Statistical Inference for Data Adaptive Target Parameters

Mark van der Laan, Alan Hubbard

Division of Biostatistics, UC Berkeley

December 13, 2013

- **Have one's cake**
  - → Define a class of parameters, and general estimation method, so that one can use the data to both define the question of interest (parameter of interest).
- **And eat it too**
  - → using the same data to estimate and draw inferences about that parameter

Methods based on Defining a class of target parameters and a method that maps the data into a choice of target parameter

Dredging with dignity

- The goal is deriving honest inference for interesting statistical parameters that are not defined before examining the data
- Based on different versions of a cross-validation approach that have different assumptions on the algorithm that generates the target parameter
  - → Use the training samples to define parameter of interest, and estimate and derive inference of this parameter on the validation samples.
- Depending on how aggressive the procedure is to find the pattern (parameter) of interest then in the extreme cases,
  - → one must keep the pattern finding, and estimation/inference completely separate to
  - → use the same (entire) data set for both

See van der Laan et al. (2013).

Examples

Prediction

- Comparing the future prediction accuracy of prediction algorithms
- Using prediction methods for global tests of the association of large collection of variables and an outcome.

Sub-group analysis

- Estimating the impacts of treatment (exposure) in non-pre-specified groups.

Parametric Regression Analysis

- Data adaptive selection of parametric regression model for estimating adjusted association.

Causal Inference

- Data adaptive estimation of treatment mechanism, for implementation of estimators, and/or defining the causal parameter (e.g., defining natural direct effect as stochastic intervention based on mediator distribution estimated from training sample).

Formal Set-up

- $O_1, \ldots, O_n$ be i.i.d. with probability distribution P_0 known to be an element of a statistical model $\mathcal{M}$.
- $B_n \in \{0,1\}^n$ 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 ease, let B_n corresponds with V-fold cross-validation scheme: i.e., $\{1, \ldots, n\}$ are divided in V equal size subgroups,
  - $\rightarrow$ estimation-sample defined by one subgroup and the parameter-generating sample is its complement.
  - $\rightarrow$ Thus, there are V such combination of parameter-generating and estimation samples.

Definition of Data Adaptive Parameter

- For split B_n,
  - $\rightarrow$ P^0_{n,B_n} = empirical distribution of the **parameter-generating sample**
  - $\rightarrow$ P^1_{n,B_n} the empirical distribution of the **estimation-sample.**
- $\Psi_{B_n,P^0_{n,B_n}} : \mathcal{M} \rightarrow \mathbb{R}$ be a parameter mapping
- $\hat{\Psi}_{B_n,P^0_{n,B_n}} : \mathcal{M}_{NP} \rightarrow \mathbb{R}$ be a corresponding **estimator of this parameter**.

The Statistical Estimand

- The data adaptive estimand $\psi_{n,0}$ is given by:

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

- **Goal of theorem** - Prove that $\sqrt{n}(\psi_n - \psi_{n,0})$ converges in distribution to mean zero normal distribution with variance σ^2, which can be consistently estimated.
- True if $\psi_n = \hat{\Psi}(P_n)$ is an asymptotically linear estimator of $\psi_{n,0}$ with influence curve $IC(P_0)$:

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

- Implies that $\sqrt{n}(\psi_n - \psi_{n,0})$ converges to a mean zero normal distribution with variance $\sigma^2 = P_0 IC(P_0)^2$, where $P_0 f \equiv \int f(o) dP_0(o)$ represents expectation operator w.r.t. P_0.

No splitting - Parameter defined and estimated on same data

Theorem

Assume that $\hat{\Psi}_P(P_n)$ 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 that $P_0(IC_{P_n}(P_0) - IC_{P_0}(P_0))^2 \to 0$ in probability and $IC_{P_n}(P_0)$ is an element of a P_0-Donsker class with probability tending to 1. Then,

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

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

Implications

- Given one is estimating a data-adaptive parameter defined by one data-adaptive operation, and not an average of V data-adaptive parameters, it is potentially easier to interpret and explain to others.
- Relatively strong assumptions on the adaptability of the algorithm, which produces data adaptive parameter.

CLT for Cross-Validated Approach

- 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 $IC_{B_n,P^0_{n,B_n}}(P_0)$:

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

- Assume: $P_0(IC_{v,P^0_{n,v}}(P_0) - IC_v(P_0))^2 \to 0$ in probability, where $IC_v(P_0)$ is a limit influence curve that can still be indexed by the split v.
- Then,

$$\sqrt{n}(\psi_n - \psi_{n,0}) = \frac{1}{V}\sum_v \sqrt{V}\sqrt{n/V}(P^1_{n,B_n} - P_0)IC_{B_n,P^0_{n,B_n}}(P_0) + o_P(1/\sqrt{n})$$

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

Algorithm and Implications of Theorem

- Use the training sample to define the parameter and then uses the corresponding estimation (validation) sample to estimate the parameter.
- Average the estimates of these data-adaptive parameters over the V estimation samples.
- Derives inference via the influence curve on each validation sample and uses these n realizations (overall all validation samples) to derive an estimate of the sample variance.
- If data adaptive parameter is *interesting*, a simple mechanism to use the data to define an interesting parameter (exploratory analysis) and then estimate and derive consistent inference (confirmatory analysis).
- Has a CV-TMLE augmentation that increases efficiency and "smoothness" of estimator.

Conclusions

- Data adaptive parameters based on this cross-validated approach open up a new set of interesting parameters.
- New definition of parameter of inference can retrieve straightforward inference whereas the equivalent non-data-adaptive parameter of full data can be difficult to derive sampling distribution.
- Covers naturally many standard practices involving model selection (e.g., type of backward selection based on change in coefficient methods).
  - → Can add formal inference to current data-adaptive techniques currently lacking them.
- Lots of interesting applications in bioinformatics, high dimensional clinical data, etc.
- But, work needs to be done on when various methods will yield trustworthy inference.

Mark J. van der Laan, Alan E Hubbard, and Sara Kherad Pajouh. Statistical inference for data adaptive target parameters. Technical Report 314, U.C. Berkeley Division of Biostatistics Working Paper Series, 2013. URL http://biostats.bepress.com/ucbbiostat/paper314.