Double Machine Learning for Static Panel Models with Fixed Effects

Paul Clarke[1] **Annalivia Polselli**[2]

[1]Institute for Social and Economic Research, University of Essex

[2]Institute of Analytics and Data Science, University of Essex

Internal seminar, Erasmus University Rotterdam

February 20, 2024

Motivation

- Consider Robinson (1988)'s partially linear regression (PLR) model for *cross-sectional* data

$$y_i = d_i\theta_0 + l_0(\mathbf{x}_i) + u_i, \quad \mathbb{E}(u_i|X_i, d_i) = 0, \tag{1}$$
$$d_i = m_0(\mathbf{x}_i) + v_i, \quad \mathbb{E}(v_i|\mathbf{x}_i) = 0 \tag{2}$$

- θ_0 is the parameter of interest to *estimate*
- Need to control for confounding factors $\mathbf{x}_i$, but l_0 and m_0 **unknown**
  - What variables? How many?
  - What functional form? Linear or nonlinear in $\mathbf{x}_i$?
- ML tools (e.g., Lasso, trees, random forests) to predict $l_0(\mathbf{x}_i)$ and $m_0(\mathbf{x}_i)$
- **But** plugging ML predictions in the model estimating equations introduce *regularisation* and *overfitting* bias (Chernozhukov et al., 2018)
  - No guarantee for $\sqrt{N}$-convergence of the target parameter

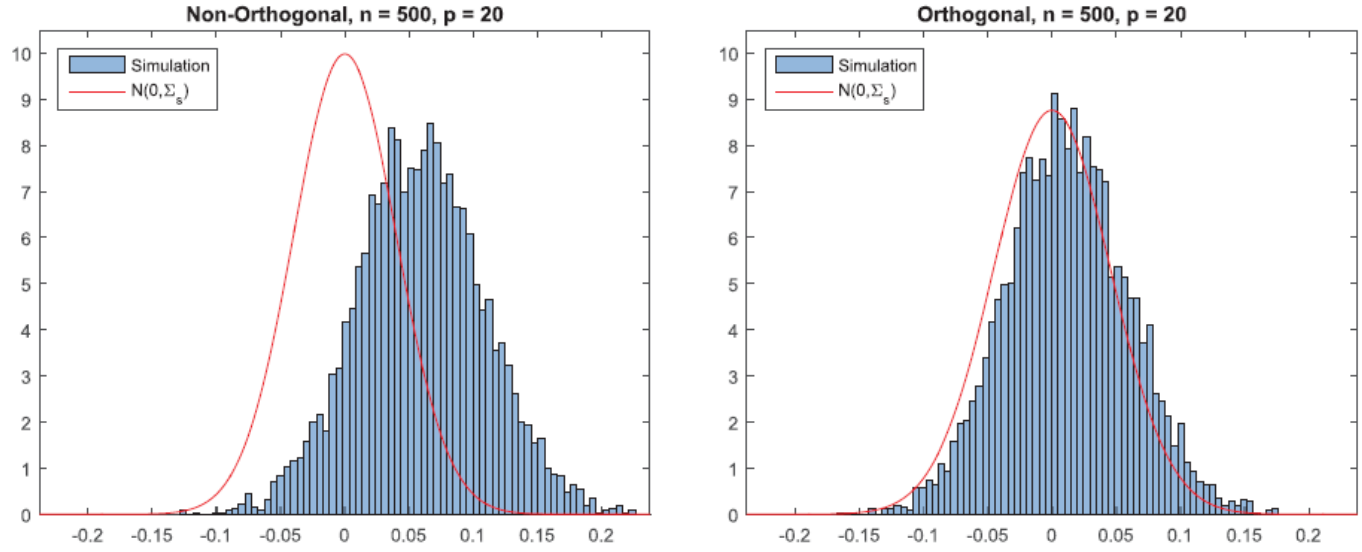
Motivation

- Consider Robinson (1988)'s partially linear regression (PLR) model for *cross-sectional* data

$$y_i = d_i\theta_0 + l_0(\mathbf{x}_i) + u_i, \quad \mathbb{E}(u_i|X_i, d_i) = 0, \tag{1}$$
$$d_i = m_0(\mathbf{x}_i) + v_i, \quad \mathbb{E}(v_i|\mathbf{x}_i) = 0 \tag{2}$$

- θ_0 is the parameter of interest to *estimate*
- Need to control for confounding factors $\mathbf{x}_i$, but l_0 and m_0 **unknown**
  - What variables? How many?
  - What functional form? Linear or nonlinear in $\mathbf{x}_i$?
- ML tools (e.g., Lasso, trees, random forests) to predict $l_0(\mathbf{x}_i)$ and $m_0(\mathbf{x}_i)$
- **But** plugging ML predictions in the model estimating equations introduce *regularisation* and *overfitting* bias (Chernozhukov et al., 2018)
  - No guarantee for $\sqrt{N}$-convergence of the target parameter

Behaviour of Conventional vs Double ML Estimators



Source: Figure 1 in Chernozhukov et al. (2018, p.C5). The figure illustrates the negative impact of regularization bias (on the left) and the benefit of orthogonalization (on the right). On the left, the sampling distribution (in blue) of a conventional (non-orthogonal) ML estimator is not centered around the theoretical mean (red curve) and, hence, upward biased. On the right, the sampling distribution of a double (orthogonal) ML estimator is not biased.

