UC Berkeley
UC Berkeley Previously Published Works

Title
Statistical Inference for Data Adaptive Target Parameters

Permalink
https://escholarship.org/uc/item/7325s077

Journal
The International Journal of Biostatistics, 12(1)

ISSN
2194-573X

Authors
Hubbard, Alan E
Kherad-Pajouh, Sara
van der Laan, Mark J

Publication Date
2016-05-01

DOI
10.1515/ijb-2015-0013

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Alan E. Hubbard*, Sara Kherad-Pajouh and Mark J. van der Laan

Statistical Inference for Data Adaptive Target Parameters

DOI 10.1515/ijb-2015-0013

Abstract: Consider one observes n i.i.d. copies of a random variable with a probability distribution that is known to be an element of a particular statistical model. In order to define our statistical target we partition the sample in V equal size sub-samples, and use this partitioning to define V splits in an estimation sample (one of the V subsamples) and corresponding complementary parameter-generating sample. For each of the V parameter-generating samples, we apply an algorithm that maps the sample to a statistical target parameter. We define our sample-split data adaptive statistical target parameter as the average of these V-sample specific target parameters. We present an estimator (and corresponding central limit theorem) of this type of data adaptive target parameter. This general methodology for generating data adaptive target parameters is demonstrated with a number of practical examples that highlight new opportunities for statistical learning from data. This new framework provides a rigorous statistical methodology for both exploratory and confirmatory analysis within the same data. Given that more research is becoming "data-driven", the theory developed within this paper provides a new impetus for a greater involvement of statistical inference into problems that are being increasingly addressed by clever, yet ad hoc pattern finding methods. To suggest such potential, and to verify the predictions of the theory, extensive simulation studies, along with a data analysis based on adaptively determined intervention rules are shown and give insight into how to structure such an approach. The results show that the data adaptive target parameter approach provides a general framework and resulting methodology for data-driven science.

Keywords: asymptotic linearity, clustering, cross-validation, data mining, influence curve, loss-function, risk, machine learning, sample splitting, sub-group analysis, super-learner, targeted maximum likelihood estimation

1 Introduction

A proliferation of statistical/data science methods have accompanied a growing systematic collection of data across many scientific fields. Progress has been made developing quantitative statistical methods well-suited to exploratory analysis, however, much remains much to be done for deriving estimators and robust inference of relevant parameters in such a context. The growth of fields such as precision medicine and high dimensional (high throughput) biology try to capitalize on the resulting "big data" by inspired pattern-finding procedures [1]; less emphasis has been given to formally defining the parameters such procedures "discover". Thus, an obvious first step necessary for driving theoretical results is to explicitly define such data adaptive parameters. The goal of previous work [2] and this paper is address the issue of rigorous inference when the target parameter is not pre-specified.

We note that there is a literature on the dangers of deriving parameters data-adaptively. The common wisdom for deriving consistent inference for a data-adaptively defined parameter is to use sample-splitting, where one of the splits is a training set used to define the parameter, and the left out then estimates this parameter on the independent "estimation" sample. Quoting Dwork et al. [3]:

> The "textbook" advice for avoiding problems of this type is to collect fresh samples from the same data distribution whenever one ends up with a procedure that depends on the existing data. Getting fresh data is usually costly and often impractical so this requires partitioning the available dataset randomly into two or more disjoint sets of data (such as a

***Corresponding author: Alan E. Hubbard,** Division of Biostatistics, University of California, Berkeley, CA, USA, E-mail: hubbard@berkeley.edu
Sara Kherad-Pajouh, Mark J. van der Laan, Division of Biostatistics, University of California, Berkeley, CA, USA

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

training and testing set) prior to the analysis. Following this approach conservatively with m adaptively chosen procedures would significantly (on average by a factor of m) reduce the amount of data available for each procedure.

Our main proposed approach aims to keep the data-adaptive part of the sample splitting algorithm described in quote, but to define an average of the data-adaptive parameter across arbitrary splits of this sort (we emphasize V-fold cross-validation below). In this way, one can still use the power of the entire dataset while avoiding strong conditions on the algorithms used to data-adaptively define the parameters.

1.1 Motivating example

The general methodology for estimation and inference for data adaptive parameters is presented below, however, we illustrate the method by a particularly challenging causal inference estimation problem. Consider data from the (W)estern (C)collaborative (G)roup (S)tudy [4], a prospective study of risk factors of coronary heart disease (CHD). The study consisted of 3,524 males (3,142 of which had complete data) aged 39–59, working in certain California corporations, who were enrolled at the outset of the study and followed for 8.5 years. Our goal is to estimate the impact on CHD from applying a treatment rule regarding cholesterol that is learned from the data. Define the data of interest to be $O=(W, A, Y)$, where W are a vector of confounders, A the treatment of interest (say cholesterol level) which is dichotomized, so that $A=I(Chol>\gamma)$, where γ is a target intervention level; Y is the indicator of a CHD event within the study period. Let $\bar{Q}_0(A, W)\equiv E_0(Y|A, W)$. We start with the ambitious goal of estimating a treatment rule for intervening on cholesterol that targets only those people that would be helped by such an intervention. To do so, we start by defining so-called counterfactuals [5], Y_a, which represents the outcome for a subject if, possibly contrary to fact, the subject had level $A=a$. This leads to a notion of potentially "full" data that includes counterfactuals, or in this case, $X=(W, A, Y_1, Y_0)$. Our goal is to estimate the impact of an intervention rule that lowers a subjects cholesterol (from $A=1$ to $A=0$) only if such a change improves the CHD outcome (that is, lowers cholesterol if $Y_1=1$ and $Y_0=0$ and $A=1$), or:

$$E\{Y-Y_{d_X},\} \tag{1}$$

with rule $d_X(Y, Y_0)=I(Y<Y_0)$. Beyond the assumptions of randomization, positivity, consistency [6], this parameter would be identifiable only under strong assumptions on the joint distribution of counterfactuals. Thus, we consider a less ambitious parameter, based on *average* impacts of intervention, conditional on W. In this case, the parameter measures the impact of only targeting those individuals that "significantly" benefit from intervening on those with high cholesterol $(A=1)$, or:

$$E\{Y-Y_{d_{0,\tau,Q}}\} \tag{2}$$

where $\bar{Q}_0(A, W)\equiv E_0(Y|A, W)$ and $(A, W)\rightarrow d_{0,\tau,\bar{Q}_0}(A, W)=I(\bar{Q}_0(A, W)-\bar{Q}_0(0, W)<\tau)$. In comparing the same individuals in a population before and after this rule is imposed, their A will stay as it was unless the original $A=1$ *and* $\bar{Q}_0(1, W)-\bar{Q}_0(0, W)>\tau$. Though this parameter is identifiable without the stronger assumptions necessary to identify (1), it still requires the strong assumption of estimating $\bar{Q}_0$ at a particular rate. Thus, we finally consider examining the impact of an empirically derived rule:

$$E\{Y-Y_{d_{n,\tau,\bar{Q}n}}\} \tag{3}$$

where the rule, $(A, W)\rightarrow d_{n,\tau,\bar{Q}_n}(A, W)$ is as above, but with the true mean function being replace by some empirically derived estimate, $\bar{Q}_n(A, W)$; it is a data adaptive parameter, as the data is used to define the parameter.

We will discuss several specific examples below, but note that the sequence of analyses often used in large scale omic studies (genomics, proteomics, metabolomics, etc.; Zhang and Chen [7], Berger et al. [8]) can be the result of a series of suggested patterns that lead to further analyses not previously considered, e. g., multiple testing, to clustering, to exploration of pathways, to more targeted analyses all with the data

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

from the same experiment. In these cases, inference that ignores that the parameters were derived data adaptively will typically be biased. Others have noted the particular dangers of high dimensional data combined with flexible methodologies to generate excessive false positive findings (Ioannidis [9], Broadhurst and Kell [10]). In many cases, even when the best intentions are to stick to a pre-specified data analysis plan, there can be feedback from the data and models chosen (e. g., covariates dropped, different basis functions tried, unplanned sub-group analyses conducted, etc.; Barraclough and Govindan [11], Marler [12]). Thus, it is important to have methods that allow such exploration and also transparent interpretation of the resulting estimates. Though there are advantages for pre-specifying the algorithm used to generate the parameter(s), the general methodology does not even require one to – that is, one can derive inference for methods of deriving patterns, even when the precise methods used to generate the parameters are not known. Thus, it can be applied in circumstances where there is little constraint on how the data is explored to generate potential parameters of interest for estimation and inference.

