On Policy Evaluation with Aggregate Time-Series Shocks[*]

Dmitry Arkhangelsky[†] Vasily Korovkin[‡]

April 6, 2021

Abstract

We propose a new algorithm for estimating treatment effects in contexts where the exogenous variation comes from aggregate time-series shocks. Our estimator combines data-driven unit-level weights with a time-series model. We use the unit weights to control for unobserved aggregate confounders and use the time-series model to extract the quasi-random variation from the observed shock. We examine our algorithm's performance in a realistic simulation based on Nakamura and Steinsson [2014]. We provide statistical guarantees for our estimator in a practically relevant regime, where both cross-sectional and time-series dimensions are large and show how to use it to conduct robust inference.

Keywords: Continuous Difference in Differences, Panel Data, Causal Effects, Instrumental Variables, Treatment Effects, Unobserved Heterogeneity, Synthetic Control.

JEL Classification: C18, C21, C23, C26.

[*] This paper benefited greatly from our discussions with Manuel Arellano, Stéphane Bonhomme, David Hirshberg, Guido Imbens, and Dmitry Mukhin. We also want to thank Martin Almuzara, Maxim Ananyev, Kirill Borusyak, Ivan Canay, Peter Hull, Alexey Makarin, Monica Martinez-Bravo, Nikolas Mittag, Eduardo Morales, Imil Nurutdinov, Christian Ochsner, and Liyang Sun as well as participants of seminars at Carlos III, CEMFI, Northwestern University, UCL, World Congress of Econometric Society, and ASSA meeting for helpful comments and suggestions. Asya Evdokimova and Gleb Kurovskiy provided excellent research assistance. Dmitry Arkhangelsky gratefully acknowledges financial support from "María de Maeztu Units of Excellence" Programme MDM-2016-0684. Vasily Korovkin gratefully acknowledges financial support from the Czech Science Foundation grant (19-25383S).

[†] CEMFI, darkhangel@cemfi.es.

[‡] CERGE-EI (a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences), vasily.korovkin@cerge-ei.cz.

1 Introduction

Changes in aggregate variables are commonly used to evaluate economic policies. The most popular design of this type is an "event study", where a one-time aggregate shock, e.g., a new law, affects some population of units but not others, and we observe both over time. To quantify the effects of these shocks, practitioners use either difference in differences or, more recently, synthetic control methodology (e.g., Ashenfelter and Card, 1984, Card and Krueger, 1993, Abadie and Gardeazabal, 2003, Bertrand et al., 2004, Abadie et al., 2010). Often, when both outcome and treatment variables are at the individual level, this approach is used as a first stage, and the aggregate change effectively plays the role of an instrument. In the absence of a single aggregate shock, researchers often employ more general time-series variation to establish causal links between unit-specific policy and outcome variables. In a typical application, outcomes and treatments are observed at some geographical level over time (e.g., Duflo and Pande, 2007, Dube and Vargas, 2013, Nakamura and Steinsson, 2014, Nunn and Qian, 2014, Guren et al., 2020b, Dippel et al., 2020, Barron et al., 2021). To address a potential endogeneity problem, researchers use aggregate time-series shocks as instruments. A standard econometric tool employed to analyze such data is a two-stage least-squares (TSLS) regression with unit and time fixed effects.[1]

Specifically, let Y_{it} be the outcome variable, W_{it} the endogenous regressor, and assume that we observe a balanced panel with n units and T periods. To establish a causal link between Y_{it} and W_{it}, the following regression is estimated by TSLS:

$$Y_{it} = \alpha_i + \mu_t + \tau W_{it} + \epsilon_{it}, \tag{1.1}$$

using $D_i Z_t$ as an instrument. Here, Z_t is an aggregate shock, D_i is a measure of "exposure" of unit i to this aggregate shock, and τ is the parameter of interest. For example, in Nunn and Qian [2014], W_{it} is the amount of food aid that country i received, and Y_{it} is a measure of local conflict, Z_t is the amount of wheat produced in the United States in the previous year, and D_i is a share of periods when country i received food aid.

In this paper, we propose a new estimator for the causal effects in applications with aggregate instruments. We demonstrate both theoretically and using simulations that our method

[1] See Arellano [2003] for a textbook treatment of TSLS with panel data.

dominates the conventional TSLS approach in a rich class of statistical models and show how to use it to conduct robust inference. To understand the motivation behind our algorithm, consider a generalization of (1.1):

$$Y_{it} = \alpha_i + \mu_t + \tau W_{it} + \theta_i H_t + \epsilon_{it}, \tag{1.2}$$

where H_t is an unobserved aggregate shock with unobserved exposure θ_i, and all other variables are the same as before. The danger of unobserved aggregate confounders with heterogeneous exposures for TSLS identification strategy is well-recognized in applied work (Nunn and Qian, 2014, Nakamura and Steinsson, 2014, Guren et al., 2020b, Chodorow-Reich et al., 2021). For example, in Nakamura and Steinsson [2014] the authors are interested in the effect of local military procurement spending (W_{it}) in the United States on the regional output growth (Y_{it}), and use the national military spending as an instrument (Z_t). In this case, H_t can represent fiscal and monetary policy changes or the general political and business cycle in the United States.

To justify TSLS regression (1.1) practitioners need to assume that either D_i is uncorrelated with θ_i, or Z_t is uncorrelated with H_t. Both of these assumptions are questionable for different reasons. In applications, D_i is rarely randomly assigned; instead, it is either a fixed characteristic of a unit or a quantity constructed directly from the data. As a result, we cannot expect it to be uncorrelated with θ_i. Similarly, aggregate instrument Z_t can be correlated with H_t for two different reasons. Either both of these variables share common trends, or their innovations are correlated. Our algorithm deals with all of these problems. First, to address potential spurious correlation caused by common trends, we use a demeaned version of Z_t. Second, we use the data to construct exposures ω_i and employ them instead of D_i. We design our exposures so that they are correlated with D_i but are approximately orthogonal to θ_i.

To understand the logic behind our construction of ω_i it is useful first to explain the mechanics of (1.1). The TSLS algorithm uses D_i to aggregate outcomes over units and then uses Z_t to construct a time-series IV estimator. In particular, we show that $\hat{\tau}_{TSLS}$ is a ratio of two OLS coefficients in the following time-series regressions:

$$\begin{aligned} Y_t &= \alpha^{(y)} + \delta Z_t + \epsilon_t^{(y)}, \\ W_t &= \alpha^{(w)} + \pi Z_t + \epsilon_t^{(w)}, \end{aligned} \tag{1.3}$$

where W_t and Y_t are weighted averages of W_{it} and Y_{it}, with the weights proportional to $D_i - \overline{D}$. Representation (1.3) is the starting point of our analysis. If unobserved H_t affects Y_{it} and W_{it} in a way that is correlated with D_i, then the aggregate errors $(\epsilon_t^{(y)}, \epsilon_t^{(w)})$ should be relatively large. Alternatively, if we construct Y_t and W_t using the weights that are uncorrelated with exposures to H_t, then the aggregate errors should be relatively small. Crucially, we need these weights to be correlated with D_i, otherwise, we can eliminate Z_t as well. To this end, we use a part of the data (the first third of the periods) to construct weights ω_i in such a way that the aggregate errors $\epsilon_t^{(w)}, \epsilon_t^{(y)}$ are small, but the covariance between ω_i and D_i is large.

To provide an interpretation of our estimator, we develop a parsimonious causal model that captures common problems researchers might face in applications. Importantly, our model allows for aggregate shocks – both observed and unobserved – to be determined in equilibrium together with the unit-level outcomes as long as they are affected by aggregate noise. This is crucial for applications in macroeconomics and related fields where it is rarely possible to find completely exogenous aggregate variables. Our setup also incorporates situations where both outcome and treatment variables are determined in a local equilibrium and are affected by unobserved aggregate shocks. An example of this problem is a study of housing wealth effects by Guren et al. [2020b]. Our model features generalized fixed effects that go far beyond the standard two-way structure imposed in (1.1) and (1.2).

We analyze the properties of our method in a high-dimensional regime where n is similar in size to T. This choice is motivated by the applications where n and T are often comparable. We prove that our algorithm delivers consistent and $\sqrt{T}$-convergent estimators even in the presence of confounding aggregate shocks. We also show that our method can be used to conduct valid inference, as long as the variance of the idiosyncratic unit-level errors is small. We demonstrate the benefits of our approach using a data-driven simulation based on Nakamura and Steinsson [2014] study of fiscal multiplier in the USA. We show that our estimator remains competitive when the model (1.1) holds and TSLS is the best estimator, and is a clear winner in more realistic situations with unobserved aggregate shocks.

Our results are connected to the recent literature on causal panel data methods. We view our estimator as a part of the broad research agenda that started with synthetic control methods (Abadie and Gardeazabal, 2003, Abadie et al., 2010, Doudchenko and Imbens, 2016), and continues with recent innovations such as Synthetic Difference in Differences (Arkhangelsky et al., 2019), and Augmented Synthetic Control (Ben-Michael et al., 2018). Our proposal allows researchers to apply these ideas to much broader contexts with endogenous unit-level variables. Our analysis is based on the combination of the design-based assumptions and a particular model for the outcomes. The benefits of considering both outcome and assignment models was emphasized recently in Arkhangelsky and Imbens [2021] in the context of strictly exogenous treatments. Finally, our demeaning procedure is directly related to the correction proposed in Borusyak and Hull [2020] for general applications with quasi-experimental shocks.

Our method is related to the literature in empirical macroeconomics that explicitly constructs D_i using the data. Recent examples of this approach include Nakamura and Steinsson [2014], and Guren et al. [2020b] (see also Guren et al. [2020a]). In this work, the authors construct D_i by running unit-specific regressions of W_{it} on aggregate shocks and possibly other individual-specific variables. Our algorithm and its analysis are quite different from these proposals. We find the weights by looking at how well they balance aggregate variation instead of looking at unit-level responses to shocks. At the same time, our method requires D_i as an input, thus making it complementary to the methods proposed in Nakamura and Steinsson [2014], and Guren et al. [2020b]. Notably, our estimator remains consistent even in the presence of correlated unobserved shocks with heterogeneous exposures.

