Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Introduction to Double Machine Learning and Uniform in High-Dimensional Additive Models

Martin Spindler
joint work with Philipp Bach, Sven Klaassen and Jannis Kück

University of Hamburg & Economic AI

When Causal Inference meets Statistical Analysis
Paris, April 18, 2023

Agenda

Introduction to Double Machine Learning

- Massive / Big Data become more and more available.
- Machine learning methods focus mostly on prediction.
- But in many situations the interest is on learning (causal) relationships and making inference.
- Bringing in statistical modelling $\rightarrow$ strength of statistics and econometrics
- Combining machine learning and (causal) inference
- Here: Estimation and inference of high-dimensional additive models

Introduction

Main goal: Provide general framework for estimating and doing inference about a low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameter (η_0) which may be estimated with the new generation of nonparametric statistical methods, "machine learning" (ML) methods, such as

- random forests,
- boosted trees,
- lasso,
- ridge,
- deep and standard neural nets,
- gradient boosting,
- their aggregations,
- and cross-hybrids.

We consider the linear regression model in a high-dimensional setting (potentially $p \geq n$)

$$Y = D\theta_0 + X_1\beta_1 + \dots X_p\beta_p + \varepsilon, \quad \mathrm{E}[\varepsilon \mid X, D] = 0,$$

- Y - outcome variable
- D - policy/treatment variable
- θ_0 - parameter of interest
- $\beta = (\beta_1, \dots, \beta_p)^t$ - nuisance parameter
- $X = (X_1, \dots, X_p)^t$ is a vector of other covariates, called “controls”or “confoundersin the sense that

$$D = \gamma^t X + \nu, \quad E[\nu|X] = 0.$$

Example: Cross-Country Growth Regression

- Relation between growth rate and initial per capita GDP, conditional on covariates, describing institutions and technological factors:

$$\underbrace{GrowthRate_i}_{Y_i} = \beta_0 + \theta_0 \underbrace{\log(GDP_i)}_{D_i} + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$$

where the model is exogenous,

$$\mathrm{E}[\varepsilon_i | D_i, X_i] = 0$$

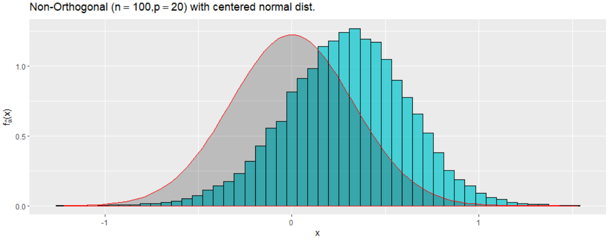
- Test the convergence hypothesis ($\theta_0 < 0$) that poorer countries catch up with richer countries, conditional on similar institutions and other factors. Prediction from the classical Solow growth model.
- In Barro-Lee data, we have $p = 60$ covariates, $n = 90$ observations. Need to do selection.

"Naive" or Prediction-Based ML Approach is Bad

Naive/Textbook Inference:

1. Select controls terms by running Lasso (or variants) of Y_i on X_i
2. Estimate θ_0 by least squares of Y_i on D_i and selected controls, apply standard inference

The distribution of $\hat{\theta}_0 - \theta_0$ looks like this:

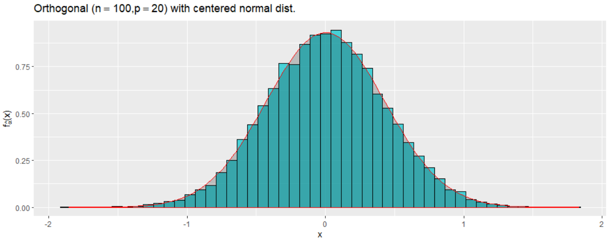


The “Double” ML Approach

1. Predict Y and D using X by $\mathrm{E}[Y|X]$ and $\mathrm{E}[D|X]$, obtained using Lasso, Random Forest or other â€œbest performingâ€ ML tools.
2. Residualize $W = Y - \mathrm{E}[Y|X]$ and $V = D - \mathrm{E}[D|X]$
3. Regress W on V to get θ_0

Frisch-Waugh-Lovell (1930s) style with ML methods

The distribution of $\hat{\theta}_0 - \theta_0$ looks like this:



Example

Method	effect	s.e.
Barro-Lee (Economic Reasoning)	-0.02	0.005
All Controls ($n = 90,\ p = 60$)	-0.02	0.031
Post-Naive Selection	-0.01	0.004
Post-Double Selection	-0.03	0.011

- Double-Selection finds 8 controls, including trade-openness and several education variables.
- Our findings support the conclusions reached in Barro and Lee and Barro and Sala-i-Martin.
- Using all controls is very imprecise.
- Using naive selection gives a biased estimate for the speed of convergence.

