Orthogonal ML for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels*

Victor Chernozhukov
MIT
vchern@mit.edu

Matt Goldman
Microsoft AI & Research
mattgold@microsoft.com

Vira Semenova
MIT
vsemen@mit.edu

Matt Taddy
Microsoft AI & Research
taddy@microsoft.com

September 8, 2018

Abstract

There has been growing interest in how economists can import machine learning tools designed for prediction to facilitate, optimize and automate the model selection process, while still retaining desirable inference properties for causal parameters. Focusing on partially linear models, we extend the *Double ML* framework to allow for (1) a number of treatments that may grow with the sample size and (2) the analysis of panel data under sequentially exogenous errors. Our *low-dimensional treatment* (LD) regime directly extends the work in Chernozhukov et al. (2016), by showing that the coefficients from a second stage, ordinary least squares estimator attain root-N convergence and desired coverage even if the dimensionality of treatment is allowed to grow at a rate of $O(N/\log N)$. Additionally we consider a *high-dimensional sparse* (HDS) regime in which we show that second stage orthogonal LASSO and debiased orthogonal LASSO have asymptotic properties equivalent to oracle estimators with known first stage estimators. We argue that these advances make Double ML methods a desirable alternative for practitioners estimating short-term demand elasticities in non-contractual settings.

*We would like to thank Vasilis Syrgkanis, Whitney Newey, Anna Mikusheva, Brian Quistorff and participants of MIT Econometrics Lunch for valuable comments.

1 Introduction

Estimation of counterfactual outcomes is a key aspect of economic policy analysis and demands a large portion of the applied economist's efforts in both industry and academia. In the absence of explicit exogenous variation – i.e., independence, randomization or experimentation – applied economists must rely on human judgment to specify a set of controls that allow them to assign causal interpretation to their estimates. This method of model selection is highly subjective and labor intensive. In response, there has been growing interest in the use of Machine Learning (ML) tools to automate and accelerate the economist's model selection process (Athey, 2017). The challenge here is to build estimators that can leverage ML prediction tools while still retaining desirable inference properties.

Recent work in econometrics and statistics has demonstrated the potential for use of ML methods in partialling out the influence of high-dimensional potential controls (Belloni and Chernozhukov (2013), Belloni et al. (2016b)). In particular, the *Double ML* framework of Chernozhukov et al. (2016) provides a general recipe for construction of control functions that can be used for inference on low dimensional treatment effects. In their partially linear model for cross-sectional data, Double ML implies first-stage estimation of both treatment and response functions of the high-dimensional controls. Using sample splitting, the out-of-sample residuals for each of these two first-stage regressions represent exogenous variation that can be used to identify a valid causal effect.

We argue that adaptations and extensions of this approach will be useful in a wide variety of common applied econometric measurement problems. In this paper, we focus on extensions that will enable the use of ML tools in firm-side demand analysis. In such settings, the econometrician works with panel data on the prices and sales of their firm's products. They will typically have available very rich item descriptions – product hierarchy information, textual descriptions and reviews, and even product images. Being on the firm side, they will also have access to the universe of demand-side variables that were used in strategic price-setting. Thus, they can confidently identify price sensitivities by assuming that after conditioning on the available demand signals, the remaining variation in price is exogenous. Moreover, the novel algorithms proposed in this paper will allow us to use the rich item descriptions to index heterogeneous product-specific price sensitivities.

Unlike most existing demand analysis frameworks, we do not require the presence of instrumental variables (e.g. cost or mark-up shifters) to identify demand elasticities. Instead, we assume that our possession of the universe of demand signals known to the firm allows us to project out the systematic component of a firm's pricing rule and 'discover' events of exogenous price variation. Of course, such an assumption may not always be realistic for economists that do not work at, or with, the firms of interest. But, given this information, the Double ML framework facilitates valid identification by allowing us to learn the control functions defined on these high-dimensional demand signals without over-fitting. Moreover, this approach allows us to use all residual price variation to learn price sensitives. We thus are likely to achieve much more precise estimation than BLP type approaches who derive identification from a small number of cost shifters. This precision will be of primary importance to an industrial practitioner looking optimize future price and promotion decisions. Finally, BLP models may often derive identification from markup shifters (e.g. sums of characteristics of competing products) that impact a firm's chosen price. However, these markup shifters (often referred to as "BLP instruments") are typically combined with an assumption of Bertrand-Nash rationality in order to generate valid moment conditions and thus are of limited utility for a firm assessing the optimality of it's own pricing decisions.

The proposed Double ML framework for estimating demand elasticities is as follows. First, we estimate reduced form relations of expected log price and sales conditional on all past realizations of demand system (lagged prices and sales). This is purely a prediction problem and we seek to maximize out of sample fit. Generally, any additionally available demand signals should also be used subject to the constraint of not using any "bad controls" that might be impacted by the realized price decision.[1] The residuals from our price model may then be interacted with indicators for location, channel, product or product type in order to produce a vector of treatments which can then be regressed against our sales residuals in a second stage estimation. The resulting coefficients then correspond to heterogeneous demand elasticities. Information about our products is organized in a hierarchical structure and we assume that the majority of product types in a particular node of the hierarchy have similar demand elasticities. Therefore, demand elasticities, modeled via such hierarchy, may have a sparse representation. Thus we suggest use of Lasso as a second stage estimator to learn these heterogeneous elasticities. When inference is desired, we suggest a debiased Lasso (similar to Javanmard and Montanari (2014)) to construct point-wise and simultaneous confidence intervals for the estimates of elasticities.

There are several methodological innovations necessary to extend from the original Double ML work to this real-world demand analysis scenario. The first concerns the large

[1] As an example, contemporaneous search volumes could be impacted by the choice to put a product on discount and thus should be excluded from our first stage models.

number of goods (and therefore treatment effects) in our setting. Chernozhukov et al. (2016) shows the validity of Double ML only for the application of a second stage OLS estimator to an asymptotically fixed number of treatments. Our first contribution is to show that such estimates yield valid inference even if the dimensionality of treatment is allowed to grow at a rate of $O(N/\log N)$. This result is presented in our *Low Dimensional* regime presented in Section 3.2. Relaxing this requirement still further, we consider a *High Dimensional Sparse* framework, where we replace second-stage OLS used in original Double ML by Lasso for estimation and a version of debiased Lasso for inference. We show that l_1 penalty is compatible with the noise incurred in previous estimation step. In particular, Lasso achieves oracle rate and debiased Lasso achieves oracle asymptotic distribution. Moreover, the latter can be used to test large number of hypotheses (which may be necessary for inference on elasticities composed from coefficients on many heterogeneous treatments). Another innovation of our method is the panel setting with item heterogeneity. We adopt correlated random effects approach and model item heterogeneity using dynamic panel Lasso (Kock and Tang (2016)), allowing for a weakly sparse unobserved heterogeneity. Given the richness of our item descriptions, we find this to be a plausible assumption. In case one is willing to make a stronger assumption of zero unobserved heterogeneity, any ML algorithm can be used at the first stage.

Finally, we extend the original Double ML framework to allow for affine modifications of a single residualized treatment variable. Applying Double ML to a high-dimensional vector of treatments would typically require a separate residualization operation for each treatment. However, in demand applications, all treatments will often be affine modifications of price. For example, interaction of price with time-invariant observables corresponds to heterogeneous own-price elasticities and leave-me-out averages may model average cross-price effects within a product category. Example 2 gives more details about such examples. As a result of affine construction, the first stage algorithm trains single price variable, instead of each affine treatment separately. This achieves better precision of the first-stage estimates and speeds up computational time.

We posit a high-dimensional log linear demand model in which a products sales may be impacted by a large number of treatments that allow for a rich pattern of heterogeneous own and cross-price elasticities. We assume that after conditioning on available demand signals, remaining variation in price is exogenous and can be used for identification of price sensitivities. Usage of ML technique and access to all of demand-side variables observed by business decision makers validates our identification. Finally, if randomized prices, if available, may provide external validity to our results.

The rest of the paper proceeds as follows. Section 2 describes our partially linear framework and gives some motivating examples. Sections 3 provide our main theoretical results for the Low Dimensional and High Dimensional Sparse regimes. Section 4 discusses strategies and results for first stage estimation of treatment and outcome.

2 Motivating Examples and The Essentials of Orthogonal Machine Learning For Structural Parameters

2.1 Motivating Examples

Our paper provides a framework for analyzing a rich variety of examples, some of which we illustrate below.

Example 1 (**Heterogeneous Treatment Effects with Modeled Heterogeneity**). Consider the following measurement model for treatment effects:

$$Y_{it} = D_{it}'\beta_0 + g_{i0}(Z_{it}) + U_{it}, \quad \mathbb{E}[U_{it}|D_{it}, Z_{it}, \Phi_t] = 0 \tag{2.1}$$

$$D_{it} = P_{it}X_{it} \tag{2.2}$$

$$P_{it} = p_{i0}(Z_{it}) + V_{it}, \quad \mathbb{E}[V_{it}|Z_{it}, \Phi_t] = 0 \tag{2.3}$$

where Y_{it} is a scalar outcome of unit i at time t, P_{it} is a "base" treatment variable, $X_{it} = (1, \tilde{X}_{it}) : \mathbb{E}\tilde{X}_{it} = 0$ is a d-vector of observable characteristics of unit i, and Z_{it} is a p-vector of controls, which includes X_{it}. The technical treatment vector D_{it} is formed by interacting the base treatment P_{it} with covariates X_{it}, creating a vector of high-dimensional treatments. The set $\Phi_t = \{Y_{i,k}, P_{i,k}, Z_{i,k}\}_{k=1}^{t-1}$ denotes the full information set available prior to period t. In practice we will assume that this set is well approximated by the several lags of outcome and base treatment variables. Equation (2.3) keeps track of confounding, that is, the effect of Z_{it} on P_{it}. The controls Z_{it} affect the treatment P_{it} through $p_{i0}(Z_{it})$ and the outcome through $g_{i0}(Z_{it})$.

In order to enable a causal interpretation for the parameters in the first measurement equation, we assume the conventional assumption of **conditional sequential exogeneity** holds, namely that the stochastic shock U_{it} governing the potential outcomes is mean independent of the past information Φ_t, controls Z_{it} and contemporaneous treatment variables P_{it} (and hence the technical treatment.) Equation (2.1) is the main measurement equation, and

$$\beta_0 = (\alpha_0, \gamma_0')$$

is the high-dimensional parameter of interest, whose components characterize the treatment effect via

$$\Delta D_{it}'\beta_0 = \Delta P_{it} X_{it}\beta_0 = \underbrace{\alpha_0}_{\text{ATE}} + \underbrace{X_{it}'\beta_0}_{\text{TME}},$$

where Δ denotes either a partial unit difference or a partial derivate with respect to the base treatment P_{it}. We see that

- α_0 is the Average Treatment/Structural Effect (ATE), and
- $X_{it}'\beta_0$ describes the so-called Treatment/Structural Modification Effect (TME).

It is useful to write the equations in the "partialled out" or "residualized" form:

$$\tilde{Y}_{it} = \tilde{D}_{it}'\beta_0 + U_{it} = \tilde{P}_{it} X_{it}'\beta_0 + U_{it} \tag{2.4}$$

where

$$\tilde{P}_{it} = P_{it} - \mathbb{E}[P_{it} \mid Z_{it}, \Phi_t] \text{ and } \tilde{Y}_{it} = Y_{it} - \mathbb{E}[Y_{it} \mid Z_{it}, \Phi_t]$$

denote the partialled out treatment and outcome, respectively. We assume that after conditioning on Z_{it},

$$\{\{\tilde{P}_{it}, \tilde{Y}_{it}\}_{t=1}^T\}_{i=1}^I$$

is an i.i.d sequence across i. For each i, $\{\tilde{P}_{it}, \tilde{Y}_{it}\}_{t=1}^T$ is a martingale difference sequence by time.

A key insight of our orthogonal/debiased machine learning is that we will be able to construct high quality point estimators and confidence parameters for both α_0 *and the high-dimensional parameter* γ_0, by essentially first estimating the residualized form of the equations above and then performing either ordinary least squared or lasso with de-biasing on the residualized form given above[2].

Example 2 (Demand Functions with Cross-Price Effects)**.** Consider the following model:

$$Y_{it} = D_{it}'\beta_0 + g_{i0}(Z_{it}) + U_{it}, \quad \mathbb{E}[U_{it}|D_{it}, Z_{it}, \Phi_t] = 0 \tag{2.5}$$

$$D_{it} = [P_{it}X_{it}, P_{-it}X_{it}] \tag{2.6}$$

$$P_{it} = p_{i0}(Z_{it}) + V_{it}, \quad \mathbb{E}[V_{it}|Z_{it}, \Phi_t] = 0 \tag{2.7}$$

[2]This partialling out approach has classical roots in econometrics, going back at least to Frisch and Waugh. In conjunction with machine learning, it was used in high-dimensional sparse linear models in Belloni et al. (2014) and with generic machine learning methods in Chernozhukov et al. (2016); in the latter paper only low-dimensional β_0's are considered, and in the former high-dimensional β_0's were considered.

where Y_{it} is log sales of product i at time t, P_{it} is a log price, $X_{it} = (1, X_{it})$ is a d-vector of observable characteristics, and Z_{it} is a p-vector of controls, which includes X_{it}. Let C_i be a set of products which have a non-zero cross-price effect on sales Y_{it}. For a product i, define the average leave-i-out price of products in C_i as:

$$P_{-it} = \frac{\sum_{j \in C_i} P_{it}}{|C_i|},$$

The technical treatment D_{it} is formed by interacting P_{it} and P_{-it} with observable product characteristics X_{it}, creating a vector of heterogeneous own and cross price effects. The $\Phi_t = \{Y_{ik}, P_{ik}, Z_{ik}\}_{k=1}^{t-1}$ denotes the full information set available prior to period t, spanned by lagged realizations of demand system. In practice we will assume that this set is well approximated by the several lags of own sales and price. Equation (2.7) keeps track of confounding, that is, the effect of Z_{it} on P_{it}. The controls Z_{it} affect the price variable P_{it} through $p_{i0}(Z_{it})$ and the sales through $g_{i0}(Z_{it})$. Conditional on observables, the sales shock U_{it} is mean independent of the past information Φ_t, controls Z_{it}, price P_{it} and P_{-it} .

Equation (2.5) defines the price effect of interest

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

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

$$\Delta D_{it}' \beta_0 = \Delta P_{it} X_{it} \beta_0^{own}$$

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

$$\Delta D_{it}' \beta_0 = \Delta P_{-it} X_{it} \beta_0^{cross}$$

Let

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

We see that

- α_0^{own} is the Average Own Elasticity and $X_{it}' \gamma_0^{own}$ is the Heterogenous Own Elasticity
- α_0^{cross} is the Average Cross-Price Elasticity and $X_{it}' \gamma_0^{cross}$ is Heterogenous Cross-Price Elasticity

It is useful to write the equations in the "partialled out" or "residualized" form:

$$\tilde{Y}_{it} = \tilde{D}'_{it}\beta_0 + U_{it} = [\tilde{P}_{it}X'_{it}, \tilde{P}_{-it}X_{it}]\beta_0 + U_{it} \tag{2.8}$$

where

$$\tilde{P}_{it} = P_{it} - \mathbb{E}[P_{it} \mid Z_{it}, \Phi_t] \text{ and } \tilde{Y}_{it} = Y_{it} - \mathbb{E}[Y_{it} \mid Z_{it}, \Phi_t]$$

denote the partialled out log price and sales, respectively.

We argue that orthogonal/debiased machine learning framework is especially appropriate for firm-side demand analysis . Specifically, we view the reduced form Equation 2.5 as best linear approximation to a true demand model in a short run, that business practitioners use for forecasting and decision making . The controls Z_{it} contain product information, lagged realizations of market quantities, and demand-side variables used for strategic price setting. Conditional on pre-determined information in Z_{it}, the residual price variation $\tilde{P}_{it}$, can be credibly used to identify price effects. A high-dimensional vector X_{it} summarizes rich product descriptions, time, and demographic information. Using the methods of this paper we are able to deliver high-quality point estimates and confidence intervals for both average effects α_0^{own} and α_0^{cross} and high-dimensional heterogeneous effects γ_0^{own} and γ_0^{cross} by first estimating the residualized form of the equations above and then performing either ordinary least squared or lasso with de-biasing on the residualized form given above.

2.2 The Econometric Model

Throughout our analysis, we consider a sequentially exogenous, partially linear panel model

$$Y_{it} = D'_{it}\beta_0 + g_{i0}(Z_{it}) + U_{it} \qquad \mathbb{E}[U_{it}|D_{it}, Z_{it}, \Phi_t] = 0, \tag{2.9}$$

$$D_{it} = d_{i0}(Z_{it}) + V_{it} \qquad \mathbb{E}[V_{it}|Z_{it}, \Phi_t] = 0 \tag{2.10}$$

where the indices i and t denote an item $i \in [I] \equiv \{1, 2, .., I\}$ and a time period $t \in [T] \equiv \{1, 2, .., T\}$, respectively. The variables Y_{it}, D_{it} and Z_{it} denote a scalar outcome, d-vector of treatments, and p-vector of controls respectively. The set $\Phi_t = \{Y_{i,k}, P_{i,k}, Z_{i,k}\}_{k=1}^{t-1}$ denotes the full information set available prior to period t. In practice we will assume that this set is well approximated by the several lags of outcome and treatment variables.