Our model is also related to the recent econometric literature on shift-share designs (Adao et al., 2019, Borusyak et al., 2018, Goldsmith-Pinkham et al., 2020). Similar to this literature, we consider situations where an instrument has a particular product structure. Our focus, however, is quite different: we propose and analyze a new estimator, while the literature has been focused on the properties of the standard IV estimator under alternative assumptions. We also relax the standard exogeneity assumption made in the shift-share literature and allow for unobserved aggregate shocks that affect different units differently. Our estimator is designed for a case with a single aggregate shock, and thus it does not directly apply to a classical shift-share setup with multiple industry-level shocks (e.g., Autor et al., 2014). However, we believe that our ideas can be extended to these applications, potentially allowing researchers to flexibly use both time and industry dimensions.

The paper proceeds as follows: in Section 2, we discuss the mechanics of TSLS regression (1.1) in more detail, present our algorithm, and show its performance in a simulation exercise. In Section 3 we introduce the causal model along with statistical restrictions and demonstrate the formal properties of our algorithm. We discuss inference in Section 3.4. Section 4 discusses possible extensions of our algorithm, connections to literature in empirical macroeconomics, and shift-share designs. Finally, Section 5 concludes.

2 Algorithm

2.1 TSLS mechanics

A common algorithm for estimating causal effects with aggregate shocks is a TSLS regression:

$$Y_{it} = \alpha_i^{(y)} + \mu_t^{(y)} + \tau W_{it} + \epsilon_{it}^{(y)}, \tag{2.1}$$

with $D_i Z_t$ as an instrument. Here W_{it} is the policy variable of interest, Y_{it} is the outcome, Z_t is the aggregate shock (instrument) and D_i is an available measure of exposure of unit i to Z_t. Regression (2.1) can be split into two parts – the reduced form and the first stage:

$$\begin{aligned} Y_{it} &= \tilde{\alpha}_i^{(y)} + \tilde{\mu}_t^{(y)} + \delta D_i Z_t + \tilde{\epsilon}_{it}^{(y)}, \\ W_{it} &= \alpha_i^{(w)} + \mu_t^{(w)} + \pi D_i Z_t + \epsilon_{it}^{(w)}, \end{aligned} \tag{2.2}$$

where $\tilde{\alpha}_i^{(y)} := \alpha_i^{(y)} + \tau\alpha_i^{(w)}$, $\tilde{\mu}_t^{(y)} := \mu_t^{(y)} + \tau\mu_t^{(w)}$, $\tilde{\epsilon}_{it} := \epsilon_{it}^{(y)} + \tau\epsilon_{it}^{(w)}$, and $\delta = \tau\pi$. Standard logic implies that a TSLS estimator is a ratio of two OLS estimators with two-way fixed effects:

$$\hat{\tau}_{TSLS} = \frac{\hat{\delta}_{OLS}^{(fe)}}{\hat{\pi}_{OLS}^{(fe)}}. \tag{2.3}$$

Alternatively, one can represent the same estimator as a ratio of two OLS estimators from the following time-series regressions:

$$Y_t = \tilde{\alpha}^{(y)} + \delta Z_t + \tilde{\epsilon}_t^{(y)}, \quad W_t = \alpha^{(w)} + \pi Z_t + \epsilon_t^{(w)}, \tag{2.4}$$

that is $\hat{\tau}_{TSLS} = \frac{\hat{\delta}^{(ts)}_{OLS}}{\hat{\pi}^{(ts)}_{OLS}}$. Here the aggregate variables are defined as follows:

$$\begin{aligned} &Y_t := \frac{1}{n}\sum_{i\leq n}\frac{Y_{it}(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]},\ W_t := \frac{1}{n}\sum_{i\leq n}\frac{W_{it}(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]},\ \tilde{\epsilon}^{(y)}_t := \frac{1}{n}\sum_{i\leq n}\frac{\tilde{\epsilon}_{it}(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]}, \\ &\epsilon^{(w)}_t := \frac{1}{n}\sum_{i\leq n}\frac{u_{it}(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]},\ \alpha^{(w)} := \frac{1}{n}\sum_{i\leq n}\frac{\alpha^{(w)}_i(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]},\ \tilde{\alpha}^{(y)} := \frac{1}{n}\sum_{i\leq n}\frac{\tilde{\alpha}^{(y)}_i(D_i-\overline{D})}{\hat{\mathbb{V}}[D_i]}. \end{aligned} \tag{2.5}$$

We aggregate Y_{it}, W_{it} with weights that sum up to zero, and thus time fixed effects are not present in (2.4). We collect these statements in the following straightforward Lemma.

Lemma 1. (Representation) *Suppose that (2.1) is estimated by TSLS regression with D_iZ_t as an instrument. Then the following numerical equivalence holds:*

$$\hat{\delta}^{(fe)}_{OLS} = \hat{\delta}^{(ts)}_{OLS}, \quad \hat{\pi}^{(fe)}_{OLS} = \hat{\pi}^{(ts)}_{OLS}, \quad \hat{\tau}_{TSLS} = \frac{\hat{\delta}^{(fe)}_{OLS}}{\hat{\pi}^{(fe)}_{OLS}} = \frac{\hat{\delta}^{(ts)}_{OLS}}{\hat{\pi}^{(ts)}_{OLS}}. \tag{2.6}$$

This result can be trivially extended to cases with time-invariant covariates (e.g., the state-level time fixed effects). The lemma's primary purpose is to demonstrate two distinct aspects of the TSLS algorithm: unit weights are used to aggregate the outcomes, and time-series IV regression is used to estimate τ. The algorithm that we propose next has a similar structure but implements both steps differently.

2.2 Description of the algorithm

Our algorithm has two steps. We use the initial part of the sample (observations from $t \leq T_0$, where T_0 is selected by the researcher) to construct the unit weights ω_i. We then use these weights to construct aggregate outcomes and estimate the reduced form and first stage coefficients by running OLS regressions.

2.2.1 Time-series model

A key input for our algorithm is the model for the mean of Z_t. To understand the need for this object consider a general decomposition:

$$\begin{aligned} &Z_t = \mu_t^{(z)} + \epsilon_t^{(z)}, \\ &\mathbb{E}[\epsilon_t^{(z)}] = 0. \end{aligned} \tag{2.7}$$

We saw that the TSLS estimator can be represented as a time-series regression of the aggregate variables Y_t, W_t on Z_t. In applications it is common to assume that Z_t is exogenous, and thus regressions (2.4) seem reasonable. However, because Y_t, W_t, Z_t are time-series we need to acknowledge the possibility of the spurious correlation caused by similar trends in Z_t and Y_t, W_t. This is a particular instance of a more general problem analyzed in Borusyak and Hull [2020].

Our solution to this problem is similar in spirit to Borusyak and Hull [2020]. We do not assume that $\mu_t^{(z)}$ is known, but instead assume that researchers have access to functions $\psi_t = (1, \dots, \psi_{pt})$ that span $\mu_t^{(z)}$:

$$\mu_t^{(z)} = \eta_z^\top \psi_t. \tag{2.8}$$

Here ψ_t can include different kinds of deterministic trends, or additional strictly exogenous aggregate variables that can explain the variation in Z_t. In the absence of any additional information and knowledge about Z_t, a natural choice for ψ_t is $\psi_t \equiv 1$ – a constant mean. In practice, Z_t might be observed at a higher frequency then the unit-level data, and this additional information can be used to construct ψ_t. For example, one can use filtering and related procedures (e.g., Hamilton [2018]) for this purpose.

2.2.2 Unit weights

Given ψ_t, the first step of our algorithm focuses on the unit weights that are later used to aggregate the outcomes. We construct these weights by solving a quadratic optimization problem:

$$
\begin{aligned}
&(\omega, \hat{\eta}_\psi^{(w)}, \hat{\eta}_z^{(w)}, \hat{\eta}_\psi^{(y)}, \hat{\eta}_z^{(y)}) = \\
&= \underset{\{w,\eta_\psi^{(w)},\eta_z^{(w)},\eta_\psi^{(y)},\eta_z^{(y)}\}}{\arg\min} \Bigg\{ \zeta^2 \frac{T_0}{n^2}\|w\|_2^2 + \frac{\sum_{t=1}^{T_0}\left(\frac{1}{n}\sum_{i=1}^n w_i Y_{it} - (\eta_\psi^{(y)})^\top \psi_t - \eta_z^{(y)} Z_t\right)^2}{\hat{\sigma}_Y^2} + \\
&\qquad \frac{\sum_{t=1}^{T_0}\left(\frac{1}{n}\sum_{i=1}^n w_i W_{it} - (\eta_\psi^{(w)})^\top \psi_t - \eta_z^{(w)} Z_t\right)^2}{\hat{\sigma}_W^2} \Bigg\}
\end{aligned}
\tag{2.9}
$$

subject to:

$$
\frac{1}{n}\sum_{i=1}^n w_i D_i = 1, \quad \frac{1}{n}\sum_{i=1}^n w_i = 0,
$$

where $\hat{\sigma}_k^2$ are scaling factors:

$$
\hat{\sigma}_Y^2 = \frac{1}{nT}\sum_{it}(Y_{it} - \overline{Y})^2, \quad \hat{\sigma}_W^2 = \frac{1}{nT}\sum_{it}(W_{it} - \overline{W})^2, \tag{2.10}
$$

and ζ is a user-specified regularization parameter.

