Regularity and adaptive debiased machine learning

Journal club'ish inspired by
Adaptive debiased machine learning using data-driven model selection techniques [van der Laan, Carone, Luedtke, and van der Laan, 2023]

Anders Munch

May 8, 2024

Short experience with ~~Twitter~~

Mark van der Laan
@mark_vdlaan

Paradigm shifting work by @Larsvanderlaan3!
It offers a new way of thinking about nonparametric causal inference and efficient estimation, particularly in challenging settings, such as limited treatment overlap.

Motivation

Build *practically* useful estimators that can be used with high-dimensional longitudinal register data. Is the targeted learning philosophy sometimes too much "asymptopia" (mathematically correct limit results, but practically useless)?

Understand of regularity.

The targeted learner and the modeler

Specify your target parameter and your model. Also pre-specify your estimator.

> I don't yet know enough about the data, I need to estimate some models and see which fit.

You shouldn't do that, use a super learner to let the computer decide these things instead.

> I am not sure if that will work in practice. I will start with a simple model, and if that looks OK, I will stick with it.

Your approach leads to invalid statistical inference!

> Your approach is too likely to lead to enormous CI or even estimates that are `NA`!

Acknowledge limitations of "off the shelf" TL

Simple stating that we estimate $\Psi : \mathcal{M} \to \mathbb{R}$ under a non-parametric model might be too optimistic/simplistic.

There can be problems that we had not thought about (e.g., positivity problems for specific subgroups).

For complex data with many variables (e.g., longitudinal register data), we might not be able to solve this problem well enough for practical sample sizes.

It might not be possible to construct an estimator which is valid across the fully nonparametric model *and* works well in practice for reasonable sample sizes.

From the abstract

Debiased machine learning estimators for nonparametric inference of smooth functionals of the data-generating distribution can suffer from excessive variability and instability. For this reason, practitioners may resort to simpler models based on parametric or semiparametric assumptions. However, such simplifying assumptions may fail to hold, and estimates may then be biased due to model misspecification. To address this problem, we propose Adaptive Debiased Machine Learning (ADML) [...]. By learning model structure directly from data, ADML avoids the bias introduced by model misspecification and remains free from the restrictions of parametric and semiparametric models. [...]

van der Laan et al. [2023]

Intermezzo on semi-parametric efficiency theory

RAL estimators and semi-parametric efficiency theory

An estimator $\hat{\Psi}_n$ of $\Psi : \mathcal{M} \to \mathbb{R}$ is *asymptotically linear* (AL) if for all $P \in \mathcal{M}$ there is a function $\varphi(\cdot; P) \in \mathcal{L}^2_P$ with $P[\varphi(\cdot; P)] = 0$ such that

$$\hat{\Psi}_n - \Psi(P) = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i; P) + o_P(n^{-1/2}).$$

The function $\varphi(\cdot; P)$ is the *influence function* (IF) of the estimator.

AL estimators can compared by comparing the norm of their IFs. The *efficient influence function* (EIF) is the IF with smallest norm.

Expect that an estimator which has the EIF as its IF is "optimal" in some sense.

A further regularity condition is needed for this to be true – the estimator should be *regular*.

Regularity

An estimator $\hat{\Psi}_n$ of $\Psi : \mathcal{M} \to \mathbb{R}$ is *regular* if its asymptotic distribution is invariant to local perturbations of the data generating mechanism.

Formally: A one-dimensional submodel $\{P_t : t \in \mathbb{R}\} \subset \mathcal{M}$ through P at $t = 0$ is *regular* if it is differentiable in quadratic mean at $t = 0$.

For $h \in \mathbb{R}$ and a regular submodel $\{P_t : t \in \mathbb{R}\}$, the sequence $P_{hn^{-1/2}}$ is a local perturbation of P.

An estimator $\hat{\Psi}_n$ is regular for the parameter Ψ with respect to the local perturbation $P_{hn^{-1/2}}$ if

$$\sqrt{n}(\hat{\Psi}_n - \Psi(P_{hn^{-1/2}})) \rightsquigarrow \mathcal{L}_P,$$

for some distribution $\mathcal{L}_P$ that does not depend on h or the path $\{P_t : t \in \mathbb{R}\}$, when $\hat{\Psi}_n$ is constructed with samples from taken from $P_{hn^{-1/2}}$.

An estimator $\hat{\Psi}_n$ is P-regular for the parameter Ψ over $\mathcal{M}$ if it is regular with respect to all local perturbation of P within $\mathcal{M}$.

