Statistical Inference for Data Adaptive Target Parameters

Mark J. van der Laan[*] Alan E. Hubbard[†]

Sara Kherad Pajouh[‡]

[*] UC Berkeley, Division of Biostatistics, laan@berkeley.edu

[†] UC Berkeley, Division of Biostatistics, hubbard@berkeley.edu

[‡] UC Berkeley, Division of Biostatistics, kherad@berkeley.edu

Statistical Inference for Data Adaptive Target Parameters

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

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 subsamples, and use this partitioning to define V splits in estimation-sample (one of the V subsamples) and corresponding complementary parameter-generating sample that is used to generate a target parameter. For each of the V parameter-generating samples, we apply an algorithm that maps the sample in a target parameter mapping which represent the statistical target parameter generated by that parameter-generating sample. We define our sample-split data-adaptive statistical target pa- rameter as the average of these V -sample specific target parameters. We present an analogue estimator of this type of data adaptive target parameter and corresponding statistical inference. This general methodology for generating data adaptive target parameters while still providing valid statistical inference is demonstrated with a number of examples. These examples demonstrate that this methodology presents new opportunities for statistical learning from data that go beyond the usual requirement that the estimand is a priori defined in order to allow for proper statistical inference. 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 - that is, the role of statisticians is being supplanted by computer scientist, deriving clever, yet typically ad hoc methods that "discover" the interesting patterns in data. The methodology presented in this paper can harness these methods, and now provide rigorous inference for the patterns, or target parameters suggested by such procedures. In this way, it returns exercises involving learning

from data back within the proper domain of rigorous statistical inference. To suggest such potential, and to verify the predictions of the theory, simulation studies based upon algorithms that map the parameter- generating sample into the desired estimand are shown. However, the methodology generalizes to situations where even these algorithms are not prespecified.

1 Introduction and motivating examples

Consider n independent and identically distributed observations $O_1, \ldots, O_n$ from a probability distribution P_0 that is known to be an element of a statistical model $\mathcal{M}$. In order to allow for formal statistical inference it is generally considered important that the statistical target parameter is a priori defined, where a statistical target parameter is defined as a mapping $\Psi : \mathcal{M} \rightarrow \mathbf{\Psi}$ from the statistical model into the parameter space $\mathbf{\Psi}$; $\psi_0 = \Psi(P_0)$ is the true parameter value. Otherwise, such analyses will typically be considered "exploratory" (data mining, data dredging) and the resulting findings typically greeted with greater skepticism. In a typical data mining exercise, a variety of target parameters, typically not pre specified, are estimated from the data. Specifically, one might apply a certain algorithm to the data that generates potential parameters of interest, but even this algorithm might be not only developed data adaptively, it is often the case that the steps taken are not well documented. For instance, the sequence of analysis often used in large scale omic studies (genomics, proteomics, metabolomics, etc.; Zhang and Chen [2011], Berger et al. [2013]) 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, leading to the highlighting of a particular pathway. In these cases, treating the resulting suggested target parameter of interest as pre-specified makes it virtually impossible to obtain honest statistical inference. Others have noted the particular dangers of high dimensional data combined with flexible methodologies to generate excessive false positive findings (Ioannidis [2008], Broadhurst and Kell [2006]). Other examples could include using a particular data adaptive method to fit a regression function only to then evaluate its fit on the same sample. In many cases, even when the best intentions are to stick to a pre-specified data analysis plan, implying targeted parameters of interest, there is feedback from the data and the types of analysis conducted: models changed (e.g., covarietes dropped) due to identifability concerns, unplanned sub-group analysis (Barraclough and Govindan [2010], Marler [2012]).

In this article, we propose methods that both utilize the potential of algorithms to find patterns in data in exploratory data analysis, but still provide meaningful inference about the resulting estimates. This paper, thus, provides a new framework for data mining methods used to generate interesting target parameters, while still providing honest statistical inference. Though there are advantages for doing so, the general methodology does not even require one to pre-specify the algorithm used to generate the target parameter(s). 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.

The method is analogous to a known commonly employed solution, which splits the sample in two parts, use one part to generate the statistical target parameter, and estimate the resulting target parameter with the other part of the sample. The great advantage of this is that honest statistical inference is obtained with standard methodology. However, the motivation for our approach is that this method comes at an enormous cost with regards to power since only half of the sample is used to estimate the statistical target parameter. Therefore, we seek extensions of this simple method, and the conditions such that the resulting inference will be consistent. Firstly, we propose to use V-fold sample splitting, use one part of the sample of size $n(1 - 1/V)$ to generate the statistical target parameter, and estimate this target parameter with the remaining sample of size n/V, and do this for each of the V splits of the sample. The target parameter is defined as the average over the V splits of the split-specific target parameters, and our estimator is defined as a corresponding average of split-specific estimators. We prove that the latter estimator allows statistical inference under very weak regularity conditions, which mean that one can derive honest inference for parameters defined by very aggressive procedures. The price one pays is that one has to accept that the statistical inference concerns the average of V target parameters. If the parameter generating algorithm is a priori specified and the same for each of the splits, then the V target parameters might be very similar, so that the interpretation of the average of the V target parameters is quite clear. On the other hand, if each split uses a different parameter-generating algorithm, then the resulting average of the V target parameters is what it is. Even in the latter case, rejecting a null hypothesis about this average of target parameters can be interesting by, for example, showing that the treatment is effective for at least one of the subgroups suggested within one of the parameter-generating samples.

In addition, we also present estimators that apply the parameter generating algorithm to the whole sample and then fits the resulting statistical target parameter on the same (full) data. This method is generally understood to provide misleading inference regarding the target parameters generated by the exploratory phase of the analysis, and, as expected, in many circumstances this understanding proves correct. However, we present the (stronger) assumptions under which such a procedure results in the asymptotically normal estimates with consistently estimable variability, noting that great caution must be exercised regarding finite sample behavior of the resulting statistical inference. In addition, in certain applications, such as estimating the true conditional risk of a candidate estimator and using this estimator to compare the performance of different candidate estimators, the finite sample bias of the resulting estimator due to overfitting might make the estimator unusable for its purpose. However, as opposed to a blanket dismal of this as a valid way to analyze data, this paper presents formal theorems for when such an approach will work asymptotically.

This article is organized as follows. In the next Section 2 we will define our general methodology with the two above mentioned estimation strategies (i.e. using sample splitting or not), present theorems establishing the asymptotic statistical performance, and discuss its implications, as well as present influence curve and bootstrap based inference. In Section 3, we discuss in detail several examples of interesting target parameters generated via the data adaptive methodology presented in Section 2. In Section 4 we demonstrate the finite sample performance (including coverage probabilities of confidence intervals) for several examples, showing the relative difference in efficiency and influence-curve based inference of three general approaches discussed in Section 2. This is followed by concluding remarks in Section 5.

2 General methodology

Let $O_1, \ldots, O_n$ be i.i.d. with probability distribution P_0 known to be an element of a 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.

For a given random split B_n, let P^0_{n,B_n} be the empirical distribution of the parameter-generating sample, and P^1_{n,B_n} be the empirical distribution of the estimation-sample. For a given B_n, $\Psi_{B_n,P^0_{n,B_n}} : \mathcal{M} \rightarrow \mathbb{R}$ is the target parameter mapping indexed by the parameter-generating sample P^0_{n,B_n}, and $\hat{\Psi}_{B_n,P^0_{n,B_n}} : \mathcal{M}_{NP} \rightarrow \mathbb{R}$ 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.

