\tilde{Q}_\epsilon^0$ is the MLE. We select $\lambda(\epsilon)$ so that $\epsilon\lambda^{-d} \to 0$. The continuity of the efficient influence function are conditions that hold if $g^0(d(\bar{L}(1)), \bar{L}(1)) > \delta > 0$ P^0-a.e., and g^0 and the conditional means under Q^0 at $\bar{A}(1) = d(\bar{L}(1))$ are continuous in $\bar{L}(1)$. This verifies the conditions of the theorem and thus establishes the validity of the method.

Let's now discuss implementation of $\Psi(\tilde{P}_\epsilon^0)$. The outer expectation w.r.t $L(0)$ can be carried out with Monte-Carlo simulation. The most inner conditional expectation of Y is just a sum over two values. To approximate the conditional expectation integrating over $L(1)$, we recommend approximating this conditional expectation integral with a Riemann sum w.r.t. a partitioning of the support of $L(1)$ as discussed earlier.

To make this implementation practical for high dimensional d, the following lemma provides a dimension reduction for $L(1)$. This would represent Step 0 in our description of the general method.

Lemma 2 *Consider the above setting with $O = (L(0), A(0), L(1), A(1), Y)$, statistical model $\mathcal{M}$ that only makes assumptions on conditional distributions $g_{A(0)}$ and $g_{A(1)}$ of $A(0)$ and $A(1)$, respectively, target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ defined by $\Psi(P) = E_P E_P(E_P(Y \mid \bar{A}(1) = d(\bar{L}(1)), \bar{L}(1)) \mid A(0) = d(L(0)), L(0))$, and its efficient influence function $D^*(P^0)$ at $P^0 \in \mathcal{M}$. Let $d(L(0)$ denote the treatment assignment for $A(0)$ under dynamic treatment d, and similarly let $d(\bar{L}(1))$ be the treatment assignment for $A(1)$.*

Define $O_r = (L_r(0) = L(0), A(0), L_r(1), A(1)), Y))$, where

$$\begin{aligned} L_r(1) &\equiv (d(\bar{L}(1)), \bar{Q}_2^0(\bar{L}(1)), \bar{g}_1^0(\bar{L}(1)) \\ \bar{g}_1^0(\bar{L}(1)) &= g_{A(1)}^0(1 \mid A(0) = d_0(L(0), \bar{L}(1)) \\ \bar{Q}_2^0(\bar{L}(1)) &= E_{P^0}(Y \mid \bar{L}(1), \bar{A}(1) = d(\bar{L}(1))). \end{aligned}$$

Note that $d(L(0))$ and $d(\bar{L}(1))$ are only functions of $L_r(0)$ and $\bar{L}_r(1)$, respectively, so that we can also write $d(L_r(0))$ and $d(\bar{L}_r(1))$.

Since O_r is a function of O, each possible probability distribution P in $\mathcal{M}$ of O implies a distribution P_r of O_r. Let $\mathcal{M}_r = \{P_r : P \in \mathcal{M}\}$ be the model for O_r induced by $\mathcal{M}$. Let $\Psi_r : \mathcal{M}_r \to \mathbb{R}$ be defined by

$$\Psi_r(P_r) = E_{P_r} E_{P_r}(E_{P_r}(Y \mid \bar{L}_r(1), \bar{A}(1) = d(\bar{L}_r(1))) \mid L_r(0), A(0) = d(L_r(0))).$$

Let $D_r^*(P_r^0)$ be efficient influence function of Ψ_r at P_r^0. We have $D_r^*(P_r^0) = D^*(P^0)$ P^0-a.e.

Proof: Let $\bar{Q}_1^0 = E_{P^0}(\bar{Q}_2^0 \mid L(0), A(0) = d(L(0)))$. The efficient influence function is given by $D^*(P^0) = \sum_{k=0}^2 D_k^*(P^0)$, where

$$\begin{aligned}
D_0^*(P^0) &= E_{P^0}(E_{P^0}(Y \mid \bar{L}(1), \bar{A}(1) = d(\bar{L}(1))) \mid L(0), A(0) = d(L(0))) - \Psi(P^0) \\
&= E_{P^0}(\bar{Q}_2^0 \mid L(0), A(0) = d(L(0))) - \Psi(P^0) \\
D_1^*(P^0) &= \frac{I(A(0) = d(L(0)))}{g_{A(0)}^0(A(0) \mid L(0))}(\bar{Q}_2^0 - \bar{Q}_1^0) \\
D_2^*(P^0) &= \frac{I(\bar{A}(1) = d(\bar{L}(1)))}{g_{A(0)}^0 g_{A(1)}^1}(Y - \bar{Q}_2^0).
\end{aligned}$$

To start with, since $L_r(1)$ includes $\bar{Q}_2^0$, we note that $\bar{Q}_{r2}^0 \equiv E_{P_r^0}(Y \mid \bar{L}_r(1), \bar{A}(1) = d(\bar{L}_r(1))) = \bar{Q}_2^0$. Here, and below, we also use that $d(\bar{L}_r(1)) = d(\bar{L}(1))$ by the fact that $L_r(1)$ includes the decision $d(\bar{L}(1))$. Because $\bar{Q}_2^0(\bar{L}(1))$ only depends on $\bar{L}(1)$ through $\bar{L}_r(1)$, and $L_r(0) = L(0)$, it also follows that

$$\bar{Q}_{r1}^0 \equiv E_{P_r^0}(\bar{Q}_{r2}^0 \mid L^r(0), A(0) = d(L_r(0))) = E_{P^0}(\bar{Q}_2^0 \mid L(0), A(0) = d(L(0)) = \bar{Q}_1^0.$$

In particular, this shows that

$$\Psi_r(P_r^0) = \Psi(P^0).$$

Since $L^r(0) = L(0)$, we also have that $g_{r,A(0)}^0 = g_{A(0)}^0$. Since $L^r(1)$ includes $g_{A(1)}^0(1 \mid \bar{L}(1), A(0) = d(L(0)))$, it follows that

$$g_{r,A(1)}^0(1 \mid \bar{L}^r(1), A(0) = d(L_r(0))) = g_{A(1)}^0(1 \mid L(1), A(0) = d(L(0))).$$