Let the set of items I belong to M independent groups of size C :[3]

$$[I] = \{(m, c), m \in \{1, 2, .., M\}, c \in \{1, 2, .., C\}\}$$

denote by $m(i)$ the index of the group of item $i = (m, c)$. When making asymptotic statements in Section 3, we assume that cluster size C is fixed and the total sample size $N = MCT \to \infty$, $d = d(N), p = p(N) \to \infty$ unless restricted otherwise. The parameter β_0 is our object of interest. We consider two regimes for β_0: a low-dimensional (LD) regime with $d = O(N/\log N)$ and a high-dimensional sparse (HDS) regime $d = d(N) > N, \|\beta_0\|_0 = s_N = o(\sqrt{N/\log p})$.

Now define the *reduced form objects*

$$\begin{aligned} l_{i0}(z) &\equiv \mathbb{E}[Y_{it}|Z_{it} = z, \Phi_t] = \mathbb{E}[Y_{it}|Z_{it} = z] \\ d_{i0}(z) &\equiv \mathbb{E}[D_{it}|Z_{it} = z, \Phi_t] = \mathbb{E}[D_{it}|Z_{it} = z] \end{aligned} \tag{2.11}$$

and the corresponding residuals

$$\begin{aligned} \tilde{D}_{it} &\equiv D_{it} - d_{i0}(Z_{it}) \\ \tilde{Y}_{it} &\equiv Y_{it} - l_{i0}(Z_{it}). \end{aligned} \tag{2.12}$$

We will also use the notation $\tilde{D}_{m,t} = [\tilde{D}_{1,t}, .., \tilde{D}_{c,t}]'$ to denote a $C \times d$ dimensional matrix of residuals and $U_{m,t} \equiv [U_{1,t}, .., U_{c,t}]'$ be a $C \times 1$ dimensional vector of disturbances, corresponding to the cluster $g \in G$.

Equation (2.9) implies a linear relationship between outcome and treatment residuals:

$$\tilde{Y}_{it} = \tilde{D}_{it}\beta_0 + U_{it}, \qquad \mathbb{E}[U_{it}|\tilde{D}_{it}] = 0. \tag{2.13}$$

The structure of all the estimators is as follows. First, we construct an estimate of the first stage reduced form $\widehat{d}, \widehat{l}$ and estimate the residuals:

$$\widehat{\tilde{D}}_{i,t} = D_{i,t} - \widehat{d}(Z_{i,t}) \quad \widehat{\tilde{Y}}_{i,t} = Y_{i,t} - \widehat{l}(Z_{i,t})$$

Second, we apply off-the-shelf LD (least squares) and HDS (Lasso, and debiased Lasso) methods, designed for linear models with exactly measures regressors and outcome. Since the true values of the residuals $\tilde{P}_{it}$ and $\tilde{Y}_{it}$ are unknown, we plug in estimated residuals that are contaminated by the first-stage approximation error. Under high-level conditions on the first stage estimators, we show that the modified estimators are asymp-

[3]The choice of the different group size $C_g \asymp C$ fits our framework.

totically equivalent to their infeasible (oracle) analogs, where the oracle knows the true value of the residual. These high-level conditions are non-primitive and require verification for panel data. However, if the observable unit descriptions are sufficiently rich to assume weak sparsity of unobserved heterogeneity as in Example 3, these conditions hold for dynamic panel Lasso estimator of Kock and Tang (2016).

Example 3 (Weakly Sparse Unobserved Heterogeneity)**.** Consider the setup of Examples 1 and 2. Assume that

$$g_{i0}(Z_{it}) = g_0(Z_{it}) + \xi_i$$
$$d_{i0}(Z_{it}) = d_0(Z_{it}) + \eta_i$$

where $g_0(\cdot)$ and $d_0(\cdot)$ are weighted averages of item-specific functions $g_{i0}(\cdot), d_{i0}(\cdot)$, and η_i, ξ_i is the time-invariant heterogeneity of item i in outcome and treatment equations. Ignoring ξ_i, η_i creates heterogeneity bias in the estimate of treatment effect. To eliminate the bias, we project ξ_i, η_i on space of time-invariant observables $\bar{Z}_i$:

$$\lambda_0(\bar{Z}_i) \equiv \mathbb{E}[\xi_i|\bar{Z}_i] \text{ and } \gamma_0(\bar{Z}_i) \equiv \mathbb{E}[\eta_i|\bar{Z}_i]$$

We assume that $\bar{Z}_i$ contains sufficiently rich item descriptions such that $a_i \equiv \xi_i - \lambda_0(\bar{Z}_i)$ and $b_i = \eta_i - \gamma_0(\bar{Z}_i)$ are small. We impose weak sparsity assumption: (see, e.g. Negahban et al. (2012)).

$$\exists \quad s < \infty \quad 0 < \nu < 1 \qquad \sum_{i=1}^{N} |a_i|^\nu \leqslant s$$
$$\sum_{i=1}^{N} |b_i|^\nu \leqslant s$$

Under this assumption, the parameters to be estimated are the functions $g_0, \lambda_0, d_0, \gamma_0$ and the vectors $a = (a_1, .., a_N)'$ and $b = (b_1, .., b_N)'$. Section 4 describes an example of an l_1-penalized method by Kock and Tang (2016) that estimates these parameters with sufficient quality under sparsity assumption on $g_0, \lambda_0, d_0, \gamma_0$.

2.3 The Panel Double ML Recipe

All the methods considered in this paper will involve some variation on the Panel Double ML Recipe outlined below.

Definition 2.1 (Panel Double ML Recipe)**.** *1. Split the data into a K-fold partition* *by time index with the indices included in each partition k are given by:*

$$I_k = \{(i,t) : \lfloor T(k-1)/K \rfloor + 1 \leqslant t \leqslant \lfloor Tk/K \rfloor\}.$$

2. *For each partition k, use a first stage estimator to estimate reduced form objects $\widehat{d}_k, \widehat{l}_k$ by excluding the data from partition k (using only I_k^c).*

3. *Compute first stage residuals according to* (2.12)*. For each data point i, use the first stage estimators who's index corresponds to their partition.*

4. *Pool the first stage residuals from all partitions and Estimate $\widehat{\beta}$ by applying a second stage estimator from Section 3.2 or 3.3, depending on its regime of β_0.*

The recipe above outlines our sample splitting strategy. Step (1) partitions the data into K folds by time indices. Steps 2 and 3 describe a cross-fitting procedure that ensures that the fitted value of the treatment $\widehat{d}_{it} = \widehat{d}_i(Z_{it})$ and outcome $\widehat{l}_{it} = \widehat{l}_i(Z_{it})$ is uncorrelated with the true residuals $\tilde{D}_{it}, \tilde{Y}_{it}$. Step (4) specifies our second stage estimation strategy. In the LD regime, we use ordinary least squares as our second stage estimator. In the HDS regime, we instead suggest Lasso for estimation and debiased Lasso for inference. [45]

3 Theoretical Results

In this section we establish the asymptotic theory of our estimators under high-level conditions, whose plausibility we discuss in Section 4. We shall use empirical process notation, adapted to panel clustered setting.

$$\mathbb{E}_N f(x_{it}) \equiv \frac{1}{N} \sum_{(i,t)} f(x_{it})$$

and

$$\mathbb{G}_N f(x_{it}) \equiv \frac{1}{\sqrt{N}} \sum_{(i,t)} (f(x_{it}) - \mathbb{E} f(x_{it}))$$

[4]This cross-fitting procedure corresponds to DML2 estimator of Chernozhukov et al. (2016) . A more popular alternative, known as DML1, requires computation of a separate estimator of β on each partition k and returns the average over K final estimators. But Remark 3.1 of Chernozhukov et al. (2016), shows that DML2 has a finite sample advantage over DML1. In addition it is more computationally efficient for large data sets. For this reason, all code and analysis in this project will use DML2, but similar results could be obtained for DML1 or other similar sample splitting and cross-fitting patterns.

[5]The goal of partition by time is to ensure that every fold contains sufficient number of observations for each item i. Alternative splitting procedures that output balanced partitions are also acceptable.

3.1 High-Level Assumptions

In this section we provide high-level restrictions of our estimators. They consist on the assumptions on the first stage estimators (Assumption 3.1 and 3.2), and standard identifiability (3.4) and light tails conditions (3.5) on the true outcome and treatment residuals $\tilde{Y}, \tilde{D}$. In addition to that, we also assume Law of Large Numbers for matrices that are sample average of a stationary process.

With high probability, they belong to a realization set D_N and L_N that are properly shrinking neighborhoods of $d_0(\cdot), l_0(\cdot)$, constrained by the following assumptions.

Assumption 3.1 (Small Bias Condition)**.** *Define the following rates:*

$$\boldsymbol{m}_N \equiv \sup_{d \in \mathcal{D}_N} \max_{1 \leqslant j \leqslant d} (\mathbb{E}(d_j(Z) - d_{0,j}(Z))^2)^{1/2}$$

$$\boldsymbol{l}_N \equiv \sup_{l \in \mathcal{L}_N} (\mathbb{E}(l(Z) - l_0(Z))^2)^{1/2}$$

$$\exists D, L \quad \sup_{d \in D_N} \max_{1 \leqslant m \leqslant d} |d_m(Z_{it})| < D \quad \sup_{l \in L_N} |l(Z_{it})| < L$$

We assume that:

$$\boldsymbol{l}_N = o(N^{-1/4}), \quad \boldsymbol{m}_N = o(N^{-1/4})$$

We shall refer to $\mathbf{m}_N$ as treatment rate and to $\mathbf{l}_N$ as the outcome rate.

Assumption 3.2 (Concentration)**.** *Let the centered out-of-sample mean squared error of treatment and outcomes exhibit a bound:*

$$\sqrt{N}\lambda_N \equiv \max_{1 \leqslant j \leqslant d} |\mathbb{G}_N(\widehat{d}_j(Z_{it}) - d_{0,j}(Z_{it}))^2| \lesssim_P o_P(1)$$

$$\sqrt{N}\lambda_N \equiv \max_{1 \leqslant j \leqslant d} |\mathbb{G}_N(\widehat{d}_j(Z_{it}) - d_{0,j}(Z_{it}))(\widehat{l}(Z_{it}) - l_0(Z_{it}))| \lesssim_P o_P(1)$$

Assumption 3.3 (LLN for Matrices for m.d.s)**.** *Let $(\tilde{D}_{mt})_{mt=(1,1)}^{G,T}$ be a stationary process with bounded realizations, whose dimension $d = d(N)$ grows. Let*

$$Q \equiv \mathbb{E}\tilde{D}'_{mt}\tilde{D}_{mt}$$

$$\|\mathbb{E}_N \tilde{D}'_{mt}\tilde{D}_{mt} - Q\| \lesssim_P \sqrt{\frac{d \log N}{N}}$$

Assumption 3.3 has been shown for the case of i.i.d case by (Rudelson (1999)). Combining his arguments with blocking methods, we can show that that it continues to hold under exponential mixing condition for panel data.

Assumption 3.4. *Let* $Q \equiv \mathbb{E}\tilde{D}'_{mt}\tilde{D}_{mt}$ *denote population covariance matrix of treatment residuals. Assume that* $\exists 0 < C_{\min} < C_{\max} < \infty$ *s.t.* $C_{\min} < \min \operatorname{eig}(Q) < \max \operatorname{eig}(Q) < C_{\max}$.

Assumption 3.5. *The following conditions hold.*

(1) $\|\tilde{D}_{mt}\| \leqslant D < \infty$

(2) *Lindeberg Condition:* $\mathbb{E}\|U_{mt}U'_{mt}\|1_{\|U_{mt}U'_{mt}\| > M} \to 0, M \to \infty$

Assumption 3.4 requires that the treatments $\tilde{D}_{mt}$ are not too collinear, allowing identification of the treatment effect β_0. Assumption 3.5 imposes technical conditions for asymptotic theory. Since a bounded treatment $\tilde{D}_{mt}$ is a plausible condition in practice, we impose it to simplify the analysis. In addition, we require the disturbances U_{mt} to have light tails as stated in Lindeberg condition.

3.2 Low Dimensional Treaments

In this section we consider Low-Dimensional (LD) case: $d = o(N/\log N)$. We define Orthogonal Least Squares and state its asymptotic theory.

Definition 3.1 (Orthogonal Least Squares)**.** *Given first stage estimators* $\widehat{d}, \widehat{l}$, *define Orthogonal Least Squares estimator:*

$$\begin{aligned}
\widehat{\beta} &= \mathbb{E}_N[D_{it} - \widehat{d}(Z_{it})][D_{it} - \widehat{d}(Z_{it})]'])^{-1}\mathbb{E}_N[D_{it} - \widehat{d}(Z_{it})][Y_{it} - \widehat{l}(Z_{it})]'].\\
&\equiv \mathbb{E}_N(\widehat{\tilde{D}}_{it}\widehat{\tilde{D}}'_{it})^{-1}\mathbb{E}_N(\widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it})\\
&\equiv \widehat{Q}^{-1}\mathbb{E}_N(\widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it}),
\end{aligned}$$

where the second and third lines implicitly define estimators of residualized vectors and matrices.

Orthogonal Least Squares is our first main estimator. As suggested by its name, it performs ordinary least squares on estimated treatment and outcome residuals, that are approximately orthogonal to the realizations of the controls. In case the dimension d is fixed, it coincides with Double Machine Learning estimator of Chernozhukov et al. (2016). Allowing dimension $d = d(N)$ to grow with sample size is a novel feature of this paper.

Assumption 3.6 (Dimensionality Restriction)**.** *(a)* $d = o\left(\frac{N}{\log N}\right)$

(b) $d\boldsymbol{m}_N\|\beta_0\| = o(1/\sqrt{N}\boldsymbol{m}_N)$

Assumption 3.6 imposes growth restrictions on the treatment dimension . The first restriction $d = o\left(\frac{N}{\log N}\right)$ ensures that covariance matrix of treatment residuals converges to the population covariance matrix. This is the weakest growth condition available in least squares series literature. The second statement $d\mathbf{m}_N\|\beta_0\| = o(1/\sqrt{N}\mathbf{m}_N)$ restricts the growth of d relative to the quality of the first stage treatment estimator.

Theorem 3.1 (Orthogonal Least Squares)**.** *Let Assumptions 3.3, 3.4, 3.5, and 3.6 hold.*

(a)

$$\|\widehat{\beta} - \beta\|_2 \lesssim_P \sqrt{\frac{d}{N}} + d\boldsymbol{m}_N^2\|\beta_0\| + \boldsymbol{l}_N\sqrt{d}\boldsymbol{m}_N + \sqrt{d/N}\sqrt{d}\boldsymbol{m}_N\|\beta_0\|$$

Let Assumption 3.1 hold for the statements below. Then,

$$\|\widehat{\beta} - \beta\|_2 \lesssim_P \sqrt{\frac{d}{N}}$$

(b) For any $\alpha \in \mathcal{S}^{d-1}$

$$\sqrt{N}\alpha'(\widehat{\beta} - \beta) = \alpha' Q^{-1}\mathbb{G}_N\tilde{D}_{it}U_{it} + R_{1,N}(\alpha)$$

where $R_{1,N}(\alpha) \lesssim_P \sqrt{N}\sqrt{d}\boldsymbol{m}_N\boldsymbol{l}_N + \sqrt{N}d\boldsymbol{m}_N^2\|\beta_0\| + \sqrt{d}\boldsymbol{l}_N + d\boldsymbol{m}_N\|\beta_0\|$

(c) Denote

$$\Omega = Q^{-1}\mathbb{E}\tilde{D}'_{mt}U_{mt}U'_{mt}\tilde{D}_{mt}Q^{-1}$$

Then, $\forall t \in \mathcal{R}$ and for any $\alpha \in \mathcal{S}^{d-1}$

$$\lim_{N\to\infty}\left|\mathrm{P}\left(\frac{\sqrt{N}\alpha'(\widehat{\beta} - \beta_0)}{\|\alpha'\Omega\|^{1/2}} < t\right) - \Phi(t)\right| = 0. \tag{3.1}$$

Theorem 3.1 is our first main result. Under small bias condition, OLS attains oracle rate and has oracle asymptotic linearity representation. Under Lindeberg condition, OLS is asymptotically normal with asymptotic variance Ω, which can be consistently estimated by White cluster-robust estimator

$$\widehat{\Omega} \equiv \widehat{Q}^{-1}\mathbb{E}_N\left[\widehat{\tilde{D}}'_{mt}\widehat{U}_{mt}\widehat{U}'_{mt}\widehat{\tilde{D}}_{mt}\right]\widehat{Q}^{-1},$$

where $\widehat{U}_{mt} \equiv (\widehat{\tilde{Y}}_{mt} - \widehat{\tilde{D}}'_{mt}\widehat{\beta})$. The asymptotic variance Ω is not affected by first stage estimation.

3.3 High Dimensional Sparse Treatments

In this section we consider a High-Dimensional Sparse case $d = d(N) > N$. We state a finite-sample bound on the rate of Orthogonal Lasso. We define a Debiased Orthogonal Lasso and provide its asymptotic linearization. This allows to conclude about its Gaussian approximation of a single coefficient (Central Limit Theorem) and many coefficients (Central Limit Theorem in High Dimension).

