Quantitative Economics 14 (2023), 471–510 1759-7331/20230471

Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence

Vira Semenova
Department of Economics, University of California, Berkeley

Matt Goldman
Meta

Victor Chernozhukov
Department of Economics, MIT

Matt Taddy
Amazon

This paper provides estimation and inference methods for conditional average treatment effects (CATE) characterized by a high-dimensional parameter in both homogeneous cross-sectional and unit-heterogeneous dynamic panel data settings. In our leading example, we model CATE by interacting the base treatment variable with explanatory variables. The first step of our procedure is orthogonalization, where we partial out the controls and unit effects from the outcome and the base treatment and take the cross-fitted residuals. This step uses a novel generic cross-fitting method that we design for weakly dependent time series and panel data. This method "leaves out the neighbors" when fitting nuisance components, and we theoretically power it by using Strassen's coupling. As a result, we can rely on any modern machine learning method in the first step, provided it learns the residuals well enough. Second, we construct an orthogonal (or residual) learner of CATE—the lasso CATE—that regresses the outcome residual on the vector of interactions of the residualized treatment with explanatory variables. If the complexity of CATE function is simpler than that of the first-stage regression, the orthogonal learner converges faster than the single-stage regression-based learner. Third, we perform simultaneous inference on parameters of the CATE function using debiasing. We also can use ordinary least squares in the last two steps when CATE is low-dimensional. In heterogeneous panel data settings, we model the unobserved unit heterogeneity as a weakly sparse deviation from Mundlak's (1978) model of correlated unit effects as a linear function of time-invariant covariates and make use of L1-penalization to estimate these models.

Vira Semenova: semenovavira@gmail.com
Matt Goldman: mattgoldman5850@gmail.com
Victor Chernozhukov: vchern@mit.edu
Matt Taddy: mataddy@gmail.com

We thank Michael Jansson, Pat Kline, Sylvia Klosin, Demian Pouzo, and two anonymous referees whose comments helped improve the paper. We thank Microsoft Research Redmond and Simons Institute for the Theory of Computing for a productive research environment.

 https://doi.org/10.3982/QE1670

We demonstrate our methods by estimating price elasticities of groceries based on scanner data. We note that our results are new even for the cross-sectional (i.i.d.) case.

KEYWORDS. Orthogonal learning, residual learning, CATE, dynamic panel data, time series, mixing, cross-fitting, neighbors-left-out.

JEL CLASSIFICATION. C14, C23, C33.

1. INTRODUCTION

Inference on heterogeneous treatment effects is an important problem (see, e.g., Athey and Imbens (2016), Chernozhukov, Demirer, Duflo, and Fernández-Val (2017), Wager and Athey (2018), Davis and Heller (2020), Banerjee et al. (2021)). Estimating these effects involve an inherent trade-off between flexibility and precision. On the one hand, discovering heterogeneity requires flexible models for the effects (e.g., by considering many groups). On the other hand, flexible models produce noisy estimates that are not useful for making decisions (e.g., the noise can result from having too few observations per group). To resolve this trade-off, empiricists decide how to create groups after making multiple attempts, a subjective, labor-intensive method that is prone to erroneous inference.

This paper contributes a method for the estimation and inference of heterogeneous treatment effects in a panel data set with many potential controls and unobserved unit heterogeneity, which addresses many of the challenges listed above. Our key results are new even to the cross-sectional settings. Thus, we first consider the cross-sectional case and the following leading example as our model to explain the key ideas. Suppose Y is an outcome, and P is a vector of treatment variables (e.g., polynomials in price), and X are controls. Suppose that the conditional expectation function is partially linear in P, as in Robinson (1988), namely,

$$Y = e_0(X) + \beta_0(X)'P + U, \qquad \mathrm{E}[U \mid P, X] = 0.$$

Here, $e_0(X)$ is the conditional average outcome in the "untreated" state ($P = 0$), and $\beta_0(X)$ describes the conditional average treatment effect under the standard conditional ignorability/exogeneity conditions. We can *orthogonalize* the outcome and treatment with respect to controls X, obtaining the residuals $\widetilde{Y} = Y - \mathrm{E}[Y \mid X]$ and $\widetilde{P} = P - \mathrm{E}[P|X]$, and then observe that the CATE function satisfies the equation:

$$\widetilde{Y} = \beta_0(X)'\widetilde{P} + U.$$

Therefore, we can learn the CATE function from this regression equation if we can learn the residuals $\widetilde{Y}$ and $\widetilde{D}$ well using modern machine learning methods. In fact, under certain conditions, we prove that our rate of learning the CATE function would be the same as if we knew the true residuals, which is an oracle-type property. Moreover, we show this under both high-level and low-level regularity conditions.

Our approach consists of approximating $\beta_0(X)$ by a linear combination of terms of a dictionary of transformation $K = K(X)$ of X, which includes a constant of 1,

$$\beta_0(X) \approx K'\beta_0,$$

If dimension $d = \dim(\beta_0)$ is low, that is, d is much smaller than the sample size n, we can learn β_0 using least squares at the rate $\sqrt{d/n}$ provided the expectation functions $\mathrm{E}[Y \mid X]$ and $\mathrm{E}[P \mid X]$ are learnable at fast enough rates. If β_0 is high-dimensional and sparse, we will rely on lasso to learn the CATE function at the rate $\sqrt{s \log d/n}$ where $s = \|\beta_0\|_0$ is the number of nonzero entries of β_0. Finally, we will use debiased lasso methods to perform Gaussian inference on the components of β, including constructing simultaneous confidence bands using fast (Gaussian) bootstrap methods. We call these new approaches above the orthogonal lasso and debiased orthogonal lasso. In addition, we also explore the use of grouped lasso methods to enforce the exclusion or inclusion of groups of variables.

Our paper considers the dynamic panel data setting arising in many empirical applications. This setup makes the problem a lot more challenging. First, all variables above will be doubly-indexed by unit $i = 1, \ldots, N$ and time $t = 1, \ldots, T$, and controls can include lagged values of outcomes, for example, and we will need to introduce unit-specific effects in the model above judiciously. We add the unit-specific effects to the conditional expectations of Y_{it} and P_{it}, and we model individual effects as linear functions of time-invariant covariates plus fixed effects that are approximately sparse. This constriction allows for the overall individual effect to be "dense" while providing enough convenience to make estimation results work. The strategy above is motivated by Mundlack's and Chamberlain's approach to viewing and modeling fixed effects as correlated random effects.

Our construction uses cross-fitting (CF) to estimate nonparametric reduced forms on a subset of data and construct the residuals (or scores) on another subset. In the i.i.d. setting, CF removes the overfitting biases that can arise from using complex nonparametric methods such as machine learning methods (see, e.g., Belloni, Chernozhukov, and Hansen (2010), Zheng and van der Laan (2010), Chernozhukov et al. (2018) for recent examples and Hasminskii and Ibragimov (1979), Schick (1986) for early, classical uses). For regular CF methods to work in a time series or unit-heterogeneous dynamic panel data, the number of periods T must be very large relative to the number of units N. We choose an alternative path and introduce a "neighbors-left-out" (NLO) cross-fitting method that applies to weakly dependent data. The NLO approach ensures that the first-stage and the second-stage samples are approximately independent. We provide exact bounds on the approximation error via Strassen's coupling. These results are of independent interest and apply beyond our context.

We use our method to estimate heterogeneous price elasticities on grocery data as an empirical application. This data set consists of textual descriptions of the products, prices, and daily aggregate sales for each (store, product, distribution channel) combination. We posit a partially linear specification where the (log) sales are the dependent variable, and lags of log prices and log sales and current product characteristics are the control variables. Assuming that the residual between the price tomorrow and its expectation today is exogenous, we use this variation to identify price elasticities. The approximate sparsity assumption helps us to rule out implausible values of price elasticities. Our estimates are broadly consistent with findings in Chevalier, Kashyap, and Rossi (2003).

All of the above constitute the principal contributions of the paper. In what follows, we describe the relations to the literature and mention some additional extensions and contributions. First, the paper contributes to modern literature on estimation and model selection in high-dimensional settings using debiased (orthogonal) machine learning (e.g., Hasminskii and Ibragimov (1979), Schick (1986), Belloni, Chernozhukov, and Hansen (2010), Zheng and van der Laan (2010), Belloni, Chernozhukov, and Hansen (2011), Belloni, Chernozhukov, and Kato (2014), Zhang and Zhang (2014), van der Geer, Bühlmann, Ritov, and Dezeure (2014), Chernozhukov et al. (2018), and references therein) by considering the high-dimensional CATE function as the focus of inference. Prior literature has mainly focused inference on low-dimensional or many target parameters without leveraging the model to help residualization (e.g., Belloni, Chernozhukov, and Kato (2014), Belloni, Chernozhukov, Chetverikov, and Wei (2019)). While our results are new, even for cross-sectional settings, our results cover the dynamic panel data settings.

We provide general theoretical guarantees for orthogonal lasso methods that apply to any case where residuals are learned well enough in a preliminary step using general machine learning methods. In cross-sectional settings, this automatically allows a wide range of high-quality machine learning tools with rigorous guarantees. In panel data settings, we rely on lasso and verify that we can learn the residuals well using lasso-based methods with weakly sparse individual effects, relying here upon in Kock and Tang (2019). We expect that other machine learning methods are potentially amenable to handling dynamic panel data settings, which is the subject of future work. In this work, we abstract away from clustering (Chiang (2018) and Chiang, Kato, Ma, and Sasaki (2019)), but it would be good to extend the present results in this direction.

In a related paper to ours, Nie and Wager (2020) establishes that the oracle rate of learning of CATE function is possible in a cross-sectional setting, proposing a similar residual regression approach. Our paper is independent, and we circulated the paper around the same time as theirs (both in December of 2017 in ArXiv). Moreover, we provide not only the oracle learning rates but also statistical inference results and also cover the dynamic panel data setting. On the other hand, rate results of Nie and Wager (2020) apply to nonlinear learners of the CATE function. A more recent work than ours is Oprescu, Syrgkanis, and Wu (2018), which develops orthogonal forest methods. Specifically, they apply generalized random forest to regress outcome residual on treatment residual interacted with a forest function of controls. They also provide some inferential results. Finally, alternative approaches to handling heterogeneous and/or continuous treatment effects are discussed in Ura (2018), Wager and Athey (2018), Semenova and Chernozhukov (2021), Jacob, Härdle, and Lessmann (2019), Fan, Hsu, Lieli, and Zhang (2019), Zimmert and Lechner (2019), Colangelo and Lee (2020), Klosin (2021).

To conduct inference on high-dimensional parameters of the CATE function, we combine the approach of Zhang and Zhang (2014) and van der Geer et al. (2014) with the Cai, Liu, and Luo's (2011) approach to matrix inversion. The inference step can also be carried out by the methods of Javanmard and Montanari (2014) and the double lasso method (Belloni, Chernozhukov, and Hansen (2014), Chernozhukov, Hansen, and Spindler (2015)), but we focus on the former. We rely on fast (Gaussian) bootstrap

to perform simultaneous inference based on many debiased lasso estimators, relying upon (Chernozhukov, Chetverikov, and Kato (2014, 2017, 2019)) and suitably extending some results to our settings. Finally, we build on (Mundlak (1978), Chamberlain (1982)) and contribute to the panel data literature that develops various approaches to handling heterogeneity, for example, Kock (2016b), Manresa (2016), Lu and Su (2016), Su, Shi, and Phillips (2016), Moon, Shum, and Weidner (2018), Kock and Tang (2019), Bonhomme, Lamadon, and Manresa (2019a,b), Gao and Li (2019), Chen, Fernandez-Val, and Weidner (2020), Lu and Su (2020) among many others; see Fernandez-Val and Weidner (2018) for a recent overview.

Structure of the paper

Section 2 introduces the model and outlines the strategy. Section 3 gives definitions of estimators and outlines some theoretical results. Section 4 states our theoretical results under general high-level conditions about the first stage. Section 5 verifies the high-level conditions focusing on the panel data settings. Section 6 gives an empirical application, and Section 7 concludes. Appendix A in the Online Supplementary Material (Semenova, Goldman, Chernozhukov, and Taddy (2023)) presents and results on independence couplings. Appendix B develops concentration results for weakly-dependent panel data. Appendix C presents the results for high-dimensional CLT for weakly dependent data. Appendix D contains proofs for Section 4, and Appendix E for Section 5. Appendix F contains tail bounds for empirical rectangular matrices in operator norm. Appendix G contains the analysis of OLS used in stage 3 of our inference procedure.

2. The set up

Here, we present the model, explain how we handle unit level heterogeneity, and outline the overall inferential strategy.

Model

Our starting point is the structural equation model

$$Y_{it} = \beta_0(X_{it}, P_{it}) + e_0(X_{it}) + \xi_i^E + U_{it}, \tag{2.1}$$

where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. Here,

- Y_{it} is a scalar outcome of unit i at time t;
- $P_{it} \in \mathrm{R}^{d_P}$ is a vector of treatment or policy variables;
- $X_{it} \in \mathrm{R}^{d_X}$ is a vector of *predetermined* variables, including possibly the lags of P_{it} and Y_{it};
- ξ_i^E is an unobserved outcome unit fixed effect;
- $M_i = \{M_{it}\}_{t=1}^T$ is a a collection fixed variables;

- X_{it} can include known functions M_i, for example, time averages $\bar{M}_i$ of M_i.

The stochastic shock U_{it} is assumed to satisfy the following sequential conditional exogeneity condition:

$$\mathrm{E}[U_{it} \mid P_{it}, X_{it}, \Phi_{it}] = 0, \quad \forall (i, t), \tag{2.2}$$

where the filtration

$$\Phi_{it} = \left\{ (X_{it'}, P_{it'}, Y_{it'})_{t'=1}^{t-1} \right\} \tag{2.3}$$

is the filtration that consists of predetermined variables for unit i prior to period t. Here, we view $M_i = \{M_{it}\}_{t=1}^{T}$'s as a fixed realization of strictly exogenous variables that can be time-varying. These variables are strictly exogenous, meaning that their entire trajectory has been predetermined relative to all other variables in the model and relative to stochastic shocks U_{it}'s.

REMARK 2.1 (Fixed Effects). Throughout the paper, we assume that

$$\{M_i, \xi_i\}_{i=1}^{N} \text{ are fixed.}$$

We view this approach as (essentially) equivalent to treating these variables as random initially and then performing the analysis conditional on their realized values.[1]

REMARK 2.2 (Important Notation Remark). Note that below we will be reassigning notation $X_{it} \leftarrow \mathrm{t}(X_{it})$ to denote variables that have been obtained as transformations of the original variables X_{it} via some mapping t. Examples of transformations include powers and their interactions. We then shall make other modeling assumptions to model the observable unit-level heterogeneity.