2 Methodology

The following is also presented in Hubbard and van der Laan [2]. Consider observed data $O_1, \ldots, O_n$, i.i.d. with probability distribution P_0, within statistical model $\mathcal{M}$. Let $B_n \in \{0,1\}^n$ be a random vector of binaries, independent of $(O_1, \ldots, O_n)$, that defines a random split into an estimation-sample $\{O_i : B_n(i) = 1\}$ and parameter-generating sample $\{O_i : B_n(i) = 0\}$. For simplicity, assume that B_n corresponds with V-fold cross-validation scheme, i. e., 1) $\{1, \ldots, n\}$ are divided in V equal size subgroups, 2) an estimation-sample is defined by one of the subgroups, 3) the parameter-generating sample is its complement resulting in V such splits of the sample. Thus, in this case B_n has only V possible values.

Given random split B_n, P^0_{n,B_n} is the empirical distribution of the parameter-generating sample, and P^1_{n,B_n} the empirical distribution of the estimation-sample. For a given B_n, $\Psi_{B_n,P^0_{n,B_n}} : \mathcal{M} \to \mathbb{R}$ is the target parameter mapping indexed by the parameter-generating sample P^0_{n,B_n}, and $\hat{\Psi}_{B_n,P^n_0,B_n} : \mathcal{M}_{NP} \to \mathbb{R}$ the corresponding estimator of this target parameter. Here $\mathcal{M}_{NP}$ is the nonparametric model and an estimator is defined as a mapping/algorithm from a nonparametric model, including the empirical distributions, to the parameter space. For simplicity, assume that the parameter is real-valued. Thus, the target parameter mapping and estimator can depend not only on parameter-generating-sample P^0_{n,B_n}, but also on the particular split B_n.

The choice of target parameter mapping and corresponding estimator can be informed by the data P^0_{n,B_n} and split B_n, but not by the estimation-sample P^1_{n,B_n}. One does not need to assume the mapping from the parameter-generating sample to the space of target parameter mappings and estimators is known, but one need only to know its realization $(\Psi_{B_n,P^0_{n,B_n}}, \hat{\Psi}_{B_n,P^0_{n,B_n}})$. Define the sample-split data adaptive statistical target parameter as $\Psi_n : \mathcal{M} \to \mathbb{R}$ with

$$\Psi_n(P) = E_{B_n} \Psi_{B_n,P^0_{n,B_n}}(P)$$

and the statistical estimand of interest is thus

$$\psi_{n,0} = \Psi_n(P_0) = E_{B_n} \Psi_{B_n,P^0_{n,B_n}}(P_0).$$

This parameter mapping depends on the data and thus it is called a *data adaptive target parameter*. A corresponding estimator of the estimand $\psi_n, 0$ is:

$$\psi_n = \hat{\Psi}(P_n) = E_{B_n} \hat{\Psi}_{B_n,P^0_{n,B_n}}(P^1_{n,B_n}).$$

The goal is to prove that $\sqrt{n}(\psi_n - \psi_{n,0})$ converges in distribution to mean zero normal distribution with variance σ^2 that can be consistently estimated, allowing the construction of confidence intervals for $\psi_{n,0}$ and also allow testing a null-hypothesis such as $H_0 : \psi_{n,0} \leq 0$. This holds if $\psi_n = \hat{\Psi}(P_n)$ is an asymptotically linear estimator of $\psi_{n,0}$ with influence curve $IC(P_0)$:

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

$$\psi_n - \psi_{n,0} = (P_n - P_0)IC(P_0) + o_P(1/\sqrt{n});$$

the notation $Pf \equiv \int f(o)dP(o)$ is used for the expectation of f (O) w.r.t. P. Since $(P_n - P_0)$ $IC(P_0) = 1/n \sum_i IC(P_0)(O_i)$ is the sum of mean zero independent random variables, the asymptotic linearity implies that $\sqrt{n}(\psi_n - \psi_{n,0})$ converges to a mean zero normal distribution with variance $\sigma^2 = P_0 IC(P_0)^2$.

Theorem 1 *Suppose that, given* (B_n, P^0_{n,B_n}), $\hat{\Psi}_{B_n, P^0_{n,B_n}}$ *is an asymptotically linear estimator of* $\Psi_{B_n, P^0_{n,B_n}}(P_0)$ *at* P_0 *with influence curve* $IC_{B_n, P^0_{n,B_n}}(P_0)$ *indexed by*(B_n, P^0_{n,B_n}):

$$\hat{\Psi}_{B_n, P^0_{n,B_n}}(P^1_{n,B_n}) - \Psi_{B_n, P^0_{n,B_n}}(P_0) = (P^1_{n,B_n} - P_0)IC_{B_n, P^0_{n,B_n}}(P_0) + R_{n,B_n},$$

where (unconditional) $R_{n,B_n} = o_P(1/\sqrt{n})$. *Assuming V-fold cross-validation, and for a given split* $B_n = v$, *assume that* $P_0 IC^2_{v, P^0_{n,v}}(P_0) - P_0 IC_v(P_0))^2 \to 0$ *in probability, where* $IC_v(P_0)$ *is a limit influence curve that can still be indexed by the split* v.

Then, $\sqrt{n}(\psi_n - \psi_{n,0}) = \frac{1}{V}\sum_v \sqrt{V}\sqrt{n/V}(P^1_{n,B_n} - P_0)IC_{v, P^0_{n,v}}(P_0) + o_P(1/\sqrt{n})$ *converges to a mean zero normal distribution with variance*

$$\sigma^2 = \frac{1}{V}\sum_{v=1}^{V} \sigma_v^2,$$

where $\sigma_v^2 = P_0 IC_v^2(P_0)$. *A consistent estimator of* σ^2 *is given by*

$$\sigma_n^2 = \frac{1}{V}\sum_{v=1}^{V} P_n IC^2_{v,n},$$

where $IC_{v,n}$ *is an* $L^2(P_0)$*-consistent estimator of* $IC_v(P_0)$. *Alternatively, one can use,*

$$\sigma_n^2 = \frac{1}{V}\sum_{v=1}^{V} P^1_{n,v} IC_{v, P^0_{n,v}}(P^0_{n,v})^2, \tag{4}$$

where $IC_{v, P^0_{n,v}}(P^0_{n,v})$ *is an* $L^2(P_0)$*-consistent estimator of* $IC_{v, P^0_{n,v}}(P_0)$*based on the sample* $P^0_{n,v}$.

The latter variance estimator avoids finite sample bias by using sample splitting and might therefore be preferable in finite samples. The proofs of theorems are provided in the Supplemental Material.

Asymptotic equivalence of standardized estimator and standardized oracle estimator Suppose that the algorithm $(B_n, P^0_{n,B_n}) \to (\Psi_{B_n, P^0_{n,B_n}}, \hat{\Psi}_{B_n, P^0_{n,B_n}})$ that maps the data and choice of sample split into an estimator and target-parameter mapping does not depend on the particular split B_n. This would be true if, for instance, a fixed algorithm was used to generate target parameters. In that case, the influence curve $IC_{B_n, P^0_{n,B_n}}(P_0)$, conditional on the parameter-generating sample P^0_{n,B_n} and split B_n, will converge to a fixed IC (P_0), which does not depend on the split. In this important case, the estimator ψ_n of $\psi_{n,0}$ is asymptotically linear with influence curve $IC(P_0)$, which is the influence curve of the estimator $\hat{\Psi}_{P_0} : \mathcal{M}_{NP} \to \mathbb{R}$ of the target parameter $\Psi_{P_0} : \mathcal{M} \to \mathbb{R}$, treating P_0 as known, leading to the limit-variance:

$$\sigma^2 = P_0 IC(P_0)^2.$$

In addition, the standardized estimator $\sqrt{n}(\psi_n - \psi_{0,n})$ has the same asymptotic variance as the standardized "oracle" estimator $\sqrt{n}(\hat{\Psi}_{P_0}(P_n) - \hat{\Psi}_{P_0}(P_0))$ (that is an estimator of an a priori specified parameter, as opposed to a data adaptive one) one would have used for the parameter $\Psi_{P_0}(P_0)$ if the parameter mapping Ψ_{P_0} is treated as known. Even though there was no loss in efficiency relative to this oracle procedure $\hat{\Psi}_{P_0}(P_n)$, we should note that this asymptotic variance is measured relative to a different target $E_{B_n}\Psi_{P^0_{n,B_n}}(P_0)$ instead of $\hat{\Psi}_{P_0}(P_0)$. Finally, we provide heuristics for choosing the number of splits in Supplemental Material.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

2.1 Splitting the sample, but using the whole sample to fit the data adaptively generated target parameter

In the above Theorem 1, one need not assume Donsker class conditions, so that the target-parameter choices $\Psi_{B_n, P^0_{n,B_n}}$ could be arbitrarily dependent on the data P^0_{n,B_n}. However, now consider an estimator $\psi^1_n \equiv E_{B_n}\hat{\Psi}_{B_n, P^0_{n,B_n}}(P_n)$ of the same "estimand" $\psi_{0,n}$ but which uses the entire sample as the estimation sample for each of the V parameter-generating samples. The asymptotics will now rely on stronger assumptions, but if the algorithm generating the target parameter and estimator is different across splits, and the stronger assumptions are satisfied, then the estimator is generally more efficient than the algorithm based on theorem 1.

