Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels

Vira Semenova, MIT, vsemen@mit.edu
Matt Goldman, Microsoft AI & Research, mattgoldman5850@gmail.com
Victor Chernozhukov, MIT, vchern@mit.edu
Matt Taddy, Microsoft AI & Research and
University of Chicago Booth School of Business, mataddy@outlook.com

Abstract

This paper provides estimation and inference methods for a large number of heterogeneous effects in the presence of a high-dimensional vector of control variables in a panel data setting. We allow the number of heterogeneous groups and the number of controls to exceed the sample size. To make informative inference possible, we assume that the treatment effect has a sparse representation and estimate it in two stages. In the first stage, we estimate the reduced form for treatment and outcome on an auxiliary sample. In the second stage, we estimate the treatment effect by ℓ_1-regularized least squares regression of outcome residuals on the treatment residuals. The proposed estimator presents an improved convergence rate over a one-stage alternative where the treatment effect and the control function are estimated jointly. In addition, we provide inference methods for single and multiple coefficients of the heterogeneous effects. We use a correlated random effects approach to model unobserved unit heterogeneity by extending the approaches from Mundlak (1978) and Chamberlain (1982) to a high-dimensional setting. Using the data from a major food distributor, we apply our method to estimate price elasticities for a large number of heterogeneous products and validate our estimates using experimental price variation.

1 Introduction

Estimation of counterfactual outcomes is a key aspect of economic analysis and calls for a large portion of the economists' efforts in both industry and academia. In the absence of explicit exogenous variation – i.e., independence or random assignment – economists must rely on human judgment to specify a set of controls that allows them to assign a causal interpretation to their estimates. This method of model selection is highly subjective and labor intensive. In recent years there has been a growing interest in the use of machine learning (ML) tools to automate and accelerate the economist's model selection process (Athey, 2017). This paper leverages ML tools to deliver a high-quality point estimate and valid confidence intervals for single and many coefficients of a high-dimensional treatment effect in a panel setting.

We identify treatment effects using the residual variation in realized treatment after conditioning on all exogenous covariates. This approach relies on the correct recovery of the treatment reduced form, that is the conditional expectation function of treatment given controls. Since its functional form is unknown, we construct a large number of technical control variables (e.g., interactions, power series) to accurately capture the conditional expectation and must apply modern ML tools to deliver satisfactory precision of the reduced form estimate. The recent Double Machine Learning (Double ML) framework of Chernozhukov et al. (2016) has shown how to leverage ML tools into high-quality estimates of a (low-dimensional) treatment effect in a cross-sectional setting. We extend this framework to allow the high dimension of the treatment effect in panel data setting.

The proposed estimator of the treatment effect consists of two stages. In the first stage, we estimate the conditional expectation of outcome and treatment conditionally on all pre-determined variables (e.g., lagged realizations of treatment and outcome as well as any other exogenous predictors). In the second stage, the original Double Machine Learning estimator computes ordinary least squares (OLS) estimates of the outcome and treatment residuals.[1] Extending this estimator to cases of high-dimensional treatment, we use Lasso in the second stage to construct point estimates and the debiased Lasso to construct confidence intervals, referring to these two-stage

[1]Residualization of both treatment and outcome (as well as sample-splitting, which is omitted from this brief description) is necessary for the orthogonality of the resulting estimator which as shown in Chernozhukov et al. (2016) guarantees valid inference even in cases where our ML algorithms converge at less than a root-N rate.

estimators as Orthogonal Lasso and Double Orthogonal Lasso, respectively. We argue that with our proposed extensions, the Double ML framework is suitable for a broad set of applied econometric problems.

We focus on firm-side demand analysis. Consider a manager trying to estimate the demand elasticities for a wide catalog of products. She has access to internal data containing rich product characteristics (textual descriptions, consumer reviews, product images) and the universe of demand signals that were used in strategic price-setting. She can then reconstruct the dynamic information set available at the time of the original pricing decision. And then when training her first stage regression onto treatment can recover both the systematic component of her firm's pricing policy as well as an idiosyncratic component that can be regarded as plausibly exogenous. In this context, our panel Double ML framework naturally facilitates identification of demand elasticities using the standard isoelastic demand specification. However, our application of modern ML methods allows us greater generality in modeling the firm's pricing policy and thus greater confidence in uncovering plausibly exogenous variation in treatment.

Additionally, our manager must set prices for a large number of products and will require precise estimates of elasticities at that level. Aggregate or category level estimates could be recovered from a model with low-dimensional treatment, but these may obscure important heterogeneity. Alternatively, unrestricted estimates for any each product may be too imprecise to be useful. To overcome this problem, the manager may instead assume that the set of products can be partitioned into a small number of groups with a constant elasticity, but is unaware of this partition. However, available product descriptions could be used to classify the products into a hierarchy that captures the important aspects of product heterogeneity. As a result, there exists a subset of these hierarchical nodes that approximates the ideal partition up to a small number of misclassified items. This structure naturally motivates a high-dimensional sparse framework for treatment effects, under which we can deliver both satisfactory precision and valid inference for product-level elasticities.

Formally, we refer to the structure described above as our high-dimensional sparse (HDS) regime. Here, we use Lasso in the final step to construct point estimates and the debiased Lasso to construct confidence intervals, referring to these two-stage estimators as Orthogonal Lasso and Double Orthogonal Lasso, respectively. We also consider a Low Dimensional (LD) regime, where the dimension of treatment is allowed

to grow (albeit at a much slower rate) and no sparsity is imposed. Here, we retain OLS as our causal estimator referring to the combined estimator as Orthogonal Least Squares. In all cases, we use the cross-fitting algorithm outlined in Chernozhukov et al. (2016) so as to avoid any complexity restrictions on our ML estimators in our first stage estimations.

The theoretical contribution of the paper consists in establishing the validity of Orthogonal Least Squares, Orthogonal Lasso and Double Orthogonal Lasso. We establish the rate of convergence and the asymptotic Gaussianity for Orthogonal Least Squares, allowing the dimension of treatment to grow with the sample size. In the HDS regime, we establish the convergence rate of the Orthogonal Lasso estimator. Under mild conditions this rate is the same as if the true values of the first-stage reduced forms were known to the econometrician and (so long as the complexity of our control function is greater than the complexity of our treatment) superior to one-stage procedures where treatments are not explicitly residualized. As for inference, we establish asymptotic Gaussianity of single and multiple coefficients of Double Orthogonal Lasso assuming that the population covariance matrix of the interacted price residuals is sufficiently sparse. In all cases, these results rely only on high-level conditions on the performance of the ML algorithms used for the first two estimations.

In order to confidently apply our methods to a panel setting, we must justify these high-level conditions for some class of underlying data generating processes and ML algorithms. To do so, we adopt a correlated random effects approach to deal with unobserved item heterogeneity and extend it in a high dimension. Specifically, we assume that the systematic component of unobserved item heterogeneity is decomposed into a item-invariant function of control variables (i.e., pooled control function) and an item-specific constant. The norm of the vector of the item-specific constants is assumed to be sufficiently small (i.e, weakly sparse as in Negahban et al. (2012)). Under these assumptions, we estimate this function of the control variables and the item-specific constants using a ML method, combined with an ℓ_1-regularization for the vector of the item-specific constants. In the case the control function is a linear and sparse, the proposed method coincides with the dynamic panel Lasso of Kock and Tang (2016). The proposed approach is an extension of the approaches of Mundlak (1978) and Chamberlain (1982) who worked in low-dimensional settings.

Finally, we use the proposed methodology to estimate demand elasticities for a major European food distributor. The data consists of the hierarchical description of

the products, prices, and daily aggregate sales for each (store, product, distribution channel) combination. We posit a partially linear log-log demand specification where the (log) sales are used as the dependent variable; lags of prices and sales as well as current product characteristics are used as the control variables acting non-linearly on the sales; and the interactions of the log price with the control variables are used as used treatment variables. At the first level of the hierarchy our estimates of own-price elasticities based on Orthogonal Least Squares are close to the results of the category-level demand studies (e.g., Chevalier et al. (2003)). At the deeper levels of hierarchy, we find benefit from imposing the sparsity constraint implied by our HDS framework. A subset of our estimated price elasticities are validated experimentally based on randomly assigned promotions for the particular group of products that we selected.

Literature review. This paper is built on two lines of research: orthogonal semi-parametric estimation and estimation of panel data models. The first line (Andrews (1994), Newey (1994), van der Vaart (1998)) is concerned with estimation of a low-dimensional target parameter defined by a moment equation that depends on a high-dimensional nuisance parameter. This line introduces the concept of Neyman-orthogonality (Neyman (1959)), that is, insensitivity of the moment equation that defines the target parameter to the biased estimation of the nuisance parameter. Combined with the sample splitting idea, Neyman-orthogonality was used in Chernozhukov et al. (2017a) and Chernozhukov et al. (2017b) to estimate the nuisance parameter by modern ML methods and deliver a high-quality, root-N consistent asymptotically normal estimator of the target parameter by solving the sample analog of the estimated moment equation. Relying on this idea, we allow the dimension of the target parameter to grow slowly with the sample size (LD regime) and exceed the sample size under the sparsity assumption (HDS regime).

The second line is concerned with the estimation of unobserved item heterogeneity in low and high-dimensional panel data models. The first approach (e.g., Belloni et al. (2016a)) models the unobserved item heterogeneity as a vector of unknown parameters and suggests partialling it out (for example, first differencing). However, this approach requires the reduced forms to be linear functions of the controls. Furthermore, the standard second-stage algorithms applied to the first-differenced residuals are sensitive to the first-stage bias of the reduced form functions. For that

reason we do not use this approach. The alternative approach by Mundlak (1978) and Chamberlain (1982), introduced in the low-dimensional regime, projects the unobserved heterogeneity onto the space of time-invariant control variables. We take this approach by allowing the projection function to be the sum of a nonlinear, highly complex function of the observables, and an item-specific constant, requiring the vector of item-specific constants to be weakly sparse (Negahban et al. (2012)). In the case the highly complex function of the observables is a linear and sparse function of the control variables, the proposed estimator coincides with Kock and Tang (2016) dynamic panel Lasso.

Structure of the paper. The rest of the paper is organized as follows. Section 2 describes our partially linear framework and gives some motivating examples. Section 3 provides our main theoretical results for the Low Dimensional and High Dimensional Sparse regimes. Section 4 discusses strategies and results for the first stage estimation of treatment and outcome. Section 5 presents the empirical results from our collaboration with the food distributor.

2 Set-Up and Motivation

Consider the following partially linear sequentially exogenous dynamic panel model by Robinson (1988):

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

$$D_{it} = d_{i0}(Z_{it}) + V_{it}, \qquad \mathbb{E}[V_{it}|Z_{it}, \Phi_t] = 0, \tag{2.2}$$

where Y_{it} is a scalar outcome of item i at time t, D_{it} is a d-vector of treatments, and Z_{it} is a p_Z-vector of control variables, respectively. The controls Z_{it} affect the treatment vector D_{it} and the outcome Y_{it} through nonlinear item-specific functions $d_{i0}(\cdot)$ and $e_{i0}(\cdot)$, respectively. The set $\Phi_t = \sigma(\{Y_{i,k}, P_{i,k}, Z_{i,k}\}_{k=1}^{t-1})$ denotes the full information set available prior to period t. In practice we will assume that this set is well approximated by several lags of outcome and treatment variables. In order to enable a causal interpretation for the treatment effect, we make the conventional assumption of **conditional sequential exogeneity**, according to which the stochastic shock U_{it} governing the potential outcomes is mean independent of the past information Φ_t,

the controls Z_{it}, and the treatment vector D_{it}.

Equation (2.1) is the main equation, and the treatment effect β_0 is the object of interest. Equation (2.2) is the secondary equation that keeps track of the confounding, that is the effect of the vector of controls Z_{it} on the treatment D_{it}. We consider two different regimes for the treatment effect β_0: a low-dimensional (LD) regime, where the dimension $d = d(N)$ grows sufficiently slow, and a high-dimensional sparse (HDS) regime, where the number of treatments d is allowed to be larger than the sample size (i.e, $d \gg N$), but the number s of the treatments whose effect is not equal to zero is constrained to be sufficiently small.

The set of items $[I]$ consists of G independent groups of size C each. Denote the set of items by:

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

where $g = g(i)$ stands for the index of the group and $c = c(i)$ is the number of item i within the group $g(i)$. We assume that the observed data $\{Y_{it}, D_{it}, Z_{it}\}_{(i,t)=(1,1)}^{I,T}$ are i.i.d across groups. For each group $g(i)$ and each $c \in \{1, 2, \ldots, C\}$, the sequences of the disturbances $\{(V_{it}, U_{it})_{g(i)=g, c(i)=c}\}_{t=1}^{T}$ are martingale difference sequences.

Now define the *treatment reduced form*:

$$d_{i0}(z) \equiv \mathbb{E}[D_{it}|Z_{it} = z, \Phi_t] = \mathbb{E}[D_{it}|Z_{it} = z] \tag{2.3}$$

and the *outcome reduced form*:

$$l_{i0}(z) \equiv \mathbb{E}[Y_{it}|Z_{it} = z, \Phi_t] = \mathbb{E}[Y_{it}|Z_{it} = z]. \tag{2.4}$$

Define the corresponding residuals of treatment and outcome:

$$\tilde{D}_{it} \equiv D_{it} - d_{i0}(Z_{it}), \quad \tilde{Y}_{it} \equiv Y_{it} - l_{i0}(Z_{it}). \tag{2.5}$$

Equation (2.1) implies a linear relationship between the outcome and the treatment residuals:

$$\tilde{Y}_{it} = \tilde{D}_{it}'\beta_0 + U_{it}, \qquad \mathbb{E}[U_{it}|\tilde{D}_{it}, Z_{it}, \Phi_t] = 0, \tag{2.6}$$

which we will also refer to as **partialled out, or residualized** form. Equation (2.1)

motivates a two-stage estimator of the target parameter β_0. First, we estimate the treatment and the outcome reduced forms $\widehat{d}_i(\cdot), \widehat{l}_i(\cdot)$ and the respective residuals:

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

using modern ML methods. Second, we apply ordinary least squares in the LD regime and Lasso and debiased Lasso in the HDS regime, where we use the estimated outcome residual $\widehat{\tilde{Y}}_{it}$ as the dependent variable and the estimated treatment residual $\widehat{\tilde{D}}_{it}$ as the vector of regressors. We will use different samples between the first and the second stage, as is further described in the Panel Double ML recipe.

A key insight of this approach is that the second-stage estimator of the treatment effect is not sensitive to the estimation error of the treatment and outcome reduced forms. The insensitivity comes from the **orthogonality property**:

$$\mathbb{E}\tilde{D}_{it}(\widehat{l}_i(Z_{it}) - l_{i0}(Z_{it})) = 0, \quad \mathbb{E}\tilde{D}_{it}(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it})) = 0, \tag{2.7}$$

that is, the residual of the treatment $\tilde{D}_{it}$ is uncorrelated with the first-stage estimation error of the reduced forms. This insight allows us to deliver a high-quality point estimator and the confidence intervals for the treatment effect β_0 in both regimes in the presence of the highly complex control functions $d_{i0}(\cdot)$ and $l_{i0}(\cdot)$.

In many relevant cases, such as in Examples 1 and 2, the high-dimensional treatment D_{it} is a technical treatment that comes from the linear transformation of some "base" treatment variable P_{it}:

$$D_{it} = A(Z_{it})P_{it}, \tag{2.8}$$

where P_{it} is a low-dimensional treatment variable and $A(\cdot) : \mathcal{R}^{dz} \to \mathcal{R}^d \bigtimes \mathcal{R}^p$ is a known matrix-valued function of the control vector Z_{it}. If this is the case, it is possible to simplify the two-stage procedure described above. Define the base treatment reduced form as:

$$P_{it} = p_{i0}(Z_{it}) + V_{it}^P, \quad \mathbb{E}[V_{it}^P | Z_{it}, \Phi_t] = 0 \tag{2.9}$$

and the base treatment residual as $\tilde{P}_{it} \equiv P_{it} - p_{i0}(Z_{it})$. By construction, the treatment reduced form $d_{i0}(Z_{it}) = A(Z_{it})p_{i0}(Z_{it})$. This relation allows us to replace the

estimation of the high-dimensional reduced form (2.3) by the low-dimensional base reduced form (2.9) and exploit the linearity relation above.

2.1 Motivating Examples

Example 1 (**Heterogeneous Treatment Effects with Modeled Heterogeneity**)**.** Consider the model described by Equations (2.1)-(2.2). Suppose that the technical treatment vector D_{it} is generated according to the following equation:

$$D_{it} = X_{it}P_{it} \tag{2.10}$$

where P_{it} is a scalar "base" treatment variable and $X_{it} = (1, \tilde{X}_{it})$ where $\mathbb{E}\tilde{X}_{it} = 0$ is a d-vector of observable characteristics of unit i that is a subset of the control vector Z_{it} (i.e, $X_{it} \subseteq Z_{it}$). Equation (2.10) is the special case of (2.9) where the matrix $A(Z_{it}) = X_{it}$. The treatment effect

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

consists of the Average Treatment Effect (ATE) α_0 and the Treatment/Structural Modification Effect (TME) γ_0. In other words, the causal effect of the unit difference in the base treatment P_{it} on the outcome Y_{it} is equal to the sum of the ATE and the TME:

$$\Delta D_{it}'\beta_0 = \Delta P_{it} X_{it}\beta_0 = \underbrace{\alpha_0}_{\text{ATE}} + \underbrace{X_{it}'\gamma_0}_{\text{TME}}.$$

Allowing the parameter γ_0 to be high-dimensional allows us to model the heterogeneity of the treatment effect for a fine granularity.

Example 2 (**Heterogeneous Own and Cross-price elasticities with Many Heterogeneous Products**)**.** Consider a firm that makes a pricing decision about a large number I of heterogenous goods. Suppose that in the short term the realizations of prices and sales can be approximated by the following partially linear model:

$$\begin{aligned} Y_{it} &= D_{it}'\beta_0 + e_{i0}(Z_{it}) + U_{it}, \quad \mathbb{E}[U_{it}|D_{it}, Z_{it}, \Phi_t] = 0, \\ D_{it} &= [P_{it}X_{it}, P_{-it}X_{it}], \\ P_{it} &= p_{i0}(Z_{it}) + V_{it}^P, \quad \mathbb{E}[V_{it}^P|Z_{it}, \Phi_t] = 0, \end{aligned} \tag{2.11}$$

where Y_{it} is the log sales of product i at time t, P_{it} is the log price, $X_{it} = (1, \tilde{X}_{it}), \quad \mathbb{E}\tilde{X}_{it} = 0$ is a d-vector of observable characteristics, and Z_{it} is a p-vector of the observable product characteristics X_{it}, the lagged realizations of market quantities, and the demand-side variables used for strategic price setting by the firm. The symbol Φ_t denotes the full information set available prior to period t, spanned by lagged realizations of the demand system. The controls Z_{it} affect the price variable P_{it} through $p_{i0}(Z_{it})$ and the sales through $e_{i0}(Z_{it})$.

Let C_i be the set of products that have a non-zero cross-price effect on the sales Y_{it}. For product i, define the average leave-i-out price of products in C_i as:

$$P_{-it} \equiv \frac{\sum_{j \in C_i} P_{jt}}{|C_i|}. \tag{2.12}$$

The technical treatment D_{it} is formed by interacting P_{it} and P_{-it} with the observable product characteristics X_{it}. The parameter β_0 stands for the vector of own and cross-price elasticities. In order to assign a causal interpretation to β_0, we assume that conditionally on all pre-determined variables the sales shock U_{it} is mean independent of the past information Φ_t, the controls Z_{it}, the own price P_{it} and the prices of co-dependent goods P_{-it}.

Equation (2.9) 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 the own and the cross-price effect, respectively. A change in the own price ΔP_{it} affects the demand via

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

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

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

Let

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

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 the Heterogenous Cross-Price Elasticity.