Moment Conditions

The two strategies rely on very different moment conditions for identifying and estimating θ_0:

$$\mathrm{E}[\varepsilon D] = \mathrm{E}[(Y - D\theta_0 - g_0(X))D] = 0 \quad (1)$$

$$\mathrm{E}[(W - V\theta_0)V] = 0, \quad (2)$$

with $W \equiv Y - \mathrm{E}[Y|X]$ and $V \equiv D - \mathrm{E}[D|X]$.

- (1) - Regression adjustment
- (2) - Neyman-orthogonal

Both approaches generate estimators of θ_0 that solve the empirical analog of the moment conditions above; unknown nuisance functions

$$g_0(X), \quad m_0(X) := \mathrm{E}[D|X], \quad \ell_0(X) = \mathrm{E}[Y|X]$$

are replaced with their ML-based estimators.

“Naive” or “Prediction-focused” ML Estimation

Suppose we use (1) with an estimator $\hat{g}_0(X)$ to estimate θ_0:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \hat{g}_0(X_i))$$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\varepsilon_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\left(g_0(X_i) - \hat{g}_0(X_i)\right)}_{:=b}$$

- $a \rightsquigarrow N(0, \bar{\Sigma})$ under standard conditions
- What about b?

Estimation Error in Nuisance Function

We will generally have $b \to \infty$:

$$b \approx (\mathrm{E}D^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i) \left(g_0(X_i) - \hat{g}_0(X_i)\right)$$

- $(g_0(X_i) - \hat{g}_0(X_i))$ error in estimating g_0

Heuristics:

- In nonparametric setting, the error is of order $n^{-\varphi}$ for $0 < \varphi < 1/2$.
- b will then look like $\sqrt{n}n^{-\varphi} \to \infty$

The "naive" or prediction-focused ML estimator $\hat{\theta}_0$ is not root-n consistent.

Orthogonalized or "Double ML" Formulation

Consider estimation based on (2)

$$\check{\theta}_0 = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{V}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{N} \hat{V}_i \hat{W}_i$$

- $\hat{V} = D - \hat{m}_0(X)$, $\hat{W} = Y - \hat{\ell}_0(X)$

Under mild conditions, we can write

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} V_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \varepsilon_i}_{:=a^*}$$

$$+ \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} V_i^2 \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_0(X_i) - \hat{m}_0(X_i)) \left(\ell_0(X_i) - \hat{\ell}_0(X_i) \right)}_{:=b^*}$$

$$+ o_p(1).$$

Heuristic Convergence Properties

- $a^* \rightsquigarrow N(0, \Sigma)$ under standard conditions
- b^* now depends on product of estimation errors in both nuisance functions
- b^* will look like $\sqrt{n} n^{-(\varphi_m + \varphi_\ell)}$ where $n^{-\varphi_m}$ and $n^{-\varphi_\ell}$ are respectively appropriate convergence rates of estimators for $m(x)$ and $\ell(x)$
- $o(n^{-1/4})$ is often an attainable rate for estimating $m(x)$ and $\ell(x)$

The Double ML estimator $\check{\theta}_0$ is a $\sqrt{n}$ consistent and approximately centered normal quite generally.

Neyman Orthogonality as Key Difference

- Key difference between estimation based on (1) and estimation based on (2) is that (2) satisfies the **Neyman orthogonality condition:**

  Let

  $$\eta_0 = (\ell_0, m_0) = (\mathrm{E}[Y|X], \mathrm{E}[D|X]) \quad , \quad \eta = (\ell, m).$$

  The partial derivative of the moment condition (2) with respect to η vanishes:

  $$\partial_\eta \mathrm{E}[\psi(W, \theta_0, \eta)]\Big|_{\eta=\eta_0} = 0,$$

  where W denotes the data (Y, D, X).
- Heuristically, the moment condition remains "valid" under "local" mistakes in the nuisance function.
- This property generally does not hold for the moment condition (1).