Theorem 2 *As above assume that conditional on* (B_n, P^0_{n,B_n}), $\hat{\Psi}_{B_n, P^0_{n,B_n}}$ *is asymptotically linear with influence curve* $IC_{B_n, P^0_{n,B_n}}(P_0)$ *so that*

$$\hat{\Psi}_{B_n, P^0_{n,B_n}}(P_n) - \Psi_{B_n, P^0_{n,B_n}}(P_0) = (P_n - P_0)IC_{B_n, P^0_{n,B_n}}(P_0) + R_{n,B_n},$$

where (unconditionally) $R_{n,B_n} = o_P(1/\sqrt{n})$. *Also, as in Theorem 1, for a given split* $B_n = v$, *assume that* $P_0 IC^2_{v, P^0_{n,v}}(P_0) - P_0 IC_v(P_0))^2 \to_{n\to\infty} 0$ *in probability, where* $IC_v(P_0)$ *is a limit that can still be indexed by the split* v.

We also assume that $IC_{v, P^0_{n,v}}(P_0)$ falls in a P_0-Donsker class with probability tending to 1.

Then,

$$\psi^1_n - \psi_{n,0} = (P_n - P_0)IC(P_0) + o_P(1/\sqrt{n}),$$

where

$$IC(P_0) \equiv \frac{1}{V}\sum_{v=1}^{V} IC_v(P_0)$$

is an average of the B_n*-specific influence curves. Thus,* $\sqrt{n}(\psi^1_n - \psi_{n,0})$ *converges to a mean zero normal distribution with variance*

$$\sigma^2_1 = P_0\left\{\frac{1}{V}\sum_{v} IC_v(P_0)\right\}^2.$$

The relative efficiency of the two estimators ψ_n and ψ^1_n is of course based on the two corresponding asymptotic variances

$$\sigma^2 = \frac{1}{V}\sum_{v=1}^{V}\sigma^2_v \text{ and } \sigma^2_1 = \frac{1}{V^2}\sum_{v_1, v_2} P_0\{IC_{v_1}(P_0)IC_{v_2}(P_0)\}.$$

In the special case that $IC_v = IC$ does not depend on the split v (i. e., the algorithm generating a target parameter and estimator is the same for each split), then $\sigma^2 = \sigma^2_1$. In the other extreme case that $P_0 IC_{v1} IC_{v2} = 0$ for $v_1 \neq v_2$, $\sigma^2 = 1/V\sum_v \sigma^2_v$ and $\sigma^2_1 = \frac{1}{V^2}\sum_v \sigma^2_v$. Thus, in the latter case $\sigma^2 = V\sigma^2_1$ and one can conclude that if the selected target parameters across the V parameter-generating samples are highly correlated, then the estimator ψ_n is almost as efficient as ψ^1_n, but if the selected target parameters across different sample splits are highly *independent/orthogonal*, then a very significant loss in efficiency up till a factor V can occur. This efficiency comparison does not take into account that ψ_n is asymptotically normally distributed under significantly weaker conditions than the conditions needed for asymptotic linearity of ψ^1_n, so that there will be cases under which the model required for asymptotic normality of ψ_n holds, but the analogue model for ψ^1_n fails to hold. This comparison also does not take into account that ψ^1_n should have better second order term behavior than ψ_n for non-linear estimators, since ψ^1_n involves using the full sample for each of the data adaptively generated target parameters.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

2.2 Using the whole sample to generate the target parameter and to subsequently estimate it: no sample splitting

Consider a mapping $P_n \rightarrow (\Psi_{P_n}, \hat{\Psi}_{P_n})$ from a sample to a target parameter mapping $\Psi_{P_n} : \mathcal{M} \rightarrow \mathbb{R}$ and corresponding estimator $\hat{\Psi}_{P_n} : \mathcal{M}_{NP} \rightarrow \mathbb{R}$. The estimand of interest is now $\Psi_{P_n}(P_0)$ and it is estimated with $\psi_n^2 = \hat{\Psi}_{P_n}(P_n)$. The possible advantage of this approach is that the estimand is a single parameter instead of an average over splits of sample-split-specific estimands, and the latter might be harder to interpret. However, as in the previous subsection, stronger conditions are needed to establish the desired asymptotic consistency and normality. In contrast to the method of the previous subsection, in which we only changed the estimator, we now actually changed the estimand as well.

Theorem 3 *Assume* $\hat{\Psi}_P(P_n)$ *is an asymptotically linear estimator of* $\Psi_P(P_0)$ *at* P_0 *with influence curve* $IC_P(P_0)$ *uniformly in the choice of parameter P in the following sense*:

$$\hat{\Psi}_{P_n}(P_n) - \hat{\Psi}_{P_n}(P_0) = (P_n - P_0)IC_{P_n} + R_n,$$

where $R_n = o_P(1/\sqrt{n})$. *In addition, assume* $P_0(IC_{P_n}(P_0) - IC_{P_0}(P_0))^2 \rightarrow 0$ *in probability and* $IC_{P_n}(P_0)$ *is an element of a* P_0*-Donsker class with probability tending to 1. Then*,

$$\hat{\Psi}_{P_n}(P_n) - \hat{\Psi}_{P_n}(P_0) = (P_n - P_0)IC_{P_0}(P_0) + o_P(1/\sqrt{n}),$$

and thus $\sqrt{n}(\psi_n^2 - \hat{\Psi}_{P_n}(P_0))$ *is asymptotically normally distributed with mean zero and variance* $\sigma^2 = P_0 IC_{p0}(P_0)$.

Again, this estimator ψ_n^2 is as efficient as the oracle estimator $\hat{\Psi}_{P_0}(P_n)$ as an estimator of $\Psi_{P_0}(P_0)$, discussed above, but one should note again that its efficiency is measured relative to a different target $\Psi_{P_n}(P_0)$ instead of $\Psi_{P_0}(P_0)$. Since the parameter Ψ_{P_0} is unknown while Ψ_{P_n} is a known target parameter mapping, one might often find the parameter $\Psi_{P_n}(P_0)$ more tangible than $\Psi_{P_0}(P_0)$, and thus perhaps easier to interpret. In essence, theorem 3 provides the conditions necessary for consistent inference of a "data-dredging" algorithm; using the same data for generating the parameter of interest, and deriving its inference.

Note, that others have examined the asymptotic of such a procedure (no sample splitting) using different theoretical approaches. For instance, Dwork et al. [3] presents asymptotic consistency for estimating a number (m) of data-adaptively derived functions of the data-generating distribution as a function of m and sample size, n.

3 Examples

In this section we showcase a few examples to demonstrate the proposed procedures for generating statistical target parameters and corresponding estimators and confidence intervals. For longer list of examples, see Supplemental Material.

3.1 Inference for the sample-split conditional risk of a data adaptive regression estimator

The set-up is identical to Dell et al. [13], Dudoit and van der Laan [14]. Let $O = (W, Y) \sim P_0$, where W is a vector of input-variables and Y is an outcome one wants to predict; P_0 is composed of the distribution of $Y|W$, $Q_0(W)$ and the distribution of W, $Q_{0,W}$. Let $\hat{Q}$ be an estimator of the true regression function $\bar{Q}_0 = E_0(Y|W)$; let $\bar{Q}_{P^0_{n,B_n}}$ be the corresponding estimate of $\bar{Q}_0 = E_0(Y|W)$ based on the parameter-generating

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

sample P^0_{n,B_n}. The target parameter generated by P^0_{n,B_n} is defined as the mean squared error $\Psi_{P^0_{n,B_n}}(P_0) = E_0(Y - \bar{Q}_{P^0_{n,B_n}}(W))^2$ or, in general, as the loss-function specific risk $E_0 L(\bar{Q}_{P^0_{n,B_n}})(W, Y)$ for some loss function $L(\bar{Q})$ satisfying $\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(\bar{Q})$.