- Double ML (DML) by Chernozhukov et al. (2018) (on the right) uses

  (a) Orthogonalised regressor of the treatment (that partials out the effect of X from D) to ↓ regularisation bias

  (b) Sample-splitting (with cross-fitting) to ↓ overfitting bias

- $\sqrt{N}$-convergence guaranteed if ML prediction estimators converge at $N^{1/4}$ rate

Related Literature

- Estimation and inference with double/debiased (orthogonal) ML (e.g., Belloni et al., 2014, 2016; Chernozhukov et al., 2018, 2022)
  - DML for panel data models (Klosin and Vilgalys, 2022; Semenova et al., 2023)
- Potential outcome framework (Neyman, 1923; Rubin, 1974)
- Panel data models with fixed effects (Mundlak, 1978; Wooldridge, 2010)
- Value of DML for policy evaluation (Baiardi and Naghi, 2021; Knaus, 2022; Strittmatter, 2023)

This Paper

- We extend PLR model to DML for panel data models by accounting for
  1. Unobserved individual heterogeneity (or fixed effects)
  2. Nonlinear functions of the covariates

  to learn the nuisance functions and consistently estimate the treatment parameter
- Extension to *panel data* models challenging because traditional panel data techniques cannot be directly used. We propose
  1. Mundlak-device for correlated random effects
  2. Approximation approach
  3. Exact (or hybrid) approach
- Main results
  - Under regularity conditions for the DGP and the convergence properties of the learner, the DML estimator for panel data is root-N consistent and converge in distribution to a normal distribution
  - We show how to learn the necessary nuisance functions in the presence of the fixed effects and nonlinear functions
  - *Any* ML learner (e.g., Lasso, CART, RF) can be used

Outline

1. Theoretical model
2. Estimation and inference with DML
3. Monte Carlo Simulations
4. Empirical application
5. Future research

Assumptions

Model assumptions:

1. (No feedback to predictors) $\mathbf{x}_{it} \perp\!\!\!\perp L_t(y_i, d_i) \mid L_t(\mathbf{x}_i), \alpha_i$,
   where $L_t(\mathbf{x}_i) \equiv \{x_{i1}, \ldots, x_{it-1}\}$
2. (Individual heterogeneity explains lag-dependence)
   $y_{it}, d_{it} \perp\!\!\!\perp L_t(y_i, \mathbf{x}_i, d_i) \mid \mathbf{x}_{it}, \alpha_i$
3. (Selection on observables and individual heterogeneity)
   $y_{it}(.) \perp\!\!\!\perp d_{it} \mid \mathbf{x}_{it}, \alpha_i$
4. (Homogeneous treatment effect) $\mathbb{E}\{y_{it}(d) - y_{it}(0) | \mathbf{x}_{it}, \alpha_i\} = d_{it}\theta_0$

Model

Under assumptions 1-4, the 'partialling-out' PLR model for panel data holds

$$y_{it} = d_{it}\theta_0 + l_0(\mathbf{x}_{it}) + \alpha_i + u_{it} \tag{3}$$

$$d_{it} = m_0(\mathbf{x}_{it}) + \gamma_i + v_{it} \tag{4}$$

- y_{it} outcome, d_{it} treatment, $\mathbf{x}_{it}$ confounders
- $l_0(\mathbf{x}_{it})$ and $m_0(\mathbf{x}_{it})$ nuisance functions *to learn* from the data
- θ_0 target parameter *to estimate*
- α_i and γ_i fixed effects
- u_{it} and v_{it} disturbances

- Fixed effects pose challenges in learning the nuisance functions with traditional panel data techniques (e.g., LSDV, WG transformation) due to

  (i) nonlinearity of the nuisance functions and

  (ii) high-dimensionality of the data

Model

Under assumptions 1-4, the 'partialling-out' PLR model for panel data holds

$$y_{it} = d_{it}\theta_0 + l_0(\mathbf{x}_{it}) + \alpha_i + u_{it} \tag{3}$$

$$d_{it} = m_0(\mathbf{x}_{it}) + \gamma_i + v_{it} \tag{4}$$

- y_{it} outcome, d_{it} treatment, $\mathbf{x}_{it}$ confounders
- $l_0(\mathbf{x}_{it})$ and $m_0(\mathbf{x}_{it})$ nuisance functions *to learn* from the data
- θ_0 target parameter *to estimate*
- α_i and γ_i fixed effects
- u_{it} and v_{it} disturbances