The structural function $p \mapsto \beta_0(x, p)$ encodes the conditional average treatment effects (CATE). Therefore, we will simply call this function the CATE function. Indeed consider the intervention policy that fixes $P_{it} = p$ in the structural equation (2.1), inducing the potential outcome:[2]

$$Y_{it}(p) := \beta_0(X_{it}, p) + e_0(X_{it}) + \xi_i^E + U_{it}.$$

Then we have that

$$\beta_0(X_{it}, p_1) - \beta_0(X_{it}, p_0) = \mathrm{E}\big[Y_{it}(p_1) \mid X_{it}\big] - \mathrm{E}\big[Y_{it}(p_0) \mid X_{it}\big]$$

is the CATE resulting from changing policy value from p_0 to p_1.

[1] As in the standard panel data modes with fixed effects, we do not formalize this conditioning to reduce presentation complexity.

[2] Here, as in Haavelmo (1944), we assume that the structural equation remains invariant under the intervention. Judea Pearl refers to the fact that the structural model implies potential outcomes as the first law of causal inference.

In what follows, we will assume that $\beta_0(X_{it}, P_{it})$ is well approximated by a linear combination of terms of a dictionary

$$D_{it} := D(X_{it}, P_{it})$$

of transformations of X_{it} and P_{it} so that

$$\beta_0(X_{it}, P_{it}) = D_{it}'\beta_0.$$

Putting things together, we arrive at the partially linear model:

$$Y_{it} = D_{it}'\beta_0 + e_0(X_{it}) + \xi_i^E + U_{it}, \tag{2.4}$$

where the key parameter β_0 is interpretable as a causal or treatment effect parameter. We will refer to P_{it} and D_{it} as *base* and *technical* treatment vectors, respectively.

In this paper, we focus on a practical case when the complexity of the control function $e_0(X_{it})$ substantially exceeds the complexity of CATE function (see Remark 5.7 for the formal comparison of complexities).

Reduced forms and orthogonalized equations

To learn the CATE function at its fastest possible rate, we need to partial out controls from treatments and outcome. Consider the treatment equation:

$$D_{it} = d_{i0}(X_{it}) + V_{it}, \qquad \mathrm{E}[V_{it} \mid X_{it}, \Phi_{it}] = 0, \tag{2.5}$$

which keeps track of confounding. We assume that the *unit-specific treatment reduced form* takes the form:

$$\mathrm{E}[D_{it} \mid X_{it}, \Phi_{it}] =: d_{0i}(X_{it}) = d_0(X_{it}; \xi_i), \tag{2.6}$$

where $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ denotes a fixed vector of unit-specific fixed treatment-selection effects. A special case $d_{0i}(X_{it}) := d_0(X_{it})$ corresponds to no unobserved unit heterogeneity in treatment. Furthermore, if the function $d_0(X_{it})$ is constant itself, there is no confounding.

Proceeding further, we model *the unit-specific outcome reduced form* as

$$\mathrm{E}[Y_{it} \mid X_{it}, \Phi_{it}] =: l_{i0}(X_{it}) = d_{i0}(X_{it})'\beta_0 + e_0(X_{it}) + \xi_i^E, \tag{2.7}$$

where $\xi^E = (\xi_1^E, \xi_2^E, \ldots, \xi_N^E)$ denotes a fixed vector of unit-specific outcome effects.

Given the outcome and treatment reduced forms, we define the treatment and outcome residuals

$$V_{it} := D_{it} - d_{i0}(X_{it}), \qquad \widetilde{Y}_{it} := Y_{it} - l_{i0}(X_{it}). \tag{2.8}$$

Equations (2.4)–(2.2) imply the following orthogonolized regression equation:

$$\widetilde{Y}_{it} = V_{it}'\beta_0 + U_{it}, \qquad \mathrm{E}[U_{it} \mid V_{it}, X_{it}, \Phi_{it}] = 0. \tag{2.9}$$

This equation identifies β_0 as the coefficient of the best linear projection of $\widetilde{Y}_{it}$ on V_{it}.

EXAMPLE 2.1 (Linear in Treatment Base Treatment Structure). Define

$$D_{it} = P_{it} K_{it}, \tag{2.10}$$

where $K_{it} := K(X_{it})$ is a collection of transformations of a subset of variables in X_{it}, including a constant of 1. Suppose there exists a low-dimensional "base" treatment variable P_{it} whose reduced form is

$$P_{it} = p_0(X_{it}) + \xi_i + V_{it}^P, \qquad \mathrm{E}\big[V_{it}^P \mid X_{it}, \Phi_{it}\big] = 0. \tag{2.11}$$

Then, (2.5), (2.10), and (2.11) imply

$$d_{i0}(X_{it}) = K(X_{it})'\big(p_0(X_{it}) + \xi_i\big), \qquad V_{it} = K(X_{it}) V_{it}^P.$$

As we will show later, the interactive structure (2.10) simplifies estimation of treatment residuals.

Unit-level effects

A standard approach to unit-level additive heterogeneity is demeaning or differencing. Because these operations introduce Nickell (1981) bias in dynamic panels, it requires an identification strategy based on instrumental variables (e.g., Arellano and Bond (1991)). Furthermore, differencing out time-invariant covariates may lead to an efficiency loss.

In this paper, we take a fixed effect approach, in which we approximate the vector of unobserved components of unit effects $\xi^E = (\xi_i^E)_{i=1}^N$ by a weakly sparse vector. Informally, the weak sparsity assumption requires ξ^E to be well approximated by a sparse vector whose number of nonzero components is small. The sparsity assumption allows us to use lasso methods to consistently estimate them (Kock and Tang (2019)).

The weak sparsity assumption may appear restrictive at the first sight. However, it does allow for rich forms of overall unit-level effects driven by time-invariant covariates and the "residual" unit effects ξ_i. To explain this better, consider the following Mundlack-style model:

$$e_0(X_{it}) + \xi_i^E = \bar{X}_{it}' \delta_0^X + \underbrace{\bar{M}_i' \delta_{M0}^E + \xi_i^E}_{a_i^E}, \tag{2.12}$$

where $\bar{X}_{it}$ are time-varying, predetermined covariates and $\bar{M}_i = \frac{1}{T}\sum_{t=1}^T M_{it}$ is time average of fixed covariates. The important difference with Mundlack's approach is that we consider ξ_i's to be weakly sparse and to condition on the realizations of $(\bar{M}_i, \xi_i)_{i=1}^N$.[3]

We note that while the residual effects ξ_i's are required to be weakly sparse, the overall unit effect a_i can actually be *dense*. Finally, the decomposition $a_i^E = \bar{M}_i' \delta_{M0}^E + \xi_i^E$ may not be unique. However, our analysis shows that this nonuniqueness does not prevent

[3]It would be interesting to consider nonweakly sparse ξ in dynamic panel models, where ξ's follow some known distribution (which is generally not compatible with the weak sparsity assumption). We leave this important direction to future research. Note that Kock (2016a) developed such results for nondynamic panel data models, which could provide a starting point for such extension.

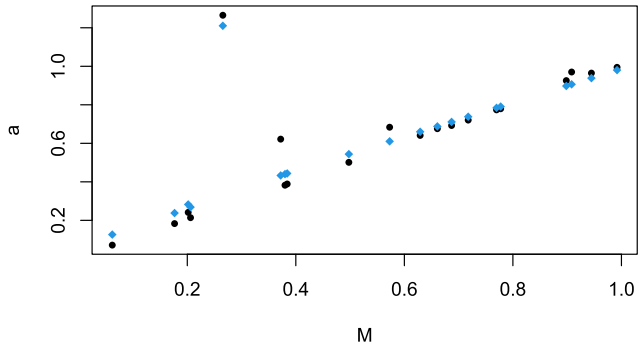


FIGURE 1. Lasso approximation of a correlated random effects model with approximately sparse deviations. *Notes*: Dots indicate (M_i, a_i), with the horizontal axis showing values of $\bar{M}_i$ and vertical axis the values of $a_i = M_i'\delta_0^M + \xi_i$. The lasso estimated unit effects are shown by rombi $(M_i, \widehat{a}_i)$, where $\widehat{a}_i := M_i\widehat{\delta}_M + \widehat{\xi}_i$. The time-invariant controls M_i are generated as i.i.d. draws from $U[0, 1]$; the sparse deviations are $\xi_i = 1/i^2$, $i = 1, 2, \ldots, N = 20$, $T = 1$; and $\delta_0^M = 1$. Here, we show the realization just for one experiment.

the overall a_i^E's be consistently estimated, as we illustrate in Figure 1, as long as there exist at least one decomposition with δ_{M0}^E and ξ^E being sufficiently sparse. We provide a more technical explanation in Remark 5.2.

For the unit-level heterogeneity in treatment, we can proceed similarly. This strategy works especially well in conjunction with linear structures such as Example 2.1, where the same approach as above applies, swapping ξ_i^E for ξ_i, so that

$$p_0(X_{it}) + \xi_i = \bar{X}_{it}'\delta_0^P + \underbrace{\bar{M}_i'\delta_{M0}^P + \xi_i}_{a_i^P}, \tag{2.13}$$

where a_i^P is the overall unit-level effect, consisting of a dense part $\bar{M}_i'\delta_{M0}^P$ plus a weakly sparse deviation ξ_i from it.

Additive unit heterogeneity works well for linear models such as in Example 2.1. On the other hand, purely additive fixed effects are not well suited for binary or discrete treatments.[4] In the latter case, empirical researchers may proceed as follows: supposing P_{it} is binary, we model

$$P_{it} = \Lambda\big(\xi_i + \bar{X}_{it}'\beta + \bar{M}_i'\delta\big) + V_{it}^P, \qquad \mathrm{E}\big[V_{it}^P \mid X_{it}, \Phi_{it}\big] = 0,$$

where $z \to \Lambda(z)$ is the link function such as logit that forces the logical range restriction on the conditional expectation function. The fixed effects here are naturally nonadditive (though additive inside the link function). Then here one can still impose approximate sparsity on $\xi = (\xi_i)_{i=1}^N$ and apply lasso-penalized logistic regression to estimate such models in practice. We expect that the results of Kock and Tang (2019) extend to this case, but this requires its own formal analysis that we leave to future work.

[4]For example, in the binary case, the conditional expectation function of P_{it} is naturally bounded by 0 and 1, but the additive fixed effects model does not naturally respect this range.

The above discussion is still somewhat abstract. We thus present the following concrete example that illustrates the flexibility of the proposed framework. In this example, the lasso-based methods are particularly helpful for both estimation of reducted forms and residuals. We will use this example to illustrate the plausibility of regularity conditions that we invoke later in the paper.

EXAMPLE 2.2 (Linear Panel Vector Autoregression With High-Dimensional Controls and Unit Effects). The following example is the special case of Example 2.1:

$$\begin{aligned} Y_{it} &= P_{it}K_{it}'\beta_0 + e_0(X_{it}) + \xi_i^E + U_{it} \\ &= P_{it}K_{it}'\beta_0 + \sum_{l=1}^{L} Y_{i,t-l}\delta_{0l}^{EE} + \sum_{l=1}^{L} P_{i,t-l}\delta_{0l}^{EP} + \bar{X}_{it}'\bar{\delta}_0^E + \bar{M}_i'\delta_{M0}^E + \xi_i^E + U_{it} \\ &= P_{it}K_{it}'\beta_0 + X_{it}'\delta_0^E + \xi_i^E + U_{it}, \end{aligned} \tag{2.14}$$

$$\begin{aligned} P_{it} &= p_0(X_{it}) + \xi_i + V_{it}^P \\ &= \sum_{l=1}^{L} P_{i,t-l}\delta_{0l}^{PP} + \sum_{l=1}^{L} Y_{i,t-l}\delta_{0l}^{PE} + \bar{X}_{it}'\bar{\delta}_0^P + \bar{M}_i'\delta_{M0}^P + \xi_i + V_{it}^P \\ &= X_{it}'\delta_0^P + \xi_i + V_{it}^P. \end{aligned} \tag{2.15}$$

In this example, the outcome responds to the current and past values of the treatment as well as past values of outcomes; a set of covariates and unit effects provide further shifts. Likewise, the treatment is assigned in response to current and past values of the treatment as well as past values of outcomes; and a set of covariates and unit effects provide further shifts.

Estimation and inference strategy

We are primarily interested in the *high-dimensional sparse* regime, where the number of technical treatments d is large

$$d = \dim(D_{it}) = \dim(\beta_0) \gg NT,$$

but only a small number $s \ll NT$ of them has nonzero effect:[5]

$$\|\beta_0\|_0 = s, \tag{2.16}$$

Importantly, the identity and the number of the nonzero coefficients is unknown.

ALGORITHM 1. In this high-dimensional regime, our estimation and inference approach have the following steps:

(1) Estimate the residuals $\widetilde{Y}_{it}$ and V_{it} using machine learning with cross-fitting.

[5]We can relax this exact sparsity assumption to approximate/weak sparsity as in Belloni, Chernozhukov, and Kato (2014). We chose a simpler assumption given the complexity of the rest of the analysis.

(2) Estimate the CATE function by lasso-penalized regression of estimated $\widetilde{Y}_{it}$ on V_{it}.

(3) Perform Gaussian inference parameters of the CATE function using debiased lasso.

In the last step, more specifically, we are performing inference on classes of the linear functionals of parameters β_0 of the CATE function $D_{it}'\beta_0$. In cross-sectional settings, a wide variety of machine learning methods provably apply to carry out step 1. In panel data settings, carrying out step 1 requires a judicious mix of modeling structures and machine learning methods that can handle fixed effects. Structures such as Example 2.1 and lasso with penalized fixed effects work provably well for this purpose. Other methods potentially apply, but this remains to be proven. Moreover, for step 1 we have to design cross-fitting to respect the panel data structure. The last step uses debiasing most similar to that of van der Geer et al. (2014), but other methods such as double lasso can also be used to carry out debiasing. The next section provides formal definitions of estimation steps focusing on the dynamic panel data case.

If D_{it} is low-dimensional, namely,

$$d \ll NT,$$

the sparsity assumption is not required. The steps (2) and (3) are replaced by linear regression estimated by ordinary least squares.

Algorithm 2. If $d \ll NT$, we perform the following steps:

(1′) Estimate the residuals $\widetilde{Y}_{it}$ and V_{it} using machine learning with cross-fitting.

(2′) Estimate the CATE function by linear regression of estimated $\widetilde{Y}_{it}$ on V_{it}.

(3′) Perform Gaussian inference on parameters of the CATE function using OLS.

This covers many practical cases, and is a very attractive applied option. We point our, however, that even in this regime, if the sparsity condition $s \ll d$ holds, the orthogonal lasso methods can outperform OLS in terms of accuracy of estimating the CATE function. Furthermore, we also note that the OLS method is not designed to handle the model selection problem. Indeed, typically researchers combine OLS with prior model selection step, which can lead to well-known inferential problems Leeb and Potcher (2005). The debiased orthogonal lasso explicitly addresses the model selection issue and provides rigorous theoretical guarantees for inference.

3. Cross-fitting with time series, estimation, and inference

In this section, we introduce the neighbor-excluding cross-fitting method that is generally applicable to weakly dependent time series or panel data. We also provide a key theoretical support for this method using Strassen's coupling. In this section, we also write down details of some estimators, focusing on the dynamic panel data case. We also informally preview theoretical results.