The estimator of $\Psi_{P^0_{n,B_n}}(P_0)$ based on the estimation sample P^1_{n,B_n} is defined as its empirical counterpart $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}) = P^1_{n,B_n} L(\bar{Q}_{P^0_{n,B_n}})$. Conditional on the sample P^0_{n,B_n}, this estimator $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})$ is asymptotically linear with influence curve $L(\bar{Q}_{P^0_{n,B_n}}) - P_0 L(\bar{Q}_{P^0_{n,B_n}})$ with no remainder. The average across sample-split data adaptive target parameters is thus defined as $\psi_{n,0} = E_{B_n} P_0 L(\bar{Q}_{P^0_{n,B_n}})$ and its corresponding estimators are $\psi_n = E_{B_n} P^1_{n,B_n} L(\bar{Q}_{P^0_{n,B_n}})$, $\psi^1_n = E_{B_n} P_n L(\bar{Q}_{P^0_{n,B_n}})$, and $\psi^2_n = P_n L(\bar{Q}_{P_n})$. Theorem 1 implies that if the loss function chosen is uniformly bounded and the estimator $\hat{\bar{Q}}(P_n)$ is consistent for a limit $\bar{Q}$ (not necessarily $\bar{Q}_0$), then $\psi_n - \psi_{n,0}$ is asymptotically linear with influence curve $L(\bar{Q}) - P_0 L(\bar{Q})$, the same influence curve as the estimator $P_n L(\bar{Q})$ of $P_0 L(\bar{Q})$ treating $\bar{Q}$ as known. This allows us to construct a confidence interval for the true conditional risk $\psi_{n,0}$, under these very weak conditions. In particular, the estimator $\hat{\bar{Q}}$ can be a highly data adaptive super learner van der Laan et al. [15].

Similarly, Theorem 2 implies a formal result for ψ^1_n, but now $L(\bar{Q}_{P_n})$ has to be an element of a P_0-Donsker class with probability tending to 1, putting some constraints on how adaptive $\bar{Q}_{P_n}$ can be. Under the same conditions, we will have that $\psi^2_n = P_n L(\bar{Q}_{P_n})$ is an asymptotically linear estimator of $P_0 L(\bar{Q})$ with the same influence curve ψ_n Even though these conditions might be satisfied for $\bar{Q}_n$, the estimator ψ^2_n is known to be wrong for the sake of using $P_n L(\bar{Q}_{P_n})$ to select among a collection of candidate estimators of $\bar{Q}_0$ since this estimator of risk will favor over-fitted estimators. Nonetheless, if the goal is to obtain confidence intervals for the asymptotic risk $P_0 L(\bar{Q}_{P_n})$ of an estimator $\bar{Q}_{P_n}$, then this method could be considered.

3.2 Inference for sample-split subgroup-specific causal effect, where the subgroups are data adaptively determined

Consider "discovering" sub-groups within the target population that have unique relationships with explanatory variable of interest (e. g., drug treatment, environmental exposure, etc.). In the case that these sub-groups are not defined apart from the data, post-hoc sub-group analysis is typically treated as purely explanatory and thus the statistical inference inherently awed, typically anti-conservatively. However, the approach we have outlined provides explicit framework for aggressively searching for interesting sub-groups, but still allows for consistent statistical inference for the resulting estimators of association parameters.

Suppose that we observe on each subject $O = (W, A, Y)$, where W are baseline covariates, A is a binary treatment, and Y a final outcome. Thus we observe n i.i.d. copies $O_1, \ldots, O_n$, and consider an algorithm that maps a data set $O_1, \ldots, O_n$ into a subgroup $W \to C(W) \in \{0, 1\}$, where $C(W) = 1$ indicates membership in the subgroup. Denote this subgroup-estimator with $\hat{C}: \mathcal{M}_{NP} \to \mathcal{C}$, where $\mathcal{C}$ is the space of functions that map a W into a binary indicator. Given a realized subgroup C, let $\Psi_C: \mathcal{M} \to \mathbb{R}$ be a desired parameter of interest such as the W-controlled effect of treatment A on Y for subgroup C, defined as

$$\Psi_C(P_0) = E_0\{E_0(Y|A = 1,\ W,\ C(W) = 1) - E_0(Y|A = 0,\ W,\ C(W) = 1)|C(W) = 1\}.$$

Let $\hat{\Psi}_C: \mathcal{M}_{NP} \to \mathbb{R}$ be an estimator of $\Psi_C(P_0)$, again where one could choose among several different estimators (including the IPTW; Robins et al. [16]), but we focus on TMLE [17, 18]: note that this is just the targeted maximum likelihood estimator for the W-controlled effect of treatment but applied to the subsample $\{i: C(W_i) = 1\}$. Assume that the regularity conditions hold so that this TMLE $\Psi_C(P_n)$ is asymptotically linear with influence curve $IC_C(P_0)$:

$$\hat{\Psi}_C(P_n) - \Psi_C(P_0) = (P_n - P_0) IC_C(P_0) + R_{C,n},$$

where $R_{C,n} = o_P(1/\sqrt{n})$.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

Define $\Psi_{P^0_{n,B_n}}: \mathcal{M} \to \mathbb{R}$ as $\Psi_{P^0_{n,B_n}} = \Psi_{\hat{C}(P^0_{n,B_n})}$, i. e., the W-controlled effect of treatment on the outcome for the data adaptively determined subgroup $\hat{C}(P^0_{n,B_n})$. Similarly, we define $\hat{\Psi}_{P^0_{n,B_n}}: \mathcal{M}_{NP} \to \mathbb{R}$ as $\hat{\Psi}_{P^0_{n,B_n}} = \hat{\Psi}_{\hat{C}(P^0_{n,B_n})}$, i. e. the TMLE of the W-controlled effect of treatment on the outcome for this data adaptively determined subgroup, treating the latter as given. The estimand of interest is thus defined as $\psi_{n,0} = E_{B_n}\Psi_{P^0_{n,B_n}}(P_0)$ and its estimator is $\psi_n = E_{B_n}\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})$. That is, for a given split B_n, we use the parameter-generating sample P^0_{n,B_n} to generate a subgroup $\hat{C}(P^0_{n,B_n})$ and corresponding TMLE of $\hat{\Psi}_{\hat{C}(P^0_{n,B_n})}(P_0)$ applied to the estimation-sample P^1_{n,B_n}, and these sample-split specific estimators are averaged across the V sample splits. By assumption we have for each split B_n

$$\hat{\Psi}_{\hat{C}(P^0_{n,B_n})}(P^1_{n,B_n}) - \Psi_{\hat{C}(P^0_{n,B_n})}(P_0) = (P^1_{n,B_n} - P_0)IC_{\hat{C}(P^0_{n,B_n})}(P_0) + R_{\hat{C}(P^0_{n,B_n})},$$

where we now assume that (unconditionally) $R_{\hat{C}(P^0_{n,B_n})} = o_P(1/\sqrt{n})$. In addition, we assume that $P_0\{IC_{\hat{C}(P^0_{n,B_n})}(P_0)\}^2$ converges to $P_0\{IC_{\hat{C}(P_0)}(P_0)\}^2$ for a limit subgroup $\hat{C}(P_0)$. Application of Theorem 1 now proves that $\psi_n - \psi_{n,0}$ is asymptotically linear with influence curve $IC_{\hat{C}(P_0)}(P_0)$ so that it is asymptotically normally distributed with mean zero and variance $\sigma^2 = P_0 IC_{\hat{C}(P_0)}(P_0)^2$.

Under the Donsker class condition on $IC_{\hat{C}(P^0_{n,B_n})}(P_0)$ we can also establish the formal results for $\psi^1_n = E_{B_n}\hat{\Psi}_{\hat{C}(P^0_{n,B_n})}(P_n)$ of $\psi_{n,0}$, and the estimator $\psi^2_n = \hat{\Psi}_{\hat{C}(P_n)}(P_n)$ of $\Psi_{\hat{C}(P_n)}(P_0)$, respectively.

4 Simulations

Simulations for different algorithms producing the data adaptive target parameters were examined for performance among the three different algorithms based on theorems 1, 2 and 3 (referred to as algorithms 1, 2 and 3). The following step provides the structure of the algorithm, but also provides some basis of understanding the data adaptive parameter.

