University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2014 *Paper* 331

Higher-order Targeted Minimum Loss-based Estimation

Marco Carone[*] Iván Díaz[†]

Mark J. van der Laan[‡]

[*]University of Washington, mcarone@uw.edu

[†]Johns Hopkins Bloomberg School of Public Health, idiaz@jhu.edu

[‡]University of California, Berkeley, laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper331

Copyright ©2014 by the authors.

Higher-order Targeted Minimum Loss-based Estimation

Marco Carone, Iván Díaz, and Mark J. van der Laan

Abstract

Common approaches to parametric statistical inference often encounter difficulties in the context of infinite-dimensional models. The framework of targeted maximum likelihood estimation (TMLE), introduced in van der Laan & Rubin (2006), is a principled approach for constructing asymptotically linear and efficient substitution estimators in rich infinite-dimensional models. The mechanics of TMLE hinge upon first-order approximations of the parameter of interest as a mapping on the space of probability distributions. For such approximations to hold, a second-order remainder term must tend to zero sufficiently fast. In practice, this means an initial estimator of the underlying data-generating distribution with a sufficiently large rate of convergence must be available – in many cases, this requirement is prohibitively difficult to satisfy. In this article, we propose a generalization of TMLE utilizing a higher-order approximation of the target parameter. This approach yields asymptotically linear and efficient estimators when a higher-order remainder term is asymptotically negligible. The latter condition is often much less stringent than that arising in a regular first-order TMLE. Beyond relaxing regularity conditions, use of a higher-order TMLE can improve inference accuracy in finite samples due to its explicit reliance on a higher-order approximation. We provide the theoretical foundations of higher-order TMLE and study its use for estimating a counterfactual mean when all potential confounders have been measured. We show, in particular, that the implementation of a higher-order TMLE is nearly identical to that of a regular first-order TMLE. Since higher-order TMLE requires higher-order differentiability of the target parameter, a requirement that often fails to hold, we also discuss and study practicable approximation strategies that allow us to circumvent this failure in applications.

1 Introduction

1.1 Motivation

Statisticians are generally concerned with making inference on a target parameter of the data-generating distribution based on draws from a probability distribution known only to lie in some model. Estimation and statistical inference is generally straightforward in parametric and small semiparametric models. Unfortunately, these models are often overly simplistic. Large semiparametric and nonparametric models usually provide a more accurate reflection of the background knowledge available on a given scientific problem. However, estimation and inference within these models is a much more subtle matter. For example, maximum likelihood estimators often do not exist, and even when they do, they may be inconsistent. If the target parameter involves the unknown data-generating distribution via an intermediate quantity whose estimation requires smoothing techniques, drawing statistical inference is especially challenging. This is the case, for example, whenever the parameter involves inherently local features of the data-generating distribution, such as a density or conditional mean function. The implementation of smoothing techniques usually involves the selection of tuning parameters that govern the bias-variance trade-off in practice. Performing an optimal bias-variance trade-off for the involved intermediate quantity is generally well understood and easy to accomplish. However, in general, this does not result in an optimal bias-variance trade-off for the target parameter. Estimators of the target parameter utilizing suboptimally tuned smoothing are usually not asymptotically linear and common approaches for quantifying their uncertainty are not valid.

Several approaches have been proposed for tackling this challenge, including one-step estimation (e.g., Bickel et al., 1997) and estimating equations methodology (e.g., van der Laan and Robins, 2003). More recently, targeted maximum likelihood (or minimum loss-based) estimation, hereafter referred to as TMLE (e.g., van der Laan and Rubin, 2006, van der Laan and Rose, 2011), was introduced as a general framework for constructing asymptotically linear and efficient substitution estimators of low-dimensional target parameters in rich infinite-dimensional models. Among other features distinguish it from earlier approaches, TMLE also seeks to guarantee that constructed estimators have sound finite-sample behavior, as discussed below. This results in a framework that facilitates the construction of estimators of target parameters that are not only asymptotically linear and efficient under much weaker conditions than required in parametric or restrictive semiparametric models but also strive to have good finite-sample performance.

The construction of these estimators heavily depends upon certain first-order asymptotic representations, as described below. For these representations to be useful, the resulting second-order remainder term must tend to zero in probability faster than $n^{-1/2}$. This ensures that the first-order approximation suffices to guide the construction of estimators and to study their asymptotic limit theory. To satisfy this condition, it must be possible to construct an estimator of the data-generating distribution, or

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

any relevant portions thereof, that converges sufficiently fast. For example, when the density of the data-generating distribution is directly involved in the target parameter, a density estimator converging in a suitable norm at a rate faster than $n^{-1/4}$ is often required to guarantee that the second-order remainder term is negligible. In many settings however, it is rather implausible for such a condition to hold, particularly when the data unit vector is high-dimensional. The remainder term will then itself contribute to the first-order asymptotic behavior of the estimator and derail it from asymptotic linearity. It is then natural to consider the construction of estimators based on higher-order asymptotic expansions, allowing us to instead require that a higher-order remainder term be asymptotically negligible. Minimal rate conditions would then be significantly relaxed. For example, if a k^{th} order expansion exists and could be utilized, an estimator of the density function converging at a rate faster than $n^{-1/[2(k+1)]}$ would suffice for the resulting remainder term to be negligible. More importantly, even when such higher-order expansions do not exist, approximate higher-order expansions can generally be constructed, leading to concrete rate gains. In the latter case however, describing the resulting minimal rate conditions in generality is more difficult since these can be somewhat context-dependent.

The objective of this paper is to describe how the TMLE framework can be generalized to explicitly utilize higher-order rather than first-order asymptotic representations. The practical significance of this is to provide guidelines for constructing estimators that have sound behavior in finite samples and are asymptotically linear and efficient under less restrictive conditions.

1.2 Brief review of the relevant literature

The building blocks of this generalization were set several decades ago, notably in the works of J. Pfanzagl [Pfanzagl, 1985], wherein the notion of higher-order gradients was introduced and higher-order expansions of finite-dimensional parameters over arbitrary model spaces were formalized. Influence functions having originated in the field of robust statistics, it is no surprise that their higher-order extensions have been used to produce a refined theory of robust statistics. La Vecchia et al. [2012] recently proposed precisely such a refinement. Our approach, however, is very different in spirit: we seek to utilize higher-order expansions to enable and guide the construction of regular and asymptotically linear estimators of statistical parameters in rich infinite-dimensional models. This is the perspective that motivated the seminal contributions of J. Robins, L. Li, E. Tchetgen & A. van der Vaart (e.g., Robins et al., 2008, 2009, Li et al., 2011, van der Vaart, forthcoming); these authors are the first to have provided a rigorous framework for precisely addressing this problem. The first exposition is that of Robins et al. [2008], where the focus resides primarily on the use of higher-order gradients to derive optimal estimators in settings where regular estimation is not possible. Subsequent works are concerned with the development of a higher-order analogue of the one-step estimator introduced early on in Levit [1975], Ibragimov and Khasminskii [1981], Pfanzagl [1982] and Bickel [1982], for example. The general approach is thus to

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

identify the dominating terms of an asymptotic expansion for a naive plug-in estimator and to perform, accordingly, an additive correction on this naive estimator. In Robins et al. [2009], the authors carefully establish the required foundations of a second-order extension of this approach and illustrate its use in the context of two problems, namely that of estimating the square of a density function and of estimating the mean response in missing data models. The latter problem is equivalent to estimating a mean counterfactual outcome in the absence of unmeasured confounders. In Li et al. [2011], the authors study in great detail the problem of inferring a treatment effect using this approach and also establish minimax rate results for certain situations in which regular inference is not possible.

An excellent review of the general higher-order extension of the one-step estimator is provided in van der Vaart [forthcoming]. The one-step estimator is simple and easy to describe. However, in finite samples, it is vulnerable to decreased performance since it does not include any safeguard ensuring that the additive correction performed on the naive plug-in estimator does not drive the estimator near and possibly beyond the natural bounds of the parameter space. Rather than performing post-hoc bias correction in the parameter space, as does the one-step procedure, the TMLE framework, which we focus on in this paper, provides guidelines for constructing an estimator of the underlying data-generating distribution, or whichever portion of it is needed to compute the parameter of interest, such that the resulting plug-in estimator enjoys regular and asymptotically linear asymptotic behavior. As such, the correction step is performed in the model space rather than in the parameter space. The appeal of devising such an approach, even decades before its actual development, was highlighted in Pfanzagl [1982]. Since the resulting estimator automatically satisfies bounds on the parameter space, it never produces nonsensical output, such as a probability estimate outside the unit interval, and can outperform in finite samples asymptotically equivalent estimators that do not have a plug-in form. This issue arises when comparing first-order TMLE and one-step estimators, but is likely to be of even greater importance in higher-order inference due to the increased complexity of the correction process.

As is highlighted by Robins et al. [2009] and further discussed in this article, higher-order gradients do not exist for several statistical parameters of interest. Nonetheless, approximate higher-order gradients can be used to produce regular estimators. In practice, the notion of approximate higher-order gradients necessarily involves the selection of certain tuning parameters. This requirement is discussed in Robins et al. [2009] but practical guidelines are neither provided nor appear particularly easy to develop. An advantage of using TMLE in this context, as we demonstrate, is that the selection of such tuning parameters can be effortlessly embedded in the framework of collaborative TMLE (e.g., van der Laan and Gruber, 2010, Gruber and van der Laan, 2010, Stitelman and van der Laan, 2010, Gruber and van der Laan, 2012, van der Laan and Rose, 2011), hereafter referred to as C-TMLE. Thus, the practical complications associated with the need for carefully-tuned approximations in the context of inference based on higher-order expansions can be tackled readily. Furthermore, as will be discussed, the framework of TMLE also provides useful tools that aim to guarantee that the resulting

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

higher-order estimator will not behave worse than the usual first-order TMLE.

1.3 Contributions and outline of this article

In this article we propose a novel k^{th} order targeted minimum loss-based estimator of a k^{th} order pathwise-differentiable target parameter, based on n independent draws from an unknown element of a given semiparametric model. We will refer to this estimator as a k-TMLE, with 1-TMLE corresponding to the usual first-order TMLE. Analogously to the large-sample properties of the usual TMLE requiring the asymptotic negligibility of a second-order remainder term, the asymptotic normality and efficiency of this k-TMLE will rely on the asymptotic negligibility of a $(k+1)^{th}$ order remainder term. As an illustration of the construction of this general k-TMLE, we develop a 2-TMLE of the mean of a counterfactual outcome under fixed treatment based on observing n independent and identically distributed copies of a random vector consisting of baseline covariates, a binary treatment and a final binary outcome. It will indeed be the case that we will often focus on the 2-TMLE to simplify the exposition. Nonetheless, a description of the extension to the general k-TMLE will usually be provided as well.

We provide a general template for constructing a 2-TMLE and for establishing its asymptotic efficiency. We show that this 2-TMLE can be constructed within the standard framework of TMLE, notably by augmenting the least-favorable parametric submodel used in the standard TMLE with an additional parameter. In addition, we demonstrate that the statistical inference of this 2-TMLE can be based on a second-order Taylor expansion, possibly providing some finite-sample improvements in the construction of confidence intervals.

As mentioned above, in many problems, the target parameter is only first-order pathwise-differentiable and a second-order gradient does not exist. As a solution, we propose a 2-TMLE based on an approximate second-order gradient and present a corresponding theory. In the context of our running example, a second-order gradient exists when the baseline covariates have finite support but not otherwise. We demonstrate how our proposed remedy is constructed and studied in the latter case, where a second-order gradient is replaced by a kernel-based approximation of the second-order gradient found in the context of finitely-supported covariates. Unfortunately, this additional approximation yields an additional bias term for the resulting 2-TMLE, referred to as the *representation error* in Robins et al. [2009]. As a result, the asymptotic linearity and efficiency theorem of the 2-TMLE require this bias term to be asymptotically negligible as well. This requires careful study in any given application, as is discussed in greater detail below.

As indicated previously, control of this bias term can be carried out within the standard TMLE framework at no risk of losing the desirable properties of the first-order TMLE. In our example, we demonstrate that, for an appropriate choice of kernel and bandwidth rate, and provided the data-generating distribution satisfies certain smoothness conditions, the asymptotic efficiency of the k-TMLE still only relies on the asymptotic negligibility of a $(k+1)^{th}$ order remainder term. We also propose a data-

http://biostats.bepress.com/ucbbiostat/paper331

adaptive bandwidth selector based on the C-TMLE framework, essentially selecting the bandwidth maximizing bias reduction in the targeting step of the TMLE.

We propose a concrete 2-TMLE for a mean counterfactual outcome, our running example, and provide an asymptotic efficiency theorem including a closed-form expression for the third-order remainder. In a companion article, we illustrate the behavior of this estimator in a simulation study and data analysis, and provide R code implementing the estimator with corresponding statistical inference.

This article is organized as follows. In Section 2, we provide the basic idea underlying the development of the higher-order TMLE in terms of higher-order gradients and we present the corresponding efficiency theory. This section should provide readers with a rather succinct overview of the various ideas discussed in this article. In Section 3, we precisely define higher-order gradients and canonical gradients, and describe how to calculate them. We perform these calculations in our running example, where we wish to learn about a treatment-specific counterfactual mean $E_{P_0}E_{P_0}(Y \mid A = 1, W)$ using n independent copies of $O = (W, A, Y) \sim P_0$, where P_0 lies in a nonparametric model $\mathcal{M}$ and W is assumed to have a finite support. We formally establish the desired second-order expansion of the target parameter in terms of the first-order canonical gradient and second-order partial canonical gradient; this serves as the basis of the asymptotic efficiency proof for the 2-TMLE. In Section 4, we propose the k-TMLE for k^{th} order pathwise-differentiable target parameters, provide a template for establishing asymptotic efficiency of this estimator, and illustrate the implementation of a 2-TMLE in the context of our running example under the assumption that W is finitely-supported. We suggest, in Section 5, a general template for the construction of a k-TMLE when higher-order partial canonical gradients do not exist but can be replaced by surrogate approximations depending on a tuning parameter h that determines the bias due to the representation error. In addition, we present a general asymptotic efficiency theorem and discuss strategies for selecting h data-adaptively using C-TMLE. We further discuss its use in the context of our running example when W is not necessarily finitely-supported. We prove formal theorems establishing asymptotic efficiency of these 2-TMLEs and explicitly provide the form of the involved remainders so that the asymptotic negligibility of the third-order remainder term can be carefully scrutinized and contrasted with the second-order remainder term arising in the usual TMLE. We provide a discussion and concluding remarks in Section 6. The proofs of our lemmas and theorems are gathered in the Appendix.

2 Overview of higher-order TMLE

2.1 Targeted maximum likelihood estimation

We consider the setting whereby n independent draws $O_1, \ldots, O_n$ are obtained from $P_0 \in \mathcal{M}$, where the statistical model $\mathcal{M}$ refers to the set of all possible probability distributions for the prototypical data structure O. Let P_n be the empirical distribution

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

of $O_1, \ldots, O_n$, defined as the discrete probability distribution assigning equal mass n^{-1} to each O_i, $i = 1, 2, \ldots, n$. Define $\mathcal{O} := \cup_{P \in \mathcal{M}} \text{supp}(P)$ with $\text{supp}(P)$ denoting the support of P; in other words, denote by $\mathcal{O}$ the union of the support of O under each $P \in \mathcal{M}$. We are also given a statistical target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ and are interested in inferring about the true parameter value $\psi_0 := \Psi(P_0)$ from the observed data. For simplicity, we focus on univariate target parameters in this article but all the developments herein can immediately be extended to Euclidean-valued target parameters with image in $\mathbb{R}^d$ as well.