- Fixed effects pose challenges in learning the nuisance functions with traditional panel data techniques (e.g., LSDV, WG transformation) due to

  (i) nonlinearity of the nuisance functions and

  (ii) high-dimensionality of the data

Handling the unobserved individual heterogeneity

Ways to handle unobserved individual heterogeneity

1. Correlated Random Effects (CRE) or Mundlak (1978)'s device
2. Within-group (WG) and First-differences (FD) transformations
   - 2.1 Approximation approach
   - 2.2 Exact (or hybrid) approach

CRE or Mundlak's Device 1/2

Proposition 1. Under Assumptions 1-4 and Mundlak-like fixed effects

$$\alpha_i = \overline{\mathbf{x}}_i'\boldsymbol{\pi}_\alpha + a_i \text{ and } \gamma_i = \overline{\mathbf{x}}_i'\boldsymbol{\pi}_\gamma + c_i, \tag{5}$$

where $\overline{\mathbf{x}}_i = T^{-1}\sum_i \mathbf{x}_{it}$, $\mathbb{E}(a_i|\overline{\mathbf{x}}_i) = \mathbb{E}(c_i|\overline{\mathbf{x}}_i) = 0$, a_i and c_i are random effects in the feasible learning problem

$$y_{it} = d_{it}\theta_0 + \widetilde{l}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + a_i + u_{it} \tag{6}$$

$$d_{it} = \widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + c_i + v_{it}, \tag{7}$$

where $\widetilde{l}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) = l_1(\mathbf{x}_{it}) + \overline{\mathbf{x}}_i'\boldsymbol{\pi}_\alpha$, $\widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) = m_1(\mathbf{x}_{it}) + \overline{\mathbf{x}}_i'\boldsymbol{\pi}_\gamma$, and

$$\mathbb{E}\big(u_{it}|v_{it}, \mathbf{x}_{it}, \overline{\mathbf{x}}_i, a_i\big) = \mathbb{E}\big(v_{it}|\mathbf{x}_{it}, \overline{\mathbf{x}}_i, c_i\big) = 0.$$

CRE or Mundlak's Device 2/2

Lemma 1. Suppose the conditions in Proposition 1 hold. The nuisance functions can be learnt from sample data:

(i) $\widetilde{l}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + a_i$ from $\{y_{it}, \mathbf{x}_{it}, \overline{\mathbf{x}}_i : t = 1, \ldots, T\}_{i=1}^N$

(ii) $\widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + c_i$ from $\{d_{it}, \mathbf{x}_{it}, \overline{\mathbf{x}}_i, \overline{d}_i : t = 1, \ldots, T\}_{i=1}^N$ if $d_{i1}, \ldots, d_{iT} | \mathbf{x}_{it}, \overline{\mathbf{x}}_i$ is *multivariate normal* to obtain the orthogonal estimator

$$v_{it} = d_{it} - \widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) - c_i$$

with $\mathbb{E}(d_{it} | \mathbf{x}_{it}, \overline{\mathbf{x}}_i) = \widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i)$

WG and FD Transformations 1/2

Let $\{w_{it}\}_{i=1}^{N}$ be a random variable and Q a panel data transformation operator

- WG transformation: $Q(w_{it}) = w_{it} - T^{-1}\sum_{t=1}^{T} w_{it}$ for all $t = 1, \ldots, T$
- FD transformation: $Q(w_{it}) = w_{it} - w_{it-1}$ for $T \geq 2$

The transformed model equations

$$Q(y_{it}) = Q(d_{it})\theta_0 + Q\big(l_0(\mathbf{x}_{it})\big) + Q(u_{it}) \tag{8}$$

$$Q(d_{it}) = Q\big(m_0(\mathbf{x}_{it})\big) + Q(v_{it}), \tag{9}$$

where $Q(\alpha_i) = Q(\gamma_i) = 0$.

WG and FD Transformations 2/2

Lemma 2 (approximation). Model (8)-(9) can be approximated by

$$Q(y_{it}) \approx Q(v_{it})\theta_0 + l_1\big(Q(\mathbf{x}_{it})\big) + Q(u_{it}) \tag{10}$$
$$Q(d_{it}) \approx m_1\big(Q(\mathbf{x}_{it})\big) + Q(v_{it}), \tag{11}$$

where l_1 and m_1 are learnt from the transformed data. Note that $Q\big(l_0(\mathbf{x}_{it})\big) \neq l_1\big(Q(\mathbf{x}_{it})\big)$ and $Q\big(m_0(\mathbf{x}_{it})\big) \neq m_1\big(Q(\mathbf{x}_{it})\big)$ when l_0 and m_0 are not linear.

