Orthogonal Machine Learning: Power and Limitations

Lester Mackey   Vasilis Syrgkanis   Ilias Zadik

Microsoft Research New England (MSR) & Massachusetts Institute of Technology (MIT)

Motivation

The increased availability of large and complex observational datasets motivates the study of **treatment effects** in the presence of high-dimensional data. As a running application, consider demand estimation from pricing and purchase data in the digital economy.

The Generalized Method of Moments (Main Tool)

Setup: For some unknown target parameter $\theta_0 \in \mathbb{R}^d$ we are given access to independent replicates $(Z_t)_{t=1}^{2n}$ of a random data vector $Z \in \mathbb{R}^\rho$ drawn from a distribution satisfying d moment conditions,

$$\mathbb{E}[m(Z, \theta_0, h_0(X)) | X] = 0, \text{ a.s.} \tag{1}$$

Here, $h_0 : \mathbb{R}^\mu \to \mathbb{R}^\ell$ is a vector of ℓ unknown nuisance functions, $X \in \mathbb{R}^\mu$ is a sub-vector of the observed data vector Z, and $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is a vector of d known moment functions.

Question: Can we find a $\sqrt{n}$-consistent and asymptotically normal ($\sqrt{n}$-a.n) estimator of θ_0, that is, an estimate $\hat{\theta}$ satisfying $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N(0, \Sigma)$ for some covariance matrix Σ?

Sample Splitting and Two-stage Estimation

We conduct a two-stage estimation procedure with sample splitting, following [1].

1. **First stage.** Form an estimate $\hat{h}$ of h_0 using $(Z_t)_{t=n+1}^{2n}$ (e.g., by running a non-parametric or high-dimensional regression procedure).
2. **Second stage.** Compute a Z-estimate $\hat{\theta}$ of θ_0 using an empirical version of the moment conditions (1) and $\hat{h}$ as a plug-in estimate of h_0:

$$\hat{\theta} \text{ solves } : \frac{1}{n}\sum_{t=1}^{n} m(Z_t, \hat{\theta}, \hat{h}(X_t)) = 0. \tag{2}$$

Main Question: How accurately do we need to learn the nuisance functions h_0 in the first stage, so that the solution of (2) is a $\sqrt{n}$-a.n estimator of θ_0? [Ideally, it would suffice to estimate h_0 at a slower than $o\left(n^{-\frac{1}{2}}\right)$ rate!]

Prior Work: Neyman Orthogonality

Definition 1 (First-Order Orthogonality) *A vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is first-order orthogonal with respect to the nuisance function if:*

$$\mathbb{E}\left[\nabla_\gamma m(Z, \theta_0, \gamma)|_{\gamma = h_0(X)} | X\right] = 0.$$

Here, $\nabla_\gamma m(Z, \theta_0, \gamma)$ is the gradient of the vector of moments with respect to its final ℓ arguments.

Theorem 1 *([1]) Suppose that*
- *m is a "smooth" enough function*
- *m satisfies the first-order orthogonality condition*
- *$\mathbb{E}[m(Z, \theta, h_0(X))] \neq 0$, when $\theta \neq \theta_0$* ***Identifiability constraint!***

Then if we can the learn each function $h_{0,i}(X), i = 1, 2, \ldots, \ell$ at rate $o\left(n^{-\frac{1}{4}}\right)$, the $\hat{\theta}$ defined by (2) is a $\sqrt{n}$-a.n. estimator of θ_0.

Main Result and k-th-Order Orthogonality

Idea: Generalize orthogonality to k-th-order derivatives to accommodate $o\left(n^{-\frac{1}{2(k+1)}}\right)$ first-stage estimation rates!

Definition 2 (k-Orthogonality of Moments) *A vector of moments $m : \mathbb{R}^\rho \times \mathbb{R}^d \times \mathbb{R}^\ell \to \mathbb{R}^d$ is called k-orthogonal if for any $\alpha \in \mathbb{N}^\ell$ with $\|\alpha\|_1 \leq k$:*

$$\mathbb{E}\left[D^\alpha m(Z, \theta_0, \gamma)|_{\gamma = h_0(X)} \middle| X\right] = 0 \tag{3}$$

where

$$D^\alpha m(Z, \theta, \gamma) := \nabla_{\gamma_1}^{\alpha_1} \nabla_{\gamma_2}^{\alpha_2} \ldots \nabla_{\gamma_\ell}^{\alpha_\ell} m(Z, \theta, \gamma). \tag{4}$$

Theorem 2 (Main Result) *Suppose that*
- *m is a "smooth" enough function*
- *m satisfies the k-th-order orthogonality condition**
- *$\mathbb{E}[m(Z, \theta, h_0(X))] \neq 0$, when $\theta \neq \theta_0$* ***Identifiability constraint!***

Then if we can the learn each function $h_{0,i}(X), i = 1, 2, \ldots, \ell$ at a $o\left(n^{-\frac{1}{2(k+1)}}\right)$ rate, the $\hat{\theta}$ defined by (2) is a $\sqrt{n}$-a.n. estimator of θ_0.