Hodges' classical example of a non-regular estimator

Following [Tsiatis, 2007, chapter 3.1], let $X_i \sim \mathcal{N}(\mu, 1)$, $\mu \in \mathbb{R}$, be iid. for $i = 1, \ldots, n$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ and define

$$\hat{\mu}_n = \begin{cases} \hat{X}_n & \text{if } |\hat{X}_n| > n^{-1/4} \\ 0 & \text{if } |\hat{X}_n| \leq n^{-1/4} \end{cases}.$$

$\sqrt{n}(\bar{X}_n - \mu)$ has limiting distribution $\mathcal{N}(0, 1)$ for all μ; as this is the MLE it is efficient.

However, $\sqrt{n}(\hat{\mu}_n - \mu)$ has the same asymptotic distribution for all $\mu \neq 0$, and asymptotic distribution $\mathcal{N}(0, 0) = 0$ for $\mu = 0$.

$\hat{\mu}_n$ appears to beat the MLE $\bar{X}_n$ – it is *super-efficient*.

Super-efficient estimators and irregularity

Picture from Wikipedia [2024]

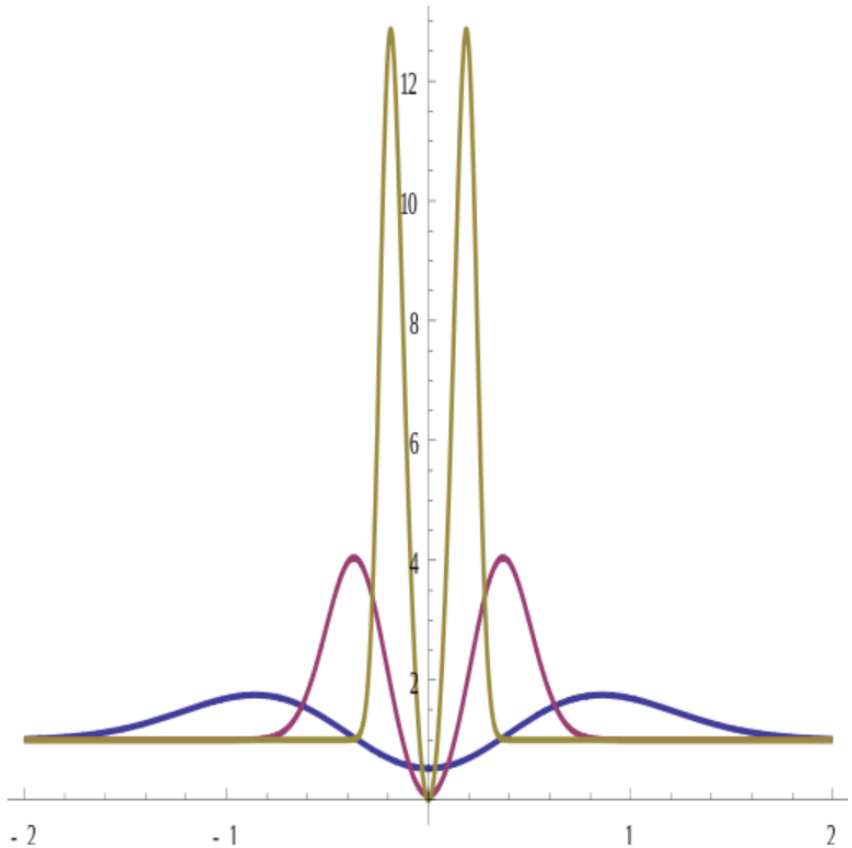

Figure: The mean square error (times n) of Hodges' estimator as a function of μ. Blue curve corresponds to $n = 5$, purple to $n = 50$, and olive to $n = 500$.

Data-adaptive estimators and irregular estimators

$$\hat{\mu}_n = \begin{cases} \hat{X}_n & \text{if } |\hat{X}_n| > n^{-1/4} \\ 0 & \text{if } |\hat{X}_n| \leq n^{-1/4} \end{cases}.$$

Think of the Hodges' estimator as a *data-adaptive estimator* that work in two steps:

1. Conduct a test for whether the mean is 0.
2. If we accept, return 0, otherwise return the empirical average.

Leeb and Pötscher [2005] argues that post-model selection estimators are versions of Hodges' estimator.

Revisit (annoying) regularity condition

Asymptotic linearity is easy to motivate. Regularity not so much: "Why should we care about data coming from a local perturbation – we usually just assume that data are iid. from some fixed P?" ...

Revisit (annoying) regularity condition

Asymptotic linearity is easy to motivate. Regularity not so much: "Why should we care about data coming from a local perturbation – we usually just assume that data are iid. from some fixed P?" ...

Quotes about regularity

Although super-efficient estimators exist, they are unnatural and have undesirable local properties associated with them. [...] From now on, we will restrict ourselves to regular estimators.