By the same formula of the above efficient influence function but applied to O_r, we also have that $D_r^*(P_r^0)$ is given by:

$$\begin{aligned}
D_{r0}^*(P_r^0) &= E_{P_r^0}(E_{P_r^0}(Y \mid \bar{L}_r(1), \bar{A}(1) = d(\bar{L}_r(1))) \mid L_r(0), A(0) = d(L_r(0))) - \Psi_r(P_r^0) \\
&= E_{P^0}(\bar{Q}_2^0 \mid L(0), A(0) = d(L(0))) - \Psi(P^0) \\
&= D_0^*(P^0) \\
D_{r1}^*(P_r^0) &= \frac{I(A(0) = d(L_r(0)))}{g_{r,A(0)}^0(A(0) \mid L_r(0))}(\bar{Q}_{r2}^0 - \bar{Q}_{r1}^0) \\
&= \frac{I(A(0) = d(L(0)))}{g_{A(0)}^0(A(0) \mid L(0))}(\bar{Q}_2^0 - \bar{Q}_1^0) \\
&= D_1^*(P^0) \\
D_{r2}^*(P_r^0) &= \frac{I(\bar{A}(1) = d(\bar{L}_r(1)))}{g_{r,A(0)}^0 g_{r,A(1)}^1}(Y - \bar{Q}_{r2}^0) \\
&= \frac{I(\bar{A}(1) = d(\bar{L}(1)))}{g_{A(0)}^0 g_{A(1)}^1}(Y - \bar{Q}_2^0) \\
&= D_2^*(P^0).
\end{aligned}$$

This proves that $D_r^*(P_r^0) = D^*(P^0)$, and completes the proof of the lemma. □

9 Numerical method for calculating second-order efficient influence function.

We refer to (Robins et al., 2008, 2009; van der Vaart, forthcoming; Carone et al., 2014). We will use the notation $P^2 f = \int f(o_1, o_2) dP(o_1) dP(o_2)$ Suppose that $\Psi : \mathcal{M} \to \mathbb{R}$ is higher-order pathwise differentiable at any $P \in \mathcal{M}$. For the sake of demonstration, lets consider the case that it is second-order pathwise differentiable at $P^0 \in \mathcal{M}$. This means that for each submodel $\{P_\delta^0 : \delta\}$ in a class of models generating the tangent space $T(P^0)$ with $dP_\delta^0 = (1 + \delta S + 0.5\delta^2 S_2) dP^0 + o(\delta^2)$, we have

$$\Psi(P_\delta^0) - \Psi(P^0) = \delta P D^{(1)}(P^0) S + 0.5\delta^2 \{P D^{(1)}(P^0) S_2 + P^2 D^{(2)}(P^0) S^2\} + o(\delta^2),$$

where $D^{(1)}(P^0)(O), D^{(2)}(P^0)(O_1, O_2)$ are the first and second-order efficient influence function at P^0, respectively. Here we used the notation $P^2 D^{(2)}(P^0) S^2 = \int D^{(2)}(P^0)(o_1, o_2) S(o_1) S(o_2) dP^0(o_1) dP^0(o_2)$. The second-order efficient influence function at P has the property that for all $(o_1, o_2) \in \mathcal{O}^2$,

$$P D^{(2)}(P)(o_1, \cdot) = P D^{(2)}(P)(\cdot, o_2) = P D^{(2)}(P) = 0.$$

In addition, we have that for all $(o_1, o_2) \in \mathcal{O}^2$, $D^{(2)}(P)(o_1, \cdot) \in T(P)$ and $D^{(2)}(P)(\cdot, o_2) \in T(P)$. The second-order pathwise differentiability typically allows for a second-order expansion of the following type: for $P, P^0 \in \mathcal{M}$,

$$\Psi(P) - \Psi(P^0) = (P - P^0) D^{(1)}(P^0) + (P - P^0)^2 D^{(2)}(P^0) + R_3(P, P^0), \quad (16)$$

where $R_3(P, P^0)$ is a third order term in the difference $P - P^0$.

The second-order efficient influence function can be used to construct a second-order one-step estimator which is asymptotically efficient under the same conditions as previously mentioned but where the second-order term condition $R_2(P_n^0, P_0) = o_P(1/\sqrt{n})$ is replaced by $R_3(P_n^0), P_0) = o_P(1/\sqrt{n})$. For example, the second-order one-step estimator, using as initial estimator $P_n^0 \in \mathcal{M}$, is defined as (Robins et al., 2008):

$$\psi_n^1 = \Psi(P_n^0) + P_n D^{(1)}(P_n^0) + P_n^2 D^{(2)}(P_n^0).$$

Alternatively, the second-order efficient influence function can be used to construct a second-order TMLE (Carone et al., 2014).

In this section we propose a numerical method that approximates $D^{(2)}(P^0)(o_1, o_2)$ at a given P^0 and pair (o_1, o_2). Given P^0 and $(o_1, o_2) \in \mathcal{O}^2$, consider the following perturbation of P^0:

$$dP^0_{\epsilon,(o_1,o_2)} = \left(1 + \epsilon \frac{d\Delta_{\lambda,o_1,o_2} - dP^0}{dP^0}\right) dP^0, \tag{17}$$

where $\Delta_{\lambda,o_1,o_2} = 0.5\Delta_{\lambda,o_1} + 0.5\Delta_{\lambda,o_2}$ is a mixture of Δ_{λ,o_1} and Δ_{λ,o_2} with weight 0.5. Alternatively, we can write $P^0_{\epsilon,(o_1,o_2)} = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o_1,o_2}$. Notice that for ϵ small enough this is indeed a density. As before, let $\tilde{P}^0_{\epsilon,(o_1,o_2)}$ be defined as the MLE projection of $P^0_{\epsilon,(o_1,o_2)}$ onto the model $\mathcal{M}$:

$$\tilde{P}^0_{\epsilon,(o_1,o_2)} = \arg \min_{P \in \mathcal{M}(P^0)} P^0_{\epsilon,(o_1,o_2)} \log \frac{dP}{dP^0},$$

assuming it exists.

We will show that, under weak regularity conditions,

$$\begin{aligned} f^{(2)}(P^0, (o_1, o_2)) &\equiv \lim_{\lambda \to 0} \frac{d^2}{d\epsilon^2} \Psi(\tilde{P}^0_{\epsilon,(o_1,o_2)})\Big|_{\epsilon=0} \\ &= 0.5D^{(2)}(P^0)(o_1, o_1) + 0.5D^{(2)}(P^0)(o_2, o_2) + D^{(2)}(P^0)(o_1, o_2). \end{aligned}$$

In addition, we have

$$\begin{aligned} f^{(2)}(P^0, o_1) &\equiv \lim_{\lambda \to 0} \frac{d^2}{d\epsilon^2} \Psi(P^0_{\epsilon,o_1})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_1, o_1) \\ f^{(2)}(P^0, o_2) &\equiv \lim_{\lambda \to 0} \frac{d^2}{d\epsilon^2} \Psi(P^0_{\epsilon,o_2})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_2, o_2) \end{aligned}$$

As a consequence,

$$D^{(2)}(P^0)(o_1, o_2) = f^{(2)}(o_1, o_2) - 0.25f^{(2)}(o_1) - 0.25f^{(2)}(o_2).$$

In the next two subsections we establish the validity of this method for calculating $D^{(2)}(P^0)(o_1, o_2)$. Firstly, we consider the easier case that the model $\mathcal{M}$ is nonparametric, and subsequently, we show that the general validity is proven by applying this result to a nonparametric model with target parameter $\tilde{\Psi}(P) = \Psi(\tilde{P})$, where $\tilde{P} = \arg\max_{P_1 \in \mathcal{M}(P^0)} P \log dP_1/dP^0$, and showing that this nonparametric extension $\tilde{\Psi}$ of Ψ has the same first and second-order efficient influence function as Ψ at a $P^0 \in \mathcal{M}$.