*A more general version of the theorem dealing with a weaker condition than k-th-order orthogonality can be found in the paper.

The Partially Linear Regression (PLR) Model

A good model for the pricing application!

Definition 3 (Partially Linear Regression (PLR))
In the partially linear regression model of observations $Z = (T, Y, X)$, $T \in \mathbb{R}$ represents a treatment or policy applied, $Y \in \mathbb{R}$ represents an outcome of interest, and $X \in \mathbb{R}^p$ is a vector of associated covariates. These observations are related via the equations

$$Y = \theta_0 T + f_0(X) + \epsilon, \quad \mathbb{E}[\epsilon \mid X, T] = 0$$
$$T = g_0(X) + \eta, \quad \mathbb{E}[\eta \mid X] = 0$$

where ϵ, η represent unobserved noise variables.

Question: Can we accommodate a slower rate than $o\left(n^{-\frac{1}{2}}\right)$ in the first stage and still be $\sqrt{n}$-a.n. in estimating θ_0 via (2)?

Literature: Yes! For nuisance $q_0(X) := f_0(X) + \theta_0 g_0(X), g_0(X)$, $o\left(n^{-\frac{1}{4}}\right)$ first stage error suffices; Theorem 1 works for

$$m(Z, \theta_0, q(X), g(X)) = (Y - q(X) - \theta_0 (T - g(X)))\,(T - g(X))\, [= \epsilon\eta]\,.$$

Main Result on PLR: We can improve our first stage error requirement using second-order orthogonality if and only if the distribution of η, conditional on X, is **not** Gaussian!

Impossibility result:

Theorem 3 (Gaussian Limitation) *Suppose η, conditional on X, follows a Gaussian distribution. There is no 2-orthogonal moment condition m the random variable $\hat{\theta}$, defined by (2), which satisfies the identifiability constraint and $|\hat{\theta} - \theta_0| = O_P(n^{-\frac{1}{2}})$.*

The result is based on the **Stein's lemma**: for ζ mean-zero Gaussian and f differentiable, $\mathbb{E}[\zeta^2]\mathbb{E}[f'(\zeta)] = \mathbb{E}[\zeta f(\zeta)]$, which **uniquely characterizes** the Gaussian distribution.

Positive result:

Theorem 4 (Non-Gaussian Power) *Suppose for some $r \in \mathbb{N}$, $\mathbb{E}[\eta^r|X]$ is known* and $\mathbb{E}[\eta^{r+1}|X] \neq r\mathbb{E}[\eta^2|X]\mathbb{E}[\eta^{r-1}|X]$ (so that $\eta|X$ does not follow a Gaussian distribution). Then the moment condition*

$$m\left(T, Y, \theta, q(X), g(X), \mathbb{E}[\eta^{r-1}|X]\right) :=$$
$$(Y - q(X) - \theta\,(T - g(X)))\left((T - g(X))^r - \mathbb{E}[\eta^r|X] - r\,(T - g(X))\,\mathbb{E}[\eta^{r-1}|X]\right)$$

is 2-orthogonal and satisfies the assumptions of Theorem 2. Hence, the random variable $\hat{\theta}$, defined by (2), is a $\sqrt{n}$-a.n. estimator of θ_0.

*Exact knowledge is not necessary; it suffices to estimate $\mathbb{E}[\eta^r|X]$ at a $o\left(n^{-\frac{1}{3}}\right)$ rate.

Application to High-Dimensional Linear Regression

$f_0(X) = \langle X, \beta\rangle$, $g_0(X) = \langle X, \gamma\rangle$ for some s-sparse $\beta, \gamma \in \mathbb{R}^p$.

Theorem 1 works when $s = o\left(\frac{\sqrt{n}}{\log p}\right)$ [1], while Theorem 4 works for $s = o\left(\frac{n^{\frac{2}{3}}}{\log p}\right)$!