Notation

In the remainder, we use notation W_{it} to refer to the data vector on unit i at time t; $W_{\cdot,t}$ the data vector on all units at time t, and so on. The data and all other random elements are defined on the underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$ that has been enriched to carry an independent standard uniform random variable. We assume that all random variables are random vectors in Euclidean spaces (the coupling result below applies to random variables taking values in Polish space). We denote the total variation of a signed measure v the space $(\Omega, \mathcal{F})$ as

$$\|v\|_{\mathrm{TV}} := \sup v(A) - v\big(A^c\big),$$

where the supremum is taken over all measurable sets A.

3.1 *Cross-fitting for weakly dependent data*

Cross-fitting (CF) reduces overfitting biases from fitting the model's nonparametric components via machine learning. In the i.i.d. settings, CF uses a one subsample to estimate nonparametric components (e.g., expectation functions) and its complement to compute the sample average of i.i.d. residuals (depending on these functions). As a result, CF plays an essential role in modern debiased inference in semiparametric models; see, e.g., Belloni, Chernozhukov, and Hansen (2010), Zheng and van der Laan (2010), Chernozhukov et al. (2018) for recent examples and Hasminskii and Ibragimov (1979) and Schick (1986) for early, classical uses of simpler sample-splitting methods for debiased inference.

In cross-sectional cases, we create partitions into folds and their complements by sampling folds randomly from the data. In the time-series case, there is only one dimension to split on. In the unit-heterogeneous panel settings, we can only split by time dimension so that unit-specific effects are estimated for all units on every partition. In both cases, two contiguous time splits may not be independent.[6] Here, we introduce a "neighbors-left-out" (NLO) cross-fitting method that applies to weakly dependent data. Whenever the data are weakly dependent, the NLO approach ensures that the first- and the second-stage samples are approximately independent. We give exact bounds on the approximation error by independent blocks via Strassen's coupling below.

Definition 3.1 (Folds and Their Quasi-Complements for Weakly Dependent Data). Consider partition of $\{1, \ldots, T\}$ into adjacent blocks $\{\mathcal{M}_k\}_{k=1}^K$,

$$\{1, \ldots, T\} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\},$$

[6]For example, if we cut a panel into two halves, we end up with two dependent data blocks. We can, in principle, make this approach work under beta-mixing by recognizing that the dependence has a vanishing effect on statistics that are sample averages of a.s. bounded random variables. However, such approach would require the number of time periods (i.e., the length of the panel) to be sufficiently large (i.e., $\log N/T = o((NT)^{-1/2})$). We avoid assuming these additional unpleasant conditions by using the NLO approach.

where each block has length $T_k \geq T_{\text{block}} := \lfloor T/(K-1) \rfloor$ for each k, such that $K \geq 3$. Let $\mathcal{N}(k)$ denote k and its immediate neighbors in $\{1, \ldots, K\}$. Define the quasi-complement of $\mathcal{M}_k$ as $\mathcal{M}_k^{\text{qc}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\} \setminus \{\mathcal{M}_l : l \in \mathcal{N}(k)\}$, and the corresponding data blocks $B_k = \{W_{\cdot,t} : t \in \mathcal{M}_k\}$ and $B_k^{\text{qc}} = \{W_{\cdot,t} : t \in \mathcal{M}_k^{\text{qc}}\}$.

The construction creates quasi-complementary sets with the left-out-neighbors. Since we use the quasi-complementary sets to fit the nonparametric nuisance functions, we recommend $K \geq 10$, to ensure that at least 70% of data is used for this task. To clarify the construction further, consider the following example: suppose we have $K = 10$ blocks $\{\mathcal{M}_k\}_{k=1}^{10}$ of adjacent time stamps $t \in \{1, \ldots, T\}$, each of size T_{block}, so that $T = KT_{\text{block}}$. Then the first quasi-complementary set $\mathcal{M}_1^{\text{qc}}$ consists of $\{\mathcal{M}_k\}_{k=3}^{10}$, the second set $\mathcal{M}_2^{\text{qc}}$ consists of $\{\mathcal{M}_k\}_{k=4}^{10}$, the third set $\mathcal{M}_3^{\text{qc}}$ consists of $\{\mathcal{M}_1\} \cup \mathcal{M}_{k=5}^{10}$, the fourth set $\mathcal{M}_4^{\text{qc}}$ consists of $\{\mathcal{M}_k\}_{k=1}^{2} \cup \mathcal{M}_{k=6}^{10}$, …, and the final set $\mathcal{M}_{10}^{\text{qc}}$ consists of $\{\mathcal{M}_k\}_{k=1}^{8}$.

The following is the application of the NLO cross-fitting method above in our context.

Algorithm 3 (NLO Cross-Fitted Residuals). (1) Construct blocks (B_k, B_k^{qc}) for $k = 1, \ldots, K$ using Definition 3.1; (2) For each k, compute estimators of reduced forms using quasi-complementary sets, namely,

$$\widehat{d}_{ik}(\cdot) = \widehat{d}_{ik}\big(\cdot, B_k^{\text{qc}}\big), \qquad \widehat{l}_{ik}(\cdot) = \widehat{l}_{ik}\big(\cdot, B_k^{\text{qc}}\big), \quad i = 1, 2, \ldots, N.$$

(3) Obtain the estimated residuals:

$$\widehat{\widetilde{Y}}_{it} := Y_{it} - \widehat{l}_{ik}(X_{it}), \qquad \widehat{V}_{it} := D_{it} - \widehat{d}_{ik}(X_{it}), \quad i = 1, 2, \ldots, N. \tag{3.1}$$

In the case of the base treatment structure of Example 2.1, the last step reduces to

$$\widehat{V}_{it}^P = P_{it} - \widehat{p}_{ik}(X_{it}), \quad \text{and} \quad \widehat{V}_{it} := K(X_{it})\widehat{V}_{it}^P, \quad i = 1, 2, \ldots, N,$$

since we first construct $\widehat{p}_{ik}(X_{it})$ and then set $\widehat{d}_{ik}(X_{it}) = K(X_{it})\widehat{p}_{ik}(X_{it})$.

3.2 *Theoretical support for the NLO cross-fitting method*

To explain the benefits of the construction, we define some notation. Suppose X and Y are random elements on the same Polish space. Define their dependence coefficient (the beta-mixing coefficient) as

$$\gamma(X, Y) = \frac{1}{2}\|P_{X,Y} - P_X \times P_Y\|_{\text{TV}},$$

where P_V denotes the distribution of the random element V. The dependence coefficient vanishes if and only if X and Y are independent.

We also make use of the following coupling result of Strassen (1965) for underlying spaces being Polish:

$$\min\big\{\mathrm{P}(X \neq Y) : X \sim P_X, Y \sim P_Y\big\} = \frac{1}{2}\|P_X - P_Y\|_{\text{TV}}. \tag{3.2}$$

(Note that the problem above is the optimal transportation problem for 0–1 cost; see Villani (2007) for discussion).

The following result follows from the application of Strassen's coupling (3.2) and Lemma 2.11 of Dudley and Philipp (1983).[7]

LEMMA 3.1 (Independent Coupling for NLO Data Blocks via Strassen). *By suitably enriching probability space, we can construct* $\widetilde{B}_k$ *and* $\widetilde{B}_k^{\mathrm{qc}}$ *that are independent of each other and that have the same marginal distributions as* B_k *and* B_k^{qc} *such that*

$$\mathrm{P}\{(B_k, B_k^{\mathrm{qc}}) \neq (\tilde{B}_k, \tilde{B}_k^{\mathrm{qc}})\} = \frac{1}{2}\|P_{B_k, B_k^{\mathrm{qc}}} - P_{B_k} \times P_{B_k^{\mathrm{qc}}}\|_{\mathrm{TV}} =: \gamma(B_k, B_k^{\mathrm{qc}}),$$

where $P_{B_k, B_k^{\mathrm{qc}}}$ *is the distribution of* (B_k, B_k^{qc}) *and* $P_{B_k} \times P_{B_k^{\mathrm{qc}}}$ *is the distribution of* $(\tilde{B}_k, \tilde{B}_k^{\mathrm{qc}})$.

If the data sequence $(W_{\cdot,t} : t \geq 1)$ is beta-mixing in t, we have that $\gamma(B_k, B_k^{\mathrm{qc}}) \to 0$, since the blocks are separated by $\lfloor T/(K-1) \rfloor \to \infty$ periods as $T \to \infty$. Thus, under beta-mixing, by using the NLO-cross-fitting, we can replace each block and its quasi-complement with independent blocks, with the probabilistic error determined by the speed of mixing of the weakly dependent time series. Since we generally obtain the nuisance parameter estimates using quasi-complements, NLO-CF allows us to treat these estimates as if (essentially) independent from the data used to compute semi-parametric scores (residuals in our context).

3.3 *First-stage estimators for learning residuals in panel data*

In this section, we give examples of the first-stage reduced form estimators for dynamic panel data, focusing on the models with base treatment structure (2.10). We rely heavily on the results of Kock and Tang (2019) in this stage.

EXAMPLE 3.1 (First-Stage Treatment Lasso and Reduced Form). Consider the model (2.11) with a single base treatment. Suppose

$$p_{i0}(X_{it}) = X_{it}'\delta_0^P + \xi_i. \tag{3.3}$$

Here, to save notation, we reassign X_{it} to denote the dictionary of transformations of original controls X_{it}, that is, $X_{it} \leftarrow \mathrm{t}(X_{it})$, where the map $\mathrm{t}(\cdot)$ generates the dictionary and $X_{it} \in \mathrm{R}^{d_X}$.

[7] Note that Strassen's couping also underlies the Berbee coupling, Berbee (1987), for real-valued random variables. We extend Berbee coupling to random vectors or, more general, random variables taking values in complete, separable metric spaces in the Appendix, and then use it to obtain concentration results.

For the *first-stage treatment* penalty level $\lambda_P = C_P\sqrt{NT\log^3(d_X+N)}$ for some constant C_P, define the k-fold specific estimator:

$$\left(\widehat{\delta}_k^P, \widehat{\xi}_k\right) = \arg\min_{\delta^P,\xi} \sum_{i=1}^{N} \sum_{t\in\mathcal{M}_k^{\mathrm{qc}}} \left(P_{it} - X_{it}'\delta^P - \xi_i\right)^2 + 2\lambda_P\left\|\delta^P\right\|_1 + 2\frac{\lambda_P}{\sqrt{N}}\|\xi\|_1, \quad k=1,\ldots,K. \tag{3.4}$$

(Note that here and below the subscript index k in $\hat{\xi}_k$ serves to indicate the k-specific estimator of the vector of fixed effects ξ, which is not to be confused with the index i that enumerates the elements of the vector ξ.)

Then, for any $t\in\mathcal{M}_k$ and any $i=1,2,\ldots,N$, the base treatment reduced form estimate is

$$\widehat{p}_{ik}(X_{it}) = X_{it}'\widehat{\delta}_k^P + \widehat{\xi}_{i,k}, \quad k=1,\ldots,K.$$

The properties of this estimator under weak sparsity assumptions on δ^P and ξ follow from Kock and Tang (2019).

Example 3.2 (First-Stage Outcome Lasso and Reduced Form). Consider the outcome model:

$$Y_{it} = D_{it}'\beta_0 + X_{it}'\delta_0^P + \xi_i^E + U_{it}. \tag{3.5}$$

Here, to save notation, we reassign X_{it} to denote the dictionary of transformations of original controls X_{it}, that is, $X_{it} \leftarrow \mathsf{t}(X_{it})$, where the map $\mathsf{t}(\cdot)$ generates the dictionary.

For the *first-stage outcome* penalty level $\lambda_E = C_E\sqrt{NT\log^3(d_X+N)}$ for some constant C_E, define the k-fold specific estimator:

$$\left(\check{\beta}_k, \widehat{\delta}_k^E, \widehat{\xi}_k^E\right) = \arg\min_{\beta,\delta^E,\xi^E} \sum_{i=1}^{N} \sum_{t\in\mathcal{M}_k^{\mathrm{qc}}} \left(Y_{it} - D_{it}'\beta - X_{it}'\delta^E - \xi_i^E\right)^2 + 2\lambda_E\left\|(\beta,\delta^E)\right\|_1 + 2\frac{\lambda_E}{\sqrt{N}}\left\|\xi^E\right\|_1. \tag{3.6}$$

Then, for any $t\in\mathcal{M}_k$, the outcome reduced form estimate is

$$\widehat{l}_{ik}(X_{it}) = \widehat{d}_{ik}(X_{it})'\check{\beta}_k + X_{it}'\widehat{\delta}_k^E + \widehat{\xi}_{ik}^E, \tag{3.7}$$

where $\widehat{d}_{ik}(\cdot)$ is the treatment reduced form estimate. In what follows, we refer to $\check{\beta}_k$ as the *preliminary*, or the *one-stage* estimator of β_0.

Lemma 5.2 establishes properties of this estimator under weak sparsity assumptions on δ^E and ξ^E, based upon Kock and Tang's (2019) analysis of dynamic panel data lasso with weakly sparse unit effects.

3.4 *The second stage: Estimating CATE functions*

Here, we describe the second stage estimators.

When D_{it} is low-dimensional, we can apply ordinary least squares to residuals.

DEFINITION 3.2 (Orthogonal Least Squares). Define

$$\widehat{\beta}_{\text{OLS}} := \arg\min_{\beta \in \mathrm{R}^d} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widehat{\widetilde{Y}}_{it} - \widehat{V}_{it}'\beta\right)^2. \tag{3.8}$$

Appendix G in the Online Supplement establishes estimation and inference results for orthogonal least squares. The rate of convergence is $\sqrt{d/NT}$.

DEFINITION 3.3 (Orthogonal Lasso). Let $\lambda_\beta = C_\beta \sqrt{\log d/NT}$ and C_β be a penalty parameter. Define

$$\widehat{\beta}_L := \arg\min_{\beta \in \mathrm{R}^d} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widehat{\widetilde{Y}}_{it} - \widehat{V}_{it}'\beta\right)^2 + \lambda_\beta \sum_{j=1}^{d} |\beta_j|. \tag{3.9}$$

Theorem 4.1 provides the near-oracle rates of convergence

$$\sqrt{s \log d/NT}$$

for the CATE function. We notice that the orthogonal lasso outperforms orthogonal least squares even in low-dimensional settings when $s \log d \ll d \ll NT$, that is, when effective dimension s of β_0 is much smaller than its nominal dimension d.

REMARK 3.1 (Key Point of Orthogonalization). By working with estimated residuals, we attain the quasi-oracle rates of convergence—the rates that result if we knew the true residuals exactly and used them instead. As a result, the orthogonal lasso also outperforms the single-stage outcome regression estimators when the CATE function is much simpler and, therefore, easier to learn the overall regression function. For example, the orthogonal lasso outperforms the first-stage outcome lasso when the CATE function is more sparse than the overall regression function, so the near-oracle rate is much better than the overall rate. Remark 5.7 below provides a formal statement. The phenomenon our paper points out is more general and is not specific to using lasso methods used in the final stage. More recent works than our paper use orthogonalization procedures like ours to learn the CATE functions for other choices of the final-stage estimator; see, for example, Kennedy (2020).