3.3.1 Lasso in High Dimensional Sparse Case

Here we introduce the basic concepts of high-dimensional sparse literature. We allow for our parameter of interest $\beta_0 \in \mathcal{R}^d$ to be high-dimensional ($d = d(N) > N$) sparse. Let s_N be the sparsity of β_0: $\|\beta_0\|_0 = s_N$. Let the set of active regressors be $T \equiv \{j \in \{1, 2, ..., d\}, \text{ s.t. } \beta_{0,j} \neq 0\}$. For a given $\bar{c} \geqslant 1$, let the set $\mathcal{RE}(\bar{c}) \equiv \{\delta \in \mathcal{R} : \|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \delta \neq 0\}$ be a restricted subset of $\mathcal{R}^d$.

Let the sample covariance matrix of true residuals be $\tilde{Q} \equiv \mathbb{E}_N \tilde{D}'_{mt}\tilde{D}_{mt}$. Let the in-sample prediction norm be: $\|\delta\|_{2,N} = (\mathbb{E}_N(\tilde{D}'_{mt}\delta)^2)^{1/2}$. Define Restricted Eigenvalue of covariance matrix of true residuals:

$$\kappa(\tilde{Q}, T, \bar{c}) := \min_{\mathcal{RE}(\bar{c})} \frac{\sqrt{s}\delta'\tilde{Q}\delta}{\|\delta_T\|_1} = \min_{\mathcal{RE}(\bar{c})} \frac{\sqrt{s}\|\delta\|_{2,N}}{\|\delta_T\|_1} \tag{3.2}$$

Assumption 3.7 (RE($\bar{c}$))**.** *For a given $\bar{c} > 1$, Restricted Eigenvalue is bounded from zero:*

$$\kappa(\tilde{Q}, T, \bar{c}) > 0$$

Assumption 3.7 has been proven for i.i.d. case by Rudelson and Zhou (2013). We assume that it holds under plausible weak dependence conditions.

Definition 3.2 (Orthogonal Lasso)**.** *Let $\lambda > 0$ be a constant to be specified.*

$$\begin{aligned} \widehat{Q}(b) &= \mathbb{E}_N(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}'_{it}b)^2 \\ \widehat{\beta}_L &= \arg\min_{b \in \mathcal{R}^k} \widehat{Q}(b) + \lambda\|\beta\|_1 \end{aligned} \tag{3.3}$$

Orthogonal Lasso is our second main estimator. It performs l_1 penalized least squares minimization using the outcome residual $\widehat{\tilde{Y}}_{it}$ as dependent variable and the treatment residuals $\widehat{\tilde{D}}_{it}$ as covariates. The regularization parameter λ controls the noise of the problem. Its choice is described below.

We summarize the noise with the help of two metrics. The first one, standard for Lasso literature, is the maximal value of the gradient coordinate of $\widehat{Q}(\cdot)$ at the true value β_0

$$\|S\|_\infty \equiv 2\|\mathbb{E}_N \widehat{\tilde{D}}'_{i,t}[\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}'_{it}\beta_0]\|_\infty$$

The second one is the maximal entry-wise difference between covariance matrices of true and estimated residuals

$$q_N \equiv \max_{1 \leqslant m,j \leqslant d} |\widehat{Q} - \tilde{Q}|_{m,j}.$$

It summarizes the noise in the covariates due to first-stage approximation error. Both quantities can be controlled by the first stage convergence and concentration rates in Assumption 3.1 and 3.2, applying Azouma-Hoeffding maximal inequality.

To control the noise, the parameter λ should satisfy: $\lambda \geqslant c\|S\|_\infty$. Asymptotically, for this to happen it suffices for λ to be determined by the following condition.

Condition 3.1 (OPT). *Fix a constant $c > 0$ to be specified. Let λ be chosen as follows:*

$$\lambda = c\big[\boldsymbol{l}_N \boldsymbol{m}_N \vee s\boldsymbol{m}_N^2 \vee \lambda_N\big]$$

Theorem 3.2 (Orthogonal Lasso). *Suppose $\exists c > 1$ such that Assumption RE($\bar{c}$) holds for $\bar{c} = (c+1)/(c-1)$. Let N be sufficiently large such that*

$$q_N(1+\bar{c})^2 s/\kappa(\tilde{Q}, T, \bar{c})^2 < 1/2$$

(a) If $\lambda \geqslant c\|S\|_\infty$, $\|\widehat{\beta}_L - \beta_0\|_{N,2} \leqslant 2\lambda \frac{\sqrt{s}}{\kappa(\tilde{Q},T,\bar{c})}$

(b) If $\lambda \geqslant c\|S\|_\infty$ and RE($2\bar{c}$) holds, $\|\widehat{\beta}_L - \beta_0\|_1 \leqslant 2\lambda \frac{s}{\kappa(\tilde{Q},T,2\bar{c})\kappa(\tilde{Q},T,\bar{c})}$

(c) Suppose λ is as in Condition OPT. As N grows,

$$\|\widehat{\beta}_L - \beta_0\|_{N,2} \lesssim_P \sqrt{s}\big[\lambda_N \vee \boldsymbol{l}_N \boldsymbol{m}_N \vee s\boldsymbol{m}_N^2 \vee \bar{\sigma}\sqrt{\frac{\log d}{N}}\big]$$

(d)

$$\|\widehat{\beta}_L - \beta_0\|_1 \lesssim_P s\big[\lambda_N \vee \boldsymbol{l}_N \boldsymbol{m}_N \vee s\boldsymbol{m}_N^2 \vee \bar{\sigma}\sqrt{\frac{\log d}{N}}\big]$$

This is our second main result in the paper. The statements (a,b) establish finite-sample bounds on $\|\widehat{\beta}_L - \beta_0\|_{N,2}$ and $\|\widehat{\beta}_L - \beta_0\|_1$ for a sufficiently large N. The statements (c,d) establish asymptotic bounds on $\|\widehat{\beta}_L - \beta_0\|_{N,2}$ and $\|\widehat{\beta}_L - \beta_0\|_1$. Under small bias and concentration conditions (Assumptions 3.1 and 3.2), the asymptotic bounds coincide with respective bounds of oracle lasso (see e.g., Belloni and Chernozhukov (2013)).

The total bias of $\widehat{\beta}_L$ scales with sparsity s, not the total dimension d. This is a remarkable property of l_1 penalization. It forces $\widehat{\beta} - \beta_0$ to belong to $\mathcal{RE}(\bar{c})$, where the total bias scales in proportion to the bias accumulated on the active regressors. This ensures convergence of Lasso in the regime $d = d(N) > N$.

Remark 3.1 (Comparison of Baseline Lasso and Orthogonal Lasso in a Linear Model)**.** Suppose the function $g(Z)$ in Equation 2.9 is a linear and sparse in Z:

$$g(Z) = Z'\gamma, \quad \|\gamma\|_0 = s_{\gamma,N} = s_\gamma < N$$

and the treatment reduced form is linear and sparse in Z

$$d_k(Z) = Z'\delta_k, \quad \|\delta\|_0 = s_\delta < s_\gamma, k \in \{1, 2, ..., d\}$$

In this problem, a researcher has a choice between running one-stage Baseline Lasso, where the covariates consist of the treatments D and the controls Z, and the Orthogonal Lasso, where the controls are partialled out first.

Let us describe an empirically relevant scenario in which Orthogonal Lasso has a faster rate. Let the complexity of treatments be smaller than the complexity of the controls:

$$\frac{s^2 \log d}{s_\gamma^2 \log p} = o(1)$$

$\widehat{\beta}_L$. Define Baseline Lasso as

$$\begin{aligned} \widehat{Q}(\beta, \gamma) &= \mathbb{E}_N (Y_{it} - D_{it}'\beta - Z_{it}'\gamma)^2 \\ \widehat{\beta}_B &= \arg\min_{(\beta,\gamma)} \widehat{Q}(\beta, \gamma) + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1 \end{aligned} \tag{3.4}$$

In case $\mathbb{E} D_{it} Z_{it}' \neq 0$, estimation error of $\widehat{\gamma}$ has a first order effect on the gradient of $\widehat{Q}(\beta, \gamma)$ with respect to β, and therefore, the bias of $\widehat{\beta} - \beta_0$ itself. Therefore, an upper bound on $\|\widehat{\beta}_B - \beta_0\|_1$ of the baseline Lasso

$$\|\widehat{\beta}_B - \beta_0\|_1 \lesssim_P \|\widehat{\gamma} - \gamma_0\|_1 \lesssim_P \sqrt{\frac{s_\gamma^2 \log p}{N}} \tag{3.5}$$

By contrast,

$$\|\widehat{\beta}_L - \beta_0\|_1 \lesssim_P \frac{s_\gamma s_\delta \log p}{N} + \frac{s s_\delta \log p}{N} + \sqrt{\frac{s^2 \log d}{N}}\bar{\sigma}$$

Therefore, the estimation error of $\widehat{\gamma}$ has a second order effect on the rate of $\widehat{\beta}_L$. Since the complexity of controls is larger than the complexity of the treatment, the error of $\widehat{\gamma}$ determines the rate of both estimators. Reducing its impact on $\widehat{\beta}_L$ from first order in $\widehat{\beta}_B$ to second order in $\widehat{\beta}_L$ by projecting the outcome and treatments on the orthocomplement of Z gives rate improvement.

3.3.2 Inference in High Dimensional Sparse case

After we have established the properties of Orthogonal Lasso, we propose a debiasing strategy that will allow us to conduct inference in HDS case. We will employ the following assumption.

Assumption 3.8 (Approximate sparsity of Q^{-1}). *Let $Q = \mathbb{E}\tilde{D}'_{mt}\tilde{D}_{mt}$ be the population covariance matrix. Assume that there exists a sparse matrix $M = [m_1, ..., m_d]'$:*

$$m_0 := \max_{1 \leqslant j \leqslant d} \|m_j\|_0 = o(1/[\sqrt{N}\boldsymbol{m}_N^2 + \sqrt{N}\boldsymbol{m}_N \boldsymbol{l}_N]) \tag{3.6}$$

that is a good approximation for inverse of matrix Q^{-1}:

$$\|Q^{-1} - M\|_\infty \lesssim \sqrt{\frac{\log d}{N}}$$

Assumption 3.8 restricts a pattern of correlations between treatment residuals. Examples of matrices Q satisfying Assumption 3.8 include Toeplitz, block diagonal, and band matrices. In addition, if dimension d and the rate $\mathbf{m}_N$ satisfy $dN\mathbf{m}_N^2 = o(1)$, any invertible matrix Q satisfies Assumption 3.8 with $M = Q^{-1}$.

Condition 3.2 (Approximate Inverse of $\widehat{Q}$). *Let a be a large enough constant, and let $\mu_N \equiv a\sqrt{\frac{\log d}{N}}$. Let $\widehat{Q} = \mathbb{E}_N \widehat{\tilde{D}}'_{mt}\widehat{\tilde{D}}_{mt}$. A matrix $M = [m_1, ..., m_d]'$ approximately inverts $\widehat{Q}$ if for each row $j \in \{1, 2, .., d\}$ $\|\widehat{Q}m_j - e_j\|_\infty \leqslant \mu_N$*

Definition 3.3 (Constrained Linear Inverse Matrix Estimation). *Let $M = M(\widehat{Q}) = [m_1, ..., m_d]'$ solve*

$$m_j^* = \arg\min \|m_j\|_1 \ \textit{s.t.} \ \|\widehat{Q}m_j - I_d\|_\infty \leqslant \mu_N \forall j \in \{1, 2, .., d\}$$

We will refer to $M(\widehat{Q})$ as CLIME.

Condition 3.2 introduces a class of matrices that approximately invert the covariance matrix $\widehat{Q}$ of estimated residuals. By Lemma C.1, this class contains all matrices that approximately invert Q (including precision matrix Q^{-1}), and therefore is non-empty. Within this class, we focus on the matrix with the smallest first norm, which we refer to as CLIME of $\widehat{Q}$. Due to proximilty of $\widehat{Q}$ to Q, CLIME of $\widehat{Q}$ consistently estimates the precision matrix Q^{-1} at rate $\sqrt{\frac{\log d}{N}}$ in elementwise norm.

Once we introduced an estimate of Q^{-1}, let us explain the debiasing strategy in the oracle case. Let $\widehat{\beta}_L$ be the (oracle) Orthogonal Lasso estimate of the treatment effect β_0 and $\tilde{U}_{it} := \tilde{Y}_{it} - \tilde{D}'_{it}\widehat{\beta}_L$ be oracle Lasso residual. Let $j \in \{1, 2, .., d\}$ be the treatment effect of interest and m_j be the j'th row of the approximate inverse M. Recognize that Lasso residual $\tilde{U}_{it}$ consists of the true residual U_{it} and the fitting error $\tilde{D}'_{it}(\beta_0 - \widehat{\beta})$

$$\tilde{U}_{it} = U_{it} + \tilde{D}'_{it}(\beta_0 - \widehat{\beta})$$

Therefore, the correction term

$$\sqrt{N} m'_j \mathbb{E}_N \tilde{D}_{it}\tilde{U}_{it} = \underbrace{m'_j \mathbb{G}_N \tilde{D}_{it} U_{it}}_{S_j} + \underbrace{\sqrt{N} m'_j \tilde{Q}(\beta_0 - \widehat{\beta})}_{\Delta}$$

where S_j is approximately normally distributed and Δ is the remainder. By definition of approximate inverse ($m'_j\tilde{Q} \approx e_j$), Δ offsets the bias of Orthogonal Lasso up to first-order: $\Delta \approx \sqrt{N}(\beta_{0,j} - \widehat{\beta}_{L,j}) + o_P(1)$. Adding this correction term to the original Lasso estimate $\widehat{\beta}_{L,j}$ returns unbiased, asymptotically normal estimate:

$$\sqrt{N}[\widehat{\beta}_{DOL,j} - \beta_{0,j}] = \sqrt{N}[m'_j \mathbb{E}_N \tilde{D}_{it}\tilde{U}_{it} + (m'_j\tilde{Q} - e_j)'(\beta_0 - \widehat{\beta}_L)] = S_j + o_P(1/\sqrt{N}) \quad (3.7)$$

Let us see that the debiasing strategy is compatible with first stage error. In presence of the latter, Equation 3.7 becomes

$$\sqrt{N}[\widehat{\beta}_{DOL,j} - \beta_{0,j}] = S_j + \underbrace{\sqrt{N} m'_j \mathbb{E}_N [\widehat{\tilde{D}}_{it}[U_{it} + R_{it}] - \tilde{D}_{it}U_{it}]}_{\Delta^{fs}} + o_P(1/\sqrt{N})$$

Since the true residual is mean independent from first-stage approximation error, the bias of the vector $\mathbb{E}_N[\widehat{\tilde{D}}_{it}[U_{it} + R_{it}] - \tilde{D}_{it}U_{it}]$ is second-order. Small bias and concentration assumptions (3.1 and 3.2) ensure that worst-case first-stage error is small

$$\max_{1 \leqslant j \leqslant d} |\mathbb{E}_N[\widehat{\tilde{D}}_{it}[U_{it} + R_{it}] - \tilde{D}_{it}U_{it}]| = \sqrt{\frac{\log d}{N}} + \lambda_N + \mathbf{m}_N^2 s \vee \mathbf{l}_N \mathbf{m}_N$$

To conclude Δ^{fs} is small, let us see that the rows of matrix M are approximately sparse. Each row m_j can be approximated by a sparse vector m_j^0 of sparsity m_0 such that $\|m_j - m_j^0\|_1 \lesssim 2m_0\sqrt{\frac{\log d}{N}}$. The sparsity of m_j^0 suffices to conclude

$$\begin{aligned}
\Delta^{fs} &= \sqrt{N}[m_j^0 + m_j - m_j^0]\mathbb{E}_N[\widehat{\tilde{D}}_{it}[U_{it} + R_{it}] - \tilde{D}_{it}U_{it}] \\
&= o_P(m_0[1 + \sqrt{\frac{\log d}{N}}][\sqrt{\frac{\log d}{N}} + \lambda_N + \mathbf{m}_N^2 s \vee \mathbf{l}_N \mathbf{m}_N]) \\
&= o_P(1)
\end{aligned}$$

After we have explained the debiasing strategy, we proceed to definition of Debiased Orthogonal Lasso.

Definition 3.4 (Debiased Orthogonal Lasso). *Let M be CLIME of $\widehat{Q}$. Then,*

$$\widehat{\beta}_{DOL} \equiv M\mathbb{E}_N\widehat{\tilde{D}}_{it}(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\widehat{\beta}_L) + \widehat{\beta}_L \tag{3.8}$$

In case dimension $d = o(1/[\sqrt{N}\mathbf{m}_N^2 + \sqrt{N}\mathbf{m}_N\mathbf{l}_N])$ grows at a small rate, Assumption 3.8 is always satisfied: one can pick $M = (\widehat{Q} + \gamma I_d)^{-1}$ to be regularized inverse of $\widehat{Q}$. This gives rise to a simple asymptotically normal estimator we refer to as Debiased Orthogonal Ridge.