We consider the case that Ψ is pathwise-differentiable at each $P \in \mathcal{M}$ with first-order canonical gradient, also known as the efficient influence function, $D^{(1)*}(P)$, so that a regular estimator of ψ_0 is asymptotically efficient if and only if it is asymptotically linear with influence function given by $D^{(1)*}(P_0)$ [Bickel et al., 1997]. Formally, a regular estimator ψ_n of ψ_0 is asymptotically efficient if and only if

$$\psi_n - \psi_0 = (P_n - P_0)D^{(1)*}(P_0) + o_P(n^{-1/2}) \ ,$$

where for any function f we write Pf to denote $\int f(o)dP(o)$.

TMLE provides a general template for constructing asymptotically efficient substitution estimators $\psi_n^* := \Psi(P_n^*)$ of ψ_0, whose asymptotic efficiency is in part a consequence of the property

$$P_n D^{(1)*}(P_n^*) = \frac{1}{n} \sum_{i=1}^{n} D^{(1)*}(P_n^*)(O_i) = 0 \ , \tag{2.1}$$

and the general first-order expansion

$$\begin{aligned} \Psi(P) - \Psi(P_0) &= (P - P_0)D^{(1)*}(P) + R_2(P, P_0) \\ &= -P_0 D^{(1)*}(P) + R_2(P, P_0) \end{aligned} \tag{2.2}$$

with $R_2(P, P_0)$ a second-order term. What is intended here as a second-order term is the obvious analogue from the finite-dimensional case. More formally though, we can define a term $R_{k+1}(P, P_0)$ to be of $(k+1)^{th}$ order if there exists a dissimilarity d such that $R_{k+1}(P, P_0)/d(P, P_0)^k$ tends to zero as P tends to P_0 in d. A second-order term will often have the form

$$\int \left[f_1(P)(o) - f_1(P_0)(o)\right]\left[f_2(P)(o) - f_2(P_0)(o)\right] f_3(P, P_0)(o) d\mu(o)$$

for some nuisance parameters f_1, f_2 and f_3 and a measure μ. Identity (2.1) can be combined with (2.2) evaluated at $P = P_n^*$ to yield that

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^{(1)*}(P_n^*) + R_2(P_n^*, P_0) \ . \tag{2.3}$$

This forms the basis of any theorem establishing the asymptotic linearity and efficiency of a TMLE. Provided it can be shown that

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

(a) $D^{(1)*}(P_n^*)$ falls in a P_0-Donsker class with probability tending to one,

(b) $P_0\left[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)\right]^2$ tends to zero in probability, and

(c) $R_2(P_n^*, P_0)$ tends to zero in probability faster than $n^{-1/2}$,

results from empirical process theory (e.g., van der Vaart and Wellner, 1996) imply the desired asymptotic efficiency, notably that

$$\psi_n^* - \psi_0 = (P_n - P_0)D^{(1)*}(P_0) + o_P(n^{-1/2}) \ .$$

To construct an estimator P_n^* solving (2.1), TMLE relies on an initial estimator $P_{n,0}$ of P_0, obtained, for example, using super learning for optimal performance, and a so-called least-favorable parametric submodel $\{P_{n,0}(\epsilon) : \epsilon\} \subset \mathcal{M}$ such that

(i) $P_{n,0}(0) = P_{n,0}$, and

(ii) $D^{(1)*}(P_{n,0})$ lies in the closure of the linear span of all scores of ϵ at $\epsilon = 0$.

A single update of TMLE is given by $P_{n,1} := P_{n,0}(\epsilon_n^0)$, where $\epsilon_n^0 := \operatorname{argmax}_\epsilon P_n \log p_{n,0}(\epsilon)$ is the MLE in the parametric submodel constructed and $p_{n,0}(\epsilon) := dP_{n,0}(\epsilon)/d\mu$ is the Radon-Nikodym derivative of $P_{n,0}$ with respect to some dominating measure μ. Subsequently, the least-favorable submodel previously constructed using $P_{n,0}$ is constructed using $P_{n,1}$ and the process is repeated to map $P_{n,1}$ into a further update $P_{n,2}$. This process is repeated until convergence, considered to have occurred at step k provided ϵ_n^k is approximately zero in some appropriate sense. The TMLE of P_0 is then defined as $P_n^* := P_{n,k}(\epsilon_n^k)$, while the TMLE of ψ_0 is the corresponding substitution estimator $\psi_n^* := \Psi(P_n^*)$. At each step j, the MLE ϵ_n^j of ϵ solves its score equation $0 = \frac{\partial}{\partial\epsilon} P_n \log p_{n,j}(\epsilon)\big|_{\epsilon=\epsilon_n^j}$, where $p_{n,j}(\epsilon) := dP_{n,j}(\epsilon)/d\mu$ is the Radon-Nikodym derivative of $P_{n,j}$ relative to μ. Therefore, at step k, we obtain that

$$P_n D^{(1)*}(P_n^*) = \left.\frac{\partial}{\partial\epsilon} P_n \log p_n^*(\epsilon)\right|_{\epsilon=0} = \left.\frac{\partial}{\partial\epsilon} P_n \log p_{n,k}(\epsilon)\right|_{\epsilon=\epsilon_n^k} \approx 0$$

since the submodel $\{P_n^*(\epsilon) : \epsilon\}$ has score $D^{(1)*}(P_n^*)$ for ϵ at $\epsilon = 0$, where $p_{n,k}$ and p_n^* are the Radon-Nikodym derivatives of $P_{n,k}$ and P_n^*, respectively, relative to μ.

2.2 Extensions of TMLE

The above iterative algorithm relies on the loglikelihood loss, which directly motivates the nomenclature *targeted maximum likelihood estimator.* However, provided an appropriate loss function can be identified, the same idea can be applied to any representation of the target parameter as $\Psi(P) = \Psi_1(Q)$ with $Q := Q(P)$, where $Q(P)$ represents the relevant portion of P for the sake of computing $\Psi(P)$, provided $D^{(1)*}(P) := D^{(1)*}(Q, g)$ can be written in terms of Q and a nuisance parameter $g := g(P)$. This generalized algorithm, referred to as *targeted minimum loss-based estimation*, requires an initial estimator $(Q_{n,0}, g_n)$ of (Q_0, g_0), a loss function $(o, Q) \mapsto L(Q)(o)$ such that

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$Q_0 = \operatorname{argmin}_Q P_0 L(Q)$, a least-favorable submodel $\{Q_{n,0}(\epsilon) : \epsilon\} \subset \{Q(P) : P \in \mathcal{M}\}$ which generally depends on g_n and is such that components of the generalized score

$$\frac{\partial}{\partial \epsilon} L(Q_{n,0}(\epsilon))\bigg|_{\epsilon=0}$$

span $D^{(1)*}(Q_{n,0}, g_n)$. An updating scheme directly analogous to that described above is then defined and yields a targeted estimator Q_n^* and corresponding targeted minimum loss-based estimator $\Psi_1(Q_n^*)$ of ψ_0 such that

$$P_n D^{(1)*}(Q_n^*, g_n) = 0 \ ,$$

an integral requirement for asymptotic linearity and efficiency.

A simple but key observation is that, for the sake of ascribing additional properties to the TMLE, a targeted estimate g_n^* of g_0 can be constructed to solve specified equations and Q_n^* can easily be tailored to solve additional equations of interest. This is accomplished notably by incorporating additional parameters in the least-favorable parametric submodel through $(Q_{n,0}, g_n)$. This generality of TMLE has been utilized in several instances before (e.g., van der Laan and Rubin, 2006, Rubin and van der Laan, 2011, van der Laan and Rose, 2011, Gruber and van der Laan, 2012, Lendle et al., 2013, van der Laan, 2014) and plays a fundamental role in the construction of a higher-order TMLE, as described in this article.

2.3 Construction of a higher-order TMLE

One of the main conditions in the proof of asymptotic efficiency for the TMLE is that the second-order term $R_2(P_n^*, P_0)$ must be $o_P(n^{-1/2})$. In this article, our goal is to construct a TMLE in which this second-order remainder condition is replaced by a third-order remainder condition, requiring instead that $R_3(P_n^*, P_0)$ be $o_P(n^{-1/2})$. More generally, with additional work, it will be possible to substitute this condition by a $(k+1)^{th}$ order remainder condition requiring that $R_{k+1}(P_n^*, P_0)$ be $o_P(n^{-1/2})$. This will facilitate the construction of estimators known to be asymptotically linear and efficient under less stringent and therefore more realistic conditions on rates of convergence of the initial estimators used in TMLE.

To achieve this, it will be necessary to include additional targeted fitting in the construction of P_n^* so that $R_2(P_n^*, P_0)$ behaves as a third-order term. Our approach will be to expand

$$R_2(P_n^*, P_0) = -\frac{1}{2} P_0^2 D^{(2)}(P_n^*) + R_3(P_n^*, P_0) \ , \tag{2.4}$$

where for a function f of a pair (o_1, o_2) defined on $\mathcal{O} \times \mathcal{O}$, with O_1 and O_2 independent variates distributed according to P_0, we define $P_0^2 f := \int f(o_1, o_2) dP_0(o_1) dP_0(o_2)$ as the expectation of f with respect to the product measure P_0^2 of (O_1, O_2), and $R_3(P_n^*, P_0)$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

is a third-order remainder term. The expansion in (2.4) is generally achieved by a so-called second-order canonical gradient $D^{(2)*}(P)$ but possibly also by other functions. In conjunction with (2.3), it yields the identity

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^{(1)*}(P_n^*) - \frac{1}{2}P_0^2 D^{(2)}(P_n^*) + R_3(P_n^*, P_0) \ .$$

Given this function $D^{(2)}(P)$, we arrange that the TMLE not only solves $P_n D^{(1)*}(P_n^*) = 0$ but also the U-statistic equation

$$0 = P_n^2 D^{(2)}(P_n^*) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D^{(2)}(P_n^*)(O_i, O_j)$$

by including additional parameters in the least-favorable parametric submodel. This allows us to obtain that

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^{(1)*}(P_n^*) + \frac{1}{2}(P_n^2 - P_0^2)D^{(2)}(P_n^*) + R_3(P_n^*, P_0) \ .$$

Lemmas 2 and 3 provide conditions under which $(P_n^2 - P_0^2)D^{(2)}(P_n^*)$ converges in probability to zero at a rate faster than $n^{-1/2}$. Asymptotic efficiency is then established as above but with the condition $R_2(P_n^*, P_0) = o_P(n^{-1/2})$ replaced by $R_3(P_n^*, P_0) = o_P(n^{-1/2})$.

Our proposed 2-TMLE is based on a remarkably simple observation: if we denote $\bar{D}_n^{(2)}(P_n^*)(O_i) := \frac{1}{n}\sum_{j=1}^{n} D^{(2)}(P_n^*)(O_i, O_j)$, then $\bar{D}_n^{(2)}(P_n^*)$ can itself be perceived as a score at P_n^* and furthermore we can write

$$P_n^2 D^{(2)}(P_n^*) = \frac{1}{n}\sum_{i=1}^{n} \bar{D}_n^{(2)}(P_n^*)(O_i) = P_n \bar{D}_n^{(2)}(P_n^*) \ .$$

As a consequence, we may simply augment our least-favorable parametric submodel $\{P(\epsilon_1) : \epsilon_1\}$ with score $D^{(1)*}(P)$ at $\epsilon_1 = 0$ in the 1-TMLE to also include a parameter ϵ_2 such that the expanded least-favorable submodel $\{P(\epsilon_1, \epsilon_2) : \epsilon_1, \epsilon_2\}$ also generates the score $\bar{D}_n^{(2)}(P)$ at $(\epsilon_1, \epsilon_2) = (0, 0)$. In this manner, the resulting 2-TMLE P_n^* solves the collection of equations

$$\begin{aligned} 0 &= P_n D^{(1)*}(P_n^*) \\ 0 &= P_n \bar{D}^{(2)}(P_n^*) \ = \ P_n^2 D^{(2)}(P_n^*) \ . \end{aligned}$$

Thus, the proposed 2-TMLE is no more than a usual 1-TMLE with least-favorable submodel extended in a particular manner.

2.4 Insufficiently differentiable target parameters

As has been noted in Robins et al. [2009], most statistical parameters of interest are unfortunately not smooth enough as functions of the data-generating distribution to

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

allow the required expansion for $R_2(P_n^*, P_0)$. These second-order remainder terms often involve squared differences between a density estimator and the true density. As a consequence, while $R_2(P_n^*, P_0)$ can often itself be expanded into second-order differences between P_n^* and P_0, these second-order terms cannot be represented as the expectation under P_0 of a second-order gradient $D^{(2)}(P_n^*)$. This is a significant hurdle for any method aiming to carry out higher-order bias reduction; in particular, it is a challenge we must face when constructing a higher-order TMLE.

In such cases, we will search for a surrogate $D_h^{(2)}$, indexed by a smoothing parameter h, such that

$$R_2(P_n^*, P_0) = -\frac{1}{2}\lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*) + R_3(P_n^*, P_0)$$

for some third-order remainder $R_3(P_n^*, P_0)$, thus leading to the representation

$$R_2(P_n^*, P_0) = -\frac{1}{2}P_0^2 D_h^{(2)}(P_n^*) + B_n(h) + R_3(P_n^*, P_0)\ .$$

Here, the representation error $B_n(h) := [P_0^2 D_h^{(2)}(P_n^*) - \lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*)]/2$ quantifies the bias resulting from this approximation of a second-order gradient. We may therefore write

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^{(1)*}(P_n^*) - \frac{1}{2}P_0^2 D_h^{(2)}(P_n^*) + B_n(h) + R_3(P_n^*, P_0)\ ,$$

which is analogous to (2.3). In order to preserve the validity of our general proof of asymptotic efficiency of the 2-TMLE above, we will require that $B_n(h)$ tend to zero faster than $n^{-1/2}$ for an appropriately chosen tuning parameter $h = h_n$ under appropriate smoothness conditions for P_0. The representation error can also be recentered as

$$B_n(h) = \frac{1}{2}\left\{P_0^2\left[D_h^{(2)}(P_n^*) - D_h^{(2)}(P_0)\right] - \lim_{h\to 0} P_0^2\left[D_h^{(2)}(P_n^*) - D_h^{(2)}(P_0)\right]\right\};$$

this form may facilitate its theoretical study. In our running example, provided higher-order kernels are utilized and the density of the covariate vector W is sufficiently smooth, $B_n(h)$ can be bounded by terms of the type $h^{m+1}\|Q_n^* - Q_0\|$, where m represents both the degree of smoothness of the underlying object and the degree of the kernel, and $\|\cdot\|$ is some norm. Thus, if $h = h_n$ tends to zero fast enough, the representation error will be $o_P(n^{-1/2})$.