Thus, assume the existence of a mapping from the parameter-generating sample P^0_{n,B_n} into a target parameter mapping and a corresponding estimator of that target parameter, where this mapping can be different for each split B_n. The choice of target parameter mapping and corresponding estimator can be informed by the data P^0_{n,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 $\Psi_n : \mathcal{M} \rightarrow \mathbb{R}$ with

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

and the statistical estimand of interest is thus

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

Note that this target parameter mapping depends on the data, which is the reason for calling it a *data-adaptive target parameter*. A corresponding estimator of the data adaptive estimand $\psi_{n,0}$ is given by:

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

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 $H_0 : \psi_{n,0} \leq 0$. In particular, this would hold if $\psi_n = \hat{\Psi}(P_n)$ is an asymptotically linear estimator of $\psi_{n,0}$ with influence curve $IC(P_0)$:

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

where we used the notation $Pf \equiv \int f(o) dP(o)$ 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 a sum of mean zero independent random variables, by the central limit theorem, this 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) $R_{n,B_n} = o_P(1/\sqrt{n})$. For a given split $B_n = v$, assume that $P_0 IC^2_{v, P^0_{n,v}}(P_0) - P_0 IC_v(P_0))^2 \rightarrow 0$ in probability, where $IC_v(P_0)$ is a limit influence curve that can still be indexed by the split v.

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

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

where $\sigma^2_v = P_0 IC^2_v(P_0)$. A consistent estimator of σ^2 is given by

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

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

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

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.

Proof: As a consequence of the asymptotic linearity assumption,

$$\begin{aligned} \psi_n - \psi_{n,0} &= E_{B_n} \hat{\Psi}_{B_n, P^0_{n,B_n}}(P^1_{n,B_n}) - E_{B_n} \Psi_{B_n, P^0_{n,B_n}}(P_0) \\ &= E_{B_n} (P^1_{n,B_n} - P_0) IC_{B_n, P^0_{n,B_n}}(P_0) + E_{B_n} R_{n,B_n}. \end{aligned}$$

$E_{B_n} R_{n,B_n} = o_P(1/\sqrt{n})$ follows from the above stated asymptotic linearity and that B_n has only a finite V values. By assumption, for a given split $B_n = v$, $P_0 IC^2_{v,P^0_{n,v}}(P_0) - P_0 IC_v(P_0))^2 \to 0$ in probability, where $IC_v(P_0)$ is a limit that might still be indexed by the split v. As a consequence, for a given split $B_n = v$, conditional on the parameter-generating sample P^0_{n,B_n}, by the standard CLT, we have that $\sqrt{n/V}(P^1_{n,B_n} - P_0) IC_{B_n,P^0_{n,B_n}}(P_0)$ converges in distribution to a normal distribution with mean zero and variables $\sigma^2_v = P_0 IC_v(P_0)^2$. Since P^1_{n,B_n} are independent across the V realizations of B_n the right-hand side $E_{B_n}(P^1_{n,B_n} - P_0) IC_{B_n,P^0_{n,B_n}}(P_0)$ is an average of V independent sums.

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

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

where $\sigma^2_v = P_0 IC^2_v(P_0)$. □

Asymptotic equivalence of standardized estimator and standardized oracle estimator: Suppose that the algorithm $(B_n, P^0_{n,B_n}) \to (\hat{\Psi}_{B_n,P^0_{n,B_n}}, \Psi_{B_n,P^0_{n,B_n}})$ 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. 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. Thus in this case the limit-variance is given by

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

We can conclude that in this important case our 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 $\hat{\Psi}_{P_0}(P_0)$.

The number of splits V. Let's consider the case above that the algorithms that generate the parameter and estimator is constant across the V splits. In that case, the asymptotic variance of ψ_n as an estimator of $\psi_{n,0}$ is not affected by V, since it equals the asymptotic variance of $\hat{\Psi}_{P_0}(P_n)$ as an estimator $\Psi_{P_0}(P_0)$. So is there any guidance in selecting between $V = 2$ or $V = 10$, for example? First, notice the estimand is affected by the choice V: it is an average over V sample-split specific target parameters. Thus, this might provide an argument to prefer one or the other. Regarding statistical behavior, if we select V large, then the parameter-generating sample is large so that one might get a relatively stable collection of V target parameters (i.e., V target parameters that are very similar to each other). However, if V is large, then the estimation-sample is relatively small (size being n/V) and even though it does not affect first order asymptotics, for non-linear estimators, it will result larger second order terms. Thus, there might be a trade-off between having more data to generate more interesting and or more stable target parameters versus having more data in the estimation-samples to control the second order terms in the sample-split specific estimators. Presently, we have no universal recommendations, but this trade-off should be considered when designing the analysis.

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

In the above Theorem 1, Donsker class conditions were assumed, 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 ψ^1_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, whereas it has the same efficiency otherwise. Formally, we define this estimator as $\psi^1_n = E_{B_n}\hat{\Psi}_{B_n,P^0_{n,B_n}}(P_n)$.

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) $R_{n,B_n} = o_P(1/\sqrt{n})$. For a given split B_n, assume that $P_0(IC_{B_n,P^0_{n,B_n}}(P_0) - IC_{B_n}(P_0))^2 \to 0$ in probability, where $IC_{B_n}(P_0)$ is a limit that can still be indexed by the split B_n. We also assume that $IC_{B_n,P^0_{n,B_n}}(P_0)$ 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

$$IC(P_0) \equiv E_{B_n}IC_{B_n}(P_0)$$

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

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

Proof: As a consequence of the stated asymptotic linearity,

$$\begin{aligned}\psi^1_n - \psi_{n,0} &= E_{B_n}\hat{\Psi}_{B_n,P^0_{n,B_n}}(P_n) - E_{B_n}\Psi_{B_n,P^0_{n,B_n}}(P_0)\\ &= E_{B_n}(P_n - P_0)IC_{B_n,P^0_{n,B_n}}(P_0) + E_{B_n}R_{n,B_n}.\end{aligned}$$

$E_{B_n}R_{n,B_n} = o_P(1/\sqrt{n})$, which follows from the above stated asymptotic linearity and that B_n has only a finite V values. For a given split B_n, we assumed $P_0(IC_{B_n,P^0_{n,B_n}}(P_0) - IC_{B_n}(P_0))^2 \to 0$ in probability, where $IC_{B_n}(P_0)$ is a limit that can still be indexed by the split B_n. We also assumed that $IC_{B_n,P^0_{n,B_n}}(P_0)$ falls in a P_0-Donsker class with probability tending to 1. Then, by van der Vaart and Wellner [1996], for a given split B_n, conditional on the parameter-generating sample P^0_{n,B_n},

$$\begin{aligned}E_{B_n}(P_n - P_0)IC_{P^0_{n,B_n}}(P_0) &= E_{B_n}(P_n - P_0)IC_{B_n}(P_0) + o_P(1/\sqrt{n})\\ &= (P_n - P_0)E_{B_n}IC_{B_n}(P_0) + o_P(1/\sqrt{n}).\end{aligned}$$

This completes the proof of Theorem 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 $\sigma^2 = \sigma_1^2$. In the other extreme case that $P_0 IC_{v_1} IC_{v_2} = 0$ for $v_1 \neq v_2$, $\sigma^2 = 1/V \sum_v \sigma_v^2$ and $\sigma_1^2 = \frac{1}{V^2} \sum_v \sigma_v^2$. Thus, in the latter case $\sigma^2 = V\sigma_1^2$ 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_n^1, 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_n^1, so that there will be cases under which the model required for asymptotic normality of ψ_n holds, but the analogue model for ψ_n^1 fails to hold. This comparison also does not take into account that ψ_n^1 should have better second order term behavior than ψ_n for non-linear estimators, since ψ_n^1 involves using the full sample for each of the data adaptively generated target parameters.