Note that the definition of P_{-it} (2.12) implies the very strong restriction that any two products j and k have the same cross-price impact on a third product i. This is certainly an unrealistic depiction of cross-price elasticities. If (for example) we believed our products to be involved in logit competition, we might prefer to construct $P_{-it} \equiv \sum_{j \neq i} \omega_j \cdot P_{jt}$, with ω_j proportional to the popularity of product j. However, there is no need to assume a particular theory of competition. Instead, we include *all possible* constructions of cross-price variables corresponding to *all plausible* theories of competition and assume that only one theory is correct. This assumption corresponds to the modeling assumption about the sparsity of the treatment effect β_0, which motivates our HDS framework. We believe that this way of modeling the cross-product effects is more flexible than the structural approach, where the substitution pattern between different products is required to follow the logit model (see., e.g., Gandhi and Houde (2016)).

A potential concern for the proposed approach is the presence of unobserved item heterogeneity that is built into the item-specific reduced form functions and the availability of the ML methods to capture this heterogeneity. We resolve this concern using the following approach.

Remark 2.1 (Weakly Sparse Unobserved Heterogeneity)**.** Suppose the high-dimensional treatment vector D_{it} is generated from a low-dimensional "base" treatment vector as in (2.8). Let $p_{i0}(\cdot)$ and $l_{i0}(\cdot)$ be the reduced forms of base treatment and outcome, defined in (2.9) and (2.4). Suppose that the reduced form functions can be decomposed into item-homogeneous and time-invariant parts:

$$p_{i0}(Z_{it}) = p_0(Z_{it}) + \xi_i, \tag{2.13}$$

$$l_{i0}(Z_{it}) = l_0(Z_{it}) + \eta_i, \tag{2.14}$$

where the functions $p_0(\cdot)$ and $l_0(\cdot)$ are functions of the controls, and ξ_i, η_i are the unobserved time-invariant random variables. Let $\bar{Z}_i$ be the set of time-invariant item characteristics that, in particular, include the averages of the control variables Z_{it}

across the observed time span. Assume that the time-invariant random variables ξ_i, η_i can be decomposed as follows:

$$\xi_i = \lambda_0^d(\bar{Z}_i) + a_i, \tag{2.15}$$

$$\eta_i = \lambda_0^g(\bar{Z}_i) + b_i, \tag{2.16}$$

where $\lambda_0^d(\cdot)$ and $\lambda_0^g(\cdot)$ are some functions of time-invariant item characteristics, $(a_i)_{i=1}^I$ is a sequence of unknown p-vectors, and $(b_i)_{i=1}^I$ is a sequence of numbers. That is, the only randomness in ξ_i and η_i comes from the randomness in $\bar{Z}_i$. We assume that the product characteristics $\bar{Z}_i$ are sufficiently rich so that the remaining unobserved heterogeneity a_i is small. We impose the following weak sparsity assumption by Negahban et al. (2012).

Assumption 2.1 (Weak Sparsity of Fixed Effects)**.** *Suppose that the relations defined in (2.13) - (2.16) hold. Assume that the I-vectors of the fixed effects are weakly sparse. That is, there exists a constant $\nu \in (0,1)$ and an upper bound on the sparsity index s such that the ν-norm of the vectors $a = (a_1, a_2, \ldots, a_I)$ and $b = (b_1, b_2, \ldots, b_I)$ is bounded by s:*

$$\sum_{i=1}^{I} |a_i|^\nu \leqslant s, \quad \sum_{i=1}^{I} |b_i|^\nu \leqslant s.$$

To sum up, the set of unknown parameters to be estimated in the first stage consists of the functions $d_0(\cdot), l_0(\cdot), \lambda_0^d(\cdot), \lambda_0^g(\cdot)$ and the vectors a and b.

Assumption 2.1 introduces a middle-ground approach between the random effects and the fixed effects approach in the panel data literature. It is built on the original approaches of Mundlak (1978) and Chamberlain (1982) and extends them in several important directions. The original idea of Mundlak (1978) was to project the unobserved heterogeneity ξ_i in (2.13) onto a linear form of the observed time-invariant control variables $\bar{Z}_i$ treating the residual a_i as a random effect. The first extension we provide consists in the functions $d_0(\cdot), l_0(\cdot), \lambda_0^d(\cdot), \lambda_0^g(\cdot)$ being nonlinear and highly complex functions of a high-dimensional control vector. The second one consists in treating the residual unobserved item heterogeneity $a_i, i \in [I]$ as a fixed unknown parameter to be estimated as opposed to the realization of an i.i.d random variable that is uncorrelated with the controls. In the case the functions $\lambda_0^d(\cdot), \lambda_0^g(\cdot)$ are linear

and sparse, this approach was used in Kock and Tang (2016).

We summarize this section with the informal description of the Panel Double ML Recipe.

Definition 2.1 (Panel Double ML Recipe)**.** *1. Split the data into a K-fold partition by time index where 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 the first stage residuals according to (2.5)*. For each data point i, use the first stage estimators whose index corresponds to their partition.*

4. Pool the first stage residuals from all partitions and estimate $\widehat{\beta}$ by applying the second stage estimator from Section 3.1 or 3.2 depending on its regime of β_0.

3 Theoretical Results

In this section we establish an asymptotic theory for our estimators under high-level conditions whose plausibility we discuss in Section 4. Denote the sample average of a function $f(\cdot)$ as:

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

and the centered and scaled sample average as:

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

In this section we provide high-level restrictions on the econometric model defined in Equations (2.1)-(2.2). They consist of assumptions on the performance of the first-stage estimators (Assumptions 3.1 and 3.2), standard identifiability (Assumption 3.4) and light tails (Assumption 3.5) conditions on the outcome disturbances and the treatment residuals. We also assume the Law of Large Numbers for the sample covariance matrices of the treatment residuals (Assumption 3.3).

Assumption 3.1 imposes the restriction of the quality of estimation of the treatment and outcome reduced forms denoted by $\mathbf{d}_N$ and $\mathbf{l}_N$, respectively. It introduces sequences of neighborhoods $\{D_N, N \geqslant 1\}$ and $\{L_N, N \geqslant 1\}$ around the respective true values of the treatment reduced form $d_{i0}(\cdot)$ and the outcome reduced form $l_{i0}(\cdot)$. These neighborhoods contain the respective first stage estimates $\widehat{d}_i(\cdot)$ and $\widehat{l}_i(\cdot)$ with probability approaching one. As the sample size N increases, the neighborhoods shrink. The quality of estimation is defined as the rate of shrinkage of the neighborgoods $\{D_N, N \geqslant 1\}$ and $\{L_N, N \geqslant 1\}$ around the respective reduced form functions $d_{i0}(\cdot)$ and $l_{i0}(\cdot)$.

Assumption 3.1 (Small Bias Condition)**.** *There exists a sequence of realization sets $D_N \subset \mathcal{R}^d$ and $L_N \subset \mathcal{R}$ such that the following conditions hold. (1) The true value $d_{i0}(\cdot)$ and $l_{i0}(\cdot)$ of the treatment and outcome reduced forms belong to D_N and L_N for all $N \geqslant 1$. (2) There exists a sequence of numbers $\phi_N = o(1)$ such that the first stage estimator $\widehat{d}_i(\cdot)$ of $d_{i0}(\cdot)$ and $\widehat{l}_i(\cdot)$ of $l_{i0}(\cdot)$ belong to D_N and L_N with probability at least $1 - \phi_N$, respectively. (3) There exist sequences $\boldsymbol{d}_N, \boldsymbol{l}_N$ such that the realization sets D_N and L_N shrink around $d_{i0}(\cdot)$ and $l_{i0}(\cdot)$ at the following rate:*

$$\boldsymbol{d}_N \equiv \sup_{d \in D_N} (\mathbb{E}\|d(Z) - d_{i0}(Z)\|^2)^{1/2},$$

$$\boldsymbol{l}_N \equiv \sup_{l \in \mathcal{L}_N} (\mathbb{E}(l(Z) - l_{i0}(Z))^2)^{1/2}.$$

(4) There exist constants D and L that bound the treatment and the outcome reduced forms almost surely:

$$\sup_{d \in D_N} \max_{1 \leqslant j \leqslant d} |d_j(Z_{it})| < D, \quad \sup_{l \in L_N} |l(Z_{it})| < L.$$

(5) The rates $\boldsymbol{d}_N$ and $\boldsymbol{l}_N$ are sufficiently small. Namely, there exists $0 < \delta < \frac{1}{2}$:

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

We assume the following bound on the convergence of the centered sample average.

Assumption 3.2 (Concentration)**.**

$$\sqrt{N}\lambda_{1N} \equiv \sup_{d\in D_N} \max_{1\leqslant j\leqslant d} |\mathbb{G}_N(d_j(Z_{it}) - d_{i0,j}(Z_{it}))^2| \lesssim_P o_P(1),$$

$$\sqrt{N}\lambda_{2N} \equiv \sup_{d\in D_N} \sup_{l\in\mathcal{L}_N} \max_{1\leqslant j\leqslant d} |\mathbb{G}_N(d_j(Z_{it}) - d_{i0,j}(Z_{it}))(l(Z_{it}) - l_{i0}(Z_{it}))| \lesssim_P o_P(1).$$

Assumption 3.3 (Law of Large Numbers for Sample Covariance Matrices)**.** *Suppose the treatment residuals* $(\tilde{D}_{gt})_{(g,t)=(1,1)}^{G,T}$ *follow a stationary process with bounded realizations. Denote their population covariance matrix as:* $Q \equiv \mathbb{E}\tilde{D}_{gt}^{\top}\tilde{D}_{gt}$*. We assume that the sample covariance matrix of the residuals* $\mathbb{E}_N\tilde{D}'_{gt}\tilde{D}_{gt}$ *converges to the population covariance matrix* Q *as the sample size* N *and the dimension* $d = d(N)$ *grow:*

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

Assumption 3.3 was shown for the i.i.d case by Rudelson (1999). We conjecture that this assumption continues to hold under the exponential mixing condition for the martingale difference sequence $(\tilde{D}_{gt})_{(g,t)=(1,1)}^{G,T}$.

Assumption 3.4 (Identification)**.** *Denote the population covariance matrix of treatment residuals by* Q*:* $Q \equiv \mathbb{E}\tilde{D}'_{gt}\tilde{D}_{gt}$*. Assume that there exists constants* C_{min}, C_{max} *such that* $0 < C_{min} < C_{max} < \infty$ *and* $C_{\min} < \min \operatorname{eig}(Q) < max \operatorname{eig}(Q) < C_{max}$*.*

Assumption 3.4 states that the treatments $\tilde{D}_{gt}$ are not too collinear in population, a requirement needed for the identification of the treatment effect β_0.

Assumption 3.5 (Lindeberg Condition)**.** *The following conditions hold. (1) The norm of the treatment residual* $\|\tilde{D}_{gt}\|$ *is bounded a.s. (2). The sequence* $\{\mathbb{E}\|U_{gt}U'_{gt}\|1_{\|U_{gt}U'_{gt}\|>M}0\}_{M=1}^{\infty}$ *converges to zero as* $M \to \infty$*. (3) There exists* $q > 2$ *such that* $(\mathbb{E}\|U_{gt}U'_{gt}\|^q)^{1/q} \leqslant \infty$*.*

Assumption 3.5 imposes technical conditions for the asymptotic theory. Since a bounded treatment $\tilde{D}_{gt}$ is a plausible condition in practice, we impose it to simplify the analysis. In addition, we require the disturbances U_{gt} to have light tails as stated in the Lindeberg condition.

3.1 Low Dimensional Regime

In this section we consider the Low-Dimensional (LD) regime. The exact bound on the growth rate of the number of treatment is determined by the quality of the first

stage estimation summarized by the treatment ($\mathbf{d}_N$) and outcome ($\mathbf{l}_N$) rates. For example, if Assumption 3.1 (5) holds, the growth rate of the number of treatments d in LD regime is required to be sufficiently small: $d = o(N^{4\delta/3})$.

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

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

where the second and third lines implicitly define the estimators of the treatment residuals $\widehat{\tilde{D}}_{it}$, their sample covariance matrix $\widehat{Q}$, and the outcome residuals $\widehat{\tilde{Y}}_{it}$.

Orthogonal Least Squares is the first estimator proposed in this paper. As suggested by its name, it performs ordinary least squares using the estimated treatment residual $\widehat{\tilde{D}}_{it}$ as the regressor and the estimated outcome residual $\widehat{\tilde{Y}}_{it}$ as the outcome variable. In the case the dimension d is treated as fixed, this estimator coincides with the Double Machine Learning estimator of Chernozhukov et al. (2017a). Allowing the dimension $d = d(N)$ to grow with the sample size N is a novel feature of this paper.

Assumption 3.6 (Dimensionality Restriction)**.**

(a) There exists a constant $C > 0$ such that $\forall j \in \{1, 2, \ldots, d\}$ the coefficient $|\beta_{0,j}|$ is bounded by C: $|\beta_{0,j}| < C$.

(b) For the quality parameter $\delta > 0$ defined in Assumption 3.1(5) the number of treatments d grows sufficiently slow: $d = o(N^{4\delta/3})$.

Assumption 3.6 imposes growth restrictions on the treatment dimension. The first restriction ensures that each component of the true treatment vector is bounded. The second restriction $d = o(N^{4\delta/3})$ defines the treatment growth rate relative to the quality of the first stage treatment estimator.

Theorem 3.1 (Orthogonal Least Squares)**.** *Suppose Assumptions 3.3, 3.4, 3.5(1) and 3.6 hold. Then the following statements hold.*

(a) The mean squared error of the Orthogonal Least Squares estimator exhibits the following bound:

$$\begin{aligned}\|\widehat{\beta} - \beta_0\|_2 &\lesssim_P \sqrt{\frac{d}{N}} + \boldsymbol{d}_N^2\|\beta_0\|_2 + \boldsymbol{l}_N \boldsymbol{d}_N \\ &\lesssim_P \sqrt{\frac{d}{N}} + \boldsymbol{d}_N^2\sqrt{d} + \boldsymbol{l}_N \boldsymbol{d}_N.\end{aligned}$$

In addition, if Assumption 3.1 (5) holds, Orthogonal Least Squares achieves the oracle rate:

$$\|\widehat{\beta} - \beta_0\|_2 \lesssim_P \sqrt{\frac{d}{N}}.$$

(b) For any vector α on a unit sphere $\mathcal{S}^{d-1} := \{q \in \mathcal{R}^d, \|q\| = 1\}$ the estimator $\alpha'\widehat{\beta}$ of the projection $\alpha'\beta_0$ is asymptotically linear:

$$\sqrt{N}\alpha'(\widehat{\beta} - \beta_0) = \alpha' Q^{-1} \mathbb{G}_N \tilde{D}_{it} U_{it} + R_{1,N}(\alpha),$$

where the remainder term $R_{1,N}(\alpha)$ exhibits the following bound: $R_{1,N}(\alpha) \lesssim_P \sqrt{N}(\boldsymbol{d}_N^2\|\beta_0\| + \boldsymbol{d}_N \boldsymbol{l}_N)$.

(c) Denote the asymptotic covariance matrix of the clustered standard errors as follows:

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

where $\tilde{D}_{gt}$ and U_{gt} are the $d \bigtimes C$ matrix and the $1 \bigtimes C$ vector of treatment residuals for the group g. Then if $R_{1,N}(\alpha) = o_P(1)$ and the Lindeberg condition (Assumption 3.5(2)) hold, the Orthogonal Least Squares estimator is asymptotically Gaussian:

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

(d) In addition, the covariance matrix Ω can be consistently estimated by its sample analog:

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

where $\widehat{U}_{gt} \equiv (\widehat{\tilde{Y}}_{gt} - \widehat{\tilde{D}}_{gt}'\widehat{\beta})$.

Theorem 3.1 is our first main result. It establishes the upper bound on the convergence rate of the Orthogonal Least Squares. The bound consists of three components: the second stage sampling error whose order is $\sqrt{\frac{d}{N}}$, the measurement error bias due the estimation error of the treatment reduced form whose order is $\mathbf{d}_N^2\|\beta_0\|_2$, and the omitted variable bias due to the correlation of the estimation error that is present in the treatment and the outcome residuals whose order is $\mathbf{d}_N\mathbf{l}_N$. In the case Assumption 3.1 (5) holds, the second-stage sampling error dominates the other terms, and the impact of the first-stage estimation error is negligible. In the case Assumption 3.1 (5) does not hold, partialling-out still improves the rate of convergence to β_0 compared to, for example, a naive one-stage procedure that jointly estimates β_0 and $e_{i0}(\cdot)$ in Equation (2.1).

In the case Assumption 3.1 (5) holds, OLS attains the oracle rate and has an oracle asymptotic linearity representation. Under the Lindeberg condition OLS is asymptotically normal with asymptotic variance matrix Ω that is not affected by the first stage estimation and can be consistently estimated by the White cluster-robust estimator $\widehat{\Omega}$.

3.2 High Dimensional Sparse Regime

Here we introduce the basic concepts of the HDS regime. Denote the sparsity $s = s_N$ of the treatment effect β_0 as the number of its non-zero entries (i.e., $s \equiv \|\beta_0\|_0$). We allow the sparsity s to grow with the sample size N. Denote the set of indices T that correspond to the non-zero coordinates of β_0 as:

$$T \equiv \{j \in \{1, 2, ..., d\} \text{ s.t. } \beta_{0,j} \neq 0\}.$$

Let T^c stand for the complement of T, δ_T stand for a d-dimensional vector such that $\delta_{j,T} = \delta_j$ on a set of indices $j \in T$, and $\delta_{j,T} = 0$ if $j \notin T$ (similarly for δ_{T^c}). Define the restricted cone as d-dimensional restricted set:

$$\mathcal{RE}(\bar{c}) \equiv \{\delta \in \mathcal{R} : \|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \delta \neq 0\}.$$

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

Eigenvalue of the sample covariance matrix of true residuals:

$$\kappa(\tilde{Q}, T, \bar{c}) := \min_{\delta \in \mathcal{RE}(\bar{c})} \frac{\sqrt{s}(\delta' \tilde{Q} \delta)^{1/2}}{\|\delta_T\|_1} = \min_{\delta \in \mathcal{RE}(\bar{c})} \frac{\sqrt{s}\|\delta\|_{2,N}}{\|\delta_T\|_1}. \tag{3.2}$$

Assumption 3.7 (Restricted Eigenvalue Assumption RE($\bar{c}$).)**.** *We assume that minimal restricted eigenvalue of $\tilde{Q}$ is bounded from zero: $\kappa(\tilde{Q}, T, \bar{c}) > 0$.*

Assumption 3.7 is the assumption about the restricted identification of the treatment residuals $\tilde{D}_{it}$ in the given sample. It states that the treatment residuals $\tilde{D}_{it}$ are not too collinear in sample, where collinearity is measured in the directions δ that belong to the restricted cone $\mathcal{RE}(\bar{c})$. In the i.i.d. case Assumption 3.7 follows from the identification in population (Assumption 3.4) as shown by Rudelson and Zhou (2013). We assume that this implication continues to hold in the panel setting under plausible weak dependence conditions on the data generating process.

Assumption 3.8 (First-Stage Rate in the HDS Regime)**.** *(1) Suppose that Assumption 3.1 (1)-(3) holds. Define the rate of shrinkage of the treatment reduced form as follows:*

$$\boldsymbol{m}_N \equiv \sup_{d \in D_N} \sup_{M \subset \{1,2,\ldots,d\}, |M| \leqslant s} (\mathbb{E}\|(d(Z) - d_0(Z))_M\|^2)^{1/2}, \tag{3.3}$$

and $\boldsymbol{l}_N$ as in Assumption 3.1. (2) The sparsity s obeys the following bound: $s\boldsymbol{m}_N = o(1)$. (3) The rates $\boldsymbol{m}_N$ and $\boldsymbol{l}_N$ obey the following restriction: $\boldsymbol{m}_N = o\left(\left(\frac{\log d}{N}\right)^{1/8}\right)$ and $\boldsymbol{m}_N \boldsymbol{l}_N = o\left(\left(\frac{\log d}{N}\right)^{1/4}\right)$.