1. Generate a random sample from the data generating distribution of size n and break into V equal size estimation samples of size $n_V = n/V$ with corresponding parameter generating samples of size $n - n/V$.(
2. For each parameter-generating sample, apply the data-adaptive algorithm to define the parameter to be estimated on the corresponding estimation sample, which defines $\Psi_{P^0_{n,B_n}}$. For instance, fit a data-adaptive regression procedure estimating the mean of outcome Y based on predictors W, say $\bar{Q}_{P^0_{n,B_n}}(W)$, and define the target parameter as the risk based on squared error loss defined as $\Psi_{P^0_{n,B_n}}(P_0) = E_{P_0}(Y - \bar{Q}_{P^0_{n,B_n}}(W))^2$, treating $\bar{Q}_{P^0_{n,B_n}}$ as fixed and known.
3. For each of the V estimation samples, estimate the data adaptive parameter. For example, in the case of the risk example described in 2., $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}) = E_{P^1_{n,B_n}}(y - \bar{Q}_{P^0_{n,B_n}}(W))^2$. In addition, derive the influence curve $IC_{B_n,P^0_{n,B_n}}(\cdot)$ of this estimator for each of the sample-splits.
4. To derive the value of the true parameter corresponding to each parameter-generating sample, we draw a very large sample using the same distribution, representing a target population (P_0). This is used to evaluate $\Psi_{P^0_{n,B_n}}(P_0) = E_0(y - \bar{Q}_{P^0_{n,B_n}}(W))^2$, where P_0 is approximated by this empirical probability distribution of this very large sample (100,000).
5. Estimate the asymptotic variance (4) of ψ_n based on the sample variance within estimation samples of $IC_{B_n,P^0_{n,B_n}}(\cdot)$ (see Theorem 1 above), and construct a corresponding Wald-type confidence interval.
6. Repeat 1–5 for 1,000 simulations, examine the distribution of standardized differences, $\sqrt{n}(\psi_n - \psi_{n,0})$, and determine the coverage probabilities for the confidence intervals.

The modifications for algorithms 2 and 3 follow from the respective theorems.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

4.1 Risk estimation of a data adaptive prediction algorithm

More background on the *conditional* risk parameter is in Section 3.1, and in this case the goal is estimation and inference regarding the "fit" of a machine learning algorithm. The data is $O=(Y,W)$, for outcome Y, predictor W, where $W\sim N(0,\sigma_W^2=4)$, $\bar{Q}_0(W)\equiv E_0(Y|W)$ is shown in Figure 1, based on a piecewise constant model, and $Y|W\sim N(\bar{Q}_0(W),\sigma_Y^2=0.25)$. For the v – th parameter-generating sample, we fit the regression with an ensemble stacking algorithm, called the SuperLearner (SL; van der Laan et al. [15]), resulting in a convex combination of a variety of algorithms ranging from very smooth to highly data adaptive: linear model, stepwise regression based on AIC (`stepAIC`; Venables and Ripley [19]), Bayesian glm (linear) model (`bayesglm`; Gelman et al. [20]), generalized additive model with smooth term for covariate Hastie and Tibshirani [21]; neural nets (nnet; Venables and Ripley [19]); and a simple null model (sample average of outcome). For the v – th parameter-generating sample the data adaptive parameter of interest was defined as the conditional risk (mean squared error; MSE), conditional on the fitted prediction function: $\Psi_{P^0_{n,B_n}}(P_0)\equiv E_0[(Y-\bar{Q}_{P^0_{n,B_n}}(W))^2]$ is the true expected squared error loss of SL fit (based on parameter-generating sample, P^0_{n,B_n}) and estimated using corresponding validation sample: $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})=E_{P^1_{n,B_n}}[(Y-\bar{Q}_{P^0_{n,B_n}}(W))^2]$. Thus, the estimand is the risk averaged over the V estimation samples: $\Psi P_n(P_0)=E_{B_n}\Psi_{P^0_{n,B_n}}(P_0)$ and the corresponding estimator is $\hat{\Psi}(P_n)=E_{B_n}\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})$. Finally, inference is derived based on (4) above, where the estimated influence curve for the v-th estimation sample is given by

$$IC_{P^0_{n,B_n},n}=(Y-\bar{Q}_{P^0_{n,B_n}}(W))^2-\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}).$$

This is repeated for sample sizes of $n=100, 500, 1000,$ using algorithm 1, 2 and 3.

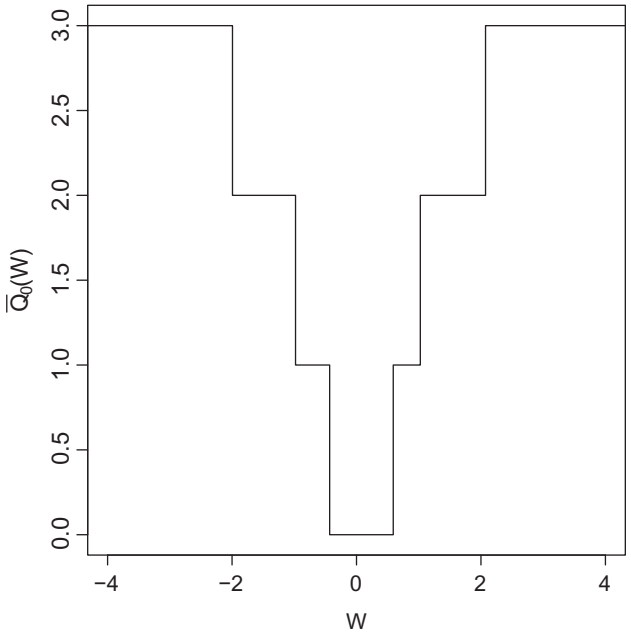

Figure 1: True model $\bar{Q}_0(W)$ for simulations of conditional risk estimation.

4.1.1 Results

We examined the empirical distribution of the standardized differences, $(\psi_n-\psi_{n,0})/se(\psi_n)$ for the risk. We observe minimal departure from normality (Figure 2), and nearly perfect coverage probability of the confidence intervals, for all sample sizes for algorithms 1 and 2 (see Table 1).However, one can see that