Donsker class condition: We try to provide more detail on the important (non-trivial) additional Donsker class conditions for theorem/algroithm 2. Specifically, the key condition in this theorem is that the random influence curve $IC_{B_n, P^0_{n,B_n}}(P_0)$, random through its dependence on the data P^0_{n,B_n}, falls with probability tending to one in a P_0-Donsker class. A Donsker class $\mathcal{F}$ is a class of functions for which the entropy integral $\int \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}, L^2(Q))} d\epsilon < \infty$ (van der Vaart and Wellner [1996]). Here $N(\epsilon, \mathcal{F}, L^2(Q))$ is the number of balls of size ϵ (w.r.t. Hilbert space norm $\| f \| = \sqrt{\int f^2 dQ}$) that is needed to cover $\mathcal{F}$, and it is called the covering number. Thus a Donsker class $\mathcal{F}$ requires this covering number $N(\epsilon, \mathcal{F}, L^2(Q))$ (which can be bounded by $N(\epsilon, \mathcal{F}, \| \cdot \|_\infty)$ w.r.t. supremum norm) to converge to infinity when $\epsilon \to 0$ at a slower rate than $\exp(1/\epsilon^2)$. Here one might keep in mind that a finite dimensional set of dimension P would have a covering number that behaves as $1/\epsilon^p$, so that Donsker classes can be much larger than finite dimensional sets.

For example, one implication is that, if the data adaptive target parameter is only random through a finite dimensional vector of coefficients, so that this influence curve $IC_{B_n, P^0_{n,B_n}}$ will only be random through a finite dimensional vector of coefficients, then this assumption will practically always hold: since in this case this random influence curve can be represented as $IC_{\beta(P^0_{n,B_n})}(P_0)$ for a finite dimensional random vector $\beta(P^0_{n,B_n})$, and such finite dimensional class of functions satisfies the entropy integral condition. However, P_0-Donsker classes are allowed to be much bigger than finite dimensional. That is $N(\epsilon, \mathcal{F}, \| \cdot \|)$ does not have to be a polynomial in $1/\epsilon$, but it can behave as $\exp(1/\epsilon^a)$ for $a < 2$. As a consequence, many large Donsker class example exist. An example of a Donsker class is the class of k-variate functions that have uniform sectional variation norm smaller than a universal constant $M < \infty$ (Van Der Laan [1996]): for example, for $k = 2$, this is a class of functions for which there exists a $M < \infty$ so that for each $f \in \mathcal{F} \int \mid df \mid < M$, $\sup_x \int \mid f(x, dy) \mid < M$, and $\sup_y \int \mid f(dx, y) \mid < M$.

One important operation that preserves the Donsker property is taking convex combinations. As a consequence, the class of convex combinations of a set of basis functions for which each basis function has a uniform sectional variation norm bounded by a $M < \infty$ is a Donsker class. For example, one might define as class of basis functions piece-wise linear functions for which the number of knot-point is bounded by a universal K. Even thought this class of functions could be infinite, the class of functions defined by all convex combinations of such basis functions would be a Donsker class. Moreover, the convex combinations can be replaced by "weighted" averages in which the sum of the absolute values of the weights is bounded by a constant. This indicates that if the influence curve $IC_{B_n, P^0_{n,B_n}}(P_0)$ can be represented as $IC_{\sum_j \beta_j(P^0_{n,B_n})\phi_j}(P_0)$, where $\sum_j \mid \beta_j \mid < M < \infty$ for some $M < \infty$ and ϕ_j are multivariate real valued functions with bounded uniform sectional variation norm, then such an influence curve will satisfy the Donsker condition. In this manner, one may show that data adaptive target parameters that use the data P^0_{n,B_n} to create, for example, a data adaptively weighted combination of outcomes and defines the target parameter as the (causal) effect of a given treatment on that outcome will result in a corresponding influence curve $IC_{B_n, P^0_{n,B_n}}$ that satisfies the Donsker class condition, allowing for infinite number of outcomes as long as the sum of the absolute value of the weights is controlled.

By a priori defining the parameter generating mapping and estimator, one can study this influence curve condition, and based on such a mathematical analysis one might feel comfortable with this Donsker class condition. Simulation studies, such as obtained by resampling data from a semi parametric fit of P_0 (i.e., semiparametric bootstrap), could be used to shed additional light on the validity of these assumptions, and to establish if normality of the standardized estimator is a reasonable assumption. In these cases, this method and the method in the next subsection are important. If one does not want to a prior specify such a parameter mapping or one wants to allow for highly adaptive mappings that potentially fall outside the relevant Donsker class, then the method based on theorem 1 should be used.

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 $\sigma^2 = P_0 IC_{P_0}(P_0)$.

The proof follows trivially.

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.

2.3 Inference via the bootstrap

The estimator $\psi_n = E_{B_n}\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})$ of $\psi_{n,0} = E_{B_n}\Psi_{P^0_{n,B_n}}(P_0)$ has a second order term that includes $E_{B_n}R_{n,P^1_{n,B_n}}$, an average over the splits B_n of the second order term $R_{n,P^1_{n,B_n}}$ of the estimator applied to the estimation sample P^1_{n,B_n}, only based on n/V observations. This suggests that the second order term in the expansion of $\psi_n - \psi_{n,0}$ might be substantial in many practical scenarios. It is well known that in such cases the bootstrap may provide important improvements by estimating a finite sample variance instead of only aiming to estimate the variance of the influence curve (which represents the asymptotic variance). On the other hand, one needs to be aware that the nonparametric bootstrap does not always work since it relies on the estimators to be smooth functions of the data. The asymptotic linearity of $\psi_n - \psi_{n,0}$ is a necessary condition for the validity of the bootstrap. Under additional smoothness conditions on the estimators $\hat{\Psi}_{P^0_{n,B_n}}$, one can establish asymptotic validity of the bootstrap. We assume these conditions to further discuss the bootstrap.

Let $O_1^{\#}, \ldots, O_n^{\#}$ be an i.i.d. sample from the empirical distribution P_n, referred to as the bootstrap-sample. Let $P_n^{\#}$ be the empirical distribution of this bootstrap sample. Let $P^{1,\#}_{n,B_n}, P^{0,\#}_{n,B_n}$ be the empirical distributions of the estimation-sample and parameter-generating-sample defined by the split B_n applied to the bootstrap sample. One can now construct $\psi_n^{\#} = E_{B_n} \hat{\Psi}_{P^{0,\#}_{n,B_n}}(P^{1,\#}_{n,B_n})$ as an estimator of $\psi^{\#}_{n,0} = E_{B_n} \Psi_{P^0_{n,B_n}}(P^1_{n,B_n})$ for each bootstrap sample $P_n^{\#}$. Given P_n, one samples a large number of draws from $P_n^{\#}$. Thus, conditional on P_n, one can construct a large number of random draws of $\sqrt{n}(\psi_n^{\#} - \psi^{\#}_{n,0})$ and use this bootstrap sampling distribution as an estimator of the distribution of $\sqrt{n}(\psi_n - \psi_{n,0})$. Specifically, the variance of $\sqrt{n}(\psi_n^{\#} - \psi^{\#}_{n,0})$ yields an estimate of the variance of $\sqrt{n}(\psi_n - \psi_{n,0})$, which can be used to construct a normal-based confidence interval.

Obviously, this bootstrap method is much more computer intensive than the asymptotic normality based methods based on the influence curve presented earlier, but should be considered when the algorithm that maps the sample P^0_{n,B_n} into the target parameter mapping and corresponding estimator, and the estimator $\hat{\Psi}_{P^0_{n,B_n}}$ applied to P^1_{n,B_n} are not too computer intensive.