Definition 3.5 (Debiased Orthogonal Ridge). *Let $d = o(1/[\sqrt{N}\boldsymbol{m}_N^2 + \sqrt{N}\boldsymbol{m}_N\boldsymbol{l}_N])$. Let $\gamma > 0, \gamma \lesssim \sqrt{\frac{\log d}{N}}$ be a regularization constant. Define debiased Ridge estimator by choosing $M \equiv (\widehat{Q} + \gamma I_d)^{-1}, \gamma \geqslant 0$ as a regularized inverse in Debiased Orthogonal Lasso:*

$$\widehat{\beta}_{Ridge} \equiv M\mathbb{E}_N\widehat{\tilde{D}}_{it}(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\widehat{\beta}_L) + \widehat{\beta}_L \tag{3.9}$$

Theorem 3.3 (Debiased Orthogonal Lasso and Debiased Orthogonal Ridge). *Let M be chosen as in Definition 3.4 or 3.5. Let Assumptions 3.5, 3.4, 3.1, 3.8 hold.*

(a) For any

$$j \in \{1, 2, .., d\}, \quad \sqrt{N}(\widehat{\beta}_j - \beta_{j,0}) = \mathbb{G}_N M\tilde{D}_{it}'U_{it} + R_{1,N,j}$$

where $\sup_{1\leqslant j\leqslant d}|R_{1,N,j}| = o_P(1)$

(b) Denote

$$\Omega = M\mathbb{E}\tilde{D}_{mt}'U_{mt}U_{mt}'\tilde{D}_{mt}M'$$

Then, $\forall t \in \mathcal{R}$

$$\lim_{N\to\infty} \left| \mathrm{P}\left(\frac{\sqrt{N}(\widehat{\beta}_j - \beta_{j,0})}{\Omega_{jj}} < t \right) - \Phi(t) \right| = 0. \tag{3.10}$$

Theorem 3.3 is our third main result. Under Assumptions 3.1 and 3.2, for all $j \in \{1, 2, ..d\}$ $\widehat{\beta}_{DOL,j}$ and $\widehat{\beta}_{Ridge,j}$ are asymptotically linear. Under Lindeberg condition, each of them is asymptotically normal with oracle covariance matrix

$$\Omega = M\mathbb{E}\tilde{D}'_{mt}U_{mt}U'_{mt}\tilde{D}_{mt}M'.$$

This matrix can be consistently estimated by

$$\widehat{\Omega} = M\mathbb{E}_N\widehat{\tilde{D}}'_{mt}\widehat{U}_{mt}\widehat{U}'_{mt}\widehat{\tilde{D}}_{mt}M$$

where $\widehat{U}_{mt} \equiv \widehat{\tilde{Y}}_{mt} - \widehat{\tilde{D}}'_{mt}\widehat{\beta}_L$

In absence of first-stage estimation error (oracle case), any matrix M that approximately inverts $\widehat{Q}$ can be used to construct a debiased, asymptotically normal estimator. For example, Javanmard and Montanari (2014) suggest variance minimizing choice of M in oracle case. In contrast to their design, we have the first-stage bias to control for. We achieve our goal by choosing sparse M.

Theorem 3.4 (Gaussian Approximation and Simultaneous Inference on Many Coefficients)**.** *Suppose the conditions in the previous theorem hold, and* $\sqrt{N}\boldsymbol{m}_N\boldsymbol{l}_N \vee \sqrt{N}\boldsymbol{m}_N^2 = o(1/\log d)$ *holds in addition. Then, we have the following Gaussian approximation result*

$$\sup_{R\in\mathcal{R}} |\mathrm{P}(\ (\text{diag }\Omega)^{-1/2}\sqrt{N}(\widehat{\beta}_{DOL} - \beta) \in R) - P(Z \in R)| \to 0$$

where $Z \sim N(0, C)$ *is a ceneterd Gaussian random vector with covariance matrix* $C = (\text{diag }\Omega)^{-1/2}\Omega(\text{diag }\Omega)^{-1/2}$ *and* $\mathcal{R}$ *denotes the collection of cubes in* $\mathbb{R}^d$ *centered at the origin. Moreover, replacing* C *with* $\widehat{C} = (\text{diag }\widehat{\Omega})^{-1/2}\widehat{\Omega}(\text{diag }\widehat{\Omega})^{-1/2}$ *we also have for* $\tilde{Z} \mid \widehat{C} \sim N(0, \widehat{C})$

$$\sup_{R\in\mathcal{R}} |\mathrm{P}((\text{diag }\widehat{\Omega})^{-1/2}\sqrt{N}(\widehat{\beta}_{DOL} - \beta) \in R) - P(\tilde{Z} \in R \mid \widehat{C})| \to_P 0.$$

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

$$\mathrm{P}(\beta_{0,j} \in [\widehat{\beta}_{DOL,j} \pm c_{1-\xi}\widehat{\Omega}_{jj}^{1/2}N^{-1/2}], j = 1, 2..., d) \to (1-\xi).$$

This first result follows as a consequence of the Gaussian approximation result of Zhang and Wu (2015) for time series, and the second by the Gaussian comparison inequalities of Chernozhukov et al. (2015), and Theorem 4.2 establishes a uniform bound on $\|\widehat{\Omega}_{jj} - \Omega_{jj}\|_\infty$. As in Chernozhukov et al. (2013a), the Gaussian approximation results above could be used not only for simultaneous confidence bands but also for multiple hypothesis testing using the step-down methods.

4 Sufficient Conditions for First Stage Estimators

In this section we describe the plausibility of first stage conditions, discussed in Section 3.

4.1 Affine Structure ofTreatments

Here we discuss a special structure of treatments that allows us to simplify high level assumptions of Section 3 and reduce computational time. Suppose there exists an observable base treatment variable P and a collection of known maps $\{\Omega^k = \Omega^k(Z) : \mathcal{Z}^p \to \mathcal{R}^{d_p}\}$, such that every treatment $D^k = \Omega^k(Z)'P, k \in \{1, 2, .., d\}$ is an affine transformation of the base treatment. In case $d > 1$, an estimate of the reduced form of the base treatment $p_0(Z)$ can be used to construct an estimate $\widehat{d}_i(Z) = \Omega^k(Z_{it})\widehat{p}_i(Z_{it})$. Lemma 4.1 shows the simplification of the conditions.

Lemma 4.1 (Affine Treatments)**.** *Suppose a first-stage estimator of $p_{i0}(Z_{it})$, denoted by $\widehat{p}(Z)$, belongs w.h.p to a realization set P_N constrained by the rates $\boldsymbol{m}_N, \boldsymbol{r}_N, \lambda_N$. Then, an estimator $\widehat{D}(Z)$ of $d_{i0}(Z)$, defined as*

$$\widehat{D}^k(Z) \equiv \Omega^k(Z)\widehat{p}(Z), \quad k \in \{1, 2, .., d\}$$

belongs to a realization set D_N that contains the true value of $d_0(Z)$ and achieves the same treatment rate, mean square rate, and concentration rate as original $\widehat{p}(Z)$, regardless of dimension d.

Lemma 4.1 shows that Assumptions 3.1 and 3.2 about the technical treatment D_{it} hold if and only if Assumptions 3.1 and 3.2 hold for the base treatment P_{it}. Therefore, the treatment rate and concentration rates are now free from dimension d. Both Examples 1 and 2 have affine treatment structure.

4.2 Dependence Structure of Observations

Here we discuss special structure of individual heterogeneity that allows us to verify Assumptions 3.1 and 3.2. For the sake of completeness, we also provide example of cross sectional data.

Example 4 (Cross-Sectional Data)**.** Let $T = 1$ and $(W_i)_{i=1}^I = (Y_i, D_i, Z_i)_{i=1}^I$ be an i.i.d sequence. Then, small bias condition (Assumption 3.1) is achievable by many ML methods under structured assumptions on the nuisance parameters, such l_1 penalized methods in sparse models (Bühlmann and van der Geer (2011), Belloni et al. (2016b)) and L_2 boosting in sparse linear models (Luo and Spindler (2016)), and other methods for classes of neural nets, regression trees, and random forests (Wager and Athey (2016)). The bound on centered out-of-sample mean squared error in Assumption 3.2 follows from Hoeffding inequality.

Example 5 (Panel Data (No Unobserved Heterogeneity))**.** Let $\{\{W_{it}\}_{t=1}^T\}_{i=1}^I$ be an i.i.d sequence. Let the reduced form of treatment and outcome be:

$$d_{i0}(Z_{it}) = d_0(Z_{it})$$
$$l_{i0}(Z_{it}) = l_0(Z_{it})$$

In other words, there is no unobserved unit heterogeneity .Then, small bias condition is achieved by many ML methods. Assumption 3.2 holds under plausible β-mixing conditions on Z_{it} (see e.g. Chernozhukov et al. (2013b).)

Remark 4.1 (Partialling out individual heterogeneity)**.** Partialling out unobserved item heterogeneity may be a desirable step in case one wants to model it in a fully flexible way. However, applying the described estimators on the partialled out data leads to the loss of their oracle properties in a dynamic panel model. In particular, plug-in estimators of asymptotic covariance matrices of Orthogonal Least Squares and debiased Ridge will be inconsistent.[6]

4.2.1 Weakly Sparse Unobserved Heterogeneity

Consider the setup of Example 2. Let the sales and price reduced form be

$$l_{i0}(Z_{it}) = l_0(Z_{it}) + \xi_i$$
$$p_{i0}(Z_{it}) = p_0(Z_{it}) + \eta_i$$

[6] Partialling out unobserved heterogeneity in a high-dimensional sparse model was considered in Belloni et al. (2016a).

where the controls $Z_{it} \equiv [Y'_{i,t-1}, Y'_{i,t-2}, ..., Y'_{i,t-L}, P'_{i,t-1}, P'_{i,t-2}, ..., P'_{i,t-L}, X_{it}]$ include all pre-determined and exogeneous observables observed for item i, relevant for predicting the reduced form [7], and ξ_i, η_i is unobserved time-invariant heterogeneity of product i. Following Example 3, we project ξ_i, η_i on space of time-invariant observables $\bar{Z}_i$:

$$\lambda_0(\bar{Z}_i) \equiv \mathbb{E}[\xi_i|\bar{Z}_i] \text{ and } \gamma_0(\bar{Z}_i) \equiv \mathbb{E}[\eta_i|\bar{Z}_i]$$

We assume that $\bar{Z}_i$ contains sufficiently rich product descriptions such that $a_i \equiv \xi_i - \lambda_0(\bar{Z}_i)$ and $b_i = \eta_i - \gamma_0(\bar{Z}_i)$ are small. We impose weak sparsity assumption: (see, e.g. Negahban et al. (2012)).

$$\exists \quad s < \infty \quad 0 < \nu < 1 \qquad \begin{aligned} \sum_{i=1}^{N} |a_i|^\nu \leqslant s \\ \sum_{i=1}^{N} |b_i|^\nu \leqslant s \end{aligned}$$

Under this assumption, the parameters to be estimated are the functions $l_0, \lambda_0, p_0, \gamma_0$ and the vectors $a = (a_1, .., a_N)'$ and $b = (b_1, .., b_N)'$. Assume that the price and sales reduced form is a linear function of observables

$$\begin{aligned} l_0(Z_{it}) &= \mathbb{E}[Y_{it}|Z_{it}] = [Z_{it}, \bar{Z}_i]'\gamma^Y + a_i \\ p_0(Z_{it}) &= \mathbb{E}[P_{it}|Z_{it}] = [Z_{it}, \bar{Z}_i]'\gamma^P + b_i \end{aligned}$$

where the parameters γ^Y, γ^P are high-dimensional sparse parameters. Consider the dynamic panel Lasso estimator of Kock-Tang:

$$(\widehat{\gamma}_k^D, \widehat{a}_k) = \sum_{(i,t) \in I_k^c} (D_{it} - Z_{it}'\gamma^D - a_i)^2 + \lambda \|\gamma^D\|_1 + \frac{\lambda}{\sqrt{N}} \|a\|_1$$

and

$$(\widehat{\gamma}_k^Y, \widehat{b}_k) = \sum_{(i,t) \in I_k^c} (Y_{it} - Z_{it}'\gamma^Y - b_i)^2 + \lambda \|\gamma^Y\|_1 + \frac{\lambda}{\sqrt{N}} \|b\|_1$$

[7] This specification of the controls implicitly assumes that conditional on own demand history, the price and sales of item i are independent from the demand history of the other members of group g_i. If this assumption is restrictive, one can re-define the controls to include all relevant information for predicting (Y_{it}, P_{it}).

Let the respective reduced form estimate be:

$$\widehat{p}(Z_{it}) = Z_{it}'\widehat{\gamma}^D + \widehat{b}_i$$
$$\widehat{l}(Z_{it}) = Z_{it}'\widehat{\gamma}^Y + \widehat{a}_i$$

Remark 4.2 (Rate of dynamic panel lasso)**.** In case $C = 1$, under mild conditions on the design of $(Y_{it}, P_{it})_{i=1,t=1}^{GC,T}$ Theorem 1 of Kock and Tang (2016) implies that first stage rates $\mathbf{m}_N = \mathbf{l}_N = \frac{\log^{3/2}(p \vee I)s_1}{\sqrt{IT}} \vee s\frac{1}{\sqrt{I}}(\frac{\lambda}{\sqrt{IT}})^{1-\nu} = o((N)^{-1/4})$, where s_1 is a bound on the sparsity of γ^P, γ^Y. If the weak sparsity measure of unobserved heterogeneity s is sufficiently small, Assumption 3.1 holds. By Corollary E.1, one can always take $\lambda_N \equiv \mathbf{m}_N^2 = o(1/\sqrt{N})$ in Assumption 3.2. We expect Assumption 3.1 to hold for any cluster size C, but proving this is left as future work.

A Notation

Let $\mathcal{S}^{d-1} = \{\alpha \in \mathcal{R}^d : \alpha'\alpha = 1\}$ denote a d-dimensional unit sphere. For a matrix A, let $\|M\| = \|A\|_2 = \sup_{\alpha \in \mathcal{S}^{d-1}} \|M\alpha\|$. For two sequences of random variables denote $a_n, b_n, n \geqslant 1 : a_n \lesssim_P b_n \equiv a_n = O_P(b_n)$. For two numeric sequences of numbers, denote $a_n, b_n, n \geqslant 1 : a_n \lesssim b_n \equiv a_n = O(b_n)$. Let $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$. The l_2 norm is denoted by $\|\cdot\|$, the l_1 norm is denoted by $\|\cdot\|_1$, the l_∞ is denoted by $\|\cdot\|_\infty$, and the l_{i0} - norm denotes the number of nonzero components of a vector. Given a vector $\delta \in \mathcal{R}^p$ and a set of indices $T \subset \{1, ..., p\}$, we denote by δ_T the vector in $\mathcal{R}^p$ in which $\delta_{Tj} = \delta_j, j \in T$ and $\delta_{Tj} = 0, j \notin T$. The cardinality of T is denoted by $|T|$. Given a covariate vector $x_{it} \in \mathcal{R}^p$, let $x_{it}[T]$ denote the vector $\{x_{it,j}, j \in T\}$. The symbol $\mathbb{E}$ denotes the expectation.

The generic index of an observation in $it, i \in [I] := \{1, 2, .., I\}, t \in \{1, 2, .., T\}$. The set I consists of pairs $\{(m, c), m \in [M] := \{1, 2, ..., M\}, c \in [C] := \{1, 2, ..., C\}\}$, where the first component $m = m(i)$ designates group number, and the second $c \in [C]$ - the cluster index within group. For a random variable V, the quantity $v_{m,t}$ denote a C-vector $v_{m,t} = [v_{1,t}, v_{2,t}, ..., v_{C,t}]$. For a random d-vector V, the quantity $v_{m,t}$ denote a $C \times d$-matrix $v_{m,t} = [v'_{1,t}, v_{2,t}, ..., v'_{C,t}]$. The observations $(Y_{mt}, D_{mt}, Z_{mt}))_{mt=1}^{MT}$ are i.i.d across $m \in [M]$. Let $N = MT$ be effective sample size. We will use empirical process notation:

$$\mathbb{E}_N f(x_{it}) \equiv \frac{1}{MT} \sum_{mt=1}^{MT} \sum_{c=1}^{C} f(x_{mct}) = \frac{1}{MT} \sum_{it=1}^{MT} f(x_{it})$$

and

$$\mathbb{G}_N f(x_{it}) \equiv \frac{1}{\sqrt{MT}} \sum_{mt=1}^{MT} \sum_{c=1}^{C} [f(x_{mct}) - \mathbb{E} f(x_{mct})] = \frac{1}{MT} \sum_{it=1}^{IT} [f(x_{it}) - \mathbb{E} f(x_{it})]$$

Recognize that this differs from regular cross-sectional empirical process notation, since we are not dividing by group size C.Let us introduce the covariance matrix of estimated residuals. Let

$$\mathbb{E}_{n,kc} f(x_{it}) := \frac{1}{N} \sum_{(m,t):(g,c,t):\in I_k^c} f(x_{it}) \text{ and } \mathbb{G}_{n,kc} f(x_{it}) := \frac{\sqrt{N}}{N} \sum_{(m,t):(g,c,t):\in I_k^c} [f(x_{it}) - \mathbb{E} f(x_{it})]$$

This operation defines sample average within a partition I_k and observations with index c within their group . This ensures that observations entering the sample average are i.i.d across groups $m \in [M]$. Let us introduce covariance matrix of estimated residuals

$$\widehat{Q} \equiv \frac{1}{MT} \sum_{it=1}^{IT} \widehat{\tilde{D}}_{it} \widehat{\tilde{D}}'_{it}$$

and true (oracle) residuals and the maximal entry-wise difference between $\widehat{Q}, \tilde{Q}$:

$$\tilde{Q} \equiv \frac{1}{MT} \sum_{it=1}^{IT} \tilde{D}_{it} \tilde{D}'_{it}$$

B Proofs

$$q_N \equiv \max_{1 \leqslant m,j \leqslant d} |\widehat{Q} - \tilde{Q}|_{m,j}$$

The following theorem follows from McLeish (1974).