Similarly as before, the construction of our proposed 2-TMLE guarantees that $P_n D^{(1)*}(P_n^*) = P_n^2 D_{h_n}^{(2)}(P_n^*) = 0$ for some value h_n of the tuning parameter. That these two score equations are solved by P_n^* follows from the use of a least-favorable submodel whose score at $\epsilon = 0$ spans both $D^{(1)*}(P)$ and $\bar{D}_{h_n}^{(2)}(P)$. In practice, selection of the tuning parameter h_n requires great care. On one hand, to ensure that $(P_n^2 - P_0^2)D_{h_n}^{(2)}(P_n^*)$ converges to zero in probability faster than $n^{-1/2}$, h_n must generally tend to zero slowly enough. On the other hand, for the representation error to be asymptotically negligible, it is required that h_n tend to zero quickly enough. This

COBRA
Collection of Biostatistics
Research Archive

constitutes the primary theoretical challenge this 2-TMLE must contend with as a result of the target parameter failing to be second-order pathwise differentiable: namely, selection of h_n necessarily involves a careful balance to ensure that

$$\frac{1}{2}(P_n^2 - P_0^2)D^{(2)}_{h_n}(P_n^*) + B_n(h_n) = o_P(n^{-1/2}) \ .$$

This involves a sensible trade-off between control of the U-statistic term and the representation error.

Of course, this theoretical challenge translates directly into a fundamental practical challenge. Indeed, one algorithm within a large collection of candidate 2-TMLE algorithms, each indexed by the corresponding choice of h, must be chosen in practice. Optimal rates for h_n can often be derived in particular applications, but these are primarily of theoretical interest and generally provide little or no practical guidance regarding the selection of h_n. Fortunately, this is precisely the kind of challenge TMLE can very naturally handle, notably by adjudicating the quality of a particular tuning value h based on the gain in fit resulting from the ensuing parametric TMLE updating step. This is accomplished formally within the C-TMLE framework previously referenced and discussed later.

3 Analytic basis for higher-order TMLE

3.1 Second-order differentiability

Let $T(P)$ be the tangent space of $\mathcal{M}$ at $P \in \mathcal{M}$, defined as the closure of the linear span of all scores of regular parametric submodels through P. This is a subspace of the Hilbert space $L_0^2(P)$ of square-integrable real-valued functions defined on the support of P, with mean zero under P, endowed with inner product $\langle h_1, h_2\rangle_P := P(h_1 h_2)$. The norm $h \mapsto \langle h, h\rangle_P^{1/2}$ will be denoted by $\|\cdot\|_{2,P}$. Let $D^{(1)*}(P) \in T(P)$ be the first-order canonical gradient of the parameter $\Psi : \mathcal{M} \to \mathbb{R}$ at P. Denote by $L_0^{2*}(P^2)$ the Hilbert space of square-integrable real-valued functions defined on the support of P^2, symmetric in its two arguments, and satisfying that

$$\int f(o_1, o)dP(o_1) = \int f(o, o_2)dP(o_2) = 0$$

for P-almost every o, equipped with the inner product

$$\langle f_1, f_2\rangle_{P^2} := P^2(f_1 f_2) = \int f_1(o_1, o_2) f_2(o_1, o_2) dP(o_1)dP(o_2) \ .$$

The norm $f \mapsto \langle f, f\rangle_{P^2}^{1/2}$ will be denoted by $\|\cdot\|_{2,P^2}$. If, for a given $P \in \mathcal{M}$ and each sufficiently smooth one-dimensional submodel $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$ through P and with first-order score $s^{(1)}(P)(o) := \frac{\partial}{\partial\epsilon} p_\epsilon(o)\big|_{\epsilon=0} / p(o)$ at $\epsilon = 0$, the representation

$$\Psi(P(\epsilon)) - \Psi(P) = \epsilon \int D^{(1)}(P)(o) s^{(1)}(o) dP(o) + o(\epsilon) \qquad (3.1)$$

Hosted by The Berkeley Electronic Press

holds for some element $D^{(1)}(P) \in L_0^2(P)$, we say that Ψ is first-order pathwise differentiable at P and $D^{(1)}(P)$ is a first-order gradient of Ψ at P (see, e.g., Pfanzagl [1982]). Defining

$$s^{(2)}(P)(o) := \left.\frac{\partial^2}{\partial \epsilon^2} p_\epsilon(o)\right|_{\epsilon=0} [p(o)]^{-1}$$

as the second-order score of ϵ in $\{P(\epsilon) : \epsilon\}$ at $\epsilon = 0$, and setting

$$\begin{aligned} A_1(s^{(1)})(P) &:= \int D^{(1)}(P)(o) s^{(1)}(o) dP(o) \\ A_2(s^{(1)}, s^{(2)})(P) &:= \int D^{(1)}(P)(o) s^{(2)}(o) dP(o) \\ &\qquad + \iint D^{(2)}(P)(o_1, o_2) s^{(1)}(o_1) s^{(1)}(o_2) dP(o_1) dP(o_2) \ , \end{aligned}$$

if the representation

$$\Psi(P(\epsilon)) - \Psi(P) = \epsilon A_1(s^{(1)})(P) + \frac{1}{2}\epsilon^2 A_2(s^{(1)}, s^{(2)})(P) + o(\epsilon^2) \tag{3.2}$$

holds for each first-order gradient $D^{(1)}(P) \in L_0^2(P)$ and an element $D^{(2)}(P) \in L_0^{2*}(P^2)$, we say that Ψ is second-order pathwise differentiable at P and $D^{(2)}(P)$ is a second-order gradient of Ψ at P (see, e.g., Pfanzagl [1985]). This definition provides a practical means of identifying candidate second-order gradients. Specifically, given a smooth one-dimensional parametric model $\{P(\epsilon) : \epsilon\} \subset \mathcal{M}$ with first- and second-order scores $s^{(1)}$ and $s^{(2)}$, respectively, any function $D^{(2)}(P) \in L_0^{2*}(P^2)$ such that

$$\iint D^{(2)}(P)(o_1, o_2) s^{(1)}(o_1) s^{(2)}(o_2) dP(o_1) dP(o_2) = \left.\frac{d^2}{d\epsilon^2}\Psi(P(\epsilon))\right|_{\epsilon=0} - P\left[D^{(1)}(P) s^{(2)}\right]$$

will be a candidate second-order gradient of Ψ at P in model $\mathcal{M}$. In view of Remark 4.4.2 of Pfanzagl [1985], given a first-order gradient $D^{(1)}(P)$ of Ψ at $P \in \mathcal{M}$, a corresponding second-order gradient can also be obtained by computing pointwise a first-order gradient of the parameter mapping $P \mapsto D^{(1)}(P)(o)$ for fixed o.

A second-order gradient often has the form

$$D^{(2)}(P)(o_1, o_2) = \int S_{1,x}(o_1) S_{2,x}(o_2) h(P)(x) d\nu(x) \tag{3.3}$$

for some $S_{1,x}$ and $S_{2,x}$ in $L_0^2(P)$ for each x, a function $x \mapsto h(P)(x)$ and a measure ν such that resulting function is symmetric in its two arguments. To see this, suppose that $\Psi(P)$ depends on P through a summary vector $(Pf : f \in \mathcal{F})$ for some finite class of functions $\mathcal{F}$. This is necessarily the case if $\mathcal{O}$ is finite, for example, as then taking $\mathcal{F} := \{u \mapsto I(u = o) : o \in \mathcal{O}\}$ will certainly do. A standard second-order Taylor

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

expansion will give that

$$
\begin{aligned}
\Psi(P(\epsilon)) - \Psi(P) \ &= \ \sum_{f\in\mathcal{F}} \frac{\partial\Psi(P)}{\partial Pf} \cdot (P(\epsilon) - P)f \\
&\quad + \frac{1}{2} \sum_{f_1, f_2 \in \mathcal{F}} \frac{\partial^2 \Psi(P)}{\partial P f_1 \partial P f_2} \cdot (P(\epsilon) - P) f_1 (P(\epsilon) - P) f_2 + o(\epsilon^2) \ .
\end{aligned}
$$

Using the representation $dP(\epsilon)/dP = 1 + \epsilon s^{(1)} + \frac{1}{2}\epsilon^2 s^{(2)} + o(\epsilon^2)$, we can write

$$
\begin{aligned}
(P(\epsilon) - P)f \ &= \ \epsilon \int (f(o) - Pf)\, s^{(1)}(o) dP(o) \\
&\quad + \frac{1}{2}\epsilon^2 \int (f(o) - Pf) s^{(2)}(o) dP(o) + o(\epsilon^2) \ , \\
(P(\epsilon) - P) f_1 (P(\epsilon) - P) f_2 \ &= \\
&\epsilon^2 \iint (f_1(o_1) - Pf_1)(f_2(o_2) - Pf_2) s^{(1)}(o_1) s^{(1)}(o_2) dP(o_1) dP(o_2) + o(\epsilon^2) \ .
\end{aligned}
$$

This motivates us to define

$$
\begin{aligned}
D^{(1),NP}(P)(o) \ &:= \ \int_{f\in\mathcal{F}} \frac{\partial\Psi(P)}{\partial Pf} \, (f(o) - Pf) \, d\mu_1(f) \\
D^{(2),NP}(P)(o_1, o_2) \ &:= \ \int_{f_1, f_2 \in \mathcal{F}} \frac{\partial^2\Psi(P)}{\partial P f_1 \partial P f_2} \, (f_1(o_1) - Pf_1) \, (f_2(o_2) - Pf_2) \, d\mu_2(f_1, f_2)
\end{aligned}
$$

with μ_1 and μ_2 the counting measures on $\mathcal{F}$ and $\mathcal{F} \times \mathcal{F}$, respectively. The heuristic derivation above yields (3.2) with $D^{(1)}(P) = D^{(1),NP}(P)$ and $D^{(2)}(P) = D^{(2),NP}(P)$. In the general case where $\mathcal{F}$ is not finite, this representation will generally still hold with other measures μ_1 and μ_2. The form of these measures will often become apparent upon taking increasingly fine discrete approximations to $\mathcal{F}$.

Representation (3.2) can be alternatively written as

$$
\begin{aligned}
\Psi(P(\epsilon)) - \Psi(P) \ &= \ (P(\epsilon) - P) D^{(1)}(P) + \frac{1}{2}(P(\epsilon) - P)^2 D^{(2)}(P) + o(\epsilon^2) \\
&= \ P(\epsilon) D^{(1)}(P) + \frac{1}{2} P(\epsilon)^2 D^{(2)}(P) + o(\epsilon^2) \ .
\end{aligned}
$$

If at each $P \in \mathcal{M}$ such second-order expansion holds uniformly along all parametric submodels through P, strong differentiability, as defined in Pfanzagl [1985], is expected to hold and then

$$
\Psi(P) - \Psi(P_0) = -P_0 D^{(1)}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0) \ , \tag{3.4}
$$

where $R_3(P, P_0)$ involves a third-order difference between P and P_0. This expansion serves as the basis for constructing a 2-TMLE in this paper, as is discussed below.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

3.2 Second-order canonical gradients

The expansions above hold for any *valid* pair of first- and second-order gradients. Here, we emphasize valid because for a given first-order gradient the set of second-order gradients for which expansion (3.2) holds is generally smaller than the set of all second-order gradients. This point was recognized by Pfanzagl [1985] and discussed in Robins et al. [2009]. Under regularity conditions, a 2-TMLE constructed using these expansions will be asymptotically linear with influence function $D^{(1)}(P_0)$. As such, semiparametric efficiency theory motivates us to select the first-order canonical gradient $D^{(1)*}(P)$ as first-order gradient. We therefore make this choice from hereon. The canonical gradient can be obtained by projecting any first-order gradient onto the first-order tangent space $T(P)$; in particular, $D^{(1)*}(P)$ is the projection of $D^{(1),NP}(P)$ onto $T(P)$. Of course, if $\mathcal{M}$ is locally saturated at P so that $T(P) = L_0^2(P)$, it follows that $D^{(1),NP}(P) = D^{(1)*}(P)$.

The notion of a second-order tangent space, particularly in the context of estimation within infinite-dimensional models, is studied very carefully in Pfanzagl [1985], Robins et al. [2009] and van der Vaart [forthcoming]. We encourage interested readers to consult these references. Representation (3.2) motivates defining the second-order tangent space at P as the closure of the linear span of cross-products of the form $(o_1, o_2) \mapsto s^{(1)}(o_1)s^{(1)}(o_2)$, where $s^{(1)}$ is a first-order score for ϵ at $\epsilon = 0$ in a one-dimensional smooth parametric submodel $\{P(\epsilon) : \epsilon\}$ such that $P(0) = P$. The second-order canonical gradient can then be defined as the projection of any second-order gradient onto the second-order tangent space. Provided an algorithm for projecting any element of $L_0^2(P)$ onto $T(P)$ is available, the latter projection can be easily computed, as the following lemma describes. This result can be alternatively obtained as a corollary of Remark 2.5.22 of Pfanzagl [1985].

Lemma 1. *Suppose that a second-order gradient $D^{(2)}(P)$ is available and defined pointwise as $D^{(2)}(P)(o_1, o_2) = \int S_{1,x}(P)(o_1)S_{2,x}(P)(o_2)h(P)(x)d\nu(x)$ for some elements $S_{1,x}(P)$ and $S_{2,x}(P)$ in $L_0^2(P)$ for each x, a function $x \mapsto h(P)(x)$ and a measure ν such that the resulting function is symmetric in its two arguments. Denoting by $S_{1,x}^*(P)$ and $S_{2,x}^*(P)$ the projection of $S_{1,x}(P)$ and $S_{2,x}(P)$, respectively, onto $T(P)$, the mapping*

$$(o_1, o_2) \mapsto D^{(2)*}(P)(o_1, o_2) := \int S_{1,x}^*(P)(o_1)S_{2,x}^*(P)(o_2)h(P)(x)d\nu(x)$$

is the second-order canonical gradient of Ψ at P relative to model $\mathcal{M}$.

For the sake of constructing a 2-TMLE, we require an element $D^{(2)}(P) \in L_0^2(P^2)$ satisfying two critical conditions:

(A) that expansion (3.4) hold, and

(B) that either $o_1 \mapsto D^{(2)}(P)(o_1, o)$ or $o_2 \mapsto D^{(2)}(P)(o, o_2)$ lie in $T(P)$ for each $o \in \mathcal{O}$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

We refer to any such element as a second-order partial canonical gradient. Let $D^{(2)}(P)$ satisfy these requirements, and suppose, for definiteness, that $o_1 \mapsto D^{(2)}(P)(o_1, o)$ is in $T(P)$ for each $o \in \mathcal{O}$. It then follows that

$$o \mapsto \bar{D}_n^{(2)}(P)(o) := \frac{1}{n}\sum_{j=1}^{n} D^{(2)}(P)(o, O_j) \in T(P)$$

and consequently, that the V-statistic

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D^{(2)}(P)(O_i, O_j) = \frac{1}{n}\sum_{i=1}^{n} \bar{D}_n^{(2)}(P)(O_i)$$

is the empirical mean of a score at P. As a consequence, a TMLE solving both $P_n D^{(1)*}(P_n^*) = 0$ and the V-statistic equation $P_n^2 D^{(2)}(P_n^*) = 0$ can be constructed by using a parametric submodel through P with score vector components spanning $D^{(1)*}(P)$ and $\bar{D}_n^{(2)}(P)$. Such a submodel can be constructed precisely because the latter are in $T(P)$. Thus, if a second-order partial canonical gradient $D^{(2)}(P)$ is available, a 2-TMLE can be devised in the same fashion as a usual 1-TMLE. It is important to note that this construction would not be possible with just any second-order gradient satisfying (3.4): for arranging a TMLE construction, it is critical that one of its coordinate projections uniformly lie in the tangent space $T(P)$. This property generally applies to second-order canonical gradients, which can therefore typically be relied upon for defining a proper 2-TMLE.