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

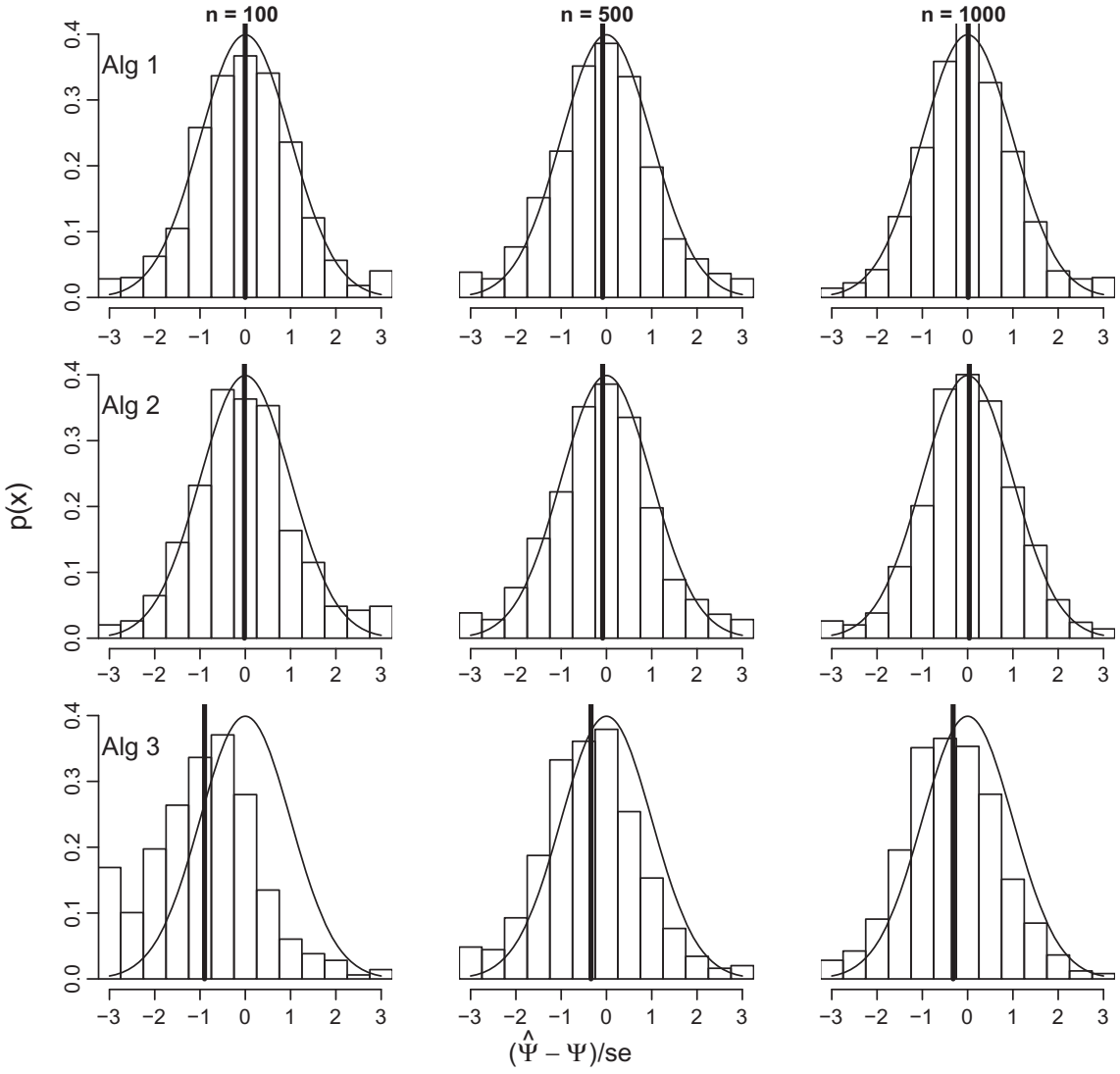

Figure 2: Distribution of $(\psi_n - \psi_{n,0})/se(\psi_n)$ (for $n = 100$, 500, 1000 with $N(0, 1)$ distribution for three algorithms (1, 2 and 3, corresponding to the 3 rows). Dark line represents the mean of these standardized values, so the difference between it and 0 is the standardized bias.

Table 1: Simulation results for risk estimation for data adaptive prediction on theorems 1–3 (ψ_n, ψ_n^1, ψ_n^2, respectively). Coverage probability is for a nominal 95 % CI.

N	Methods	Average true parameter	Average estimated	Cov. Prob.
$n = 100$	ψ_n	0.55	0.55	0.93
	ψ_n^1	0.54	0.54	0.92
	ψ_n^2	0.45	0.40	0.78
$n = 500$	ψ_n	0.55	0.55	0.94
	ψ_n^1	0.55	0.55	0.94
	ψ_n^2	0.46	0.46	0.92
$n = 1000$	ψ_n	0.40	0.40	0.95
	ψ_n^2	0.40	0.41	0.95
	ψ_n^2	0.36	0.36	0.93

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

algorithm 3, though resulting in a lower "average" risk, results in biased inference (and non-normal sampling distribution for the relatively modest sample sizes. This implies the Donsker conditions are not met. by a sample size of $n = 1000$, though there is still some standardized bias, the coverage probability is nearly perfect. Thus, even for a highly adaptive algorithm, where overfitting (underestimation of risk) seems particularly troublesome, at modest sample sizes, one begins achieving the conditions of theorem 3.

We also examined the same procedure for estimating the risk difference using algorithm 2. In this case, we observe slower convergence, but still relatively good coverage for an estimate that is particularly sensitive to over-fitting.

4.2 Average treatment effect for given prediction model

Average Treatment Effect, or ATE, is commonly the parameter of interest in applications of causal inference methods (Rubin [5]). We use the same set-up as we did for the example in the introduction, but for the ATE: $E(Y_1 - Y_0)$, where (Y_1, Y_0) are the counterfactual outcomes for an individual unit if they have $A = 1$ and $A = 0$, respectively. Consider n i.i.d. observations of $O = (W, A, Y) \sim P_0$, where Y is an outcome, A is a binary treatment of interest, and W a set of potential confounders. Under the assumption, the ATE equals the following statistical estimand:

$$ATE - E_{0,W}\{E_0(Y|A = 1, W) - E_0(Y|A = 0, W)\}.$$

Let $\bar{Q}_0(a, W) \equiv E_0(Y|A = a, W)$, and assume that $\bar{Q}_0$ is known. Then, the estimate of the ATE would be:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_0(1, W_i) - \bar{Q}_0(0, W_i)\}.$$

Given an estimator of $\bar{Q}_0$ on each of the training samples, we calculate the resulting data-adaptive ATE on the corresponding validation samples, and take the average, to derive our parameter of interest:

$$\Psi_{P_n}(P_0) = E_{B_n}E_{P_0}\{\bar{Q}_{P^0_{n,B_n}}(1, W) - \bar{Q}_{P^0_{n,B_n}}(0, W)\}, \tag{5}$$

The data generating distribution for this simulation is defined by $W \sim N\{0, var_0(W) = 4\}$, $A|W$ is binomial with $logit\{P(A = 1|W)\} = -4 + 2 * W$ and $Y|(W, A) \sim N\{\bar{Q}_0(A, W), var_0(Y|A, W) = 0.25\}$, where $\bar{Q}_0(a, W)$ is shown in Figure 3.

To derive $\bar{Q}_{P^0_{n,B_n}}$, we use the SuperLearner (SL) based upon the following learners: linear model, stepwise regression based on AIC (`stepAIC`; Venables and Ripley [19]), Bayesian glm (linear) model (`bayesglm`; Gelman et al. [20]), generalized additive model with smooth term for covariate Hastie and Tibshirani [21]; neural nets Venables and Ripley [19]; and a null model (intercept only).

We applied both algorithms 1–3 for sample size of $n = 500$.

4.2.1 Results

Examining the empirical distribution of the standardized differences, $(\psi_n - \psi_{n,0})/se(\psi_n)$ for the ATE parameter looked close to standard normal sampling distributions for both all algorithms 1–3 (not shown). Table 2 shows the results of the simulations based on both algorithms 1 and 3, and as one can see, the estimation is unbiased, and the coverage of confidence intervals based IC-based estimates of the standard errors is close to perfect. Though algorithms 1 and 3 produced different data adaptive target parameters and corresponding estimators, due to the linearity of the estimator $\Psi_{P^0_{n,B_n}}$ (i. e., it is just a difference in sample means), ψ_n and ψ_n^2 have the same MSE. This implies that even in this case, where very adaptive estimators are used, the Donsker class assumptions hold, as the confidence intervals have the nominal coverage.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

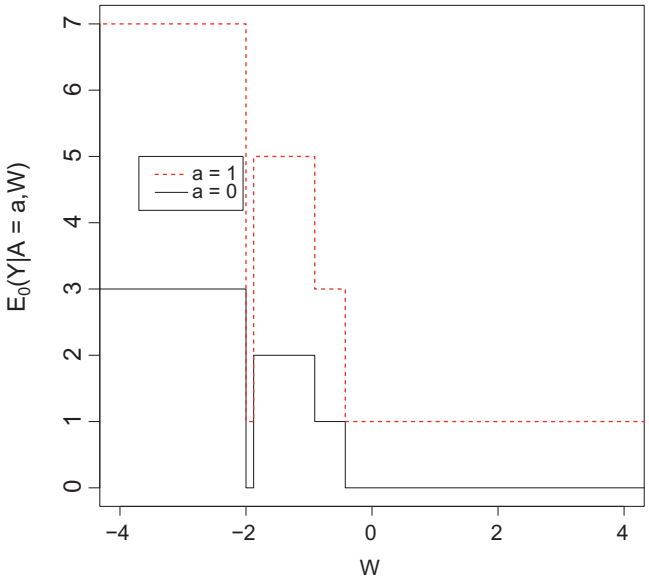

Figure 3: $E_0(Y|A=a, W)$ for the ATE simulations.

Table 2: Simulation results for *ATE* for the algorithms based on theorems 1–3 (ψ_n, ψ_n^1, ψ_n^2, respectively). Coverage probability is for a nominal 95 % CI.

Methods	Average true parameter	Average estimated	MSE	Cov. Prob.
ψ_n	1.71	0.89	0.82	1.19
ψ_n^1	1.71	0.89	0.82	0.51
ψ_n^2	1.68	0.89	0.79	0.63

4.3 Variable reduction

We consider a situation that has an analogue in high dimensional 'omic data, where multiple testing is often done to target a relatively small subset of (for instance) genes among the tens of thousands of candidates. The method evaluated in this simulation uses the parameter-generating sample to selects a small subset of the original genes, and subsequently it uses the estimation sample to estimate the effect of these genes on some phenotype. In this manner, it avoids the need to apply multiple testing procedures that control a type-I error rate among very large number of tests.