9.1 Validity of the method for nonparametric models

The following theorem establishes the validity of the proposed method for a nonparametric model.

Theorem 12 *Suppose $\mathcal{M}(P^0)$ is nonparametric so that $\tilde{P}^0_{\epsilon,(o_1,o_2)} = P^0_{\epsilon,(o_1,o_2)}$ for all $\epsilon \in (-\delta, \delta)$ for some $\delta > 0$.*

Assume (16) at $P^0_{\epsilon,(o_1,o_2)}$ and P^0:

$$\begin{aligned}\Psi(P^0_{\epsilon,(o_1,o_2)}) - \Psi(P^0) &= (P^0_{\epsilon,(o_1,o_2)} - P^0)D^{(1)}(P^0) + (P^0_{\epsilon,(o_1,o_2)} - P^0)^2 D^{(2)}(P^0) \\ &\quad + R_3(P^0_{\epsilon,(o_1,o_2)}, P^0).\end{aligned}$$

In addition, assume $D^{(1)}(P^0)$ is continuous at o_1 and o_2, $D^{(2)}(P^0)$ is continuous at (o_1,o_2), (o_1,o_1) and (o_2,o_2). Finally, assume that $R_3(P^0_{\epsilon,(o_1,o_2)}, P^0) = o(\epsilon^2)$.

We have

$$\begin{aligned}f^{(2)}(P^0,(o_1,o_2)) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(\tilde{P}^0_{\epsilon,(o_1,o_2)})\Big|_{\epsilon=0} \\ &= 0.5D^{(2)}(P^0)(o_1,o_1) + 0.5D^{(2)}(P^0)(o_2,o_2) + D^{(2)}(P^0)(o_1,o_2).\end{aligned}$$

In addition, we have

$$\begin{aligned}f^{(2)}(P^0,o_1) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(P^0_{\epsilon,o_1})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_1,o_1) \\ f^{(2)}(P^0,o_2) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(P^0_{\epsilon,o_2})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_2,o_2)\end{aligned}$$

As a consequence, we have

$$D^{(2)}(P^0)(o_1,o_2) = f^{(2)}(o_1,o_2) - 0.25f^{(2)}(o_1) - 0.25f^{(2)}(o_2).$$

Proof of Theorem 12: For notational convenience, in this proof let $P^0_\epsilon = P^0_{\epsilon,(o_1,o_2)}$. (16) at $P = P^0_\epsilon$ yields:

$$\Psi(P^0_\epsilon) - \Psi(P^0) = (P^0_\epsilon - P^0)D^{(1)}(P^0) + (P^0_\epsilon - P^0)^2 D^{(2)}(P^0) + R_3(P^0_\epsilon, P^0).$$

We now use that

$$\begin{aligned}(P^0_\epsilon - P^0)D^{(1)}(P^0) &= \epsilon\Delta_{\lambda,(o_1,o_2)}D^{(1)}(P^0) \\ &= 0.5\epsilon\Delta_{\lambda,o_1}D^{(1)}(P^0) + 0.5\epsilon\Delta_{\lambda,o_2}D^{(1)}(P^0) \\ (P^0_\epsilon - P^0)^2D^{(2)}(P^0) &= \epsilon^2\Delta^2_{\lambda,(o_1,o_2)}D^{(2)}(P^0) \\ &= \epsilon^2 0.25\Delta^2_{\lambda,o_1}D^{(2)}(P^0) + \epsilon^2 0.25\Delta^2_{\lambda,o_2}D^{(2)}(P^0) \\ &\quad +\epsilon^2 0.5\Delta_{\lambda,o_1}\Delta_{\lambda,o_2}D^{(2)}(P^0).\end{aligned}$$

The continuity assumptions imply that the expectations w.r.t. Δ^2_{λ,o_1}, Δ^2_{λ,o_2} and Δ_{λ,o_1,o_2} of $D^{(2)}(P^0)$ converge to the pointless evaluation at (o_1,o_1), (o_2,o_2), (o_1,o_2), respectively, as $\lambda \to 0$. Thus, we conclude that

$$\begin{aligned}\Psi(P^0_\epsilon) - \Psi(P^0) &= 0.5\epsilon\Delta_{\lambda,o_1}D^{(1)}(P^0) + 0.5\epsilon\Delta_{\lambda,o_2}D^{(1)}(P^0) \\ &\quad +\epsilon^2 0.25\Delta^2_{\lambda,o_1}D^{(2)}(P^0) + \epsilon^2 0.25\Delta^2_{\lambda,o_2}D^{(2)}(P^0) \\ &\quad +\epsilon^2 0.5\Delta_{\lambda,o_1}\Delta_{\lambda,o_2}D^{(2)}(P^0) + R_3(P^0_\epsilon, P^0).\end{aligned}$$

Thus,

$$\begin{aligned}f^{(2)}(P^0,(o_1,o_2)) &\equiv \lim_{\lambda\to 0}\frac{d^2}{d\epsilon^2}\Psi(P^0_\epsilon)\Big|_{\epsilon=0} \\ &= 0.5D^{(2)}(P^0)(o_1,o_1) + 0.5D^{(2)}(P^0)(o_2,o_2) + D^{(2)}(P^0)(o_1,o_2).\end{aligned}$$

Similarly, we can show $f^{(2)}(P^0,o) = 2D^{(2)}(P^0)(o,o)$. Using that $R_3(P^0_\epsilon, P^0) = o(\epsilon^2)$, it follows:

$$f^{(2)}(o_1,o_2) - 0.25f^{(2)}(o_1) - 0.25f^{(2)}(o_2) = D^{(2)}(P^0)(o_1,o_2).$$

This completes the proof of Theorem 12. □

9.2 Validity of method for general models

We will now generalize the proof to arbitrary models. The key insight is that 1) we can apply the above proof to $\tilde{\Psi} : \mathcal{M}_{NP}(P^0) \to \mathbb{R}$ to obtain the numerical approximation result for the second-order efficient influence function $\tilde{D}^{(2)}(P^0)$, and 2), we will show below that the first and second-order efficient influence function of $\tilde{\Psi}$ at a $P^0 \in \mathcal{M}$ are identical to the first and second-order efficient influence function of $\Psi : \mathcal{M} \to \mathbb{R}$ at P^0.