The rate $\mathbf{m}_N$ is the analog of the convergence rate $\mathbf{d}_N$, which is defined in Assumption 3.1, in the HDS Regime. It controls the rate of shrinkage of the vector-valued functions $d_M(\cdot)$ that are the projections of the treatment reduced form $d(Z)$ onto the set $M \subset \{1, 2, \ldots, d\}$ of the coordinates. The size of M is at most equal to the sparsity s of the treatment effect β_0. The rate $\mathbf{m}_N$ provides the upper bound on the first-stage bias accumulated on the set of the regressors T whose effect is non zero.

Without additional assumptions about the treatment vector D_{it}, Assumptions 3.8(2) and (3) are non-trivial. In particular, in the worst case the rate $\mathbf{m}_N$ scales

with the dimension d:

$$\mathbf{m}_N \leqslant \sup_{d \in D_N} (\mathbb{E}\|d(Z) - d_0(Z)\|^2)^{1/2} = \mathbf{d}_N.$$

However, it is possible to achieve a satisfactory bound on $\mathbf{m}_N$ if the high-dimensional treatment is generated by a linear transformation from a low-dimensional "base" treatment (i.e., Equation (2.8) holds).

Remark 3.1 (Plausibility of Assumption 3.8)**.** (1) Suppose there exists a low-dimensional base treatment P_{it} of dimension p such that the treatment vector D_{it} is equal to the image of the linear transformation $A = A(\cdot)$ of the base treatment P_{it}. The rows of the map $A(\cdot)$ are bounded in the second norm:

$$\sup_{j \in \{1,2,\ldots,d\}} (\mathbb{E}\|A_{j,\cdot}(Z)\|^2)^{1/2} \leqslant \bar{A} < \infty.$$

(2) Suppose there exists a sequence of realization sets $\{P_N, N \geqslant 1\}$ such that the following conditions hold. For each $N \geqslant 1$ the set P_N contains the reduced form function $p_{i0}(\cdot)$ defined in (2.9). There exists a sequence of numbers $\phi_N = o(1)$ such that the estimator $\widehat{p}_i(\cdot)$ of the reduced form $p_{i0}(\cdot)$ belongs to P_N with probability at least $1 - \phi_N$. The sets P_N converge around $p_{i0}(Z)$ at the rate $\mathbf{p}_N$ defined as:

$$\mathbf{p}_N := \sup_{p \in P_N} (\mathbb{E}\|p_{i0}(Z_{it}) - p(Z_{it})\|^2)^{1/2}.$$

Then the rate $\mathbf{m}_N$ is bounded by the rate $\mathbf{p}_N$: $\mathbf{m}_N \lesssim s\mathbf{p}_N$.

Assumption 3.9 (Tail Bound in HDS Regime)**.** *There exists $\bar{\sigma} \in (0, \infty)$ such that the outcome disturbance U_{it} is a subGaussian random variable with variance proxy $\bar{\sigma}$.*

Assumption 3.9 is a mild assumption on the outcome disturbance U_{it}, requiring it to have lighter tails that Gaussian random variable.

Definition 3.2 (Orthogonal Lasso)**.** *Let $\lambda > 0$ be a penalty parameter to be specified. Define the Orthogonal Lasso estimator as the minimizer of the ℓ_1-penalized least squares criterion function:*

$$\widehat{\beta}_L \equiv \arg\min_{b \in \mathcal{R}^k} \widehat{Q}(b) + \lambda\|\beta\|_1 \equiv \arg\min_{b \in \mathcal{R}^k} \mathbb{E}_N(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}_{it}' b)^2 + \lambda\|\beta\|_1.$$

Orthogonal Lasso is our second proposed 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 choice of the regularization parameter λ is described in Condition 3.1.

We summarize the noise of the problem with the help of the two quantities. The first quantity is equal to the gradient S of the empirical loss function $\widehat{Q}(\cdot)$ evaluated at the true value of the treatment effect β_0 as:

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

This gradient S contains (a) the second-stage sampling error and (b) the omitted variable bias due to the correlation between the estimation errors that are present in the estimated treatment and outcome residuals. To control the first source of noise we choose the parameter λ to dominate every coordinate of the gradient of the empirical loss function: $\lambda \geqslant c\|S\|_\infty$. Asymptotically, for this to happen it suffices for λ has to satisfy Condition 3.1.

Condition 3.1 (Choice of λ)**.** *For some universal constant c let the regularization parameter λ be chosen as follows:*

$$\lambda = c\big[\boldsymbol{l}_N \boldsymbol{m}_N \vee \boldsymbol{m}_N^2 \vee \lambda_{1N} \vee \lambda_{2N}\big],$$

where $\mathbf{m}_N$ and $\boldsymbol{l}_N$ are defined in Assumptions 3.1, 3.8 and $\lambda_{1N}, \lambda_{2N}$ are as in Assumption 3.2.

The second quantity q_N is defined as the maximal entrywise difference of the covariance matrices of the treatment residuals $\tilde{Q}$ and the estimated treatment residuals $\widehat{Q}$:

$$q_N \equiv \max_{1\leqslant k,j\leqslant d} |\widehat{Q} - \tilde{Q}|_{k,j}. \tag{3.4}$$

In the case Assumptions 3.2(1) and 3.8 hold, q_N admits the following bound:

$$q_N \lesssim_P \mathbf{m}_N + \lambda_{1N}.$$

Lemma 3.1 (Relation Between Restricted Eigenvalues of Sample Covariance Matrices)**.** *Let $\bar{c} > 1$ be a constant such that Assumption 3.7 holds. Then the difference of*

the restricted eigenvalues of these covariance matrices obeys the following bound:

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

Lemma 3.1 is an important building block in establishing the validity of Orthogonal Lasso. Combined with the bound (3.4), Lemma 3.1 implies that the restricted eigenvalue $\kappa(\widehat{Q}, T, \bar{c})$ based on the estimated treatment residuals is bounded away from zero for a sufficiently large N. In particular, if the sample covariance matrix of the **true** treatment residuals $\tilde{Q}$ obeys the restricted identification condition (Assumption 3.7), the sample covariance matrix of the **estimated** treatment residuals $\widehat{Q}$ also obeys this condition for a sufficiently large N.

For the vector $\delta = \widehat{\beta} - \beta_0, \delta \in R^d$ denote the in-sample prediction error in terms of the treatment residuals as: $\|\delta\|_{2,N} = (\mathbb{E}_N(\tilde{D}_{it}'\delta)^2)^{1/2}$ and the in-sample prediction error in terms of the estimated treatment residuals: $\|\delta\|_{\widehat{d},2,N} = (\mathbb{E}_N(\widehat{\tilde{D}}_{it}'\delta)^2)^{1/2}$.

Theorem 3.2 (Orthogonal Lasso)**.** *Suppose Assumptions 3.2, 3.5, 3.7, 3.8 (1)-(2), 3.9 hold and N is sufficiently large so that $sq_N(1+\bar{c})^2 < \frac{1}{2}$. Then the following statements hold:*

(a) On the event $\{\lambda \geqslant c\|S\|_\infty\}$ the (infeasible) in-sample prediction error is bounded as:

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

(b) Let RE($2\bar{c}$) hold. On the event $\{\lambda \geqslant c\|S\|_\infty\}$ the ℓ_1 norm of the Lasso error is bounded as:

$$\|\widehat{\beta}_L - \beta_0\|_1 \leqslant 2\lambda \frac{s}{\kappa(\tilde{Q}, T, 2\bar{c})\kappa(\tilde{Q}, T, \bar{c})}.$$

(c) Suppose λ is as in Condition 3.1. As N grows, the (infeasible) in-sample prediction error is bounded as:

$$\|\widehat{\beta}_L - \beta_0\|_{N,2} \lesssim_P \sqrt{s}\big[\boldsymbol{l}_N \boldsymbol{m}_N \vee \boldsymbol{m}_N^2 \vee \lambda_{1N} \vee \lambda_{2N} \vee \bar{\sigma}\sqrt{\frac{\log d}{N}}\big],$$

(d) and the ℓ_1 norm of the Lasso error is bounded as:

$$\|\widehat{\beta}_L - \beta_0\|_1 \lesssim_P s\big[\boldsymbol{l}_N \boldsymbol{m}_N \vee \boldsymbol{m}_N^2 \vee \lambda_{1N} \vee \lambda_{2N} \vee \bar{\sigma}\sqrt{\frac{\log d}{N}}\big].$$

Theorem 3.2 is our second main result. Statements (a) and (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. Statements (c) and (d) establish asymptotic bounds on $\|\widehat{\beta}_L - \beta_0\|_{N,2}$ and $\|\widehat{\beta}_L - \beta_0\|_1$. Under Assumption 3.8(3) the asymptotic bounds coincide with the respective infeasible bounds corresponding to the case where the treatment and outcome residuals are observed.

The total bias of $\widehat{\beta}_L$ scales with the sparsity s rather than with the total dimension d. This is a remarkable property of ℓ_1 regularization penalty. The convexity of the empirical loss function forces $\widehat{\beta} - \beta_0$ to belong to the restricted cone $\mathcal{RE}(\bar{c})$. On this restricted subset of the d-dimensional Eucledian space $\mathcal{R}^d$ the total bias scales in proportion to the bias accumulated on the set of active regressors T, rather than on the whole set of regressors. This property enables the convergence of Orthogonal Lasso in the HDS regime.

Comparison of the one-stage Lasso and Orthogonal Lasso in a Linear Model. Let the function $e_{i0}(Z)$ given in (2.1) be a linear function of the control vector Z: $e_{i0}(Z) = Z'\omega$. Let the coefficient ω be sparse with a sparsity index $s_\omega := \|\omega\|_0$ that obeys $s_\omega = o(N)$. In addition, suppose that for each treatment $j \in \{1, 2, \ldots, d\}$ the reduced form $d_{j,i0}(Z)$ of treatment j is linear in Z: $d_{j,i0}(Z) = Z'\delta_j$. Finally, let the coefficient δ_j be sparse with a sparsity index $s_{\delta_j} \leqslant s_\omega$. In this problem a researcher has the choice between the Orthogonal Lasso defined above and the one-stage Lasso defined as follows:

$$\begin{aligned} &\widehat{Q}(\beta, \omega) \equiv \mathbb{E}_N(Y_{it} - D'_{it}\beta - Z'_{it}\omega)^2, \\ &(\widehat{\beta}_B, \widehat{\omega}_B) \equiv \arg \min_{(\beta,\omega)\in\mathcal{R}^{d+d_Z}} \widehat{Q}(\beta, \omega) + \lambda_B(\|\beta\|_1 + \|\gamma\|_1). \end{aligned} \tag{3.5}$$

The one-stage Lasso algorithm uses Y as the dependent variable and the vector (D, Z) of treatments and controls as the regressors without distinguishing their roles. The coefficient $\widehat{\beta}_B$ is the one-stage Lasso coefficient of the treatment subvector D.

Consider the empirically relevant case where the complexity of the vector of controls, measured by $s_\omega^2 \log d_Z$, is larger than the complexity of the treatment vector, measured by $s^2 \log d$. For example, this case occurs if the number of controls d_Z is larger than the number of treatments d, (i.e., $d_Z \gg d$), and the sparsity indices of the treatment (s) and the control (s_γ) subvectors are of the same order of magnitude. If this is the case, the regularization parameter λ_B has to be chosen aggressively to

prevent the variance of $\widehat{\beta}_B$ from exploding. As a result, the regularization bias of $\widehat{\beta}_B$ is proportional to $\lambda_B \sim \sqrt{\frac{\log d_Z}{N}}$:

$$\|\widehat{\beta}_B - \beta_0\|_1 \lesssim_P \|(\widehat{\beta}_B, \widehat{\omega}_B) - (\beta_0, \omega_0)\|_1 \tag{3.6}$$

$$\lesssim_P (s + s_\omega)\lambda_B \lesssim_P \sqrt{\frac{s_\omega^2 \log d_Z}{N}}. \tag{3.7}$$

By contrast, the first stage bias $\mathbf{l}_N$ has only a second-order effect on the choice of λ for Orthogonal Lasso. As a result, partialling-out dampens the translation of the regularization bias associated with the large complexity of the control vector into the bias of the estimator of the treatment effect.

3.3 Double Orthogonal Lasso

In this section we consider the HDS regime where $d = d(N)$ may exceed N. We will refer to Q^{-1} as the precision matrix.

Assumption 3.10 below is a sufficient condition required for consistent estimation of the precision matrix Q^{-1} in high dimension. It requires the existence of a sufficiently sparse matrix Σ^{sp} that is a good approximation of Q^{-1} in the ℓ_1 norm. This assumption restricts the pattern of correlations between the treatment residuals. Examples of matrices Q satisfying Assumption 3.10 include Toeplitz, block diagonal, and band matrices.

Assumption 3.10 (Approximate Sparsity of the Precision Matrix)**.** *Assume that there exists a sparse matrix $\Sigma^{sp} \in \mathcal{R}^{d\times d}$ such that the following conditions hold. (1) The sparsity index s_Σ defined as:*

$$s_\Sigma := \max_{1\leqslant j\leqslant d} \|\Sigma^{sp}_{j,\cdot}\|_0$$

is sufficiently small: $s_\Sigma\sqrt{N}(\boldsymbol{m}_N^2 \vee \boldsymbol{m}_N \boldsymbol{l}_N) = o(1)$. (2) There exists a rate $r_N = o(\sqrt{\frac{\log d}{N}})$ such that the matrix Σ^{sp} is a good approximation of the precision matrix:

$$\max_{1\leqslant j\leqslant d} \|Q^{-1}_{j,\cdot} - \Sigma^{sp}_{j,\cdot}\|_1 \lesssim o(r_N).$$

Below we introduce an estimator $\widehat{\Sigma}$ of the precision matrix Q^{-1}. Among all possible matrices that approximately invert $\widehat{Q}$, which we refer to as orthogonalization set of

$\widehat{Q}$, we choose the matrix with the smallest ℓ_1 norm. Javanmard and Montanari (2014) showed that the orthogonalization set of $\tilde{Q}$, the sample covariance matrix based on the true residuals, contains the precision matrix Q^{-1}. In addition, the approximate sparsity of Q^{-1} (Assumption 3.10) ensures that the orthogonalization sets of $\widehat{Q}$ and $\tilde{Q}$ coincide. Therefore, the orthogonalization set of $\widehat{Q}$ is non-empty, and the estimator below is well-defined.

Definition 3.3 (Constrained Linear Inverse Matrix Estimator (CLIME))**.** *Let $c_\mu > 0$ be a universal constant. Let $\mu_N := c_\mu \sqrt{\frac{\log d}{N}} = o(1)$ be a sequence of numbers. Define the CLIME estimator $\widehat{\Sigma}$ of the precision matrix Σ as: $\widehat{\Sigma} := [\widehat{\Sigma}_1, \widehat{\Sigma}_2, \ldots, \widehat{\Sigma}_d]'$, where the j'th row is*

$$\widehat{\Sigma}_j := \arg\min_{w \in \mathcal{R}^d} \|w\|_1 \ s.t. \ \|\widehat{Q}_j w - e_j\|_\infty \leqslant \mu_N. \tag{3.8}$$

Definition 3.4 (Double Orthogonal Lasso)**.** *Let $\widehat{\Sigma}$ be the CLIME estimator of the precision matrix Q^{-1}. Define the Double Orthogonal Lasso estimator as:*

$$\widehat{\beta}_{DOL} := \underbrace{\widehat{\Sigma}\mathbb{E}_N \widehat{\tilde{D}}_{it} \widehat{\tilde{Y}}_{it}}_{\widehat{\beta}_{naive}} + (-\widehat{\Sigma}\widehat{Q} + I)\widehat{\beta}_L = \widehat{\Sigma}\mathbb{E}_N \widehat{\tilde{D}}_{it}(\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}'_{it}\widehat{\beta}_L) + \widehat{\beta}_L, \tag{3.9}$$

where $\widehat{\beta}_L$ is Orthogonal Lasso.

Double Orthogonal Lasso is our third main estimator. Its name reflects the two orthogonalization steps that are present in its construction. The first step refers to the construction of the treatment and the outcome residuals in (2.5). At the second-step we start with a naive estimator $\widehat{\beta}_{naive}$ of the treatment effect β_0 that is equal to the de-correlated sample covariance matrix of the estimated treatment and the outcome residuals, where $\widehat{\Sigma}$ is used as the decorrelation matrix. This estimator is sensitive to the biased estimation of $\widehat{\Sigma}$ and is of low quality. Adding the bias correction term $(-\widehat{\Sigma}\widehat{Q} + I)\widehat{\beta}_L$ makes the original estimator insensitive to the biased estimation of $\widehat{\Sigma}$ and Orthogonal Lasso $\widehat{\beta}_L$.

Lemma 3.2 (Bound on the Error of the CLIME Estimator)**.** *Let Σ^{sp} be a sparse approximation of the precision matrix Q^{-1} and $\widehat{\Sigma}$ be the CLIME estimator. Then the*

ℓ_1 norm of the difference of Σ^{sp} and $\widehat{\Sigma}$ is bounded as follows:

$$\max_{1 \leqslant j \leqslant d} \|\Sigma^{sp}_{j,\cdot} - \widehat{\Sigma}_{j,\cdot}\|_1 \leqslant s_\Sigma \mu_N. \tag{3.10}$$

Combining (3.10) with Assumption 3.10 yields the bound: $\max_{1 \leqslant j \leqslant d} \|\Sigma^{sp}_{j,\cdot} - \widehat{\Sigma}_{j,\cdot}\|_1 \leqslant s_\Sigma \mu_N + o(\mu_N)$.

Lemma 3.2 states that the estimator $\widehat{\Sigma}$ approximates the sparse approximation Σ^{sp} of the precision matrix Q^{-1} sufficiently well in the ℓ_1 norm. Since Σ^{sp} has sparse rows, the product of Σ^{sp} and the sample covariance $\mathbb{E}_N \widehat{\tilde{D}}_{it} \widehat{\tilde{Y}}_{it}$ has small bias accumulated from the first-stage. The first-stage bias scales in proportion to the sparsity index s_Σ rather than the dimension of the treatment vector d. This property enables the asymptotic linearity of the Double Orthogonal Lasso established further.

The gain of the sparsity of the treatment effect β_0 is present not only when the dimension d exceeds the sample size. Consider the empirically relevant case where the sparsity of the treatment vector s is smaller than the dimension $d = d(N)$ of the treatment vector, which in its turn is smaller than the sample size:

$$s \ll d(N), \quad d = o(N^{4\delta/3}).$$

Here $\delta > 0$ is the quality parameter of Assumption 3.1(5). While Orthogonal Least Squares of Definition 3.1 is well-defined in this regime, there may be a finite-sample gain from the exploitation of the sparsity of the treatment effect β_0. We describe an estimator that exhibits such a gain below.

Definition 3.5 (Double Orthogonal Ridge)**.** *For some universal constant $c_\tau > 0$ define the rate $\tau_N := c_\tau \sqrt{\frac{\log d}{N}}$. Define Double Orthogonal Ridge as:*

$$\widehat{\beta}_{DOR} := (\widehat{Q} + \tau_N I_d)^{-1} \mathbb{E}_N \widehat{\tilde{D}}_{it} (\widehat{\tilde{Y}}_{it} - \widehat{\tilde{D}}'_{it} \widehat{\beta}_L) + \widehat{\beta}_L, \tag{3.11}$$

where $\widehat{\beta}_L$ is Orthogonal Lasso.