Lemma 3 (exact). Suppose the conditions in Proposition 1 hold, and Mundlak-like fixed effects (5) holds. The nuisance parameters can be learnt from the data,

(i) $\widetilde{l}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) = l_1(\mathbf{x}_{it}) + \overline{\mathbf{x}}_i$ satisfy $Q\big(\widetilde{l}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i)\big) = Q\big(l_0(\mathbf{x}_{it})\big)$ and

(ii) $\widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) = m_1(\mathbf{x}_{it}) + \overline{\mathbf{x}}_i$ satisfy $Q\big(\widetilde{m}_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i)\big) = Q\big(m_0(\mathbf{x}_{it})\big)$.

Orthogonal Score Function

- Estimating θ_0 requires to calculate the *orthogonal score function* with the ML plug-in predictions
- We *adapt* Chernozhukov et al. (2018)'s orthogonal score function to the panel data framework as

$$\boldsymbol{\psi}^{\perp}\Big(W; \theta_0, \boldsymbol{\eta}_0\Big) = \mathbf{v}_i' \mathbf{u}_i, \tag{12}$$

where $\mathbf{v}_i = (v_{i1}, \ldots, v_{iT})'$ and $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$ are defined as shown above, and with moment condition

$$\mathbb{E}\big\{\boldsymbol{\psi}^{\perp}(W; \theta_0, \boldsymbol{\eta}_0)\big\} = \mathbf{0} \tag{13}$$

- Use method of moments (MM) to estimate θ_0 based on the sample analogue of (13)

Estimation and inference with DML

Proposition 2. Suppose that conditions in Proposition 1 hold, the nuisance functions are Lipschitz continuous (i.e. defined on an interval and with bounded first derivatives), and suitable machine learning algorithms are available to *learn* the nuisance functions $\boldsymbol{\eta} = \{l_1, m_1\}$ at rate $N^{1/4}$.

The DML estimator solves the finite-sample analog of $\mathbb{E}\{\psi^{\perp}(W; \theta, \boldsymbol{\eta})\} = 0$

$$\frac{1}{N_k} \sum_{i \in \mathcal{W}_k} \psi(W_k; \theta, \widehat{\boldsymbol{\eta}}_k) = 0 \tag{14}$$

and has closed-form solution

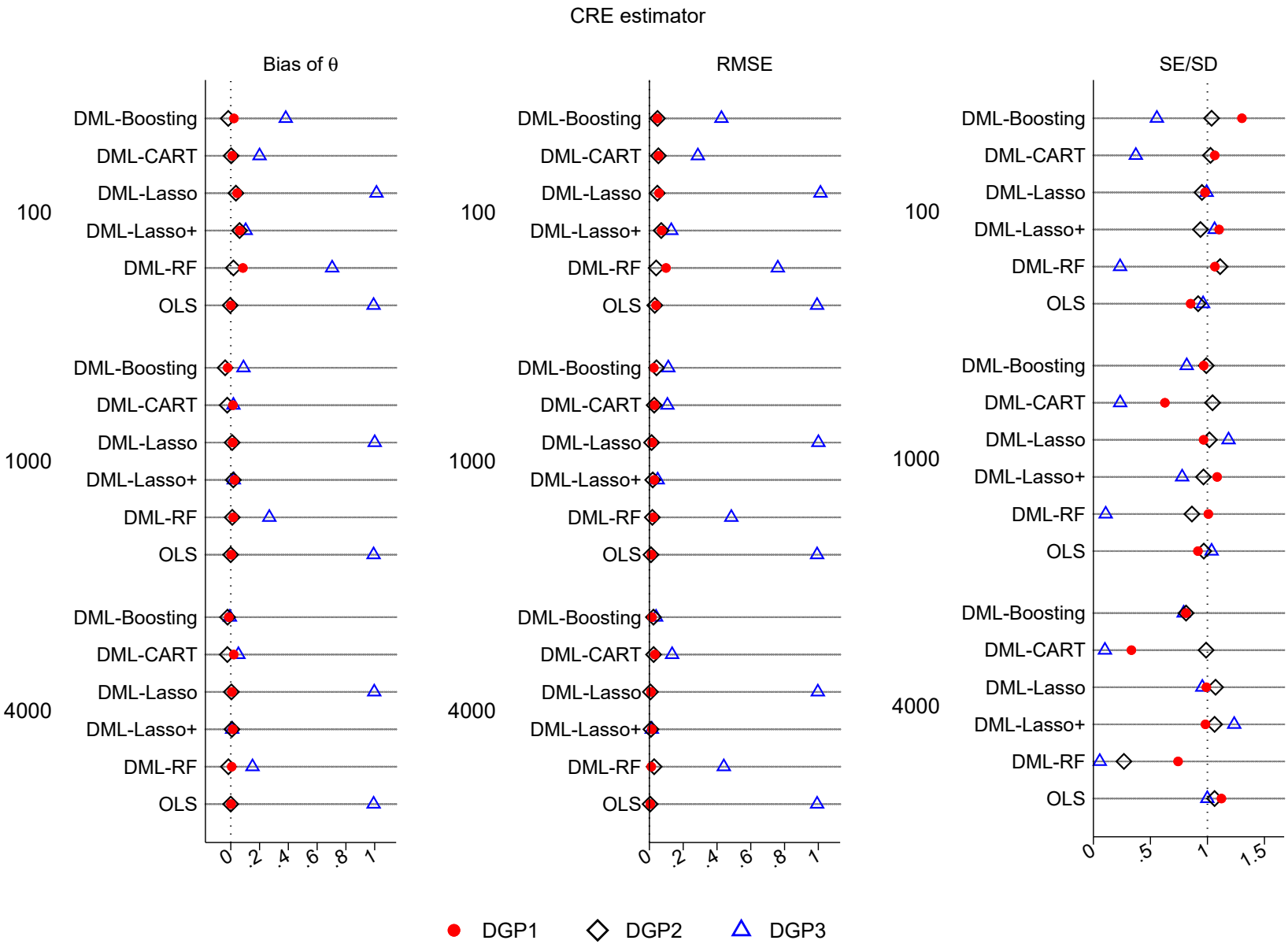
$$\widehat{\theta}_{DML} = \left(\frac{1}{N} \sum_{i \in W} \widehat{\mathbf{v}}_i' \widehat{\mathbf{v}}_i \right)^{-1} \frac{1}{N} \sum_{i \in W} \widehat{\mathbf{v}}_i' \widehat{\mathbf{u}}_i \tag{15}$$