REMARK 3.2 (Data-Adaptive Penalty Levels). We choose the penalty level for first and stage estimators of the stated simple form above to simplify theoretical arguments. Rigorous and data-adaptive choices of penalty levels λ's, in particular of the constants C_P and C_P, as well as the generalization of ℓ_1-penalty to its weighted analog, are discussed

in, for example, Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017) for the cross-sectional case, and implemented in the hdm *R* package by Chernozhukov, Hansen, and Spindler (2016). Their choices likely carry over to the dynamic panel data settings under the conditional sequential exogeneity condition.

3.5 *The third stage: Debiased inference on parameters of CATE functions*

Here, we describe the third stage that performs debiased inference. Due to the bias induced by ℓ_1-shrinkage, penalized estimators cannot be used for inference based on the standard Gaussian approximation. We construct a debiased estimator based on a variant of van der Geer et al. (2014) and Zhang and Zhang (2014) with a new choice of the debiasing matrix.

Consider the covariance matrix of residuals

$$Q = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathrm{E} V_{it} V_{it}' \tag{3.10}$$

and its inverse Q^{-1}. Define the sample covariance matrix of the residuals as

$$\widehat{Q} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{V}_{it}\widehat{V}_{it}'. \tag{3.11}$$

Estimate approximate inverse of $\widehat{Q}$ by

$$\widehat{\Omega} = \arg\min_{\Omega\in\mathrm{R}^{d\times d}} \|\Omega\|_1 : \|\widehat{Q}\Omega - I_d\|_\infty \le \lambda_Q, \tag{3.12}$$

where

$$\lambda_Q := C_Q \kappa_{NT}, \qquad \kappa_{NT} := \sqrt{\log^3\bigl(d^2\log(NT)\bigr)\log NT/NT}, \tag{3.13}$$

where C_Q is a tuning constant. Finally, symmetrize the approximate inverse $\widehat{\Omega}$ as

$$\widehat{\Omega}^{\mathrm{CLIME}} = \bigl(\widehat{\omega}_{ij}^{\mathrm{CLIME}}\bigr), \qquad \widehat{\omega}_{ij}^{\mathrm{CLIME}} = \widehat{\omega}_{ij}1_{\{|\widehat{\omega}_{ij}|<|\widehat{\omega}_{ji}|\}} + \widehat{\omega}_{ji}1_{\{|\widehat{\omega}_{ij}|>|\widehat{\omega}_{ji}|\}}. \tag{3.14}$$

In other words, between $\widehat{\omega}_{ij}$ and $\widehat{\omega}_{ji}$, we take the one with smaller absolute value to obtain a symmetric matrix $\widehat{\Omega}^{\mathrm{CLIME}}$, as in Cai, Liu, and Luo (2011).

Definition 3.4 (Debiased Orthogonal Lasso). Define

$$\widehat{\beta}_{\mathrm{DL}} := \widehat{\beta}_L + \widehat{\Omega}^{\mathrm{CLIME}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{V}_{it}\bigl(\widehat{\widetilde{Y}}_{it} - \widehat{V}_{it}'\widehat{\beta}_L\bigr). \tag{3.15}$$

Theorems 4.2 and 4.3 show that $\sqrt{NT}(\widehat{\beta}_{\mathrm{DL}} - \beta_0)$ is approximately distributed as $N(0, \Sigma)$ over rectangular regions. The covariance matrix

$$\Sigma := Q^{-1}\Gamma Q^{-1} = Q^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}V_{it}V_{it}'U_{it}^2Q^{-1} \tag{3.16}$$

is estimated by its sample analog

$$\widehat{\Sigma}(\widehat{\beta}_L) := \widehat{Q}^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{V}_{it}\widehat{V}_{it}'\big(\widehat{\widetilde{Y}}_{it} - \widehat{V}_{it}'\widehat{\beta}_L\big)^2\widehat{Q}^{-1} =: \widehat{Q}^{-1}\widehat{\Gamma}(\widehat{\beta}_L)\widehat{Q}^{-1}. \tag{3.17}$$

This method allows for constructing componentwise and simultaneous confidence intervals for all components of β_0. This method also allows for performing inference on linear functionals $a'\beta_0$ of β_0 (provided that the ℓ_1-norm of a is bounded).

4. Theoretical results on orthogonal lasso

4.1 *Consistency of orthogonal lasso*

The following assumptions impose regularity conditions on weak dependence, tail behavior, and the reduced form estimators.

Assumption 4.1 (Sampling and Asymptotics). (1) *The data sequence* $\{\{W_{it}\}_{t=1}^{T}\}_{i=1}^{N}$ *obeys the model* (2.4) *of Section* 2. (2) *The data on units* $W_{i,\cdot}$ *are independent across* i*, and beta-mixing at geometric speed with respect to time* t*, uniformly in* i:

$$\gamma(q) := \sup_{\bar{t}\le T, i\le N}\gamma\big(\{W_{it}\}_{t\le\bar{t}}, \{W_{it}\}_{t\ge\bar{t}+q}\big) \le C_\kappa\exp(-\kappa q) \tag{4.1}$$

for all $q \ge 1$*, and for some constants* $C_\kappa \ge 0$ *and* $\kappa > 0$. (3) *The number of time periods* T *is large enough,* $T^{-1}\log(N) = o(1)$.

Assumption 4.1 limits the data dependence across time periods with the exponential mixing step. It is a standard weak dependence condition in the literature (Hahn and Kuersteiner (2011), Fernandez-Val and Lee (2013)). We incur it to ensure the validity of inference based on the panel cross-fitting of Definition 3.1. Note that the requirement on N comes from the relation

$$\gamma\big(\{W_{\cdot,t}\}_{t\le\bar{t}}, \{W_{\cdot,t}\}_{t\ge\bar{t}+q}\big) \le N\max_{i\le N}\gamma\big(\{W_{i,t}\}_{t\le\bar{t}}, \{W_{i,t}\}_{t\ge\bar{t}+q}\big), \tag{4.2}$$

which can be found using the union bound.

The next condition ensures identification of the coefficients of the CATE function.

Assumption 4.2 (Identification). *Let* $Q = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}V_{it}V_{it}^{\top}$ *denote the population covariance matrix of treatment residuals. Assume that there exist constants* $C_{\min}$, $C_{\max}$ *such that* $0 < C_{\min} \le \min\mathrm{eig}(Q) \le \max\mathrm{eig}(Q) \le C_{\max} < \infty$.

A collection of centered random variables $\{X_j\} \in \mathrm{R}$ is said to be uniformly σ^2-sub-Gaussian if

$$\mathrm{E}\exp(\lambda X_j) \leq \exp\bigl(\lambda^2\sigma^2/2\bigr), \quad \forall \lambda \in \mathrm{R}\, \forall j. \tag{4.3}$$

Assumption 4.3 (Sub-Gaussian Tails). *The following conditions hold for some constants* $0 < \underline{\sigma}^2 < \bar{\sigma}^2 < \infty$. (1) *For* $j = 1, 2, \ldots, d$, $(V_{it})_j$ *are* $\bar{\sigma}^2$*-sub-Gaussian conditional on* X_{it}, Φ_{it}. (2) U_{it} *is* $\bar{\sigma}^2$*-sub-Gaussian conditional on* V_{it}, X_{it}, Φ_{it}. (3) U_{it} *is conditionally non-degenerate, namely,* $\inf_{it} \mathbb{E}[U_{it}^2|V_{it}, X_{it}, \Phi_{it}] \geq \underline{\sigma}^2$ *with probability* 1.

Assumption 4.4 (Additional Regularity Conditions). *We suppose that the true parameter vector has bounded* ℓ_1*-norm*: (*a*)

$$\|\beta_0\|_1 \leq \bar{C}_\beta$$

for some finite constant $\bar{C}_\beta$; (*b*) *and that the number of nonzero coefficients does not increase too quickly*:

$$(s \vee 1)\kappa_{NT} = (s \vee 1)\sqrt{\log^3\bigl(d^2 \log(NT)\bigr) \log NT/NT} = o(1).$$

(*c*) *The tuning constants* C_β *and* C_Q *in the penalty levels* λ_β *and* λ_Q *are sufficiently large.* (*d*) *The number* $d \to \infty$.

Let $1 \leq i \leq N$ be a unit index, and let $1 \leq j \leq d_P$ be a component index. We define a generic nuisance function to be

$$\mathbf{g}(\cdot) = \bigl\{g_{ij}(\cdot)\bigr\} : \mathcal{X} \to \mathrm{R}^{N \times d_P},$$

a generic $N \times d_P$-matrix, and we let

$$\mathbf{g}_0(\cdot) = \bigl\{g_{ij0}(\cdot)\bigr\} : \mathcal{X} \to \mathrm{R}^{N \times d_P}$$

be its true value; here, $\mathcal{X}$ is a subset of $\mathrm{R}^{N \times d_X}$. Let G_{NT} be a sequence of neighborhoods around $\mathbf{g}_0(\cdot)$ containing realizations of its generic machine learning estimators $\widehat{\mathbf{g}}(\cdot)$ with probability approaching 1. As the sample size NT increases, we expect the sets G_{NT} to converge toward $\mathbf{g}_0$ in suitable norms. We denote rate of convergence in the mean square norm as

$$\mathbf{g}_{NT} := \max_{1 \leq j \leq d_P} \sup_{\mathbf{g} \in G_{NT}} \left((NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathrm{E}\bigl(g_{ij}(X_{it}) - g_{ij0}(X_{it})\bigr)^2 \right)^{1/2} \tag{4.4}$$

and let $\mathbf{g}_{NT,\infty}$ be a sequence of nonnegative constants such that with probability $1 - o(1)$:

$$\max_{1 \leq j \leq d_P} \sup_{\mathbf{g} \in G_{NT}} \sup_{it} \bigl|g_{ij}(X_{it}) - g_{ij0}(X_{it})\bigr| \leq \mathbf{g}_{NT,\infty}. \tag{4.5}$$

In particular, we specialize the notion as follows:

1. For the technical treatment reduced form, we replace the letters g with d:
   - $\mathbf{d}(\cdot)$ denotes the parameter and $\mathbf{d}_0(\cdot)$ the true reduced form for treatment;
   - D_{NT} denotes the set containing first-stage estimates $\widehat{\mathbf{d}}(\cdot)$ of $\mathbf{d}_0(\cdot)$ w.p. $1-o(1)$
   - $\mathbf{d}_{NT}$ and $\mathbf{d}_{NT,\infty}$ are rates of convergence of D_{NT} to $\mathbf{d}_0(\cdot)$;
2. For the outcome reduced form, we replace the letters g with l:
   - $\mathbf{l}(\cdot)$ denotes the parameter and $\mathbf{l}_0(\cdot)$ the true value reduced form for outcome;
   - L_{NT} denotes the set containing first-stage estimates $\widehat{\mathbf{l}}(\cdot)$ of $\mathbf{l}_0(\cdot)$ w.p. $1-o(1)$
   - $\mathbf{l}_{NT}$ and $\mathbf{l}_{NT,\infty}$ are rates of convergence of L_{NT} to $\mathbf{l}_0(\cdot)$.

The key assumption on the quality of reduced form estimators is as follows.

ASSUMPTION 4.5 (Regularity Conditions and Convergence Rates for Residual Learners). *We suppose that the reduced form estimators obey*: $\widehat{\mathbf{l}}(\cdot) \in L_{NT}$ *and* $\widehat{\mathbf{d}}(\cdot) \in D_{NT}$ *such that* $\mathbf{d}_{NT}, \mathbf{d}_{NT,\infty}, \mathbf{l}_{NT}, \mathbf{l}_{NT,\infty}$ *decay sufficiently fast*:

$$\mathbf{d}_{NT}^2 + \mathbf{d}_{NT}\mathbf{l}_{NT} = o\big((NT)^{-1/2}\big), \tag{4.6}$$

$$\mathbf{l}_{NT,\infty} = o\big(\log^{-1/2}(dNT)\big), \qquad \mathbf{d}_{NT,\infty} = o\big(\log^{-1/2}(dNT)\big). \tag{4.7}$$

As we discuss in Section 5, this assumption is plausible for the lasso-based first-stage estimators that we consider and may be plausible for others, too. The condition of bounded ℓ_1-norm can be relaxed to allow for an increasing norm at the cost of somewhat more complicated regularity conditions, as can be seen from the proofs.

Theorem 4.1 establishes the convergence rate of orthogonal lasso in ℓ_2 and ℓ_1 norm.

THEOREM 4.1 (Oracle Rates for Orthogonal Lasso). *Suppose Assumptions 4.1–4.5 hold. Then the orthogonal lasso possesses the following oracle rate guarantees*:

$$\big(\mathbb{E}_{NT}\big(V_{it}'(\widehat{\beta}_L - \beta_0)\big)^2\big)^{1/2} = O_P\left(\sqrt{\frac{s\log d}{NT}}\right), \qquad \|\widehat{\beta}_L - \beta_0\|_1 = O_P\left(\sqrt{\frac{s^2\log d}{NT}}\right). \tag{4.8}$$

Theorem 4.1 is our first main result. It establishes the convergence rate of orthogonal lasso. This rate coincides with the oracle convergence rate, where oracle knows the first-stage function $e_0(\cdot)$ and the unobserved unit effects $\{\xi_i^E\}_{i=1}^N$ in the model (2.4) and, therefore, knows the residuals and uses lasso on these true residuals.

4.2 *Estimation of* Q^{-1} *and* Σ *in high-dimensional setting*

To perform statistical inference, we will need to assume approximate sparsity for the inverse Q^{-1} of the covariance matrix of the residuals Q. We will use the following notation.

For a square matrix $A = (a_{ij})$, denote

$$\|A\|_{1,\infty} = \max_{1\le j\le d}\sum_{i=1}^{d}|a_{ij}|, \qquad \|A\|_{\infty,1} = \max_{1\le i\le d}\sum_{j=1}^{d}|a_{ij}|.$$

ASSUMPTION 4.6 (Regularity for Estimating Q^{-1}). (*a*) *Let A_Q and $a_Q > 1$ be finite constants such that for any column j,*

$$\big(Q^{-1}_{mj}\big)^* \le A_Q m^{-a_Q}, \quad m, j = 1, 2, \ldots, d,$$

where $(Q_j^{-1})^*$ is a nonincreasing rearrangement of $(|Q^{-1}_{mj}|)_{m=1}^d$. Furthermore, for $\lambda_Q = C_Q\kappa_{NT}$,

$$\lambda_Q^{1-1/a_Q} = o\big(s^{-1}\log^{-1/2} d\big). \tag{4.9}$$

Assumption 4.6(a) ensures that the CLIME estimator of Q^{-1} defined in the equation (3.14) converges sufficiently fast. If $d \gg NT$, it requires Q^{-1} to be approximately sparse so that it can be consistently estimated with only NT observations. Examples of high-dimensional matrices Q with an approximately sparse inverse include block diagonal, Toeplitz, and band matrices.

The following lemma establishes the rate bound for the CLIME estimator of Q^{-1} for $\widehat{Q}$ in (3.11). The result holds under mixing dependence and approximate sparsity of Q, which may be of independent interest.