Theorem 3.3 (Double Orthogonal Lasso and Double Orthogonal Ridge)**.** *Suppose Assumptions 3.2, 3.5(1), 3.7, 3.8 (1)-(3), 3.9, 3.10 hold. Then the following statements hold.*

(a) For any coordinate $j \in \{1, 2, \ldots, d\}$ the estimator $\widehat{\beta}_{DOL,j}$ of the j'th coordinate $\beta_{0,j}$ is asymptotically linear:

$$\sqrt{N}(\widehat{\beta}_{DOL,j} - \beta_{j,0}) = \mathbb{G}_N Q^{-1} \mathbb{E}_N \tilde{D}_{it} U_{it} + R_{j,N}, \tag{3.12}$$

where the remainder term $R_{j,N}$ is sufficiently small: $\max_{1 \leqslant j \leqslant d} |R_{j,N}| = o_P(1)$.

(b) Denote the asymptotic covariance matrix of the clustered standard errors as follows:

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

If the Lindeberg condition (Assumption 3.5 (2)) holds, Double Orthogonal Lasso is asymptotically Gaussian:

$$\lim_{N \to \infty} \sup_{t \in \mathcal{R}} |\mathrm{P}(\frac{\sqrt{N}(\widehat{\beta}_{DOL,j} - \beta_{0,j})}{\Omega_{jj}} < t) - \Phi(t)| = 0.$$

(c) In addition, if Assumptions 3.3 and 3.4 hold, statements (a)-(c) hold for Double Orthogonal Ridge.

Theorem 3.3 is our third main result. It establishes an asymptotically Gaussian approximation of each coefficient of the Double Orthogonal Lasso $\widehat{\beta}_{DOL}$ and Double Orthogonal Ridge $\widehat{\beta}_{DOR}$. The asymptotic covariance matrix Ω is estimated consistently by its cluster-robust sample analog $\widehat{\Omega} := \widehat{\Sigma} \mathbb{E}_N \widehat{\tilde{D}}_{gt} \widehat{\tilde{D}}'_{gt} \widehat{\tilde{U}}_{gt} \widehat{\tilde{U}}_{gt} \widehat{\Sigma}$.

Consider the infeasible case where the treatment and the outcome residuals are observed. Then any matrix $\widehat{\Sigma}$ that approximately inverts Q can serve as a decorrelation matrix to construct a naive estimator of the treatment effect. For example, Javanmard and Montanari (2014) suggest a variance minimizing choice of $\widehat{\Sigma}$ in a linear model with deterministic regressors. In contrast to their design, we establish asymptotic normality with random covariates and need a consistent estimator of the precision matrix Q^{-1}, which exists by Assumption 3.10. Going back to the feasible case, Assumptions 3.8 and 3.10 ensure that the first-stage bias does not affect the asymptotic sample average representation of Double Orthogonal Lasso and Double Orthogonal Ridge.

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. Let* $\widehat{\beta} \in \{\widehat{\beta}_{DOL}, \widehat{\beta}_{DOR}\}$. *Then the following Gaussian approximation result holds:*

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

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

$$C = (\mathrm{diag}\ \Omega)^{-1/2}\Omega(\mathrm{diag}\ \Omega)^{-1/2}$$

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

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

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

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

Theorem 3.4 is our fourth main result. It establishes the Gaussian approximation of Double Orthogonal Lasso and Double Orthogonal Ridge in high dimension. The statement of the theorem follows from the Gaussian approximation result of Zhang and Wu (2015) for time series and the Gaussian comparison inequalities of Chernozhukov et al. (2015). As discussed in Chernozhukov et al. (2013a), this Gaussian approximation result could be used not only for simultaneous confidence bands, but also for multiple hypothesis testing using step-down methods.

4 Verification of High-Level Conditions on the Reduced Form Estimators

Here we discuss a special structure of time-invariant heterogeneity that allows us to verify Assumptions 3.1, 3.2, and 3.8. For the sake of completeness, we also provide an example of cross-sectional data and panel data with no unobserved heterogeneity.

Example 3 (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, the small bias conditions (Assumptions 3.1 and 3.8) are achievable by many ML methods under structured assumptions on the nuisance parameters such as: ℓ_1 penalized methods in sparse models (Bühlmann and van der Geer (2011), Belloni et al. (2016b)); ℓ_2 boosting in sparse linear models (Luo and Spindler (2016)); other methods for classes of neural nets, regression trees, and random forests (Wager and Athey (2016)). The bound on the centered out-of-sample mean squared error in Assumption 3.2 follows from the Hoeffding inequality.

Example 4 (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, Assumption 3.1 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 gives an example of the ML estimator that satisfies Assumptions 3.1, 3.2, and Assumption 3.8. This is the estimator we use for price and sales reduced forms in our empirical application.

Remark 4.1 (Example 2, continued)**.** Assume that the price and sales reduced forms are linear functions of the control variables:

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

where the parameters π^Y, π^P are high-dimensional sparse parameters whose sparsity indices are bounded by s_π. Fix a partition $k \in [K]$ in the Panel Double ML recipe. On the partition k define the dynamic panel lasso of Kock and Tang (2016) as follows:

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

and

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

For every observation (i,t) in the partition I_k the estimators of the reduced form are defined as:

$$\widehat{p}_i(Z_{it}) = Z_{it}'\widehat{\pi}_k^D + \widehat{a}_i, \quad \widehat{l}_i(Z_{it}) = Z_{it}'\widehat{\pi}_k^Y + \widehat{b}_i.$$

In the case group size $C = 1$, Theorem 1 of Kock and Tang (2016) establishes the upper bound on the first stage rates:

Remark 4.2 (Rate of dynamic panel lasso)**.** In the case $C = 1$, under mild conditions on the design of $(Y_{it}, P_{it})_{(i,t)=(1,1)}^{I,T}$ Theorem 1 of Kock and Tang (2016) implies that the first stage rates

$$\mathbf{m}_N = \mathbf{l}_N = \frac{\log^{3/2}(p \vee I)s_\pi}{\sqrt{IT}} \vee s_\pi \frac{1}{\sqrt{I}}(\frac{\lambda}{\sqrt{IT}})^{1-\nu} = o((N)^{-1/4}),$$

where s_π is a bound on the sparsity of π^D, π^Y.

If the bound s_π is sufficiently small, Assumptions 3.1 and Assumption 3.8 hold. By Corollary 7.1, one can always take $\lambda_{1N} \equiv \mathbf{m}_N^2 = o(1/\sqrt{N})$ and $\lambda_{2N} \equiv \mathbf{m}_N\mathbf{l}_N = o(1/\sqrt{N})$ in Assumption 3.2. We expect Assumptions 3.1, 3.2, and 3.8 to hold for any cluster size C, but proving this is left for future work.

5 Empirical Application To Demand Estimation

In this section, we apply our estimators to measure own and cross-price elasticities faced by a major food distributor that sells to retailers. This distributor provided us with a sample of their transactional data containing all sales data from a number of major branches and spanning approximately four years. The data consist of a weekly time series of price and units sold for each of the 4,673 unique products in each of the eleven locations, and for each of the three delivery channels.[2] In total, our data include almost two million weekly observations.

[2]Customers can shop online or via telesales for same-day collection or next-day delivery, with each such combination constituting a separate channel.

Level 0 Category	Drinks	Household Items	Other Food	Protein
Level 1 Categories	Water	Tableware	Sweets	Dairy
	Soda	Sanitation	Snacks	Seafood
	Adult Beverages	Boxes	Sugar	Red Meat
		Stationary	Veggies	Chicken

Table 1: First two levels of hierarchical categorization for products used in this analysis.

Furthermore, we have access to detailed product descriptions that we used to construct a hierarchical categorization for each product, which was then included in our dataset. Our hierarchy is five levels deep, but we will only provide names for the first two levels (which we refer to as Level 1 and Level 2 categories). They are presented in Table 1[3].

The data exhibit a frequent variation in price as products cycle between "promotion on" and "promotion off" on a regular cadence. Such a variation in price may be correlated with demand expectations. **So it is critical that our first stage estimators accurately capture forward looking expectations of price setters to avoid omitted variable bias. The concern of the omitted variable bias is generally untestable.** However, we also have access to a subset of data in which the distributor agreed to randomize prices across two locations. This randomization allows us to experimentally validate the elasticities learned in the broader data set. This final analysis is presented in Section 5.3.

5.1 Demand Model

Let each unique combination of product, delivery channel, and store be indexed by i and let the corresponding log sales and log price in week t be denoted by $Q_{i,t}$ and $P_{i,t}$, respectively. Let $\{H_k\}_{k=1}^{K}$ be a collection of sets corresponding to our hierarchical categorization of the products. See Figure 1 as an example. Here, H_1 might correspond to the set "Drinks" and H_2 to the set "Soda". Additionally, we may refer to different levels of our hierarchy to identify some sub-collection of these sets. For example,

[3]In order to avoid any threat to the anonymity of the distributor, we have altered the names of some of these categories (without changing their meaning) and we will not report the names of any lower level categories.

Drinks are a *Level 1* category, whereas Water and Soda are *Level 2* categories. The leaf nodes (e.g., S. Pellegrino) are not individual products; rather, they are the finest level of categorization in which multiple products are still included.[4]

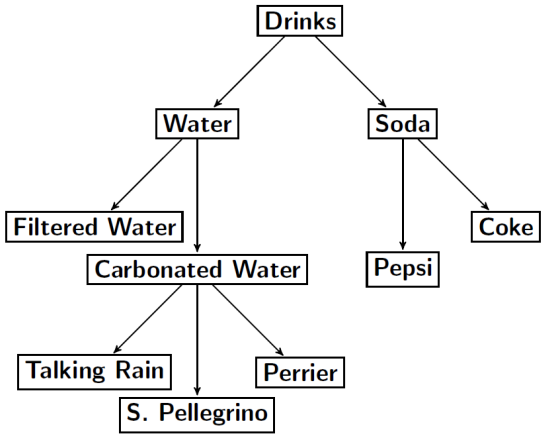


Figure 1: An example of a hierarchical categorization that is used to classify products. The leaf nodes should be viewed as individual products and intermediate nodes at various levels of categorization.

Our parameters of interest will be own-price elasticities (ϵ^o), which will be estimated heterogeneously over some subset of our hierarchy, and cross-price elasticities (ϵ^{cp}), which will correspond to impacts of the average non-self price within various subsets of the hierarchy.[5] **We decide later on which subsets of our hierarchy to include in any given specification.** Let Ξ_{op} and Ξ_{cp} denote the set of indices k used to model own and cross-price elasticities, respectively, in any given specification. Our demand model is:

$$Q_{i,t} = P_{i,t}\left(\sum_{k\in\Xi_{op}} 1_{i\in H_k}\cdot\epsilon_k^o\right) + P_{-i,k,t}\left(\sum_{k\in\Xi_{cp}} 1_{i\in H_k}\cdot\epsilon_k^{cp}\right) + g_0(Z_{i,t}) + U_{i,t}, \qquad (5.1)$$

[4]Individual products might then be (for example) different sizes or packagings of S. Pellegrino bottled water. Pricing is done at the level of individual products, which is also the level of our modeling. However, we will not model heterogeneous elasticities at the level of individual products in this empirical exercise due to computational constraints.

[5]This choice of the treatment variable reflects the intuition that products which share many common levels of hierarchy are most likely to have significant cross-price effects. It further imposes that the strength of cross-price effects is constant within a group. Alternative treatments could capture different proposed structures. For example, a revenue-weighted average price would correspond to a model in which consumers made choices based on the independence of irrelevant alternatives within each group.

where

$$P_{-i,k,t} \equiv \frac{\sum_{j \neq i, j \in H_k} P_{j,t}}{|H_k| - 1}$$

is the average non-self price in the group H_k. The controls $Z_{i,t}$ include time, store and product fixed effects, L lagged realizations of the demand system $(Y_{i,t-l}, P_{i,t-l})_{i \in [I], l \in \{1,2,..,L\}}$, and suitably chosen interactions to maximize the predictive performance of the first stage ML models. Denote the reduced form of log sales and log price, respectively, by

$$l_{i0}(z) \equiv \mathbb{E}[Q_{i,t} | Z_{i,t} = z],$$
$$p_{i0}(z) \equiv \mathbb{E}[P_{i,t} | Z_{i,t} = z].$$

Let $\tilde{P}_{i,t} \equiv P_{i,t} - p_{i0}(Z_{i,t})$ and $\tilde{Q}_{i,t} \equiv Q_{i,t} - l_{i0}(Z_{i,t})$ be the corresponding residuals. Intuitively, l_{i0} and p_{i0} may be thought of as one-period ahead, price-blind forecasts, and $\tilde{Q}$ and $\tilde{P}$ may be thought of as the corresponding deviations. Equation ((5.1)) implies a model linear in these deviations:

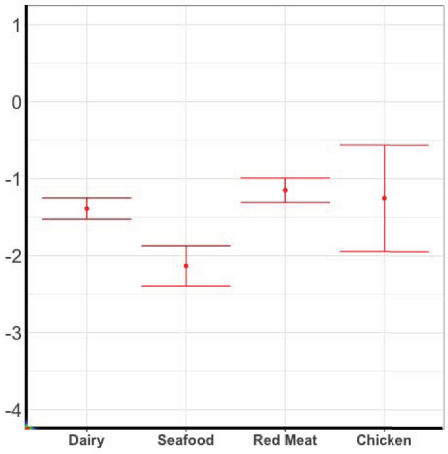
$$\tilde{Q}_{i,t} = \tilde{P}_{i,t} \left(\sum_{k \in \Xi_{op}} 1_{i \in H_k} \cdot \epsilon_k^o \right) + \tilde{P}_{-i,k,t} \left(\sum_{k \in \Xi_{cp}} 1_{i \in H_k} \cdot \epsilon_k^{cp} \right) + U_{i,t}. \tag{5.2}$$

We estimate various specifications of this model using Orthogonal Least Squares, Orthogonal Lasso, and Double Orthogonal Ridge as described in Sections 3.1 and 3.2. Across specifications we will vary the number of hierarchical categorizations included in Ξ_{op} and Ξ_{cp} in order to gauge the relative performance of our estimators at various dimensions of treatment. We will also add terms to the regression which enable us to measure heterogeneity in own-price elasticity at the monthly level in order to check for seasonal patters in price sensitivity. The results for own price elasticity are presented in Section 5.2, while the results on cross-price elasticity are presented in 5.4.

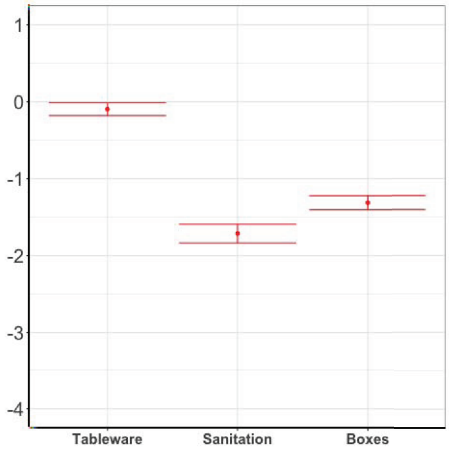
5.2 Own-Price Elasticity Results

In our first specification, we estimate average elasticities across our Level 1 product categories. We run a separate estimation on each Level 1 group, and the only included treatment variables are the heterogeneous own-price elasticities that correspond to

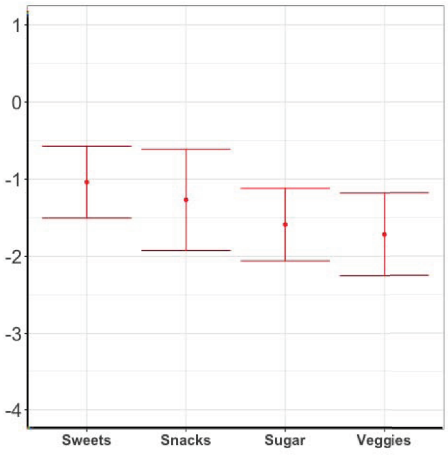
Level 2 categories. Since the dimension of treatment is quite small ($d \leqslant 4$ in all cases), we appeal to the results of our LD framework and use Orthogonal Least Squares. The resulting estimated elasticities along with 95% confidence intervals are presented in Figure 2.



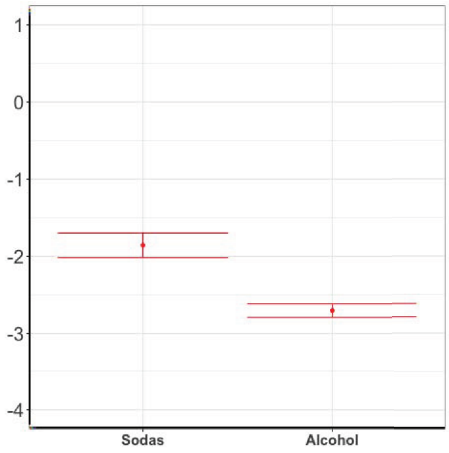
(a) Level 1 category: Protein



(b) Level 1 category: Household Items



(c) Level 1 category: Other Food



(d) Level 1 category: Drinks

Figure 2: Average price elasticities by level 1 category as estimated by orthogonal least squares.

Estimates range from the lowest $(-2.71)^{***}$ for Sodas and $(-2.12)^{***}$ for Seafood to a meager $(-0.4)^{**}$ for Tableware.[6] All product categories have elasticities that

[6]***, ** and * indicate the statistical significance at $0.99, 0.95, 0.90$ level, respectively.

are statistically less than zero, and, except for Tableware, all product categories have average elasticities less than -1.[7]

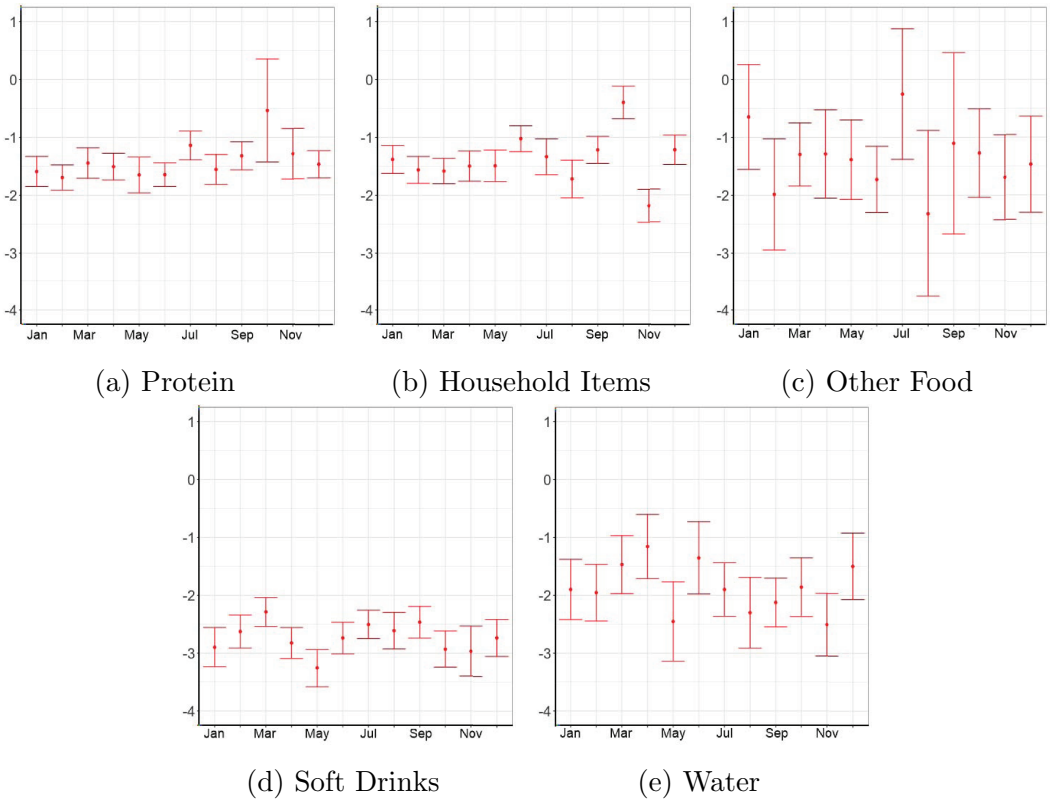
In our next specification, we estimate the heterogeneous own-price elasticity across a calendar year. Thus our treatments consist only of the own-price variable interacted with dummies for each month. Figure 3 shows the resulting estimates. Unsurprisingly, these estimates are significantly noisier and reveal only a few departures from a baseline of constant price sensitivity.[8] In particular, we do not see any strong evidence of bargain-hunting behavior during holiday seasons. This is broadly consistent with the findings of Chevalier et al. (2003); however. that paper studies consumer purchases in a grocery store rather than purchases from a distributor as we do. Somewhat intuitively, we did find that the elasticity of sodas is slightly closer to zero during warm months as compared to the rest of the year.[9]

Finally, we consider estimation of own-price elasticities at finer levels granularity within our hierarchy. Each of our four Level 0 groups has between 40 and 80 leaf nodes along which we might wish to estimate heterogeneity, with the number of total observations per leaf node ranging from as many as $5,000$ to as few as 100. One option is to use Orthogonal Least Squares with a separate treatment interaction for each leaf node, thus enabling us to learn independently estimated price elasticities. This would ensure unbiasedness (ignoring the upstream error from our estimation of reduced forms). However, this would make no use of our hierarchical categorization and would result in very noisy estimates for leaf nodes with few observations or little idiosyncratic variation in price. If we instead suppose that the true impact of our hierarchy on product elasticity is sparse (i.e. the presence in the majority of product categories has a zero added impact on elasticity), we have the very sparsity that is needed to motivate our HDS framework. As such, we may prefer to use Orthogonal Lasso or Double Orthogonal Ridge estimators.