Let $O=(A, Y=(Y_1, Y_2, \ldots, Y_p))$ where A is binary vector of zeros and ones (indicating, for instance, phenotype), and Y is a multivariate outcome. Consider an algorithm that maps a data set $O_1, \ldots, O_n$ into a subset $C \subset \{1, \ldots, p\}$ of genes, where we denote this subset-estimator as $\hat{C}: \mathcal{M}_{NP} \rightarrow \mathcal{C}$, where $\mathcal{C}$ is the set of p–dim vectors with components in $\{0, 1\}$, so that $\mathcal{C}=\{0, 1\}^p$. We define our data adaptive parameter as:

$$\Psi_n(P_0) = E_{B_n}\{E_0(Y^*(\hat{C}(P^0_{n,B_n}))|A=1) - E_0(Y^*(\hat{C}(P^0_{n,B_n}))|A=0)\}, \tag{6}$$

where

$$Y^*(\hat{C}(P^0_{n,B_n})) = \frac{1}{|\hat{C}|}\sum_j I(\hat{C}(j)=1)Y_j$$

is an average of the gene-expression across a subset (cluster) of genes, where this subset is determined by a procedure on the parameter-generating sample.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

The estimator of (6) based on the estimation sample is simply

$$\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}) = E_{B_n}\{E_{P^1_{n,B_n}}(Y^*_{P^0_{n,B_n}}|A=1) - E(Y^*_{P^0_{n,B_n}}|A=0)\}$$

and its influence curve is estimated as follows

$$IC_{P^0_{n,B_n}}(Y^*, A) = \left[\frac{I(A=1)}{P^1_{n,B_n}(A=1)} - \frac{I(A=0)}{P^1_{n,B_n}(A=0)}\right](Y^* - E_{P^1_{n,B_n}}(Y^*|A)). \quad (7)$$

To investigate this estimator, we simulate based on a design where there are equal numbers of $A=0$ and $A=1$; for each (gene) j, the distribution of Y_j, given A, is defined by the following regression equation

$$Y_j = B_{0j} + B_{1j}A + e_j \quad j = 1, \ldots, p. \quad (8)$$

The coefficients (the B_{0j},B_{1j} were generated by a multivariate normal distribution with $E(B_0) = E(B_1) = 0$ and a variance covariance matrix with $Cov(B_0, B_1) = 1 \quad i = j$ and $Cov(B_0, B_1) = .2 \quad i \neq j$. Note, that these coefficients are fixed in the simulation, not random, so this is just a convenient mechanism to generate a distribution of effect sizes, B_{1j} for which there is a true ranking based on the resulting P_0. The errors e_j were independent draws from a random $N(0, \sigma^2_e)$ distribution, and we repeated the simulation both for different magnitudes of the residual error, (different σ^2_e) but also for increasing sample sizes. The function that defines the subset of genes is simply based on ranking the genes by $\hat{B}_{1j} = E_{P^0_{n,B_n}}(Y_j|A=1) - E_{P^0_{n,B_n}}(Y_j|A=0)$, and then $\hat{C}(P^0_{n,B_n}))$ is the indicator that a genes is in the top 15. Thus, this is a large variable-reduction exercise, where we examine the association of the average gene expression and phenotype of a data-adaptively selected subset of genes. The same procedure for deriving the data adaptive target parameter and estimator is repeated for all 3 algorithms, with the corresponding methods for deriving the inference via the influence curve carried out as described above.

4.3.1 Results

The results of the simulation are shown in Table 3 for the set with $\sigma^2_e = 2$. In this case, we observe very good performance with regards to coverage probability for algorithm 1, even at relatively modest sample sizes. On

Table 3: Simulation results for *Variable Reduction* for the algorithms based on theorems 1–3 ($\psi_n, \psi^1_n, \psi^2_n$, respectively). Coverage probability is for a nominal 95 % CI.

N	Methods	Average true parameter	Average estimated	Cov. Prob.
$n=100$	ψ_n	2.24	2.25	0.83
	ψ^1_n	2.24	2.58	0.0060
	ψ^2_n	2.63	2.57	0.014
$n=500$	ψ_n	2.48	2.48	0.91
	ψ^1_n	2.48	2.55	0.66
	ψ^2_n	2.49	2.56	0.69
$n=1,000$	ψ_n	2.51	2.51	0.92
	ψ^1_n	2.51	2.58	0.77
	ψ^2_n	2.51	2.57	0.80
$n=2,000$	ψ_n	2.56	2.56	0.94
	ψ^1_n	2.56	2.58	0.86
	ψ^2_n	2.56	2.58	0.88
$n=10,000$	ψ_n	2.51	2.51	0.96
	ψ^1_n	2.51	2.52	0.95
	ψ^2_n	2.51	2.52	0.94

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

the other hand, for algorithms two and three, which have an overlap in their parameter-generating and estimation samples, the confidence intervals have nominal at larger sample sizes $(n = 10,000)$, so the apparent violation of the conditions at smaller samples sizes$(n \le 2,000)$ for this very adaptive procedure, is not a violation at still relatively modest sample sizes. However, algorithm 1 shows very good performance at all by the smallest sample size, with regards to statistical inference, while not having greater sampling variability; thus, algorithm 1, all things being equal, is the safer choice.

5 Data analysis: Data adaptive estimation of the impact of interventions on cholesterol in WCGS study

We re-visit the original example discussed in the introduction: estimating the impact of the targeted treatment of cholesterol on rates of coronary heart disease (CHD) (3). Again, data is $O = (W, A, Y)$ with outcome Y (CHD), variable of interest A (total cholesterol; A is indicator of total cholesterol >180 mg/DL) and covariates W (age, weight, height, smoking, behavior type, etc.). We estimate a parameter akin to (3), using algorithm 1. Like the ATE example in Section 4.2, our parameter and corresponding estimator (on the validation sample) is defined via an estimate of the estimated regression of Y on A, W using the training sample to define the rule, $d_{\tau, \bar{Q}_{P^0_{n,B_n}}}$:

$$\Psi_{n, P^0_{n,B_n}}(P_0) = E(Y - Y_{d_{\tau, \bar{Q}_{P^0_{n,B_n}}}}) = EY - EY_{d_{\tau, \bar{Q}_{P^0_{n,B_n}}}} \tag{9}$$

where, as above, $\bar{Q}_{P^0_{n,B_n}}$ is the regression estimator on training sample. The estimator for (9) is the sample average minus the targeted maximum likelihood estimator (TMLE; van der Laan and Rose [17]) of the rule-specific mean. For the estimate of regression $\bar{Q}_{P^0_{n,B_n}}$ used to define the treatment rule, we used the ensemble machine learning method, *SuperLearner* [22, 23]. The learners included both very simple and more potentially complex, adaptive models: 1) fixed mean model, 2) main-terms logistic regression, 3) Stepwise logistic Regression with 2-way interactions, 4) Generalized Additive Model [21] with smooths for all non-factor covariates, 5) neural nets [24], 6) penalized regression using *glmnet* [25], and 7) nearest neighbor [24]. SL is itself based 10-fold cross-validation *within* the parameter-generating sample. On the corresponding estimation sample, the TMLE estimator also requires a regression of Y on (A, W) and we also used the SuperLearner with a similar set of learners; an estimate of the so-called treatment mechanism $(g_0(W) \equiv P(A = 1|W))$ is also required and main terms logistic regression was used. To derive inference, we need the plug-in influence curve (IC) for the estimation-sample-specific estimator. In this case:

$$IC_{P^0_{n,B_n}}(P^1_{n,B_n}) = Y - \left[\frac{I\{A = d_{\tau, \bar{Q}_{P^0_{n,B_n}}}(A, W)\}}{g_{P^1_{n,B_n}}(W)} \{Y - \bar{Q}_{P^1_{n,B_n}}(A, W)\} + \bar{Q}_{P^1_{n,B_n}}\{d_{\tau, \bar{Q}_{P^0_{n,B_n}}}(A, W), W\} \right] - \Psi_{P^0_{n,B_n}}(P^1_{n,B_n})$$

where $(\bar{Q}_{P^1_{n,B_n}}, g_{P^1_n, B_n})$ represent the estimators of $(\bar{Q}_0, g_0)$ on the validation sample. For our specific implementation, we used an arbitrary cut-off for "significant" improvement from lowering cholesterol from the current level as an reduction in risk of CHD of greater than 2.5 % $(\tau = 0.025)$. Besides the estimate of each of the training-sample specific parameters, we also estimate the average of these across the $V = 10$ folds as: $\Psi_n(P_0) = E_{B_n}\Psi_{P^0_{n,B_n}}(P_0)$, where the standard error of the estimate was calculated as (4). In addition, an equivalent estimate of the change in CHD rate if cholesterol was lowered in all subjects was done for comparison. The goal of the estimation is to determine whether one can target fewer people, but still not sacrifice much increase in the overall CHD rate.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

5.1 Results

The results (Figure 4) suggest one would reduce the risk of CHD by 3.1 % (95 % CI = 2.3 – 3.9%), by using the derived rule (which targets about 44 % of the population to reduce cholesterol from the observed value). However, estimating the impact of targeting all those with cholesterol $>180mg/dL$ (based on equivalent data-adaptive estimator) results in intervention in nearly double the population ($A=1$ in 86 % of sample), and in reduction of CHD rate of 4.6 %, 95 % CI = (3.7 – 5.5%). We then followed-up by estimating (9) using algorithms based on theorem 2 and 3, resulting in the estimates of 3.9 %, (95 % CI = (2.2 – 5.5%)) and 3.0 %, (95 % CI = (1.1 – 4.5%)), respectively, both which result in similar estimates and inference. This shows the potential of using the data-adaptive parameter approach when one has parameters that are complicated functions of unknown parts of the data-generating distribution. The fact that one also gets trustworthy inference for such an adaptive parameter makes this general approach a very compelling option to consider for circumstances where a non-adaptive parameter is impractical to estimate or requires large parametric assumptions to identify.