LEMMA 4.1 (Consistency of the CLIME Estimator). *Suppose Assumptions 4.1–4.6 hold. The CLIME estimator converges in ℓ_∞-norm and $\ell_{\infty,1}$-norm,*

$$\big\|\widehat{\Omega}^{\mathrm{CLIME}} - Q^{-1}\big\|_\infty = \big\|\widehat{\Omega} - Q^{-1}\big\|_\infty = O_P(\lambda_Q), \tag{4.10}$$

$$\big\|\widehat{\Omega}^{\mathrm{CLIME}} - Q^{-1}\big\|_{1,\infty} = \big\|\widehat{\Omega}^{\mathrm{CLIME}} - Q^{-1}\big\|_{\infty,1} = O_P\big(\lambda_Q{}^{1-1/a_Q}\big), \tag{4.11}$$

$$\big\|I_d - \widehat{\Omega}^{\mathrm{CLIME}}\widehat{Q}\big\|_\infty = \big\|I_d - \widehat{Q}\widehat{\Omega}^{\mathrm{CLIME}}\big\|_\infty = O_P\big(\lambda_Q{}^{1-1/a_Q}\big). \tag{4.12}$$

4.3 *Pointwise Gaussian inference with debiased orthogonal lasso*

The following theorem establishes validity of Gaussian inference for parameters $\alpha'\beta_0$, where α is a fixed vector with bounded ℓ_1 norm. This is our second main result.

THEOREM 4.2. *Let K_α be a finite constant. Suppose Assumptions 4.1–4.6 hold, and the Lindeberg condition holds for each $m > 0$:*

$$\limsup_{NT\to\infty}\sup_{\|\alpha\|_2=1,\,\|\alpha\|_1\le K_\alpha}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}[\big(\alpha' V_{it}U_{it}\big)^2 1\big\{\big|\alpha' V_{it}U_{it}\big| > m\sqrt{NT}\big\}] = 0.$$

Then the debiased lasso estimator is asymptotically Gaussian:

$$\lim_{NT\to\infty}\sup_{\|\alpha\|_2=1,\,\|\alpha\|_1\le K_\alpha}\sup_{t\in\mathrm{R}}\left|\mathrm{P}\left(\frac{\sqrt{NT}\alpha'(\widehat{\beta}_{\mathrm{DL}} - \beta_0)}{\sqrt{\alpha'\Sigma\alpha}} < t\right) - \Phi(t)\right| = 0,$$

where $\Phi(t)$ *is the CDF of* $N(0,1)$. *Moreover, the result continues to hold when* Σ *is replaced by* $\widehat{\Sigma}$ *such that* $\|\widehat{\Sigma}-\Sigma\|_\infty = o_P(1)$.

The following lemma establishes consistency of the high-dimensional covariance matrix for the approximate Gaussian distribution of the debiased lasso estimator. Notably, the dimension of this matrix can exceed the sample size. Define

$$\gamma_{NT} := (NT)^{-1/4} + \mathbf{l}_{NT} + \sqrt{s\log d/NT} + \mathbf{l}_{NT}^2 \log\big(d^2 NT\big). \tag{4.13}$$

Assumption 4.7 (Conditions for Σ Estimation). *We suppose that the following condition on the growth of the dimension* d *holds*:

$$\kappa_{NT}\log^2\big(d^2NT\big) = o(1). \tag{4.14}$$

and $\gamma_{NT} = o(1)$.

Lemma 4.2 (Consistency of Variance Matrix Estimator). *Suppose Assumptions 4.1–4.7 hold. Then the estimator* $\widehat{\Sigma}(\widehat{\beta}_L)$ *converges in* ℓ_∞ *norm,*

$$\big\|\widehat{\Sigma}(\widehat{\beta}_L) - \Sigma\big\|_\infty = O_P\big(\gamma_{NT} + \lambda_Q^{1-1/a_Q}\big) =: O_P(\zeta_{NT}) = o_P(1). \tag{4.15}$$

4.4 *Simultaneous inference*

We next present theoretical results on simultaneous inference on many structural coefficients.

Define the following rates:

$$\rho_{NT} := \sqrt{\log(dNT)/NT}(\mathbf{d}_{NT,\infty} + \mathbf{l}_{NT,\infty}) + r_{NT}, \tag{4.16}$$

$$r_{NT} := \mathbf{d}_{NT}^2 + \mathbf{d}_{NT}\mathbf{l}_{NT} + \big(\mathbf{d}_{NT,\infty}^2 + \mathbf{d}_{NT,\infty}\mathbf{l}_{NT,\infty}\big)\sqrt{(NT)^{-1}\log(NT)\log d}. \tag{4.17}$$

Assumption 4.8 (Regularity Conditions for Simultaneous Inference on Many Coefficients). (1) *There exists a sequence* $\pi_{NT}^{UV} \geq 1$ *so that* $\sup_{it}\|V_{it}U_{it}\|_\infty \leq \pi_{NT}^{UV}$ *a.s. and*

$$0 < \min_{it}\big\|\mathrm{E}V_{it}V_{it}'\big\|_\infty < \infty.$$

(2) *For some constant* $c_2 : 0 < c_2 < 1/4$, *the following rate conditions hold*:

$$\pi_{NT}^{UV}\log d\log(NT)\log^{7/2}(dNT) \lesssim (NT)^{1/2-2c_2} \tag{4.18}$$

and $\log^4 d\log^2(NT) = o(\sqrt{NT})$. (3) *The following rate conditions hold*:

$$\sqrt{NT}\rho_{NT} + \lambda_Q^{1-1/a_Q} s\log^{1/2} d = o\big(\log^{-1/2} d\log^{-1/2} NT\big). \tag{4.19}$$

Theorem 4.3 establishes high-dimensional Gaussian approximation for a treatment effect vector β_0 and allows to conduct simultaneous inference on its coefficients.

THEOREM 4.3 (Simultaneous Inference on Many Coefficients). *Suppose Assumptions 4.1–4.8 with Σ as in (3.16) and $\widehat{\Sigma}$ as in (3.17). Then the following Gaussian approximation result holds for $\widehat{\beta}_{\mathrm{DL}}$:*

$$\sup_{R\in\mathcal{R}}\left|\mathrm{P}\big((\operatorname{diag}\Sigma)^{-1/2}\sqrt{NT}(\widehat{\beta}_{\mathrm{DL}}-\beta_0)\in R\big)-\mathrm{P}(Z\in R)\right|\to 0, \tag{4.20}$$

where $Z\sim N(0,\mathcal{C})$ is a centered Gaussian random vector with the covariance matrix

$$\mathcal{C}=(\operatorname{diag}\Sigma)^{-1/2}\Sigma(\operatorname{diag}\Sigma)^{-1/2}$$

and $\mathcal{R}$ denotes a collection of cubes in $\mathbb{R}^d$ centered at the origin. In addition, if

$$\gamma_{NT}+\lambda_Q^{1-1/a_Q}=o\big(\log^{-2}d\log^{-1}NT\big), \tag{4.21}$$

then, replacing $\mathcal{C}$ with $\widehat{\mathcal{C}}=(\operatorname{diag}\widehat{\Sigma})^{-1/2}\widehat{\Sigma}(\operatorname{diag}\widehat{\Sigma})^{-1/2}$, we also have for $\widehat{Z}\mid\widehat{\mathcal{C}}\sim N(0,\widehat{\mathcal{C}})$,

$$\sup_{R\in\mathcal{R}}\left|\mathrm{P}\big((\operatorname{diag}\widehat{\Sigma})^{-1/2}\sqrt{NT}(\widehat{\beta}_{\mathrm{DL}}-\beta_0)\in R\big)-\mathrm{P}(\widehat{Z}\in R\mid\widehat{\mathcal{C}})\right|\to_P 0. \tag{4.22}$$

Consequently, for the $c_{1-\xi}=(1-\xi)$-quantile of $\|\widehat{Z}\|_\infty\mid\widehat{\mathcal{C}}$, we have

$$\mathrm{P}\big(\beta_{0,j}\in\big[\widehat{\beta}_{\mathrm{DL},j}\pm c_{1-\xi}\widehat{\Sigma}_{jj}^{1/2}(NT)^{-1/2}\big],\ j=1,2,\ldots,d\big)\to(1-\xi).$$

Theorem 4.3 is our third main result. It extends the high-dimensional Gaussian approximations of Chernozhukov, Chetverikov, and Kato (2013), Zhang and Wu (2017), Chernozhukov, Chetverikov, and Kato (2019) to a panel setting.

4.5 *Orthogonal group lasso*

In this section, we focus on Example 2.1 with a linear control function $e_0(\cdot)$ in (2.4). Applied economists fitting (2.4) often would like to include the variable $(K_{it})_j$ whenever the interaction of the base treatment P_{it} and the control $(K_{it})_j$ is selected in the second stage. However, in the case of the orthogonal lasso, the sets of controls selected in the stage 2 may not be a subset of the controls selected in the stage 1. To address this concern, we *group* the main and interaction effects of controls K_{it} to attain the desired model selection pattern.

Decompose the covariate vector

$$X_{it}=(K_{it},Z_{it}),$$

where K_{it} is a vector of heterogeneity-relevant controls and Z_{it} is its complement. The function can be written as

$$e_0(X_{it})=K_{it}'\rho_0+Z_{it}'\delta_{0Z}^E$$

and the linear model (2.4),

$$Y_{it}=\underbrace{(P_{it}K_{it},K_{it})'}_{D_{it}}\underbrace{(\beta_0,\rho_0)}_{\tilde{\beta}_0}+Z_{it}'\delta_{0Z}^E+\xi_i^E+U_{it}. \tag{4.23}$$

Assuming both the interaction effect β_0 and the main effect ρ_0 are s-sparse, the vector $\bar{\beta}_0$ obeys *group sparsity assumption*,

$$\|\beta_0, \rho_0\|_{2,0} := \sum_{j=1}^{d} 1\{(\beta_{0j}, \rho_{0j}) \neq (0,0)\} \leq \sum_{j=1}^{d} 1\{\beta_{0j} \neq 0\} + \sum_{j=1}^{d} 1\{\rho_{0j} \neq 0\} = 2s \ll d.$$

The unit-specific treatment reduced form is

$$D_{it} = d_{i0}(Z_{it}) = d_0(Z_{it}; \xi_i)$$

and the unit-specific outcome reduced form is

$$\mathrm{E}[Y_{it} \mid Z_{it}] = d_{i0}(Z_{it})'\bar{\beta}_0 + Z_{it}'\delta_{0Z}^E + \xi_i^E.$$

The residualized form is

$$\ddot{\bar{Y}}_{it} = \bar{V}_{it}'\bar{\beta}_0 + U_{it},$$

where

$$\ddot{\bar{Y}}_{it} := Y_{it} - \mathrm{E}[Y_{it} \mid Z_{it}], \qquad \bar{V}_{it} := D_{it} - \mathrm{E}[D_{it} \mid Z_{it}] =: D_{it} - d_{i0}(Z_{it}). \tag{4.24}$$

The *orthogonal group lasso* estimator is the first component $\widehat{\beta}_{\mathrm{GL}}$ of the following minimization problem:

$$\widehat{\beta}_{\mathrm{GL}} := \arg \min_{\bar{\beta}=(\beta,\rho)\in \mathrm{R}^{2d}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(\ddot{\bar{Y}}_{it} - \widehat{\bar{V}}_{it}'\bar{\beta}\big)^2 + \lambda_\beta \sum_{j=1}^{d} \big\|(\beta_j, \rho_j)\big\|_2. \tag{4.25}$$

The *debiased orthogonal group lasso* estimator is

$$\widehat{\beta}_{\mathrm{DGL}} := \widehat{\beta}_{\mathrm{GL}} + \widehat{\Omega}^{\mathrm{CLIME}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{V}_{it}\big(\ddot{\bar{Y}}_{it} - \widehat{V}_{it}'\widehat{\beta}_{\mathrm{GL}}\big). \tag{4.26}$$

Lemma 4.3 (Orthogonal Group Lasso). *Under Assumptions 4.1–4.5 for $\bar{V}_{it}$ and $d_{i0}(Z_{it})$ as in (4.24), the orthogonal group lasso attains the following rate*:

$$\big(\mathbb{E}_{NT}\big(\bar{V}_{it}'(\widehat{\beta}_{\mathrm{GL}} - \beta_0)\big)^2\big)^{1/2} = O_P\left(\sqrt{\frac{s \log d}{NT}}\right), \qquad \|\widehat{\beta}_{\mathrm{GL}} - \beta_0\|_1 = O_P\left(\sqrt{\frac{s^2 \log d}{NT}}\right). \tag{4.27}$$

Furthermore, the statements of Theorems 4.2 and 4.3 hold for the debiased orthogonal group lasso.

5. Verification of assumptions on the first-stage estimators of residuals

The purpose of this section is to verify Assumption 4.5 in i.i.d. and panel data settings. For the lasso-based methods of Examples 3.1–3.2, we give examples of nuisance parameter estimates, nuisance realization sets, and the low-level assumptions that suffice for Assumption 4.5 to hold. Unless proven immediately, all numbered remarks are formally proven in the Online Supplement, Appendix E.

5.1 *No unobserved unit heterogeneity: General ML*

Suppose the unit-specific vector function in (2.11) obeys $p_{i0}(\cdot) = p_{j0}(\cdot)$, $1 \le i, j \le N$. Let $p_0(\cdot) = p_{i0}(\cdot) = p_{j0}(\cdot)$ be the single coordinate of the N-vector $\mathbf{p}_0(\cdot)$ entering in (4.4). If the covariates X_{it} are i.i.d. over i and t, the mean square rate $\mathbf{p}_{NT}$ reduces to

$$\mathbf{p}_{NT} = \sup_{p \in P_{NT}} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(\mathrm{E}\big(p(X_{it}) - p_0(X_{it})\big)^2\big)^{1/2} = \sup_{p \in P_{NT}} \big(\mathrm{E}\big(p(X) - p_0(X)\big)^2\big)^{1/2}.$$

Furthermore, one can split by unit index to construct independent partitions and use regular cross-fitting instead of the NLO one. In this case, the condition (2) of Assumption 4.1 redundant.

The upper bound on $\mathbf{p}_{NT}$ are available for i.i.d. data (across time) for many regularized methods under structured assumptions on the functions $p_0(x)$ and $e_0(x)$, such as random forest, neural networks, or boosting. Specifically, the bound on $\mathbf{p}_{NT}$ is achievable by ℓ_1 penalized methods in sparse models (van der Geer et al. (2014), Belloni, Chernozhukov, and Wei (2016), Belloni and Chernozhukov (2013)), ℓ_2 boosting in sparse linear models (Luo and Spindler (2016)), neural networks (Schmidt-Hieber (2017), Farrell, Liang, and Misra (2021)), and random forest in small (Wager and Walther (2015)) and high (Syrganis and Zampetakis (2020)) dimensions with the sparsity structure. While most of these results are established in an i.i.d. setting, we conjecture that similar rates could be established under weak dependence, by relying on Berbee coupling, exponential mixing conditions, and/or the martingale-difference property of regression errors.

5.2 *Unobserved unit heterogeneity: Lasso*

In this section, we verify Assumption 4.5 for the first-stage lasso estimator.