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.

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

One of the fundamental statistical parameters relevant to prediction models (e.g., diagnostic models) is the performance of such model fits to future data. Thus, estimation and *inference* regarding risk estimates are crucial for determining expectations for future performance. In this case, the parameter is clearly a data adaptive target parameter, since of interest is not the performance of repeated experiments where the model is re-fit, but the performance of an actual model, fit to data, that will be used in the future.

Let $O = (W, Y)$, where W is a vector of input-variables and Y is an outcome one wants to predict. Let P_0 be its probability distribution and let the statistical model $\mathcal{M}$ be nonparametric. Let $\hat{\bar{Q}}$ be an estimator of the true regression function $\bar{Q}_0 = E_0(Y \mid W)$, and let $\bar{Q}_{P^0_{n,B_n}}$ be the corresponding estimate of $\bar{Q}_0 = E_0(Y \mid W)$ based on the parameter-generating sample P^0_{n,B_n}. The target parameter $\Psi_{P^0_{n,B_n}}(P_0)$ generated by P^0_{n,B_n} is defined as the mean squared error $E_0(Y - \bar{Q}_{P^0_{n,B_n}}(W))^2$ 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 sample-split data adaptive target parameter 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 $\psi_n = E_{B_n} P^1_{n,B_n} L(\bar{Q}_{P^0_{n,B_n}})$, $\psi_n^1 = E_{B_n} P_n L(\bar{Q}_{P^0_{n,B_n}})$, and $\psi_n^2 = P_n L(\bar{Q}_{P_n})$. 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., 2007].These results were presented earlier (Dell et al. [2012], Dudoit and van der Laan [2005]).

Similarly, Theorem 2 implies a formal result for ψ_n^1, 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 $\psi_n^2 = P_n L(\bar{Q}_{P_n})$ is an asymptotically linear estimator of $P_0 L(\bar{Q})$ with the same influence curve $L(\bar{Q}) - P_0 L(\bar{Q})$. Even though these conditions might be satisfied for $\bar{Q}_n$, the estimator ψ_n^2 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 the sample-split AUC of a data adaptive regression estimator.

The motivation for this parameter is identical to above, it simply represents a different measure of performance for diagnostic models.

Let $\hat{\bar{Q}}$ be an estimator of $\bar{Q}_0 = E_0(Y \mid W)$, and let $\bar{Q}_{P^0_{n,B_n}}$ be the corresponding estimate of $\bar{Q}_0 = E_0(Y \mid W)$ based on the sample P^0_{n,B_n}. The target parameter $\Psi_{P^0_{n,B_n}}(P_0)$ generated by P^0_{n,B_n} is defined as the true area under the curve $AUC(P_0, \bar{Q}_{P^0_{n,B_n}})$, where

$$\begin{aligned} AUC(P_0, \bar{Q}) &= \int_0^1 P_0\left(\bar{Q}(W) > c \mid Y = 1\right) P_0\left(\bar{Q}(W) = c \mid Y = 0\right) dc \\ &= P_0\left(\bar{Q}(W_1) > \bar{Q}(W_2) \mid Y_1 = 1, Y_2 = 0\right), \end{aligned}$$

where the latter equivalent representation is in terms of two independent observations $(W_1, Y_1), (W_2, Y_2)$. The estimator of $\Psi_{P^0_{n,B_n}}(P_0)$ based on P^1_{n,B_n} is defined as its empirical counterpart $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}) = AUC(P^1_{n,B_n}, \bar{Q}_{P^0_{n,B_n}})$. The sample-split data adaptive target parameter is thus defined as $\psi_{n,0} = E_{B_n} AUC(P_0, \bar{Q}_{P^0_{n,B_n}})$ and its corresponding estimators are $\psi_n = E_{B_n} AUC(P^1_{n,B_n}, \bar{Q}_{P^0_{n,B_n}})$, $\psi_n^1 = E_{B_n} AUC(P_n, \bar{Q}_{P^0_{n,B_n}})$, and $\psi_n^2 = AUC(P_n, \bar{Q}_{P_n})$. In Dell et al. [2012] we show that $AUC(P_n, \bar{Q})$ is an asymptotically linear estimator of $AUC(P_0, \bar{Q})$ with influence curve

$$\begin{aligned} IC_{AUC}(P_0, \bar{Q})(O) = & \frac{I(Y=1)}{P_0(Y=1)} P_0\left(\bar{Q}(W) < x \mid Y = 0\right)|_{x=\bar{Q}(W)} \\ & + \frac{I(Y=0)}{P_0(Y=0)} P_0\left(\bar{Q}(W) > x \mid Y = 1\right)|_{x=\bar{Q}(W)} \\ & - \left\{ \frac{I(Y=0)}{P_0(Y=0)} + \frac{I(Y=1)}{P_0(Y=1)} \right\} AUC(P_0, \psi). \end{aligned}$$

The theorem 1 implies that if 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 $IC(P_0) = IC_{AUC}(P_0, \bar{Q})$. This allows one to construct a confidence interval for the true sample-split AUC $\psi_{n,0}$. In particular, an estimator $\hat{\bar{Q}}$ can be a highly data adaptive super learner. These results for the cross-validated area under the curve for a given estimator of $\bar{Q}_0$ were earlier presented in Dell et al. [2012], which also provide a simulation and data example to demonstrate the finite sample coverage of the confidence interval.

Similar remarks as in the previous example apply to the alternative estimator $\psi_n^1 = E_{B_n} AUC(P_n, \bar{Q}_{P^0_{n,B_n}})$ of the same estimand (sample split area under the curve) $\psi_{n,0} = E_{B_n} AUC(P_0, \bar{Q}_{P^0_{n,B_n}})$, and the estimator $\psi_n^2 = AUC(P_n, \bar{Q}_{P_n})$ of $AUC(P_0, \bar{Q}_{P_n})$, but these two estimators rely on the Donsker class condition putting some constraints on how adaptive the estimator $\bar{Q}_{P_n}$ can be.

3.3 Inference for sample-split cluster-specific target parameters, where the clusters are data adaptively determined

Consider a situation where one has a very high dimensional set of variables, potentially correlated in relatively distinct groups, but for which the definition of such groups is not known a priori, and thus must be determined empirically. Furthermore, that summaries of the values of the variables in these blocks represent meaningful summaries of their joint relationship to an explanatory variable. One such situation might be genomic experiments, where the expression of highly correlated genes represent distinct pathways that might be simultaneously triggered by the same intervention (e.g., drug). Thus, one might derive significant variable reduction by creating summaries of the highly correlated gene expression in these blocks, as well as exploratory summaries of potential pathways to more efficiently measure the impact of the intervention of interest. Because this involves both an exploratory part (forming the blocks/clusters), but still with the need for formal statistical inference to determine the significance of the relationship of these clusters to the intervention, it is an ideal application for the methodologies in this paper.