The optimization problem (2.9) is not quite standard and to gain intuition it is useful to consider an edge case. If ζ is equal to infinity, then the last two parts of the optimized function do not matter and it is straightforward to show that $\omega_i = \frac{D_i - \overline{D}}{\hat{\mathbb{V}}[D_i]}$. In other words, we get the same unit weights that were used in Lemma 1. Once we start to decrease ζ the second two terms start to play a role, thus forcing the following approximate equalities (for $t \le T_0$):

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \omega_i Y_{it} \approx \left(\hat{\eta}_\psi^{(y)}\right)^\top \psi_t + \hat{\eta}_z^{(y)} Z_t, \\
&\frac{1}{n}\sum_{i=1}^n \omega_i W_{it} \approx \left(\hat{\eta}_\psi^{(w)}\right)^\top \psi_t + \hat{\eta}_z^{(w)} Z_t.
\end{aligned}
\tag{2.11}
$$

The motivation for enforcing (2.11) comes from applications in which (2.1) is estimated. The common concern in practice, is that while Z_t is an exogenous shock, it might not be the only aggregate variable that affects both the outcome and the endogenous policy variable. The conventional assumption is that such unobserved shocks, if present, either do not affect units differently – and thus are captured by time fixed effects – or affect them in a way that is unrelated to D_i. In either of these situations, one should expect (2.11) to be satisfied for $\omega_i = \frac{D_i - \overline{D}}{\hat{\mathbb{V}}[D_i]}$. In this case, our algorithm should produce the unit weights that are similar to those that are currently used.

In practice, D_i is a characteristic of a unit, and one cannot expect it to be randomly assigned across units. As a result, a priori, there is no reason to believe that the unobserved shocks affect units in a way that is unrelated to D_i. However, it is natural to assume that these shocks do not entirely mimic Z_t and ψ_t, and there is a combination of units that it is affected by Z_t, ψ_t, and nothing else. An empirical manifestation of such combination is a property like (2.11), where most of the variation in aggregate variables can be attributed to observed variables. As a result, by using the unit weights that enforce (2.11) we might hope to balance out the unobserved shocks. This is the motivation behind the optimization problem (2.9). In the coming sections, we formalize this intuition and provide formal statistical guarantees for this property of the unit weights ω.

2.2.3 Aggregate regressions

The second step of our algorithm consists of two time-series regressions. To this end we construct aggregate variables for $t > T_0$:

$$Y_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}\omega_i, \qquad W_t = \frac{1}{n}\sum_{i=1}^{n} W_{it}\omega_i, \tag{2.12}$$

and use them to estimate the first stage and reduced form coefficient by OLS. In particular, we run the following regressions for $t > T_0$:

$$\begin{aligned} Y_t &= \beta^{(y)} + (\eta_\psi^{(y)})^\top \psi_t + \delta Z_t + \varepsilon_t^{(y)}, \\ W_t &= \beta^{(w)} + (\eta_\psi^{(w)})^\top \psi_t + \pi Z_t + \varepsilon_t^{(w)}, \end{aligned} \tag{2.13}$$

and use $\hat{\delta}, \hat{\pi}$ either to construct the usual IV ratio $\hat{\tau} := \frac{\hat{\delta}}{\hat{\pi}}$, or conduct inference (see Section 3.4).

One can immediately see that our estimator has three important differences compared to the conventional algorithm described in Section 2.1. The critical difference is that we construct the unit weights by solving (2.9) and use these weights to aggregate the outcomes. The second difference is that we include functions ψ_t in the aggregate regressions (2.13) to address potential

```
Algorithm 1: Estimation algorithm
  Data: $\{Y_{it}, W_{it}\}_{it}, \{D_i\}_{i=1}^n, \{Z_t, \psi_t\}_{t=1}^T, T_0, \zeta$
  Result: First-stage and reduced form estimates $(\hat{\pi}, \hat{\delta})$
  Construct the unit weights $\{\omega_i\}_{i=1}^n$ by solving optimization problem (2.9);
  for $t \leftarrow T_0 + 1$ to $T$ do
      Construct $Y_t = \frac{1}{n}\sum_{i=1}^n Y_{it}\omega_i$, and $W_t = \frac{1}{n}\sum_{i=1}^n W_{it}\omega_i$.
  end
  Using the data for $t > T_0$ estimate two regressions by OLS:

      $Y_t = \beta^{(y)} + (\eta_\psi^{(y)})^\top \psi_t + \delta Z_t + \varepsilon_t^{(y)},$
      $W_t = \beta^{(w)} + (\eta_\psi^{(w)})^\top \psi_t + \pi Z_t + \varepsilon_t^{(w)}$

    and report $\hat{\delta}$, $\hat{\pi}$;
```

spurious correlation. Finally, we use sample splitting and estimate (2.13) only using the data for $t > T_0$.

Our algorithm has two tuning parameters ζ and T_0. Theoretical results in Section 3 show how the resulting error depends on T_0 if ζ^2 is of constant order. In practice, we recommend setting T_0 to $\frac{T}{3}$, i.e., using a third of the available data to learn the weights and the rest to estimate the parameters. We use the following expression to choose the regularization parameter:

$$\zeta := \frac{\min\left\{\sigma_{T/2}(Y), \sigma_{T/2}(W)\right\}}{\sqrt{n+T}}, \tag{2.14}$$

where $\sigma_k(\cdot)$ corresponds to the k-th largest singular value of the matrix.

2.3 Illustration

We illustrate the performance of our algorithm in a Monte-Carlo experiment. To construct this simulation, we rely on the data and analysis from Nakamura and Steinsson [2014], where the authors investigate the relationship between government spending and state GDP growth. They use state data on total military procurement for 1966 through 2006 and combine it with U.S. Bureau of Economic Analysis state GDP and state employment datasets. The authors complement these data with the oil prices data from the St. Louis Federal Reserve's FRED database and state-level inflation series constructed by Del Negro [1998] and their inflation calculations for after 1995.

A crucial quantity that Nakamura and Steinsson [2014] want to capture by estimating growth-spending relationship is an open economy relative multiplier. They compare different U.S. states and study their reaction to aggregate military spending fluctuations in a panel setting. They argue that this strategy allows them to control for common shocks (such as monetary policy). It also allows them to account for the potential endogeneity of local procurement spending.

To illustrate their approach, we introduce some notation — also used in our simulations below. For a generic observation – a state i, and a generic period t, denote per capita output growth in state i from year $t-2$ to t by Y_{it}.[2] Similarly, denote two-year growth in per capita military procurement spending in state i and year t, normalized by output, in year $t-2$, by W_{it}. Finally, let Z_t be the change in total national procurement from year $t-2$ to t. This leads to a dataset with $n = 51$ states and $T = 39$ periods.

In their baseline specification, the authors interact state fixed effects with the fluctuations in aggregate military spending and use this interaction as an IV for state-level military procurement. This exercise is equivalent to running state-by-state regressions to estimate the exposures and then use the weights based on these estimated exposures in a TSLS. More preciesly, the authors first construct D_i by estimating the regression for every unit i:

$$W_{it} = \alpha_i + \pi_i Z_t + u_{it}, \tag{2.15}$$

and then estimate equation

$$Y_{it} = \beta_i + \mu_t + \tau W_{it} + \varepsilon_{it} \tag{2.16}$$

by TSLS using $\hat{\pi}_i Z_t$ as an instrument.

In our experiments we try to capture the spirit of this empirical exercise and investigate how different features of the data generating process affect the performance of the algorithms. Formally, our simulations are based on the following model:

$$\begin{aligned} Y_{it} &= \beta_i^{(y)} + \mu_t^{(y)} + L_{it}^{(y)} + \tau W_{it} + \theta_i^{(y)} H_t + \epsilon_{it}^{(y)}, \\ W_{it} &= \beta_i^{(w)} + \mu_t^{(w)} + L_{it}^{(w)} + \pi_i Z_t + \theta_i^{(w)} H_t + \epsilon_{it}^{(w)}. \end{aligned} \tag{2.17}$$

[2] The authors advocate for using two-year changes instead of one-year changes together with leads and lags.

Here parameters $\{\beta_i^{(y)}, \beta_i^{(w)}, \mu_t^{(y)}, \mu_t^{(w)}, L_{it}^{(y)}, L_{it}^{(w)}, \tau, \pi_i, \theta_i^{(w)}, \theta_i^{(y)}\}_{i,t}$ are fixed, while $\epsilon_{it}^{(y)}, \epsilon_{it}^{(w)}$ and $\{Z_t, H_t\}_{t=1}^T$ are random (see Section 3 for the discussion of this model). For our simulation to be realistic we use the data described above to construct $\{L_{it}^{(y)}, L_{it}^{(w)}, \pi_i\}_{it}$, and the models for $\{Z_t\}_{t=1}^T$ and $\{\epsilon_{it}^{(y)}, \epsilon_{it}^{(w)}\}_{it}$.[3] The data are not directly informative about H_t and $\{\theta_i^{(w)}, \theta_i^{(y)}\}_i$ and we need to make ad hoc choices that we describe below.

First, we eliminate the time fixed effects from both the outcome and the policy variables by demeaning the data for each period t:

$$\tilde{W}_{it} := W_{it} - \frac{1}{n}\sum_{i=1}^n W_{it}, \quad \tilde{Y}_{it} := Y_{it} - \frac{1}{n}\sum_{i=1}^n Y_{it}. \tag{2.18}$$

Then we run the following regressions for each unit using all the periods:

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i^{(y)} + \delta_i Z_t + \varepsilon_{it}^{(y)}, \\ \tilde{W}_{it} &= \alpha_i^{(w)} + \pi_i^{(0)} Z_t + \varepsilon_{it}^{(w)}, \end{aligned} \tag{2.19}$$

define $\pi_i := \hat{\pi}_i^{(0)}$, and use it in (2.17).