We will make use of weak sparsity assumptions on fixed effects. Weak sparsity generalizes the exact sparsity restriction to accommodate small deviations from sparsity. Given a constant $\nu : 0 < \nu < 1$, the vector $u \in \mathrm{R}^{d_u}$ is said to be (ν, S)-weakly sparse (Negahban, Ravikumar, Wainwright, and Yu (2012)) if there exists a bound $S = S(N, T)$, that may depend on N, T, such that

$$\sum_{j=1}^{d_u} |u_j|^\nu \le S. \tag{5.1}$$

Lemma A.1 in Kock and Tang (2019) gives examples of distributions that generate weakly sparse draws with probability $1 - o(1)$. For one example, if ξ_i are independent Gaussian draws

$$\xi_i \sim N\big(0, \sigma_i^2\big), \qquad \max_{1 \le i \le N} \sigma_i^2 = O\big(\log^3(d_X + N)/\big(N^{1/\nu} T\big)\big), \quad i = 1, 2, \ldots, N. \tag{5.2}$$

Then the vector $\xi = (\xi_i)_{i=1}^N$ obeys (5.1) with probability $1 - o(1)$ with $S^P = \sqrt{N \log^{3\nu}(d_X + N)/T^\nu}$.

In what follows, we will rely on regressors and treatments being sub-Gaussian. In dynamic models where regressors include lagged values of outcomes and treatments, this assumption is nontrivial. We verify it from the model primitives in Remark 5.1 below.

REMARK 5.1 (Plausibility of Sub-Gaussian Assumption on Treatments and Outcomes and Their Lags). Consider Example 2.2. Substituting the treatment equation into the outcome equation gives

$$\begin{pmatrix} Y_{it} \\ P_{it} \end{pmatrix} = \sum_{l=1}^{L} A_{l,it} \begin{pmatrix} Y_{i,t-l} \\ P_{i,t-l} \end{pmatrix} + T_{it}, \tag{5.3}$$

where

$$A_{l,it} = \begin{bmatrix} \delta_{0l}^{EE} + K_{it}'\beta_0\delta_{0l}^{PE} & \delta_{0l}^{EP} + K_{it}'\beta_0\delta_{0l}^{PP} \\ \delta_{0l}^{PE} & \delta_{0l}^{PP} \end{bmatrix}, \tag{5.4}$$

$$T_{it} := \begin{bmatrix} \bar{X}_{it}'\bar{\delta}_0^E + \bar{M}_i'\delta_{M0}^E + \xi_i^E + U_{it} + K_{it}'\beta_0\big(\bar{X}_{it}'\bar{\delta}_0^P + \bar{M}_i'\delta_{M0}^P + \xi_i + V_{it}^p\big) \\ \bar{X}_{it}'\bar{\delta}_0^P + \bar{M}_i'\delta_{M0}^P + \xi_i + V_{it}^p \end{bmatrix}. \tag{5.5}$$

We can represent the reduced form as the canonical vector autoregression of order 1:

$$F_{it} = \Pi_{it}F_{i,t-1} + \varphi_{it}, \qquad F_{it} = \big[(Y_{i,t-l}, P_{i,t-l})_{l=0}^{L-1}\big]',$$

where

$$\underset{(2L\times 2L)}{\Pi_{it}} := \begin{bmatrix} A_{1,it} & A_{1,it} & \dots & A_{L-1,it} & A_{L,it} \\ I_2 & 0_2 & \dots & 0_2 & 0_2 \\ 0_2 & I_2 & 0_2 & \dots & 0_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0_2 & \dots & 0_2 & I_2 & 0_2 \end{bmatrix}, \qquad \underset{(2L\times 1)}{\varphi_{it}} := \begin{bmatrix} T_{it} \\ 0_2 \\ 0_2 \\ \vdots \\ 0_2 \end{bmatrix}.$$

We note that there are no restrictions on L here, but our conditions implicitly restrict $K_{it}'\beta_0$ to be bounded. Assume that the following conditions holds uniformly in (i, t): (1) The initial condition $F_{i,0}$ and T_{it}'s are $\bar{\sigma}^2$-sub-Gaussian vectors. (2) The singular values $\lambda(\Pi_{it})$ of Π_{it} obey $\|\lambda(\Pi_{it})\|_\infty \leq 1 - \delta$ for some constant $\delta > 0$. Then $\|F_{it}\|$ is $A\bar{\sigma}^2/(1-\delta)$- sub-Gaussian, for some numerical constant A.

Consider the following condition for learning the treatment reduced form:

(FS-TL) Consider the model

$$P_{it} = X_{it}'\delta_0^P + \xi_i + V_{it}^P,$$

for each i, the residuals V_{it}^P are a martingale difference sequence with respect to filtration Φ_{it}. Suppose (a) the vectors δ_0^P and ξ are (ν, S)-weakly sparse with $S = S^P$ and $S = N^{-\nu/2}S^P$, respectively, and $S^P = O(N^{1/2}\log^{3\nu/2}(d_X + N)T^{-\nu/2})$; $\|\delta_0^P\|_1$ is

bounded; (b) The Gram matrix

$$\Psi_X := (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\big[X_{it}X_{it}'\big]$$

has all of its eigenvalues bounded from above and below by $B_{\max}$ and $B_{\min}$, where $0 < B_{\min} \le B_{\max}$ are finite constants; (c) Each element of X_{it} and V_{it}^P is $\bar{\sigma}^2$-sub-Gaussian, where $\bar{\sigma}^2$ is a finite constant; (d) $\log(d_X NT)/N = o(1)$.

All the constants are understood to be independent of (N, T). The condition FS-TL summarizes Assumptions A1–A3 in Kock and Tang (2019). Plugging $\lambda_P = C_P \times \sqrt{NT\log^3(d_X + N)}$ into the stochastic bounds in Corollary A.1, page 332, Kock and Tang (2019) results in the lemma below, which establishes the properties of the first-stage treatment lasso.

Lemma 5.1 (First-Stage Treatment Lasso). *Under Condition* (*FS-TL*), *the lasso estimator in Example* 3.1 *with* $\lambda_P = C_P\sqrt{NT\log^3(d_X + N)}$ *large enough obeys the following bounds* $wp \to 1$:

$$\big\|\widehat{\delta}^P - \delta_0^P\big\|_1 \le N^{-1/2}\zeta_{NT,\infty}, \qquad \|\widehat{\xi} - \xi_0\|_1 \le \zeta_{NT,\infty}, \tag{5.6}$$

where, for some large enough constant $\bar{C}_P$,

$$\zeta_{NT,\infty} = \bar{C}^P S^P\big(T^{-1/2}\log^{3/2}(d_X + N)\big)^{(1-\nu)}. \tag{5.7}$$

Remark 5.2 (Time-Invariant Covariates). Consider Example 3.1. Suppose the condition FS-TL holds with time-invariant fixed covariates $(\bar{M}_i)_{i=1}^N$. There are (infinitely) many ways to decompose the total effect vector $a = (a_1, a_2, \ldots, a_N)$ into the observable part $M_i'\delta$ and remainder part ξ_i:

$$a_i = M_i'\delta + \xi_i, \quad \forall i : i = 1, 2, \ldots, N.$$

Given the sparsity parameters (ν, S^P), the minimizer $(\widehat{\delta}, \widehat{\xi})$ of the lasso optimization problem (3.4) obeys the bound (5.6) expressed in terms of (ν, S^P). Thus, we are free to choose (δ, ξ) whose sparsity parameters ν, S imply the tightest bound on $\zeta_{NT,\infty}$ in (5.7).

Multiple sparse decompositions must be equivalent in the following sense. Consider two possible decompositions

$$a_i = M_i'\delta^1 + \xi_i^1 = M_i'\delta^2 + \xi_i^2, \quad \forall i,$$

with $\zeta_{NT,\infty}^1$ and $\zeta_{NT,\infty}^2$ rates, respectively, determined by the weak sparsity parameters of (δ^1, ξ^1) and (δ^2, ξ^2). Let $\widehat{\xi}$ be any given minimizer to the lasso problem. Then these decompositions must be equivalent in the following sense:

$$\big\|\xi^1 - \xi^2\big\|_1 \le \big\|\xi^1 - \widehat{\xi}\big\|_1 + \big\|\widehat{\xi} - \xi^2\big\|_1 \le 2\max\big(\zeta_{NT,\infty}^1, \zeta_{NT,\infty}^2\big).$$

Further, consider the following example with exact sparsity. Suppose ξ is s-exactly sparse and M is a one-dimensional covariate drawn from Bernoulli distribution. $M_i \sim \text{Bern}(p_M)$ such that $s \ll Np_M$. Consider

$$a_i = M_i\delta^1 + \xi_i^1 = M_i\delta^2 + \xi_i^2, \quad \forall i,$$

such that $\xi_i^1 \neq \xi_i^2$ for at least one i (i.e., $\xi^1 \neq \xi^2$ as vectors). Then $\delta^1 \neq \delta^2$, and the vector $M_i(\delta^1 - \delta^2) \neq 0$ for at least $Np_M/2$ entries, w.p. $1 - o(1)$. Since $s \ll Np_M$, it cannot be the case that ξ^1 and ξ^2 are s-sparse at the same time. In this case, there exists a single decomposition (δ, ξ) obeying sparsity assumption.

Consider the following condition:

(FS-OL) Consider the model of Example 3.2:

$$Y_{it} = D_{it}'\beta_0 + X_{it}'\delta_0^P + \xi_i^E + U_{it},$$

where the residuals U_{it} are an m.d.s with respect to the filtration Φ_{it}. Suppose (a) $\xi^E = (\xi_i^E)_{i=1}^N \in \mathrm{R}^N$ and $(\delta_0^E, \beta_0) \in \mathrm{R}^{d_{DX}}$ are (ν^E, S^E) and $(\nu^E, N^{-\nu^E/2}S^E)$-weakly sparse vectors with $S^E = O(N^{1/2}\log^{3\nu^E/2}(d_X + N)T^{-\nu^E/2})$; $\|(\beta_0, \delta_0^E)\|_1$ is bounded (b) The Gram matrix

$$\Psi_{DX} := (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\big[(D_{it}, X_{it})(D_{it}, X_{it})'\big]$$

has all of its eigenvalues bounded from above and below by $B_{\max}$ and $B_{\min}$, where $0 < B_{\min} \leq B_{\max} < \infty$, w.p. $1 - o(1)$. (c) Each element of D_{it}, X_{it} and U_{it} is $\bar{\sigma}$-sub-Gaussian, where $\bar{\sigma}$ is a finite constant.

The following lemma follows from Kock and Tang (2019), and establishes the properties of the first-stage outcome lasso.

Lemma 5.2 (First-Stage Outcome Lasso). *Under the condition (FS-OL), the estimator* $(\check{\beta}, \widehat{\delta}^E, \widehat{\xi}^E)$ *defined in Example 3.2 obeys the following bounds wp* $1 - o(1)$:

$$\left\|\left(\check{\beta}, \widehat{\delta}^E\right) - \left(\beta_0, \delta_E^P\right)\right\|_1 \leq N^{-1/2}\zeta_{NT,\infty}^E, \qquad \left\|\widehat{\xi}^E - \xi_0^E\right\|_1 \leq \zeta_{NT,\infty}^E, \tag{5.8}$$

where, for some sufficiently large constant $\bar{C}_E$,

$$\zeta_{NT,\infty}^E = \bar{C}_E S^E\left(T^{-1/2}\log^{3/2}(d_X + N)\right)^{(1-\nu^E)}. \tag{5.9}$$

Next, we proceed to the construction of nuisance realization sets for the treatment and outcome models.

Remark 5.3 (Realization Sets for Reduced Form for Base Treatment). Define P_{NT} as a collection of N-vector functions:

$$P_{NT} = \left\{ \begin{array}{c} \mathbf{p}(x) = \left\{p_i(x_i)\right\}_{i=1}^N = \left\{x_i'\delta^P + \xi_i\right\}_{i=1}^N : \\ \left\|\delta^P - \delta_0^P\right\|_1 \leq N^{-1/2}\zeta_{NT,\infty}, \|\xi - \xi_0\|_1 \leq \zeta_{NT,\infty} \end{array} \right\}. \tag{5.10}$$

Under Condition (FS-TL), the mean square rate $\mathbf{p}_{NT}$ in (4.4) obeys $\mathbf{p}_{NT} = O(N^{-1/2}\zeta_{NT,\infty})$, and the sup-rate upper bound chosen as $\mathbf{p}_{NT,\infty} := 2\zeta_{NT,\infty}$ satisfies the sup-rate condition (4.5).

In the context of Example 2.1, define D_{NT} as a collection of N-vector functions

$$D_{NT} = \{\mathbf{d}(x) = \{d_i(x_i)\}_{i=1}^N = \{p_i(x_i)K(x_i)\}_{i=1}^n : \mathbf{p}(x) = \{p_i(x_i)\}_{i=1}^N \in P_{NT}\}.$$

The following remark helps verifying Assumption 4.5 in models with base treatment structure.

REMARK 5.4 (Deducing Rates in Base Treatment Cases). Consider the models (2.4)–(2.5) with single base treatment structure (2.10). Suppose the matrix $K(\cdot)$ in (2.10) has a.s. bounded entries. Suppose the base treatment reduced form vector $\mathbf{p}_0(\cdot)$ in (2.11) converges at rates $\mathbf{p}_{NT}$ and $\mathbf{p}_{NT,\infty}$. Then the worst-case rates $\mathbf{d}_{NT}$ and $\mathbf{d}_{NT,\infty}$ of the technical treatment reduced form in (2.10) obey $\mathbf{d}_{NT} = O(\mathbf{p}_{NT})$ and $\mathbf{d}_{NT,\infty} = O(\mathbf{p}_{NT,\infty})$. This follows from the Cauchy–Schwarz inequality.

REMARK 5.5 (Realization Sets for Reduced Form for Outcome). Suppose the matrix

$$\Psi_D := (NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T \mathrm{E}\big[\big(P_{it}K_{it}'\beta_0\big)^2 X_{it}X_{it}'\big]$$

has all of its eigenvalues bounded from above and below by $B_{\max}$ and $B_{\min}$, where $0 < B_{\min} \le B_{\max} < \infty$, w.p. $1 - o(1)$ and suppose $\|K(X_{it})\|_\infty \le \bar{K} < \infty$ a.s. for some constant $\bar{K}$. Define the outcome nuisance realization set

$$L_{NT} = \left\{ \begin{array}{l} \mathbf{l}(x) = \big(l_i(x_i)\big)_{i=1}^N = \{d_i(x_i)'\beta + x_i'\delta^E + \xi_i^E\}_{i=1}^N : \mathbf{d}(x) \in D_{NT}, \\ \|\beta - \beta_0\|_1 + \big\|\delta^E - \delta_0^E\big\|_1 \le N^{-1/2}\zeta_{NT,\infty}^E, \big\|\xi^E - \xi_0^E\big\|_1 \le \zeta_{NT,\infty}^E \end{array} \right\}. \tag{5.11}$$

Under Condition (FS-OL), the mean square rate $\mathbf{l}_{NT}$ in (4.4) obeys

$$\mathbf{l}_{NT} = O\big(N^{-1/2}\big(\zeta_{NT,\infty} + \zeta_{NT,\infty}^E\big)\big),$$

and the sup-rate upper bound chosen as $\mathbf{l}_{NT,\infty} := 2(\bar{K}\|\beta_0\|_1\zeta_{NT,\infty} + \zeta_{NT,\infty}^E)$ satisfies the sup-rate condition (4.5).