Literature and Generalization

- Literature
  - Linear model: Belloni, Chernozhukov, Hansen (2015), Zhang and Zhang (2015), Bühlmann et al. (2015)
  - Instrumental variable estimation: Belloni, Chen, Chernozhukov, Hansen (2012), Chernozhukov, Hansen, Spindler (2015)
  - Various treatment effecs: Belloni, Chernozhukov, Fernandez-Val, Hansen (2017)
- Software implementation: R package hdm (Chernozhukov, Hansen, Spindler, 2016), R / Python package doubleML (Bach, Chernozhukov, Kurz, Spindler, 2021a, 2021b)

Literature and Generalization

- Inference about low-dimensional parameters in high-dimensional (linear) models:
  - Belloni, Chernozhukov, Hansen, and coauthors (in a series of papers)
- Inference about high-dimensional parameters by allowing the number of moment condition to grow with sample size:
  - Belloni et al. (2018) "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework"
  - Chernozhukov et al. (2017) "Central Limit Theorems and Bootstrap in High Dimensions"

High-dimensional Additive Models

- Additive models are quite popular in statistics, imposing an additive structure to evade curse of dimensionality

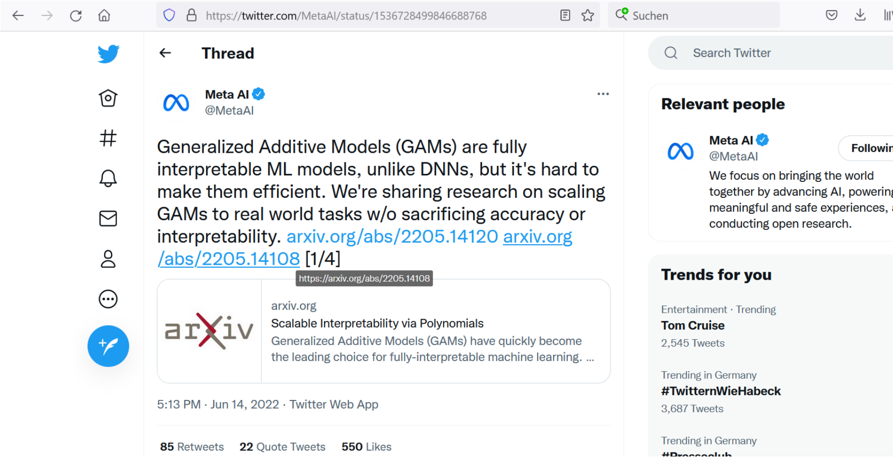
$$Y = \beta + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] = 0$$

  where β denotes a constant and $f_j(\cdot)$ univariate functions.
- We rewrite the model as

$$Y = f_1(X_1) + f_{-1}(X_{-1}) + \varepsilon,$$

  where $f_1(X_1)$ with $\mathbb{E}[f_1(X_1)] = 0$ denotes the target component and $f_{-1}(X_{-1})$ is a nuisance function.

Motivation


https://twitter.com/MetaAI/status/1536728499846688768
Suchen
Thread
Search Twitter
Meta AI
@MetaAI
Generalized Additive Models (GAMs) are fully interpretable ML models, unlike DNNs, but it's hard to make them efficient. We're sharing research on scaling GAMs to real world tasks w/o sacrificing accuracy or interpretability. arxiv.org/abs/2205.14120 arxiv.org/abs/2205.14108 [1/4]
https://arxiv.org/abs/2205.14108
arXiv
arxiv.org
Scalable Interpretability via Polynomials
Generalized Additive Models (GAMs) have quickly become the leading choice for fully-interpretable machine learning. ...
5:13 PM · Jun 14, 2022 · Twitter Web App
85 Retweets 22 Quote Tweets 550 Likes
Relevant people
Meta AI
@MetaAI
We focus on bringing the world together by advancing AI, powering meaningful and safe experiences, conducting open research.
Trends for you
Entertainment · Trending
Tom Cruise
2,545 Tweets
Trending in Germany
#TwitternWieHabeck
3,687 Tweets
Trending in Germany

Literature Review

- Confidence bands under fixed dimension: Härdle (1989), Sun and Loader (1994), Fan and Zhang (2000) and many others
- Estimation rates in high dimensions: Lin and Zhang (2006), Meier et al. (2009), Huang et al. (2010), Kato (2012), Lou et al. (2014) and many others
- Confidence bands in high dimensions: Kozbur (2020), Lu, Kolar and Liu (2020), Gregory, Mammen and Wahl (2021)