Denoting by f° the symmetrization

$$(o_1, o_2) \mapsto f^\circ(o_1, o_2) := \frac{f(o_1, o_2) + f(o_2, o_1)}{2}$$

of a given function $(o_1, o_2) \mapsto f(o_1, o_2) \in \mathbb{R}$, we note that $P^2 f = P^2 f^\circ$ for any measure $P \in \mathcal{M}$. Since in the construction of a 2-TMLE the second-order gradient $D^{(2)}(P)$ appears only via empirical moments of the form $P_n^2 D^{(2)}(P)$ for $P \in \mathcal{M}$, any element $D_+^{(2)}(P) \in L_0^2(P^2)$ such that $P_n^2 D_+^{(2)}(P) = P_n^2 D^{(2)}(P)$ for each $P \in \mathcal{M}$ could be used in practice. In view of this observation and to simplify our presentation, hereafter we allow abuse of definition and refer as second-order gradient or canonical gradient any element of $L_0^2(P^2)$ whose symmetrization is a second-order gradient or canonical gradient of Ψ at $P \in \mathcal{M}$.

3.3 Example: estimating a counterfactual mean

Our motivating example regards the estimation of $\psi_0 := \Psi(P_0)$ using n independent copies of $O = (W, A, Y) \sim P_0$, where $\Psi(P) = E_P E_P(Y \mid A = 1, W)$ is the counterfactual mean under certain causal assumptions. The first-order canonical gradient is known in this case to be

$$D^{(1)*}(P)(o) := \frac{a}{\bar{g}(w)}\left[y - \bar{Q}(w)\right] + \bar{Q}(w) - \Psi(P)\ ,$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

where $o := (w, a, y)$ is a possible realization of O, $\bar{g}(w) := P(A = 1 \mid W = w)$, $\bar{Q}(w) := E_P(Y \mid A = 1, W = w)$ and $Q_W(w) := P(W = w)$. With this definition, we find that $\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) + R_2(P, P_0)$ with

$$R_2(P, P_0) := P_0 \left[\left(\frac{\bar{g} - \bar{g}_0}{\bar{g}} \right) (\bar{Q} - \bar{Q}_0) \right],$$

where $\bar{g}_0$ and $\bar{Q}_0$ denote $\bar{g}$ and $\bar{Q}$, respectively, under P_0. Our goal is to find a representation $R_2(P, P_0) = -\frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0)$. For this purpose, we assume W has a finite support; this will be relaxed in later sections. It can be shown that the second-order canonical gradient at (o_1, o_2), with $o_1 := (w_1, a_1, y_1)$ and $o_2 := (w_2, a_2, y_2)$, is given by (the symmetrization of – see note above)

$$D^{(2)}(P)(o_1, o_2) := H(P)(w_1, a_1, w_2, a_2)[y_1 - \bar{Q}(w_1)] \ , \tag{3.5}$$

where

$$H(P)(w_1, a_1, w_2, a_2) := \frac{2 a_1 I(w_1 = w_2)}{\bar{g}(w_1) Q_W(w_1)} \left[1 - \frac{a_2}{\bar{g}(w_1)} \right].$$

It is not difficult to directly verify that indeed

$$\frac{1}{2} P_0^2 D^{(2)}(P) = -P_0 \left[\left(\frac{\bar{g} - \bar{g}_0}{\bar{g}} \right) (\bar{Q} - \bar{Q}_0) \right] + R_3(P, P_0) \ , \tag{3.6}$$

where $R_3(P, P_0)$ is given by

$$P_0 \left[\left(1 - \frac{\bar{g}_0 Q_{W,0}}{\bar{g} Q_W} \right) \left(\frac{\bar{g} - \bar{g}_0}{\bar{g}} \right) (\bar{Q} - \bar{Q}_0) \right].$$

Thus, it holds that $\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \frac{1}{2} P_0 D^{(2)}(P) + R_3(P, P_0)$ for a third-order term $R_3(P, P_0)$, as desired. To ensure $D^{(2)}(P)$ is not degenerate, W must finitely-supported so that the event $\{W_1 = W_2\}$ has positive probability.

4 Inference using higher-order TMLE

4.1 Asymptotic linearity and efficiency

We assume, in this section, that the target parameter is second-order pathwise differentiable and that (3.4) holds for some second-order partial canonical gradient $D^{(2)}(P)$. Under prescribed conditions, a 2-TMLE will be asymptotically linear and efficient irrespective of the particular second-order partial canonical gradient selected. A precise enumeration of these conditions are given in the below theorem.

Theorem 1. *Suppose that the target parameter Ψ admits the second-order expansion $\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0)$, and that $P_n^* \in \mathcal{M}$ satisfies the equations*

$$P_n D^{(1)*}(P_n^*) = 0 \quad \textit{and} \quad P_n^2 D^{(2)}(P_n^*) = P_n \bar{D}_n^{(2)}(P_n^*) = 0$$

with $\bar{D}_n^{(2)}(P_n^*)(o) := \frac{1}{n} \sum_{i=1}^n D^{(2)}(P_n^*)(o, O_i)$. Then, provided that

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

1. *there exists a P_0-Donsker class $\mathcal{F}$ such that $D^{(1)*}(P_n^*)$ is in $\mathcal{F}$ with probability tending to one, and $P_0\left[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)\right]^2 = o_P(1)$;*
2. *$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = o_P(n^{-1/2})$;*
3. *$R_3(P_n^*, P_0) = o_P(n^{-1/2})$;*

ψ_n^* is an asymptotically linear estimator of ψ_0 with influence function $D^{(1)*}(P_0)$. It is thus also asymptotically efficient.

Verification of condition 2 in this theorem may seem particularly daunting. The following lemma provides a set of conditions which suffice to establish condition 2 and may be easier to verify in practice.

Lemma 2. *Provided it can be established that*

(i) there exists a P_0-Donsker class $\mathcal{G}$ such that $o \mapsto \int D^{(2)}(P_n^*)(o, o_2)dP_0(o_2)$ and $o \mapsto \int D^{(2)}(P_n^*)(o_1, o)dP_0(o_1)$ are in $\mathcal{G}$ with probability tending to one;

(ii) the integrals defined as $J_{n,1}^* := \int \left[\int D^{(2)}(P_n^*)(o_1, o_2)dP_0(o_1)\right]^2 dP_0(o_2)$ and $J_{n,2}^* := \int \left[\int D^{(2)}(P_n^*)(o_1, o_2)dP_0(o_2)\right]^2 dP_0(o_1)$ both tend to zero in probability;

(iii) $(P_n - P_0)^2 D^{(2)}(P_n^*) = o_P(n^{-1/2})$,

then condition 2 of Theorem 1 holds.

The following lemma, as presented in Gill et al. [1995] and using the concept of *uniform sectional variation norm*, is particularly useful to verify (iii). For a given function $f : \mathbb{R}^d \to \mathbb{R}$, the uniform sectional variation norm $\|f\|_v^*$ of f is defined as $\|f\|_v^* := \sup_s \sup_{x_s} \int |f(dx_s, x_{-s})|$, where the supremum is over all possible sections of f and $\int |f(dx_s, x_{-s})|$ represents the variation norm of the section $x_s \mapsto f(x_s, x_{-s})$. Here, the latter section is defined by a given subset $s \subseteq \{1, \ldots, d\}$, $x_s := (x_j : j \in s)$ and $x_{-s} := (x_j : j \notin s)$.

Lemma 3. *Suppose that $\mathcal{O} \subseteq \mathbb{R}^m$ for some integer m. Provided $D^{(2)}(P_n^*)$ is right-continuous with left-hand limits, there exists a real number $C < \infty$ such that*

$$(P_n - P_0)^2 D^{(2)}(P_n^*) \ \leq \ C\|F_n - F_0\|_\infty^2 \|D^{(2)}(P_n^*)\|_v^* \ ,$$

where F_n and F_0 are the distributions functions associated to P_n and P_0, respectively.

Most importantly, as a consequence of this lemma and in view of Donsker's Theorem, we have that $(P_n - P_0)^2 D^{(2)}(P_n^*) = O_P(n^{-1}\|D^{(2)}(P_n^*)\|_v^*)$.

The theorem provided above can be readily generalized to arbitrarily higher orders based on the existence of a higher-order expansion

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \sum_{j=2}^{k} \frac{1}{j!} P_0^j D^{(j)}(P) + R_{k+1}(P, P_0)$$

Hosted by The Berkeley Electronic Press

with appropriate higher-order partial canonical gradients $D^{(j)}$ for $j = 1, 2, \ldots, k$, and an estimator $P_n^* \in \mathcal{M}$ such that $P_n \bar{D}^{(j)}(P_n^*) = 0$ for each $j = 1, 2, \ldots, k$, where

$$\bar{D}_n^{(j)}(P)(o) := \frac{1}{n^{j-1}} \sum_{\ell_1, \ell_2, \ldots, \ell_{j-1}} D^{(j)}(P)(o, O_{\ell_1}, O_{\ell_2}, \ldots, O_{\ell_{j-1}})$$

and $\bar{D}_n^{(j)}(P)$ is in $T(P)$. In this generalized context, condition 1 of Theorem 1 remains intact, while condition 2 requires additionally that $(P_n^j - P_0^j)D^{(j)}(P_n^*) = o_P(n^{-1/2})$ for each $j = 3, 4, \ldots, k$, and condition 3 is relaxed to $R_{k+1}(P_n^*, P_0) = o_P(n^{-1/2})$.

4.2 Constructing confidence intervals

The same techniques for constructing confidence intervals based on the usual TMLE can be utilized in the context of a higher-order TMLE. Since a k-TMLE ψ_n^* is asymptotically linear with influence function $D^{(1)*}(P_0)$ irrespective of k, it follows that $n^{1/2}(\psi_n^* - \psi_0)$ converges in law to a normal variate with mean zero and variance $\sigma_0^2 := P_0[D^{(1)*}(P_0)]^2$. This suggests that the Wald-type interval

$$\left(\psi_n^* - z_{1-\alpha/2}\sigma_n n^{-1/2},\ \psi_n^* + z_{1-\alpha/2}\sigma_n n^{-1/2}\right), \tag{4.1}$$

where $\sigma_n^2 := P_n[D^{(1)*}(P_n^*)]^2$ and z_β is the β-quantile of the standard normal distribution, has asymptotic coverage level $(1 - \alpha)$. TMLE procedures of different order exhibit identical first-order behavior, although inclusion of higher-order terms will generally guarantee such behavior holds under a wider range of scenarios. The interval (4.1) can therefore be utilized with any higher-order targeted estimator P_n^* of P_0.

The above approach provides asymptotically correct inference. However, it does not explicitly utilize the higher-order expansion upon which a k-TMLE is constructed. It is plausible that confidence intervals with improved finite-sample performance may be obtained by incorporating higher-order terms from this expansion. For this purpose, a simple bootstrap approach can be devised based on the fact that, provided a random sample $O_1^{\#}, \ldots, O_n^{\#} \sim P_n^\circ$ for some consistent estimator P_n° of P_0, the conditional distribution of the bootstrapped statistic

$$\begin{aligned} Z_n^{\#} &:= n^{-1/2} \sum_{i=1}^{n} \left[D^{(1)*}(P_n^*)(O_i^{\#}) - P_n D^{(1)*}(P_n^*)\right] \\ &\qquad + \frac{n^{-3/2}}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[D^{(2)}(P_n^*)(O_i^{\#}, O_j^{\#}) - P_n^2 D^{(2)}(P_n^*)\right] \end{aligned}$$

given P_n and the distribution of $n^{1/2}(\psi_n - \psi_0)$ approximate each other arbitrarily well for large n and for almost every P_n. Obvious choices for P_n° include, for example, the empirical measure P_n and the targeted estimator P_n^*. This suggests the confidence interval

$$\left(\psi_n^* - q_{1-\alpha/2,n} n^{-1/2},\ \psi_n^* - q_{\alpha/2,n} n^{-1/2}\right)$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

with $q_{\beta,n}$ the β-quantile of the conditional distribution of $Z_n^{\#}$ given P_n. Of course, the involved quantiles can be estimated arbitrarily well by simulation. This confidence interval is easy to implement and takes into account the second-order variability explained by the U-statistic process implied by the second-order partial canonical gradient $D^{(2)}(P)$.

4.3 Implementing a higher-order TMLE

The practical implementation of a higher-order TMLE is no more difficult than a regular TMLE since the former can indeed be seen as an example of the latter. The additional effort involved, in reality, lies in the computation of higher-order partial canonical gradients, or of suitable approximations to such if need be, as discussed in the next section. Below, we describe the implementation of a second-order targeted minimum loss-based estimator.

Given independent variates $O_1, O_2, \ldots, O_n$ distributed according to $P_0 \in \mathcal{M}$, we wish to estimate $\psi_0 := \Psi(P_0)$ for a given target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ of interest. Suppose that the parameter $P \mapsto Q(P)$ satisfies that

i) $\Psi = \Psi_1 \circ Q$ for some mapping $\Psi_1 : Q(\mathcal{M}) \to \mathbb{R}$, and

ii) $Q_0 = \text{argmin}_{Q \in Q(\mathcal{M})} P_0 L(Q)$ for some loss function $(o, Q) \mapsto L(Q)(o)$.

Here, Q_0 represents the true summary $Q(P_0)$. With slight abuse of notation, for the sake of notational convenience, we take $\Psi(Q)$ to mean $\Psi_1(Q)$ hereafter. Several parametrizations may be possible and it will generally be preferable to choose the least complex such parametrization for which we can find an appropriate loss function and parametric fluctuation submodels with scores at Q spanning the appropriate empirical gradients, as described below.

Let $g = g(P)$ be a nuisance parameter and write $g_0 := g(P_0)$. Suppose that the first-order canonical gradient $D^{(1)*}(P)$ of Ψ can be represented as $D^{(1)*}(Q(P), g(P))$ and that $D^{(2)}(P) := D^{(2)}(Q(P), g(P))$ is any associated second-order partial canonical gradient of Ψ. Similar abuse of notation is tolerated here as well. Suppose that for each $(o_1, o_2) \in \mathcal{O} \times \mathcal{O}$ we can write $D^{(2)}(P)(o_1, o_2) = \int_x S_x^*(P)(o_1) f_x(P)(o_2) d\nu(x)$ for $S_x^*(P) \in T(P)$ for each x, for some arbitrary function $(x, o) \mapsto f_x(P)(o)$ and for some measure ν. Then, setting

$$\bar{D}_n^{(2)}(Q, g)(o) := \int_x S_x^*(P)(o) \frac{1}{n} \sum_{j=1}^{n} f_x(P)(O_j) d\nu(x) \ ,$$

we have that $\bar{D}_n^{(2)}(Q, g) \in T(P)$ and will play the role of a second-order score at P.

Let $\mathcal{Q}(Q, g) := \{Q_g(\epsilon) : \epsilon\} \subset Q(\mathcal{M})$ be a second-order least-favorable submodel, in the sense that $Q_g(0) = Q$ and both $D^{(1)*}(Q, g)$ and $\bar{D}_n^{(2)}(Q, g)$ lie in the closure of the linear span

$$\left\{ z^T \frac{\partial}{\partial \epsilon} L(Q_g(\epsilon)) \bigg|_{\epsilon=0} : z \in \mathbb{R}^p \right\}$$

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

of the generalized score vector at $\epsilon = 0$, where p denotes the dimension of ϵ.