Suppose that one observes on each subject a p-dimensional gene-expression profile $Y \in \mathbb{R}^p$, a binary treatment/exposure A, and a vector of baseline characteristics W. Thus $O = (W, A, Y)$ and we observe n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P_0$. Suppose that the statistical model $\mathcal{M}$ is nonparametric. Consider an algorithm that maps a data set $O_1, \ldots, O_n$ into a cluster $C \subset \{1, \ldots, p\}$ of genes. Denote this cluster-estimator with $\hat{C} : \mathcal{M}_{NP} \rightarrow \mathcal{C}$, where $\mathcal{C}$ is the space of possible cluster values. Given a realized cluster C, let $\Psi_C : \mathcal{M} \rightarrow \mathbb{R}$ be a desired parameter of interest such as the effect of treatment A on $Y(M(C))$, controlling for the baseline covariates W, where $M(C)$ is the medoid/center of the cluster C, defined as

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

Alternatively, one might define $\Psi_C(P_0)$ as the average over all genes j in cluster C of the effect of treatment A on $Y(j)$, controlling for W, defined as

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

Let $\hat{\Psi}_C : \mathcal{M}_{NP} \rightarrow \mathbb{R}$ be an estimator of $\Psi_C(P_0)$ such as a targeted maximum likelihood estimator as presented in van der Laan and Rubin [2006] and van der Laan and Rose [2011]. Assume that the regularity conditions hold so that this TMLE $\hat{\Psi}_C(P_n)$ 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 $R_{C,n} = o_P(1/\sqrt{n})$. We define $\Psi_{P^0_{n,B_n}} : \mathcal{M} \rightarrow \mathbb{R}$ as $\Psi_{P^0_{n,B_n}} = \Psi_{\hat{C}(P^0_{n,B_n})}$, i.e., the causal effect of treatment on the data adaptively determined cluster $\hat{C}(P^0_{n,B_n})$. Similarly, we define $\hat{\Psi}_{P^0_{n,B_n}} : \mathcal{M}_{NP} \rightarrow \mathbb{R}$ 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 of this data adaptively determined cluster, 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 cluster $\hat{C}(P^0_{n,B_n})$ 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

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

where we now assume that (unconditionally) $R_{\hat{C}(P^0_{n,B_n}),n} = o_P(1/\sqrt{n})$. 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 cluster $\hat{C}(P_0)$. 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.

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

On obvious subject of interest in many studies, including drug trials, is the identification of subgroups within the target population that have unique relationships with explanatory variable of interest (e.g., drug treatment, environmental exposure, etc.). Often, these sub-groups are not a priori known, and so sub-group analysis is typically treated as purely explanatory and thus the statistical inference inherently flawed, typically anti-conservative. However, the approach we have outlined above now allows both an aggressive search for interesting sub-groups, and formal, statistical inference concerning parameters related to the association of an explanatory and outcome variable of interest within the same data set. This is one of the exciting new opportunities this approach now permits.

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 $O_1, \ldots, O_n$ of $O \sim P_0$. Suppose that the statistical model $\mathcal{M}$ is nonparametric. Consider an algorithm that maps a data set $O_1, \ldots, O_n$ into a subgroup $W \rightarrow C(W) \in \{0, 1\}$, where $C(W) = 1$ indicates membership in the subgroup. Denote this subgroup-estimator with $\hat{C} : \mathcal{M}_{NP} \rightarrow \mathcal{C}$, where $\mathcal{C}$ is the space of functions that map a W into a binary indicator. Given a realized subgroup C, let $\Psi_C : \mathcal{M} \rightarrow \mathrm{I\!R}$ be a desired parameter of interest such as the W-controlled effect of treatment A on Y for subgroup C, defined as

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

Let $\hat{\Psi}_C : \mathcal{M}_{NP} \rightarrow \mathrm{I\!R}$ be an estimator of $\Psi_C(P_0)$ such as a targeted maximum likelihood estimator or TMLE [van der Laan and Rubin, 2006, van der Laan and Rose, 2011]: 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 $\hat{\Psi}_C(P_n)$ 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 $R_{C,n} = o_P(1/\sqrt{n})$.

Define $\Psi_{P^0_{n,B_n}} : \mathcal{M} \rightarrow \mathrm{I\!R}$ 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 $\hat{\Psi}_{P^0_{n,B_n}} : \mathcal{M}_{NP} \rightarrow \mathrm{I\!R}$ 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

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

where we now assume that (unconditionally) $R_{\hat{C}(P^0_{n,B_n},n} = o_P(1/\sqrt{n})$. 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).

- Algorithm 1, in the context of V-fold cross validation, defines the parameter on the parameter generating sample, $\Psi_{P^0_{n,B_n}}$, and estimated on the corresponding estimation samples, $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n})$ which are then averaged to produce an estimate of the data adaptive parameter of interest.
- Algorithm 2 uses the same procedure to define the target parameter, but the estimation for each parameter defined by a particular parameter generating sample is estimated on the full sample.
- Algorithm 3 which generates the target parameter based on the whole sample, and subsequently estimates this target parameter on the whole sample: $\hat{\Psi}_{P_n}(P_n)$.

We will report three sets of simulations which will be explained in detail below.

4.1 General Simulation Structure

Description of the general simulation structure provides concreteness to what we refer to as data adaptive target parameters, as well as how the three algorithms differ in their details.

(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, this defines $\Psi_{P^0_{n,B_n}}$. For instance, fit a data-adaptive regression procedure estimating the mean of outcome Y based on predictors X, say $\hat{m}_v(X) \equiv m_{P^0_{n,B_n}}(X)$, and define the target parameter as the risk based on squared error loss defined as $\Psi_{P^0_{n,B_n}}(P_0) = E_{P_0}(Y - \hat{m}_v(X))^2$, treating $\hat{m}_v$ 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., $\hat{\Psi}_{P^0_{n,B_n}}(P^1_{n,B_n}) = E_{P^1_{n,B_n}}(Y - \hat{m}_v(X))^2$. 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 $\Psi_{P^0_{n,B_n}}(P_0) = E_{P_0}(Y - \hat{m}_v(X))^2$, where P_0 is approximated by this empirical probability distribution of this very large sample (specifically, a sample of 100000).

(5) Estimate the asymptotic variance (1) 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 1000 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.

4.2 Risk Estimation of a Data Adaptive Prediction Algorithm

For this simulation, the motivating question concerns estimation of the risk of a machine learning algorithm. For this, we have the following set-up:

- The data is $O = (Y, X)$, for outcome Y, predictor X, where $X \sim N(0, \sigma^2_X = 4)$, $m_{true}(X) \equiv E_0(Y \mid X)$ is shown in Figure 1, based on a piecewise constant model, and $Y|X \sim N(m_{true}(X), \sigma^2_Y = 0.25)$.
- 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. [2007]), 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 [2002]), Bayesian glm (linear) model (`bayesglm`; Gelman et al. [2012]), generalized additive model with smooth term for covariate Hastie and Tibshirani [1990]; neural nets (`nnet`; Venables and Ripley [2002]); 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: $\Psi_{P^0_{n,v}}(P_0) \equiv E_{P_0}[(Y - \hat{m}_v(X))^2]$ (treating $\hat{m}_v$ as given) is the expected squared error loss of SL fit on parameter-generating sample, where $\hat{m}_v = \hat{m}(P^0_{n,v})$ is estimated as $\hat{\Psi}_{P^0_{n,v}}(P^1_{n,v}) = E_{P^1_{n,v}}[(Y - \hat{m}_v(X))^2]$.
- The data adaptive parameter of interest is thus defined as the risk averaged over the V estimation samples: $\Psi_{P_n}(P_0) = \frac{1}{V}\sum_{v=1}^{V} \Psi_{P^0_{n,v}}(P_0)$.
- The corresponding estimator is defined as $\hat{\Psi}(P_n) = \frac{1}{V}\sum_{v=1}^{V} \hat{\Psi}_{P^0_{n,v}}(P^1_{n,v})$.
- Finally, inference is derived based on (1) above, where the estimated influence curve for the v-th estimation sample is given by

$$IC_{P^0_{n,v},n} = (Y - \hat{m}_v(X))^2 - \hat{\Psi}_{P^0_{n,v}}(P^1_{n,v}).$$

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

4.2.1 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, and for both algorithms 1 and 2 (see Table 1).