which is $\sqrt{N}$-consistent for θ_0 with a normal limiting distribution as in Chernozhukov et al. (2018, Theorems 3.1 and 3.2) and approximate variance

$$\sigma^2 \equiv J_0^{-1} \mathbb{E}\big[\psi(W; \theta, \boldsymbol{\eta}) \psi(W; \theta, \boldsymbol{\eta})'\big] J_0^{-1}, \text{ where } J_0 = \mathbb{E}(\mathbf{v}_i' \mathbf{v}_i).$$

MC Simulations: CRE

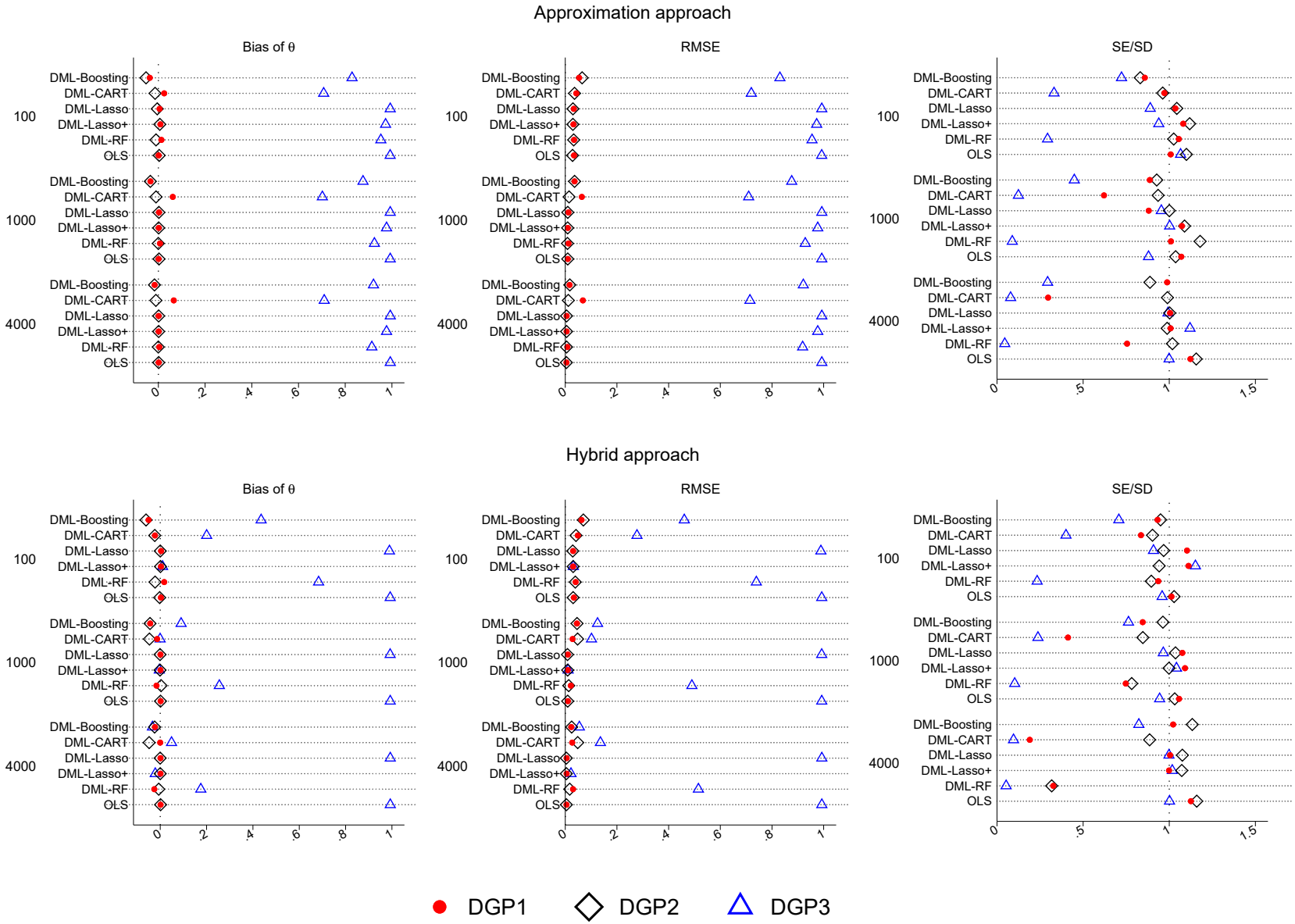
DGP



Note: DGP1 is linear in the covariates; DGP2 smooth non-linear; DGP3 non-smooth non-linear. $T = 10$, *100 MC reps.*

MC Simulations: WG

▸ DGP



Note: DGP1 is linear in the covariates; DGP2 smooth non-linear; DGP3 non-smooth non-linear. $T = 10$, *100 MC reps.*

Empirical Application

- The effect of the introduction of the National Minimum Wage (NMW) on voting behaviour in the UK as in Fazio and Reggiani (2023, EER)
- We replicate Specification (2) of Table (5) with DML for PLR model

$$Vote_{it} = NMW_{it}\theta + l(\mathbf{x}_{it}) + \alpha_i + u_{it} \tag{16}$$

$$NMW_{it} = m(\mathbf{x}_{it}) + \gamma_i + v_{it} \tag{17}$$

$Vote_{it}$ is binary outcome for voting for a conservative party; NMW_{it} is the treatment variable equal to one if the respondent was paid the NMW in 1999; $\mathbf{x}_{it}$ includes age, age squared education, marital status, household size, income of other members

- Base learners: Lasso w/dictionary, regression tree, random forest
- Data: British Household Panel Survey (BHPS)
- Like Fazio and Reggiani (2023), we find that benefiting from NMW increases the vote for conservative parties **but** with different magnitudes

			Approximation approach			*Hybrid approach*		
	OLS	OLS	DML-Lasso	DML-CART	DML-RF	DML-Lasso	DML-CART	DML-RF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: "Vote conservative"							
NMW	0.097**	0.088**	0.086***	0.091***	0.098***	0.079**	0.091**	0.095***
	(0.045)	(0.045)	(0.045)	(0.044)	(0.045)	(0.045)	(0.042)	(0.048)
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wave x Region FE	No	Yes	Yes	No	No	Yes	No	No
