A Proof of Orthogonal Double Machine Learning with 
Z-Estimators

Vasilis Syrgkanis
Microsoft Research
April 18, 2017

Abstract

We consider two stage estimation with a non-parametric first stage and a generalized method of moments second stage, in a simpler setting than [CCD+16]. We give an alternative proof of the theorem given in Chernozhukov et al. [CCD+16] that orthogonal second stage moments, sample splitting and $n^{1/4}$-consistency of the first stage, imply $\sqrt{n}$-consistency and asymptotic normality of second stage estimates. Our proof is for a variant of their estimator, which is based on the empirical version of the moment condition (Z-estimator), rather than a minimization of a norm of the empirical vector of moments (M-estimator). This note is meant primarily for expository purposes, rather than as a new technical contribution.

1 Two-Stage Estimation

Suppose we have a model which predicts the following set of moment conditions:

$$E[m(Z, \theta_0, h_0(X))] = 0$$

where $\theta_0 \in R^d$ is a finite dimensional parameter of interest, $h_0: S \to R^d$ is a nuisance function we do not know, $Z$ are the observed data which are drawn from some distribution and $X \in S$ is a subvector of the observed data.

We want to understand the asymptotic properties of the following two-stage estimation process:

1. First stage. Estimate $h_0(\cdot)$ from an auxiliary data set (e.g. running some non-parametric regression) yielding an estimate $\hat{h}$.

2. Second stage. Use the first stage estimate $\hat{h}$ and compute an estimate $\hat{\theta}$ of $\theta_0$ from an empirical version of the moment condition: i.e.

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

The question we want to ask is: is $\hat{\theta}$ $\sqrt{n}$-consistent. More formally, is it true that:

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma)$$

for some constant co-variance matrix $\Sigma$. We will assume that the moment conditions that we use satisfy the following orthogonality property:

**Definition 1** (Orthogonality). For any fixed estimate $\hat{h}$ that can be the outcome of the first stage estimation, the moment conditions are orthogonal if:

$$E \left[ \nabla_{\gamma} m(Z, \theta_0, h_0(X)) \cdot (\hat{h}(X) - h_0(X)) \right] = 0$$

where $\nabla_{\gamma} m(\cdot, \cdot, \cdot)$ denotes the gradient of $m$ with respect to its third argument.
2 Orthogonality Implies Root-$n$ Consistency

Assumption 1. We will make the following regularity assumptions:

- **Rate of First Stage.** The first stage estimation is $n^{-1/4}$-consistent in the squared mean-square-error sense, i.e.
  \[ n^{1/2} E_X \left[ \| \hat{h}(X) - h_0(X) \|^2 \right] \to_p 0 \]  
  (5)
  where the convergence in probability statement is with respect to the auxiliary data set.

- **Regularity of First Stage.** The first stage estimate and the nuisance function are uniformly bounded by a constant, i.e.: $\| \hat{h}(x) \|, \| h_0(x) \| \leq C$ for all $x \in S$.

- **Regularity of Moments.** The following smoothness conditions hold for the moments
  1. For any $z, x, \gamma$ the function $m(z, \theta, \gamma)$ is continuous in $\theta$. Also $m(z, \theta, \gamma) \leq d(z)$ and $E[d(Z)] < \infty$.
  2. Similarly, the same conditions hold for $\nabla \theta m(z, \theta, \gamma)$.
  3. $E[\nabla \theta m(z, \theta_0, h_0(x))]$ is non-singular.
  4. the Hessian $\nabla \theta \gamma m(z, \theta, \gamma)$ has the largest eigenvalue bounded by some constant $\lambda$ uniformly for all $\theta$ and $\gamma$.
  5. the derivative $\nabla \theta \gamma m(z, \theta, \gamma)$ has norm, uniformly bounded by $\sigma$

Theorem 2. Under Assumption 1 and assuming that $\hat{\theta}$ is consistent, if the moment conditions satisfy the orthogonality property then $\hat{\theta}$ is also $\sqrt{n}$-consistent and asymptotically normal.

Proof. By doing a first-order Taylor expansion of the empirical moment condition around $\theta_0$ and by the mean value theorem, we have:

\[
\sqrt{n}(\hat{\theta} - \theta_0) = \left[ \frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X_t)) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} m(Z_t, \theta_0, \hat{h}(X_t)) \]  

where $\hat{\theta}$ is convex combination of $\theta_0$ and $\hat{\theta}$. We will show that $A$ converges in probability to a constant $J^{-1}$ and that $B$ converges in distribution to a normal $N(0, V)$, for some constant co-variance matrix $V$. Then the theorem follows by invoking Slutzky’s theorem, which shows convergence in distribution to $N(0, J^{-1}V)$.