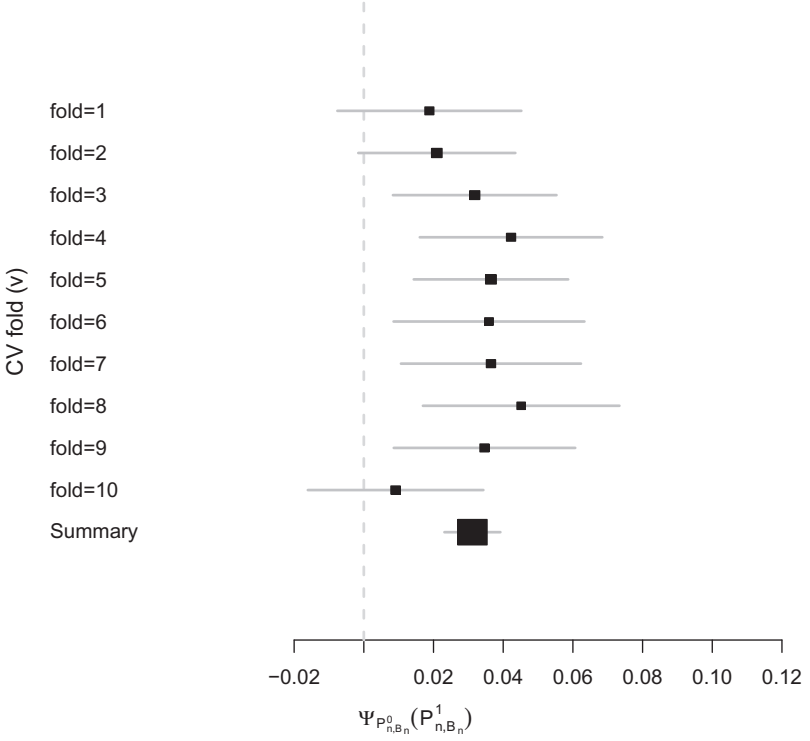

Figure 4: Forest plot of estimates (with 95 % CI's using influence curve-based standard errors) of validation-sample-specific and average data adaptive parameter (9), estimated using the WCGS data, with the intervention on cholesterol, and targeted group being those with an estimated average benefit due to lower cholesterol of a >2.5% reduction ($\tau=0.025$) in estimated CHD. "Summary" is the estimate of the average across the validation-specific estimates, or $E_{B_n}\Psi_{P^0_{n,B_n}}(P^0_{n,B_n})$.

6 Conclusion

Significant scientific progress can be made by generating target parameters based on past studies, and evaluating them on future, independent data. We discussed above, however, how such costly splitting of data is potentially unnecessary; the proposed data adaptive target parameter and corresponding statistical procedure studied in this article allows for general sample splits, and averaging the results across such

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

splits. The theoretical and simulation results demonstrate that statistical inference is preserved under minimal conditions, even though the estimators are now based on all the data. To obtain valid finite sample inference it is important to utilize our corresponding variance estimator (4), and that the sample size for the estimation sample is chosen large enough so that the second order terms of a possible non-linear estimator are controlled.

We also showed that if the algorithm that generates the target parameter is not too adaptive to small changes in the data, then no sample splitting is necessary. Specifically, if the set of influence curves generated by this parameter-generating algorithm when applied to an empirical distribution is a P_0-Donsker class, then statistical inference based on the method ψ_n^2 hat uses all the data to both generate the parameter and the estimate it is asymptotically valid. Thus, it provides a theorem for estimation and inference for so-called data-dredging. There are a large variety of data-mining applications where consistent estimation and inference are possible, including using the data to fit a finite dimension vector of coefficients that deterministically identifies a target parameter of interest. If the sample size is large and/or the parameter generating algorithm is well understood so that our Theorem 3 can be formally applied, then algorithm 3 should be considered as an important method.

We have demonstrate that data adaptive target parameter framework provides a formalized approach for estimating target parameters that are either very hard or impossible to pre-specify. There are many examples of interest that have not been highlighted in this article, where the motivation can come from dimension reduction, or complex causal parameters. There are few constraints on how one uses the data to define interesting parameters and we expect their are many applications in Big Data situations for which this approach is particularly well-suited.

Funding: Patient-Centered Outcomes Research Institute (PCORI) Pilot Project Program Award, (Grant/Award Number: 'ME-1306-02735. DISCLAIMER: All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology. Committee); National Institute of Health, (Grant/Award Number: 'R01AI074345-06A1'); National Institutes of Environmental Health Sciences, (Grant/Award Number: 'Grant number P42ES004705).

References

1. Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 3rd ed. Burlington, Massachusetts: Morgan Kaufmann, 2011.
2. Hubbard A, van der Laan M. Mining with inference: data-adaptive target parameters. In: P Buhlmann, P Drineas, M Kane, M van der Laan, editors. Handbook of big data, Handbook of Modern Statistical Methods. Boca Raton, FL: CRC Press, 2016:439–52.
3. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. Preserving statistical validity in adaptive data analysis. *arXiv preprint arXiv:1411.2664*, 2014.
4. Ragland DR, Brand RJ. Coronary heart disease mortality in the Western Collaborative Group Study. Follow-up experience of 22 years. Am J Epidemiol 1988;127:462–75.
5. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat 1978;6:34–58.
6. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer-Verlag, 2003.
7. Zhang F, Chen JY. Data mining methods in omics-based biomarker discovery. Methods Mol Biol 2011;719:511–26. doi: 10.1007/978-1-61779-027-0_24.
8. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet May 2013;14:333–46. doi: 10.1038/nrg3433.
9. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology Sep 2008;19:640–8. doi: 10.1097/EDE.0b013e31818131e7.
10. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2006;2:171–96.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM

11. Barraclough H, Govindan R. Biostatistics primer: what a clinician ought to know: subgroup analyses. J Thor Oncol 2010;5:741.
12. Marler JR. Secondary analysis of clinical trials – a cautionary note. Prog Cardiovas Dis 2012;54:335–7.
13. Le Dell E, Petersen M, van der Laan MJ. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. Technical report, U.C. Berkeley Division of Biostatistics Working Paper Series, http://www.bepress.com/ucbbiostat/paper304, 2012.
14. Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Stat Methodol 2005;2:131–54.
15. van der Laan MJ, Polley EC, Hubbard AE. Superlearner. Stat Appl Genet Mol Biol 2007a;6.
16. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
17. van der Laan M, Rose S. Targeted learning: causal inference for observational and experimental data. New York: Springer, 2011.
18. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. Int J Biostat 2006;2. URL http: //www.bepress.com/ijb/vol2/iss1/11. Article 11.
19. Venables WN, Ripley BD. Modern applied statistics with S, 4th ed. New York: Springer, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.
20. Gelman A, Su Yu.-S, Yajima M, Hill J, Grazia Pittau M, Kerman J, Zheng T. Arm: data analysis using regression and multilevel/hierarchical models, 2012. URL http://CRAN. R-project.org/package-arm. R package version 1.5-08.
21. Hastie TJ, Tibshirani RJ. Generalized additive models. New York: Chapman and Hall, 1990.
22. Polley E, van der Laan M. *SuperLearner: Super Learner Prediction*, 2012. URL http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-6.
23. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007b;6:Article25. doi: 10.2202/1544-6115.1309.
24. Ripley BD. Pattern recognition and neural networks. Cambridge, New York: Cambridge University Press, 1996.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw Feb 2010;33:1–22. ISSN 1548–7660.

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2015-0013) offers supplementary material, available to authorized users.

Brought to you by | University of California - Berkeley
Authenticated
Download Date | 6/21/16 7:25 PM