For $k \in \{y, w\}$ let $E^{(k)}$ be the $n \times T$ matrix of residuals from (2.19): $(E^{(k)})_{it} := \hat{\varepsilon}_{it}^{(k)}$. We construct $L_{it}^{(k)}$ by solving the following problem:

$$L^{(k)} := \underset{M, \text{rank}(M) = 11}{\arg\min} \sum_{it} \left(E_{it}^{(k)} - M_{it}\right)^2 \tag{2.20}$$

which implies that $L^{(k)}$ simply sets all but 11 largest singular values of $E^{(k)}$ to zero. We use the residuals $E^{(k)} - L^{(k)}$ to construct the covariance matrix:

$$\Sigma := \frac{1}{nT}\sum_{it} \begin{pmatrix} \left(E_{it}^{(y)} - L^{(y)_{it}}\right)^2 & \left(E_{it}^{(y)} - L^{(y)_{it}}\right)\left(E_{it}^{(w)} - L_{it}^{(w)}\right) \\ \left(E_{it}^{(y)} - L^{(y)_{it}}\right)\left(E_{it}^{(w)} - L_{it}^{(w)}\right) & \left(E_{it}^{(w)} - L^{(w)_{it}}\right)^2 \end{pmatrix}, \tag{2.21}$$

and generate $(\epsilon_{it}^{(y)}, \epsilon_{it}^{(w)})$ from $\mathcal{N}(0, \Sigma)$. Finally, we estimate the model for Z_t by fitting an ARIMA model to the data $\{Z_t\}_{t=1}^T$ (scaled to have unit variance) using the automatic model selection package in R.

[3] There is no need to construct $\{\beta_i^{(y)}, \beta_i^{(w)}, \mu_t^{(y)}, \mu_t^{(w)}\}_{it}$ because the algorithms we consider are invariant with respect to unit and time fixed effects.

	Design 1		Design 2		Design 3		Design 4	
Estimator	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
$\hat{\pi}$	0.014	-0.000	0.038	0.000	0.028	0.010	0.128	0.078
$\hat{\pi}_{TSLS}$	0.010	0.000	0.044	-0.001	0.310	0.271	0.280	0.239
$\hat{\delta}$	0.067	-0.000	0.329	-0.005	0.108	0.030	0.284	0.030
$\hat{\delta}_{TSLS}$	0.050	0.001	0.371	-0.000	0.135	0.113	0.359	0.098
$\hat{\tau}$	0.063	-0.000	0.348	-0.001	0.102	0.029	0.325	0.038
$\hat{\tau}_{TSLS}$	0.047	0.001	0.414	0.007	0.117	0.101	0.407	0.103

Table 1: Each simulations has 2000 replications; first design: no generalized FE, no unobserved shock; second design: generalized FE, no unobserved shock; third design: no generalized FE, unobserved shock; fourth design: generalized FE, unobserved shock.

We construct H_t as a linear combination of Z_t and an independent random process that has the same distribution as Z_t. We set $\theta_i^{(w)}$ to be equal to a linear combination of $\hat{\pi}_i$ and an independent standard normal variable, and do the same for $\theta_i^{(y)}$. Parameters of these combinations, elements of matrix Σ, and parameters of the model for Z_t are presented in Appendix D. Notably, we choose these parameters to make the corresponding unobserved components similar in size to $\pi_i Z_t$.

We compare the performance of our estimator (as described by Algorithm 1) with the standard TSLS algorithm from Section 2.1. In both cases we use the data to construct D_i by running the following regressions using for $t \leq \frac{T}{3}$:

$$W_{it} = \alpha_i + \pi_i Z_t + \varepsilon_{it}, \tag{2.22}$$

and set $D_i = \hat{\pi}_i$. We consider four different designs. In the first design we drop $L_{it}^{(w)}, L_{it}^{(y)}$, as well as H_t from the model (2.17). In this case, TSLS algorithm should perform better than ours, because it uses the optimal weights. With the second design we start to increase the complexity and add $L_{it}^{(w)}, L_{it}^{(y)}$ back to the model. One can think of this design as a DGP for the data from Nakamura and Steinsson [2014] under which the TSLS approach is justified. Here we should expect both algorithms to perform well in terms of bias, but potentially differ in terms of variance. In the third design we drop $L_{it}^{(w)}, L_{it}^{(y)}$ but add H_t, finally in the fourth case we have both components.

In Table 1 we report results over 2000 for simulations for the case of $\tau = 1.43$ that corresponds to the estimate obtained in Nakamura and Steinsson [2014]. The results confirm the intuition

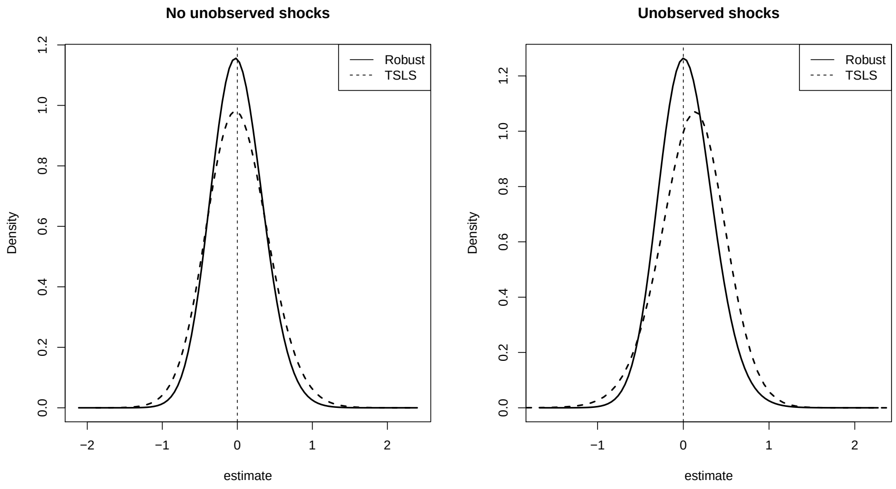

Figure 1: Distribution of errors $\hat{\tau} - \tau$ for the second and the fourth design of Table 1.

discussed above: in the simplest case, our estimator for τ is less precise than $\hat{\tau}_{TSLS}$, although the difference is small. We see sizable gains in RMSE (20%) for the second design. Notably, all parts of this design come directly from data and are not driven by our choices. In the third case our estimator eliminates most of the bias, while the TSLS error is dominated by it. Finally in the most general design our estimator is nearly unbiased and dominates the TSLS in terms of RMSE. In Figure 1 we plot the densities of $\hat{\tau} - \tau$ over the simulations for the second and the fourth design. These plots demonstrate the gains in both variance and bias and show the estimator's overall behavior. Once again, we see that even when TSLS is approximately unbiased, there are gains from using our approach that come from the increase in precision.

3 Formal analysis

We observe n units (i being a generic one) over T periods (t is a generic period). For each unit, we observe an outcome variable Y_{it}, an endogenous policy variable (treatment) W_{it}, an aggregate shock Z_t, and a measure of exposure of unit i to this shock D_i. Our goal is to estimate a causal relationship between Y_{it} and W_{it}. We abstract away from additional unit-specific time-invariant covariates, but they can be incorporated in a straightforward way.

3.1 Causal model

In this section, we present a causal model that we will later use to interpret the output of our algorithm. We view this model as a parsimonious framework that allows us to discuss the central problems researchers face in applications in the simplest possible form. We start with a model of potential outcomes. In addition to w_t (potential value of W_{it}) and z_t (potential value of Z_t), we also introduce h_t – an unobserved aggregate shock that causally affects both the outcome and the treatment variable. We define $w^t := (\dots, w_1, \dots, w_t)$, $z^t := (\dots, z_1, \dots, z_t)$, and $h^t := (\dots, h_1, \dots, h_t)$, and make the following assumption:

Assumption 3.1. (Potential outcomes)
Potential outcomes are generated as follows:

$$\begin{aligned} Y_{it}(w^t, h^t) &= \alpha_{it}^{(y)} + \tau w_t + \theta_i^{(y)} h_t, \\ W_{it}(h^t, z^t) &= \alpha_{it}^{(w)} + \pi_i z_t + \theta_i^{(w)} h_t. \end{aligned} \tag{3.1}$$

As a result, the observed outcomes behave in the following way:

$$\begin{aligned} Y_{it} &= \alpha_{it}^{(y)} + \tau W_{it} + \theta_i^{(y)} H_t, \\ W_{it} &= \alpha_{it}^{(w)} + \pi_i Z_t + \theta_i^{(w)} H_t. \end{aligned} \tag{3.2}$$

The critical part of this assumption and our setup overall is the unobserved aggregate variable H_t. The danger such shocks present for identification is well-recognized in applied work (e.g., Chodorow-Reich et al. [2021]). The typical restriction made in the literature is to assume that all units are affected by unobserved variables in the same way, or, in other words, assume that $\theta_i^{(w)}, \theta_i^{(y)}$ do not vary over i or at least are unrelated to π_i. We do not make this assumption and instead allow for rich heterogeneity in exposures (see also a discussion in Section 4). Following most empirical applications, we focus on contemporaneous treatment effects and assume that only current quantities affect the outcomes. Finally, we assume away heterogeneity in treatment effects mainly to simplify the exposition. As we discuss in Appendix E, our theoretical results can be extended to allow for such heterogeneity, and under additional assumptions, the resulting estimand can be interpreted as a weighted average of individual treatment effects.

Our next assumption describes the relation between the aggregate shocks and the potential outcomes:

Assumption 3.2. (Independence)
Aggregate shocks are independent of potential outcomes:

$$\{Z_t, H_t\}_{t=1}^T \perp\!\!\!\perp \{\alpha_{it}^{(w)}, \alpha_{it}^{(y)}, \theta_i^{(y)}, \theta_i^{(w)}, \pi_i\}_{it}. \tag{3.3}$$

To interpret this restriction we consider two different scenarios where it might hold: applications with exogenous aggregate shocks, and equilibrium models.

Models with exogenous shocks. The most natural case for Assumption 3.2 arises in applications where Z_t, H_t can be plausibly considered exogenous, i.e., determined outside of the relevant model for the unit-level outcomes. Such situations are common in development literature, where these aggregate shocks emerge in developed countries and then directly or indirectly affect the outcomes in the developing countries [Nunn and Qian, 2014].