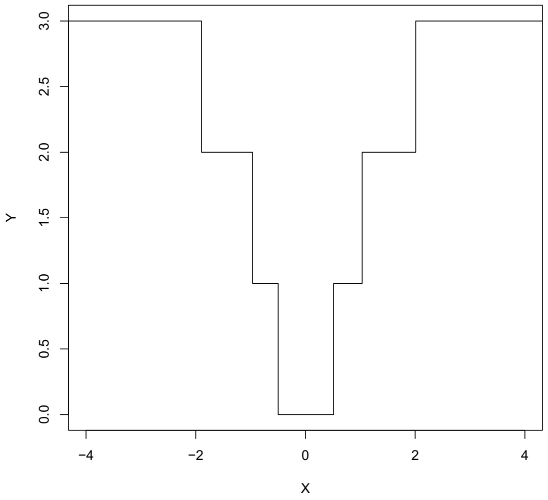

Figure 1: True model $m_{true}(X)$ for simulations of conditional risk estimation

Table 1: Simulation results for Estimating Conditional Risk for Methods based on Theorem 1 and 2. Coverage probability is for a nominative 95% CI

Algorithm	n	Ave Est. $E_{B_n}\hat{\Psi}$	Ave. True $E_{B_n}\Psi(P_0)$	MSE	Variance	Cov. Prob.
1	100	0.59	0.59	2.43	2.43	0.92
1	500	0.55	0.55	1.39	1.39	0.94
1	1000	0.40	0.40	0.38	0.38	0.94
2	500	0.83	0.84	1.11	1.05	0.94
2	1000	0.83	0.84	1.12	1.07	0.95

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

4.3 The average effect of treatment/variable for a given regression fit

Average Treatment Effect, or ATE, is commonly the parameter of interest in applications of causal inference methods, such as propensity score methods (Rubin [1978]). In the potential outcomes framework, the ATE is defined as $E(Y(1)-Y(0))$, 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 $O = (W, A, Y)$, where Y is an outcome, A is a binary treatment of interest, and W a set of potential confounders. Under the randomization assumption and a positivity assumption, the ATE equals the following statistical estimand:

$$ATE = E_W\{E(Y \mid A = 1, W) - E(Y \mid A = 0, W)\}.$$

Let $Q(a, W) \equiv E(Y \mid A = a, W)$, and assume that Q is known. Then, the estimate of the ATE would be:

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

Given an estimator Q_{P_n} of Q, we will estimate Q on the parameter-generating samples, and then calculate the ATE on the corresponding estimation sample, resulting in the following data adaptive

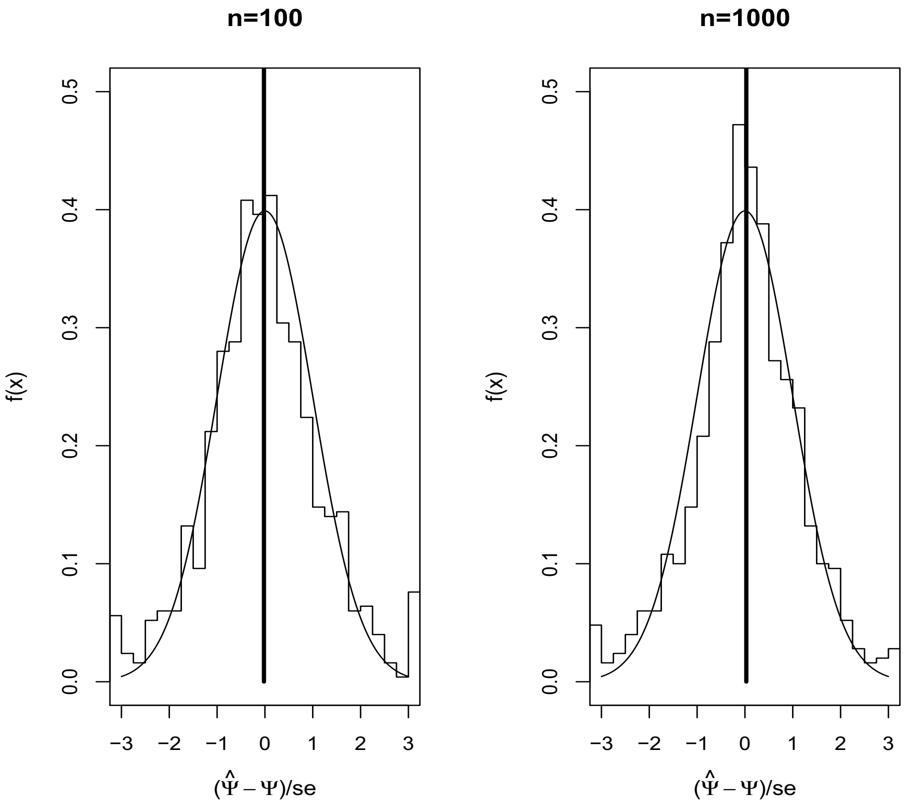

Figure 2: Distribution of $(\psi_n - \psi_{n,0})/se(\psi_n)$ (for $n = 100$ and 1000) with N(0,1) distribution for comparison based on theorem 1. Dark line represents the mean of these standardized values, so the difference between it and 0 is the standardized bias

.

target parameter:

$$\Psi_{P_n}(P_0) = E_{B_n} E_{P_0}\{Q_{P^0_{n,B_n}}(1, W) - Q_{P^0_{n,B_n}}(0, W)\}.$$

Here $Q_{P^0_{n,B_n}}$ is the estimate of the regression of Y on (A, W).

The data generating distribution is defined by $W \sim N(0, var(W) = 4)$, $A \mid W$ is binomial with $logit\{P(A = 1 \mid W)\} = -4 + 2 * W$ and $Y \mid (W, A) \sim N(Q(A, W), Var(Y \mid A, W) = 0.25)$, where $Q(a, W) = E(Y \mid A = a, W)$ is shown in Figure 4.

The data-adaptive target parameter is defined in terms of an estimator Q_{P_n}. As above for the risk estimation simulations we used the SuperLearner (SL) based upon the following learners: linear model, stepwise regression based on AIC (`stepAIC`; Venables and Ripley [2002]), Bayesian glm (linear) model (`bayesglm`; Gelman et al. [2012]), generalized additive model with smooth term for covariate Hastie and Tibshirani [1990]; neural nets Venables and Ripley [2002]; and a null model (intercept only).

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

4.3.1 Results

Examining the empirical distribution of the standardized differences, $(\psi_n - \psi_{n,0})/se(\psi_n)$ for the ATE parameter, we see convergence to normal sampling distributions for both algorithms 1 and 3