Suppose that an initial estimator $(Q_{n,0}, g_n)$ of (Q_0, g_0) is at our disposal. As with a regular TMLE, a 2-TMLE will generally be an iterative procedure, though analytic convergence after a single step can be demonstrated in important examples. A 2-TMLE updating step will be identical to that of a regular TMLE, except for the use of a second-order rather than first-order least-favorable submodel. Specifically, given the current estimate $Q_{n,m}$ of Q_0, the updated estimate $Q_{n,m+1}$ is defined as

$$Q_{n,m+1} := \underset{Q \in \mathcal{Q}(Q_{n,m}, g_n)}{\text{argmin}} \; P_n L(Q) \; .$$

Alternatively, setting $\epsilon_n^m := \text{argmin}_\epsilon P_n L(Q_{n,m}(\epsilon))$, we can express the updated estimate of Q_0 as $Q_{n,m+1} := Q_{n,m,g_n}(\epsilon_n^m)$. This iterative updating step will generally be repeated until convergence, adjudicated by ϵ_n^m being sufficiently close to zero. Denoting by Q_n^* the limit of this iterative procedure, the 2-TMLE of ψ_0 is given by $\psi_n^* := \Psi(Q_n^*)$. The desirable asymptotic properties of ψ_n^* are in large part a consequence of the fact that

$$P_n D^{(1)*}(Q_n^*, g_n) = P_n \bar{D}_n^{(2)}(Q_n^*, g_n) = 0$$

by construction.

The implementation of a k-TMLE proceeds almost identically, except for the use of an appropriate higher-order least-favorable submodel. The construction of such a submodel requires the computation of higher-order partial canonical gradients – the effort involved is therefore primarily analytic. Denoting by $D^{(j)}(Q, g)$ an available j^{th} order partial canonical gradient, suppose

$$D^{(j)}(P)(o_1, o_2, \ldots, o_j) = \int_x S_x^*(P)(o_1) f_x(P)(o_2, o_3, \ldots, o_j) d\nu(x)$$

for some $S_x^*(P) \in T(P)$ for each x, a function $(x, o_2, o_3, \ldots, o_j) \mapsto f_x(P)(o_2, o_3, \ldots, o_j)$ and a measure ν. Setting

$$\bar{D}_n^{(j)}(Q, g)(o) := \int_x S_x^*(P)(o) \frac{1}{n^{j-1}} \sum_{\ell_1, \ldots, \ell_{j-1}}^{n} f_x(P)(O_{\ell_1}, O_{\ell_2}, \ldots, O_{\ell_{j-1}}) d\nu(x) \; ,$$

we have that $\bar{D}_n^{(j)}(Q, g)$ lies in $T(P)$ and now plays the role of a j^{th} order score at P. For $\mathcal{Q}(Q, g) := \{Q_g(\epsilon) : \epsilon\} \subset Q(\mathcal{M})$ to be a k^{th} order least-favorable submodel, each of $D^{(1)*}(Q, g)$ and $D^{(j)}(Q, g)$, $j = 2, 3, \ldots, k$, must be contained in the closure of the linear span of the generalized score vector $\frac{\partial}{\partial \epsilon} L(Q_g(\epsilon))$ at $\epsilon = 0$. Given an initial estimate $(Q_{n,0}, g_n)$ of (Q_0, g_0), applying the same iterative updating algorithm described above will generate a revised estimate Q_n^* such that

$$P_n D^{(1)*}(Q_n^*, g_n) = P_n \bar{D}_n^{(2)}(Q_n^*, g_n) = \ldots = P_n \bar{D}_n^{(k)}(Q_n^*, g_n) = 0 \; ,$$

from which asymptotic properties of the k-TMLE $\Psi(Q_n^*)$ can be derived.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

4.4 Example: estimating a counterfactual mean

We provide a concrete example of the above algorithm by developing a 2-TMLE in the context of our motivating example, namely that of estimating $E_{P_0}E_{P_0}(Y \mid A=1,W)$. To ensure that a second-order gradient exists, we focus on the case where the covariate vector W has finite support; in this case, a second-order gradient is given by (3.5).

The target parameter $\Psi(P) := E_P E_P(Y \mid A=1,W)$ depends on P via $Q(P) = (\bar{Q}(P), Q_W(P))$, where $\bar{Q}(P)(w) := E_P(Y \mid A=1, W=w)$ and $Q_W(P)(w) := P(W=w)$ describes the marginal distribution of W. It is straightforward to verify that the loss function $L(Q) := L_1(\bar{Q}) + L_2(Q_W)$, where we define

$$\begin{aligned} L_1(\bar{Q})(o) &:= -a\left[y \log \bar{Q}(w) + (1-y)\log(1-\bar{Q}(w))\right] \quad \text{and} \\ L_2(Q_W)(o) &:= -\log Q_W(w) \end{aligned}$$

pointwise for each $o=(w,a,y)$, is such that $(\bar{Q}_0, Q_{W,0}) := (\bar{Q}(P_0), Q_W(P_0))$ indeed minimizes the true risk $P_0L(Q)$.

As previously stated, the first-order canonical gradient of Ψ is given by $D^{(1)*}(P) = D_Y^{(1)*}(P) + D_W^{(1)*}(P)$, where

$$\begin{aligned} D_Y^{(1)*}(P)(o) &:= H^{(1)}(\bar{g})(a,w)\left[y - \bar{Q}(w)\right] \text{ and} \\ D_W^{(1)*}(P)(o) &:= \bar{Q}(w) - \Psi(P)\ , \end{aligned}$$

with $H^{(1)}(\bar{g})(a,w) := a/\bar{g}(w)$ and $\bar{g}(w) := P(A=1 \mid W=w)$, as before. Furthermore, $D^{(2)}(P)(o_1,o_2) := H^{(2)}(P)(w_1,a_1,w_2,a_2)\left[y_1 - \bar{Q}(w_1)\right]$ is a second-order gradient upon symmetrization, with

$$H^{(2)}(P)(w_1,a_1,w_2,a_2) \ := \ \frac{2a_1 I(w_1=w_2)}{\bar{g}(w_1)Q_W(w_1)}\left[1 - \frac{a_2}{\bar{g}(w_1)}\right].$$

Setting $g(P) := \bar{g}$, both $D^{(1)*}(P)$ and $D^{(2)}(P)$ depend on P through $(Q(P), g(P))$, and the components of $(Q(P), g(P))$ are variationally independent of each other since they involve orthogonal portions of the likelihood. This provides greater flexibility in constructing appropriate fluctuation submodels. Taking $S_x^*(P)(o) := 2aI(w=x)[y - \bar{Q}(w)]$, $f_x(P)(o) := -I(w=x)[a-\bar{g}(w)]/[\bar{g}(x)^2Q_W(x)]$ and ν the counting measure on the support of Q_W, we have that $D^{(2)}(P)(o_1,o_2) = \int_x S_x^*(P)(o_1)f_x(P)(o_2)d\nu(x)$ with $S_x^*(P) \in T(P)$ for each x, as expected. Defining

$$\bar{D}_n^{(2)}(P)(o) := \bar{H}_n^{(2)}(P)(w,a)\left[y - \bar{Q}(w)\right]$$

with $\bar{H}_n^{(2)}(P)(w,a) := \frac{1}{n}\sum_{j=1}^n H^{(2)}(P)(w,a,W_j,A_j)$, we note that $\bar{D}_n^{(2)}(P) \in T(P)$ and in fact is a score of the conditional distribution of Y given A and W. Furthermore, we observe that

$$P_n\bar{D}_n^{(2)}(P) = P_n^2 D^{(2)}(P)\ .$$

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

We also note that, at $Q_W = Q_{W,n}$, the empirical distribution of W,

$$\bar{H}_n^{(2)}(P)(w,a) = \frac{2a}{\bar{g}(w)}\left[1 - \frac{\bar{g}_{n,NP}(w)}{\bar{g}(w)}\right] ,$$

where $\bar{g}_{n,NP}(w) := \sum_{i=1}^n I(A_i = 1, W_i = w) / \sum_{i=1}^n I(W_i = w)$ is the nonparametric maximum likelihood estimator of $\bar{g}_0(w)$.

We can readily verify that $P_n D_W^{(1)*}(P) = 0$ if $Q_W(P) = Q_{W,n}$. Thus, if the initial estimator of $Q_{W,0}$ is taken to be $Q_{W,n}$, to achieve our objective of solving the requisite first- and second-order estimating equations, it will suffice to produce an estimate $\bar{Q}_n^*$ of $\bar{Q}_0$ satisfying

$$P_n D_Y^{(1)*}(\bar{Q}_n^*, Q_{W,n}, \bar{g}_n) = P_n \bar{D}_n^{(2)}(\bar{Q}_n^*, Q_{W,n}, \bar{g}_n) = 0 \tag{4.2}$$

with $\bar{g}_n$ our initial estimate of $\bar{g}_0$. Given any particular $\bar{Q}$, Q_W and $\bar{g}$, the submodel determined by

$$\bar{Q}_{\bar{g},Q_W}(\epsilon) := \text{expit}\left[\text{logit}(\bar{Q}) + \epsilon_1 H^{(1)}(\bar{g}) + \epsilon_2 \bar{H}_n^{(2)}(\bar{g}, Q_W)\right]$$

with $\epsilon := (\epsilon_1, \epsilon_2)$ satisfies that $\bar{Q}_{\bar{g},Q_W}(0) = \bar{Q}$ as well as

$$\begin{aligned}
\frac{\partial}{\partial \epsilon_1} L(\bar{Q}_{\bar{g},Q_W}(\epsilon), Q_W)(o)\Big|_{\epsilon=0} &= H^{(1)}(\bar{g})(o)\left[y - \bar{Q}(w)\right] = D_Y^{(1)*}(P)(o) , \\
\frac{\partial}{\partial \epsilon_2} L(\bar{Q}_{\bar{g},Q_W}(\epsilon), Q_W)(o)\Big|_{\epsilon=0} &= \bar{H}_n^{(2)}(\bar{g}, Q_W)(o)\left[y - \bar{Q}(w)\right] = \bar{D}_n^{(2)}(P)(o) .
\end{aligned}$$

As such, given an initial estimate $(\bar{Q}_{n,0}, \bar{g}_n)$ of $(\bar{Q}_0, \bar{g}_0)$, a first updated estimate $\bar{Q}_{n,1}$ of $\bar{Q}_0$ is obtained by selecting the minimizer of the empirical risk $P_n L_1(\bar{Q})$ with $\bar{Q}$ ranging over the parametric submodel determined by

$$\bar{Q}_{n,0}(\epsilon) := \text{expit}\left[\text{logit}(\bar{Q}_{n,0}) + \epsilon_1 H^{(1)}(\bar{g}_n) + \epsilon_2 \bar{H}_n^{(2)}(\bar{g}_n, Q_{W,n})\right]$$

and $-\infty < \epsilon_1, \epsilon_2 < +\infty$. The optimal values of ϵ_1 and ϵ_2 can be readily obtained as estimated regression coefficients in the fit of a logistic regression model with outcome Y_i, covariates $H^{(1)}(\bar{g}_n)(O_i) = A_i / \bar{g}_n(W_i)$ and

$$\bar{H}_n^{(2)}(\bar{g}_n, Q_{W,n})(O_i) = \frac{2A_i}{\bar{g}_n(W_i)}\left[1 - \frac{\bar{g}_{n,NP}(W_i)}{\bar{g}_n(W_i)}\right] ,$$

and offset $\text{logit}(\bar{Q}_{n,0}(W_i))$ restricted to the subset of data points for which $A_i = 1$. It is not difficult to see that any further attempt to update $\bar{Q}_{n,1}$ by considering the second-order least-favorable submodel through it will not produce any change. As such, in this case, the algorithm terminates in a single step, so that $\bar{Q}_n^* = \bar{Q}_{n,1}$ and (4.2) must then hold. The resulting 2-TMLE of ψ_0 is finally given by

$$\Psi(\bar{Q}_n^*, Q_{W,n}) = \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^*(W_i) .$$

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper331

This estimator allows concrete theoretical gains relative to the 1-TMLE described previously and preliminary simulation results suggest these indeed translate well into practical gains.

In the algorithm described above, there was no need to update the estimate of $Q_{W,0}$ at all. This resulted from the selection of the NPMLE $Q_{W,n}$ as initial estimator and of the log-likelihood loss for Q_W. Had any of these two choices differed, it would generally have been necessary to iteratively update the estimate of $Q_{W,0}$ using an appropriate fluctuation submodel to ensure that the resulting targeted estimate $Q^*_{W,n}$ satisfy that

$$P_n D_W^{(1)*}(\bar{Q}^*_n, Q^*_{W,n}) = 0 \ ,$$

leading then to the 2-TMLE of ψ_0 given by $\sum_w \bar{Q}^*_n(w) Q^*_{W,n}(w)$.

As transpires from the developments above, if the support of W is finite but nonetheless rich, large samples will be required to ensure that $\bar{g}_{n,NP}$ behaves sufficiently well as an estimator of $\bar{g}_0$. Given the sufficiency property of the propensity score as a summary of potential confounders, it is natural to inquire whether the use of a second-order partial canonical gradient based on the propensity score may allow us to circumvent the dimensionality of W. Suppose that W is finitely supported and consider the second-order partial canonical gradient given by

$$D_0^{(2)a}(P)(o_1, o_2) := H_0^a(P)(w_1, a_1, w_2, a_2)[y_1 - \bar{Q}(w_1)] \ ,$$

where we define pointwise

$$H_0^a(P)(w_1, a_1, w_2, a_2) := \frac{2a_1 I(\bar{g}_0(w_1) = \bar{g}_0(w_2))}{\bar{g}(w_1) Q_{\bar{g}_0(W),0}(\bar{g}_0(w_1))} \left[1 - \frac{a_2}{\bar{g}(w_1)}\right]$$

and $Q_{\bar{g}_0(W),0}$ denotes the probability mass function of $\bar{g}_0(W)$ under P_0. Clearly, the important benefit of this choice is that the event $\{W_1 = W_2\}$ is replaced by the better-supported event $\{\bar{g}_0(W_1) = \bar{g}_0(W_2)\}$. Unsurprisingly, (3.6) holds identically when using $D_0^{(2)a}$ instead of $D^{(2)}$. Of course, the limitation of this result, in practice, is that use of this alternate second-order partial canonical gradient presupposes knowledge of $\bar{g}_0$ and $Q_{\bar{g}_0(W),0}$. To render this approach practically useful, we may replace $\bar{g}_0$ by $\bar{g}$ in the definition of $D_0^{(2)a}$ and hope that this substitution contributes no more than an additional third-order term in (3.6). Unfortunately, a careful study of the resulting remainder reveals that it is only a second-order term, essentially behaving asymptotically like $R_2(P, P_0)$. In particular, for the sake of achieving asymptotic linearity and efficiency, the conditions required on the initial estimators in such a 2-TMLE procedure are no different than those arising in the study of the standard 1-TMLE. While it is plausible that finite-sample gains would be derived from the use of a 2-TMLE based on this candidate second-order partial canonical gradient, these benefits cannot be expected to persist asymptotically.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

5 Inference using approximate second-order gradients

5.1 Asymptotic linearity and efficiency

The theory of higher-order TMLE outlined above applies to settings wherein the parameter is higher-order pathwise differentiable. However, as indicated before, in realistic models, many parameters of interest are not smooth enough as functionals of the data-generating distribution to admit even a second-order gradient. Nonetheless, a useful higher-order theory can still be developed with the introduction of approximate higher-order partial canonical gradients, as the following theorem indicates.