More generally, in such applications one can interpret Assumption 3.2 as a strict exogeneity assumption (e.g., Arellano [2003]) that is routinely made in the difference-in-differences applications. This assumption implies that one can safely condition on the aggregate variables without being concerned that such variables are affected by past, present or future outcomes.

Equilibrium models. In some applications, e.g., in macroeconomics, Z_t, H_t are determined in general equilibrium and cannot be treated as exogenous. To explain how our framework can fit such situations, we consider stylized examples. First, we put restrictions on the potential outcomes:

$$\begin{aligned} \alpha_{it}^{(y)} &= \breve{\alpha}_{it}^{(y)} + \epsilon_{it}^{(y)}, \\ \alpha_{it}^{(w)} &= \breve{\alpha}_{it}^{(w)} + \epsilon_{it}^{(w)}. \end{aligned} \tag{3.4}$$

Here we treat $\{\breve{\alpha}_{it}^{(y)}, \breve{\alpha}_{it}^{(w)}\}_{it}$ as fixed numbers (condition on them) while the variables $\{\epsilon_{it}^{(y)}, \epsilon_{it}^{(w)}\}_{it}$ are random. We interpret them as measurement errors. Similarly, we treat individual loadings $\{\theta_i^{(w)}, \theta_i^{(y)}, \pi_i\}_{i=1}^n$ as fixed.

In the first example we consider situations where Z_t and H_t are policy variables that are determined in the equilibrium:

$$\begin{aligned} Z_t &= \sum_{i=1}^{n} \phi_i^{(z)} (\breve{\alpha}_{it}^{(w)} + \pi_i Z_t + \theta_i^{(w)} H_t) + \breve{\epsilon}_t^{(z)}, \\ H_t &= \sum_{i=1}^{n} \phi_i^{(h)} (\breve{\alpha}_{it}^{(w)} + \pi_i Z_t + \theta_i^{(w)} H_t) + \breve{\epsilon}_t^{(h)}, \end{aligned} \tag{3.5}$$

where coefficients $\{\phi_i^{(z)}, \phi_i^{(h)}\}_i$ are fixed and aggregate errors $(\breve{\epsilon}_t^{(z)}, \breve{\epsilon}_t^{(h)})$ are random. Solving these equations we express Z_t, H_t in the following way:

$$\begin{aligned} Z_t &= \mu_t^{(z)} + \epsilon_t^{(z)}, \\ H_t &= \mu_t^{(h)} + \epsilon_t^{(h)}, \end{aligned} \tag{3.6}$$

where now $\epsilon_t^{(z)}$ and $\epsilon_t^{(h)}$ are correlated. It is immediate that once we assume that aggregate errors $(\epsilon_t^{(z)}, \epsilon_t^{(h)})$ are independent of $\{\epsilon_{it}^{(w)}, \epsilon_{it}^{(y)}\}_{it}$ Assumption 3.2 holds despite the fact that Z_t, H_t are endogenous.

Our setup also allows for local equilibrium models of the type considered in Guren et al. [2020b]. Let the outcomes and the treatments be determined by the following equations:

$$\begin{aligned} Y_{it} &= \breve{\alpha}_{it}^{(y)} + \tau W_{it} + \theta_i^{(y)} H_t + \epsilon_{it}^{(y)}, \\ W_{it} &= \breve{\alpha}_{it}^{(w)} + \gamma Y_{it} + \theta_i^{(w)} \nu_t + \epsilon_{it}^{(w)}. \end{aligned} \tag{3.7}$$

In the typical example Y_{it} can be the retail employment in location i, period t, while W_{it} is the house price. The aggregate shocks ν_t, H_t are exogenous and unobserved (and possibly correlated). Following Guren et al. [2020b] define two aggregate variables:

$$\begin{aligned} W_t &= \gamma Y_t + \frac{1}{n} \sum_{i=1}^{n} \left(\breve{\alpha}_{it}^{(w)} + \theta_i^{(w)} \nu_t \right), \\ Y_t &= \tau W_t + \frac{1}{n} \sum_{i=1}^{n} \left(\breve{\alpha}_{it}^{(y)} + \theta_i^{(y)} H_t \right). \end{aligned} \tag{3.8}$$

Substituting the value for Y_{it} in the equation for W_{it}, the value for Y_t in the equation for W_t, and rearranging the terms, we get the representation for W_{it} and W_t:

$$\begin{aligned} W_{it} &= \frac{1}{1-\gamma\tau}\left(\check{\alpha}^{(w)}_{it} + \gamma\check{\alpha}^{(w)}_{it}\right) + \frac{\gamma\theta^{(y)}_i}{1-\gamma\tau}H_t + \frac{\theta^{(w)}_i}{1-\gamma\tau}\nu_t + \frac{1}{1-\gamma\tau}\left(\epsilon^{(w)}_{it} + \gamma\epsilon^{(y)}_{it}\right), \\ W_t &= \mu^{(w)}_t + \theta^{(y)}H_t + \theta^{(w)}\nu_t, \end{aligned} \tag{3.9}$$

where $\theta^{(y)}$ and $\theta^{(w)}$ are averages of $\frac{\gamma\theta^{(y)}_i}{1-\gamma\tau}$ and $\frac{\theta^{(w)}_i}{1-\gamma\tau}$ respectively, and $\mu^{(w)}_t$ is an average of $\frac{1}{1-\gamma\tau}\left(\check{\alpha}^{(w)}_{it} + \gamma\check{\alpha}^{(w)}_{it}\right)$. Observe that in this model W_t and H_t are correlated by construction unless $\theta^{(y)}$ is equal to zero and ν_t and H_t are uncorrelated.

Expressing ν_t in terms of H_t and W_t and going back to the original equation for Y_{it} we get a particular version of (3.1):

$$\begin{aligned} Y_{it} &= \check{\alpha}^{(y)}_{it} + \tau W_{it} + \theta^{(y)}_i H_t + \epsilon^{(y)}_{it}, \\ W_{it} &= \tilde{\alpha}^{(w)}_{it} + \pi_i W_t + \tilde{\theta}^{(w)}_i H_t + \frac{1}{1-\gamma\tau}\left(\epsilon^{(w)}_{it} + \gamma\epsilon^{(y)}_{it}\right), \end{aligned} \tag{3.10}$$

where $\tilde{\alpha}^{(w)}_{it} = \frac{1}{1-\gamma\tau}\left(\check{\alpha}^{(w)}_{it} + \gamma\check{\alpha}^{(w)}_{it}\right) - \frac{1}{\theta^{(w)}(1-\gamma\tau)}\theta^{(w)}_i\mu^{(w)}_t$, $\pi_i = \frac{1}{\theta^{(w)}(1-\gamma\tau)}\theta^{(w)}_i$, and $\tilde{\theta}^{(w)}_i = \frac{1}{1-\gamma\tau}(\gamma\theta^{(y)}_i - \frac{\theta^{(y)}}{\theta^{(w)}}\theta^{(w)}_i)$. Crucially, unlike Guren et al. [2020b], we would not assume that π_i and $\theta^{(y)}_i$ are uncorrelated. We return to this example in Section 4.

3.2 Econometric model

This section describes the key statistical restrictions we impose on the causal model from the previous section. In our formal analysis, we derive asymptotic properties as n, T_0, and T_1 go to infinity. Formally, all objects in our analysis are allowed to change with n and T_0, T_1, and thus they should be indexed with n and T_0, T_1. For brevity, we omit these indices.

We start with the model for the baseline outcomes:

Assumption 3.3. (Outcome Model) *Assume that $\{(\theta^{(y)}_i, \theta^{(w)}_i, \pi_i)\}_i$ are deterministic, and $\alpha^{(y)}_{it}, \alpha^{(w)}_{it}$ have the following decomposition:*

$$\begin{aligned} \alpha^{(y)}_{it} &= \beta^{(y)}_i + \mu^{(y)}_t + L^{(y)}_{it} + \epsilon^{(y)}_{it}, \\ \alpha^{(w)}_{it} &= \beta^{(w)}_i + \mu^{(w)}_t + L^{(w)}_{it} + \epsilon^{(w)}_{it}, \end{aligned} \tag{3.11}$$

where $\{(\beta_i^{(y)}, \mu_t^{(y)}, \beta_i^{(w)}, \mu_t^{(w)}, L_{it}^{(y)}, L_{it}^{(w)})\}_{it}$ *are deterministic, and for any* (i,t) *the idiosyncratic shocks are jointly normal:*

$$\begin{pmatrix} \epsilon_{it}^{(w)} \\ \epsilon_{it}^{(y)} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma_w^2 & \rho_{id}\sigma_w\sigma_y \\ \rho_{id}\sigma_w\sigma_y & \sigma_y^2 \end{pmatrix}\right), \tag{3.12}$$

independent over units and periods, and $\max\{\sigma_y, \sigma_w\} < c_\sigma$ *for some constant* $c_\sigma > 0$*.*

This assumption generalizes the conventional two-way fixed effects model, allowing for additional fixed effects captured by $L_{it}^{(y)}$ and $L_{it}^{(w)}$. As a simple example, one can think of interactive fixed effects (e.g., Holtz-Eakin et al. [1988], Chamberlain [1992], Bai [2009], Moon and Weidner [2015, 2017]):

$$L_{it}^{(w)} = a_i^{(w)} b_t^{(w)}, \quad L_{it}^{(y)} = a_i^{(y)} b_t^{(y)}. \tag{3.13}$$

In our analysis, we allow this part to be much more general; in particular, we do not require the rank of corresponding matrices to be fixed and all its singular values to be large. This is practically important, because in applications $(L_{it}^{(w)}, L_{it}^{(y)})$ can have a complicated structure (see Arkhangelsky et al. [2019] for a related discussion).