Theorem 5 *Suppose*
- *ϵ, η are independent of X, T and almost surely bounded by a constant $C > 0$,*
- *$\mathbb{E}[\eta^3] \neq 0$ or $\mathbb{E}[\eta^4] \neq 3\mathbb{E}[\eta^2]$,*
  Suffices for non-Gaussianity!
- *$\theta_0 \in [-M, M]$ for some $M > 0$, and*
- *X has iid standard Gaussian entries.*

Then if

$$s = o\left(n^{\frac{2}{3}}/\log p\right),$$

and in the first stage estimation

(a) We create estimators $\hat{q}, \hat{\gamma}$ of $q := \theta_0\gamma + \beta, \gamma$ via LASSO by regressing Y, X and T, X respectively.

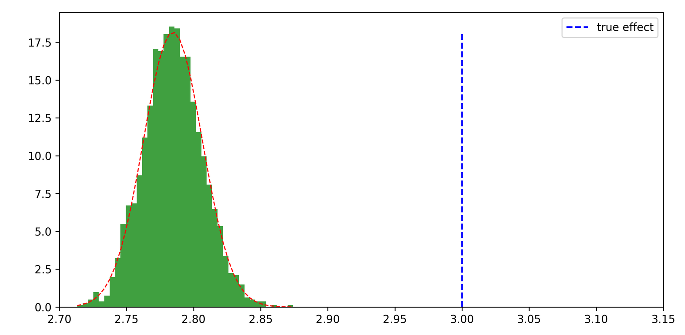
(b) Based on a split sample and our estimator $\hat{\gamma}$ of γ, we use $\mu^{(2)}$ and $\mu^{(3)}$ to estimate $\mathbb{E}[\eta^2]$ and $\mathbb{E}[\eta^3]$, where $\mu^{(2)} = \frac{1}{n}\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)^2$ and

$$\mu^{(3)} = \frac{1}{n}\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)^3 - 3\frac{1}{n}\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)\mu^{(2)}$$

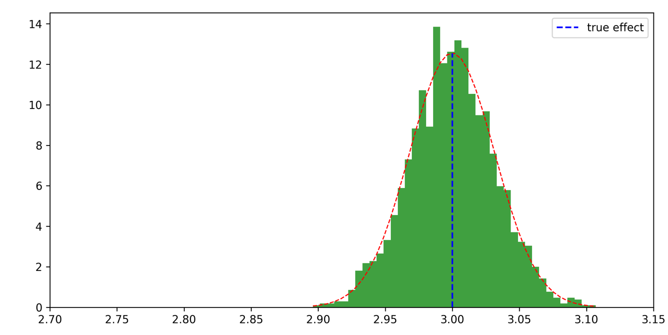
for $(T'_t, X'_t)_{t=1}^n$ an i.i.d. sample independent of $\hat{\gamma}$.

Then, using the moments of Theorem 4, the $\hat{\theta}$ defined by (2) is a $\sqrt{n}$-a.n. estimator of θ_0.

Experiments: First order orthogonal vs Second order orthogonal

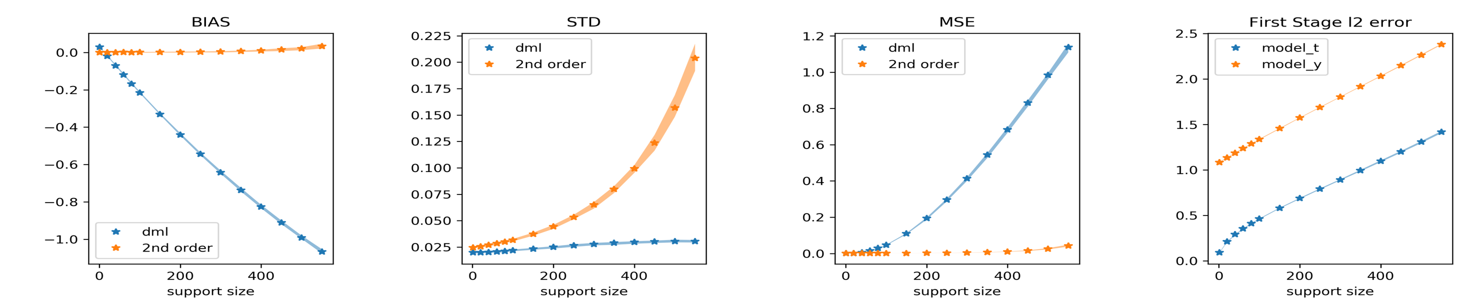


(a)



(b)

100 Monte Carlo experiments, $\theta_0 = 3$ and $p = 1000, n = 5000, s = 100$.



100 Monte Carlo experiments, where $\theta_0 = 3$ and $p = 1000, n = 5000$ with varying sparsity.

Bibliography

1. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/Debiased/Neyman Machine Learning of Treatment Effects. American Economic Review, 2017.