Theorem 2. *Suppose that for each h the element $D_h^{(2)}(P) \in L_0^2(P^2)$ satisfies that*

$$o_1 \mapsto D_h^{(2)}(P)(o_1, o)$$

lies in $T(P)$ for each $o \in \mathcal{O}$, that the target parameter Ψ admits the second-order expansion $\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \frac{1}{2} \lim_{h \to 0} P_0^2 D_h^{(2)}(P) + R_3(P, P_0)$, and that $P_n^* \in \mathcal{M}$ satisfies the equations

$$P_n D^{(1)*}(P_n^*) = 0 \quad \text{and} \quad P_n^2 D_{h_n}^{(2)}(P_n^*) = P_n \bar{D}_{h_n,n}^{(2)}(P_n^*) = 0$$

with $\bar{D}_{h_n,n}^{(2)}(P_n^*)(o) := \frac{1}{n} \sum_{i=1}^{n} D_{h_n}^{(2)}(P_n^*)(o, O_i)$. Then, provided that

1. *there exists a P_0-Donsker class $\mathcal{F}$ such that $D^{(1)*}(P_n^*)$ is in $\mathcal{F}$ with probability tending to one, and $P_0 \left[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)\right]^2 = o_P(1)$;*
2. *$(P_n^2 - P_0^2) D_{h_n}^{(2)}(P_n^*) = o_P(n^{-1/2})$;*
3. *$B_n(h_n) := \frac{1}{2} \left[P_0^2 D_{h_n}^{(2)}(P_n^*) - \lim_{h \to 0} P_0^2 D_h^{(2)}(P_n^*)\right] = o_P(n^{-1/2})$;*
4. *$R_3(P_n^*, P_0) = o_P(n^{-1/2})$;*

ψ_n^* is an asymptotically linear estimator of ψ_0 with influence function $D^{(1)*}(P_0)$. It is thus also asymptotically efficient.

Sufficient conditions for establishing the validity of Condition 2 in the above theorem are identical to those discussed in Lemma 2 upon replacing the second-order partial canonical gradient $D^{(2)}$ by its approximation $D_{h_n}^{(2)}$. Again, Lemma 3 implies that

$$(P_n - P_0)^2 D_{h_n}^{(2)}(P_n^*) = O_P\left(\frac{\|D_{h_n}^{(2)}(P_n^*)\|_v^*}{n}\right),$$

which provides a basis for establishing that this term is $o_P(n^{-1/2})$ for sequences h_n that do not converge to zero too quickly. In contrast, it will be necessary for h_n to tend to zero quickly enough for the representation error $B_n(h_n)$ to be asymptotically

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

negligible. A careful trade-off between these opposing constraints must therefore be achieved.

The above theorem is a direct generalization of Theorem 1 allowing for the lack of existence of a second-order partial canonical gradient. This theorem can readily be generalized to higher orders without any additional effort. For this purpose, it will be required that the representation errors $B_n^j(h_n) := \frac{1}{2}[P_0^j D_{h_n}^{(j)}(P_n^*) - \lim_{h\to 0} P_0^j D_h^{(j)}(P_n^*)]$ be themselves $o_P(n^{-1/2})$ for $j = 2, 3, \ldots, k$. This will ensure that the approximation used does not itself generate higher-order terms which then would have to be taken into account as well. If relevant portions of the underlying data-generating distribution P_0 are sufficiently smooth, this condition will be achievable using appropriate smoothing techniques, for example. Provided the higher-order expansion

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \sum_{j=2}^{k} \frac{1}{j!} \lim_{h\to 0} P_0^j D_h^{(j)}(P) + R_{k+1}(P, P_0) \ ,$$

Theorem 2 holds if conditions 2 and 3 above are replaced by the requirement that $(P_n^j - P_0^j) D_{h_n}^{(j)}(P_n^*)$ and $B_n^j(h_n)$ both be $o_P(n^{-1/2})$ for each $j = 2, 3, \ldots, k$, and condition 4 instead involves the $(k+1)^{th}$ order remainder term, so that $R_{k+1}(P_n^*, P_0) = o_P(n^{-1/2})$.

5.2 Implementation and selection of tuning parameter

A proper k-TMLE procedure will need to yield an estimate $P_n^* \in \mathcal{M}$ of P_0 such that

$$P_n D^{(1)*}(P_n^*) = P_n \bar{D}_{h_n,n}^{(2)}(P_n^*) = \ldots = P_n \bar{D}_{h_n,n}^{(k)}(P_n^*) = 0 \ ,$$

where $\bar{D}_{h_n,n}^{(j)}(P_n^*)$ is defined similarly as $\bar{D}_n^{(j)}(P_n^*)$ but with $D^{(j)}$ replaced by $D_{h_n}^{(j)}$. Thus, a k-TMLE procedure can be defined in exactly the same way as in Subsection 4.3, where the scores from higher-order terms are replaced by their corresponding approximation. As such, except for the selection of an appropriate tuning parameter value h_n, the algorithm is no more difficult to implement than in settings where higher-order gradients exist. Of course, the key challenge we must face in practice is the data-adaptive determination of an appropriate value for h_n.

The choice of h in this problem determines the least-favorable parametric submodel used in the TMLE algorithm. Thus, the selection of h can be seen as completely analogous to the selection of an estimator of the treatment or censoring mechanism, principal determinants of the first-order least-favorable submodel, in the problem of estimating a counterfactual mean when all potential confounders have been recorded or when the available data are subject to censoring. A principled solution to this problem, referred to as the collaborative TMLE algorithm and abbreviated C-TMLE, has been described, thoroughly discussed and implemented in van der Laan and Gruber [2010], Gruber and van der Laan [2010], Stitelman and van der Laan [2010], van der Laan and Rose [2011] and Gruber and van der Laan [2012]. This approach can readily be utilized here as well.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

C-TMLE consists of a TMLE algorithm that automatically selects among a collection of candidate least-favorable parametric submodels in its updating step. For each candidate submodel, the resulting TMLE algorithm yields a decrease in empirical risk; the magnitude of this change can serve as a criterion for adjudicating the value of this particular submodel. C-TMLE uses precisely this criterion for data-adaptively building or selecting a least-favorable parametric submodel, and thereby a corresponding TMLE. Under certain regularity conditions, the resulting estimator is asymptotically linear and efficient provided at least one of the candidate submodels allows for complete bias reduction asymptotically.

In the setting of a 2-TMLE, denoting by $(Q_{n,0}, g_n)$ an initial estimate of (Q_0, g_0) and letting h index a candidate second-order least-favorable submodel determined by $D^{(1)*}(Q_{n,0}, g_n)$ and $D_h^{(2)}(Q_{n,0}, g_n)$, a C-TMLE solution would first select the h-value for which the possible decrease in the empirical risk $P_n L(Q_{h,n,0}(\epsilon))$ along the corresponding h-specific parametric submodel $\{Q_{h,n,0}(\epsilon) : \epsilon\}$ through $Q_{n,0}$ is maximal and perform a single updating step along this submodel. In other words, h_n^0 would be selected as the minimizer of

$$h \mapsto P_n L(Q_{h,n,0}(\epsilon_n^0(h)))$$

with $\epsilon_n^0(h) := \text{argmin}_\epsilon\, P_n L(Q_{h,n,0}(\epsilon))$ for each h and the first update would then be $Q_{n,1} := Q_{h_n^0,n,0}(\epsilon_n^0(h_n^0))$. The next targeting step would be carried out similarly, with subsequent choices of h constrained to be non-increasing. Specifically, letting h_n^1 denote the minimizer of

$$h \mapsto P_n L(Q_{h,n,1}(\epsilon_n^1(h)))$$

over the interval $[0, h_n^0]$, where $\epsilon_n^1(h) := \text{argmin}_\epsilon\, P_n L(Q_{h,n,1}(\epsilon))$ for each h, the updated estimate of Q_0 would then be $Q_{n,2} := Q_{h_n^1,n,1}(\epsilon_n^1(h_n^1))$. This iterative process would proceed until the decrease in empirical risk from an additional step is no longer significant according to some pre-specified criterion, such as BIC or some cross-validation risk, for example. Denoting the final estimate of Q_0 and the final value of h_n at convergence by Q_n^* and h_n^*, respectively, the TMLE $\Psi(Q_n^*)$ could be directly used, or the 2-TMLE based on the initial estimate $Q_{n,0}$ and final tuning parameter choice $h = h_n^*$ could be constructed anew.

In the scheme proposed above, the performance of a particular h-value is adjudicated on the basis of the decrease in empirical risk resulting from a single targeting step. While this seems sensible in settings where convergence of the TMLE updating process occurs analytically in a single step, in other settings, it may not be an optimal way to proceed. As an alternative, it would be possible to carry out a fully iterated TMLE until convergence for every choice of h, to select the optimal h-value on this basis, and then, to repeat until overall convergence. Of course, this variant of the algorithm would be potentially much more computationally intensive.

The approach described above is used to fully automate the selection of the tuning parameter required in the setting of a 2-TMLE based on an approximate second-order partial canonical gradient. The idea could directly be applied to a k-TMLE as well. More importantly though, the same principle could be used to construct more involved

COBRA
Collection of Biostatistics
Research Archive

collaborative TMLE algorithms that not only select tuning parameters but also other choices that define the approximate second-order partial canonical gradient.

5.3 Example: estimating a counterfactual mean

When the distribution of the vector of potential confounders has infinite support, the counterfactual mean generally does not admit a second-order gradient in the nonparametric model we have been considering. An approximate second-order canonical gradient can nonetheless be considered, leading to a well-defined 2-TMLE, which we describe and study in this subsection.

5.3.1 An approximate second-order canonical gradient

Suppose that $W := (W(1), \dots, W(d))$ and each $W(j)$ is real-valued. Defining pointwise $D_h^{(2)}(\bar{Q}, \bar{g}, Q_W)(o_1, o_2) := H_h^{(2)}(\bar{g}, Q_W)(w_1, a_1, w_2, a_2)[y_1 - \bar{Q}(w_1)]$, where

$$H_h^{(2)}(\bar{g}, Q_W)(w_1, a_1, w_2, a_2) := \frac{1}{h^d} K\left(\frac{w_1 - w_2}{h}\right) \frac{2a_1}{\bar{g}(w_1) Q_W(w_1)} \left[1 - \frac{a_2}{\bar{g}(w_1)}\right]$$

with K a compactly-supported multivariate kernel function and h a positive bandwidth, $D_h^{(2)}$ is seen to be a kernel approximation of the second-order partial canonical gradient $D^{(2)}$ defined in Subsection 4.4. For each w, we henceforth denote $h^{-1}K(h^{-1}w)$ by $K_h(w)$. The following lemma establishes conditions under which the bias term arising in a TMLE due to the use of this approximate second-order canonical gradient is negligible.

Lemma 4. *Suppose that the distribution of W has compact support and is absolutely continuous with respect to Lebesgue measure with density $Q_{W,0}$. Suppose that $Q^{\circ}_{W,n}$ is a working estimate of $Q_{W,0}$. If*

1. *both $\bar{g}_0$ and $Q_{W,0}$ are $(m_0 + 1)$-times continuously differentiable almost surely;*
2. *K is orthogonal to all polynomial powers up until m_0;*
3. *there exists some $\delta > 0$ such that $\bar{g}_0$ is bounded below by δ, and both $\bar{g}_n$ and $Q^{\circ}_{W,n}$ are bounded below by δ with probability tending to one,*

then we have that

$$P_0^2 D_h^{(2)}(\bar{Q}_n^*, \bar{g}_n, Q^{\circ}_{W,n}) - \lim_{h \to 0} P_0^2 D_h^{(2)}(\bar{Q}_n^*, \bar{g}_n, Q^{\circ}_{W,n}) = O_P\left(h^{m_0+1} \|\bar{Q}_n^* - \bar{Q}_0\|\right),$$

where $\|\bar{Q}_n^* - \bar{Q}_0\|^2 := \int (\bar{Q}_n^* - \bar{Q}_0)^2(w) Q_{W,0}(w) dw$.

The result above explicitly deals with kernel smoothing with an equal bandwidth in all dimensions. The lemma also holds, however, if a multivariate bandwidth is utilized, with h substituted by $\max_j h_j$ in the statement of the lemma.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

5.3.2 A corresponding 2-TMLE

Computation of the counterfactual mean does not in principle require an estimate of the density function of W - indeed, an estimate of the distribution function of W suffices. Nonetheless, the second-order targeting process explicitly requires this density function, as is apparent from the form of the approximate second-order canonical gradient. Thus, to construct a 2-TMLE, an estimate of the density of W must first be obtained. At least two approaches can be considered to tackle this issue:

1. a fixed smooth estimate $Q^\circ_{W,n}$ of $Q_{W,0}$ is used whenever required in the second-order targeting process to yield a targeted estimate of $\bar{Q}^*_n$ but final computation of the targeted estimate of ψ_0 is based on the empirical measure $Q_{W,n}$;
2. the same smooth estimate $Q^\circ_{W,n}$ of $Q_{W,0}$ is used both in the second-order targeting process and in the computation of the targeted estimate of ψ_0.

The second option requires a substantial amount of additional work since in order for the smooth estimator $Q^\circ_{W,n}$ to be appropriate for the sake of constructing the targeted estimate of ψ_0 it must itself be targeted. As such, it will need to be iteratively updated along with estimates of $\bar{Q}_0$. This issue is circumvented in the first option since the empirical distribution $Q_{W,n}$ is a NPMLE and therefore already solves the relevant score equation. For this reason, we recommend the first option in practice and restrict our theoretical study below to this option alone.

Consider the 2-TMLE using the empirical distribution $Q_{W,n}$ as initial estimator of $Q_{W,0}$ in the TMLE algorithm but using a fixed smooth estimator $Q^\circ_{W,n}$ in the quantity

$$H^{(2)}_h(\bar{g}_n, Q^\circ_{W,n})$$

used to construct the logistic regression-based least-favorable submodel through a current estimate $\bar{Q}_{n,m}$ of $\bar{Q}_0$. This is then precisely an example of the 2-TMLE of Subsection 4.4. The implementation of this algorithm is particularly simple because a single step of targeting suffices to achieve analytic convergence. To perform this single updating step, the maximum likelihood estimate ϵ_n of $\epsilon = (\epsilon_1, \epsilon_2)$ in the logistic regression model

$$\bar{Q}_{n,0,h}(\epsilon) := \text{expit}\left[\text{logit}(\bar{Q}_{n,0}) + \epsilon_1 H^{(1)}(\bar{g}_n) + \epsilon_2 \bar{H}^{(2)}_{h,n}(\bar{g}_n, Q^\circ_{W,n})\right]$$

is obtained, with $H^{(1)}(\bar{g}_n)(o) = a/\bar{g}_n(w)$ and

$$\begin{aligned}\bar{H}^{(2)}_{h,n}(\bar{g}_n, Q^\circ_{W,n})(o) &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h^d} K\left(\frac{w - W_j}{h}\right) \frac{2a}{\bar{g}_n(w) Q^\circ_{W,n}(w)} \left[1 - \frac{A_j}{\bar{g}_n(w)}\right] \\ &= \frac{2a}{\bar{g}_n(w) Q^\circ_{W,n}(w)} \left[\frac{1}{n}\sum_j K_h(w - W_j) - \frac{\frac{1}{n}\sum_j K_h(w - W_j) A_j}{\bar{g}_n(w)}\right].\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

If the kernel estimate $Q^{\circ}_{W,n}(w) := \frac{1}{n}\sum_{i=1}^{n} K_h(w - W_i)$ is used, it is easy to verify that the simplification

$$\bar{H}^{(2)}_{h,n}(\bar{g}_n, Q^{\circ}_{W,n})(o) = \frac{2a}{\bar{g}_n(w)}\left[1 - \frac{\bar{g}_{n,h}(w)}{\bar{g}_n(w)}\right]$$

ensues, where $\bar{g}_{n,h}(w) := \sum_j K_h(w - W_j)A_j / \sum_j K_h(w - W_j)$ is the Nadaraya-Watson estimator of $\bar{g}_0(w)$ indexed by bandwidth h.