We impose normality of the errors to simplify exposition; similar results would hold for subgaussian noise under additional technical assumptions. Independence over time might appear at odds with the empirical practice that emphasizes the danger of persistence in errors (e.g., Bertrand et al. [2004]). In our setup, we attribute such persistence to generalized fixed effects $L_{it}^{(w)}, L_{it}^{(y)}$ (condition on it). Our analysis can be generalized to account for finite dependence in idiosyncratic errors.

Our next assumption restricts the distribution of both observed and unobserved time shocks. Since these shocks essentially provide quasi-experimental variation in our setup, we call this assumption a design model.

Assumption 3.4. (Design model) *The aggregate shocks* (Z_t, H_t) *have the following representation for known vectors* $\{\psi_t\}_{t=1}^T$*:*

$$\begin{aligned} Z_t &= \eta_z^\top \psi_t + \epsilon_t^{(z)}, \\ H_t &= \eta_h^\top \psi_t + \epsilon_t^{(h)}, \end{aligned} \tag{3.14}$$

where the first component of each ψ_t is equal to 1*, and* dim$\{\psi_t\} = p < c_p$ *for some constant* $c_p > 0$*. For* $k \in \{z, h\}$ *define* $\epsilon^{(k)} := (\epsilon_T^{(k)}, \dots, \epsilon_1^{(k)})$*; there exist* T*-dimensional vectors* $\nu^{(z)}, \nu^{(h)}$*, two upper-triangular matrices* $\Lambda^{(z)}, \Lambda^{(h)}$ *with non-zero diagonal elements, and* $\rho_{ag} \in (-c_{ag}, c_{ag})$ *for* $c_{ag} < 1$ *such that the following holds:*

$$\begin{aligned}
&\epsilon^{(z)} = \Lambda^{(z)}\nu^{(z)}, \\
&\epsilon^{(h)} = \Lambda^{(h)}(\rho_{ag}\nu^{(z)} + \sqrt{(1-\rho_{ag}^2)}\nu^{(h)}).
\end{aligned} \tag{3.15}$$

Vectors $\nu^{(z)}, \nu^{(h)}$ *are independent, have independent components with uniformly bounded subgaussian norm, and* $\mathbb{E}\left[(\nu_t^{(z)})^2\right] = \mathbb{E}\left[(\nu_t^{(h)})^2\right] = 1$*.*

This assumption describes a rich class of linear time series models. We do not impose stationarity of the errors $\epsilon_t^{(z)}$, allowing coefficients of matrices $\Lambda^{(z)}, \Lambda^{(h)}$ to vary over time in a general way. For example, the variance in each period can be different. In Appendix A.2 we impose additional technical restrictions on matrices $\Lambda^{(z)}, \Lambda^{(h)}$ that exclude very persistent cases (e.g., random walks), but still allow for other forms of non-stationarity. The dependence between $\epsilon_t^{(z)}$ and $\epsilon_t^{(h)}$ arises from the correlation between the underlying shocks. Notably, the correlation coefficient ρ_{ag} is bounded away from one, implying that there is variation in $\epsilon_t^{(h)}$ that is not entirely explained by $\epsilon_t^{(z)}$ (and vice versa). We assume that the means of Z_t and H_t are equal to linear combinations of vectors ψ_t. This is without loss of generality for H_t, because its mean can be treated as a part of $L_{it}^{(w)}$ and $L_{it}^{(y)}$.

Our final assumption describes the size and the complexity of the fixed effects $L_{it}^{(w)}, L_{it}^{(y)}$ and their connection with $\theta_i^{(w)}, \theta_i^{(y)}$ and π_i. To state it we introduce additional notation:

$$\begin{aligned}
&\tilde{L}_{it}^{(y)} := \tau L_{it}^{(w)} + L_{it}^{(y)}, \\
&\tilde{\theta}_i^{(y)} := \tau\theta_i^{(w)} + \theta_i^{(y)}.
\end{aligned} \tag{3.16}$$

Define two $n \times T_0$ matrices $L^{(w),(0)}, L^{(y),(0)}$, such that for $k \in \{y, w\}$ $\left(L^{(k),(0)}\right)_{it} = L_{it}^{(k)}$. The next assumption requires that there exists a vector of weights that approximately eliminates the fixed effects and the unobserved shocks and is correlated with π_i:

Assumption 3.5. (Complexity and size of fixed effects) *There exist* $\breve{\omega}$ *and constants*

$c_{\breve{\omega}}, c_L$ such that as n and T_0 go to infinity the following holds for some $c_{fe} = o(1)$:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\breve{\omega}_i = 0,\ \frac{1}{n}\sum_{i=1}^{n}\breve{\omega}_i\pi_i = 1,\ \|\breve{\omega}\|_2 \le c_{\breve{\omega}}\sqrt{n},\\
&\min_{\eta_\psi^{(y)},\eta_z^{(y)}}\left\{\sum_{t=1}^{T_0}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\breve{\omega}_i(\tilde{L}_{it}^{(y)} + \tilde{\theta}_i^{(y)}\epsilon_t^{(h)}) - \eta_\psi^{(y)}\psi_t - \eta_z^{(y)}\epsilon_t^{(z)}\right]^2\right\} \le c_{fe}^2,\\
&\min_{\eta_\psi^{(w)},\eta_z^{(w)}}\left\{\sum_{t=1}^{T_0}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\breve{\omega}_i(L_{it}^{(w)} + \theta_i^{(w)}\epsilon_t^{(h)}) - \eta_\psi^{(w)}\psi_t - \eta_z^{(w)}\epsilon_t^{(z)}\right]^2\right\} \le c_{fe}^2.
\end{aligned}
\tag{3.17}
$$

Also, the following is satisfied for some fixed constant c_L:

$$
\begin{aligned}
&\max_{i,t}\left|L_{it}^{(w)}\right| \le c_L,\quad \max_{i,t}\left|\tilde{L}_{it}^{(y)}\right| \le c_L,\\
&\max_{i}\left|\theta_i^{(w)}\right| \le c_L,\quad \max_{i}\left|\tilde{\theta}_i^{(y)}\right| \le c_L,
\end{aligned}
\tag{3.18}
$$

and $\max\{\mathit{rank}(L^{(w),(0)}), \mathit{rank}(L^{(y),(0)})\} = o\left(\min\{n, T_0\}\right)$.

This assumptions allows for very general $L_{it}^{(w)}, L_{it}^{(y)}$, in particular it imposes only a mild rank restriction compared with those commonly assumed in the literature (e.g., Bai [2009], Moon and Weidner [2015, 2017]). Alternatively, one can formulate this restriction in terms of approximate rank (number of sufficiently large singular values) to allow for non-degenerate matrices $L^{(w),(0)}, L^{(y),(0)}$. We also require fixed effects to be bounded and restrict their relationship with π_i. To better understand this part, consider a simple situation where $L_{it}^{(w)}$ and $\tilde{L}_{it}^{(y)}$ are described by interactive fixed effects as in (3.13). The first part of the Assumption 3.5 is satisfied as long as π_i is bounded and is not spanned by $(1, \theta_i^{(w)}, \theta^{(y)}, a_i^{(w)}, a_i^{(y)})$, or, in other words, R^2 in the following regression is bounded away from 1:

$$
\pi_i = c_0 + c_{\theta,w}\theta_i^{(w)} + c_{\theta,y}\theta_i^{(y)} + c_{a,y}a_i^{(y)} + c_{a,w}a_i^{(w)} + e_i \tag{3.19}
$$

Indeed, in this case, the normalized residuals from this regression can play the role of $\breve{\omega}_i$ and c_{fe} in (3.17) can be set equal to zero. The second part of the Assumption 3.5 is satisfied as long as fixed effects $(a_i^{(w)}, a_i^{(w)}, b_t^{(w)}, b_t^{(y)})$ are bounded. In more general models it might be infeasible to set c_{fe} in Assumption 3.5 to zero, but one can guarantee that it approaches zero as the size of the data becomes larger.

3.3 Statistical guarantees

We start our analysis by looking at weights $\tilde{\omega}$ that potentially depend on the data from the first part of the sample – periods 1 to T_0 – in a completely general way. For any such weights we can construct the aggregate outcomes and estimate the first stage, and the reduced form coefficients. We demonstrate that the error of such estimator has a deterministic component proportional to the correlation between the weights and exposures to unobserved shocks. This component is unaffected by T_1 – the size of the second part of the sample. The additional random element of the error is decreasing with T_1. We then specialize these results for ω described in Section 2.2.2 and show the gains from using our weights. All the proofs are collected in the Appendix.

Let $\{\tilde{\omega}_i\}_{i=1}^n$ be the sequence of weights such that $\sum_{i=1}^n \tilde{\omega}_i = 0$. For $t \in (T_0, T]$ define the aggregate variables:

$$\begin{aligned} Y_t(\tilde{\omega}) &:= \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i Y_{it}, \\ W_t(\tilde{\omega}) &:= \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i W_{it}. \end{aligned} \tag{3.20}$$

We then estimate coefficients by OLS in the following regressions:

$$\begin{aligned} Y_t(\tilde{\omega}) &= \beta^{(y)}(\tilde{\omega}) + (\eta_\psi^{(y)}(\tilde{\omega}))^\top \psi_t + \delta(\tilde{\omega}) Z_t + \varepsilon_t^{(y)}, \\ W_t(\tilde{\omega}) &= \beta^{(w)}(\tilde{\omega}) + (\eta_\psi^{(w)}(\tilde{\omega}))^\top \psi_t + \pi(\tilde{\omega}) Z_t + \varepsilon_t^{(w)}, \end{aligned} \tag{3.21}$$

and focus our attention on $\hat{\pi}(\tilde{\omega})$ and $\hat{\delta}(\tilde{\omega})$ – the first stage, and the reduced form coefficient, respectively.