Motivation

- **Goal:** Providing uniformly valid confidence bands for the target function $f_1(\cdot)$ in a high-dimensional setting.
- Main idea:
  - Approximation of each component with sieves:

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1)$$
$$f_{-1}(X_{-1}) = \beta_0^T h(X_{-1}) + b_2(X_{-1}),$$

    for a suitable set of approximating functions $g(x) = (g_1(x), \dots, g_{d_1}(x))^T$ and $h(x) = (h_1(x), \dots, h_{d_2}(x))^T$ (e.g. b-splines,...).
- The number of approximating functions d_1 may grow with sample size.

Double Machine Learning Framework (1/2)

- Given the approximations, we consider the very high-dimensional regression model

$$Y = \theta_0^T g(X_1) + \beta_0^T h(X_{-1}) + b_1(X_1) + b_2(X_{-1}) + \varepsilon.$$

- Further, assume that

$$g_l(X_1) = (\gamma_0^{(l)})^T Z_{-l} + \nu^{(l)}, \quad \mathbb{E}[\nu^{(l)} Z_{-l}] = 0$$

with $Z := (g_1(X_1), \dots, g_{d_1}(X_1), h_1(X_{-1}), \dots, h_{d_2}(X_{-1}))^T$.

- This partially linear model is well known and estimating $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,d_1})$ can be recast into a general Z-estimation problem:

$$\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0 \quad l = 1, \dots, d_1.$$

Double Machine Learning Framework (2/2)

- The score is given by

$$\psi_l(W,\theta,\eta) = \Big(Y - \theta g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\Big) \cdot \Big(g_l(X_1) - (\eta^{(2)})^T Z_{-l}\Big).$$

with $\psi_l(W,\theta_{0,l},\eta_{0,l}) = \varepsilon \cdot \nu^{(l)}$.

- The nuisance parameters are

$$\eta_{0,l}^{(1)} := \beta_0^{(l)} = (\theta_{0,1},\dots,\theta_{0,l-1},\theta_{0,l+1},\dots,\theta_{0,d_1},\beta_{0,1},\dots,\beta_{0,d_2})^T$$
$$\eta_{0,l}^{(2)} := \gamma_0^{(l)}, \quad \eta_{0,l}^{(3)}(X) := b_1(X_1) + b_2(X_{-1}).$$

- The score fulfills the near *Neyman orthogonality* condition

$$\partial_\eta \mathbb{E}[\psi(W,\theta_{0,l},\eta)]|_{\eta=\eta_{0,l}} = o(n^{-1/2}).$$

Challenges (1/2)

- **Main Challenge** DML provide valid inference on θ_0, but we are interested in
$$f_1(\cdot) \approx \theta_0^T g(\cdot).$$
- Non-trivial extension of the DML Framework is needed.
- Using b-splines or other local estimators, we have a problem of vanishing eigenvalues:

$$cd_1^{-1} \leq \inf_{\|\xi\|_2=1} \mathbb{E}\left[\left(\xi^T g(X_1)\right)^2\right] \leq \sup_{\|\xi\|_2=1} \mathbb{E}\left[\left(\xi^T g(X_1)\right)^2\right] = Cd_1^{-1}$$

since $E[g_l(X_l)^2] = O\left(\frac{t_1}{(d_1 - t_1 + 2)}\right)$ where t_1 denotes the number of non-zero elements of g.

Challenges (2/2)

- The lasso estimators need to fulfill

$$\left\|\hat{\beta}_0^{(l)} - \beta_0^{(l)}\right\|_2 = o(n^{-1/4}) \quad \text{and} \quad \left\|\hat{\gamma}_0^{(l)} - \gamma_0^{(l)}\right\|_2 = o(n^{-1/4})$$

- On the other hand, we rely on the two approximations

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1)$$
$$f_{-1}(X_{-1}) = \beta_0^T h(X_{-1}) + b_2(X_{-1}),$$

with $\theta_0 \in \mathbb{R}^{d_1}$ and $\beta_0 \in \mathbb{R}^{d_2}$.

- We need to ensure that

$$\sup_x \left(b_1(x_1) + b_2(x_{-l})\right) = o(n^{-1/4}).$$

- There is a trade-off regarding the number of approximating function d_1 and d_2.