Given the targeted estimate $\bar{Q}^*_{n,h_n} := \bar{Q}_{n,0,h_n}(\epsilon_n)$ of $\bar{Q}_0$ based on this single-step procedure, the resulting 2-TMLE of ψ_0 is $\psi^*_n := \Psi(\bar{Q}^*_{n,h_n}, Q_{W,n}) = \frac{1}{n}\sum_{i=1}^{n}\bar{Q}^*_{n,h_n}(W_i)$. The following theorem provides conditions under which ψ^*_n is an asymptotically linear and efficient estimator of ψ_0. Below, we will make reference to the third-order remainder term R_3 defined as

$$R_3(P, P_0) := P_0\left[\left(1 - \frac{Q_{W,0}\bar{g}_0}{Q_W\bar{g}}\right)\left(\frac{\bar{g} - \bar{g}_0}{\bar{g}}\right)(\bar{Q} - \bar{Q}_0)\right],$$

as before.

Theorem 3. *Under the conditions of Lemma 4, and provided that*

1. *each of $\bar{g}_n - \bar{g}_0$, $\bar{Q}^*_n - \bar{Q}_0$ and $Q^{\circ}_{W,n} - Q_{W,0}$ tend to zero in $L^2(Q_{W,0})$-norm;*
2. *there exists some $\delta > 0$ such that $\bar{g}_0$, $\bar{g}_n$ and $Q^{\circ}_{W,n} \cdot \bar{g}_n$ are bounded below by δ with probability tending to one;*
3. *each of $\bar{g}_n$, $\bar{Q}^*_n$ and $Q^{\circ}_{W,n}$ have uniform sectional variation norm bounded by some $M < \infty$ with probability tending to one;*
4. *the kernel function K is $2d$-times differentiable and $h_n^{2d} n \to +\infty$,*

and either of

5a. $R_2(P^*_n, P_0) = o_P(n^{-1/2})$;
5b. $R_3(P^*_n, P_0) = o_P(n^{-1/2})$ *and* $\|\bar{Q}^*_n - \bar{Q}_0\| h_n^{m_0+1} = o_P(n^{-1/2})$

holds, ψ^*_n is an asymptotically linear estimator of ψ_0 with influence function $D^{(1)*}(P_0)$. It is thus also asymptotically efficient.

As is evident from the above conditions, the rate at which the bandwidth h_n decreases plays a critical role in the asymptotic behavior of the 2-TMLE described. On one hand, condition 2 of the theorem requires that the bandwidth converge to zero sufficiently quickly in order for $n^{1/2}\|\bar{Q}^*_n - \bar{Q}_0\| h^{m_0+1}$ to itself converge to zero, where m_0 is the order of the kernel K used. This ensures that the representation error is negligible. On the other hand, condition 6 requires h_n to converge to zero slowly enough to allow control of the term $(P^2_n - P^2_0)D_{h_n}(P^*_n)$.

We remark that when K is indeed taken to be a tensor product of uniform kernels over $(-0.5, +0.5)$, $m_0 = 1$ and condition 2 becomes more restrictive. While for $d \leq 2$, there does indeed exist a rate h_n for which both conditions are true, this is not the case if $d > 2$. Therefore, even though higher-order kernels increase the variation norm

Hosted by The Berkeley Electronic Press

and thereby lead to more stringent conditions on h_n, when the vector of potential confounders includes more than two components, it is necessary to use higher-order kernels that fully exploit the underlying smoothness of the distribution of (A, W) in order to control the representation error.

Scrutiny of the theorem above reveals that a 2-TMLE will indeed generally be asymptotically linear and efficient in a larger model compared to a corresponding 1-TMLE. On one hand, as explicitly reflected in Theorem 3, for example, it is generally true that whenever a 1-TMLE is efficient, so will be a 2-TMLE. This formalizes our earlier claim that 2-TMLE operates in a safe haven wherein we expect not to hurt a 1-TMLE by performing the additional targeting required to construct a 2-TMLE. On the other hand, we note that 2-TMLE will be efficient in many instances in which 1-TMLE is not. As an illustration, suppose in the setting of our motivating example that W is a univariate random variable with a sufficiently smooth density function. Suppose also that $\bar{g}_0$ is smooth enough so that an optimal univariate second-order kernel smoother can be utilized to produce an estimate of $\bar{g}_0$. In this case, efficiency of a 1-TMLE requires that $\bar{Q}_n$ tends to Q_0 at a rate faster than $n^{-1/10}$. In contrast, the corresponding 2-TMLE built upon a second-order canonical gradient approximated using an optimal second-order kernel smoother will be efficient provided that $\bar{Q}_n$ is consistent for $\bar{Q}_0$, irrespective of the actual rate of convergence. The difference between these requirements may not seem drastic in settings where $\bar{Q}_0$ is sufficiently smooth since then constructing an estimator $\bar{Q}_n$ which satisfies both requirements is easy. This is certainly not so if $\bar{Q}_0$ fails to be smooth, in which case achieving convergence even at $n^{-1/10}$-rate may be a challenge. This problem is exacerbated further if W has several components. For example, if W is 5-dimensional, a 1-TMLE requires that $\bar{Q}_n$ tend to $\bar{Q}_0$ faster than $n^{-5/18}$, whereas the corresponding 2-TMLE based on a third-order kernel-smoothed approximation of the second-order canonical gradient requires that $\bar{Q}_n$ tend to $\bar{Q}_0$ faster than $n^{-1/5}$. While the latter is achievable using an optimal second-order kernel smoother, the former is not, and without further smoothness assumptions on $\bar{Q}_0$, a 1-TMLE will generally not be efficient.

As in the last paragraph of Section 4, we may hope that systematic dimension reduction may be performed by replacing $K_h(w_1 - w_2)$ in the definition of $D_h^{(2)}$ by a kernel-based discrepancy based on the propensity score at w_1 and w_2. This can be accomplished if $\bar{g}_0$ is known, in which case $K_h(\bar{g}_0(w_1) - \bar{g}_0(w_2))$ can be used to define a dimension-reduced approximate second-order partial canonical gradient without sacrificing the order of the remainder in the associated expansion. As before, replacing $\bar{g}_0$ by $\bar{g}$ in this kernel discrepancy unfortunately introduces a second-order term in the remainder, thereby invalidating the theoretical justification for using such a second-order partial canonical gradient. Again, this does not preclude the possibility that finite-sample benefits might be derived from taking this path.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

6 Concluding remarks

If the target parameter is higher-order pathwise differentiable, there appears to be no reason not to implement a higher-order TMLE using appropriate higher-order partial canonical gradients. Compared to its first-order counterpart, a higher-order TMLE will be an asymptotically linear and efficient substitution estimator under weaker conditions. This article provides a clear template describing how to construct such a higher-order TMLE. As discussed at length, its implementation is identical to that of a regular TMLE, save for the use of a higher-dimensional least-favorable parametric submodel that generates a set of scores defined by the partial canonical gradients of all orders.

We explicitly described this construction of a second-order TMLE for estimating a counterfactual mean under the assumption that the vector W of confounders has finite support. The latter ensured the presence of ties $W_i = W_j$ for $i \neq j$, a necessary ingredient for the existence of a non-degenerate second-order gradient. If W includes continuous components, for example, such ties have null probability of occurring and a second-order gradient does not exist. Since a second-order canonical gradient is obtained by projecting any candidate second-order gradient into a second-order tangent space, it may well be that the second-order canonical gradient involves a smoothed version of the indicator $I(W_i = W_j)$ if the semiparametric model is sufficiently smaller than the saturated nonparametric model.

Inspired by the fact that $E_P E_P(Y \mid A = 1, W) = E_P E_P(Y \mid A = 1, \bar{g}(W))$, we note that it may be of interest to instead focus on estimation of the data-adaptively-defined parameter value $\Psi_n(P_0)$, where $\Psi_n(P) := E_P E_P(Y \mid A = 1, \bar{g}_n(W))$ and $\bar{g}_n$ is considered deterministic. The obvious benefit from adopting such an approach is that the set of confounders we must adjust for is reduced to a univariate measure, and as such, only univariate smoothing is required in computing the approximate second-order canonical gradient. Nonetheless, this simplification is achieved by altering the target of inference – this nullifies the need to consider the bias due to adjusting for $\bar{g}_n(W)$ rather than $\bar{g}_0(W)$, but also modifies the interpretation of the resulting inference, since the target parameter can no longer be considered as a bona fide mean counterfactual outcome under causal assumptions.

To tackle target parameters that are not higher-order pathwise differentiable, we make use of approximate higher-order gradients in this article. This approximation strategy impacts only the higher-order bias reduction performed by the higher-order partial canonical gradients. Since such a higher-order TMLE is tailored to eliminate the higher-order terms in the Taylor expansion for the TMLE, a meaningful bias reduction relative to a standard first-order TMLE can be expected in practice if higher-order partial canonical gradients are reasonably approximated. Unfortunately, these approximations rely on possibly high-dimensional smoothing and one may wonder whether we have simply replaced a difficult problem by another difficult problem. Fortunately, approximate higher-order TMLEs operate in a safe haven in which the desirable properties of the first-order TMLE are preserved. Adding extra parameters to the least-favorable

COBRA
Collection of Biostatistics
Research Archive

submodel always improve the estimator asymptotically and can be very reasonably expected not to harm the estimator in finite samples.

In practice, the higher-order TMLE, based on actual or approximate higher-order partial canonical gradients, should aim to achieve as much bias reduction as possible with the higher-order least-favorable submodel. In this paper, we provided theoretically-sound and practicable building blocks for achieving this. Since any higher-order TMLE is a substitution estimator, the finite-sample variability induced by the need to fit a higher-order least-favorable submodel is controlled by global bounds on the model and target parameter mapping. This is likely even more crucial in higher-order procedures given the need to correct a greater number of terms. Even more importantly, the seemingly daunting task of selecting tuning parameters and other potential inputs of the algorithm can be seamlessly overcome using an implementation of C-TMLE. The latter provides concrete tools for data-adaptively fine-tuning the selection of a higher-order least-favorable submodel with the objective of maximizing its effectiveness in reducing bias.

As highlighted in this article, the second-order TMLE in our motivating example provides a concrete demonstration of a gain relative to a first-order TMLE: it is asymptotically linear and efficient in a significantly larger statistical model than an analogue first-order TMLE. This is directly parallel to the advances made in the seminal work of Robins et al. [2009] in the context of the one-step estimators. Another advantage of including higher-order partial canonical gradients into the TMLE framework as carried out in this article is that they yield contributions to the influence function of the higher-order TMLE that can be directly incorporated in the construction of confidence intervals, thereby possibly leading to finite-sample performance improvements.

Finally, since the Donsker class conditions imposed in our theorems can be restrictive in some settings, it is certainly of interest to develop a higher-order cross-validated TMLE. The latter would use a cross-validated version of the empirical risk in the iterative updating procedure and could be shown to lead to an asymptotically linear and efficient estimator even when such Donsker class conditions fail to hold. Such a development would be a direct extension of the work of Zheng and van der Laan [2011] in the first-order case, particularly since, as we have illustrated, a higher-order TMLE can be framed as a first-order TMLE with an augmented least-favorable submodel.

Acknowledgement

This research was supported in part by NIH grant 2R01AI074345 and by an endowment generously provided by Genentech.

References

P.J. Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, 1997.

R.D. Gill, M.J. van der Laan, and J.A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré*, 31(3):545–597, 1995.

S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1), 2010.

S. Gruber and M.J. van der Laan. Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics*, 8(1), 2012.

I.A. Ibragimov and R.Z. Khasminskii. *Statistical estimation*. Springer, 1981.

D. La Vecchia, E. Ronchetti, and F. Trojani. Higher-order infinitesimal robustness. *Journal of the American Statistical Association*, 107(500):1546–1557, 2012.

S.D. Lendle, B. Fireman, and M.J. van der Laan. Balancing score adjusted targeted minimum loss-based estimation. 2013.

B.Y. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740, 1975.

L. Li, E. Tchetgen Tchetgen, A.W. van der Vaart, and J.M. Robins. Higher order inference on a treatment effect under low regularity conditions. *Statistics & Probability Letters*, 81 (7):821–828, 2011.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*. Springer, 1982.

J. Pfanzagl. *Asymptotic expansions for general statistical models*. Springer, 1985.

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.

D.B. Rubin and M.J. van der Laan. Targeted ancova estimator in rcts. In *Targeted Learning*, pages 201–215. Springer, 2011.

O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1), 2010.

M.J. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

M.J. van der Laan and S. Gruber. Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), 2010.

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

M.J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data.* Springer, 2011.

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

A.W. van der Vaart. *Asymptotic statistics.* Cambridge University Press, 2000.

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

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer, 1996.