[7]The estimated elasticity of Soft Drinks is close to the elasticities of orange juice found in the analysis of publicly available data from Dominick's Finer Foods. However, these data relate to consumer purchases from a retailer as opposed to the current analysis involving retailer purchases from a distributor.

[8]The biggest apparent departure is that Household Items appear to be very inelastic during the month of October. However, Household Items are a composite of two elastic Level 1 categories and one inelastic Level 1 category (Tableware), and we believe this pattern is driven by a larger than normal fraction of Tableware promotions in the October months of our data.

[9]The estimates presented above are obtained without accounting for cross-price effects. Accounting for cross-price effects returned estimates within one standard error of the original ones. For that reason, we exclude cross-price effects from the analysis of deeper levels of the hierarchy.



(a) Protein (b) Household Items (c) Other Food

(d) Soft Drinks (e) Water

Figure 3: Average price elasticities across the months of a calendar year as estimated by orthogonal least squares.

For comparison purposes, we use both of these estimators as well as Orthogonal Least Squares to estimate the heterogeneous own-price elasticities within the Level 1 category of Protein. We consider three different specifications in which we vary the number of levels of the hierarchy used to estimate the own-price heterogeneity. The results are presented in Figure 4. Going from left to right, we start with Orthogonal Lasso, which has the greatest level of shrinkage (and therefore bias), then Double Orthogonal Ridge (less shrinkage), and finally Orthogonal Least Squares (no shrinkage). The first row of this figure shows the distribution of estimated elasticities when only Levels 1 and 2 are used to estimate the heterogeneity. Here, the dimension of treatment is relatively small ($d = 22$); as a result, we see that the estimates of Orthogonal Least Squares are relatively plausible and that our LASSO estimators are only slightly more compressed. However, as we increase the dimension of treatment by adding all Level 3 dummies ($d = 62$; see the second row) and then all Level 4 categories ($d = 77$; third row), we see that the distribution of Orthogonal Least

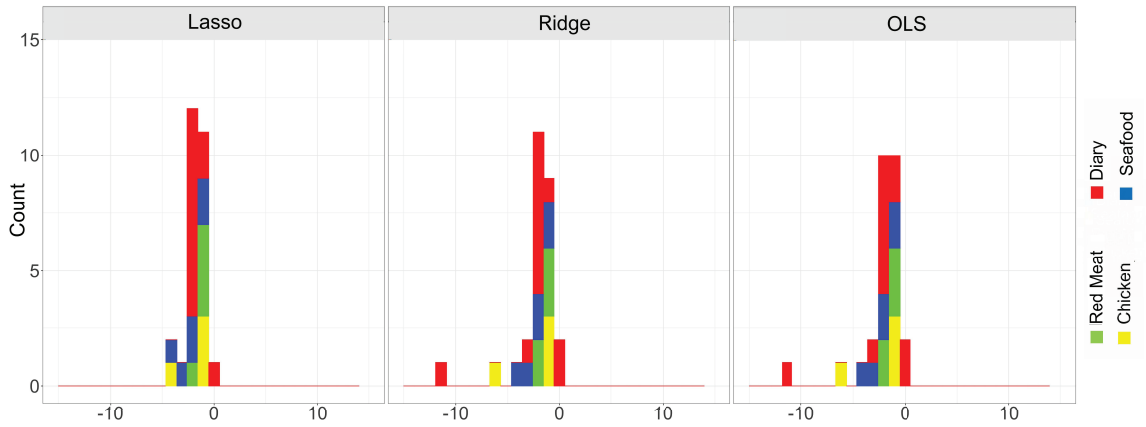
Squares estimates becomes increasingly dispersed and a significant number of positive (and therefore implausible) estimated elasticities are observed. By contrast, the distribution of estimated elasticities changes much less as the dimension of treatment is increased and even in the third row does not show any positive estimated elasticities. This stability is driven by the progressively higher level of shrinkage selected by our second stage Lasso estimator. By contrast, our Double Orthogonal Ridge strikes a middle ground. It features a significant shrinkage yielding less noisy (and therefore often more plausible) estimates than Orthogonal Least Squares, but it must restrict the shrinkage, as compared to Orthogonal Lasso, so as to guarantee a small asymptotic bias and allow for valid confidence intervals.

To better visualize how the shrinkage of Orthogonal Lasso and Double Orthogonal Ridge impacts on our estimates, Figure 5 shows the estimated elasticities and (except for the case of Orthogonal Lasso) associated confidence intervals for 11 selected Dairy products. Note that in all cases the Double Orthogonal Ridge point estimate is between the point estimates of Orthogonal Least Squares and Orthogonal Lasso. This reflects the fact that Double Orthogonal Ridge is essentially a matrix-weighted combination of OLS and Lasso as can be seen from (3.11). Moving from left to right, these products are sorted in descending order of the width of their Orthogonal Least Squares confidence interval. Note that when that confidence interval is wide, the Double Orthogonal Ridge confidence interval is clustered around the Lasso point estimate, but as the OLS confidence intervals shrink, the Double Orthogonal Ridge estimate and confidence interval are pulled progressively towards it.

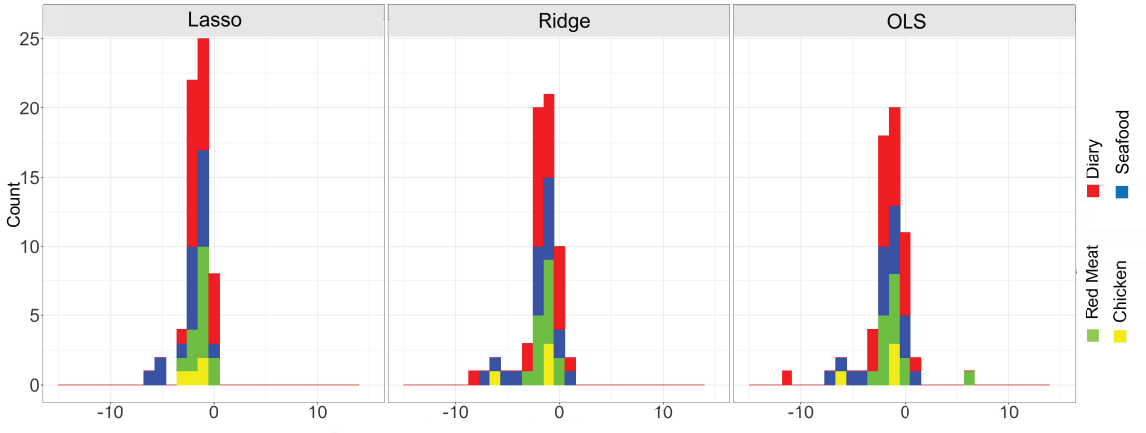
5.3 Experimental Validation of Own-Price Elasticities

Our food distributor agreed to run a two week promotion on 40 unique product, channel combinations that we selected. We will refer to these combinations as products. These 40 products were selected as the products for which a price cut was estimated to result in the greatest potential increase in profit, while maintaining constraints that they span all major product categories and that any two products in the chosen group do not have significant relationship, as measured by cross-price elasticity, between each other. For each product, the location of the promotion was randomly determined, and the other location maintained prices at a baseline level.

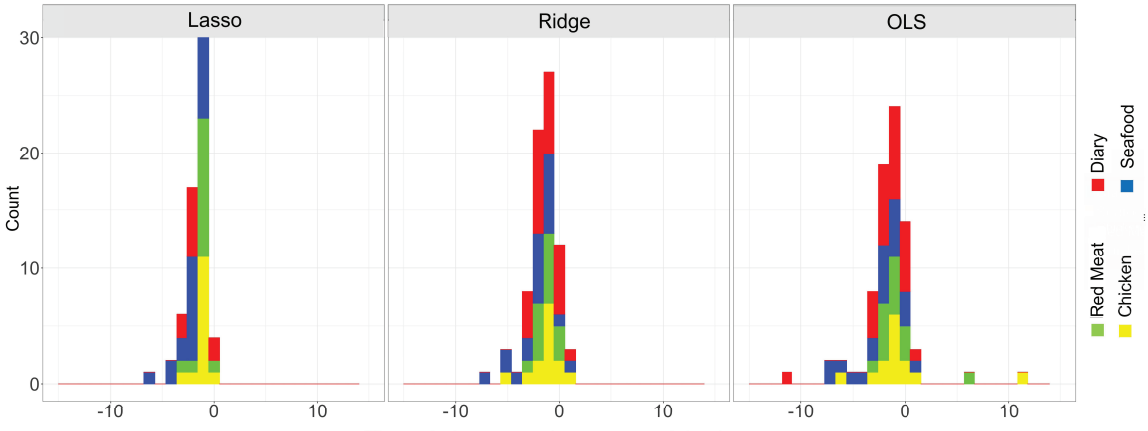
For each of these 40 products, we then computed the own price elasticity implied



(a) second level of the hierarchy at a number of groups equal to 27 and a Ridge penalty of 0.1



(b) third level of the hierarchy at a number of groups equal to 62 and a Ridge penalty of 0.5



(c) fourth level of the hierarchy at a number of groups equal to 77 and a Ridge penalty of 0.9

Figure 4: Histograms of the own-price elasticity computed for various estimators and dimensions of treatment at a total sample size of 193,144.

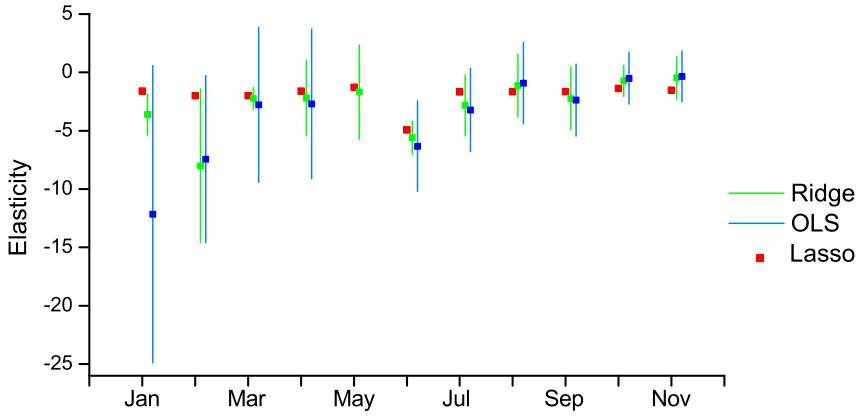


Figure 5. Estimated elasticities for selected Protein Products.

by this experiment as given by

$$\widehat{\epsilon}_{exp} = \frac{(\log Q_1 - \log \widehat{Q}_1) - (\log Q_2 - \log \widehat{Q}_2)}{\log P_1 - \log P_2},$$

where Q_s, P_s are the level of sales and price in location s, $s \in \{1, 2\}$. The estimated $\widehat{Q}_s$ is the price-blind forecast of expected sales. We also computed $\widehat{\epsilon}_{DML}$ as the fitted elasticities from our Double ML model. The two sets of elasticities are compared in Figure 6. As you can see, the experimentally learned elasticities have a much greater dispersion as they are learned from only a single biweekly sales outcome. However, they have the advantage of being learned from randomly assigned prices and thus can be seen as a source of ground truth to validate our broader estimates. The means of the two groups on Figure 6 are statistically indistinct. The the slope of the red line is not different from one, so we cannot reject the null that our estimated elasticities are the true ones.

5.4 Cross Price Elasticities

In this section, we present estimates of the average cross-elasticity within our Level 2 categories. *A priori* it is not obvious whether we should expect to primarily find positive or negative cross-price elasticities. Surely, price cuts on one product may have competitive impacts on similar products, but the loss-leader effect occurs where price cuts on one (for example) dairy product draw customers who may buy a number of other dairy products. Quantifying which products serve as loss leaders is beyond the scope of this chapter. Instead, we focus on the aggregate question of which effect

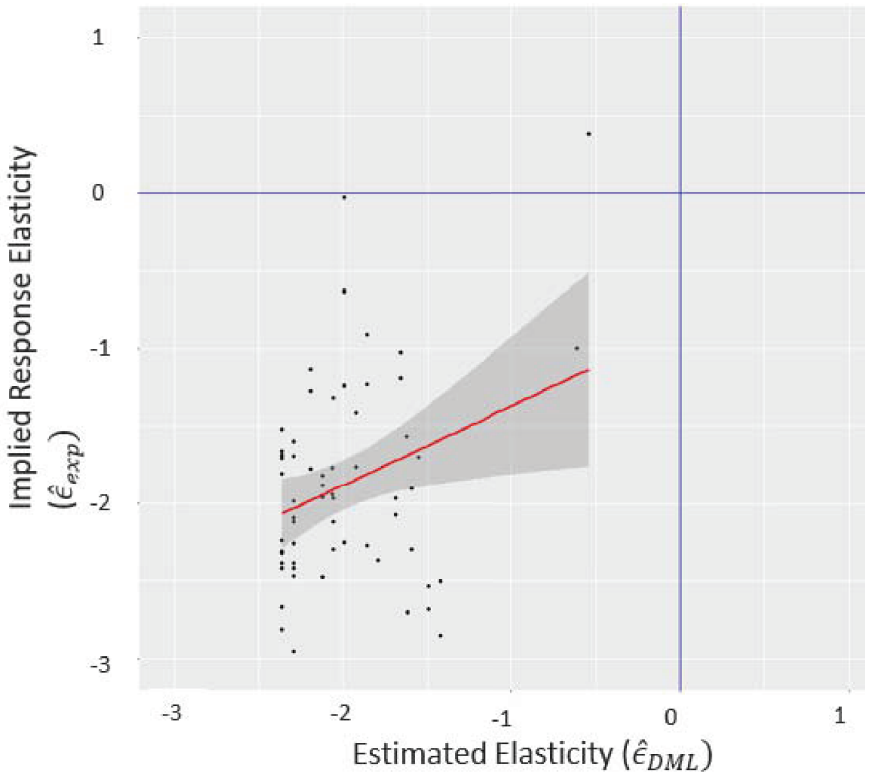


Figure 6: Scatterplot comparing our estimated elasticities to experimentally validated demand elasticities.

predominates in a given product category.

We do this by running a separate specification of for the Level 1 categories: Drinks, Protein, and Other Food. Referring back to ((5.1)), we choose Ξ_{cp} to contain all Level 2 categories, and we choose Ξ_{op} to contain all Level 4 categories in order to ensure that we control appropriately for own-price effects.

The results for our cross-price coefficients are presented in Table 2. Interpreting these figures requires some care. Recall that our cross-price treatments are the average non-self price in any particular set within our hierarchy. As such, these figures can be interpreted to give the expected percentage change in sales of a particular Soft Drinks product if *every other* Soft Drinks product saw a 1% price increase. As such, the actual cross-price elasticities between any two products can be calculated by taking the coefficient from Table 2 and dividing it by the number of products in the corresponding Level 2 category.

Level 1 Category	Level 2 Category	Estimated Average Cross-Price Effect	s.e.
Drinks	Adult Beverages	0.741***	0.246
	Water	1.041***	0.149
	Soft Drinks	0.637**	0.257
Protein	Red Meat	-0.582***	0.196
	Dairy	0.018	0.210
	Fish	-0.520*	0.286
	Poultry	-0.429***	0.175
Other Food	Sugar	-0.705*	0.382
	Sweets	-0.458	0.397
	Veggies	-1.181**	0.506
	Snacks	-0.847***	0.354

Table 2: Cross Price Elasticity, Drinks

6 Conclusion

We extend the Double ML framework to panel data where we enable estimation and inference on a high-dimensional treatment parameter, even while using modern ML methods to allow for arbitrarily complex control functions. After partialling out the estimated impact of these controls, we use Lasso and debiased Lasso for estimation and inference (respectively) and provide conditions under which the convergence rate of Lasso and asymptotic linearity of debiased Lasso are not affected by error from estimating the control functions. Additionally, we extend existing results to show that OLS may achieve valid inference on a high-dimensional vector of treatments so long as the rate of growth is sufficiently slow. All of our results require only that the ML methods employed for partialling out treatment and outcome achieve a sufficient rate of convergence. Furthermore, for the special case of a Dynamic Panel Lasso we provide low-level conditions under which we can achieve these required rates.

The resulting estimation algorithms – termed Orthogonal Least Squares, Orthogonal Lasso, and Double Orthogonal Lasso (for the debiased lasso) – are then applied to the problem of firm-side demand analysis. Here our methods offer a number of important advantages over other demand analysis frameworks. In particular, they leverage internal data to uncover natural experiments in the firm's historical pricing decisions and facilitate estimation and inference on demand elasticities for a wide catalog of products.

A final advantage of our method is that it should be well-suited to wide implementation. In this context, the estimand of ML methods used in the first two steps correspond closely to time-series forecasts of future sales and price realizations – an object already frequently computed by industrial data scientists. From such a starting point, our methods offer a straightforward approach to estimation and inference on demand parameters.

Finally, we demonstrate the utility of our methods by applying them to the estimation of demand elasticities for a major European food distributor. At the first level of the hierarchy our estimates of own-price elasticities based on Orthogonal Least Squares are close to the results of the category-level demand studies (e.g., Chevalier et al. (2003)). At the deeper levels of hierarchy, we find benefit from imposing the sparsity constraint implied by our HDS framework. A subset of our estimated price elasticities are validated experimentally based on randomly assigned promotions for a selected subset of products and we are unable to statistically reject the accuracy of estimates obtained from our HDS framework.

7 Appendix

Notation. We use the standard notation for vector and matrix norms. For a vector $v \in \mathcal{R}^d$, denote the ℓ_2 norm of a as $\|v\|_2 := \sqrt{\sum_{j=1}^d v_j^2}$. Denote the ℓ_1 norm of v as $\|v\|_1 := \sum_{j=1}^d |v_j|$, the ℓ_∞ norm of v as $\|v\|_\infty := \max_{1 \leqslant j \leqslant d} |v|_j$, and ℓ_0 norm of v as $\|v\|_0 := \sum_{j=1}^d 1_{\{a_j \neq 0\}}$. Denote a unit sphere as $\mathcal{S}^{d-1} = \{\alpha \in \mathcal{R}^d : \|\alpha\| = 1\}$. For a matrix M, denote its operator norm by $\|M\|_2 = \sup_{\alpha \in \mathcal{S}^{d-1}} \|M\alpha\|$. We use standard notation for numeric and stochastic dominance. For two numeric sequences $\{a_n, b_n\}, n \geqslant 1$ $a_n \lesssim b_n$ stands for $a_n = O(b_n)$. For two sequences of random variables $\{a_n, b_n, n \geqslant 1\}$: $a_n \lesssim_P b_n$ stands for $a_n = O_P(b_n)$. Finally, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

Every observation is indexed by a pair (i, t) where $i \in [I] := \{1, 2, .., I\}$ stands for the cross-sectional index and $t \in \{1, 2, .., T\}$ stands for the time index. The set of cross-sectional indices I consists of the pairs $\{(g, c), g \in [G] := \{1, 2, ..., G\}, c \in [C] := \{1, 2, ..., C\}\}$ where the first component $g = g(i)$ designates the group number, and the second $c \in [C]$ designates the cluster index within the group. For a random variable V, the symbol V_{gt} stands for the C-vector $V_{gt} = [V_{1t}, V_{2t}, ..., V_{Ct}]$. For a random d-vector D, the symbol D_{gt} stands for the $C \bigtimes d$ matrix: $D_{gt} = [D_{1t}, D_{2t}, ..., D_{Ct}]$.