Tsiatis [2007]

This type of regularity is common and is often considered desirable: A small change in the parameter should not change the distribution of the estimator too much; a disappearing small change should not change the (limit) distribution at all. However, some estimator sequences of interest, such as shrinkage estimators, are not regular.

van der Vaart [2000]

[...] the suggested estimator [...] will – although being consistent – not be close to the finite-sample distribution uniformly in the unknown parameters, thus providing a rather useless estimator.

Leeb and Pötscher [2005]

Such criticisms of superefficient estimators may not be as applicable in situations in which regular nonparametric estimators do not exist or are too variable for reliable inference.

van der Laan et al. [2023]

Back to the paper

ADMLE – the central idea

Let Ψ be a target parameter defined on a collection of probability measures $\mathcal{M}$.

We assume that $P_0 \in \mathcal{M}_0$ for some *oracle submodel* $\mathcal{M}_0 \subset \mathcal{M}$, but we don't know $\mathcal{M}_0$.

Estimate the submodel $\mathcal{M}_0$ from data with $\mathcal{M}_n$ and define a projected target parameter Ψ_n data-adaptively using $\mathcal{M}_n$.

Construct an efficient estimator of the data-adaptive parameter Ψ_n.

If $\mathcal{M}_0$ is much smaller than $\mathcal{M}$ we expect a sizable decrease in variance.

Example of oracle model

Assume that the data is $X = (Y, A, W)$ and that we want to estimate the average treatment effect

$$\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, W] - \mathbb{E}_P[Y \mid A = 0, W]]$$

An oracle model could be

$$\mathcal{M}_0 = \left\{ \begin{array}{c} \mathbb{E}_P[Y \mid A = a, W = w] = \alpha a + \beta^T w : \alpha \in \mathbb{R}, \beta \in \mathbb{R}^d, \\ P(\mathrm{d}a, \mathrm{d}w) \text{ unspecified} \end{array} \right\}$$

Important point is that we do not assume a known model $\mathcal{M}_0$, but only that assume that the data-generating distribution P_0 actually belongs to some (unknown) smaller model $\mathcal{M}_0 \subset \mathcal{M}$.

ADMLE – more formally

1. Specify data-adaptive working models $\mathcal{M}_n \subset \mathcal{M}$ intended to approximate $\mathcal{M}_0$.
2. For some loss function $\ell \colon \mathbb{R}^d \times \mathcal{M} \to \mathbb{R}$, define the projection of P onto the working model $\mathcal{M}_n$ as

$$\Pi_n(P) = \underset{Q \in \mathcal{M}_n}{\operatorname{argmin}} \int \ell(x; Q) P(\mathrm{d}x).$$

3. Define the data-adaptive target parameter $\Psi_n = \Psi \circ \Pi_n \colon \mathcal{M} \to \mathbb{R}$.
4. Construct an efficient estimator of Ψ_n.

An ADMLE $\hat{\Psi}_n$ is an estimator that satisfies the asymptotic expansion

$$\hat{\Psi}_n = \Psi(P_0) + (\mathbb{P}_n - P_0)[D_{n,P_0}] + o_P(n^{-1/2})$$

where D_{n,P_0} is the nonparametric efficient influence function of the data-adaptive target parameter Ψ_n.

Data-adaptive and *oracle* target parameters

Different target parameters are in play:

$\Psi : \mathcal{M} \to \mathbb{R}$ The original target parameter.

$\Psi_n = \Psi \circ \Pi_n$ Data-adaptive target parameter

$\Psi_0 = \Psi \circ \Pi$ Oracle project-based target parameter.

The operator Π is defined as the projection onto the true (but unknown) oracle submodel,

$$\Pi(P) = \underset{Q \in \mathcal{M}_0}{\operatorname{argmin}} \int \ell(x; Q) P(\mathrm{d}x).$$

The oracle target parameter

Model selection is known to produce irregular estimators, even when the model selection step is consistent [Leeb and Pötscher, 2005].

van der Laan et al. [2023] circumvent this issue by redefining the target to be the *oracle target parameter* $\Psi_0 = \Psi \circ \Pi$.

While an ADMLE is irregular for the original parameter Ψ, they show that it is RAL for Ψ_0 (at any $P_0 \in \mathcal{M}_0$).

Ψ_0 is involves the projection onto the *unknown* oracle model $\mathcal{M}_0$. If we don't know $\mathcal{M}_0$ we don't really know Ψ_0 – is that weird?

Maybe not if we accept that we are trying to "learn what can be learned" from the data.