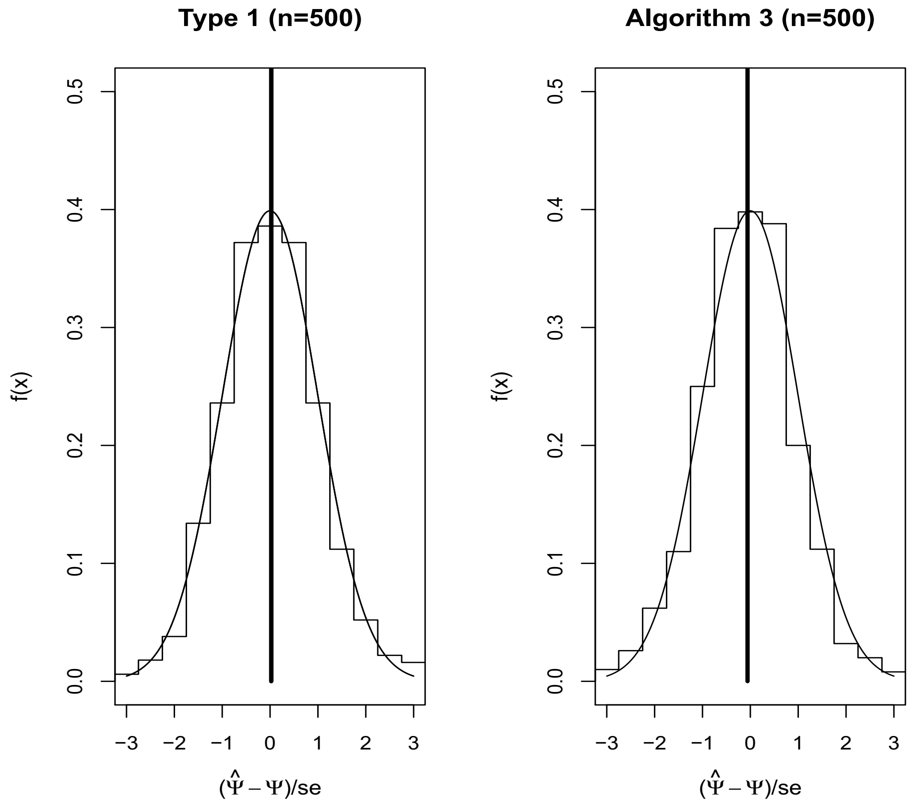

Figure 3: Results for estimation of ATE. Distribution of $(\psi_n - \psi_{n,0})/se(\psi_n)$ (for $n = 500$) with N(0,1) distribution for comparison, with estimates based on both algorithm 1 and 3. Dark line represents the mean of these standardized values, so the difference between it and 0 is the standardized bias

.

at $n = 500$.

Table 2: Simulation results for Estimating ATE using algorithms based on Theorem 1 and 3

type	Ave Est. $E_{B_n}\hat{\Psi}$	Ave. True $E_{B_n}\Psi(P_0)$	MSE	Variance	Coverage Prob (95% CI)
1	0.81	0.81	0.90	0.90	0.95
3	0.81	0.81	1.17	1.17	0.94

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 $\hat{\Psi}_{P^0_{n,B_n}}$ (i.e., it is just a difference in sample means), ψ_n and ψ_n^2 have the same MSE.

4.4 Variable Reduction

We consider a situation that has an analogue in high dimensional omit data, where multiple testing is often done to highlight a relatively small subset of say genes for further study, among the tens of

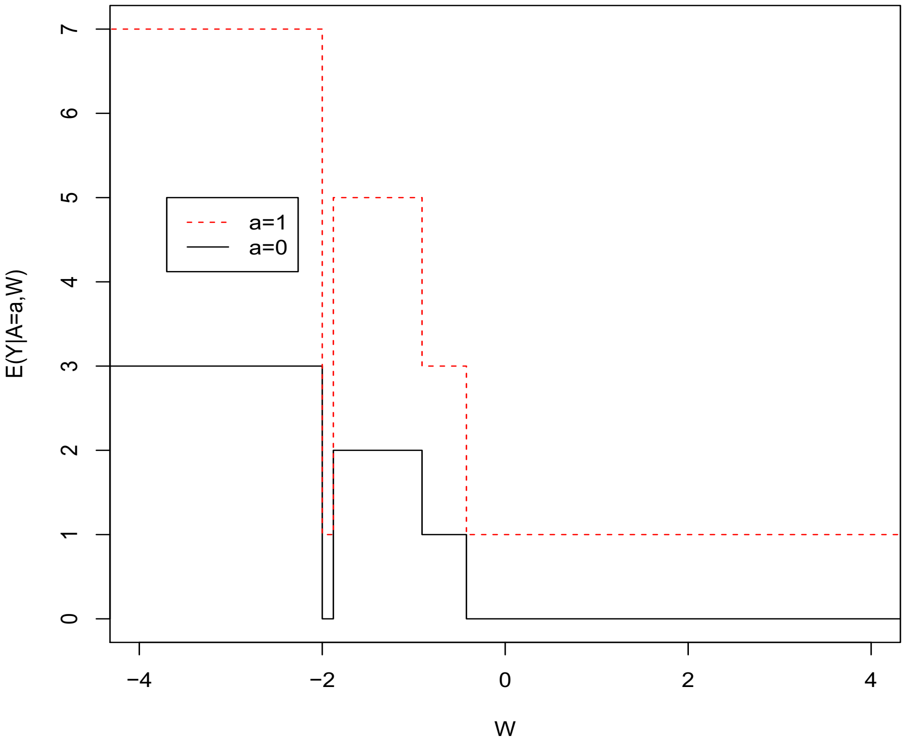

Figure 4: $E(Y \mid A = a, W)$ for the ATE simulations

thousands of candidates in the data. 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 $O = (A, Y = (Y_1, Y_2, ..., Y_p))$ where A is a binary vector of zeros and ones, and Y is a multivariate outcome. The true distribution, P_0 is generated based on a design where there are equal numbers of $A = 0$ and $A = 1$, and for each (gene) j, the distribution of Y_j, given A, is defined by the following regression equation

$$Y_j = B_{0j} + B_{1j}A + e_j \quad j = 1, \ldots, p. \tag{2}$$

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 $\text{Cov}(B_0, B_1) = 1 \quad i = j$ and $\text{Cov}(B_0, B_1) = .2 \quad i \neq j$. 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 $N(0, \sigma_e^2)$ distribution, and we repeated the simulation both for different magnitudes of the residual error, (different σ_e^2) but also for increasing sample sizes.

We define our data adaptive parameter as:

$$\Psi_{P^0_{n,B_n}}(P_0) = E_{B_n} E_{P_0}(Y^*_{P^0_{n,B_n}} \mid A = 1) - E(Y^*_{P^0_{n,B_n}} \mid A = 0), \tag{3}$$

where

$$Y^*_{P^0_{n,B_n}} = \frac{1}{\sum_j I(j \in \mathcal{S}_{P^0_{n,B_n}})} \sum_j I(j \in \mathcal{S}_{P^0_{n,B_n}}) Y_j$$

is an average of the gene-expression across a subset of genes, where this subset, $\mathcal{S}^0_{P_{n,B_n}}$ is determined by a procedure on the parameter-generating sample. Specifically, for each parameter-generating sample we simply rank the genes by $\hat{B}_{1j} = E_{P^0_{n,B_n}}(Y_j \mid A = 1) - E_{P^0_{n,B_n}}(Y_j \mid A = 0)$, and the set $\mathcal{S}_{P^0_{n,B_n}}$ is defined the top 15 j's according to this ranking.

The estimator of (3) based on the estimation sample is simply

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

and its influence curve is estimated as follows

$$IC_{P^0_{n,B_n}}(Y^*, A) = \left[\frac{I(A = 1)}{P^1_{n,B_n}(A = 1)} - \frac{I(A = 0)}{P^1_{n,B_n}(A = 0)}\right](Y^* - E_{P^1_{n,B_n}}(Y^* \mid A)) \qquad (4)$$

where in this case $P^1_{n,B_n}(A = 1) = 0.5$ by design.

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. Note that the stability in the sets $\mathcal{S}_{P^0_{n,B_n}}$ across samples will decrease as σ^2_e increases. Thus, this provides a choice of algorithm for generating the parameter and a data generating distribution that allows us to examine the relative performance of the three algorithms as the parameter generating algorithm becomes more variable.

4.4.1 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 the other hand, for algorithms two and three, that have an overlap in their parameter-generating and estimation samples, proper coverage is not obtained until relatively large sample sizes. Part of this has to do with bias, and that can be seen in the distribution of standardized estimates for the different algorithms (see Figure 5). The bias of algorithm 3 only becomes more severe as the procedure defining the target parameter becomes more variable (see Table 3). For all simulations, algorithm 1 shows very good performance with regards to statistical inference, while apparently not having greater sampling variability.