Theorem B.1. *Let $\{\tilde{D}_{mt}, \tilde{U}_{mt}\}_{mt=1}^{N}$ be a m.d.s. of d-vectors. Let Assumption 3.4 and 3.5 hold. Let*

$$Q \equiv \mathbb{E} \tilde{D}'_{mt} \tilde{D}_{mt}$$

and

$$\Gamma \equiv \mathbb{E} \tilde{D}'_{mt} U_{mt} U'_{mt} \tilde{D}_{mt}$$

and

$$\Omega \equiv Q^{-1} \Gamma Q^{-1}$$

and $\Phi(t)$ be a $N(0,1)$ c.d.f. Then, $\forall t \in \mathcal{R}$ and for any $\alpha \in \mathcal{S}^{d-1}$, we have

$$\lim_{N \to \infty} \left| \mathrm{P} \left(\frac{\sqrt{N} \alpha' Q^{-1} \mathbb{E}_N \tilde{D}'_{mt} U_{mt}}{\|\alpha' \Omega\|^{1/2}} < t \right) - \Phi(t) \right| = 0. \tag{B.1}$$

Proof. Let $\xi_{mt} \equiv \frac{\alpha' Q^{-1} \tilde{D}'_{mt} U_{mt}}{\sqrt{N} \|\alpha' \Omega\|^{1/2}}$. Let us check conditions of Theorem 3.2 from McLeish (1974):

(a) $\max_{1 \leqslant mt \leqslant N} |\xi_{mt}| = O_P(1)$

(b) $\mathbb{E} \xi^2_{mt} 1_{\|U_{mt} U'_{mt}\| > \epsilon} \lesssim \mathbb{E} \|U_{mt} U'_{mt}\| 1_{\|U_{mt} U'_{mt}\| > \epsilon} \to 0, \epsilon \to 0$ by Assumption 3.5

(c) $\sum_{mt=1}^{N} \xi^2_{mt} = \mathbb{E}_N \xi^2_{mt} \to_p \frac{\alpha' Q^{-1} \mathbb{E} \tilde{D}'_{mt} U_{mt} U'_{mt} \tilde{D}_{mt} Q^{-1} \alpha}{\alpha' \Omega \alpha} = 1$

■

Proof of Theorem 3.1.

$$\begin{aligned}
\|\widehat{\beta}-\beta_0\| &= \widehat{Q}^{-1}\mathbb{E}_N \widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it} - \beta_0 \\
&\leqslant \widehat{Q}^{-1}\mathbb{E}_N \widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it} \pm \widehat{Q}^{-1}\mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it} \pm \tilde{Q}^{-1}\mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it} - \beta_0 \\
&\leqslant \underbrace{\|\widehat{Q}^{-1}\|\|\mathbb{E}_N \widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it} - \mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it}\|}_{a} + \underbrace{\|\widehat{Q}^{-1}-\tilde{Q}^{-1}\|\|\mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it}\|}_{b} \\
&+ \underbrace{\|\tilde{Q}^{-1}\mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it} - \beta_0\|}_{c}
\end{aligned}$$

Under Assumption 3.3, $\|\tilde{Q}-Q\| \lesssim_P \sqrt{\frac{d\log N}{N}}$. Therefore, wp $\to 1$, all eigenvalues of $\tilde{Q}^{-1}$ are bounded away from zero. Indeed, suppose $\tilde{Q}$ has an eigenvalue less then $C_{\min}/2$. Then, there exists a vector $a \in \mathcal{S}^{d-1}$, such that $a'\tilde{Q}a < C_{\min}/2$. Then,

$$\|\tilde{Q}-Q\| \geqslant |a'(\tilde{Q}-Q)|a \geqslant C_{\min}/2$$

Therefore, w.h.p. the eigenvalues of $\tilde{Q}$ are bounded away from zero. By Lemma D.1 $\|\widehat{Q}-\tilde{Q}\| \lesssim_P d\mathbf{m}_N^2$. Therefore, w.h.p. the eigenvalues of $\widehat{Q}$ are bounded away from zero. By Lemma D.1

$$\|a\| \lesssim_P [\sqrt{d}\mathbf{m}_N\mathbf{l}_N + d\mathbf{m}_N^2\|\beta_0\| + \lambda_N]$$

$$\begin{aligned}
\|b\|\|\mathbb{E}_N \tilde{D}_{it}\tilde{Y}_{it}\| &\lesssim_P \|\widehat{Q}^{-1}-\tilde{Q}^{-1}\|(\|Q\|\|\beta_0\| + O_P(1/N)) \\
&\lesssim_P \|\widehat{Q}^{-1}\|\|\widehat{Q}-\tilde{Q}\|\|\tilde{Q}^{-1}\| \\
&\lesssim_P [d\mathbf{m}_N^2 + \sqrt{d}\lambda_N]\|\beta_0\| \\
\|c\| &\lesssim_P \sqrt{\frac{d}{N}}
\end{aligned}$$

$$\|\widehat{\beta}-\beta_0\| \lesssim_P \sqrt{d}\mathbf{m}_N\mathbf{l}_N \vee (d\mathbf{m}_N^2 + \sqrt{d}\lambda_N)\|\beta_0\| \vee \sqrt{d/N} \tag{B.2}$$

Step 2: Asymptotic Linearity Let

$$\widehat{\tilde{Y}}_{it} = \widehat{\tilde{D}}_{it}'\beta_0 + R_{it} + U_{it}$$

where

$$R_{it} = (\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))'\beta_0 + (l_{i0}(Z_{it}) - \widehat{l}_i(Z_{it})), i \in \{1,2,...,N\}$$

summarizes first stage approximation error.

$$\sqrt{N}\alpha'(\widehat{\beta}-\beta) = \sqrt{N}\alpha'(\widehat{Q}^{-1}\mathbb{E}_N\widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it} - \beta_0) \tag{B.3}$$

$$= \sqrt{N}\alpha'\widehat{Q}^{-1}\mathbb{E}_N\widehat{\tilde{D}}_{it}(R_{it}+U_{it}) \tag{B.4}$$

$$= \sqrt{N}\alpha' Q^{-1}\mathbb{E}_N\tilde{D}_{it}U_{it} + R_{1,N}(\alpha) \tag{B.5}$$

where

$$R_{1,N}(\alpha) = \underbrace{\sqrt{N}\alpha'\widehat{Q}^{-1}[\mathbb{E}_N\widehat{\tilde{D}}_{it}(R_{it}+U_{it}) - \mathbb{E}_N\tilde{D}_{it}U_{it}]}_{S_1} + \underbrace{\sqrt{N}\alpha'(\widehat{Q}^{-1}-Q^{-1})\mathbb{E}_N\tilde{D}_{it}U_{it}}_{S_2}$$

In Step 1 it was shown that the eigenvalues of $\widehat{Q}^{-1}$ are bounded away from zero. By Lemma D.1,

$$|S_1| \leqslant \|\alpha\|\|\widehat{Q}^{-1}\|\|\sqrt{N}[\mathbb{E}_N\widehat{\tilde{D}}_{it}(R_{it}+U_{it}) - \mathbb{E}_N\tilde{D}_{it}U_{it}]\| \lesssim_P \sqrt{N}[\sqrt{d}\mathbf{m}_N\mathbf{l}_N + d\mathbf{m}_N^2\|\beta_0\| + \lambda_N]$$

By Lemma D.1,

$$|S_2| \leqslant \|\alpha\|\|\widehat{Q}^{-1}-Q^{-1}\|\|\sqrt{N}\mathbb{E}_N\tilde{D}_{it}U_{it}\| \lesssim_P \|\widehat{Q}^{-1}\|\|\widehat{Q}-Q\|\|\widehat{Q}^{-1}\|\|\sqrt{N}\mathbb{E}_N\tilde{D}_{it}U_{it}\|$$
$$\lesssim_P [d\mathbf{m}_N^2 + \lambda_N + \sqrt{\frac{d\log N}{N}}]\bar{\sigma}O_P(1)$$

Equation B.5 establishes Asymptotic Linearity representation of $\widehat{\beta}$. Theorem B.1 implies Asymptotic Normality of $\widehat{\beta}$ with asymptotic variance

$$\Omega = Q^{-1}\underbrace{\mathbb{E}\tilde{D}'_{mt}U_{mt}U'_{mt}\tilde{D}'_{mt}}_{\Gamma}Q^{-1}$$

Step 3: Asymptotic Variance Let $\widehat{U}_{mt} = \widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{mt}\widehat{\beta}$ be estimated outcome disturbances. Then, asymptotic variance $\Omega = Q^{-1}\Gamma Q^{-1}$ can be consistently estimated by $\Omega = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1}$ where

$$\widehat{\Gamma} = \mathbb{E}_{MT}\widehat{\tilde{D}}'_{mt}\widehat{U}_{mt}\widehat{U}'_{mt}\widehat{\tilde{D}}_{mt}$$

Let $\tilde{\Gamma} = \mathbb{E}_{MT}\tilde{D}'_{mt}U_{mt}U'_{mt}\tilde{D}_{it}$ be a oracle estimate of Γ, where oracle knows β_0 and the first stage estimates. Let $\xi_{it} := \tilde{D}'_{it}U_{it}$ and $\widehat{\xi}_{it} := \widehat{\tilde{D}}'_{it}\widehat{U}_{it}$ be d-vectors. Recognize that

$\widehat{\Gamma} = \mathbb{E}_N \widehat{\xi}_{it}\widehat{\xi}_{it}'$ and $\tilde{\Gamma} = \mathbb{E}_N \xi_{it}\xi_{it}'$. Recognize that

$$\begin{aligned}
\|\widehat{\Gamma} - \tilde{\Gamma}\| &= \|\mathbb{E}_{MT}[\widehat{\xi}_{it}\widehat{\xi}_{it}' - \xi_{it}\xi_{it}']\| \\
&\leqslant \|\mathbb{E}_{MT}[\widehat{\xi}_{it} - \xi_{it}]\widehat{\xi}_{it}'\| + \|\mathbb{E}_{MT}[\widehat{\xi}_{it} - \xi_{it}]\xi_{it}'\| \\
&\leqslant \underbrace{\|\mathbb{E}_{MT}[\widehat{\xi}_{it} - \xi_{it}]^2\|}_{a} \sup_{\alpha \in \mathcal{S}^{d-1}} (\mathbb{E}_{MT}(\alpha'\widehat{\xi}_{it})^2)^{1/2} \\
&+ \underbrace{\|\mathbb{E}_{MT}[\widehat{\xi}_{it} - \xi_{it}]^2\|}_{a} \sup_{\alpha \in \mathcal{S}^{d-1}} (\mathbb{E}_{MT}(\alpha'\xi_{it})^2)^{1/2}\| \\
&\lesssim_P aO_P(1)
\end{aligned}$$

Recognize that both $\widehat{\xi}_{it}$ and ξ_{it} are inner products of C summands.

$$\begin{aligned}
\widehat{\xi}_{it} - \xi_{it} &= [\sum_{c=1}^{C} \widehat{\tilde{D}}_{it}[\widehat{Y}_{it} - \widehat{\tilde{D}}'_{it}\beta_0 + \widehat{\tilde{D}}'_{it}\beta_0 - \widehat{\tilde{D}}'_{it}\widehat{\beta}] \\
&- \sum_{c=1}^{C} \tilde{D}_{it}U_{it}] \\
a = \|\mathbb{E}_{MT}(\widehat{\xi}_{it} - \xi_{it})^2\| &\leqslant C\|\mathbb{E}_N[\widehat{\tilde{D}}_{it}[\widehat{Y}_{it} - \widehat{\tilde{D}}'_{it}\beta_0 + \widehat{\tilde{D}}'_{it}\beta_0 - \widehat{\tilde{D}}'_{it}\widehat{\beta}] \\
&- \sum_{c=1}^{C} \tilde{D}_{it}U_{it}]^2\| \\
&\lesssim_P \underbrace{\|\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}[U_{it}]]^2\|}_{o_P(1) \text{ by Lemma D.1}} + \underbrace{\|\widehat{\beta} - \beta_0\|^2}_{o_P(1)} \quad = o_P(1)
\end{aligned}$$

By Lemma D.1, $\|\widehat{Q} - Q\| \lesssim_P d\mathbf{m}_N^2 + \sqrt{d}\lambda_N = o_P(1)$. By Assumption 3.3, $\|\tilde{Q} - Q\| = o_P(1)$.

$$\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1} = (Q^{-1} + o_P(1))(\Gamma + o_P(1))(Q^{-1} + o_P(1))$$

■

Proof of Theorem 3.2. For every $\delta = \widehat{\beta} - \beta_0, \delta \in R^d$ we use the notation:

$$\|\delta\|_{2,N} = (\mathbb{E}_N(\tilde{D}_{it}'\delta)^2)^{1/2}$$

and

$$\|\delta\|_{\widehat{d},2,N} = (\mathbb{E}_N(\widehat{\tilde{D}}'_{it}\delta)^2)^{1/2}$$

$$\widehat{Q}(\widehat{\beta}_L) - \widehat{Q}(\beta_0) - \mathbb{E}_N(\widehat{\tilde{D}}'_{it}\delta)^2 = -2\underbrace{\mathbb{E}_N\left[U_{it}\tilde{D}'_{it}\delta\right]}_{a} \tag{B.6}$$

$$-2\underbrace{\mathbb{E}_N\left[U_{it}(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\delta\right]}_{b} \tag{B.7}$$

$$-2\underbrace{\mathbb{E}_N\left[(l_{i0}(Z_{it}) - \widehat{l}_i(Z_{it}) + (d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\beta_0)(\tilde{D}_{it})'\delta\right]}_{c} \tag{B.8}$$

$$-2\underbrace{\mathbb{E}_N\left[(l_{i0}(Z_{it}) - \widehat{l}_i(Z_{it}) + (d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\beta_0)(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it})))'\delta\right]}_{d} \tag{B.9}$$

By Lemma D.4, $|b+c| \lesssim_P D^2\sqrt{\frac{\log(2d)}{N}} + \mathbf{m}_N^2 s\|\beta_0\|^2 + \mathbf{m}_N\mathbf{l}_N$ and $|d| \lesssim_P \lambda_N + \mathbf{m}_N^2 s\|\beta_0\|^2 + \mathbf{m}_N\mathbf{l}_N$. Since a is a sample average of bounded martingale difference sequences $a \lesssim_P \sqrt{\frac{s\log d}{N}}$ by Azouma-Hoeffding inequality. Therefore, with high probability $\exists c > 1 \quad \lambda \geqslant c[\sqrt{\frac{s\log d}{N}} + \sqrt{\frac{\log(2d)}{N}} + \mathbf{m}_N^2 s\|\beta_0\|^2 + \mathbf{m}_N\mathbf{l}_N + \lambda_N]$. Optimality of $\widehat{\beta}_L$ and the choice of λ imply:

$$\lambda(\|\beta_0\|_1 - \|\widehat{\beta}\|_1) \geqslant \|\delta\|^2_{\widehat{d},2,N} \geqslant -\lambda/c\|\delta\|_1 \tag{B.10}$$

Triangle inequality implies:

$$\begin{aligned} -\lambda/c\|\delta\|_1 &\leqslant \lambda(\|\beta_0\|_1 - \|\widehat{\beta}\|_1) \leqslant \lambda(\|\delta_T\|_1 - \|\delta_{T^c}\|_1) \\ \|\delta_{T^c}\|_1 &\leqslant \frac{c+1}{c-1}\|\delta_T\|_1 \end{aligned}$$

Therefore, δ belongs to the restricted set in the RE($\bar{c}$), where $\bar{c} = \frac{c+1}{c-1}$. By Lemma D.3,

$$\delta \in \mathcal{RE}(\bar{c}) \Rightarrow$$

$$(1 - \frac{(1+\bar{c})^2}{\kappa(\tilde{Q},T,\bar{c})^2})\|\delta\|^2_{2,N} \leqslant \|\delta\|^2_{\widehat{d},2,N} \leqslant \|\delta\|^2_{2,N}(1 + \frac{(1+\bar{c})^2}{\kappa(\tilde{Q},T,\bar{c})^2})$$

$$\begin{aligned} \|\delta\|^2_{2,N} &\leqslant \frac{\|\delta\|^2_{\widehat{d},2,N}}{(1 - q_N(1+\bar{c})^2 s/\kappa(\tilde{Q},T,\bar{c})^2)} \\ &\leqslant \lambda\|\delta_T\|_1\frac{1}{(1 - q_N(1+\bar{c})^2 s/\kappa(\tilde{Q},T,\bar{c})^2)} \\ &\leqslant \lambda\frac{\sqrt{s}\|\delta\|_{2,N}}{\kappa(\tilde{Q},T,\bar{c})}\frac{1}{(1 - q_N(1+\bar{c})^2 s/\kappa(\tilde{Q},T,\bar{c})^2)} \end{aligned}$$

$$
\begin{aligned}
\|\delta\|_1 &\leqslant \|\delta\|_{2,N}\frac{\sqrt{s}}{\kappa(\tilde{Q},T,2\bar{c})} \\
&\leqslant \lambda\frac{s}{\kappa(\tilde{Q},T,2\bar{c})\kappa(\tilde{Q},T,\bar{c})}\frac{1}{(1-q_N(1+\bar{c})^2 s/\kappa(\tilde{Q},T,\bar{c})^2)}
\end{aligned}
$$

■

C Inference in High-Dimensional Sparse Models

Definition C.1 (Orthogonalization matrix)**.** *Let $\mu_N : N \geqslant 1$ be an $o(1)$ sequence. We say that a $d \times d$-dimensional matrix $M = [m_1, ..., m_d]'$ orthogonalizes a given matrix Q at rate μ_N if:*

$$
\|MQ - I\|_\infty \leqslant \mu_N \tag{C.1}
$$

Denote by $M_{\mu_N}(Q)$ the set of all matrices that orthogonalize Q at rate μ_N. We will refer to it as orthogonalization set of M.