One of the main results

Theorem (Theorem 5)

Suppose that the working model $\mathcal{M}_n$ approximates $\mathcal{M}_0$ fast enough (sort of like at $\approx n^{-1/4}$ rate) and that additional regularity conditions hold. Then, the ADMLE $\hat{\Psi}_n$ is a P_0-asymptotically linear estimator for Ψ_0 with influence function equal to the efficient influence function of $\Psi_0 : \mathcal{M}_{\text{np}} \to \mathbb{R}$ at P_0 relative to the nonparametric model $\mathcal{M}_{\text{np}}$.

> *An important consequence of Theorem 5 is that an ADMLE is a P_0-regular estimator for Ψ_0 relative to the nonparametric model $\mathcal{M}$. Hence, even under sampling from a worst-case local perturbation of P_0, an ADMLE allows locally uniformly valid nonparametric inference on the oracle parameter Ψ_0. This implies that, at least in a local asymptotic sense, there is no loss in performance of the ADMLE from empirically learning $\mathcal{M}_0$ compared to the oracle that knows $\mathcal{M}_0$ or Ψ_0.*
>
> *van der Laan et al. [2023]*

The estimator – details are lacking

Formally, an ADMLE $\hat{\Psi}_n$ is an estimator that satisfies the asymptotic expansion

$$\hat{\Psi}_n = \Psi(P_0) + (\mathbb{P}_n - P_0)[D_{n,P_0}] + o_P(n^{-1/2})$$

where D_{n,P_0} is the nonparametric efficient influence function of Ψ_n.

van der Laan et al. [2023]

Not so clear how to implement an ADML in practice. Example with relaxed lasso in + standard model robust sandwich estimator of the variance.

Model selection

> *Suppose that we can employ data-driven model selection techniques to learn a working statistical model $\mathcal{M}_n \subset \mathcal{M}$ that sufficiently approximates some unknown submodel $\mathcal{M}_0 \subset \mathcal{M}$. Although the working model $\mathcal{M}_n$ may not contain the true data-generating distribution P_0 for any n, we assume that $\mathcal{M}_0$ is a smooth statistical model containing P_0. The smoothness condition on $\mathcal{M}_0$ rules out degenerate models such as $\mathcal{M}_0 = \{P_0\}$.*
>
> *van der Laan et al. [2023]*

When to stop? How do we make sure not to be to aggressive in the model selection step? Nor to conservative?

What to do if we are willing to accept an approximate model $\mathcal{M}_n$ that does not contain P_0?

For any P_0 there are many smooth models (nested and non-nested) which can be assumed to contain P_0.

What is the cost of giving up regularity?

Imposing a smaller submodel $\mathcal{M}_0$ can introduce bias but reduces variance $\rightarrow$ classic bias-variance trade-off.

Working with the projection-based oracle parameter Ψ_0 brings the deal back to a classic bias-variance trade-off.

If we consider the original target parameter Ψ, the ADMLE will be irregular $\rightarrow$ here we are trading variance for something else – irregularity.

> *Sacrificing some regularity can be justifiable to achieve efficiency gains, especially when nonparametric regular estimators for Ψ are unavailable, such as when the ATE is nonparametrically unidentifiable.*
>
> *van der Laan et al. [2023]*

What is the cost? What deal are we making?

Is regularity a binary concept?

Theorem (Theorem 6)

... $\hat{\Psi}_n$ (the ADMLE) is P_0-regular for Ψ (the original target parameter) over all local alternatives $P_{0,hn^{-1/2}}$ in the oracle submodel $\mathcal{M}_0$. Consequently, $\sqrt{n}(\hat{\Psi}_n - \Psi_0) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}[D_{0,P_0}(O)])$, even under sampling from local perturbations of P_0 remaining in $\mathcal{M}_0$.

> *Theorem 6 shows that the regularity and superefficiency of ADMLEs fall in a continuous spectrum driven by the size of the oracle model.*
>
> *van der Laan et al. [2023]*

Maybe Hodge's estimator is extreme with $\mathcal{M}_0 = \{P_0\}$. Can we expect less irregularity if $\mathcal{M}_0$ is large? Could it be OK to give up some regularity, i.e., not be stable under all local perturbations but under some?

Conclusions

- It might be "asymptopia" to say that we estimate ATE in the non-parametric model when we have high-dimensional longitudinal data.
- Sometimes we will be forced to impose restrictions/constraints to make things work in practice (and maybe even in theory).
- Can we ask the data how complex the questions we ask can be?
- Is it a good idea to try to automate an formalize this process – or should we just work heuristically?
- Should we be willing to give up (some degree of) regularity?

References

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

L. van der Laan, M. Carone, A. Luedtke, and M. van der Laan. Adaptive debiased machine learning using data-driven model selection techniques. *arXiv preprint arXiv:2307.12544*, 2023.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wikipedia. Hodges' estimator — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Hodges'%20estimator&oldid=1216101801`, 2024. [Online; accessed 07-May-2024].