To understand our first result, suppose that Y_{it} and W_{it} satisfy the restrictions described in Assumptions 3.1 and 3.3. Then we get the following aggregate series for $t \in (T_0, T]$:

$$\begin{aligned} Y_t(\tilde{\omega}) &= \beta^{(y)}(\tilde{\omega}) + L_t^{(y)}(\tilde{\omega}) + \theta^{(y)}(\tilde{\omega}) H_t + \tau\pi(\tilde{\omega}) Z_t + \epsilon_t^{(y)}(\tilde{\omega}), \\ W_t(\tilde{\omega}) &= \beta^{(w)}(\tilde{\omega}) + L_t^{(w)}(\tilde{\omega}) + \theta^{(w)}(\tilde{\omega}) H_t + \pi(\tilde{\omega}) Z_t + \epsilon_t^{(w)}(\tilde{\omega}). \end{aligned} \tag{3.22}$$

Here aggregate quantities are weighted averages of the corresponding unit-level parameters, their precise definition is given in Appendix A.1.

Since we condition on the first part of the sample, the aggregate trends $L_t^{(y)}(\tilde{\omega})$, $L_t^{(w)}(\tilde{\omega})$ are deterministic and thus should be uncorrelated with innovations in Z_t. As a result, they would not attribute to the bias of the final estimator, only to its variance. The same, however, does not hold for H_t that remains random and by assumption can be correlated with Z_t. This correlation would lead to a bias of the constant size, unaffected by T_1. Finally, we expect the aggregate idiosyncratic errors $\epsilon_t^{(y)}(\tilde{\omega}), \epsilon_t^{(w)}(\tilde{\omega})$ to be small (average of n idiosyncratic shocks).

We formalize this intuition for a simplified class of models described by the next assumption. We impose these restrictions only to simplify the presentation of the results. General case is presented in Appendix (Theorem B.1).

Assumption 3.6. (Autoregressive case) *Assume that* $\psi_t \equiv 1$, $\Lambda^{(z)} = \Lambda^{(h)}$, *and* $\Lambda^{(z)}_{jl} = \{l \geq j\}\rho^{l-j}$, *where* $|\rho| < c < 1$.

This assumption says that the mean of both processes does not change over time and the innovations follow an AR(1) process with autoregresion coefficient ρ. For $k \in \{y, w\}$ define the following quantities:

$$\begin{aligned}
\tilde{L}_t^{(k)}(\tilde{\omega}) &:= \sum_{T_0 < l \leq T} \left(L_l^{(k)}(\tilde{\omega}) - \frac{1}{T_1} \sum_{j > T_0} L_j^{(k)}(\tilde{\omega}) \right) \{l \geq t\}\rho^{l-t} \text{ for } t > T_0, \\
l^{(k)}(\tilde{\omega}) &:= \sqrt{\frac{(1-\rho_{ag}^2)\sum_{t>0}(\tilde{L}_t^{(k)}(\tilde{\omega}))^2}{T_1}}.
\end{aligned} \tag{3.23}$$

Define the correlation coefficient between $\tilde{L}_t^{(y)}(\tilde{\omega}), \tilde{L}_t^{(w)}(\tilde{\omega})$:

$$\rho_l(\tilde{\omega}) := \frac{\sum_{t>0} \tilde{L}_t^{(w)}(\tilde{\omega})\tilde{L}_w^{(k)}(\tilde{\omega})}{l^{(w)}(\tilde{\omega})l^{(y)}(\tilde{\omega})}, \tag{3.24}$$

and two matrices:

$$\begin{aligned}
\Sigma &:= \begin{pmatrix} (\sigma_y^2 + \tau^2\sigma_w^2 + 2\tau\rho_{id}\sigma_y\sigma_w) & \sigma_w(\rho_{id}\sigma_y + \tau\sigma_w) \\ \sigma_w(\rho_{id}\sigma_y + \tau\sigma_w) & \sigma_w^2 \end{pmatrix}, \\
\Sigma_{ag}(\tilde{\omega}) &:= \begin{pmatrix} \left(l^{(y)}(\tilde{\omega})\right)^2 & \rho_l(\tilde{\omega})l^{(y)}(\tilde{\omega})l^{(w)}(\tilde{\omega}) \\ \rho_l(\tilde{\omega})l^{(y)}(\tilde{\omega})l^{(w)}(\tilde{\omega}) & \left(l^{(w)}(\tilde{\omega})\right)^2 \end{pmatrix}.
\end{aligned} \tag{3.25}$$

We now are ready to state our first formal result.

Theorem 1. (Arbitrary weights) *Suppose Assumptions 3.1,3.2,3.3,3.4,3.6 hold; Let the* *weights $\{\tilde{\omega}_i\}_{i=1}^n$ be such that $\frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i = 0$, and for $k \in \{y, w\}$ $\max_t \left\{|\tilde{L}_t^{(k)}(\tilde{\omega})|\right\} = o_p\left(l^{(k)}(\tilde{\omega})\right)$. Then as n and T_1 approach infinity we have the following result:*

$$\begin{pmatrix} \hat{\delta}(\tilde{\omega}) - \frac{\tau}{n}\sum_{i=1}^n \tilde{\omega}_i \pi_i \\ \hat{\pi}(\tilde{\omega}) - \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \pi_i \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \tilde{\theta}_i^{(y)} \\ \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \theta_i^{(w)} \end{pmatrix} \left(\rho_{ag} + \mathcal{O}_p\left(\frac{1}{\sqrt{T_1}}\right)\right) + \\ \sqrt{\frac{1-\rho^2}{T_1}} \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega})(\xi_z + o_p(1)) + \frac{\|\tilde{\omega}\|_2 \sqrt{1-\rho^2}}{n\sqrt{T_1}}(\xi_{cr} + o_p(1)), \quad (3.26)$$

where $\xi_{cr} \sim \mathcal{N}(0, \Sigma)$, ξ_z is independent of ξ_{cr}, $\mathbb{E}[\xi_{cr}] = 0, \mathbb{V}[\xi_z] = \mathcal{I}_2$, and it converges in distribution to a standard normal vector.

This result applies to any weights including those that depend on the first part of the dataset as long as they average to zero and satisfy condition $\max_t \left\{|\tilde{L}_t^{(k)}(\tilde{\omega})|\right\} = o_p\left(l^{(k)}(\tilde{\omega})\right)$ for $k \in \{y, w\}$. The latter simply requires the aggregate trends $\tilde{L}_t^{(y)}(\tilde{\omega}), \tilde{L}_t^{(y)}(\tilde{\omega})$ to be sufficiently spread-out over time so that no single period dominates. For any such weights, the theorem states that the estimation error has three components:

$$\begin{aligned} &\text{bias} := \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \tilde{\theta}_i^{(y)} \\ \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \theta_i^{(w)} \end{pmatrix} \left(\rho_{ag} + \mathcal{O}_p\left(\frac{1}{\sqrt{T_1}}\right)\right), \\ &\text{aggregate noise} := \sqrt{\frac{1-\rho^2}{T_1}} \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega})(\xi_z + o_p(1)), \\ &\text{cross-sectional noise} := \frac{\|\tilde{\omega}\|_2 \sqrt{1-\rho^2}}{n\sqrt{T_1}}(\xi_{cr} + o_p(1)). \end{aligned} \quad (3.27)$$

The bias is proportional to ρ_{ag} and the covariance between the weights and exposures to unobserved shocks. The latter can be of constant order, unless $\tilde{\omega}_i$ are independent of $\theta_i^{(w)}, \theta_i^{(y)}$ by design. Thus the bias does not go away unless ρ_{ag} is small (converges to zero). Similarly, we can expect $l^{(y)}(\tilde{\omega})$ and $l^{(w)}(\tilde{\omega})$ to be of constant order making the aggregate noise behave as $\frac{1}{\sqrt{T_1}}$. In this case, the cross-sectional noise is dominated by the aggregate one for any bounded weighs $\tilde{\omega}$. This is natural because the fundamental exogenous variation comes from the time-series dimension. Notably, the resulting variance directly depends on the properties of the aggregate shocks.

With the weights that depend on the first part of the data in a systematic way, we can hope to reduce the bias. To do so, we need to find the weights that "balance out" $\theta_i^{(w)}, \theta_i^{(y)}$. Assumptions 3.1, 3.5 suggest that this might be possible – the exposures to H_t do not change over time and the initial periods are informative about them. Our second formal result shows that this is indeed possible, and describes the behavior of the estimator for the weights proposed in Section 2.2.2. To state it we need to connect the variables $\{D_i\}_{i=1}^n$ that we use to construct ω to unobserved exposures π_i. We make the following assumption:

Assumption 3.7. (Proportional exposures) *There exist numbers (η_0, η_π) such that $\eta_\pi \neq 0$, and for every i we have that $D_i = \eta_0 + \eta_\pi \pi_i$.*

This restriction is motivated by the empirical work where it is commonly imposed [Nunn and Qian, 2014]. We make this assumption to simplify exposition and consider a more general version in Appendix (Theorem B.2). We also discuss data-dependent D_i in Section 4.

Theorem 2. (Systematic weights) *Suppose Assumptions 3.1,3.2,3.3,3.4,3.5,3.6,3.7 hold, and $\max_t \left\{|\tilde{L}_t^{(k)}(\omega)|\right\} = o_p\left(l^{(k)}(\omega)\right)$; additionally, suppose that as T_0, T_1, and n approach infinity we have $\frac{T_0}{n} = c_{asp} + o(1)$ for $c_{asp} \in (0,1)$, and $\zeta = c \max\left\{\max\{\sigma_w, \sigma_y, c_{fe}\}, \frac{\max\{\sigma_w^2, \sigma_y^2, c_{fe}^2\}}{\sqrt{c_{asp}}}\right\}$. Then the following holds:*

$$\begin{pmatrix} \hat{\delta}(\omega) - \frac{\tau}{\eta_\pi} \\ \hat{\pi}(\omega) - \frac{1}{\eta_\pi} \end{pmatrix} = \frac{\xi_{bias}}{\sqrt{T_0}}\left(\rho_{ag} + \mathcal{O}_p\left(\frac{1}{\sqrt{T_1}}\right)\right) + \sqrt{\frac{1-\rho^2}{T_1}\Sigma^{\frac{1}{2}}(\omega)(\xi_z + o_p(1))} + \frac{\|\omega\|_2\sqrt{1-\rho^2}}{n\sqrt{T_1}}(\xi_{cr} + o_p(1)), \quad (3.28)$$

where ξ_{cr}, ξ_z are the same as in Theorem 1, and ξ_{bias} is a tight two-dimensional random vector independent of ξ_{cr}, ξ_z.