Lemma C.1 (Relation between orthogonalization sets of true and estimated residuals)**.** *Let $\widehat{Q}$ and $\tilde{Q}$ be a sample covariance matrix of estimated and true residuals. Let and $M_{\mu_N}(\widehat{Q})$, $M_{\mu_N}(\tilde{Q})$ be their respective orthogonalization sets with common rate μ_N. Then, $\forall M \in M_{\mu_N}(\tilde{Q})$ that satisfy Assumption 3.8, with high probability*

$$
\mathrm{P}(M \in M(\widehat{Q})) \to 1, N \to \infty, d \to \infty
$$

Moreover, since $Q^{-1} \in M(\tilde{Q}))$ by Lemma 6.2 of Javanmard and Montanari (2014), $Q^{-1} \in M(\widehat{Q}))$. Therefore, $M(\widehat{Q}))$ is non-empty w.h.p.

Lemma C.2 (Asymptotic Linearity of $\widehat{\beta}_{DOL}$ and $\widehat{\beta}_{Ridge}$)**.**

$$
\sqrt{N}(\widehat{\beta}_{DOL} - \beta_0) = \sqrt{N}M\mathbb{E}_N\widehat{\tilde{D}}_{it}(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\widehat{\beta}_L) + \widehat{\beta}_L - \beta_0 \tag{C.2}
$$

where $\|R_{1,N}\|_\infty \lesssim_P \sqrt{N}\lambda\mu_N \vee \sqrt{N}[\sqrt{s}\boldsymbol{m}_N + \boldsymbol{l}_N](\boldsymbol{m}_N m_0 \vee \lambda_N)|T|$

Proof of Lemma C.1.

$$
|Q^{-1}\widehat{Q} - I|_\infty \leqslant |Q^{-1}\tilde{Q} - I|_\infty + |Q^{-1}(\widehat{Q} - \tilde{Q})|_\infty
$$

$$
\begin{aligned}
|Q^{-1}(\widehat{Q}-\tilde{Q})|_\infty &\leqslant \max_{1\leqslant m,j\leqslant d} |Q^{-1}_{j,\cdot}(\widehat{Q}-\tilde{Q})_{\cdot,m}| \\
&\leqslant \max_{1\leqslant j\leqslant d} \sum_{i=1}^{d} |Q^{-1}|_{j,i} \max_{1\leqslant m,j\leqslant d} |(\widehat{Q}-\tilde{Q})_{i,m}| \\
&\lesssim m_0 q_N
\end{aligned}
$$

where the last inequality follows from Assumption 3.8 and Lemma D.2. By Lemma 6.2 of Javanmard and Montanari (2014) ,

$$
\mathrm{P}(|Q^{-1}\tilde{Q}-I|_\infty \geqslant a\sqrt{\frac{\log d}{N}}) \leqslant 2d^{-c_2} \tag{C.3}
$$

which finishes the proof.

■

Proof of Lemma C.2. Let

$$
R_{it} = (d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\beta_0 + (\widehat{l}_i(Z_{it}) - l_{i0}(Z_{it})), i \in \{1,2,...,N\} \tag{C.4}
$$

summarize the first-stage approximation error, that contaminates the outcome . Assumption 3.1 implies a rate on the bias of R_{it}

$$
(\mathbb{E}[(d_{i,0}(Z_{it}) - d(Z_{it}))'\beta_0 + (l(Z_{it}) - l_{i0}(Z_{it}))]^2)^{1/2} \leqslant \sqrt{s}\mathbf{m}_N \vee \mathbf{l}_N
$$

$$
\begin{aligned}
\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\beta_0 &= \tilde{Y}_{it} - \tilde{D}_{it}\beta_0 + (\widehat{\tilde{Y}}_{it} - \tilde{Y}_{it}) - (\widehat{\tilde{D}}_{it} - \tilde{D}_{it})\beta_0 \\
&= U_{it} + R_{it}
\end{aligned}
$$

and

$$
\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\widehat{\beta}_L = U_{it} + R_{it} + (\widehat{\tilde{D}}_{it})'(\beta_0 - \widehat{\beta}_L)
$$

$$
\begin{aligned}
\widehat{\beta}_{DOL} &= M\mathbb{E}_N \widehat{\tilde{D}}_{it}'(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}\widehat{\beta}_L) + \widehat{\beta}_L \\
&= M\mathbb{E}_N[\tilde{D}_{it} \pm [\widehat{\tilde{D}}_{it} - \tilde{D}_{it}]]'(U_{it} + R_{it} + \widehat{\tilde{D}}_{it}(\beta_0 - \widehat{\beta}_L) + \widehat{\beta}_L \\
&= \beta_0 + M\mathbb{E}_N\tilde{D}_{it}'U_{it} + \Delta_U + \Delta_D + \Delta_R + \Delta
\end{aligned}
$$

$$
\begin{aligned}
\Delta_U &= M\mathbb{E}_N[\widehat{\tilde{D}}_{it} - \tilde{D}_{it}]' U_{it} = M \sum_{k=1}^{K} \sum_{c=1}^{C} \mathbb{E}_{n,kc} \Delta_{U,kc} \\
\Delta_D &= M\mathbb{E}_N[\widehat{\tilde{D}}_{it} - \tilde{D}_{it}]' R_{it} = M \sum_{k=1}^{K} \sum_{c=1}^{C} \mathbb{E}_{n,kc} \Delta_{D,kc} \\
\Delta_R &= M\mathbb{E}_N \tilde{D}'_{it} R_{it} = M \sum_{k=1}^{K} \sum_{c=1}^{C} \mathbb{E}_{n,kc} \Delta_{R,kc} \\
\Delta &= (M\widehat{Q} - I)(\beta_0 - \widehat{\beta}_L)
\end{aligned}
$$

C.0.1 Step 1

Let m_j^0 be j'the row of a sparse approximation of Q^{-1} and let $T := \{k, (m_j^0)_k \neq 0\}$ be the set of its active coordinates: $|T| = O(1)$. Let $\delta = m_j - m_j^0$. Since $Q^{-1} \in M_{\mu_N}(\widehat{Q})$, triangular inequality implies:

$$
\begin{aligned}
\|m_j\|_1 &\leqslant \|(m_j^0)_T\|_1 \leqslant \|(m_j)_T\|_1 + \|(\delta)_T\|_1 \\
\|(m_j)_T\|_1 + \|(m_j)_{T^c}\|_1 &\leqslant \|(m_j)_T\|_1 + \|(\delta)_T\|_1 \\
\|(\delta_j)_{T^c}\|_1 = \|(m_j)_{T^c}\|_1 &\leqslant \|(\delta)_T\|_1
\end{aligned}
$$

Therefore, $\delta \in \mathcal{RE}(2)$, that is $\|\delta\|_1 \leqslant 2\|\delta_T\|_1$.

C.0.2 Step 2: Covariance of Approximation Errors $\Delta_{D,kc}$

For any vector $m_j = m_j^0 + \delta, m_j \in \mathcal{R}^d$,

$$
\begin{aligned}
\underbrace{(m_j'(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it})))^2}_{(a+b)^2} \leqslant 2\,&[\underbrace{((m_j^0)'(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it})))^2}_{a^2} \\
&+ \underbrace{(\delta'(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it})))^2]}_{b^2}
\end{aligned}
$$

Applying $\mathbb{E}_{n,kc}(\cdot)$ to both sides of inequality:

$$
\begin{aligned}
\mathbb{E}_{n,kc}(m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2 \leqslant\, &2[\mathbb{E}_{n,kc}(m_j^0)'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2] \\
&+ 2[\mathbb{E}_{n,kc}(\delta'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2]
\end{aligned}
$$

Let $\Gamma := \mathbb{E}_N(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))'$.

$$
\begin{aligned}
\mathbb{E}_{n,kc}(\delta'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2 = \delta'\Gamma\delta &\leqslant \max_{1\leqslant j,k\leqslant d} |\Gamma|_{j,k}\|\delta\|_1^2 \\
&\leqslant^{ii} q_N\|\delta\|_1^2 \\
&\leqslant^{iii} q_N 4\|(\delta)_T\|_1^2 \\
&=^{iv} O_P(4q_N|T|) = O_P(4[\lambda_N + \mathbf{m}_N^2]]|T|) \qquad \text{(C.5)}
\end{aligned}
$$

where (ii) is by Lemma D.2 , (iii) is by $\delta \in \mathcal{RE}(2)$ (Step 1), and (iv) is by T is finite. Assumption 3.2 implies:

$$
\mathbb{E}_N(m_j^0)'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2 \leqslant m_0[\lambda_N + \mathbf{m}_N^2]
$$

we obtain a bound for any $j \in \{1, 2, .., d\}$:

$$
\begin{aligned}
\sqrt{N}|\Delta_{D,kc}| &= \sqrt{N}|\mathbb{E}_{n,kc}[m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))]R_{it}| \\
&\leqslant \sqrt{N}((\mathbb{E}_{n,kc}m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2)^{1/2}(\mathbb{E}_{n,kc}R_{it}^2)^{1/2} \\
&\lesssim_P (\sqrt{N}\mathbf{m}_N m_0 \vee \lambda_N|T|)(\sqrt{s}\mathbf{m}_N \vee \mathbf{l}_N \vee \lambda_N)
\end{aligned}
$$

C.0.3 Step 3: Covariance of Approximation Error and Sampling Error $\Delta_{U,kc}$

Recall that $\Delta_{U,kc} = \mathbb{E}_{n,kc}[m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))]|U_{it}$. Since $U_{it}|(D_{it}, Z_{it})_{it=1}^N] = 0$, $\Delta_{U,kc}$ is mean zero.

$$
\mathbb{E}[\mathbb{E}_{n,kc}[m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))]U_{it}|(D_{it}, Z_{it})_{it=1}^N] = 0
$$

$$
\begin{aligned}
\sqrt{N}|\Delta_{U,kc}| &\lesssim_P^i [N\mathbb{E}[\Delta_{U,kc}^2|(D_{it}, Z_{it})_{it=1}^N]]^{1/2} \lesssim_P^{ii} [\mathbb{E}_{n,kc}(m_j'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))U_{it})^2]^{1/2} \\
&\lesssim_P^{iii} 4q_N m_0\bar{\sigma}
\end{aligned}
$$

where i is by Markov inequality conditionally on $(D_{it}, Z_{it})_{it=1}^N$, ii is by uncorrelatedness of $(U_{it})_{it=1}^N$ and iii is by Step 2 (Equation C.5).

C.0.4 Step 4: $\Delta_{R,kc}$

Fix a partition $k \in [K]$. Conditionally on I_k^c,

$$\begin{aligned}
&\mathbb{E}[\sqrt{n}\mathbb{E}_{n,kc}(m_j^0)'\tilde{D}_{it}R_{it}|I_k^c] = 0 \\
&(\sqrt{N}|\Delta_{R,kc}|)^2 \lesssim_P^i n\mathbb{E}[\mathbb{E}_{n,kc}(m_j^0)'\tilde{D}_{it}R_{it}|I_k^c]^2 \qquad \leqslant \mathbb{E}[(m_j^0)'\tilde{D}_{it}R_{it}]^2 \\
&\leqslant^{ii} m_0\mathbb{E}[\|\tilde{D}_{it}\|^2|I_k^c, Z_{it}]\mathbb{E}_{n,kc}R_{it}^2 \\
&\lesssim^{iii} m_0^2(sd\mathbf{m}_N^2 + \mathbf{l}_N^2 + \lambda_N)
\end{aligned}$$

where i by Markov conditionally on $I_k^c, (Z_{it})_{it=1}^N$ and iii is by Equation C.4.

$$\mathbb{E}[\mathbb{E}_{n,kc}\tilde{D}_{it}R_{it}|I_k^c] = \mathbb{E}_{n,kc}[\mathbb{E}[D_{it}|I_k^c, Z_{it}]R_{it}|I_k^c] = 0$$

Azouma-Hoeffding inequality implies

$$\max_{1\leqslant j\leqslant d}|\mathbb{E}_{n,kc}\tilde{D}_{j,it}R_{it}||I_k^c \lesssim_P \sqrt{\frac{\log d}{N}}$$

$$\begin{aligned}
\sqrt{N}|\delta'\mathbb{E}_{n,kc}\tilde{D}_{j,it}R_{it}| &\leqslant \sqrt{N}\|\delta\|_1 \max_{1\leqslant j\leqslant d}|\mathbb{E}_{n,kc}\tilde{D}_{j,it}R_{it}|| \\
&\leqslant \sqrt{N}2\|\delta_T\|_1 D\sqrt{\frac{s\log d}{N}} \\
&\leqslant \sqrt{N}2|T|\sqrt{\frac{\log d}{N}}D\sqrt{\frac{s\log d}{N}}
\end{aligned}$$

Therefore, $\sqrt{N}|\Delta_{R,kc}| = o_P(1)$

C.0.5 Step 4: Δ

$$\begin{aligned}
\sqrt{N}\|\Delta\|_\infty &\leqslant \|M\widehat{Q} - I\|_\infty\|\beta_0 - \widehat{\beta}_L\|_1 \\
&\lesssim_P 2\sqrt{N}\mu_N\lambda s
\end{aligned}$$

■

C.0.6 Consistency of $\widehat{\Omega}$

Let $\widehat{U}_{mt} = \widehat{\tilde{Y}}_{mt} - \widehat{\tilde{D}}_{mt}\widehat{\beta}_L$ be estimated outcome disturbances. Then, asymptotic variance $\Omega = M\Gamma M'$ can be consistently estimated by

$$\widehat{\Omega} = M\mathbb{E}_N \widehat{\tilde{D}}_{mt}' \widehat{U}_{mt} \widehat{U}_{mt}' \widehat{\tilde{D}}_{mt} M'$$

Let $\xi_{it} := \tilde{D}_{it}' U_{it}$ and $\widehat{\xi}_{it} := \widehat{\tilde{D}}_{it}' \widehat{U}_{it}$ be d-vectors. Recognize that $\widehat{\Gamma} = \mathbb{E}_N \widehat{\xi}_{it} \widehat{\xi}_{it}'$ and $\tilde{\Gamma} = \mathbb{E}_N \xi_{it}\xi_{it}'$. Let m_j be the j'th row of M. Recognize that:

$$\begin{aligned}
m_j'[\widehat{\Gamma} - \tilde{\Gamma}]m_j &= \mathbb{E}_{MT} m_j'[\widehat{\xi}_{it}\widehat{\xi}_{it}' - \xi_{it}\xi_{it}']m_j \\
&\leqslant \mathbb{E}_{MT} m_j'[\widehat{\xi}_{it} - \xi_{it}]\widehat{\xi}_{it}' m_j' + \mathbb{E}_{MT} m_j'[\widehat{\xi}_{it} - \xi_{it}]\xi_{it}' m_j \\
&\leqslant (\underbrace{\mathbb{E}_{MT}[m_j'(\widehat{\xi}_{it} - \xi_{it})]^2}_{a})^{1/2} (\mathbb{E}_{MT}[m_j'(\widehat{\xi}_{it})]^2)^{1/2} \\
&+ (\underbrace{\mathbb{E}_{MT}[m_j'(\widehat{\xi}_{it} - \xi_{it})]^2}_{a})^{1/2} (\mathbb{E}_{MT}[m_j'(\xi_{it})]^2)^{1/2} \\
&\lesssim_P aO_P(1)
\end{aligned}$$

Proof: Recognize that both $\widehat{\xi}_{it}$ and ξ_{it} are inner products of C summands.

$$\begin{aligned}
\widehat{\xi}_{it} - \xi_{it} &= [\sum_{c=1}^{C} \widehat{\tilde{D}}_{it}[\widehat{Y}_{it} - \widehat{\tilde{D}}_{it}'\beta_0 + \widehat{\tilde{D}}_{it}'\beta_0 - \widehat{\tilde{D}}_{it}'\widehat{\beta}_L] \\
&- \sum_{c=1}^{C} \tilde{D}_{it} U_{it}] \\
a = \mathbb{E}_{MT}[m_j'(\widehat{\xi}_{it} - \xi_{it})]^2\| &\leqslant C\mathbb{E}_N[m_j' \big[\widehat{\tilde{D}}_{it}[\widehat{Y}_{it} - \widehat{\tilde{D}}_{it}'\beta_0] + m_j'[\widehat{\tilde{D}}_{it}'\beta_0 - \widehat{\tilde{D}}_{it}'\widehat{\beta}_L]] \\
&- \sum_{c=1}^{C} m_j' \tilde{D}_{it} U_{it}]\big]^2 \\
&\lesssim_P \underbrace{\|\mathbb{E}_N(m_j'[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}[U_{it}]])^2\|}_{o_P(1)} \\
&+ \underbrace{m_0 \max_{1 \leqslant k,j \leqslant d} \|\widehat{Q}_{kj}\| \|\widehat{\beta}_L - \beta_0\|_1^2}_{o_P(1)} \qquad = o_P(1)
\end{aligned}$$