The following lemma provides an important building block of the proof and is itself of interest. It shows that given a one dimensional parametric

model $\{P_\epsilon : \epsilon\} \subset \mathcal{M}_{NP}$ through $P \in \mathcal{M}$ with score S^*, its corresponding MLE-submodel $\{\tilde{P}_\epsilon : \epsilon\} \subset \mathcal{M}$, where $\tilde{P}_\epsilon$ is the Kullback-Leibler projection of P_ϵ onto $\mathcal{M}$, has score $\tilde{S} = \Pi(S^* \mid T(P))$ (i.e., the projection of S^* on the tangent space $T(P)$ in the Hilbert space $L_0^2(P)$).

Lemma 3 *Consider a submodel $\{P_\epsilon : \epsilon\} \subset \mathcal{M}_{NP}$ through $P \in \mathcal{M}$ at $\epsilon = 0$ with score $S^* \in L_0^2(P)$. Here $S^* = \lim_{\epsilon \to 0} \frac{dP_\epsilon - dP}{dP\epsilon}$ with this limit defined in the Hilbert space $L_0^2(P)$.*

Let $\tilde{P}_\epsilon = \arg\max_{P_1 \in \mathcal{M}(P)} P_\epsilon \log dP_1/dP$, and recall that $T(P)$ is the tangent space at P for model $\mathcal{M}(P) \ll P$. Note that $\{\tilde{P}_\epsilon : \epsilon\} \subset \mathcal{M}(P)$.

Regularity conditions: *Assume the score $\tilde{S} \equiv \lim_{\epsilon \to 0} \frac{d\tilde{P}_\epsilon - dP}{\epsilon dP}$ of $\{\tilde{P}_\epsilon : \epsilon\}$ at $\epsilon = 0$ exists as a limit in $L_0^2(P)$. Assume that $\tilde{P}_\epsilon$ solves score equations $P_\epsilon S_h(\tilde{P}_\epsilon) = 0$ for $h \in \mathcal{H}$, where, for each $P_1 \in \mathcal{M}$, the closure of the linear span of $\{S_h(P_1) : h \in \mathcal{H}\} \subset T(P_1)$ in $L_0^2(P_1)$ equals $T(P_1)$. Assume that*

$$\begin{aligned} \sup_{h \in \mathcal{H}} \mid (\tilde{P}_\epsilon - P)\{S_h(\tilde{P}_\epsilon) - S_h(P)\} \mid &= o(\epsilon) \\ \sup_{h \in \mathcal{H}} \mid (P_\epsilon - P)\{S_h(\tilde{P}_\epsilon) - S_h(P)\} \mid &= o(\epsilon). \end{aligned}$$

Then, $\tilde{S}$ is given by

$$\tilde{S} = \Pi(S^* \mid T(P)).$$

Proof: Let $S(P) = (S_h(P) : h \in \mathcal{H})$. We have $\tilde{P}_\epsilon S(\tilde{P}_\epsilon) = 0$ and $P_\epsilon S(\tilde{P}_\epsilon) = 0$. Thus:

$$(\tilde{P}_\epsilon - P)S(\tilde{P}_\epsilon) = (P_\epsilon - P)S(\tilde{P}_\epsilon).$$

The left-hand side can be written as:

$$P\frac{d\tilde{P}_\epsilon - dP}{dP}S(\tilde{P}_\epsilon).$$

The right-hand side can be written as:

$$P\frac{dP_\epsilon - dP}{dP}S(\tilde{P}_\epsilon).$$

By assumption, $(\tilde{P}_\epsilon - P)\{S(\tilde{P}_\epsilon) - S(P)\} = o(\epsilon)$ and $(P_\epsilon - P)\{S(\tilde{P}_\epsilon) - S(P)\} = o(\epsilon)$. Thus,

$$P\frac{d\tilde{P}_\epsilon - dP}{\epsilon dP}S(P) = P\frac{dP_\epsilon - dP}{\epsilon dP}S(P) + o(1).$$

Taking the limit for $\epsilon \to 0$ on both sides, and using that, by assumption, the limit for $\epsilon \to 0$ of the integrand converges in $L_0^2(P)$ to the desired score, we obtain

$$P\tilde{S}S(P) = PS^*S(P).$$

Thus $P(\tilde{S} - S^*)S_h(P) = 0$ for all $h \in \mathcal{H}$. Since the linear span of $\{S_h(P) : h \in \mathcal{H}\}$ equals $T(P)$ this implies $S^* - \tilde{S} \perp T(P)$. Since $\tilde{S} \in T(P)$, this proves $\tilde{S} = \Pi(S^*|T(P))$. $\square$

Building on this lemma, the following theorem establishes that the first and second-order efficient influence function of the nonparametric extension $\tilde{\Psi}$ of Ψ at a $P^0 \in \mathcal{M}$ is equal to the first and second-order efficient influence function Ψ at P^0.

Theorem 13 *Consider a model $\mathcal{M}$ and $\Psi : \mathcal{M} \to \mathbb{R}$. Let $P^0 \in \mathcal{M}$ be given. Let $\mathcal{M}(P^0) = \{P_1 \in \mathcal{M} : P_1 \ll P^0\}$ and let $T(P^0) \subset L_0^2(P^0)$ be the tangent space at P^0.*

Consider now a locally nonparametric model $\mathcal{M}_{NP}(P^0) \supset \mathcal{M}(P^0)$ with tangent space at P^0 equal to $L_0^2(P^0)$ and dominated by P^0, and the nonparametric extension $\tilde{\Psi} : \mathcal{M}_{NP}(P^0) \to \mathbb{R}$ of Ψ defined by

$$\tilde{\Psi}(P) = \Psi(\tilde{P}),$$

where

$$\tilde{P} = \arg \max_{P_1 \in \mathcal{M}(P^0)} P \log dP_1/dP^0.$$

Here, for each $P \in \mathcal{M}_{NP}(P^0)$, $\tilde{P} \in \mathcal{M}(P^0)$ solves $PS_h(\tilde{P}) = 0$ for $h \in \mathcal{H}$, where the closure of the linear span of $\{S_h(\tilde{P}) : h \in \mathcal{H}\} \subset L_0^2(\tilde{P})$ equals $T(\tilde{P})$.