No. Observations	19,961	19,961	19,961	19,961	19,961	19,961	19,961	19,961
No. Groups	4,927	4,927	4,927	4,927	4,927	4,927	4,927	4,927

*Note: Column (1) reports the original figures of Specification (2) in Table 5 in Fazio and Reggiani (2023) estimated using least squares; Column (2) adds the interaction between wave and region fixed effects to Column (1); remaining columns use DML with different learners. Base control variables include: age, age squared (not for CART and RF), education, marital status, household size, income of other members. DML-Lasso uses an extended dictionary of non-linear terms of the control variables and fixed effects (i.e., polynomials of order three and interactions of the control variables). The hybrid approach includes the individual means of the control variables and fixed effects. Standard errors (in parenthesis) are clustered at the individual level. Standard errors (in parenthesis) are clustered at the individual level. Resampling information: 5 folds, cross-fitting, PO score, DML algorithm 2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Summary

- Simple problem (homogeneous effects, no lag dependence) but raised many challenges for ML
- Approximate transformation approaches can work but not as reliable as 'exact' approach; DML-CRE reliable estimates
- Bias reduction with DML-CART/RF/Lasso with extended dictionary even for very non-linear functions, unlike OLS and DML-Lasso w/t dictionary
- Tree-based approaches under-estimate SD of sampling distribution

Future research

- Future extension of DML within the panel data framework
  - Inclusion of instrumental variables (IV) in PLR model
  - Non-linear models (e.g., IRM by Chernozhukov et al., 2018)
  - Heterogeneous treatment effects following Nie and Wager (2021)
  - Dynamic panel models (e.g., Semenova et al., 2023)
- We will allowing for the use of ensemble/super-learning that is expected to outperform Lasso-based algorithms

Thank you for your attention!