Table 3: Simulation results for *Variable Reduction* for the algorithms based on theorems 1-3 ($\psi_n, \psi_n^1, \psi_n^2$, repsectively).

Samples	Methods	True average	Estimated average	Cov Prob
n=100	ψ_n	2.243	2.247	0.826
	ψ_n^1	2.243	2.584	0.006
	ψ_n^2	2.632	2.568	0.014
n=500	ψ_n	2.483	2.484	0.912
	ψ_n^1	2.483	2.553	0.662
	ψ_n^2	2.487	2.557	0.686
n=1000	ψ_n	2.515	2.511	0.916
	ψ_n^1	2.515	2.584	0.774
	ψ_n^2	2.515	2.568	0.798
n=2000	ψ_n	2.556	2.558	0.945
	ψ_n^1	2.556	2.577	0.864
	ψ_n^2	2.557	2.578	0.876
n=10000	ψ_n	2.515	2.514	0.956
	ψ_n^1	2.515	2.519	0.949
	ψ_n^2	2.515	2.519	0.941

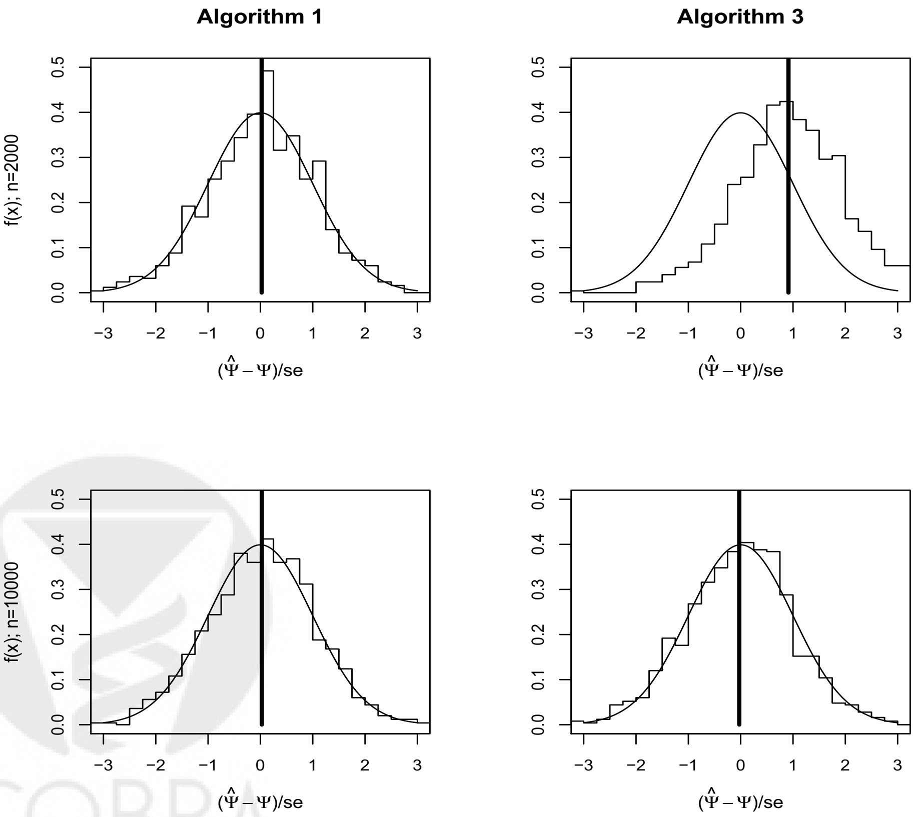

Figure 5: Results of the simulation based on variable reduction, for $\sigma_e^2 = 2$. Distribution of $(\psi_n - \psi_{n,0})/se(\psi_n)$ for $n = 2000$ and $n = 10000$, algorithms 1 and 3. Also displayed is the N(0,1) distribution for comparison; the dark line represents the mean of these standardized values, so the difference between it and 0 is the standardized bias.

5 Some concluding remarks

Much scientific progress has been obtained by generating target parameters based on past studies, and evaluating them on future studies. However, such costly splitting of a stream of data is by no means necessary, and 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 splits. Our formal results show that statistical inference is preserved under minimal conditions, even though the estimators are now based on all the data. The price one pays is that the statistical target parameter is an average of parameters generated by the different sample splits. To obtain valid finite sample inference it is is important to utilize our corresponding variance estimator (1), 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 the influx of 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. There are a large variety of such applications, including ones that use 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, this method should be considered as an important method.

Thus, gains in efficiency can be obtained with algorithm 2 (ψ_n^1), while further gains in interpretability of the target parameter can be obtained with algorithm 3 (ψ_n^2), but the asymptotic validity of both of these algorithms now relies on this Donsker class condition. One expects that a practical implication of this Donsker class condition is that the asymptotic normality will kick in at later sample sizes than it would for algorithm 1, but this requires further study as well. In particular, we observed in our simulation examples that these two algorithms 2 and 3 might suffer from higher levels of finite sample bias. Nonetheless, given the big data era, all three methods are important and should be considered. In future work, we plan to address diagnostic tools that will assist the user to make this choice between these algorithms, possibly based on running a semi parametric bootstrap.

In this article we demonstrated the methodology and theory for a few examples. These examples demonstrate that data adaptive target parameters provide new approaches for analyzing data sets, allowing us to target parameters one would not be able to a priori specify, allowing to tackle multiple testing problems by allowing the data adaptive generation of relatively few null hypotheses of interest. Instead of having to define a target parameter mapping, one can focus on thinking about a mapping that maps data into a target parameter. There are many examples of interest that have not been highlighted in this article. For example, one may select a target parameter based on its practical level of identifiability, thereby arraigning that no power is wasted on target parameters that are not supported by the data.

References

Helen Barraclough and Ramaswamy Govindan. Biostatistics primer: what a clinician ought to know: subgroup analyses. *Journal of Thoracic Oncology*, 5(5):741, 2010.

Bonnie Berger, Jian Peng, and Mona Singh. Computational solutions for omics data. *Nat Rev Genet*, 14(5):333–46, May 2013. doi: 10.1038/nrg3433.

David I Broadhurst and Douglas B Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171–196, 2006.

Erin Le Dell, M. Petersen, and M.J. van der Laan. 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.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2:131–154, 2005.

Andrew Gelman, Yu-Sung Su, Masanao Yajima, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman, and Tian Zheng. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2012. URL `http://CRAN.R-project.org/package=arm`. R package version 1.5-08.

T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman and Hall, New York, 1990.

John P A Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–8, Sep 2008. doi: 10.1097/EDE.0b013e31818131e7.

John R Marler. Secondary analysis of clinical trials—a cautionary note. *Progress in cardiovascular diseases*, 54(4):335–337, 2012.

D.B. Rubin. Bayesian inference for causal effects: the role of randomization. *Ann. Statisti.*, 6:34–58, 1978.

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

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Superlearner. *Statistical Applications in Genetics & Molecular Biology*, 6, 2007.

Mark J Van Der Laan. Efficient estimation in the bivariate censoring model and repairing npmle. *The Annals of Statistics*, 24(2):596–627, 1996.

M.J. van der Laan and D.B. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2006. URL `http://www.bepress.com/ijb/vol2/iss1/11`. Article 11.

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

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

Fan Zhang and Jake Y Chen. Data mining methods in omics-based biomarker discovery. *Methods Mol Biol*, 719:511–26, 2011. doi: 10.1007/978-1-61779-027-0_24.