Assume that the regularity conditions of the previous Lemma 3 hold for each submodel $\{P_\epsilon^0 : \epsilon\} \subset \mathcal{M}_{NP}(P^0)$ over a class $\mathcal{J}^*$ of submodels whose tangent space equals $L_0^2(P^0)$. In addition, assume that Ψ is pathwise differentiable along a class of submodels that includes (or equals) the corresponding class $\mathcal{J}$ of submodels $\{\tilde{P}_\epsilon^0 : \epsilon\} \subset \mathcal{M}$ with tangent space $T(P^0)$.

Let $D^{(1)}(P^0)$ be the efficient influence function of $\Psi : \mathcal{M} \to \mathbb{R}$ at P^0. Then, $\tilde{\Psi}$ is pathwise differentiable at P^0 with efficient influence function

$$\tilde{D}^1(P^0) = D^{(1)}(P^0).$$

Suppose now that Ψ and $\tilde{\Psi}$ are second-order pathwise differentiable at P^0 along the classes $\mathcal{J}$ and $\mathcal{J}^*$ of submodels, respectively. Let $D^{(2)}(P^0)$ be the second-order efficient influence function of $\Psi : \mathcal{M} \to \mathbb{R}$ at P^0, and let $\tilde{D}^{(2)}(P^0)$ be the second-order efficient influence function of $\tilde{\Psi} : \mathcal{M}_{NP}(P^0) \to \mathbb{R}$ at P^0. Then, we also have

$$\tilde{D}^{(2)}(P^0) = D^{(2)}(P^0).$$

Proof: We provide the proof of the last statement which assumes Ψ is second-order pathwise differentiable at P^0. The first statement is a direct consequence of this proof. By definition of second-order pathwise differentiability of Ψ at P^0 w.r.t. the class $\mathcal{J}$ of submodels, for each submodel $\{P^0_\epsilon : \epsilon\}$ in $\mathcal{J}^*$ with score S^*, we have

$$\begin{aligned}\tilde{\Psi}(P^0_\epsilon) &= \Psi(\tilde{P}_\epsilon) \\ &= \Psi(P^0) + \epsilon P^0 D^{(1)}(P^0)\tilde{S} + 0.5\epsilon^2\{P^0 D^{(1)}(P^0)\tilde{S}_2 + P^{02}D^{(2)}(P^0)\tilde{S}^2\} + o(\epsilon^2).\end{aligned}$$

However, by Lemma 3, $\tilde{S} = \Pi(S^* \mid T(P^0))$ and $D^{(1)}(P^0) \in T(P^0)$ so that $P^0 D^{(1)}(P^0)\tilde{S} = P^0 D^{(1)}(P^0)S^*$. We also have $\Psi(P^0) = \tilde{\Psi}(P^0)$ since $P^0 \in \mathcal{M}$. Since $D^{(2)}(P^0)(o_1, \cdot) \in T(P^0)$, and $D^{(2)}(P^0)(\cdot, o_2) \in T(P^0)$, for all $(o_1, o_2) \in \mathcal{O}^2$, we have

$$\int D^{(2)}(P^0)(o_1, o_2)\tilde{S}(o_1)dP^0(o_1) = \int D^{(2)}(P^0)(o_1, o_2)S^*(o_1)dP^0(o_1),$$

and applying this again to the integral over o_2, we obtain

$$\begin{aligned}&\int \left\{\int D^{(2)}(P^0)(o_1, o_2)S^*(o_1)dP^0(o_1)\right\}\tilde{S}(o_2)dP^0(o_2) \\ &= \int \left\{\int D^{(2)}(P^0)(o_1, o_2)\tilde{S}(o_2)dP^0(o_2)\right\}S^*(o_1)dP^0(o_1) \\ &= \int\int D^{(2)}(P^0)(o_1, o_2)S^*(o_2)dP^0(o_2)S^*(o_1)dP^0(o_1) \\ &= P^{02}D^{(2)}(P^0)S^{*2}.\end{aligned}$$

Thus, we have shown:

$$\tilde{\Psi}(P^0_\epsilon) = \tilde{\Psi}(P^0) + P^0 D^{(1)}(P^0)S^* + 0.5\epsilon^2 P^{02}D^{(2)}(P^0)S^{*2}\} + 0.5\epsilon^2 P^0 D^{(1)}(P^0)\tilde{S}_2 + o(\epsilon^2).$$

Note that the last second-order expansion of $\tilde{\Psi}$ was implied by the second-order pathwise differentiability of Ψ at P^0. If $\tilde{S}_2$ can be replaced by S^*_2, then this proves second-order pathwise differentiability of $\tilde{\Psi}$ at P^0 with the same first and second-order efficient influence function. So we still need to show that $P^0 D^{(1)}(P^0)(\tilde{S}_2 - S^*_2) = 0$. Since $\tilde{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}$ is assumed to be second-order pathwise differentiable, we also have

$$\tilde{\Psi}(P^0_\epsilon) = \tilde{\Psi}(P^0) + P^0 D^{(1)}(P^0)S^* + 0.5\epsilon^2 P^{02}D^{(2)}(P^0)S^{*2}\} + 0.5\epsilon^2 P^0 D^{(1)}(P^0)S^*_2 + o(\epsilon^2).$$

Thus, taking the difference between the two second-order expansions of $\tilde{\Psi}(P^0_\epsilon)$, dividing by ϵ^2, and taking the limit as $\epsilon \to 0$, shows that $P^0 D^{(1)}(P^0)\tilde{S}_2 = P^0 D^{(1)}(P^0)S^*_2$. This completes the proof. As a remark, we strongly suspect that $P^0 D^{(1)}(P^0)(\tilde{S}_2 - S^*_2) = 0$ without having to assume second-order pathwise differentiability of $\tilde{\Psi}$. □

Due to this theorem we only need to develop a method for determining the first and second-order efficient influence function of $\tilde{\Psi} : \mathcal{M}_{NP}(P^0) \to \mathbb{R}$ for a nonparametric model. Theorem 12 provides the proof of its validity for general target parameters defined on a nonparametric model, which thus also applies to this particular $\tilde{\Psi}$. This results in the following final theorem.