✉ annalivia.polselli[at]essex.ac.uk

https://github.com/POLSEAN/XTDML

MiSoC ESRC Research Centre on Micro-Social Change
University of Essex
UKRI Economic and Social Research Council
ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

References I

Baiardi, A. and Naghi, A. A. (2021). The value added of machine learning to causal inference: Evidence from revisited studies. *arXiv preprint arXiv:2101.00878*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Fazio, A. and Reggiani, T. (2023). Minimum wage and tolerance for high incomes. *European Economic Review*, 155:104445.

Klosin, S. and Vilgalys, M. (2022). Estimating continuous treatment effects in panel data using machine learning with an agricultural application. *arXiv preprint arXiv:2207.08789*.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3):602–627.

References II

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, pages 69–85.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczyc*, 10:1–51.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, pages 299–319.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510.

Strittmatter, A. (2023). What is the value added by using causal machine learning methods in a welfare experiment evaluation? *Labour Economics*, 84:102412.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

MC Simulation Design

Back

- Three designs for the data generating process (DGP)
  - DGP 1: Linear in the confounders
  - DGP 2: Nonlinear but smooth (exponential and trigonometric functions)
  - DGP 3: Non-smooth and nonlinear (discontinuous function)

DGP

- $N = \{100, 1000, 4000\}$, $T = 10$
- $p = 30$ control variables, but $s = 2$ relevant (sparsity)
- Estimators: WG, FD, CRE
- Compare learners (Lasso, CART, RF) within the DML context wrt OLS and Oracle
- Hyperparameters of ML learners are tuned with random grid search

Tuning

DGP

Back

$$y_{it} = d_{it}\theta + l(\mathbf{x}_{it}) + \alpha_i + u_{it}$$
$$d_{it} = m(\mathbf{x}_{it}) + \zeta_i + v_{it}$$
$$\mathbf{x}_{it} \sim N(0, 5)$$
$$\alpha_i = 0.25 \left(\frac{1}{T} \sum_{t=1}^{T} d_{it} - \bar{d} \right) + 0.25 \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{it,k} + c_i, \text{ for } k = \{1, 3\}$$
$$\zeta_i \sim N(0, 1),\ c_i \sim N(0, 0.95)$$
$$a = 0.25,\ b = 0.5$$

Design 1	**Design 2**	**Design 3**
$m_0 = \mathbf{x}_{it,1} + a \cdot \mathbf{x}_{it,3}$	$m_0 = cos(\mathbf{x}_{it,1}) + a \cdot \frac{exp(\mathbf{x}_{it,3})}{1+exp(\mathbf{x}_{it,3})}$	$m_0 = a\left(\mathbf{x}_{it,1} \cdot \mathbb{1}[\mathbf{x}_{it,1} > 0]\right) + b\left(\mathbf{x}_{it,1} \cdot \mathbf{x}_{it,3}\right)$
$l_0 = \mathbf{x}_{it,1} + a \cdot \mathbf{x}_{it,3}$	$l_0 = \frac{exp(\mathbf{x}_{it,1})}{(1+exp(\mathbf{x}_{it,1}))} + a \cdot cos(\mathbf{x}_{it,3})$	$l_0 = b\left(\mathbf{x}_{it,1} \cdot \mathbf{x}_{it,3}\right) + a\left(\mathbf{x}_{it,3} \cdot \mathbb{1}[\mathbf{x}_{it,3} > 0]\right)$

$N = \{100, 1000, 10000\}$
$T = 10$
$p = 30,\ s = 2$ (sparsity)
$R = 100$ MC replications