We break all observations into $K \cdot C$ boxes such that the observations are weakly dependent within each box. A generic box $(k, c), k \in [K], c \in [C]$ consists of the indices of the partition I_k whose cluster index is equal to c. Each box contains GT indices. Denote the sample average within the box (k, c) as:

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

Since the number of boxes $K \cdot C$ is fixed, it suffices to establish a probabilistic bound on the sample average within each box.

Denote also the sample average within the whole sample as:

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

and

$$\mathbb{G}_N f(x_{it}) \equiv \frac{1}{\sqrt{GT}} \sum_{gt=1}^{GT} \sum_{c=1}^{C} [f(x_{gct}) - \mathbb{E} f(x_{gct})] = \frac{1}{\sqrt{GT}} \sum_{it=1}^{IT} [f(x_{it}) - \mathbb{E} f(x_{it})].$$

Denote the sample covariance matrix of the residuals as $\tilde{Q} \equiv \frac{1}{GT} \sum_{it=1}^{IT} \tilde{D}_{it} \tilde{D}'_{it}$, the sample covariance matrix of the estimated residuals as $\widehat{Q} \equiv \frac{1}{GT} \sum_{it=1}^{IT} \widehat{\tilde{D}}_{it} \widehat{\tilde{D}}'_{it}$, and the maximal entrywise difference of the two matrices as: $q_N \equiv \max_{1 \leqslant m,j \leqslant d} |\widehat{Q} - \tilde{Q}|_{m,j}$. For the vector $\delta = \widehat{\beta} - \beta_0, \delta \in R^d$, denote the in-sample prediction error in terms of the treatment residuals as $\|\delta\|_{2,N} = (\mathbb{E}_N(\tilde{D}'_{it}\delta)^2)^{1/2}$ and the in-sample prediction error in terms of the estimated treatment residuals as $\|\delta\|_{\widehat{d},2,N} = (\mathbb{E}_N(\widehat{\tilde{D}}'_{it}\delta)^2)^{1/2}$.

7.1 Proof of Theorems from Main Text

Proof of Theorem 3.1. **Theorem 3.1 (a).** Step 1. Let us show that the minimal eigenvalues of the sample covariance matrices $\tilde{Q}$ and $\widehat{Q}$ are bounded from zero. Under Assumption 3.3, $\|\tilde{Q} - Q\| \lesssim_P \sqrt{\frac{d \log N}{N}}$. Therefore, wp $\to 1$, and all eigenvalues of $\tilde{Q}^{-1}$ are bounded away from zero. To see this, suppose $\tilde{Q}$ has an eigenvalue less than $C_{\min}/2$. 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 7.2 (1) $\|\widehat{Q}-\tilde{Q}\| \lesssim_P \mathbf{d}_N^2+\sqrt{d}\lambda_{1N}$. Therefore, w.h.p. the eigenvalues of $\widehat{Q}$ are also bounded away from $C_{\min}/4$.

Step 2. Let us show that $\|\mathbb{E}_N \tilde{D}_{it}\tilde{U}_{it}\| \lesssim_P \sqrt{d/N}$. This can be seen as follows:

$$\mathbb{E}[\|\mathbb{E}_N \tilde{D}_{it}\tilde{U}_{it}\|^2] = \mathbb{E}[\tilde{D}_{it}\tilde{U}_{it}'\tilde{U}_{it}\tilde{D}_{it}]/N \leqslant \sigma_U^2 d/N.$$

Step 3. Consider the following expansion:

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

Lemma 7.2 (2) shows that $\|a\| \lesssim_P [\mathbf{d}_N\mathbf{l}_N + \mathbf{d}_N^2\|\beta_0\| + \sqrt{d}(\lambda_{1N}+\lambda_{2N})]$.

$$\begin{aligned}
\|b\| &= \|\widehat{Q}^{-1}-\tilde{Q}^{-1}\|\|\mathbb{E}_N\tilde{D}_{it}\tilde{Y}_{it}\| && (\tilde{Y}_{it} = D_{it}\beta_0 + U_{it}) \\
&= \|\widehat{Q}^{-1}-\tilde{Q}^{-1}\|\|\tilde{Q}\beta_0 + \mathbb{E}_N\tilde{D}_{it}\tilde{U}_{it}\| \\
&\lesssim_P \|\widehat{Q}^{-1}\|\|\widehat{Q}-\tilde{Q}\|\|\tilde{Q}^{-1}\|\big(\|\beta_0\|O_P(\sqrt{d\log N/N}) + O_P(\sqrt{d/N})\big) \\
&\lesssim_P (C_{\min}/4)^{-1}[\mathbf{d}_N^2 + \sqrt{d}\lambda_{1N}](C_{\min}/2)^{-1}\|\beta_0\|, && \text{(Lemma 7.2 (a))}
\end{aligned}$$

where the third line follows from Steps 1 and 2.

Step 4.

$$\begin{aligned}
\|c\| &= \|\tilde{Q}^{-1}\mathbb{E}_N\tilde{D}_{it}\tilde{Y}_{it} - \beta_0\| = \|\tilde{Q}^{-1}\mathbb{E}_N\tilde{D}_{it}\tilde{U}_{it}\| \\
&\lesssim_P \|\mathbb{E}_N\tilde{D}_{it}\tilde{U}_{it}\| \lesssim_P \sqrt{d/N}.
\end{aligned}$$

To conclude,

$$\|\widehat{\beta}-\beta_0\| \lesssim_P \mathbf{d}_N\mathbf{l}_N \vee (\mathbf{d}_N^2 + \sqrt{d}\lambda_N)\|\beta_0\| \vee \sqrt{d/N}. \tag{7.1}$$

Theorem 3.1 (b,c) Let

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

where

$$R_{it} \equiv (\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))'\beta_0 + (l_{i0}(Z_{it}) - \widehat{l}_i(Z_{it})), \quad (i,t) \in [I,T]$$

summarizes the first stage approximation error. Consider the following expansion:

$$\begin{aligned}
\sqrt{N}\alpha'(\widehat{\beta} - \beta) &= \sqrt{N}\alpha'(\widehat{Q}^{-1}\mathbb{E}_N\widehat{\tilde{D}}_{it}\widehat{\tilde{Y}}_{it} - \beta_0) && (7.2)\\
&= \sqrt{N}\alpha'\widehat{Q}^{-1}\mathbb{E}_N\widehat{\tilde{D}}_{it}(R_{it} + U_{it}) && (7.3)\\
&= \sqrt{N}\alpha'Q^{-1}\mathbb{E}_N\tilde{D}_{it}U_{it} + R_{1,N}(\alpha), && (7.4)
\end{aligned}$$

where the remainder term $R_{1,N}(\alpha)$ summarizes the approximation error as follows:

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

As shown at Step 1, the eigenvalues of $\widehat{Q}^{-1}$ are bounded away from zero. By Lemma 7.2(2),

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

By Lemma 7.2(b),

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

Equation (7.4) establishes an Asymptotic Linearity representation of $\widehat{\beta}$. Therefore, the leading term of equality (7.4) is equal to $\sqrt{N}\alpha'Q^{-1}\mathbb{E}_N\tilde{D}_{it}U_{it} = \sqrt{N}\alpha'Q^{-1}\mathbb{E}_N\tilde{D}'_{gt}U_{gt} := \sqrt{N}\mathbb{E}_N\xi_{gt}$ where ξ_{gt} is defined in Theorem 7.1. Therefore, the asymptotic normality of $\widehat{\beta}$ follows from Theorem 7.1 with the asymptotic variance matrix $\Omega = Q^{-1}\Gamma Q^{-1}$ where $\Gamma = \mathbb{E}\tilde{D}'_{gt}U_{gt}U'_{gt}\tilde{D}'_{gt}$. **Theorem 3.1 (d)** Step 1. Let $\Gamma := \mathbb{E}\tilde{D}'_{gt}\tilde{D}_{gt}U_{gt}U'_{gt}$. We will show that $\widehat{\Gamma} := \mathbb{E}_N\widehat{\tilde{D}}'_{gt}\widehat{\tilde{D}}_{gt}\widehat{U}_{gt}\widehat{U}'_{gt}$ converges to Γ. By Assumption 3.3 and Lemma 7.2, $\|\widehat{Q} - Q\|_P \lesssim_P \mathbf{d}_N^2 + \sqrt{d}\lambda_{1d}$. Therefore,

$$\|\widehat{\Omega} - \Omega\| = \|\widehat{Q}^{-1}\widehat{\Gamma}\widehat{Q}^{-1} - Q^{-1}\Gamma Q^{-1}\| \lesssim_P o(1).$$

Step 2. Convergence of $\widehat{\Gamma}$.

$$\begin{aligned}\|\widehat{\Gamma}-\Gamma\| = &\underbrace{\mathbb{E}_N\widehat{\tilde{D}}_{gt}'\widehat{\tilde{D}}_{gt}R_{gt}R_{gt}'}_{J_1}+\underbrace{\mathbb{E}_N\tilde{D}_{gt}\tilde{D}_{gt}'U_{gt}'U_{gt}-\Gamma}_{J_2}\\ &+\underbrace{\mathbb{E}_N(\widehat{\tilde{D}}_{gt}'\widehat{\tilde{D}}_{gt}-\tilde{D}_{gt}\tilde{D}_{gt}')U_{gt}'U_{gt}}_{J_3}\\ &+\underbrace{\mathbb{E}_N\widehat{\tilde{D}}_{gt}'\widehat{\tilde{D}}_{gt}\tilde{D}_{gt}'(\beta_0-\widehat{\beta})}_{J_4}.\end{aligned}$$

Step 2.a. J_1. Define an event $\mathcal{E}_N := \{\widehat{d}_k \in D_N \quad \& \quad \widehat{l}_k \in L_N \quad \forall k \in [K]\}$, such that the reduced form estimates $\widehat{d}_k$ and $\widehat{l}_k$ belong to the realization sets D_N, L_N for each fold $k \in [K]$. By union bound, this event holds w.p $1-o(1)$: $\mathrm{P}_{P_N}(\mathcal{E}_N) \geqslant 1 - K\phi_N = 1-o(1)$.

$$\mathbb{E}_N\widehat{\tilde{D}}_{gt}'\widehat{\tilde{D}}_{gt}R_{gt}R_{gt}' \equiv \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}J_{1,k} := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\zeta_{gt}\zeta_{gt}',$$

where $J_{1,k}$ is the analog of J_1 on partition $k \in [K]$ and $\zeta_{gt} := \widehat{\tilde{D}}_{gt}R_{gt}$. On the event $\mathcal{E}_N$ let us apply Lemma 7.1 for ζ_{gt}. The first assumption of Lemma 7.1 follows from Assumptions 3.5(1) and 3.6: $\sup_{d\in D_N, l\in L_N}(\mathbb{E}\|\zeta_{gt}\|^q)^{1/q} \lesssim d^{1/q}$. Then for o_N defined in Lemma 7.1,

$$\mathbb{E}_{n,kc}J_{1,k} - \mathbb{E}[J_{1,k}|I_k^c] \lesssim_P o_N.$$

Decomposing ζ_{gt} into a function of Z_{gt} and a residual $\tilde{D}_{gt}$ gives:

$$\zeta_{gt} = (\tilde{D}_{gt} + \widehat{d}(Z_{gt}) - d_{i0}(Z_{gt}))((\widehat{d}(Z_{gt}) - d_{i0}(Z_{gt}))'\beta_0 + \widehat{l}(Z_{gt}) - l_{i0}(Z_{gt})).$$

The conditional mean of $J_{1,k}$ is equal to:

$$\begin{aligned}\mathbb{E}[J_{1,k}|I_k^c] &= \mathbb{E}[\zeta_{gt}\zeta_{gt}'|I_k^c]\\ &= \underbrace{\mathbb{E}[(\widehat{d}(Z_{gt}) - d_{i0}(Z_{gt}))R_{gt}((\widehat{d}(Z_{gt}) - d_{i0}(Z_{gt}))R_{gt})'|I_k^c]}_{P_1} + \underbrace{\mathbb{E}[\tilde{D}_{gt}'\tilde{D}_{gt}R_{gt}R_{gt}'|I_k^c]}_{P_2}.\end{aligned}$$

since each term involving a cross-product of $\tilde{D}_{gt}(\widehat{d}(Z_{gt}) - d_{i0}(Z_{gt}))$ has a zero mean.

Conditionally on the event $\mathcal{E}_N$

$$\begin{aligned}\|P_1\| &\leqslant \sup_{\alpha\in\mathcal{S}^{d-1},d\in D_N,l\in L_N} \|\mathbb{E}\zeta_{gt}\zeta_{gt}'\alpha\| \\ &\leqslant^{i} \sup_{\alpha\in\mathcal{S}^{d-1},d\in D_N,l\in L_N} \mathbb{E}\|\zeta_{gt}\|\mathbb{E}\|\alpha'\zeta_{gt}\| \\ &\leqslant^{ii} \sup_{d\in D_N,l\in L_N} \mathbb{E}\|R_{gt}\| \lesssim \mathbf{d}_N\|\beta_0\| + \mathbf{l}_N,\end{aligned}$$

where i holds by Cauchy Scwartz.

By Lemma 6.1 of Chernozhukov et al. (2017a), textcolorbluechange 6.1 by its label the conditional convergence implies unconditional. The term P_2 is bounded similarly:

$$\begin{aligned}\|P_2\| &\leqslant \sup_{\alpha\in\mathcal{S}^{d-1},d\in D_N,l\in L_N} \|\mathbb{E}\tilde{D}_{gt}'\tilde{D}_{gt}R_{gt}R_{gt}'\alpha\| \\ &\leqslant \sup_{\alpha\in\mathcal{S}^{d-1},d\in D_N,l\in L_N} \mathbb{E}\|\tilde{D}_{gt}R_{gt}\|\mathbb{E}\|\alpha'\tilde{D}_{gt}R_{gt}\| \\ &\leqslant \sup_{d\in D_N,l\in L_N} \mathbb{E}\|R_{gt}\| \lesssim \mathbf{d}_N\|\beta_0\| + \mathbf{l}_N.\end{aligned}$$

Step 2.b. J_2. Let $\zeta_{2,gt} := \tilde{D}_{gt}U_{gt}$. The first assumption of Lemma 7.1 follows from Assumptions 3.5(1),(3) and 3.6: $(\mathbb{E}\|\zeta_{2,gt}\|^q)^{1/q} \lesssim d^{1/q}$. Then for o_{2N} defined in Lemma 7.1,

$$\|J_2 - \Gamma\| \lesssim_P o_{2N}.$$

Step 2.c. J_3. Let $\zeta_{3,gt} := (\widehat{\tilde{D}}_{gt}\widehat{\tilde{D}}_{gt}' - \tilde{D}_{gt}\tilde{D}_{gt})U_{gt}U_{gt}'$. The first assumption of Lemma 7.1 follows from Assumptions 3.5(1),(3) and 3.6: $(\mathbb{E}\|\zeta_{3,gt}\|^q)^{1/q} \lesssim d^{1/q}$. Then for o_{3N} defined in Lemma 7.1,

$$\|J_{3,k} - \mathbb{E}[J_3|I_k^c]\| \lesssim_P o_{3N},$$

where $J_{3,k}$ is the analog of J_3 on partition $k \in [K]$. Conditionally on the event $\mathcal{E}_N$

$$\begin{aligned}\mathbb{E}[J_{3,k}|I_k^c] &= \mathbb{E}[(\widehat{\tilde{D}}_{gt}\widehat{\tilde{D}}_{gt}' - \tilde{D}_{gt}\tilde{D}_{gt})U_{gt}U_{gt}'|I_k^c] \\ &\lesssim_P \mathbb{E}[(\widehat{d}_g(Z_{it}) - d_{i0}(Z_{it}))(\widehat{d}_g(Z_{it}) - d_{i0}(Z_{it}))'U_{gt}U_{gt}'|I_k^c]\end{aligned}$$

Step 3.d J_4. To bound J_4, recognize that

$$\begin{aligned}
J_4 &\lesssim_P \|\mathbb{E}_N \widehat{\tilde{D}}'_{gt}\widehat{\tilde{D}}_{gt}\| \max_{1\leqslant gt\leqslant N} \|\tilde{D}'_{gt}(\widehat{\beta}-\beta_0)\|^2 \\
&\lesssim_P \|\mathbb{E}_N \widehat{\tilde{D}}'_{gt}\widehat{\tilde{D}}_{gt}\| \max_{1\leqslant gt\leqslant N} (\widehat{\beta}-\beta_0)'\tilde{D}_{gt}\tilde{D}'_{gt}(\widehat{\beta}-\beta_0) \\
&\lesssim_P \|\mathbb{E}_N \widehat{\tilde{D}}'_{gt}\widehat{\tilde{D}}_{gt}\|\|\widehat{\beta}-\beta_0\|^2 \max_{1\leqslant gt\leqslant N} \|\tilde{D}_{gt}\tilde{D}'_{gt}\| \\
&\lesssim_P [C_{\max} + O_P(\sqrt{\frac{d\log N}{N}}) + \mathbf{d}_N^2 + \sqrt{d}\lambda_{1N}]O_P(\frac{d}{N}\max_{1\leqslant gt\leqslant N}\|\tilde{D}_{gt}\tilde{D}'_{gt}\|) = o_P(1).
\end{aligned}$$

■

Proof of Remark 3.1.

$$\begin{aligned}
\mathbf{m}_N^2 &\equiv \sup_{d\in D_N} \sup_{M\subset\{1,2,\ldots,d\},|M|\leqslant s} \mathbb{E}\|(d_i(Z_{it}) - d_{i0}(Z_{it}))_M\|^2 \\
&\leqslant \sup_{p(\cdot)\in P_N} \sup_{M\subset\{1,2,\ldots,d\},|M|\leqslant s} \sum_{j\in M} \mathbb{E}(A_j(Z_{it})(p_i(Z_{it}) - p_{i0}(Z_{it})))^2 \\
&\leqslant s \sup_{p(\cdot)\in P_N} \bar{A}\mathbf{p}_N,
\end{aligned}$$

where the last inequality follows from Cauchy-Scwartz and the assumptions of Remark 3.1. ■

Proof of Lemma 3.1. 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 7.2(b). The bound on the difference of $\|\delta\|^2_{\widehat{d},2,N}$ and $\|\delta\|^2_{2,N}$ is as follows:

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

By the definition of $\mathcal{RE}(\bar{c})$, for any $\delta \in \mathcal{RE}(\bar{c})$ 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 $\kappa(\widehat{Q},T,\bar{c})$:

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

Taking the squares of both sides of the inequality and re-arranging gives:

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

■

Proof of Theorem 3.2.

$$\widehat{Q}(\widehat{\beta}_L) - \widehat{Q}(\beta_0) - \mathbb{E}_N(\widehat{\tilde{D}}_{it}'\delta)^2 = -2\underbrace{\mathbb{E}_N\big[U_{it}\tilde{D}_{it}'\delta\big]}_{I_1} \tag{7.5}$$

$$-2\underbrace{\mathbb{E}_N\big[U_{it}(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\delta\big]}_{I_2} \tag{7.6}$$

$$-2\underbrace{\mathbb{E}_N\big[(l_{i0}(Z_{it}) - \widehat{l}_i(Z_{it}) + (d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'\beta_0)(\tilde{D}_{it})'\delta\big]}_{I_3} \tag{7.7}$$

$$-2\underbrace{\mathbb{E}_N\left[(l_{i0}(Z_{it})-\widehat{l}_i(Z_{it})+(d_{i0}(Z_{it})-\widehat{d}_i(Z_{it}))'\beta_0)(d_{i0}(Z_{it})-\widehat{d}_i(Z_{it}))'\delta\right]}_{I_4}. \quad (7.8)$$

By Lemma 7.3, $|I_2+I_3| \lesssim_P D^2\sqrt{\frac{\log(2d)}{N}}+\mathbf{m}_N^2\|\beta_0\|^2+\mathbf{m}_N\mathbf{l}_N$ and $|I_4| \lesssim_P \sqrt{s}\lambda_{1N}+\lambda_{2N}+\mathbf{m}_N^2\|\beta_0\|^2+\mathbf{m}_N\mathbf{l}_N$. Since I_1 is the sample average of bounded martingale difference sequences, $a \lesssim_P \sqrt{\frac{s\log d}{N}}$ by the Azouma-Hoeffding inequality. Therefore, with high probability $\exists c>1 \quad \lambda \geqslant c[\sqrt{\frac{s\log d}{N}}+\sqrt{\frac{\log(2d)}{N}}+\mathbf{m}_N^2\|\beta_0\|^2+\mathbf{m}_N\mathbf{l}_N+\lambda_{1N}\sqrt{s}+\lambda_{2N}]$. The optimality of $\widehat{\beta}_L$ and the choice of λ imply:

$$\lambda(\|\beta_0\|_1-\|\widehat{\beta}\|_1) \geqslant \|\delta\|^2_{\widehat{d},2,N} \geqslant -\lambda/c\|\delta\|_1. \quad (7.9)$$

The triangle inequality implies:

$$-\lambda/c\|\delta\|_1 \leqslant \lambda(\|\beta_0\|_1-\|\widehat{\beta}\|_1) \leqslant \lambda(\|\delta_T\|_1-\|\delta_{T^c}\|_1) \leqslant \lambda\|\delta_T\|_1 \quad (7.10)$$

$$\|\delta_{T^c}\|_1 \leqslant \frac{c+1}{c-1}\|\delta_T\|_1. \quad (7.11)$$

Therefore, δ belongs to a restricted set in the RE($\bar{c}$) where $\bar{c}=\frac{c+1}{c-1}$. Lemma 3.1 implies:

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

Therefore, the statement of Theorem 3.2 (a) holds:

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

where i holds by Lemma 3.1, ii holds by (7.10) , and iii holds by Assumption 3.7.