Theorem 14 *Assume the conditions of Theorem 13, including second-order pathwise differentiability of both Ψ and $\tilde{\Psi}$ at $P^0 \in \mathcal{M}$. Then, $\tilde{\Psi} : \mathcal{M}_{NP}(P^0) \to \mathbb{R}$, defined by $\tilde{\Psi}(P) = \Psi(\tilde{P})$ with $\tilde{P} = \arg\max_{P_1 \in \mathcal{M}(P^0)} P \log dP_1/dP^0$, has the same first and second-order efficient influence functions $D^{(1)}(P^0)$ and $D^{(2)}(P^0)$, respectively, as Ψ.*

Consider the submodels $P_{\epsilon,o_1} = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o_1}$, $P_{\epsilon,o_2} = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,o_2}$, and $P_{\epsilon,(o_1,o_2)} = (1-\epsilon)P^0 + \epsilon\Delta_{\lambda,(o_1,o_2)}$.

Assume (16) for $\tilde{\Psi}$ at $P^0_{\epsilon,(o_1,o_2)}$ and P^0:

$$\begin{aligned}\tilde{\Psi}(P^0_{\epsilon,(o_1,o_2)}) - \tilde{\Psi}(P^0) &= (P^0_{\epsilon,(o_1,o_2)} - P^0)D^{(1)}(P^0) + (P^0_{\epsilon,(o_1,o_2)} - P^0)^2 D^{(2)}(P^0) \\ &\quad + R_3(P^0_{\epsilon,(o_1,o_2)}, P^0).\end{aligned}$$

In addition, assume $D^{(1)}(P^0)$ is continuous at o_1 and o_2, $D^{(2)}(P^0)$ is continuous at (o_1,o_2), (o_1,o_1) and (o_2,o_2). Finally, assume that $R_3(P^0_{\epsilon,(o_1,o_2)}, P^0) = o(\epsilon^2)$.

We have

$$\begin{aligned}f^{(2)}(P^0,(o_1,o_2)) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(\tilde{P}^0_{\epsilon,(o_1,o_2)})\Big|_{\epsilon=0} \\ &= 0.5D^{(2)}(P^0)(o_1,o_1) + 0.5D^{(2)}(P^0)(o_2,o_2) + D^{(2)}(P^0)(o_1,o_2).\end{aligned}$$

In addition, we have

$$\begin{aligned}f^{(2)}(P^0,o_1) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(\tilde{P}^0_{\epsilon,o_1})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_1,o_1) \\ f^{(2)}(P^0,o_2) &\equiv \lim_{\lambda\to 0} \frac{d^2}{d\epsilon^2}\Psi(P^0_{\epsilon,o_2})\Big|_{\epsilon=0} \\ &= 2D^{(2)}(P^0)(o_2,o_2)\end{aligned}$$

In particular,

$$D^{(2)}(P^0)(o_1,o_2) = f^{(2)}(o_1,o_2) - 0.25f^{(2)}(o_1) - 0.25f^{(2)}(o_2).$$

10 Using a targeted perturbation to compute efficient influence function as a function

In the previously discussed methods, we used a perturbation $P^0_{\epsilon,o}$ at $P^0 \in \mathcal{M}$ in the direction of a possible realization o of O that provided an MLE $\tilde{P}^0_\epsilon$ so that

$$D^*_\Psi(P^0)(o) = \lim_{\epsilon \to 0} \frac{\Psi(\tilde{P}^0_\epsilon) - \Psi(P^0)}{\epsilon}$$

is the efficient influence function of $\Psi : \mathcal{M} \to \mathbb{R}$ at P^0 evaluated at o. Here we used the notation $D^*_\Psi(P)$ for the efficient influence function of $\Psi : \mathcal{M} \to \mathbb{R}$ at P. Thus, even though the algorithm only yields the efficient influence function at a single value o, the same $\tilde{P}^0_\epsilon$ can be used to obtain the efficient influence function at o of any pathwise differentiable target parameter Ψ defined on the model $\mathcal{M}$. Nonetheless, if our goal is an efficient estimator of a particular $\Psi(P_0)$, then, the fact that $\tilde{P}^0_\epsilon$ can be universally applied across all target parameters is not very helpful. In that case the disadvantage of the proposed algorithm is that for the purpose of constructing an efficient estimator of $\Psi(P_0)$ with corresponding influence function based inference, one needs to rerun it for each observation $o = O_i$, $i = 1, \ldots, n$. Another disadvantage of the perturbation $P^0_{\epsilon,o}$ is that it typically requires a smoothing parameter λ.

In this section we propose another perturbation P^0_ϵ that is specifically targeted towards our specific Ψ, and as a result the corresponding $\tilde{P}^0_\epsilon$ will now yield the whole function $o \to D^*(P^0)(o)$. This perturbation will not rely on a smoothing parameter, but will rely on having an initial gradient at P^0 of Ψ. To construct an efficient estimator of $\Psi(P_0)$ and a corresponding confidence interval one only needs to run this algorithm once.

Let $D(P^0)(O)$ be a gradient of $\Psi : \mathcal{M} \to \mathbb{R}$. A gradient is often much easier to find than the actual canonical gradient and can be found by representing the pathwise derivative along a path as a covariance of the score of the path with a particular fixed function (same for all paths), where this latter function is now a gradient of the pathwise derivative. Alternatively, one might have a simple estimator available that is known to be asymptotically linear so that $D(P^0)$ can be defined as the influence function of that estimator. One can also define a nonparametric extension $\Psi_{NP} : \mathcal{M}_{NP} \to \mathbb{R}$ so that $\Psi_{NP}(P) = \Psi(P)$ for $P \in \mathcal{M}$, and define $D(P^0)$ as the gradient of the pathwise derivative of Ψ_{NP}. The latter gradient can be computed as a standard gradient as in the Appendix A2 of Rose and van der Laan (2011). Let $\{P^0_\epsilon : \epsilon\} \subset \mathcal{M}_{NP}$ be a parametric model through P^0 at $\epsilon = 0$ with score equal to $D(P^0)$. It is assumed that this model is chosen so that all probability distributions in this parametric family

are dominated by P^0. For example, we might use the exponential model:

$$dP_\epsilon^0(o) = C(\epsilon, P^0)\exp(\epsilon D(P^0))dP^0(o),$$

where $C(\epsilon, P^0) = \{\int_o \exp(\epsilon D(P^0)(o))dP^0(o)\}^{-1}$ is the normalizing constant. If ϵ is forced to be small enough, then one can also select the easier parametric family:

$$dP_\epsilon^0(o) = (1 + \epsilon D(P^0)(o))dP^0(o).$$

Suppose now that the regularity conditions of Lemma 3 apply to $\{P_\epsilon^0 : \epsilon\}$ and its corresponding MLE

$$\tilde{P}_\epsilon^0 = \arg\max_{P_1 \in \mathcal{M}(P^0)} P_\epsilon^0 \log dP_1/dP^0,$$

defined as earlier. Then, by application of Lemma 3, we have that the score of $\{\tilde{P}_\epsilon^0 : \epsilon\}$ at $\epsilon = 0$ equals the projection of $D(P^0)$ onto the tangent space $T(P^0)$, which thus equals the efficient influence function $D^*(P^0)$:

$$D^*(P^0) = \lim_{\epsilon \to 0} \frac{d\tilde{P}_\epsilon^0 - dP^0}{\epsilon dP^0}.$$