W. Zheng and M.J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

Appendix: proof of theorems and lemmas

Proof of Lemma 1

We first establish the fact that, for any orthonormal basis $\{e_1, e_2, \ldots\}$ of the first-order tangent space $T(P) \subseteq L_0^2(P)$, the second-order tangent space $T^{(2)}(P)$ is given by

$$H := \left\{ (o_1, o_2) \mapsto \sum_{i_1, i_2} C(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2) : C \in \mathscr{C} \right\}$$

with $\mathscr{C}$ denoting the set of all symmetric mappings from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{R}$. By definition, $T^{(2)}(P)$ is the closure of the linear span of $\mathcal{S}_2(P)$, the set of linear combinations of score cross-products at P, formally defined as

$$\left\{ (o_1, o_2) \mapsto \int S_x(o_1) S_x(o_2) d\mu(x) : S_x \in T(P) \text{ for each } x,\ \mu \text{ has finite support} \right\}.$$

Take any element $s \in \mathcal{S}_2(P)$, say $(o_1, o_2) \mapsto s(o_1, o_2) := \int_x S_x(o_1) S_x(o_2) d\mu(x)$. For each x, $S_x \in T(P)$ can be written as $\sum_i C_x(i) e_i$ for some $C_x : \mathbb{N} \to \mathbb{R}$. Thus, we can write $s(o_1, o_2) = \sum_{i_1, i_2} C(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2)$ with $C(i_1, i_2) := \int C_x(i_1) C_x(i_2) d\mu(x)$. Hence, we see that $\mathcal{S}_2(P) \subseteq H$. Since H equals its closed linear span, it follows that $T^{(2)}(P) \subseteq H$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Suppose an arbitrary element h of H is characterized by $C_0 \in \mathscr{C}$. Given $\delta > 0$, it is always possible to find $K(\delta) \in \mathbb{N}$ such that

$$\left\| (o_1, o_2) \mapsto \sum_{i_1, i_2 > K(\delta)} C_0(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2) \right\| < \delta$$

in $L_2(P)$-norm. Set $x(i_1, i_2) = (i_1, i_2)$ for each $i_1, i_2 \in \{1, 2, \ldots, K(\delta)\}$ and let μ^δ be a measure putting mass $\text{sgn}(C_0(i_1, i_2))/2$ at $x(i_1, i_2)$ for $i_1, i_2 \in \{1, 2, \ldots, K(\delta)\}$ and no mass elsewhere. For $x = x(i_1, i_2)$ with $i_1, i_2 \in \{1, 2, \ldots, K(\delta)\}$ and for $j \in \{i_1, i_2\}$, set $C_x^\delta(j) := |C_0(i_1, i_2)|^{1/2}$. Otherwise, set $C_x^\delta(j) = 0$. Constructing $S_x := \sum_j C_x^\delta(j) e_j$, it is not difficult to verify that

$$\begin{aligned}
\int_x S_x^\delta(o_1) S_x^\delta(o_2) d\mu^\delta(x) &= \sum_{j_1, j_2} \int C_x^\delta(j_1) C_x^\delta(j_2) d\mu^\delta(x) e_{j_1}(o_1) e_{j_2}(o_2) \\
&= \sum_{j_1, j_2} \sum_{1 \le i_1, i_2 \le K(\delta)} C_{x(i_1,i_2)}^\delta(j_1) C_{x(i_1,i_2)}^\delta(j_2) d\mu^\delta(x(i_1, i_2)) e_{j_1}(o_1) e_{j_2}(o_2) \\
&= \sum_{1 \le j_1, j_2 \le K(\delta)} C_0(j_1, j_2) e_{j_1}(o_1) e_{j_2}(o_2) \ .
\end{aligned}$$

Since we have that

$$\begin{aligned}
\int_x S_x^\delta(o_1) S_x^\delta(o_2) d\mu^\delta(x) - \sum_{i_1, i_2} C_0(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2) \\
= \int_x S_x^\delta(o_1) S_x^\delta(o_2) d\mu^\delta(x) - \sum_{i_1, i_2 \le K(\delta)} C_0(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2) \\
- \sum_{i_1, i_2 > K(\delta)} C_0(i_1, i_2) e_{i_1}(o_1) e_{i_2}(o_2) \ ,
\end{aligned}$$

we have constructed an element of $\mathcal{S}_2(P) \subseteq T^{(2)}(P)$ which approximates h to a prescribed level of accuracy. Since $T^{(2)}(P)$ is closed, h must necessarily lie in $T^{(2)}(P)$, and so, $H \subseteq T^{(2)}(P)$. This therefore proves that $T^{(2)}(P) = H$.

In view of the representation of $T^{(2)}(P)$ derived above and the fact that $T^{(2)}(P)$ is itself a Hilbert space, for each x, we can write

$$\begin{aligned}
&\Pi \left\{ (o_1, o_2) \mapsto S_{1,x}(o_1) S_{2,x}(o_2) \mid T^{(2)}(P) \right\} (o_1, o_2) \\
&\quad = \textstyle\sum_{j_1, j_2} E_P \left[S_{1,x}(O_1) S_{2,x}(O_2) e_{j_1}(O_1) e_{j_2}(O_2) \right] e_{j_1}(o_1) e_{j_2}(o_2) \\
&\quad = \textstyle\sum_{j_1, j_2} E_P \left[S_{1,x}(O_1) e_{j_1}(O_1) \right] E_P \left[S_{2,x}(O_2) e_{j_2}(O_2) \right] e_{j_1}(o_1) e_{j_2}(o_2) \\
&\quad = \textstyle\sum_{j_1} E_P \left[S_{1,x}(O_1) e_{j_1}(O_1) \right] e_{j_1}(o_1) \sum_{j_2} E_P \left[S_{2,x}(O_2) e_{j_2}(O_2) \right] e_{j_2}(o_2) \\
&\quad = \Pi \left\{ o \mapsto S_{1,x}(o) \mid T(P) \right\} (o_1) \cdot \Pi \left\{ o \mapsto S_{2,x}(o) \mid T(P) \right\} (o_2) \ .
\end{aligned}$$

The lemma thus follows since the projection operator is linear.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Proof of Theorem 1

Theorem 1 is a special case of Theorem 2.

Proof of Lemma 2

This follows upon noting, using Fubini's theorem, that

$$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = (P_n - P_0)\left[D_1^{(2)}(P_n^*) + D_2^{(2)}(P_n^*)\right] + (P_n - P_0)^2 D^{(2)}(P_n^*)\ ,$$

where $D_1^{(2)}(P)(o) := \int D^{(2)}(P)(o_1, o)dP_0(o_1)$ and $D_2^{(2)}(P)(o) := \int D^{(2)}(P)(o, o_2)dP_0(o_2)$ for each $P \in \mathcal{M}$. Under conditions (i) and (ii), Lemma 19.24 of van der Vaart [2000] can be used to conclude that the first summand on the right-hand side of the expression above is $o_P(n^{-1/2})$. The result follows then from condition (iii).

Proof of Lemma 3

A proof can be found in Gill et al. [1995].

Proof of Theorem 2

Based on the assumed second-order expansion, we have that

$$\begin{aligned}\Psi(P_n^*) - \Psi(P_0) &= -P_0 D^{(1)*}(P_n^*) - \tfrac{1}{2}\lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*) + R_3(P_n^*, P_0)\\ &= -P_0 D^{(1)*}(P_n^*) - \tfrac{1}{2}P_0^2 D_{h_n}^{(2)}(P_n^*) + B_n(h_n) + R_3(P_n^*, P_0)\\ &= (P_n - P_0)D^{(1)*}(P_n^*) + \tfrac{1}{2}(P_n^2 - P_0^2)D_{h_n}^{(2)}(P_n^*) + B_n(h_n) + R_3(P_n^*, P_0)\ ,\end{aligned}$$

where we have used that $P_n^* \in \mathcal{M}$ and h_n are selected so that

$$P_n D^{(1)*}(P_n^*) = P_n^2 D_{h_n}^{(2)}(P_n^*) = 0\ .$$

The 2-TMLE algorithm we propose in this paper will produce such a P_n^*. Under condition 1, Lemma 19.24 of van der Vaart [2000] implies that $(P_n - P_0)D^{(1)*}(P_n^*) = (P_n - P_0)D^{(1)*}(P_0) + o_P(n^{-1/2})$. The result then follows directly in view of conditions 2, 3 and 4.

Proof of Lemma 4

The bias term $B_n(h) = \frac{1}{2}[P_0^2 D_h^{(2)}(P_n^*) - \lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*)]$ can be studied similarly as one studies the bias of a kernel density estimator. Defining pointwise

$$\begin{aligned}f_{1n}(w_1, w_2) &:= \bar{g}_n(w_1) - \bar{g}_0(w_2)\\ f_{2n}(w_1) &:= \frac{\bar{g}_0(w_1)}{\bar{g}_n^2(w_1)Q^{\circ}_{W,n}(w_1)}\left[\bar{Q}_0(w_1) - \bar{Q}_n^*(w_1)\right]\ ,\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

the identity $\frac{1}{2}P_0^2 D_h^{(2)}(P_n^*) = \iint K_h(w_1 - w_2) f_{1n}(w_1, w_2) f_{2n}(w_1) dP_{W,0}(w_1) dP_{W,0}(w_2)$ can be directly verified. Writing $h_{1n}(w_1, w_2) := f_{1n}(w_1, w_2) Q_{W,0}(w_2)$, we can express this as

$$\begin{aligned}
&\iint K_h(w_1 - w_2) f_{1n}(w_1, w_2) f_{2n}(w_1) dP_{W,0}(w_1) dP_{W,0}(w_2) \\
&\quad = \int \left[\int K_h(w_1 - w_2) h_{1n}(w_1, w_2) dw_2 \right] f_{2n}(w_1) dP_{W,0}(w_1) \\
&\quad = \int \left[\int K(u) h_{1n}(w_1, w_1 - uh) du \right] f_{2n}(w_1) dP_{W,0}(w_1) \\
&\quad = \iint K(u) \left[h_{1n}(w_1, w_1) + \sum_{\mathbf{m} \in A(m_0)} \prod_{j=1}^{d} \frac{(-1)^{m_j} (u_j h_j)^{m_j}}{m_j!} h_{1n}^{(\mathbf{m})}(w_1, w_1) \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \left. + O(h^{m_0+1}) \right] du f_{2n}(w_1) dP_{W,0}(w_1)
\end{aligned}$$

with $A(m_0) := \{(m_1, m_2, \ldots, m_d) \in \{0, 1, 2, \ldots, m_0\}^d : \sum_{j=1}^d m_j \leq m_0\}$ and

$$h_{1n}^{(\mathbf{m})}(w_1, w_1) := \left. \frac{\partial^{\sum_{j=1}^d m_j}}{\partial w_{21}^{m_1} \partial w_{22}^{m_2} \ldots \partial w_{2d}^{m_d}} h_{1n}(w_1, w_2) \right|_{w_2 = w_1},$$

where w_{2j} denotes the j^{th} component of w_2, provided $w_2 \mapsto h_{1n}(w_1, w_2)$ is $(m_0 + 1)$ times continuously differentiable at $w_2 = w_1$ on the support of W, assumed compact, and each of its $(m_0 + 1)^{th}$ order partial derivatives are uniformly bounded by a fixed $M < \infty$ with probability tending to one. We have used here the orthogonality of K to polynomials of all degrees up until m_0. It follows then that

$$\begin{aligned}
\frac{1}{2} P_0^2 D_h^{(2)}(P_n^*) &= \int h_{1n}(w_1, w_1) f_{2n}(w_1) dP_{W,0}(w_1) + O_P\left(h^{m_0+1} \|f_{2n}\|\right) \\
&= \frac{1}{2} \lim_{h \to 0} P_0^2 D_h^{(2)}(P_n^*) + O_P\left(h^{m_0+1} \|\bar{Q}_n^* - \bar{Q}_0\|\right),
\end{aligned}$$

where we have now used that $\bar{g}_n$ and $Q_{W,n}^{\circ}$ are bounded away from zero with probability tending to one.

Proof of Theorem 3

Since our 2-TMLE can be perceived as a particular 1-TMLE, when condition 5a holds, the result follows from the asymptotic efficiency and linearity of a 1-TMLE. Under conditions 1–4, $\bar{Q}_n^*$ will be sufficiently well-behaved as to allow a usual analysis involving Donsker theorems, even despite the additional targeting in 2-TMLE relative to a standard 1-TMLE. We omit further details in this simple case and instead focus on the case where condition 5b holds.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

We note that Lemma 4 allows us to write

$$\begin{aligned} R_2(P_n^*, P_0) &= -\tfrac{1}{2} \lim_{h \to 0} P_0^2 D_h^{(2)}(P_n^*) + R_3(P_n^*, P_0) \\ &= -\tfrac{1}{2} P_0 D_h^{(2)}(P_n^*) + O_P(h^{m_0+1} \|\bar{Q}_n^* - \bar{Q}_0\|) + R_3(P_n^*, P_0) \end{aligned}$$

under smoothness conditions on the distribution of (A, W), where

$$R_3(P, P_0) := P_0 \left[\left(\frac{Q_{W,0}\bar{g}_0}{Q_W \bar{g}} - 1 \right) \left(\frac{\bar{g} - \bar{g}_0}{\bar{g}} \right) (\bar{Q} - \bar{Q}_0) \right],$$

thereby giving us the expansion

$$\Psi(P_n^*) - \Psi(P_0) = -P_0 D^{(1)*}(P_n^*) - \frac{1}{2} P_0^2 D_{h_n}^{(2)}(P_n^*) + R_3(P_n^*, P_0) + O_P(h_n^{m_0+1} \|\bar{Q}_n^* - \bar{Q}_0\|) \ .$$

By construction, our 2-TMLE P_n^* solves the desired first- and second-order equations:

$$P_n D^{(1)*}(P_n^*) = 0 \ \text{ and } \ P_n^2 D_{h_n}^{(2)}(P_n^*) = 0 \ .$$

Thus, under condition 5a, we obtain that

$$\Psi(Q_n^*) - \Psi(Q_0) = (P_n - P_0) D^{(1)*}(P_n^*) + \frac{1}{2}(P_n^2 - P_0^2) D_{h_n}^{(2)}(P_n^*) + o_P(n^{-1/2}) \ .$$

If conditions 3 and 4 hold, $D^{(1)*}(P_n^*)$ is consistent in the sense that $P_0[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)]^2$ tends to zero in probability. Additionally, if condition 4 holds, $D^{(1)*}(P_n^*)$ has uniform sectional variation norm bounded by some fixed constant with probability tending to one. Thus, we can conclude by Lemma 19.24 of van der Vaart [2000] that $(P_n - P_0) D^{(1)*}(P_n^*) = (P_n - P_0) D^{(1)*}(P_0) + o_P(n^{-1/2})$. As per the V-statistic term, we can use Lemma 2 to write

$$(P_n^2 - P_0^2) D_{h_n}^{(2)}(P_n^*) = (P_n - P_0)^2 D_{h_n}^{(2)}(P_n^*) + o_P(n^{-1/2})$$

since we have that $o_1 \mapsto \int D_{h_n}^{(2)}(o_1, o_2) dP_0(o_2)$ and $o_2 \mapsto \int D_{h_n}^{(2)}(o_1, o_2) dP_0(o_1)$ both have uniformly bounded uniform sectional variation norm asymptotically under conditions 3 and 4 and hence fall in a P_0-Donsker class with probability tending to one. This is shown in a similar fashion as the fact that the variation norm of $w_2 \mapsto \int K_h(w_1 - w_2) dP_0(w_1) = \int K(u) p_0(uh + w_2) du$ is bounded uniformly in h. As such, integrating over each of the variables the kernel $K_h(w_1 - w_2)$ does not pose any problem anymore. The consistency condition is a straightforward consequence of conditions 3 and 4.

The term $(P_n - P_0)^2 D_{h_n}^{(2)}(P_n^*)$ is bounded above by $O_P(n^{-1} \|D_{h_n}^{(2)}(P_n^*)\|_v^*)$ according to Lemma 3. Under conditions 4 and 5, this uniform sectional variation norm is uniformly bounded by a constant multiple of the uniform sectional variation norm of $(w_1, w_2) \mapsto K_h(w_1 - w_2)$ with probability tending to one. The latter is bounded by $O(h^{-d})$. Indeed, for a typical smooth multivariate kernel with compact support, while differentiation with respect to the $2d$ components of $(w_1, w_2) \mapsto K_h(w_1 - w_2)$ will

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

generate a factor h^{-2d}, the integral of the absolute value of the resulting function, and therefore of the uniform sectional variation norm of interest, will be of order $O(h^{-d})$ since the support is assumed to be compact. This motivates condition 6 of the theorem. While this condition is sufficient, it is unlikely to be necessary and may possibly be weakened using alternative techniques of proof.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press