D Supplementary Lemmas

Lemma D.1 (First Stage Error)**.**

$$\|\widehat{Q} - \tilde{Q}\|_2 \lesssim_P d\boldsymbol{m}_N^2 + \sqrt{d}\lambda_N$$

$$\sqrt{N}\|\mathbb{E}_N[\widehat{\tilde{D}}_{i,t}[R_{i,t} + U_{i,t}] - \tilde{D}_{i,t}U_{i,t}]\|_2 \lesssim_P \sqrt{N}\sqrt{d}\boldsymbol{m}_N\boldsymbol{l}_N + d\boldsymbol{m}_N^2\|\beta_0\| + \sqrt{N}\lambda_N$$

$$\|\mathbb{E}_N[\widehat{\tilde{D}}_{i,t}[R_{i,t} + U_{i,t}] - \tilde{D}_{i,t}U_{i,t}]^2\|_2 \lesssim_P d^2\boldsymbol{m}_N^2\|\beta_0\|^2 + d\boldsymbol{l}_N^2$$

Lemma D.2 (Bound on Restricted Eigenvalue of Treatment Residuals)**.** *Let*

$$q_N = \max_{1 \leqslant m,j \leqslant d} |\mathbb{E}_N[\widehat{Q} - \tilde{Q}]|_{m,j}$$

$$q_N \leqslant \boldsymbol{m}_N^2$$

Lemma D.3 (In-Sample Prediction Norm: True and Estimated Residuals)**.** *Let $\bar{c} > 1$ be a constant. Let $\delta \in R^p$ belong to the set $\mathcal{RE}(\bar{c})$*

$$\|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1$$

and assume 3.7($\bar{c}$) holds. Let q_N be defined in Lemma D.2. If $q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2} < 1$,

$$\sqrt{1 - q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}} \leqslant \frac{\|\delta\|_{\widehat{d},2,N}}{\|\delta\|_{2,N}} \leqslant \sqrt{1 + q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}} \tag{D.1}$$

Lemma D.4 (Maximal Inequality for First Stage Approximation Errors)**.** *Let $\widehat{d}_i(Z_{it}), \widehat{l}_i(Z_{it})$ be the first-stage estimate of the treatment and outcome reduced form, and U_{it} is the sampling error . Then, the following bounds hold w.h.p:*

$$\max_{1 \leqslant m \leqslant d} \mathbb{E}_N\|(\widehat{d}_{m,0}(Z_{it}) - d_{m,0}(Z_{it}))(\widehat{l}_i(Z_{it}) - l(Z_{it}))\| \leqslant (\lambda_N + \boldsymbol{m}_N\boldsymbol{l}_N) \tag{D.2}$$

$$\max_{1 \leqslant m \leqslant d} \mathbb{E}_N\|(\widehat{d}_{m,0}(Z_{it}) - d_{m,0}(Z_{it}))(\widehat{d}_i(Z_{it}) - d(Z_{it}))'\beta_0\| \leqslant (\lambda_N + \boldsymbol{m}_N^2 s\|\beta_0\|^2) \tag{D.3}$$

In addition, by Azouma-Hoeffding inequality

$$\max_{1\leqslant m\leqslant d} \mathbb{E}_N \|\tilde{D}_{i,m}(\widehat{l}_i(Z_{it}) - l(Z_{it}))\| \leqslant (DL\sqrt{\frac{\log(2d)}{N}}) \tag{D.4}$$

$$\max_{1\leqslant m\leqslant d} \mathbb{E}_N \|\tilde{D}_{i,m}(\widehat{d}_i(Z_{it}) - d(Z_{it}))'\beta_0\| \leqslant (D^2\sqrt{\frac{\log(2d)}{N}}) \tag{D.5}$$

and

$$\max_{1\leqslant m\leqslant d} \|\mathbb{E}_N|(\widehat{d}_{m,0}(Z_{it}) - d_{m,0}(Z_{it}))U_{it}\| \leqslant (D^2\bar{\sigma}^2\sqrt{\frac{\log 2d}{N}})$$

E Proofs for Section D

The Proofs of Lemmas D.1, D.1, D.2 in cross-sectional case follow the steps below:

1. Decompose a term into KC summands, corresponding to K partitions and C clusters. Within each cluster $\tilde{D}_{it}, \tilde{Y}_{it}$ are m.d.s.

2. Equate the first-order bias to zero by orthogonality and conditional independence

3. Bound the first-order term by Markov inequality and conditional independence

4. Bound the second-order out-of-sample error by Assumption 3.2

Proof of Lemma D.1. **Step 1**

$$\widehat{Q} - \tilde{Q} = \underbrace{\mathbb{E}_N(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{a} + \underbrace{(\mathbb{E}_N(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))')'}_{a'}$$
$$+ \underbrace{\mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{b}$$

Let

$$\mathbb{E}_{n,kc} f(x_{it}) := \frac{1}{N} \sum_{(m,t):(m,c,t):\in I_k^c} f(x_{it}) \text{ and } \mathbb{G}_{n,kc} f(x_{it}) := \frac{1}{\sqrt{N}} \sum_{(m,t):(m,c,t):\in I_k^c} [f(x_{it}) - \mathbb{E}f(x_{it})]$$

The summation in each a_{kc}, b_{kc} is by i.i.d groups $m \in [M]$ and time $t \in [T]$.

$$
\begin{aligned}
a &= \mathbb{E}_N(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))' = \frac{1}{KC}\sum_{k=1}^{K}\sum_{c=1}^{C} \underbrace{\mathbb{E}_{n,kc}(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{a_{kc}} \\
b &= \mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))' \qquad \text{(E.1)} \\
&= \frac{1}{KC}\sum_{k=1}^{K}\sum_{c=1}^{C} \underbrace{\mathbb{E}_{n,kc}(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{b_{kc}}
\end{aligned}
$$

Step 2

$$
\mathbb{E}[a_{kc}|(W_{it})_{i\in I_k^c}] = \mathbb{E}_{n,kc}\mathbb{E}_{Z_{it}}\mathbb{E}[\tilde{D}_{it}|Z_{it},(W_{it})_{i\in I_k^c}](\widehat{d}_i(Z_{it}) - d(Z_{it})) \tag{E.2}
$$

$$
\mathbb{E}[\tilde{D}_{it}|Z_{it}] = 0
$$

Step 3

$$
\begin{aligned}
\mathbb{E}[\|\alpha' a_{kc}\|^2|(W_{it})_{i\in I_k^c}] &= \mathbb{E}[\|\alpha'\mathbb{E}_{n,kc}(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))\|^2|(W_{it})_{i\in I_k^c}] \\
&=^i \frac{1}{n}\mathbb{E}[\|\alpha'(\tilde{D}_{it})(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))\|^2|(W_{it})_{i\in I_k^c}] \\
&= \frac{1}{n}\mathbb{E}_{Z_{it}}\mathbb{E}[(\alpha'\tilde{D}_{it})^2|Z_{it},(W_{it})_{i\in I_k^c}][\|d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it})\|^2|(W_{it})_{i\in I_k^c}] \\
&\leqslant n^{-1}C_{\max} d\mathbf{m}_N^2
\end{aligned}
$$

where equality i follows from conditional exogeneity of errors $\tilde{D}_{it}$ across $(m,t) \in [G,T]$. Markov inequality implies:

$$
\|a\| \leqslant \sum_{k=1}^{K}\|a_{kc}\| \leqslant \sum_{k=1}^{K}\sup_{\alpha\in\mathcal{R}^k:\|\alpha\|=1}\|\alpha' a_{kc}\| = O_P(\sqrt{d}\mathbf{m}_N/\sqrt{N})
$$

Step 4

The bias of b_{kc} attains the following bound:

$$
\begin{aligned}
\|\mathbb{E}[\alpha' b_{kc}|(W_{it})_{i\in I_k^c}]\|^2 &= \sum_{j=1}^{d}[\mathbb{E}\underbrace{(\alpha'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))}_{A}\underbrace{(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))_j}_{B}]^2 \\
&\leqslant \sum_{j=1}^{d}\underbrace{\mathbb{E}(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))_j^2}_{\mathbb{E}A^2}\underbrace{\mathbb{E}(\alpha'(\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it})))^2}_{\mathbb{E}B^2} \\
&\leqslant (d\mathbf{m}_N^2)^2
\end{aligned}
$$

Under Assumption 3.2

$$\|\alpha'(b_{kc} - \mathbb{E}[b_{kc}|(W_{it})_{i\in I_k^c}])\|| (W_{it})_{i\in I_k^c}] \lesssim_P \sqrt{d}\lambda_N$$

In case $T = 1$ (no time dependence), the Step 4(b) can be shown as follows.

Conditional variance of b_{kc} attains the following bound:

$$\begin{aligned}
\mathbb{E}[\|\alpha'(b_{kc} - \mathbb{E}[b_{kc}|(W_{it})_{i\in I_k^c}])\|^2|(W_{it})_{i\in I_k^c}] &\leqslant n^{-1}\mathbb{E}[\|(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})' \\
&\quad - \mathbb{E}(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})'|(W_{it})_{i\in I_k^c}\|^2|(W_{it})_{i\in I_k^c}] \\
&\lesssim d/N
\end{aligned}$$

Therefore, $\lambda_N := \sqrt{1/N}$ in Assumption 3.2.

$$\|\widehat{Q} - \tilde{Q}\| = \|a + a' + b\| \lesssim_P (d\mathbf{m}_N^2 + \sqrt{d}\lambda_N)$$

$\blacksquare$

Step 1

$$\begin{aligned}
\left[\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}\tilde{U}_{it}]\right] &= \underbrace{\mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))U_{it}}_{e} \\
&+ \underbrace{\mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))R_{it}}_{f} \\
&+ \underbrace{\mathbb{E}_N\tilde{D}_{it}R_{it}}_{g}
\end{aligned}$$

Let

$$\mathbb{E}_{n,kc}f(x_{it}) := \frac{1}{N}\sum_{(m,t):(g,c,t):\in I_k^c} f(x_{it}) \text{ and } \mathbb{G}_{n,kc}f(x_{it}) := \frac{\sqrt{N}}{N}\sum_{(m,t):(g,c,t):\in I_k^c}[f(x_{it}) - \mathbb{E}f(x_{it})]$$

The summation in each a_{kc}, b_{kc} is by i.i.d groups $g \in [G]$ and time $t \in [T]$.

$$e = \mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))U_{it} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\mathbb{E}_{n,kc}(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))U_{it}}_{e_{kc}}$$

$$f = \mathbb{E}_N(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))R_{it} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\mathbb{E}_{n,kc}(d_{i,0}(Z_{it}) - \widehat{d}_i(Z_{it}))R_{it}}_{f_{kc}}$$

$$g = \mathbb{E}_N \tilde{D}_{it} R_{it} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\mathbb{E}_{n,kc}\tilde{D}_{it}R_{it}}_{g_{kc}}$$

Step 2. Conditionally on I_k^c,

$$\mathbb{E}[e_{kc}|I_k^c] = 0, \quad \mathbb{E}[g_{kc}|I_k^c] = 0$$

Step 3.

$$n\mathbb{E}\|e_{kc}\|^2|((W_{it})_{i\in I_k^c}) = \mathbb{E}_{Z_{it}}[\mathbb{E}[(\tilde{U}_{it})^2|Z_{it},(W_{it})_{i\in I_k^c}]\|(\widehat{d}_k(Z_{it}) - d_k(Z_{it}))^2\||(W_{it})_{i\in I_k^c}] \leqslant \bar{\sigma}^2 d\mathbf{m}_N^2$$
$$n\mathbb{E}\|g_{kc}\|^2|((W_{it})_{i\in I_k^c}) = \mathbb{E}_{Z_{it}}[\mathbb{E}\|\tilde{D}_{it}\|^2|Z_{it},(W_{it})_{i\in I_k^c}](R_{it})^2|(W_{it})_{i\in I_k^c}] \leqslant d\mathbf{l}_N^2 + d^2\mathbf{m}_N^2\|\beta_0\|^2$$

Step 4. Conditionally on I_k^c,

$$\mathbb{E}[\|f_{kc}\||I_k^c] \leqslant d\mathbf{m}_N^2\|\beta_0\| + \sqrt{d}\mathbf{m}_N\mathbf{l}_N$$

$$\sqrt{n}(f_{kc} - \mathbb{E}[f_{kc}|((W_{it})_{i\in I_k^c}]|((W_{it})_{i\in I_k^c}] \leqslant \sqrt{n}\lambda_N\sqrt{d}$$

Markov inequality implies:

$$\begin{aligned}
\sqrt{n}e_{kc} &= o_P(\bar{\sigma}\sqrt{d}\mathbf{m}_N),\\
\sqrt{n}f_{kc} &= o_P(\sqrt{N}\sqrt{d}\mathbf{m}_N\mathbf{l}_N + \sqrt{N}d\mathbf{m}_N^2\|\beta_0\| + \sqrt{d}\lambda_N'),\\
\sqrt{n}g_{kc} &= o_P(\sqrt{d}\mathbf{l}_N + d\mathbf{m}_N\|\beta_0\|)
\end{aligned}$$

By Markov inequality,

$$\begin{aligned}
\|\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}U_{it}]^2\| &\lesssim_P \|\mathbb{E}_N[[\widehat{\tilde{D}}_{it} - \tilde{D}_{it}]R_{it}]^2\| + \|\mathbb{E}_N[[\widehat{\tilde{D}}_{it} - \tilde{D}_{it}]U_{it}]^2\| \\
&+ \|\mathbb{E}_N\tilde{D}_{it}^2R_{it}^2\| \\
&\lesssim [d\mathbf{m}_N^2\|\beta_0\|^2 + \mathbf{l}_N^2]dD^2 + d\mathbf{m}_N^2\bar{\sigma}^2
\end{aligned}$$

Proof of Lemma D.2. Let $(a_{kc})_{k=1}^K, (b_{kc})_{k=1}^K$ be as defined in (Proof E). The errors are computed on $k \in [K]$ partition for the cluster $c \in [C]$

Step 1

$$\widehat{Q}-\tilde{Q}=\frac{1}{KC}\sum_{k=1}^{K}\sum_{c=1}^{C}[\underbrace{\mathbb{E}_{n,kc}(\tilde{D}_{it})(d_{i,0}(Z_{it})-\widehat{d}_i(Z_{it}))'}_{a_{kc}}+a'_{kc} \tag{E.3}$$

$$+\underbrace{\mathbb{E}_{n,kc}(d_{i,0}(Z_{it})-\widehat{d}_i(Z_{it}))(d_{i,0}(Z_{it})-\widehat{d}_i(Z_{it}))']}_{b_{kc}} \tag{E.4}$$

Step 2

$$\mathbb{E}[a_{kc}|\{Z_{it},(W_{it})_{i\in I_k^c}\}]=0$$

Step 3 By definition of $\mathbf{m}_N$,

$$\begin{aligned}\max_{1\leqslant m,j\leqslant d}\mathbb{E}[|b_{kc}||(W_{it})_{i\in I_k^c}]_{m,j}&=\max_{1\leqslant m,j\leqslant d}|\mathbb{E}[(d_{i,0}(Z_{it})-\widehat{d}_i(Z_{it}))(d_{i,0}(Z_{it})-\widehat{d}_i(Z_{it}))'|(W_{it})_{i\in I_k^c}]|_{m,j}\\&\leqslant \mathbf{m}_N^2\end{aligned}$$

Step 4 Fix the hold-out sample I_k^c. Conditionally on I_k^c, a_{kc} is mean zero $d\times d$ matrix with bounded entries. Azouma-Hoeffding inequality for martingale difference sequence implies:

$$\mathbb{E}[\max_{1\leqslant m,j\leqslant d^2}|\mathbb{E}_n a_{kc}|_{m,j}|(W_{it})_{i\in I_k^c}]\leqslant D^2\sqrt{\frac{\log(2d^2)}{N}}$$

Assumption 3.2 implies

$$\mathbb{E}[\max_{1\leqslant m,j\leqslant d^2}|\mathbb{E}_n b_{kc}-\mathbb{E}[b_{kc}|(W_{it})_{i\in I_k^c}]|_{m,j}|(W_{it})_{i\in I_k^c}]\leqslant\lambda_N$$

$$\max_{1\leqslant m,j\leqslant d^2}|\mathbb{E}_n a_{kc}|_{m,j}=O_P(D^2\sqrt{\frac{\log(2d^2)}{N}})$$

$$\max_{1\leqslant m,j\leqslant d^2}|\mathbb{E}_n b_{kc}-\mathbb{E}[b_{kc}|(W_{it})_{i\in I_k^c}]|_{m,j}=\lambda_N$$

$$\begin{aligned}\max_{1\leqslant m,j\leqslant d}|a+a'+b|_{m,j}&\leqslant\sum_{k=1}^{K}\max_{1\leqslant m,j\leqslant d}|a_{kc}+a'_{kc}+b_{kc}|_{m,j}\\&\lesssim_P K[\mathbf{m}_N^2+D^2\sqrt{\log(2d^2)/N}+\lambda_N]\end{aligned}$$