Thus, by selecting ϵ as a small value, for any $o \in \mathcal{O}$, we can approximate $D^*(P^0)(o)$ with

$$\frac{d\tilde{P}_\epsilon^0 - dP^0}{\epsilon dP^0}(o).$$

Theorem 15 *Let $D(P^0)$ be a gradient of $\Psi : \mathcal{M} \to \mathbb{R}$. Let $\{P_\epsilon^0 : \epsilon\} \ll P^0$ be a parametric model through P^0 at $\epsilon = 0$ with score at $\epsilon = 0$ equal to $D(P^0)$. Suppose that the regularity conditions of Lemma 3 apply to $\{P_\epsilon^0 : \epsilon\}$ and its corresponding MLE*

$$\tilde{P}_\epsilon^0 = \arg\max_{P_1 \in \mathcal{M}(P^0)} P_\epsilon^0 \log dP_1/dP^0,$$

where $\mathcal{M}(P^0)$ is a submodel of $\mathcal{M}$, dominated by P^0, whose tangent space at P^0 equals the tangent space of $\mathcal{M}$ at P^0. Then, the efficient influence function at P^0 can be represented as:

$$D^*(P^0) = \lim_{\epsilon \to 0} \frac{d\tilde{P}_\epsilon^0 - dP^0}{\epsilon dP^0}.$$

10.1 TMLE

The parametric submodel $\{\tilde{P}^0_\epsilon : \epsilon\} \subset \mathcal{M}$ is a least favorable parametric model through P^0 at $\epsilon = 0$, since its score at $\epsilon = 0$ equals the efficient influence function $D^*(P^0)$. As a consequence, it can be used to compute a TMLE. So let

$$\epsilon^0_n = \arg\max_\epsilon P_n \log \frac{d\tilde{P}^0_\epsilon}{dP^0},$$

and define the update of P^0 as $P^1_n = \tilde{P}^0_{\epsilon^0_n}$. Since ϵ^0_n is an interior maximum, it can be shown that under regularity conditions and under the assumption that P^0 approximates P_0, this one-step update already satisfies

$$P_n D^*(P^1_n) = o_P(1/\sqrt{n}).$$

As a consequence, under the usual regularity conditions of the TMLE, we have that $\Psi(P^1_n)$ is asymptotically efficient. Instead of stopping at the first step, one could also iterate the updating process till convergence: $P^{k+1}_n = \tilde{P}^k_{n,\epsilon^k_n}$, where $\epsilon^k_n = \arg\max_\epsilon P_n \log d\tilde{P}^k_{n,\epsilon}/dP^k$, $k = 1, 2, \ldots$, till $\epsilon^K_n \approx 0$. Then, we will have $P_n D^*(P^K) \approx 0$.

Let's consider the one-step TMLE. Note that this might require computing $\tilde{P}^0_\epsilon$ for a grid of ϵ-values, but one can also implement Newton-Raphson type algorithm by using that the derivative of the log likelihood in ϵ at a particular ϵ^0 is defined by the score of $\tilde{P}^0_\epsilon$ at ϵ^0 , and the latter score can be estimated with the numerical approximation $(d\tilde{P}^0_{\epsilon+\delta} - dP^0/(\delta dP^0)$ for some small δ. Since ϵ_n will be of the order of the Kullback-Leibler divergence between P^0 and P_0, the MLE ϵ_n can be expected to be larger than $1/\sqrt{n}$, so that this TMLE will thus not require searching for very small values ϵ.

10.2 Example: Bivariate right-censored data

Consider the bivariate right-censored data model in which we assume that (C_1, C_2) is independent of (T_1, T_2), and our target parameter $\Psi(P) = P(T_1 > t_{10}, T_2 > t_{20})$ is the survival probability at (t_{10}, t_{20}). Recall that the density of P factorizes into $p_Q p_G$. The efficient influence function $D^*(P)$ is the same in the model $\mathcal{M}(G_0)$ that assumes that the censoring cumulative distribution function G is known as in the actual model $\mathcal{M}$. Thus, we can select a gradient of $\Psi : \mathcal{M}(G_0) \to \mathbb{R}$. We can use the inverse probability of censoring weighted gradient:

$$D(P^0) = \frac{I(\tilde{T}_1 > t_{10}, \tilde{T}_2 > t_{20})\Delta_1\Delta_2}{\bar{G}^0(\tilde{T}_1, \tilde{T}_2)} - \Psi(P^0).$$

Consider now the following submodel through P^0:

$$dP^0_\epsilon(O) = dP^0(O)(1 + \epsilon D(P^0)).$$

We now define the MLE

$$\tilde{P}^0_\epsilon = \arg \max_{P_1 \in \mathcal{M}(G^0)} P^0_\epsilon \log dP_1/dP^0.$$

Note that $dP_1/dP^0 = p_{Q_1}/p_{Q^0}$ so that this MLE only involves maximizing over the parameter space of bivariate cumulative distributions Q. The implementation of $\tilde{P}^0_\epsilon$ can be implemented using discretization techniques as earlier described for the single observation perturbation $P^0_{\epsilon,o}$. Thus the MLE $\tilde{P}^0_\epsilon$ is determined by a $\tilde{Q}^0_\epsilon$ and G^0. For small $\epsilon > 0$, we now have that

$$D^*(P^0) \approx \frac{d\tilde{P}^0_\epsilon - dP^0}{dP^0\epsilon} = \frac{p_{\tilde{Q}^0_\epsilon} - p_{Q^0}}{p_{Q^0}\epsilon}.$$

One could use $\epsilon = 1/\sqrt{n}$ since then the approximation error for the efficient influence function will be $O(1/\sqrt{n})$, and that is still smaller or equal than the rate at which the initial estimator P^0 converges to P_0 and thereby $D^*(P^0)$ converges to $D^*(P_0)$, so that the finite sample behavior of the one-step estimator or TMLE is not meaningfully affected by this approximation error. As described above, we can now also immediately implement a TMLE by solving for ϵ that maximizes $P_n \log d\tilde{P}^0_\epsilon/dP^0$, so that the TMLE of $\Psi(P_0)$ can be defined as $\Psi(\tilde{P}^0_{\epsilon_n})$.