The statement of Theorem 3.2 (b) holds:

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

where i holds by the definition of $\mathrm{RE}(2\bar{c})$ and ii holds by Assumption 3.7 for $2\bar{c}$. The proof of Theorem 3.2 (c,d) follows from the choice of λ and statements (a,b). ∎

Definition 7.1 (Orthogonalization Set)**.** *Let $\mu_N : N \geqslant 1$ be a $o(1)$ sequence. We say that a $d \bigtimes d$-dimensional matrix $M = [m_1, ..., m_d]'$ orthogonalizes a given matrix Q at rate μ_N if:*

$$\|MQ - I\|_\infty \lesssim \mu_N. \tag{7.12}$$

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

Proof of Lemma 3.2. Step 1. Fix a coordinate $j \in \{1, 2, \ldots, d\}$. We will show in Step 2 that $\delta_j := \widehat{\Sigma}_j - \Sigma_j^{sp}$ belongs to the restricted set $\mathcal{RE}(2)$. Let $T_j := \{l \in \{1, 2, \ldots, d\} \quad \Sigma_{j,l}^{sp} \neq 0\}$ be the set of rows where the vector Σ_j^{sp} is not equal to zero. Lemma 7.4 shows that Σ_j^{sp} belongs to the orthogonalization set $M_N(\widehat{Q})$. Therefore, $\|\widehat{Q}_j\Sigma_j^{sp} - e_j\|_\infty \leqslant \mu_N$. The following bound holds:

$$\begin{aligned}\max_{1\leqslant j\leqslant d}\|\Sigma_j^{sp} - \widehat{\Sigma}_j\|_1 &\leqslant 2\max_{1\leqslant j\leqslant d}\|\Sigma_j^{sp} - \widehat{\Sigma}_j\|_{T_j}\\ &\leqslant 2s_\Sigma\mu_N,\end{aligned}$$

where the first inequality holds by Step 2 and the last inequality holds by Step 2 and Assumption 3.10.

Step 2. Lemma 7.4 shows that Σ_j^{sp} belongs to the orthogonalization set $M_N(\widehat{Q})$. By the definition of CLIME

$$\|\widehat{\Sigma}_j\|_1 \leqslant^{i} \|\Sigma_j^{sp}\|_1 = \|(\Sigma_j^{sp})_{T_j}\|_1 \leqslant^{ii} \|(\widehat{\Sigma}_j)_{T_j}\|_1 + \|(\delta_j)_{T_j}\|_1,$$

where i follows by the definition of CLIME and Lemma 7.4, ii follows from the

property of the norm. Therefore,

$$\|(\widehat{\Sigma}_j)_{T_j}\|_1 + \|(\widehat{\Sigma}_j)_{T_j^c}\|_1 \leqslant \|(\widehat{\Sigma}_j)_{T_j}\|_1 + \|(\delta_j)_{T_j^c}\|_1 \Rightarrow$$
$$\|(\delta_j)_{T_j^c}\|_1 = \|(\widehat{\Sigma}_j)_{T_j^c}\|_1 \leqslant \|(\delta)_{T_j}\|_1.$$

Therefore, $\delta_j \in \mathcal{RE})(2)$. ∎

Proof of Theorem 3.11. We will write $M = \widehat{\Sigma}$ in the case of $\widehat{\beta}_{DOL}$ and $M = (\widehat{Q}+\tau I)^{-1}$ in the case of $\widehat{\beta}_{DOR}$. Denote the sparse approximation to the population precision matrix as $M^{sp} = \Sigma^{sp}$ in the case of $\widehat{\beta}_{DOL}$ and as $M^{sp} = Q^{-1}$ in the case of $\widehat{\beta}_{DOR}$. Let $s_M := \max_{1\leqslant j\leqslant d} |M_j^{sp}|$ be the sparsity index of M. Finally, we write

$$\mathbf{m}_N(x) := \sup_{S\subset\{1,2,\ldots,d\},\quad |S|\leqslant x} (\mathbb{E}\|\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it})\|^2)^{1/2}.$$

$$\sqrt{N}(\widehat{\beta} - \beta_0) = Q^{-1}\mathbb{E}_N\tilde{D}_{it}U_{it} + \underbrace{M\mathbb{E}_N\tilde{D}_{it}R_{it}}_{R_1} + \underbrace{M\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it}}_{R_2} + \underbrace{M(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it}}_{R_3}$$
$$+ \underbrace{(M\widehat{Q} - I)(-\widehat{\beta}_L + \beta_0)}_{R_4} + \underbrace{(M - Q^{-1})\mathbb{E}_N\tilde{D}_{it}U_{it}}_{R_5}.$$

Term R_1. Recognize that $\mathbb{E}[\tilde{D}_{it}R_{it}] = 0$ by orthogonality (2.7). Fix a coordinate $j \in [d]$.

$$R_{1,j} = M_j^{sp}\mathbb{E}_N\tilde{D}_{it}R_{it} + (M_j - M_j^{sp})\mathbb{E}_N\tilde{D}_{it}R_{it}.$$

We focus on the box (k, c) of size n. Conditionally on the event $\mathcal{E}_N$

$$\mathbb{E}[M_j^{sp}\mathbb{E}_{n,kc}\tilde{D}_{it}R_{it}|\mathcal{E}_N]^2 \leqslant n^{-1}\sup_{d(\cdot)\in D_N, l(\cdot)\in L_N}\mathbb{E}_Z\mathbb{E}[((M_j^{sp})'\tilde{D}_{it})^2R_{it}^2|Z_{it}, I_k^c, \mathcal{E}_N]$$
$$\lesssim n^{-1}C_{\max}(\mathbf{m}_N\|\beta_0\| + \mathbf{l}_N)^2$$

Therefore $\sqrt{n}M_j^{sp}\mathbb{E}_N\tilde{D}_{it}R_{it} = o_P(\mathbf{m}_N\|\beta_0\| + \mathbf{l}_N)$. As for the second summand,

$$\begin{aligned}|(M_j - M_j^{sp})\mathbb{E}_N\tilde{D}_{it}R_{it}| &\leqslant \|(M_j - M_j^{sp})\|_1 \max_{1\leqslant j\leqslant d}\|\mathbb{E}_N\tilde{D}_{it}R_{it}\| \\ &\leqslant 2\|(M_j - M_j^{sp})_{T_j}\|_1 \max_{1\leqslant j\leqslant d}\|\mathbb{E}_N\tilde{D}_{it}R_{it}\| \\ &\leqslant 2s_M\mu_N\sqrt{\frac{\log d}{N}} = o(1/\sqrt{N}).\end{aligned}$$

Term R_2. Decompose the term:

$$\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it} = \mathbb{E}[(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it}] + \mathbb{E}_N\left((\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it}\right)^0.$$

where the second term denotes the demeaned value. Conditionally on $\mathcal{E}_N$ the bias of the first summand is of second-order:

$$\begin{aligned}\max_{S\subset\{1,2,\ldots,d\};|S|\leqslant s_M}|M_j^{sp}\mathbb{E}[(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it}]| &\leqslant \sup_{d\in D_N, l\in L_N}\mathbb{E}\|M_j^{sp}(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})\|\mathbb{E}\|R_{it}\| \\ &\leqslant \mathbf{m}_N(s_M)(\mathbf{m}_N\|\beta_0\| \vee \mathbf{l}_N).\end{aligned}$$

$$\max_{1\leqslant j\leqslant d}|\mathbb{E}_N\left((\widehat{\tilde{D}}_{it} - \tilde{D}_{it})R_{it}\right)^0| \lesssim \sqrt{\frac{\log d}{N}},$$

Term R_3. Recognize that $\mathbb{E}[(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it}] = 0$ by the orthogonality (2.7). Fix a coordinate $j \in [d]$.

$$R_{3,j} = M_j^{sp}\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it} + (M_j - M_j^{sp})\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it}.$$

We focus on the box (k, c) of size n. Conditionally on the event $\mathcal{E}_N$

$$\begin{aligned}\mathbb{E}[M_j^{sp}\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it}]^2 &\leqslant n^{-1}\sup_{d(\cdot)\in D_N, l(\cdot)\in L_N}\mathbb{E}((M_j^{sp})'(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))M_j^{sp})U_{it}^2 \\ &\leqslant n^{-1}\mathbf{m}_N^2(s_\Sigma).\end{aligned}$$

Therefore $\sqrt{n}M_j^{sp}\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it} = o_P(\mathbf{m}_N(s_\Sigma))$. As for the second summand,

$$\begin{aligned}
|(M_j - M_j^{sp})\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})U_{it}| &\leqslant \|M_j - M_j^{sp}\|_1 \max_{1\leqslant j\leqslant d} \|\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})_j U_{it}\| \\
&\leqslant 2\|(M_j - M_j^{sp})_{T_j}\|_1 \max_{1\leqslant j\leqslant d} \|\mathbb{E}_N(\widehat{\tilde{D}}_{it} - \tilde{D}_{it})_j U_{it}\| \\
&\leqslant 2s_M\mu_N\sqrt{\frac{\log d}{N}} = o(1/\sqrt{N}).
\end{aligned}$$

Term R_4.

$$\begin{aligned}
\max_{1\leqslant j\leqslant d} |(M\widehat{Q} - I)_j(\beta_0 - \widehat{\beta}_L)| &\leqslant \max_{1\leqslant j\leqslant d} \|(M\widehat{Q} - I)_j\|_\infty \|\beta_0 - \widehat{\beta}_L\|_1 \\
&\leqslant \mu_N \left(\frac{s^2 \log d}{N}\right)^{1/2} = o(1/\sqrt{N}).
\end{aligned}$$

Term R_5.

$$\begin{aligned}
|R_{5,j}| &\lesssim_P^i \max_{1\leqslant j\leqslant d} \|(M - Q^{-1})_j\|_1 \max_{1\leqslant j\leqslant d} |\mathbb{E}_N \tilde{D}_{it} U_{it}| \\
&\lesssim_P^{ii} \max_{1\leqslant j\leqslant d} \|(M - Q^{-1})_j\|_1 \sqrt{\frac{\log d}{N}} \\
&\lesssim_P^{iii} \left(\max_{1\leqslant j\leqslant d} \|(M - M^{sp})_j\|_1 + \max_{1\leqslant j\leqslant d} \|(M^{sp} - Q^{-1})_j\|_1\right) \|\sqrt{\frac{\log d}{N}}.
\end{aligned}$$

where i holds by Cauchy Schwartz, ii holds by the Azouma-Hoeffding inequality for a martingale difference sequence, and iii holds the by triangle inequality for the norm. In the case $\widehat{\beta} = \widehat{\beta}_{DOL}$ the last term can be bounded by:

$$\lesssim_P (s_\Sigma\mu_N + r_N)\sqrt{\frac{\log d}{N}} = o(1/\sqrt{N})$$

by Lemma 3.2 ($s_\Sigma\mu_N$ term) and Assumption 3.10 (r_N term). In the case $\widehat{\beta} = \widehat{\beta}_{DOR}$ the last term can be bounded by:

$$\lesssim_P (\tau_N\sqrt{d}).$$

iv holds by the choice of τ_N in Definition 3.11 and the choice of $M^{sp} = Q^{-1}$. ∎

Corollary 7.1 (Convergence of approximation error)**.** *Suppose the coordinates of*

the vector Z_{it} are a.s. bounded (i.e., there exists $C_Z < \infty$ such that $\|Z_{it}\|_\infty \leqslant C_Z$ a.s.).Then the out-of-sample squared approximation error of the treatment and the outcome reduced form exhibits the following bound:

$$\mathbb{E}_N(\widehat{p}_i(Z_{it}) - p_{i0}(Z_{it}))^2 \lesssim_P \frac{\max(s_\pi^2, s_a^2)\log^3(p \vee N)}{N},$$
$$\mathbb{E}_N(\widehat{l}_i(Z_{it}) - l_{i0}(Z_{it}))^2 \lesssim_P \frac{\max(s_\pi^2, s_b^2)\log^3(p \vee N)}{N}.$$

Denote

$$\boldsymbol{m}_N := \frac{\max(s_\pi, s_a)\log^{3/2}(p \vee N)}{\sqrt{N}}, \quad \boldsymbol{l}_N := \frac{\max(s_\pi, s_b)\log^{3/2}(p \vee N)}{\sqrt{N}}.$$

Proof of Corollary 7.1. Fix a partition $k \in [K]$.The choice of $\lambda^D = \lambda^Y = \sqrt{N\log(p \vee N)}$ yields the following bound:

$$\begin{aligned}
a_{kc} &= \frac{1}{N}\sum_{(i,t)\in I_k}(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))^2 = \frac{1}{N}\sum_{(i,t)\in I_k}(Z_{it}(\widehat{\pi}^D - \pi_0^D) + \widehat{a}_i - a_i)^2 \\
&\leqslant \frac{2}{N}\sum_{(i,t)\in I_k}(Z_{it}(\widehat{\pi}^D - \pi_0^D))^2 + 2\|\widehat{a} - a_0\|_2^2/N \\
&\leqslant 2\|Z_{it}\|_\infty\|\widehat{\pi}^D - \pi_0^D\|_1^2 + 2\|\widehat{a} - a_0\|_2^2/N \\
&\lesssim \|Z_{it}\|_\infty\lambda^D s_\pi^2 \lesssim_P \mathbf{m}_N^2.
\end{aligned}$$

$$\begin{aligned}
b_{kc} &= \sum_{(i,t)\in I_k}(\widehat{l_k}(Z_{it}) - l_0(Z_{it}))^2 = \frac{1}{N}\sum_{(i,t)\in I_k}(Z_{it}(\widehat{\pi}^Y - \pi_0^Y) + \widehat{b}_i - b_i)^2 \\
&\leqslant \frac{2}{N}\sum_{(i,t)\in I_k}(Z_{it}(\widehat{\pi}^Y - \pi_0^Y))^2 + 2\|\widehat{b} - b_0\|_2^2/N \\
&\leqslant 2\|Z_{it}\|_\infty\|\widehat{\pi}^Y - \pi_0^Y\|_1^2 + 2\|\widehat{b} - b_0\|_2^2/N \\
&\lesssim \lambda^Y s_\pi^2 \lesssim_P \mathbf{l}_N^2.
\end{aligned}$$

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

7.2 Supplementary Lemmas

Lemma 7.1 (Lemma 2, Matrix Convergence Theory Hansen (2014))**.** *Let $(w_i)_{i=1}^N$ be an i.i.d sequence of d-vectors such that $\|w_i\| \leqslant \xi_d^2$ a.s. Suppose that for some $p > 2$*

$$(\mathbb{E}\|w_i\|^p)^{1/p} \leqslant \xi_d$$

$$o_N = \begin{cases} N^{-1/2}\xi_d^{p/(p-2)}(\log d)^{(p-4)/(2p-4)}, & \text{if } p > 4 \\ N^{-(1-2/p)}\xi_d^2 d^{4/p-1} & \text{if } 2 < p \leqslant 4 \end{cases}$$

Then, w.p. $\to 1$,

$$\|\mathbb{E}_N w_i w_i' - \mathbb{E} w_i w_i'\| \lesssim_P o_N$$

Lemma 7.2 (First Stage Error)**.** *Suppose Assumptions 3.1, 3.2 and 3.4 hold with treatment rate $\boldsymbol{d}_N$, outcome rate $\boldsymbol{l}_N$, and concentration rates $\lambda_{1N}, \lambda_{2N}$. Then the following statements hold. (1) The operator norm of the difference of $\tilde{Q}$ and $\widehat{Q}$ is bounded as:*

$$\|\tilde{Q} - \widehat{Q}\|_2 \lesssim_P \boldsymbol{d}_N^2 + \sqrt{d}\lambda_{1N}. \tag{7.13}$$

(2). If Assumption 3.8 holds, the entrywise norm of the difference of $\tilde{Q}$ and $\widehat{Q}$ is bounded as:

$$q_N := \|\tilde{Q} - \widehat{Q}\|_\infty \lesssim_P \boldsymbol{m}_N^2 + \lambda_{1N}. \tag{7.14}$$

(3). The remainder terms of Orthogonal Least Squares is bounded as:

$$\sqrt{N}\|\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}U_{it}]\|_2 \lesssim_P \sqrt{N}\boldsymbol{d}_N\boldsymbol{l}_N + \boldsymbol{d}_N^2\|\beta_0\| + \sqrt{N}\sqrt{d}(\lambda_{1N} + \lambda_{2N}) \tag{7.15}$$

and

$$\|\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it} + U_{it}] - \tilde{D}_{it}U_{it}]^2\|_2 \lesssim_P \boldsymbol{d}_N^2\|\beta_0\|^2 + d\boldsymbol{l}_N^2. \tag{7.16}$$

Proof of Lemma 7.2. **Step 1.** The difference of the sample covariance matrices of the residuals $\tilde{Q}$ and the estimated treatment residuals $\widehat{Q}$ is decomposed as:

$$\widehat{Q} - \tilde{Q} = \mathbb{E}_N \underbrace{(\tilde{D}_{it})(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{a_{it}} + (\mathbb{E}_N \underbrace{(\tilde{D}_{it})(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))')'}_{a'_{it}}$$
$$+ \mathbb{E}_N \underbrace{(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))'}_{b_{it}}.$$