Main Theorem

Modifying the Double Machine Learning Framework enables us to provide uniformly valid confidence bands for the target function

$$f_1(x) \approx \theta_0^T g(x).$$

Define

$$\hat{u}(x) := \hat{\theta}^T g(x) + \frac{(g(x)^T \hat{\Sigma}_n g(x))^{1/2} c_\alpha}{\sqrt{n}}$$

$$\hat{l}(x) := \hat{\theta}^T g(x) - \frac{(g(x)^T \hat{\Sigma}_n g(x))^{1/2} c_\alpha}{\sqrt{n}}.$$

Theorem

Under regularity assumptions, it holds

$$P\left(\hat{l}(x) \leq f_1(x) \leq \hat{u}(x), \forall x \in \mathcal{I}\right) \to 1 - \alpha.$$

Simulation Study (1/3)

- Data generating process based on Gregory et al. (2021)

$$Y_i = \sum_{j=1}^{p} f_j(X_{j,i}) + \varepsilon_i, \quad i = 1, \ldots, n$$

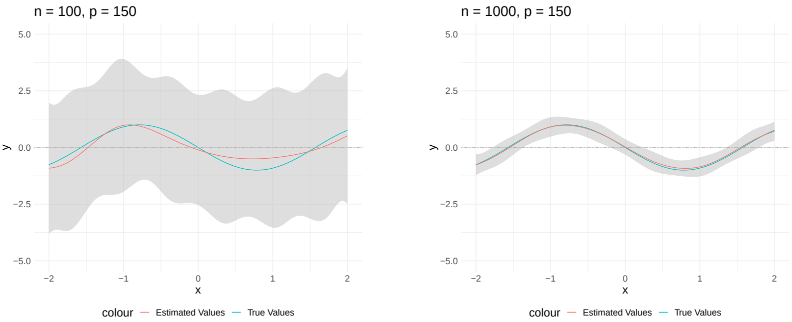
with 4 non-zero components and $\varepsilon \sim N(0,1)$.

- For each component, we use cubic B-Splines with nine degrees of freedom for approximation.
- $X_j \sim \mathcal{U}[-2.5, 2.5]$ and $Cov(X_k, X_l) = 0,5^{|k-l|}$.

DGP 1 (sine)	$f_1(x) = -\sin(2 \cdot x)$
DGP 2 (quad)	$f_2(x) = x^2 - 25/12$
DGP 3 (line)	$f_3(x) = x$
DGP 4 (expo)	$f_4(x) = \exp(-x) - \frac{2}{5}\sinh(\frac{5}{2})$

Table: Data generating processes in simulation study.

Simulation Study (2/3)



(a) $n = 100$ (b) $n = 1000$

Figure: Exemplary simulation results for $f_1(x) = -\sin(2 \cdot x)$.

Simulation Study (3/3)

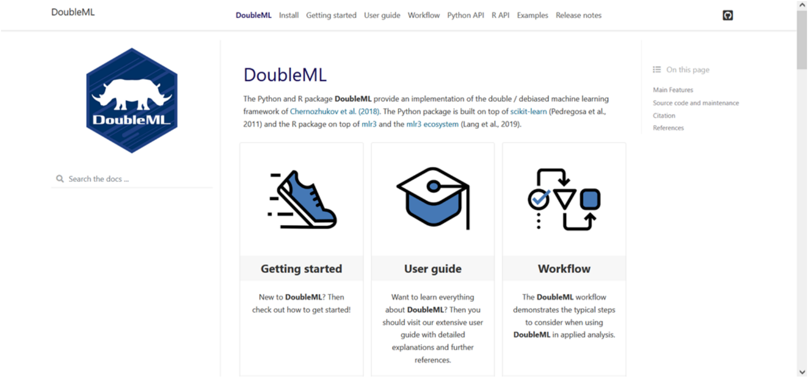
n	p	sine	quad	line	expo
100	150	0.954	0.926	0.942	0.962
100	50	0.912	0.93	0.942	0.948
1000	150	0.932	0.946	0.926	0.938
1000	50	0.936	0.948	0.92	0.954

Table: Simulation results: Coverage achieved by simultaneous confidence bands for $\alpha = 0.05$ over the interval $[-1.5, 1.5]$ in $R = 500$ repetitions.