Hyperparameter tuning

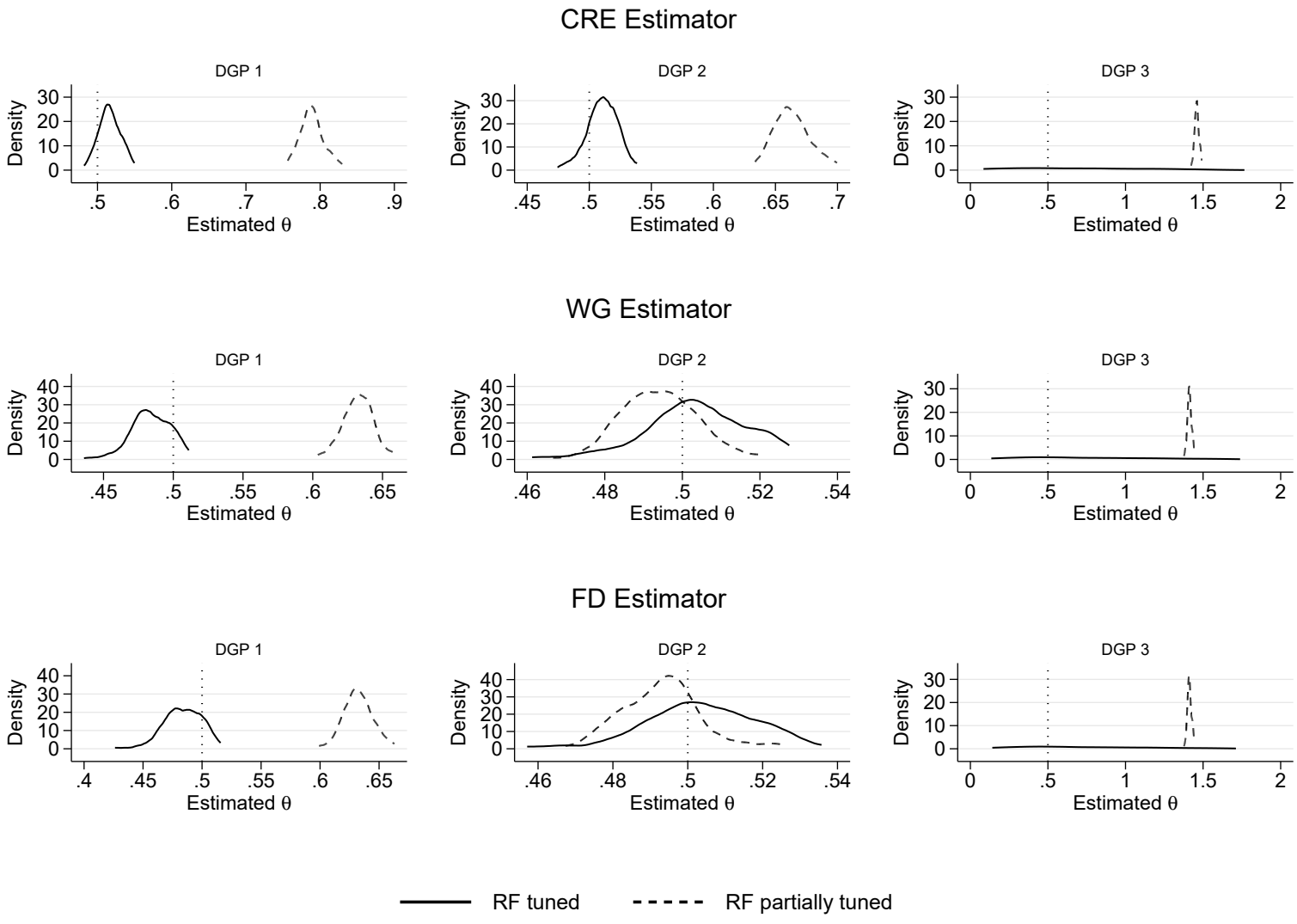
Back

Table 1: Hyperparameter tuning

Learner	Hyperparamters	Value of parameter in set	Description
Lasso	`lambda.min`	–	λ equivalent to minimum mean cross-validated error
CART	`cp`	real value in $\{0.01,0.02\}$	Prune all nodes with a complexity less than cp from the printout.
	`minbucket`	integer in $\{5, \lceil N/2 \rceil\}$	M inimum number of observations in any terminal leaf node.
	`maxdepth`	integer in $\{1,10\}$	Maximum depth of any node of the final tree.
Boosting	`lambda`	real value in $\{0,10\}$	L2 regularization term on weights.
	`maxdepth`	integer in $\{5,10\}$	Maximum depth of any node of the final tree.
	`nrounds`	100	Number of decision trees in the final model
RF	`num.trees`	integer in $\{5,100\}$	Number of trees in the forest.
	`min.node.size`	integer in $\{5, \lceil N/2 \rceil)\}$	Minimal node size to split at.
	`max.depth`	integer in $\{1,10\}$	Maximum depth of any node of the final tree.

Note: Hyperparameter tuning for CART, boosting and RF is conducted with a random grid search. For RF, nodes with size smaller than min.node.size can occur.

Sampling distribution of $\widehat{\theta}$



Note: 'Tuned' RF use the optimal configuration of hyperparameters from the ranges reported in Table 1. 'Partially tuned' RF uses 100 trees, maximum depth is 100, and the minimum node size is tuned. Hyperparameters are tuned via grid search. $N = 1,000$.