Denote the sample averages of $\{a_{it}\}_{(g,t)=(1,1)}^{GT}$ and $\{b_{it}\}_{(g,t)=(1,1)}^{GT}$ within the box (k, c) by

$$\bar{a}_{kc} := \mathbb{E}_{n,kc} a_{it}, \quad \bar{b}_{kc} := \mathbb{E}_{n,kc} b_{it}. \tag{7.17}$$

Since the number of the boxes KC is fixed, it suffices to establish a probabilistic bound for the sample average of each box and apply the union bound. Recognize that $\bar{a}_{kc}$ is mean zero:

$$\mathbb{E}[\bar{a}_{kc} | I_k^c, (Z_{it})] = \mathbb{E}[\tilde{D}_{it}(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))' | I_k^c, (Z_{it})_{i,t \in I_k}] = 0$$

by the definition of the residual $\tilde{D}_{it}$ in (2.5) and the partition scheme described in the Panel Double ML Recipe. For any $\alpha \in \mathcal{S}^{d-1}$ the second moment of the vector norm of $\|\alpha' \bar{a}_{kc}\|$ is bounded as:

$$\mathbb{E}[\|\alpha' \bar{a}_{kc}\|^2 | I_k^c, (Z_{it})] = n^{-1} \sup_{d(\cdot) \in D_N} \mathbb{E}[(\alpha' \tilde{D}_{it})^2 \|d_{i0}(Z_{it}) - d(Z_{it})\|^2 | I_k^c, (Z_{it})]$$
$$\leqslant (\alpha' Q \alpha)/N \mathbf{d}_N^2 \leqslant C_{\max}^2 \mathbf{d}_N^2 / N,$$

where the first equality follows from the weak dependence structure within the box, the second inequality follows from Assumption 3.1 and the last one follows from Assumption 3.4. Applying the Markov inequality conditionally on the partition I_k^c, the controls Z_{it} on the indices of the partition I_k and the event $\widehat{d}(\cdot) \in D_N$, we get:

$$\|\bar{a}_{kc}\|_2 \leqslant \sup_{\alpha \in \mathcal{S}^{d-1}} \|\alpha' \bar{a}_{kc}\| \lesssim_P \mathbf{d}_N / \sqrt{N}. \tag{7.18}$$

Fix $\alpha \in \mathcal{S}^{d-1}$. The bias of the vector $\alpha' b_{kc}$ attains the following bound:

$$\begin{aligned}\|\mathbb{E}[\alpha' b_{kc}|I_k^c, \widehat{d}(\cdot) \in D_N]\|^2 &= \sum_{j=1}^{d} \sup_{d(\cdot)\in D_N} [\mathbb{E} \underbrace{(\alpha'(d(Z_{it}) - d_{i0}(Z_{it})))}_{A} \underbrace{(d(Z_{it}) - d_{i0}(Z_{it}))_j}_{B}]^2 \\ &\leqslant \sum_{j=1}^{d} \underbrace{\mathbb{E}(d(Z_{it}) - d_{i0}(Z_{it}))_j^2}_{\mathbb{E}A^2} \underbrace{\mathbb{E}(\alpha'(d(Z_{it}) - d_{i0}(Z_{it})))^2}_{\mathbb{E}B^2} \\ &\leqslant (\mathbf{d}_N^2)^2. \end{aligned} \tag{7.19}$$

Assumption 3.2(1) implies the following bound:

$$\|\alpha'(b_{kc} - \mathbb{E}[b_{kc}|I_k^c])\||I_k^c, \widehat{d}(\cdot)D_N] \lesssim_P \sqrt{d}\lambda_{1N}. \tag{7.20}$$

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

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

Therefore, in the case $T = 1$ λ_{1N} can be taken to be $N^{-1/2}$. Adding the bounds (7.18), (7.19), (7.20) yields:

$$\|\widehat{Q} - \tilde{Q}\| = \|a + a' + b\| \lesssim_P (\mathbf{d}_N^2 + \sqrt{d}\lambda_{1N}).$$

Step 2. By the definition of $\mathbf{m}_N$,

$$\begin{aligned}&\max_{1\leqslant m,j\leqslant d} \mathbb{E}[|b_{kc}||I_k^c, \widehat{d}(\cdot) \in D_N]_{m,j} \\ &= \max_{1\leqslant m,j\leqslant d} \sup_{d\in D_N} |\mathbb{E}[(d_{i0}(Z_{it}) - d(Z_{it}))(d_{i0}(Z_{it}) - d(Z_{it}))'|I_k^c, \widehat{d}(\cdot) \in D_N]|_{m,j} \qquad (7.21) \\ &\leqslant \mathbf{m}_N^2. \qquad (7.22)\end{aligned}$$

Fix a hold-out sample I_k^c. Conditionally on I_k^c, a_{kc} is a mean zero $d \bigtimes d$ matrix with bounded entries. The Azouma-Hoeffding inequality for a martingale difference

sequence implies:

$$\mathbb{E}[\max_{1\leqslant m,j\leqslant d^2} |\mathbb{E}_n a_{kc}|_{m,j}|I_k^c] \leqslant D^2\sqrt{\frac{\log(2d^2)}{N}}. \tag{7.23}$$

$$\mathbb{E}[\max_{1\leqslant m,j\leqslant d^2} |\mathbb{E}_n b_{kc} - \mathbb{E}[b_{kc}|I_k^c]|_{m,j}|I_k^c] \leqslant \lambda_{1N}. \tag{7.24}$$

Adding the bounds (7.21) and (7.24) delivers the result.

Step 3.

$$\begin{aligned}\left[\mathbb{E}_N[\widehat{\tilde{D}}_{it}[R_{it}+U_{it}] - \tilde{D}_{it}\tilde{U}_{it}]\right] &= \mathbb{E}_N \underbrace{(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))U_{it}}_{e_{it}} \\ &\quad + \mathbb{E}_N \underbrace{(d_{i0}(Z_{it}) - \widehat{d}_i(Z_{it}))R_{it}}_{f_{it}} + \mathbb{E}_N \underbrace{\tilde{D}_{it}R_{it}}_{g_{it}}.\end{aligned}$$

Let $\bar{e}_{kc}, \bar{f}_{kc}$, and $\bar{g}_{kc}$ be as in Equation (7.17). By conditional exogeneity (Equations (2.6) and (2.5)), $\mathbb{E}[e_{kc}|I_k^c] = 0$, $\mathbb{E}[g_{kc}|I_k^c, (Z_{it})_{i,t\in I_k}] = 0$. Furthermore, the weak dependence within the box (k,c) and Assumption 3.1 imply:

$$\begin{aligned}n\mathbb{E}[\|\bar{e}_{kc}\|^2|I_k^c] &\leqslant \sup_{l(\cdot)\in L_N}\sup_{d(\cdot)\in D_N} \mathbb{E}_{Z_{it}}[\mathbb{E}[(\tilde{U}_{it})^2|Z_{it}, I_k^c]\|(d_k(Z_{it}) - d_{k,0}(Z_{it}))^2\||I_k^c] \\ &\leqslant \bar{\sigma}^2\mathbf{d}_N^2,\end{aligned}$$

$$\begin{aligned}n\mathbb{E}[\|\bar{g}_{kc}\|^2|(I_k^c, l(\cdot)\in L_N, d(\cdot)\in D_N)] &\leqslant \sup_{l(\cdot)\in L_N}\sup_{d(\cdot)\in D_N} \mathbb{E}_{Z_{it}}[\mathbb{E}\|\tilde{D}_{it}\|^2|Z_{it}, I_k^c](R_{it})^2|I_k^c] \\ &\leqslant d\mathbf{l}_N^2 + \mathbf{d}_N^2\|\beta_0\|^2.\end{aligned}$$

The bound on $\bar{f}_{kc}$ holds:

$$\begin{aligned}\mathbb{E}[\|f_{kc}\||I_k^c] &\leqslant \mathbf{d}_N^2\|\beta_0\| + \mathbf{d}_N\mathbf{l}_N \\ \sqrt{n}(f_{kc} - \mathbb{E}[f_{kc}|I_k^c]|I_k^c) &\leqslant \sqrt{n}\sqrt{d}(\lambda_{1N} + \lambda_{2N})\end{aligned}$$

The Markov inequality implies:

$$\begin{aligned}\sqrt{n}\bar{e}_{kc} &= o_P(\bar{\sigma}\mathbf{d}_N),\\ \sqrt{n}\bar{f}_{kc} &= o_P(\sqrt{N}\mathbf{d}_N\mathbf{l}_N + \sqrt{N}\mathbf{d}_N^2\|\beta_0\| + \sqrt{d}\lambda_{1N} + \sqrt{d}\lambda_{2N}),\\ \sqrt{n}\bar{g}_{kc} &= o_P(\sqrt{d}\mathbf{l}_N + \mathbf{d}_N\|\beta_0\|)\end{aligned}$$

■

Theorem 7.1. *Let $\{\tilde{D}_{gt}, \tilde{U}_{gt}\}_{gt=1}^N$ be a martingale difference sequence of d-vectors. Let Assumptions 3.4 and 3.5 hold. Let $Q \equiv \mathbb{E}\tilde{D}'_{gt}\tilde{D}_{gt}$ be the population covariance matrix of treatment residuals, $\Gamma \equiv \mathbb{E}\tilde{D}'_{gt}U_{gt}U'_{gt}\tilde{D}_{gt}$ stand for the covariance matrix of the product of the treatment residuals and the standard errors, $\Omega \equiv Q^{-1}\Gamma Q^{-1}$, and $\Phi(t) : \mathcal{R} \to [0,1]$ be the cumulative distribution function of $N(0,1)$ random variable. Then $\forall t \in \mathcal{R}$ and for any vector $\alpha \in \mathcal{S}^{d-1}$ we have*

$$\lim_{N\to\infty}\left|\mathrm{P}\left(\frac{\sqrt{N}\alpha'Q^{-1}\mathbb{E}_N\tilde{D}'_{gt}U_{gt}}{\|\alpha'\Omega\|^{1/2}} < t\right) - \Phi(t)\right| = 0. \tag{7.25}$$

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

$$\frac{1}{N}\sum_{gt=1}^N \xi_{gt}^2 = \mathbb{E}_N\xi_{gt}^2 \to_p \frac{\alpha'Q^{-1}\mathbb{E}\tilde{D}'_{gt}U_{gt}U'_{gt}\tilde{D}_{gt}Q^{-1}\alpha}{\alpha'\Omega\alpha} = 1$$

holds by the Law of Large Numbers. The second condition

$$\mathbb{E}\xi_{gt}^2 1_{\|U_{gt}U'_{gt}\|>\epsilon} \lesssim \mathbb{E}\|U_{gt}U'_{gt}\|1_{\|U_{gt}U'_{gt}\|>\epsilon} \to 0, \epsilon \to 0$$

follows from Assumption 3.5. To see the third condition $\max_{1\leqslant gt\leqslant N}|\xi_{gt}| = O_P(1)$,

recognize that:

$$\begin{aligned}
\mathbb{E}\max_{1\leqslant gt\leqslant N}|\xi_{gt}| &\leqslant (\mathbb{E}\big(\max_{1\leqslant gt\leqslant N}|\xi_{gt}|\big)^2)^{1/2} = (\mathbb{E}\max_{1\leqslant gt\leqslant N}|\xi_{gt}|^2)^{1/2} \qquad \text{(Cauchy Scwartz)} \\
&\leqslant (\mathbb{E}\sum_{k=1}^{K}|\xi_{gt}|^2)^{1/2} \qquad (\max_{1\leqslant gt\leqslant N}|\xi_{gt}|^2 \leqslant \textstyle\sum_{k=1}^{K}|\xi_{gt}|^2) \\
&\leqslant \sqrt{N}\frac{1}{\sqrt{N}}\frac{(\alpha' Q^{-1}\mathbb{E}\tilde{D}'_{gt}U_{gt}U_{gt'}\tilde{D}'_{gt}Q^{-1}\alpha)^{1/2}}{\|\alpha'\Omega\|} = 1,
\end{aligned}$$

where the first line follows form the Cauchy–Schwartz inequality, the second line replaces the maximum $\max_{1\leqslant gt\leqslant N}|\xi_{gt}|^2$ of N positive random variables by their sum a $\sum_{k=1}^{K}|\xi_{gt}|^2$, and the third line simplifies the expression. The application of the Azouma-Hoeffding inequality delivers:

$$\mathrm{P}(\big|\max_{1\leqslant gt\leqslant N}|\xi_{gt}| - \mathbb{E}\max_{1\leqslant gt\leqslant N}|\xi_{gt}|\big| > t) \leqslant \sum_{k=1}^{K}\mathrm{P}(|\xi_{gt}| - \mathbb{E}\max_{1\leqslant gt\leqslant N}|\xi_{gt}|\big| > t) \leqslant N\exp^{-t^2/2\sigma^2}.$$

■

Lemma 7.3 (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 the outcome reduced form and U_{it} be the sampling error . Then, the following bounds hold w.h.p:*

$$\max_{1\leqslant m\leqslant d}\mathbb{E}_N\|(\widehat{d}_{m,i0}(Z_{it}) - d_{m,i0}(Z_{it}))(\widehat{l}_i(Z_{it}) - l_{i0}(Z_{it}))\| \leqslant (\lambda_{2N} + \boldsymbol{m}_N\boldsymbol{l}_N) \tag{7.26}$$

$$\max_{1\leqslant m\leqslant d}\mathbb{E}_N\|(\widehat{d}_{m,i0}(Z_{it}) - d_{m,i0}(Z_{it}))(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))'\beta_0\| \leqslant (\lambda_{1N}\sqrt{s} + \boldsymbol{m}_N^2\|\beta_0\|^2) \tag{7.27}$$

In addition, by the Azouma-Hoeffding inequality

$$\max_{1\leqslant m\leqslant d}\mathbb{E}_N\|\tilde{D}_{i,m}(\widehat{l}_i(Z_{it}) - l_{i0}(Z_{it}))\| \leqslant (DL\sqrt{\frac{\log(2d)}{N}}) \tag{7.28}$$

$$\max_{1\leqslant m\leqslant d}\mathbb{E}_N\|\tilde{D}_{i,m}(\widehat{d}_i(Z_{it}) - d_{i0}(Z_{it}))'\beta_0\| \leqslant (D^2\sqrt{\frac{\log(2d)}{N}}) \tag{7.29}$$

and

$$\max_{1\leqslant m\leqslant d}\|\mathbb{E}_N|(\widehat{d}_{m,i0}(Z_{it}) - d_{m,i0}(Z_{it}))U_{it}\| \leqslant (D^2\bar{\sigma}^2\sqrt{\frac{\log 2d}{N}})$$

Proof of Lemma 7.3. Let

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

where

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

is the first stage approximation error.

Step 1. Fix a coordinate number $m \in \{1, 2, \ldots, d\}$. Define the following quantities:

$$e_{kc,m} = \mathbb{E}_{n,kc}(\widehat{d}_{m,i0}(Z_{it}) - d_{m,i0}(Z_{it}))U_{it},$$

$$f_{kc,m} = \mathbb{E}_{n,kc}(\widehat{d}_{m,i0}(Z_{it}) - d_{m,i0}(Z_{it}))R_{it},$$

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

Conditionally on I_k^c, $\mathbb{E}[e_{kc,m}|I_k^c] = 0$ and $\mathbb{E}[g_{kc,m}|I_k^c] = 0$ for any partition $k \in [K]$, cluster index $c \in [C]$, and coordinate $m \in [d]$. Conditionally on I_k^c,

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

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

$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|e_{kc,m}\||I_k^c] = O(\bar{\sigma}D\sqrt{\frac{\log d}{N}}),$$

$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|(f_{kc,m})^0\||I_k^c] = O(\sqrt{s}\lambda_{1N} + \lambda_{2N}),$$

$$\mathbb{E}[\max_{1\leqslant m\leqslant d} \|g_{kc,m}\||I_k^c] = O([D^2 s + DL]\sqrt{\frac{\log d}{N}}).$$

Therefore,

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

■

Lemma 7.4 (Relation between the orthogonalization sets of true and estimated residuals)**.** *Let $\widehat{Q}$ and $\tilde{Q}$ be the sample covariance matrices of estimated treatment residuals and treatment residuals, respectively. Let $M_{\mu_N}(\widehat{Q})$ and $M_{\mu_N}(\tilde{Q})$ be their respective orthogonalization sets with the common rate $\mu_N = \sqrt{\frac{\log d}{N}}$. (1) Then $\forall M \in M_{\mu_N}(\tilde{Q})$ that satisfy Assumption 3.10, with high probability*

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

(2) The precision matrix Q^{-1} and its sparse approximation Σ^{sp} belong to $M_{\mu_N}(\tilde{Q}$.

Proof of Lemma 7.4. We show (1). Let M be any matrix that satisfies Assumption 3.10. Then

$$|M\widehat{Q} - I|_\infty \leqslant |M\tilde{Q} - I|_\infty + |M(\widehat{Q} - \tilde{Q})|_\infty \leqslant \mu_N + |M(\widehat{Q} - \tilde{Q})|_\infty,$$

where the last inequality is assumed in the statement of the Lemma.

$$\begin{aligned}
|M(\widehat{Q} - \tilde{Q})|_\infty &\leqslant \max_{1\leqslant m,j\leqslant d}|M_{j,\cdot}(\widehat{Q} - \tilde{Q})_{\cdot,m}| \\
&\leqslant \max_{1\leqslant j\leqslant d}\sum_{i=1}^{d}|M|_{j,i}\max_{1\leqslant m\leqslant d}|(\widehat{Q} - \tilde{Q})_{i,m}| \\
&\lesssim (s_\Sigma B_{\max} + \mu_N c_\mu)q_N.
\end{aligned}$$

where the last inequality follows from Assumption 3.10 and Lemma 7.2(b). Q.E.D. (2). By Lemma 6.2 of Javanmard and Montanari (2014) there exists a constant $c_2 > 0$

such that:

$$\mathrm{P}(|Q^{-1}\tilde{Q} - I|_\infty \geqslant c_\mu \sqrt{\frac{\log d}{N}}) \leqslant 2d^{-c_2}. \tag{7.30}$$

Therefore $Q^{-1} \in M_N(\tilde{Q})$. By (1) $Q^{-1} \in M_N(\widehat{Q})$. Consider the sparse approximation Σ^{sp} of Q.

$$\begin{aligned}
|\Sigma^{sp}\tilde{Q} - I|_\infty &\leqslant |Q^{-1}\tilde{Q} - I|_\infty + |(Q^{-1} - \Sigma^{sp})\tilde{Q}|_\infty \\
&\leqslant c_\mu \sqrt{\frac{\log d}{N}} + \max_{1 \leqslant j \leqslant d} \|Q^{-1}_{j,\cdot} - \Sigma^{sp}_j\|_1 \|\tilde{Q}\|_\infty \\
&\leqslant c_\mu \sqrt{\frac{\log d}{N}} + o(\sqrt{\frac{\log d}{N}}).
\end{aligned}$$

Therefore, Σ^{sp} belongs to $M_{\mu_N}(\tilde{Q})$ and $M_{\mu_N}(\widehat{Q})$.

■

References

Andrews, D. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 1(62):43–72.

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

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

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18:5–46.

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., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017a). Double/debiased machine learning for treatment and causal parameters.

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*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. (2017b). Locally robust semiparametric estimation.

Chevalier, J., Kashyap, A., and Rossi, P. (2003). Why don't prices rise during periods of peak demand? evidence from scanner data. *American Economic Review*, 93(1):15–37.

Gandhi, A. and Houde, J.-F. (2016). Measuring substitution patterns in differentiated products industries. *University of Wisconsin-Madison and Wharton School*.

Hansen, B. (2014). A unified asymptotic distribution theory for parametric and non-parametric least squares.

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.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

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.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. 62:245–271.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, 213(57).

Robinson, P. M. (1988). Root-n consistent semiparametric regression.

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.

van der Vaart, A. (1998). Asymptotic statistics.

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*.