Summary of the Paper

- Methodology for uniformly valid confidence bands for a nonparametric function $f_1(X_1)$ in a high-dimensional additive model.
- Non-trivial extension of the DML Framework (details).
- Analysis of regression models in high-dimensions without imposing the strong assumptions of linearity.
- We provide simulation studies (details) and an empirical illustration of the estimation procedure (details).

More on Double Machine Learning



More on Double Machine Learning

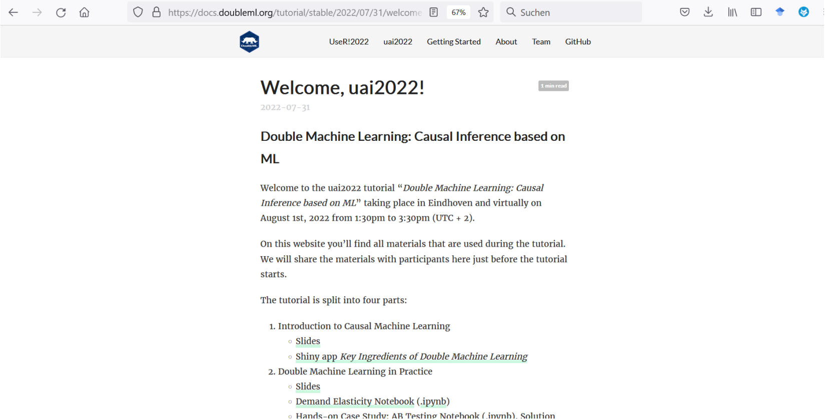


Figure: https://www.auai.org/uai2022/tutorials

More on Double Machine Learning

Applied Causal Inference Powered by ML and AI

Victor Chernozhukov[*], Christian Hansen[†], Nathan Kallus[‡], Martin Spindler[§], Vasilis Syrgkanis[¶]

August 9, 2022

Publisher: Online

Bibliography

- Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y. (2018). "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework." *Annals of statistics* 46 (6B): 3643.
- Belloni, A., Chernozhukov V. and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81 (2): 608-650.
- Bühlmann, P. and Meinshausen, N. (2006). "High-dimensional graphs and variable selection with the lasso." *The annals of statistics* 34 (3): 1436-1462.
- Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). "Central limit theorems and bootstrap in high dimensions." *The Annals of Probability* 45 (4): 2309-2352.
- Fan, J. and Zhang, W. (2000). "Simultaneous confidence bands and hypothesis testing in varying-coefficient models." *Scandinavian Journal of Statistics* 27 (4): 715-731.
- Gregory, K., Mammen, E. and Wahl, M. (2021). "Statistical inference in sparse high-dimensional additive models." *The Annals of Statistics* 49 (3), 1514-1536.
- Härdle, W. (1989). "Asymptotic maximal deviation of M-smoothers." *Journal of Multivariate Analysis* 29 (2): 163-179.
- Kozbur, D. (2020). "Inference in additively separable models with a high-dimensional set of conditioning variables." *Journal of Business and Economic Statistics*: 1-17.
- Lu, J., Kolar, M. and Liu, H. (2020). "Kernel meets sieve: Post-regularization confidence bands for sparse additive model." *Journal of the American Statistical Association*: 1-16.
- Sun, J. and Loader, C. R. (1994). "Simultaneous confidence bands for linear regression and smoothing." *The Annals of Statistics* 22 (3): 1328-1345.
- Van de Geer, S., Bühlmann, P., Ritov, Y. A. and Dezeure, R. (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." *The Annals of Statistics* 42 (3): 1166-1202.
- Zhang, C. H. and Zhang, S. S. (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society* Series B: Statistical Methodology: 217-242.

Thank you for your attention!

Martin Spindler

University of Hamburg

martin.spindler@uni-hamburg.de

The Role of Neyman Orthogonality

- We have the taylor expansion

$$J\sqrt{n}(\hat{\theta} - \theta_0) = A_n + \sqrt{n}\mathrm{D}O(\|\hat{\eta} - \eta_0\|) + C\sqrt{n}O(\|\hat{\eta} - \eta_0\|^2) + o_p(1),$$

where A_n is approximately Gaussian under weak conditions.

- Under Neyman orthogonality,

$$D := \partial_\eta \mathbb{E}[\psi(W, \theta_0, \eta)]|_{\eta=\eta_0} = 0$$

and thus we only need

$$C\sqrt{n}O(\|\hat{\eta} - \eta_0\|^2) \to 0$$

for $\sqrt{n}$ consistency which requires $\|\hat{\eta} - \eta_0\| = o_P(n^{-1/4})$.