**Convergence of $A$ to inverse derivative.** By the regularity of the moments, we have a uniform law of large numbers for the quantity $\frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X_t))$, i.e.:

\[
\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X)) - E[\nabla \theta m(Z, \hat{\theta}(X))] \right\| \to_p 0 \]  

(7)

Since $\hat{\theta}$ is consistent, we also have that $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to_p \theta$. Combining the latter two properties, we get that conditional on the auxiliary data set:

\[
\frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X)) \to E \left[ \nabla \theta m(Z, \theta_0, \hat{h}(X)) \right] \]  

(8)

Moreover, since $\hat{h}$ is consistent we get that:

\[
\frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X)) \to E \left[ \nabla \theta m(Z, \theta_0, h_0(X)) \right] \]  

(9)

Since the matrix $E[\nabla \theta m(z, \theta_0, h_0(x))]$ is non-singular, by continuity of the inverse we get:

\[
\left[ \frac{1}{n} \sum_{t=1}^{n} \nabla \theta m(Z_t, \hat{\theta}(X)) \right]^{-1} \to [E[\nabla \theta m(Z, \theta_0, h_0(X))]^{-1} = J^{-1} \]  

(10)

2
Asymptotic normality of $B$. To argue asymptotic normality of $B$ we take a second-order Taylor expansion of $B$ around $h_0(X_t)$ for each $X_t$:

$$B = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} m(Z_t, \theta_j, h_0(X_t)) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \nabla_{\gamma} m(Z_t, \theta_j, h_0(X_t)) \cdot \left( \hat{h}(X_t) - h_0(X_t) \right)$$

First we observe that $C$ is the sum of $n$ i.i.d. random variables, divided by $\sqrt{n}$. Thus by the Central Limit Theorem, we get that $C \to N(0, V)$, for some constant co-variance matrix $V$. Then we conclude by showing that $D, E \to_p 0$.

Second we argue that $n^{1/4}$ consistency of the first stage, implies that $E \to_p 0$. Since $\nabla_{\gamma} m(z, \theta, \gamma)$ has a largest eigenvalue uniformly bounded by $\lambda^*$, we have that the quantity $E$ is bounded by

$$|E| \leq \frac{\lambda^*}{2} \sqrt{n} \left( \frac{1}{n} \sum_{t=1}^{n} \| \hat{h}(X_t) - h_0(X_t) \|^2 \right)$$

Fixing the auxiliary data set, the quantity $\frac{1}{n} \sum_{t=1}^{n} \| \hat{h}(X_t) - h_0(X_t) \|^2$ converges to $E[\| \hat{h}(X_t) - h_0(X_t) \|^2]$. Subsequently by $n^{1/4}$-consistency of the first stage, and regularity of the first stage, we get that $E \to_p 0$.

Finally, we argue that orthogonality implies that $D \to_p 0$. We show that both the mean and the trace of the co-variance of $D$ converge to 0. The mean conditional on the auxiliary data set is:

$$E[D \mid \hat{h}] = \sqrt{n} E \left[ \nabla_{\gamma} m(Z, \theta_j, h_0(X)) \cdot \left( \hat{h}(X) - h_0(X) \right) \mid \hat{h} \right] = 0$$

The diagonal entries of the co-variance conditional on the auxiliary dataset is:

$$E[D^2 \mid \hat{h}] = \frac{1}{n} \sum_{t \neq t'} E \left[ \nabla_{\gamma} m(Z, \theta_j, h_0(X_t)) \cdot \left( \hat{h}(X) - h_0(X) \right) \mid \hat{h} \right]^2$$

$$+ \frac{1}{n} \sum_{t \neq t'} E \left[ \| \nabla_{\gamma} m(Z, \theta_j, h_0(X)) \cdot \left( \hat{h}(X) - h_0(X) \right) \|^2 \mid \hat{h} \right]$$

All the cross terms are zero by orthogonality, giving:

$$E[D^2 \mid \hat{h}] = E \left[ \| \nabla_{\gamma} m(Z, \theta_j, h_0(X)) \|^2 \cdot \left( \hat{h}(X) - h_0(X) \right) \|^2 \right] \leq \sigma^2 E \left[ \| \hat{h}(X) - h_0(X) \|^2 \right]$$

Since $\hat{h}$ is consistent, we get that the latter converges to zero. Since the mean of $D$ and the trace of its co-variance converge to zero, we get that $D \to_p 0$.

Consistency of the estimator also follows easily from standard arguments, if one makes Assumption[4] and the extra condition that the moment condition in the limit is satisfied only for the true parameters, which is needed for identification (see e.g. [NM94] for the formal set of extra regularity assumptions needed for consistency).

3 Orthogonal Moments for Conditional Moment Problems

One special case of when the orthogonality condition is satisfied is the following stronger, but easier to check property of conditional orthogonality:

**Definition 2** (Conditional Orthogonality). The moment conditions are conditionally orthogonal if:

$$E[\nabla_{\gamma} m(Z, \theta_j, h_0(X))|X] = 0$$

(15)
Lemma 3. Conditional orthogonality implies orthogonality, when an auxiliary data set is used to estimate \( \hat{h} \).

Proof. By the law of iterated expectations we have:

\[
\mathbb{E} \left[ \nabla \gamma m(Z, \theta_0, h_0(X)) \cdot (\hat{h}(X) - h_0(X)) \right] = \mathbb{E} \left[ \mathbb{E} \left[ \nabla \gamma m(Z, \theta_0, h_0(X)) \cdot (\hat{h}(X) - h_0(X)) \mid \hat{h}, X \right] \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \nabla \gamma m(Z, \theta_0, h_0(X)) \mid \hat{h}, X \right] \cdot (\hat{h}(X) - \hat{h}(X)) \right] = 0
\]

Where in the last part we used the conditional orthogonality property.

For conditional moment problems studied in [Cha92], [CCD+16] shows how one can transform in an algorithmic manner an initial set of moments to a vector of orthogonal moments.

References

[CCD+16] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and a. W. Newey. Double Machine Learning for Treatment and Causal Parameters. ArXiv e-prints, July 2016.

[Cha92] Gary Chamberlain. Efficiency bounds for semiparametric regression. Econometrica, 60(3):567–596, 1992.

[NM94] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. Handbook of Econometrics, 4:2111 – 2245, 1994.