Combining the results from Remarks 5.1–5.5, we provide sufficient conditions to verify Assumption 4.5.

REMARK 5.6 (Verification of Assumption 4.5 for First-Stage Lasso Estimators). Consider the setup of Remarks 5.1–5.5 with $\nu, \nu^E < 1$. Suppose the scales S and S^E are not too big, namely,

$$\big(S^P\big)^2 N^{-1/2}T^{\nu-1/2}\log^{3(1-\nu)}(d_X + N) = o(1),$$
$$S^P \cdot S^E N^{-1/2}T^{(\nu+\nu^E)/2-1}\log^{3(1-(\nu+\nu^E)/2)}(d_X + N) = o(1).$$

Adding the equations above and multiplying by $(NT)^{1/2}$ gives

$$\sqrt{NT}\left(\mathbf{p}_{NT}^2 + \mathbf{p}_{NT}\mathbf{l}_{NT}\right) = o(1),$$

which suffices for (4.6). Likewise, assuming

$$\zeta_{NT,\infty}^P = \bar{C}^P S^P \left(T^{-1/2}\log^{3/2}(d_X + N)\right)^{(1-\nu)} = o\left(\log^{-1/2}(dNT)\right),$$
$$\zeta_{NT,\infty}^E = \bar{C}^E S^E \left(T^{-1/2}\log^{3/2}(d_X + N)\right)^{(1-\nu_E)} = o\left(\log^{-1/2}(dNT)\right)$$

directly verifies (4.7).

Orthogonal lasso achieves an oracle rate for CATE estimation, which can be strictly better than the nonorthogonal approach. The comparison is provided in terms of the upper bounds on the rates.

Remark 5.7 (Improvement of Orthogonal Lasso Upon One-Stage Lasso). Suppose the treatment effect vector β_0 is "less complex" than the first-stage parameter (δ_0^E, ξ_0^E), that is,

$$s\log d \ll \left(S^E\right)^2 T^{\nu^E}\log^{3(1-\nu^E)}(d_X + N). \tag{5.12}$$

Then, under Assumption 4.5, the upper bound on the oracle lasso rate is attained. Dividing (5.12) by NT and taking square root gives

$$\sqrt{s\log d/NT} = o\left(N^{-1/2}\zeta_{NT,\infty}^E\right),$$

where $N^{-1/2}\zeta_{NT,\infty}^E$ is the mean square rate bound of the preliminary nonorthogonal estimator $\check{\beta}$ in (3.6).

6. Empirical application

To show the immediate usefulness of the method, we consider the problem of inference on demand elasticities for grocery products. Our transactional data come from a major European food distributor that sells to retailers. The identifier of each observation consists of the cross-sectional index—the product code, the store location, the distribution channel (i.e., collection or delivery)—and the timestamp. For each value of the index, we compute weekly averages of the price and the quantity sold. Overall, we have 1163 unique products, sold at 8 site locations via 2 delivery channels, at $T = 208$ time periods (weeks). In addition to the transactional data, we have access to the product catalog, which classifies products into a tree. For example, the product code *Vanilla Soft Scoop Ice Cream 4 Ltr package* is classified into a hierarchy whose Level 1 is Sweets, Level 2 is Ice Cream & Shakes & Syrups, and Level 3 is Ice Cream. We filter out observations whose either price or sales is zero, which constitutes less than 5% of the sample.

The next step is to convert the categorical covariates representing classification into a vector of binary covariates. For each node j and the product i, the binary indicator for

the node j is equal to one if the product i belongs to the node j and zero otherwise. Since a binary indicator for a parent and all its children creates a linearly dependent covariate set, we exclude one child category for each parent. In the absence of any restrictions, different excluded categories yield numerically equivalent results. Under the sparsity assumption (2.16), this is no longer the case. The sparsity assumption requires that most siblings have similar treatment effects, albeit for a small number of exceptions whose identities are unknown. To obtain the sparse treatment modification effect, one has to exclude category that belongs to the majority (i.e., is not an exception). We assume that the store brand belongs to the majority, and can be taken as the baseline (excluded) category.

We postulate a partially linear dynamic panel model for weekly log demand

$$\begin{aligned}\log(Q)_{it} = \log(\mathrm{P})_{it} \cdot \Bigg(\sum_{h\in\mathcal{H}} 1\big\{h(i)=h\big\}\cdot\beta_{0h}\Bigg) \\ + \big(\log(\mathrm{P})_{it-1}, \log(Q)_{it-1}\big)'\big(\alpha_1^P, \alpha_1^S\big) \\ + \gamma^E_{h(i)} + a^E_{pc(i)} + \rho^E_{s(i)} + \zeta^E_{c(i)} + \gamma^E_{m(t)} + U_{it},\end{aligned} \tag{6.1}$$

where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ with $T = 208$ time periods (weeks). The cross-sectional unit index i indicates the combination of the product $pc(i)$ at the store $s = s(i)$ offered via $c = c(i)$ channel. The outcome variable $Y_{it} := \log(Q)_{it}$ is total log demand for unit i, and the base treatment $P_{it} := \log(\mathrm{P})_{it}$ is the log price. The hierarchy depth $\mathcal{H}$ varies between Level 2 (Figure 2a) and Level 3 (Figure 2b), and notation $h(i)$ denotes the hierarchical encoding of the product code $pc(i)$.

The model is a special case of Example 2.1. Here, the interaction covariates are the hierarchy fixed effects

$$K_i = \bigcup_{h\in\mathcal{H}} 1\big\{h(i)=h\big\}$$

and the parameter β_0 is

$$\beta_0 = \Bigg(\bigcup_{h\in\mathcal{H}} \beta_{0h}\Bigg),$$

which results in the CATE function

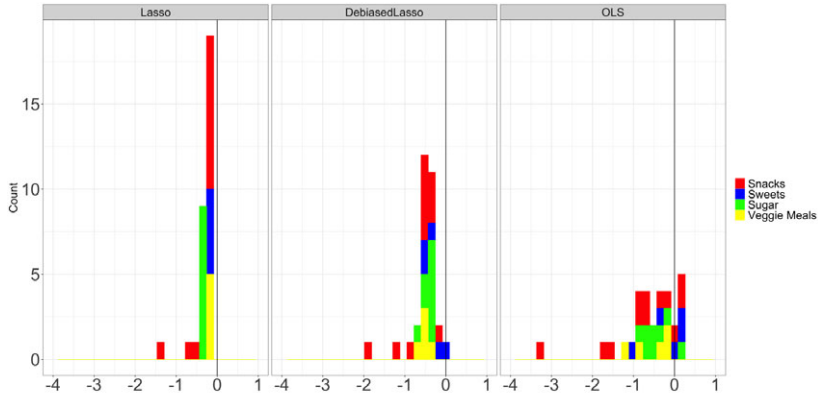
$$\epsilon_i(\beta_0) = K_i'\beta_0 = \sum_{h\in\mathcal{H}} 1\big\{h(i)=h\big\}\cdot\beta_{0h}$$

being equal to the heterogeneous elasticity $\epsilon_i(\beta_0)$ of unit i. In addition to the hierarchy fixed effects K_i, the first-stage controls include the product, store, channel fixed effects,

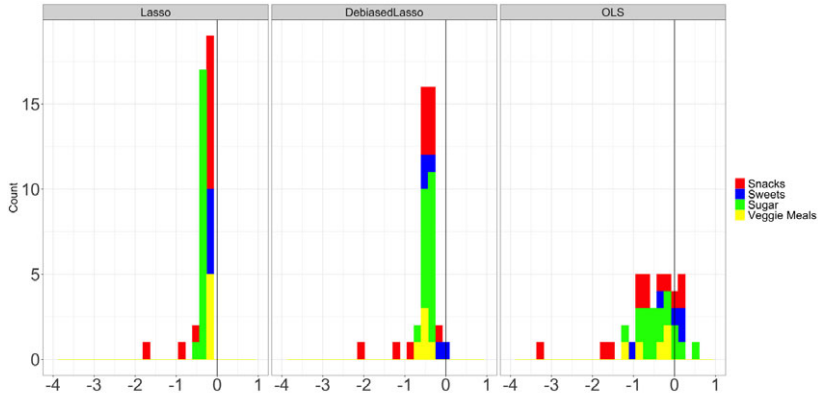
$$Z_i = \Bigg(\bigcup_{pc\in PC} 1\big\{pc(i)=pc\big\}, \bigcup_{s\in S} 1\big\{s(i)=s\big\}, \bigcup_{c\in C} 1\big\{c(i)=c\big\}\Bigg)$$

and the time effects $M_t = \bigcup_{m=1}^{12}\{m(t)=m\}$. Thus, the first-stage controls are

$$X^E_{it} = \big(\log(\mathrm{P})_{it-1}, \log(Q)_{it-1}, K_i, Z_i, M_t\big). \tag{6.2}$$



(a) Level 2 groups whose total is $d = 31$



(b) Level 3 groups whose total is $d = 40$

FIGURE 2. Histogram of estimated price elasticities for each category of Level 2 (Figure 2a) and Level 3 (Figure 2b) for Snacks. Estimates: Orthogonal lasso (left panel), debiased orthogonal lasso (middle panel), OLS (right panel). See the text for details.

Therefore, (6.1) is a special case of (2.4) with $X_{it} = X_{it}^E$ in (6.2) and $\xi_i^E = 0\ \forall i$:

$$Y_{it} = P_{it} \cdot (K_i' \beta_0) + (X_{it}^E)' \delta_0^E + 0 + U_{it}.$$

To estimate (β_0, δ_0^E), we invoke the lasso estimator of Example 3.2 with the outcome $Y_{it} = \log(\mathrm{Q})_{it}$ and the covariate vector $X_{it} = X_{it}^E$ in (6.2), restricting $\xi^E = (\xi_i^E)_{i=1}^N$ to be equal to zero.

The price equation takes the form

$$\log(\mathrm{P})_{it} = \log(\mathrm{P})_{it-1} \cdot \left(\sum_{h \in \mathcal{H}} h(i) \cdot \zeta_{0h}^P \right) + \gamma_{h(i)}^P + a_{pc(i)}^P + \rho_{s(i)}^P + \zeta_{c(i)}^P + \gamma_{m(t)}^E + V_{it}^P. \quad (6.3)$$

Taking

$$X_{it}^P = \left(\log(\mathrm{P})_{it-1} \cdot \bigcup_{h\in\mathcal{H}} 1\{h(i)=h\}, K_i, Z_i, M_t\right) \tag{6.4}$$

and $\xi_i = 0$ gives

$$P_{it} = \left(X_{it}^P\right)'\delta_0^P + 0 + V_{it}^P.$$

To estimate δ_0^P, we invoke the lasso estimator of Example 3.1 with the outcome P_{it} and the covariate vector $X^P it$ as in (6.4), restricting $\xi = (\xi_i)_{i=1}^N$ to be equal to zero.

In the second stage, we interact the first-stage price residuals $\widehat{V}_{it}^P$ with hierarchy fixed effects K_i to obtain treatment residual

$$\widehat{V}_{it} = \widehat{V}_{it}^P K_i.$$

Next, we regress the outcome residual $\widetilde{Y}_{it} = \widetilde{\log(Q)}_{it}$ onto V_{it},

$$\widetilde{Y}_{it} = V_{it}'\beta_0 + U_{it}.$$

The lasso estimator $\widehat{\beta}_L$ is as in Definition 3.3 with λ_β chosen by cross-validation. To simplify computation, the debiasing matrix is taken to be the Ridge inverse, which has similar properties to the CLIME estimator in the moderate-dimensional case. For each estimate $\widehat{\beta} \in \{\widehat{\beta}_{\mathrm{OLS}}, \widehat{\beta}_L, \widehat{\beta}_{\mathrm{DL}}\}$, we report a d-vector of distinct heterogeneous elasticities $(\bigcup_{h\in\mathcal{H}} \widehat{\epsilon}_h(\widehat{\beta}))'$, that is,

$$\widehat{\epsilon}_h(\widehat{\beta}) = \sum_{\bar{h}\in\mathcal{H}} 1\{\bar{h}=h\}\cdot\widehat{\beta}_h, \quad h\in\mathcal{H}.$$

We consider two choices of the partition $\mathcal{H}$: Level 2 partition (Figure 2a) with $d = 31$ and Level 3 partition (Figure 2b) with $d = 40$, respectively.

Figure 2 qualitatively summarizes our results. On each panel, the histogram shows estimated heterogeneous elasticities. The total number of points is equal to the total number of heterogeneous groups (i.e., the cardinality of $\mathcal{H}$). It is $d = 31$ for Figure 2a and $d = 40$ for Figure 2b. A single vertical bar represents a collection of heterogeneous groups with the same value of estimated elasticity, and its height shows the number of such groups. The distinct parts of the bar are grouped by Level 1 (Snacks, Sweets, Sugar, and Veggie Meals), as marked by color. A small number of distinct bars on Figure 2(b, left plot) indicates that the vector of heterogeneous elasticities' difference $\widehat{\beta}_h$ has many zeroes. As expected, the lasso elasticity estimate $\widehat{\beta}_L$ is sparse, which makes the histogram of $\widehat{\epsilon}_h(\widehat{\beta}_L)$ very concentrated. In contrast, the debiased lasso $\widehat{\beta}_{\mathrm{DOL}}$ is not sparse, and the histogram of $\widehat{\epsilon}_h(\widehat{\beta}_{\mathrm{DOL}})$ is more dispersed.

We find the Snacks category to be relatively homogenous. For example, lasso estimates suggest that all Sugar products (Figure 2 b, left panel, green bar) have the same elasticity value regardless of sugar type or packaging. As a result, all $d = 40$ heterogeneous groups can be pooled into $s = 7$ distinct ones. Second-stage shrinkage helps to reduce noise, which proves useful to obtain elasticities consistent with economic theory. For example, 7 out of $d = 40$ groups have positive OLS estimates, while neither lasso

nor debiased lasso have any. We find the debiased lasso elasticity estimates to be the ones most consistent with economic theory predictions.

7. Extensions

The following extensions are not formally covered by theoretical framework of Section 4. Nevertheless, we expect the results would extend to these settings with suitable treatment of clustering, given the recent developments of Chiang et al. (2019).