Back

Double Machine Learning Estimator

The score ψ is linear in θ, meaning

$$\psi_l(W, \theta, \eta) = \psi_l^a(X, \eta^{(2)})\theta + \psi_l^b(X, \eta)$$

with

$$\psi_l^a(X, \eta^{(2)}) = -g_l(X_1)(g_l(X_1) - (\eta^{(2)})^T Z_{-l})$$

and

$$\psi_l^b(X, \eta) = (Y - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X))(g_l(X_1) - (\eta^{(2)})^T Z_{-l}).$$

Thus, the estimator is given by

$$\hat{\theta}_l = -\mathbb{E}_n[\psi_l^a(X_i, \hat{\eta}^{(2)})]^{-1}\mathbb{E}_n[\psi_l^b(X_i, \hat{\eta})]$$

for all $l = 1, \ldots, d_1$.

Proof Snippet

We prove the following Bahadur representation

$$\sup_{x\in I}\left|\sqrt{n}(g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T\left(\hat{\theta}-\theta_0\right)\right| = \sup_{x\in I}\left|\mathbb{G}_n(\psi_x)\right| + o_P(1)$$

with

$$\psi_x(\cdot) := (g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T J_0^{-1}\psi(\cdot,\theta_0,\eta_0)$$

where $J_{0,l} = -\mathbb{E}[(\nu^{(l)})^2]$ and

$$\Sigma_n = \begin{pmatrix} \frac{\mathbb{E}[(\varepsilon\nu^{(1)})^2]}{\mathbb{E}[(\nu^{(1)})^2]^2} & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(2)})^2]} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\ \frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}[(\varepsilon\nu^{(2)})^2]}{\mathbb{E}[(\nu^{(2)})^2]^2} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \cdots & \frac{\mathbb{E}[(\varepsilon\nu^{(d_1)})^2]}{\mathbb{E}[(\nu^{(d_1)})^2]^2} \end{pmatrix}.$$

Bootstrap

The critical value c_α can be determined by the following multiplier bootstrap method introduced in Chernozhukov et al. (2017). Define

$$\hat{\psi}_x(\cdot) := (g(x)^T \hat{\Sigma}_n g(x))^{-1/2} g(x)^T \hat{J}_0^{-1} \psi(\cdot, \hat{\theta}_0, \hat{\eta}_0)$$

and let

$$\hat{\mathcal{G}} = \left(\hat{\mathcal{G}}_x\right)_{x \in I} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^{(i)} \hat{\psi}_x\left(W^{(i)}\right)\right)_{x \in I},$$

where $(\xi^{(i)})_{i=1}^n$ are independent standard normal random variables. The multiplier bootstrap critical value c_α is given by the $(1-\alpha)$-quantile of the conditional distribution of $\sup_{x \in I} |\hat{\mathcal{G}}_x|$ given $(W^{(i)})_{i=1}^n$.

Empirical Application: Boston Housing Prices (1/2)

- Method applied on the well-known Boston Housing data, with $n = 506$ observations and $p = 11$ (continuous) covariates.

$$
\begin{aligned}
MEDV_i = {} & f_1(LSTAT_i) + f_2(CRIM_i) + f_3(ZN_i) + f_4(INDUS_i) + f_5(RM_i) \\
& + f_6(AGE_i) + f_7(DIS_i) + f_8(TAX_i) + f_9(PTRATIO_i) \\
& + f_{10}(ETHN_i) + f_{11}(NOX_i) + \epsilon_i .
\end{aligned}
$$

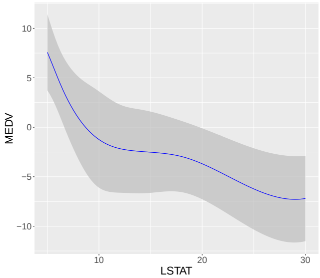


Figure: Estimated $f_1(x)$ with simultaneous 95%-confidence bands.

Empirical Application: Boston Housing Prices (2/2)

MEDV	median value of owner-occupied homes in USD 1000's
NOX	nitric oxides
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft
INDUS	proportion of non-retail business acres per town
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
TAX	full-value property-tax rate per USD 10,000
PTRATIO	pupil-teacher ratio by town
BLACK	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
LSTAT	percentage of lower status of the population

Table: List of variables: Boston Housing Data.

Back