$\blacksquare$

Proof of Lemma D.3. For every $\delta = \widehat{\beta} - \beta_0, \delta \in R^d$ we use the notation:

$$\|\delta\|_{2,N} = (\mathbb{E}_N(\tilde{D}_{it}'\delta)^2)^{1/2}$$

and the in-sample predition error with respect to estimated residuals by

$$\|\delta\|_{\widehat{d},2,N} = (\mathbb{E}_N(\widehat{\tilde{D}}_{it}'\delta)^2)^{1/2}$$

The bound on the difference between $\|\delta\|_{\widehat{d},2,N}^2$ and $\|\delta\|_{2,N}^2$ is as follows:

$$\begin{aligned}
|\|\delta\|_{\widehat{d},2,N}^2 - \|\delta\|_{2,N}^2| &= \delta' |\mathbb{E}_N \widehat{\tilde{D}}_{it} \widehat{\tilde{D}}_{it}' - \mathbb{E}_N \tilde{D}_{it}\tilde{D}_{it}'|\delta \\
&\geqslant -q_N \|\delta\|_1^2
\end{aligned}$$

By definition of $\mathcal{RE}(\bar{c})$, forall $\delta \in \mathcal{RE}(\bar{c})$ the following holds:

$$\begin{aligned}
|\|\delta\|_{\widehat{d},2,N}^2 - \|\delta\|_{2,N}^2| &\geqslant -q_N \|\delta\|_1^2 \\
&\geqslant -q_N((1+\bar{c})|\delta_T|_1)^2 \\
&\geqslant -q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}\|\delta\|_{2,N}^2
\end{aligned}$$

Therefore,

$$\sqrt{1 - q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}} \leqslant \frac{\|\delta\|_{\widehat{d},2,N}}{\|\delta\|_{2,N}} \leqslant \sqrt{1 + q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}}$$

This implies a bound on $\kappa(\widehat{Q},T,\bar{c})$:

$$\begin{aligned}
\kappa(\widehat{Q},T,\bar{c}) := &\min_{\delta \in \mathcal{RE}(\bar{c})} \frac{\sqrt{s}\|\delta\|_{\widehat{d},2,N}}{\|\delta_T\|_1} \\
&\leqslant \min_{\delta \in \mathcal{RE}(\bar{c})} \frac{\sqrt{s}\|\delta\|_{2,N}}{\|\delta_T\|_1}\sqrt{1 + q_N \frac{(1+\bar{c})^2 s}{\kappa(\tilde{Q},T,\bar{c})^2}} \Rightarrow
\end{aligned}$$

$$\sqrt{\kappa(\tilde{Q},T,\bar{c})^2 - (1+\bar{c})^2 s} \leqslant \kappa(\widehat{Q},T,\bar{c}) \leqslant \sqrt{\kappa(\tilde{Q},T,\bar{c})^2 + (1+\bar{c})^2 s}$$

$\blacksquare$

Proof of Lemma D.4. Let

$$\widehat{\tilde{Y}}_{it} = \widehat{\tilde{D}}_{it}'\beta_0 + R_{it} + U_{it}$$

where

$$R_{it} = (\widehat{d}_i(Z_{it}) - d_{i,0}(Z_{it}))'\beta_0 + (l_0(Z_{it}) - \widehat{l}_i(Z_{it})), i \in \{1, 2, ..., N\}$$

summarizes first stage approximation error.

Step 1 Define the following quantities:

$$e_{k,m} = \mathbb{E}_{n,kc}(\widehat{d}_{m,0}(Z_{it}) - d_{m,0}(Z_{it}))U_{it}$$

$$f_{k,m} = \mathbb{E}_{n,kc}(\widehat{d}_{m,0}(Z_{it}) - d_{m,0}(Z_{it}))R_{it}$$

$$g_{k,m} = \mathbb{E}_{n,kc}\tilde{D}_{it}R_{it}$$

Step 2 Conditionally on I_k^c, $\mathbb{E}[e_{k,m}|I_k^c] = 0 \quad \forall k \in [K], m \in [d]$, $\mathbb{E}[g_{k,m}|I_k^c] = 0$.

Step 3 Conditionally on I_k^c,

$$\mathbb{E}[f_{k,m}|I_k^c] \leqslant \sup_{(d,l)\in(D_N,L_N)} \max_{1\leqslant m\leqslant d} (\mathbb{E}(d_m(Z_{it}) - d_{i,0}(Z_{it}))^2)^{1/2}(\mathbb{E}(R_{it})^2)^{1/2}$$
$$\leqslant \mathbf{m}_N[\mathbf{m}_N\sqrt{s} \vee \mathbf{l}_N]$$

Step 4 Conditionally on I_k^c, the terms e_{kc}, g_{kc} and demeaned term $(f_{k,m})^0 = f_{k,m} - \mathbb{E}[f_{k,m}|I_k^c]$ are bounded by maximal inequality for conditional expectation. Since the bound in RHS does not depend on I_k^c, the bound is also unconditional.

$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|e_{k,m}\||I_k^c] = O(\bar{\sigma}D\sqrt{\frac{\log d}{N}}),$$
$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|(f_{k,m})^0\||I_k^c] = O(\lambda_N),$$
$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|g_{k,m}\||I_k^c] = O([D^2 s + DL]\sqrt{\frac{\log d}{N}})$$

Therefore,

$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|e_m\|] = O(\bar{\sigma}D\sqrt{\frac{\log d}{N}}),$$
$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|(f_m)^0\|] = O([D^2 s + DL]\sqrt{\frac{\log d}{N}}),$$
$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|g_m\||] = O(\lambda_N)$$

■

Definition E.1 (First Stage Lasso-Panel). *Let $\eta^Y = [\eta_1^Y, ..., \eta_N^Y]'$ and $\eta^D = [\eta_1^D, ..., \eta_N^D]'$ be vector of individual heterogeneity parameters in outcome (Equation 2.1) and treatment (Equation 2.3), respectively. For every index k in the set of partition indices $[K]$, let*

$$\widehat{Q}_k(\alpha^D, \gamma^D, \eta^D) = \sum_{(i,t)\in I_k^c} (D_{i,t} - Z_{i,t}'\gamma^D - \eta_{it}^D)^2 \tag{E.5}$$

$$(\widehat{\alpha}_k^D, \widehat{\gamma}_k^D, \widehat{\eta}_k^D) = \arg\min \widehat{Q}_k(\alpha^D, \gamma^D, \eta^D) + \lambda_D \|\gamma^D\|_1 + \frac{\lambda_D}{\sqrt{N}} \|\eta^D\|_1 \tag{E.6}$$

$$\widehat{Q}_k(\alpha^Y, \gamma^Y, \eta^Y) = \sum_{(i,t)\in I_k^c} (Y_{i,t} - Z_{i,t}'\gamma^Y - \eta_{it}^Y)^2 \tag{E.7}$$

$$(\widehat{\alpha}_k^Y, \widehat{\gamma}_k^Y, \widehat{\eta}_k^Y, \widehat{\beta}) = \arg\min \widehat{Q}_k(\alpha^Y, \gamma^Y, \eta^Y, \beta) + \lambda_Y \|\gamma^Y\|_1 + \frac{\lambda_Y}{\sqrt{N}} \|\eta^Y\|_1 \tag{E.8}$$

Let $s_{\eta^Y}, s_{\eta^D}, s_{\gamma^Y}, s_{\gamma^D}$ denote the sparsity indices of $\eta^Y, \eta^D, \gamma^Y, \gamma^D$, respectively.

Theorem E.1 (Rate of First Stage Lasso-Panel). *Let $\lambda^D = \lambda^Y = ((4MN \log(p \vee N))^3)^{1/2}$ for each partition $k \in [K]$. By Theorem 1 from Kock and Tang (2016) the following rates hold with high probability:*

$$\|\widehat{\eta}_k^D - \eta_0^D\|_1 \lesssim_P \frac{\lambda^D s_{\eta^D}}{\kappa_2^2 \sqrt{N}T}$$
$$\|\widehat{\eta}_k^Y - \eta_0^Y\|_1 \lesssim_P \frac{\lambda^Y s_{\eta^Y}}{\kappa_2^2 \sqrt{N}T}$$
$$\|\widehat{\gamma}_k^Y - \gamma_0^Y\|_1 \lesssim_P \frac{\lambda^Y s_{\gamma^Y}}{\kappa_2^2 N}$$
$$\|\widehat{\gamma}_k^D - \gamma_0^D\|_1 \lesssim_P \frac{\lambda^D s_{\gamma^D}}{\kappa_2^2 N}$$

Corollary E.1 (Convergence of approximation error). *Suppose the components controls $Z_{i,t}$ are a.s. bounded:*

$$\|Z_{i,t}\|_\infty \leqslant C_Z$$

Then, the following out-of-sample squared approximation error of treatment and outcome reduced form exhibits the following bound:

$$\mathbb{E}_N(\widehat{p}_k(Z_{i,t}) - p_0(Z_{i,t}))^2 \lesssim_P \frac{\max(s_{\gamma^D}^2, s_{\eta^D}^2) \log^3(p \vee N)}{N}$$

and

$$\mathbb{E}_N(\widehat{l}_k(Z_{i,t}) - l_0(Z_{i,t}))^2 \lesssim_P \frac{\max(s^2_{\gamma^Y}, s^2_{\eta^Y}) \log^3(p \vee N)}{N}$$

Denote

$$\sqrt{d}\boldsymbol{m}_N = \frac{\max(s_{\gamma^D}, s_{\eta^D}) \log^{3/2}(p \vee N)}{\sqrt{N}}$$

and

$$\boldsymbol{l}_N = \frac{\max(s_{\gamma^Y}, s_{\eta^Y}) \log^{3/2}(p \vee N)}{\sqrt{N}}$$

Proof of Corollary E.1. Fix a partition $k \in [K]$.The choice of $\lambda^D = \lambda^Y = \sqrt{N \log(p \vee N)}$ yields the following bound:

$$\begin{aligned}
a_{kc} &= \frac{1}{N} \sum_{(i,t) \in I_k} (\widehat{d}_k(Z_{i,t}) - d_0(Z_{i,t}))^2 = \frac{1}{N} \sum_{(i,t) \in I_k} (Z_{i,t}(\widehat{\gamma}^D - \gamma^D_0) + \widehat{\eta}_{it} - \eta_{it})^2 \\
&\leqslant \frac{2}{N} \sum_{(i,t) \in I_k} (Z_{i,t}(\widehat{\gamma}^D - \gamma^D_0))^2 + \|\widehat{\eta}^D - \eta^D_0\|_2^2/N \\
&\leqslant 2\|Z_{i,t}\|_\infty \|\widehat{\gamma}^D - \gamma^D_0\|_1^2 + \|\widehat{\eta}^D - \eta^D_0\|_2^2/N \\
&\leqslant \|Z_{i,t}\|_\infty \frac{2\lambda^D s^2_{\gamma^D}}{\kappa_2^2(N)^2} + \frac{\lambda^D s^2_{\gamma^D}}{\kappa_2^2(N)^2} \lesssim_P d\mathbf{m}_N^2
\end{aligned}$$

$$\begin{aligned}
b_{kc} &= \sum_{(i,t) \in I_k} (\widehat{l}_k(Z_{i,t}) - l_0(Z_{i,t}))^2 = \frac{1}{N} \sum_{(i,t) \in I_k} (Z_{i,t}(\widehat{\gamma}^Y - \gamma^Y_0) + \widehat{\eta}^Y_{it} - \eta^Y_{it})^2 \\
&\leqslant \frac{2}{N} \sum_{(i,t) \in I_k} (Z_{i,t}(\widehat{\gamma}^Y - \gamma^Y_0))^2 + \|\widehat{\eta}^Y - \eta^Y_0\|_2^2/N \\
&\leqslant 2\|Z_{i,t}\|_\infty \|\widehat{\gamma}^Y - \gamma^Y_0\|_1^2 + \|\widehat{\eta}^Y - \eta^Y_0\|_2^2/N \\
&\leqslant \|Z_{i,t}\|_\infty \frac{2\lambda^Y s^2_{\gamma^Y}}{\kappa_2^2(N)^2} + \frac{\lambda^Y s^2_{\gamma^Y}}{\kappa_2^2(N)^2} \lesssim_P \mathbf{l}_N^2
\end{aligned}$$

Since K is a fixed finite number, $\sum_{k=1}^K a_{kc} \lesssim_P d\mathbf{m}_N^2$ and $\sum_{k=1}^K b_{kc} \lesssim_P \mathbf{l}_N^2$

∎

F Supplementary Statements without Proof

Example 6 (Smooth Function)**.** Let D be a scalar variable. Suppose the regression function $\mathbb{E}[Y|Z,D]$ is additively separable in D and controls Z:

$$Y = m_0(D) + g_0(Z) + U, \quad \mathbb{E}[U|Z,D] = 0$$

where the target function $m_0 \in \Sigma(\mathcal{X}, s)$ belongs to Holder s-smoothness class. Let the series terms $\{p_m(\cdot)\}_{m=1}^d$ with the sup-norm $\xi_d \equiv \sup_{D \in \mathcal{D}} \|p(D)\|_2$. Define β_0 as best linear predimor of m_0:

$$m_0(D) = \sum_{m=1}^{d} p_m(D)'\beta_0 + V, \quad \mathbb{E}p(D_{i,t})V_{i,t} = 0$$

where V is design approximation error with L^2 rate $r_d \to 0, d \to \infty$. Then, replacing Assumption 3.6 by a modified growth condition $\frac{\xi_d^2 \log N}{N} = o(1)$ yields the following L^2 rate on $\widehat{m}(D) = \sum_{m=1}^d p_m(D)'\widehat{\beta}$:

$$\|\widehat{\beta} - \beta\|_2 \lesssim_P \sqrt{\frac{d}{N}} + d\mathbf{m}_N^2\|\beta_0\| + \mathbf{l}_N\sqrt{d}\mathbf{m}_N + \sqrt{d/N}\sqrt{d}\mathbf{m}_N\|\beta_0\| + r_d$$
$$\|\widehat{m} - m\|_{F,2} \lesssim_P \sqrt{\frac{d}{N}} + d\mathbf{m}_N^2\|\beta_0\| + \mathbf{l}_N\sqrt{d}\mathbf{m}_N + \sqrt{d/N}\sqrt{d}\mathbf{m}_N\|\beta_0\| + r_d$$

Remark F.1 (Double Robustness in DML framework)**.** Orthogonal Least Squares is not doubly robust, since it places quality requirements on each of the treatment and outcome rates. In case one of the treatment or outcome reduced form is misspecified ($\sqrt{d}\mathbf{m}_N \not\to 0$ or $\mathbf{l}_N \not\to 0$), the estimator is inconsistent. A less stringent requirement is to ask for at least one regression to be correctly specified, namely:

$$\mathbf{l}_N\sqrt{d}\mathbf{m}_N = o(1)$$

This property is called double robustness . A doubly robust version of DML estimator can be obtained as follows:

Definition F.1 (Doubly Robust DML (DRDML))**.** *Let $(W_{i,t})_{i=1}^N = (Y_{i,t}, D_{i,t}, Z_{i,t})_{i=1}^N$ be a random sample from law P_N. Let the estimated values $(\widehat{d}(Z_{i,t}), \widehat{l}(Z_{i,t}))_{i=1}^N$ satisfy 3.1. Define Doubly Robust DML estimator:*

$$\widehat{\beta}_{DR} = \widehat{\beta}_{DR,(\widehat{d},\widehat{l})} = (\mathbb{E}_N[D_{i,t} - \widehat{d}(Z_{i,t})]D_{i,t}]')^{-1}\mathbb{E}_N[D_{i,t} - \widehat{d}(Z_{i,t})][Y_{i,t} - \widehat{l}(Z_{i,t})]']$$

Theorem F.1 (Rate of DRDML)**.** *Let Assumptions 3.1,3.2, 3.4, 3.3, 3.5 hold. Assume there is no approximation error $R = 0$. Then, w.p. $\to 1$,*

$$\|\widehat{\beta}_{DR} - \beta\|_2 \lesssim_P \boldsymbol{l}_N \sqrt{d} \boldsymbol{m}_N + \sqrt{\frac{d}{N}}$$

References

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016a). Inference in high dimensional panel models with an application to gun control. *Journal of Business and Economic Statistics*.

Belloni, A., Chernozhukov, V., and Kato, K. (2014). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016b). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Bühlmann, P. and van der Geer, S. (2011). Statistics for high-dimensional data. *Springer Series in Statistics*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et al. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6).

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013b). Testing many moment inequalities. *arXiv preprint arXiv:1312.7614*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *https://arxiv.org/pdf/1306.3171.pdf*.

Kock, A. B. and Tang, H. (2016). Uniform inference in high-dimensional dynamic panel data models. *Econometric Theory*.

Luo, Y. and Spindler, M. (2016). High-dimensional l2 boosting: Rate of convergence. *arXiv:1602.08927*.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *The Annals of Probability*, 2(4):620–628.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447.

Wager, S. and Athey, S. (2016). Estimation and inference of heterogeneous treatment effects using random forests. *https://arxiv.org/abs/1510.04342*.

Zhang, D. and Wu, W. B. (2015). Gaussian approximation for high dimensional time series. *https://arxiv.org/pdf/1508.07036.pdf*.