Heterogeneous own and cross-price elasticities with many heterogeneous products

Consider a firm that makes a pricing decision about a large number N of heterogeneous goods. Let $\mathcal{C} : \{1, 2, \dots, N\} \to \{1, 2, \dots G\}$ be a known partition of the products into the set of G independent clusters. The notation $\mathcal{C}(i)$ stands for all members of the i'th cluster. For any two products i and j from distinct clusters $\mathcal{C}(i) \neq \mathcal{C}(j)$, the cross-price elasticity between i and j is assumed to be zero. Define the average leave-i-out price of products in the i'th cluster as

$$P_{-it} := \frac{\sum_{j \in \mathcal{C}(i), j \neq i} P_{jt}}{|\mathcal{C}(i) - 1|}. \tag{7.1}$$

Suppose that in the short term the realizations of prices and sales can be approximated by the following partially linear model:

$$Y_{it} = D_{it}'\beta_0 + e_0(X_{it}) + \xi_i^E + U_{it}, \qquad \mathrm{E}\big[U_{it} | (X_{jt}, P_{jt}, \Phi_{jt})_{j \in \mathcal{C}(i)}\big] = 0, \tag{7.2}$$

$$D_{it} = \big[K_{it}' P_{it}, K_{it}' P_{-it}\big], \tag{7.3}$$

$$P_{it} = p_0(X_{it}) + \xi_i + V_{it}^P, \qquad \mathrm{E}\big[V_{it}^P | X_{it}, \Phi_{it}\big] = 0, \tag{7.4}$$

where Y_{it} is the log sales of product i at time t, P_{it} is the log price, X_{it} is a p_X-vector of the observable product characteristics, the lagged realizations of market quantities Y_{it}, P_{it}, and the demand-side variables used for strategic price setting by the firm. The symbol Φ_{it} denotes the full information set available for unit i prior to period t, spanned by lagged realizations of the demand system. The controls X_{it} affect the price variable P_{it} through $p_0(\cdot)$ and the sales through $e_0(\cdot)$.

The technical treatment D_{it} is formed by interacting P_{it} and P_{-it} with the observable product characteristics K_{it} such that $\mathbb{E}K_{it} = 0$. The parameter β_0 stands for the vector of own and cross-price elasticities. In order to assign a causal interpretation to β_0, we assume that, after conditioning on all predetermined variables, the sales shock U_{it} is mean independent of all the information $(X_{jt}, \Phi_{jt}, P_{jt})_{j \in \mathcal{C}(i)}$ about members of the i'th cluster (i.e., U_{it} is dissociated from $(X_{jt}, \Phi_{jt}, P_{jt})_{j \in \mathcal{C}(i)}$, Chiang et al. (2019)). The asymptotic results should be clustered at the level of independent clusters G rather than individual products N.

Equation (2.11) defines the price effect of interest

$$\beta_0 = \big(\beta_0^{\text{own}}, \beta_0^{\text{cross}}\big),$$

where β_0^{own} and β_0^{cross} are $d/2$ dimensional vectors of the own- and the cross-price effect, respectively. A change in the own price ΔP_{it} affects demand via

$$\Delta D_{it}' \beta_0 = \Delta P_{it} K_{it}' \beta_0^{\text{own}},$$

and a change in the average price ΔP_{-it} affects demand via

$$\Delta D_{it}' \beta_0 = \Delta P_{-it} K_{it}' \beta_0^{\text{cross}}.$$

Let

$$\beta_0^{\text{own}} := \left(\alpha_0^{\text{own}}, \gamma_0^{\text{own}}\right) \quad \text{and} \quad \beta_0^{\text{cross}} := \left(\alpha_0^{\text{cross}}, \gamma_0^{\text{cross}}\right).$$

We see that

- α_0^{own} is the baseline own elasticity, and $K_{it}' \gamma_0^{\text{own}}$ is the heterogeneous own elasticity;
- α_0^{cross} is the baseline cross-price elasticity, and $K_{it}' \gamma_0^{\text{cross}}$ is the heterogeneous cross-price elasticity.

References

Arellano, Manuel and Stephen Bond (1991), "Some tests of specification for panel data. Monte Carlo evidence and an application to employment equations." *Review of Economic Studies*, 58, 277–297. [478]

Athey, Susan and Guido Imbens (2016), "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, 113, 7353–7460. Available at https://www.pnas.org/content/113/27/7353. [472]

Banerjee, Abhijit, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, Francine Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha (2021), "Selecting the most effective nudge: Evidence from a large-scale experiment on immunization." Available at https://arxiv.org/abs/2104.09645. [472]

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato (2014), "Uniform post-selection inference for least absolute deviation regression and other z-estimation problems." *Biometrika*, 102, 77–94. doi:10.1093/biomet/asu056. [480]

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012), "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica*, 80, 2369–2429. [487]

Belloni, Alexandre and Victor Chernozhukov (2013), "Least squares after model selection in high-dimensional sparse models." *Bernoulli*, 19, 521–547. [495]

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Ying Wei (2019), "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework." *Annals of Statistics*, 46, 3643–3675. [474]

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen (2017), "Program evaluation and causal inference with high-dimensional data." *Econometrica*, 85, 233–298. [487]

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2010), "Lasso methods for Gaussian instrumental variables models." arXiv preprint, arXiv:1012.1297. [473, 474, 482]

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2011), "Inference for high-dimensional sparse econometric models." arXiv preprint, arXiv:1201.0220. [474]

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014), "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies*, 81, 608–650. [474]

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato (2014), "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems." *Biometrika*, 102, 77–94. [474]

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei (2016), "Post-selection inference for generalized linear models with many controls." *Journal of Business & Economic Statistics*, 34, 606–619. [495]

Berbee, Henry (1987), "Convergence laws in the strong law for bounded mixing sequences." *Probability Theory and Related Fields*, 74, 255–270. [484]

Bonhomme, Stephane, Thibaut Lamadon, and Elena Manresa (2019a), "Discretizing unobserved heterogeneity." [475]

Bonhomme, Stephane, Thibaut Lamadon, and Elena Manresa (2019b), "A distributional framework for matched employer employee data." *Econometrica*, 87, 699–739. [475]

Cai, Tony, Weidong Liu, and Xi Luo (2011), "A constrained l_1 minimization approach to sparse precision matrix estimation." *Journal of the American Statistical Association*, 106, 594–607. [474, 487]

Chamberlain, Gary (1982), "Multivariate regression models for panel data." *Journal of Econometrics*, 18, 5–46. [475]

Chen, Mingli, Ivan Fernandez-Val, and Martin Weidner (2020), "Nonlinear factor models for network and panel data." *Journal of Econometrics*. [475]

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018), "Double/debiased machine learning for treatment and structural parameters." *Econometrics Journal*, 21, C1–C68. [473, 474, 482]

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2013), "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors." *Annals of Statistics*, 41, 2786–2819. [493]

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2014), "Gaussian approximation of suprema of empirical processes." *Annals of Statistics*, 42, 1564–1597. [475]

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2017), "Central limit theorems and bootstrap in high dimensions." *Annals of Probability*, 45, 2309–2352. [475]

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2019), "Inference on causal and structural parameters using many moment inequalities." *Review of Economic Studies*, 86, 1867–1900. [475, 493]

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val (2017), "Generic machine learning inference on heterogenous treatment effects in randomized experiments." arXiv e-prints, arXiv:1712.04802. [472]

Chernozhukov, Victor, Chris Hansen, and Martin Spindler (2016), "High-dimensional metrics in R." Available at https://arxiv.org/abs/1603.01700. [487]

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (2015), "Valid post-selection and post-regularization inference: An elementary, general approach." *Annual Review of Economics*, 7, 649–688. [474]

Chevalier, Judith, Anil Kashyap, and Peter Rossi (2003), "Why don't prices rise during periods of peak demand? Evidence from scanner data." *American Economic Review*, 93, 15–37. [473]

Chiang, Harold D. (2018), "Many average partial effects: With an application to text regression." arXiv e-prints, arXiv:1812.09397. [474]

Chiang, Harold D., Kengo Kato, Yukun Ma, and Yuya Sasaki (2019), "Multiway cluster robust double/debiased machine learning." arXiv e-prints, arXiv:1909.03489. [474, 504]

Colangelo, Kyle and Ying-Ying Lee (2020), "Double debiased machine learning nonparametric inference with continuous treatments." arXiv e-prints, arXiv:2004.03036. [474]

Davis, Jonathan M. V. and Sara B. Heller (2020), "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs." *Review of Economics and Statistics*, 102, 664–677. doi:10.1162/rest_a_00850. [472]

Dudley, Richard Mansfield and Walter Philipp (1983), "Invariance principles for sums of Banach space valued random elements and empirical processes." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62, 509–552. [484]

Fan, Qingliang, Yu-Chin Hsu, Robert P. Lieli, and Yichong Zhang (2019), "Estimation of conditional average treatment effects with high-dimensional data." arXiv e-prints, arXiv:1908.02399. [474]

Farrell, Max, Tengyuan Liang, and Sanjog Misra (2021), "Deep neural networks for estimation and inference." *Econometrica*, 89, 181–213. [495]

Fernandez-Val, Ivan and Joonhwan Lee (2013), "Panel data models with non-additive unobserved heterogeneity: Estimation and inference." *Quantitative Economics*, 4, 453–481. [488]

Fernandez-Val, Ivan and Martin Weidner (2018), "Fixed effect estimation of large t panel data model." *Annual Review of Economics*, 10, 109–138. [475]

Gao, Wayne Yuan, and Ming Li (2019), "Robust semiparametric estimation in panel multinomial choice models." [475]

Haavelmo, Trygve (1944), "The probability approach in econometrics." *Econometrica*, 12, 1–115. [476]

Hahn, Jinyong and Guido Kuersteiner (2011), "Bias reduction for dynamic nonlinear panel models with fixed effects." *Econometric Theory*, 27, 1152–1191. [488]

Hasminskii, Rafail Z. and Ildar A. Ibragimov (1979), "On the nonparametric estimation of functionals." In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*. [473, 474, 482]

Jacob, Daniel, Wolfgang Karl Härdle, and Stefan Lessmann (2019), "Group average treatment effects for observational studies." [474]

Javanmard, Adel and Andrea Montanari (2014), "Confidence intervals and hypothesis testing for high-dimensional regression." *Journal of Machine Learning Research*, 2, 2869–2909. [474]

Kennedy, Edward, (2020), "Towards optimal doubly robust estimation of heterogeneous causal effects." arXiv preprint, arXiv:2004.14497. [486]

Klosin, Sylvia (2021), "Automatic double machine learning for continuous treatment effects." [474]

Kock, Anders Bredahl (2016a), "Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions." *Econometric Theory*, 32, 243–259. [478]

Kock, Anders Bredahl (2016b), "Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models." *Journal of Econometrics*, 195, 71–85. [475]

Kock, Anders Bredahl and Haihan Tang (2019), "Uniform inference in high-dimensional dynamic panel data models." *Econometric Theory*, 35, 295–359. [474, 475, 478, 479, 484, 485, 495, 497, 498]

Leeb, Hannes and Benedikt Potcher (2005), "Model selection and inference: Facts and fiction." *Econometric Theory*, 21, 21–59. [481]

Lu, Xun and Liangjun Su (2016), "Shrinkage estimation of dynamic panel data models with interactive fixed effects." *Journal of Econometrics*, 190, 148–175. [475]

Lu, Xun and Liangjun Su (2020), "Determining individual or time effects in panel data models." *Journal of Econometrics*, 215, 60–83. [475]

Luo, Ye and Martin Spindler (2016), "High-dimensional L_2-boosting: Rate of convergence." arXiv e-prints, arXiv:1602.08927. [495]

Manresa, Elena (2016), "Estimating the structure of social interactions using panel data." [475]

Moon, Hyungsik Roger, Matthew Shum, and Martin Weidner (2018), "Estimation of random coefficients logit demand models with interactive fixed effects." *Journal of Econometrics*, 206, 613–644. [475]

Mundlak, Yair (1978), "On the pooling of time series and cross section data." *Econometrica*, 46, 69–85. [471, 475]

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu (2012), "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers." *Statistical Science*, 27, 538–557. [495]

Nickell, Stephen (1981), "Biases in dynamic models with fixed effects." *Econometrica*, 49, 1824–1851. [478]

Nie, Xinkun and Stefan Wager (2020), "Quasi-oracle estimation of heterogeneous treatment effects." Biometrika. [474]

Oprescu, Miruna, Vasilis Syrgkanis, and Zhiwei Steven Wu (2018), "Orthogonal random forest for causal inference." arXiv e-prints, arXiv:1806.03467. [474]

Robinson, Peter M. (1988), "Root-n-consistent semiparametric regression." *Econometrica*, 56, 931–954. [472]

Schick, Anton (1986), "On asymptotically efficient estimation in semiparametric models." *Annals of Statistics*, 14, 1139–1151. [473, 474, 482]

Schmidt-Hieber, Johannes (2017), "Nonparametric regression using deep neural networks with ReLU activation function." arXiv e-prints, arXiv:1708.06633. [495]

Semenova, Vira and Victor Chernozhukov (2021), "Debiased machine learning of conditional average treatment effects and other causal functions." *Econometrics Journal*, 24, 2, 264–289. [474]

Semenova, Vira, Matt Goldman, Victor Chernozhukov, and Matt Taddy (2023), "Supplement to 'Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence'." *Quantitative Economics Supplemental Material*, 14, https://doi.org/10.3982/QE1670. [475]

Strassen, Volker (1965), "The existence of probability measures with given marginals." *Annals of Mathematical Statistics*, 36, 423–439. [483]

Su, Liangjun, Zhentao Shi, and Peter Phillips (2016), "Identifying latent structures in panel data." *Econometrica*, 84, 1824–1851. [475]

Syrganis, Vasilis and Manolis Zampetakis (2020), "Estimation and inference with trees and forests in high dimensions." arXiv e-prints, arXiv:2007.03210. [495]

Ura, Takuya (2018), "Heterogeneous treatment effects with mismeasured endogenous treatment." *Quantitative Economics*, 9, 1335–1370. [474]

van der Geer, Sara, Peter Bühlmann, Yaakov Ritov, and Ruben Dezeure (2014), "On asymptotically optimal confidence regions and tests for high-dimensional models." *Annals of Statistics*, 42, 1166–1202. [474, 481, 487, 495]

Villani, Cédric (2007), *Topics in Optimal Transportation*, Vol. 58. American Mathematical Soc. [484]

Wager, Stefan and Susan Athey (2018), "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113, 1228–1242. [472, 474]

Wager, Stefan and Guenther Walther (2015), "Adaptive concentration of regression trees, with application to random forests." arXiv e-prints, arXiv:1503.06388. [495]

Zhang, Cun-Hui and Stefanie Zhang (2014), "Confidence intervals for low-dimensional parameters in high-dimensional linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217–242. [474, 487]

Zhang, Danna and Wei Biao Wu (2017), "Gaussian approximation for high dimensional time series." *Annals of Statistics*, 45, 1895–1919. [493]

Zheng, Wenjing and Mark J. van der Laan (2010), "Asymptotic theory for cross-validated targeted maximum likelihood estimation." Technical report, UC Berkeley Division of Biostatistics. [473, 474, 482]

Zimmert, Michael and Michael Lechner (2019), "Nonparametric estimation of causal heterogeneity under high-dimensional confounding." arXiv e-prints, arXiv:1908.08779. [474]

Co-editor Christopher Taber handled this manuscript.

Manuscript received 27 June, 2020; final version accepted 3 December, 2022; available online 20 December, 2022.