10.3 Generalization to minimum loss-based mapping

Consider the case that $\Psi(P) = \Psi_1(Q(P))$, $Q(P) = \arg\min_{Q \in Q(\mathcal{M})} PL(Q)$ for a loss function $(O, Q) \to L(Q)(O)$, and $D^*(P) = D^*(Q(P), G(P))$ for some nuisance parameter $G(P)$. Let (Q^0, G^0) be given and consider a mapping $(Q, G) \to P_{Q,G} \in \mathcal{M}$, so that we can define $P^0 = P_{Q^0,G^0} \in \mathcal{M}$ as a uniquely defined data distribution consistent with (Q^0, G^0). Let $Q^0_\epsilon \in Q(\mathcal{M})$ be a perturbation of Q^0 chosen so that $P^0_\epsilon = P_{Q^0_\epsilon,G^0}$ is a perturbation of P^0 with score at $\epsilon = 0$ equal to an initial gradient $D(Q^0, G^0)$. We now define the minimum loss projection $\tilde{Q}^0_\epsilon$ of Q^0_ϵ onto the model space $\mathcal{Q}(P^0) \subset Q(\mathcal{M})$ as follows:

$$\tilde{Q}^0_\epsilon = \arg \min_{Q \in \mathcal{Q}(P^0)} P^0_\epsilon L(Q),$$

where $\mathcal{Q}(P^0)$ is an appropriate subspace of the parameter space $\{Q(P) : P \in \mathcal{M}\}$. The key assumption is that the set of score equations $P_\epsilon S_h(\tilde{Q}^0_\epsilon) = 0$, $h \in$

$\mathcal{H}$, solved by this MLE $\tilde{Q}_\epsilon^0$ is rich enough so that solving these score equations implies $P_\epsilon^0 D^*(\tilde{Q}_\epsilon^0, G^0) = 0$. Here $S_h(Q) = \frac{d}{d\delta} L(Q_{h,\delta})\big|_{\delta=0}$ is the generalized score at Q using a path $\{Q_{h,\delta} : \delta\} \subset Q(\mathcal{M})$ through Q at $\delta = 0$, indexed by a choice h ranging over an index set $\mathcal{H}$. Under appropriate regularity conditions, we have that for small $\epsilon > 0$

$$D^*(Q^0, G^0) \approx \frac{dP_{\tilde{Q}_\epsilon^0, G^0} - dP_{Q^0, G^0}}{dP_{Q^0, G^0}\epsilon}.$$

The result follows immediately if one can show that the mapping $(Q, G) \to P_{Q,G}$ is chosen so that for all $h \in \mathcal{H}$

$$\frac{d}{d\delta} L(Q_{h,\delta})\bigg|_{\delta=0} = -\frac{d}{d\delta} \log \frac{dP_{Q_{h,\delta}^0, G^0}}{dP_{Q^0, G^0}}\bigg|_{\delta=0}, \text{ at } Q = \tilde{Q}_\epsilon^0.$$

11 Discussion

In this article we demonstrated that one can compute the efficient influence function at a data distribution P^0 in the statistical model by computing a maximum likelihood estimator $\tilde{P}_\epsilon^0$, or, more generally, a minimum loss-based estimator $\tilde{Q}_\epsilon^0$, but with the usual empirical distribution replaced by a perturbation P_ϵ^0 of P^0 and selecting $\epsilon \approx 0$. We proposed two types of perturbation, a perturbation in the direction of a single observation and a targeted perturbation defined by its score at $\epsilon = 0$ being an initial gradient. The first perturbation can be inputted in the target parameter mapping and results in that way in an approximation for the efficient influence function at P^0 for any target parameter. The targeted perturbation directly generates the whole efficient influence function at P^0 as a function in o. The first perturbation relies on a regularization/smoothing parameter which needs to be tuned as a function of ϵ ,while the targeted perturbation does not depend on such a tuning parameter. We developed formal conditions under which these methods for computing the efficient influence function are valid, which appear to be very weak. We also demonstrated the methods with three examples. In future work we plan to implement these methods and corresponding one-step and TML estimators in order to evaluate and develop its practical feasibility in realistic estimation problems. Our results promise the development of efficient TMLE of pathwise differentiable target parameters that are only based on the capability to compute an MLE over the statistical model at a smooth perturbation of initial estimator P^0, thereby making efficient estimation accessible to computer savvy scientists that are good in implementing algorithms that maximize a criterion.

Acknowledgements

Mark van der Laan was supported by NIH grant R01 AI074345-06. Marco Carone was supported by a Genentech Endowed Professorship at the University of Washington. Alex Luedtke was supported by the NDSEG Fellowship Program of the U.S. Department of Defense.

References

H. Bang and J.M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

M. Carone, I. Díaz, and M.J. van der Laan. Higher-order targeted minimum loss-based estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 2014.

M.N. Chang. Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Annals of Statistics*, 18:391–404, 1990.

M.N. Chang and G. Yang. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Annals of Statistics*, 15: 1536–1547, 1987.

C. Frangakis, T. Qian, Z. Wu, and I. Diaz. Deductive derivation and turing-computerization of semiparametric efficient estimation. *to appear in Biometrics*, 2015.

P. Groeneboom and J.A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*. Birkhauser verlag, 1992.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Statistics*, 27:887–906, 1956.

A.R. Luedtke, M. Carone, and M.J. van der Laan. Discussion of deductive derivation and turing-computerization of semiparametric efficient estimation. *to appear in Biometrics*, 2015.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M.J. van der Laan. Targeted minimum loss based estimation of marginal structural working models. *Journal of Causal Inference*, 2:DOI: 10.1515/jci–2013–0007, also available at: http://biostats.bepress.com/ucbbiostat/paper312/, 2014.

C.M. Quale, M.J. van der Laan, and J.M. Robins. Locally efficient estimation with bivariate right censored data. *Journal of the American Statistical Association*, 101:1076–1084, 2006.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

J.M. Robins, L. Li, E. Tchetgen Tchetgen, and A.W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

J.M. Robins, L. Li, E. Tchetgen Tchetgen, and A.W. van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69(2-3):227–247, 2009.

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

A. Rotnitzky, Q. Lei, M. Sued, and J. M. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2): 439–456, doi: 10.1093/biomet/ass013, 2012.

M.J. van der Laan. *Efficient and Inefficient Estimation in Semiparametric Models*. Center for Mathematics and Computer Science, CWI-tract 114, 1996a.

M.J. van der Laan. *Efficient and Inefficient Estimation in Semiparametric Models*. Centre of Computer Science and Mathematics, Amsterdam, cwi tract 114 edition, 1996b.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *International Journal of Biostatistics*, 8:doi: 10.1515/1557–4679.1370, 2012.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996.

A.W. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, forthcoming.

A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.