The first implication of this result is that with the weights ω, the estimator is consistent as long as n, T_0, T_1 go to infinity, and $n \sim T_0$. The restriction for n and T_0 is natural: we are looking for n different weights ω_i that are only required to satisfy two restrictions. Each $t \in [1, T_0]$ provides additional information and it is intuitive that we require T_0 that is similar in size to n to find reasonable weights. In Section 4 we discuss practical means of reducing this requirement.

There are two differences between Theorems 1 and 2. The first one is in the behavior of the bias. With the weights ω the estimator is not only consistent, but also has the bias of the order

Algorithm 2: Estimation of variance

```
Data: $\{Y_{it}, W_{it}\}_{it}, \{\omega_i\}_{i=1}^n, \{Z_t, \psi_t\}_{t=1}^T, \hat{\Lambda}^{(z),(1)}, T_0$
Result: Variance estimate $\hat{\Sigma}(\omega)$
for $t \leftarrow T_0 + 1$ to $T$ do
    Construct $Y_t = \frac{1}{n}\sum_{i=1}^n Y_{it}\omega_i$, and $W_t = \frac{1}{n}\sum_{i=1}^n W_{it}\omega_i$.
end
Construct OLS residuals $\{\hat{\epsilon}_t^{(y)}, \hat{\epsilon}_t^{(w)}\}_{t=T_0+1}^T$ in the following regressions (for $t > T_0$):

    $Y_t = \beta^{(y)} + (\eta_\psi^{(y)})^\top \psi_t + \delta Z_t + \varepsilon_t^{(y)},$
    $W_t = \beta^{(w)} + (\eta_\psi^{(w)})^\top \psi_t + \pi Z_t + \varepsilon_t^{(w)},$
    $Z_t = \eta_0 + (\eta_\psi^{(z)})^\top \psi_t + \varepsilon_t^{(z)},$

  and for $k \in \{y, w, z\}$ define $\hat{\varepsilon}^{(k)} := (\hat{\varepsilon}_T^{(k)}, \dots, \hat{\varepsilon}_{T_0+1}^{(k)})$;
Compute and report $\hat{\Sigma}(\omega)$:

    $\hat{\Sigma}(\omega) := \begin{pmatrix} \frac{\|\hat{\varepsilon}^{(y)}\hat{\Lambda}^{(z),(1)}\|_2^2}{\|\hat{\varepsilon}^{(z)}\|_2^4} & \frac{\hat{\varepsilon}^{(y)}\hat{\Lambda}^{(z),(1)}(\hat{\Lambda}^{(z),(1)})^\top(\varepsilon^{(w)})^\top}{\|\hat{\varepsilon}^{(z)}\|_2^4} \\ \frac{\hat{\varepsilon}^{(y)}\hat{\Lambda}^{(z),(1)}(\hat{\Lambda}^{(z),(1)})^\top(\varepsilon^{(w)})^\top}{\|\hat{\varepsilon}^{(z)}\|_2^4} & \frac{\|\hat{\varepsilon}^{(w)}\hat{\Lambda}^{(z),(1)}\|_2^2}{\|\hat{\varepsilon}^{(z)}\|_2^4} \end{pmatrix}$
```

$\mathcal{O}_p\left(\frac{1}{\sqrt{T_0}}\right)$. If $T_0 \sim T_1$, this implies that the estimator is asymptotically normal, albeit biased. The second difference is that both the first stage and the reduced form estimands are well-defined deterministic objects that depend on the relationship between D_i and π_i (Assumption 3.7). We view this result as the main theoretical justification for using the algorithm from Section 2.2 instead of the conventional TSLS regression.

To put Theorem 2 in context, it is useful to benchmark it against the ideal situation where π_i is completely random, and thus is uncorrelated with any of the fixed effects. In this case, if we use weights $\tilde{\omega}$ that are proportional to $\pi_i - \overline{\pi}$ the resulting covariances $\frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \theta_i^{(w)}, \frac{1}{n}\sum_{i=1}^n \tilde{\omega}_i \tilde{\theta}_i^{(y)}$ are of the order $\frac{1}{\sqrt{n}}$ and have zero mean. Theorem 2 delivers the same order (but not zero mean), at the expense of using the first T_0 periods to find the weights. The random weights also imply that the variance coming from $L_{it}^{(w)}, \tilde{L}_{it}^{(y)}$ is of the order $\frac{1}{\sqrt{nT}}$. Our construction cannot guarantee that: the restrictions that we impose on $L_{it}^{(w)}, \tilde{L}_{it}^{(y)}$ in Assumption 3.5 do not say anything about periods beyond T_0.

3.4 Inference

Theorem 2 cannot be immediately used for inference because we do not know the distribution of ξ_{bias} and do not have an estimator for ρ_{ag}. A standard theoretical tool to avoid this problem, is to assume that $\frac{T_1}{T_0} = o(1)$. In this case, the bias is dominated by the variance that under certain assumptions can be estimated leading to asymptotically valid inference. However, such "undersmoothing" technique provides little guidance for empirical research.

In practice, we recommend setting $T_1 \sim T_0$, estimating the variance, and using normal approximation conditional on ω to conduct the inference. Variance can be computed in multiple ways, Algorithm 2 provides a particular realization. It uses an estimated matrix $\hat{\Lambda}^{(z)}$ as an input, in particular its $T_1 \times T$ submatrix $\hat{\Lambda}^{(z),(1)}$ that describes the distribution of $(\epsilon^{(z)}_T, \dots, \epsilon^{(z)}_{T_0+1})$. Let $\hat{\Sigma}(\omega)$ be the resulting variance matrix. We suggest that users conduct inference using the following approximation under $\mathbf{H}_0 : \tau = \tau_0$:

$$(\hat{\delta} - \tau_0 \hat{\pi}) \approx \mathcal{N}\left(0, (1, -\tau_0)\hat{\Sigma}(\omega)(1, -\tau_0)^\top\right). \tag{3.29}$$

Practically this means that τ_0 is rejected at level α by the following decision rule:

$$\{\tau_0 \text{ is rejected}\} = \left\{\left|\hat{\delta} - \tau_0 \hat{\pi}\right| \geq \sqrt{(1, -\tau_0)\hat{\Sigma}(\omega)(1, -\tau_0)^\top} z_{1-\alpha/2}\right\}, \tag{3.30}$$

where z_α is a α-quantile of the standard normal distribution. The confidence set can be constructed by collecting all values of τ_0 that are not rejected. This "Anderson-Rubin"-type construction is robust to small first stage coefficients (see Andrews et al. [2019] for a recent survey).

Using (3.30) for inference is natural in our context, in fact, Borusyak and Hull [2020] recommend a similar, albeit non-asymptotic, procedure for a general class of causal problems with exogenous shocks (see also references therein). Theorem 1 can be used to show that such inference is valid for arbitrary weights $\tilde{\omega}$ as long as $\rho_{ag} = o\left(\frac{1}{\sqrt{T_1}}\right)$. This requirement reduces to a much weaker condition $\rho_{ag} = o(1)$ if instead of generic weights researchers use ω and Theorem 2. Practically, this means that with our weights the inference based on (3.30) is accurate if the correlation between H_t and Z_t is small.

Condition $\rho_{ag} = o(1)$ is strong because it excludes the practically relevant case of strong (but not perfect) correlation between the aggregate shocks. If $T_0 \sim T_1$ Theorem 2 is not helpful in this case. To deal with such situations we impose an additional assumption that the idiosyncratic errors $\epsilon_{it}^{(w)}, \epsilon_{it}^{(y)}$ are small. Similar assumption has been used extensively in the non-linear measurement error literature (e.g., Chesher [1991], Evdokimov and Zeleneev [2016], see also Schennach [2016]). It might be surprising that such a condition is helpful because in Theorem 2 the error from the cross-sectional noise is dominated by the aggregate variation. However, when the variance of $\epsilon_{it}^{(w)}, \epsilon_{it}^{(y)}$ is small we can construct better weights ω_i. Our next theorem provides formal guarantees in the low-noise regime (for a general version see Theorems B.2, B.3):

Theorem 3. (Inference) *Suppose conditions of Theorem 2 hold; in addition, suppose that the following restrictions are satisfied for some constants $c_l, c_{\rho_l}, c_{int} \in (0,1)$ as n, T_0, T_1 approach infinity:*

$$\begin{aligned} &\max\{\sigma_y, \sigma_w\} = o(1), \quad \frac{T_0}{T_1} = c_{int} + o(1), \\ &\min\{l^{(y)}(\omega), l^{(w)}(\omega)\} > c_l > 0, \quad |\rho_l(\omega)| < c_{\rho_l} < 1. \end{aligned} \tag{3.31}$$

Then the following holds:

$$\begin{pmatrix} \hat{\delta}(\omega) - \frac{\tau}{\eta_\pi} \\ \hat{\pi}(\omega) - \frac{1}{\eta_\pi} \end{pmatrix} = \sqrt{\frac{1-\rho^2}{T_1}} \Sigma^{\frac{1}{2}}(\omega)(\xi_z + o_p(1)) \tag{3.32}$$

where ξ_z is the same as in Theorem 1 and converges in distribution to a standard two-dimensional normal random vector. In addition, suppose that $\hat{\Lambda}^{(z)}$ is constructed using an estimator $\hat{\rho} = \rho + o_p(1)$. Then the test described in (3.30) is consistent.

This result provides an alternative to Theorem 2 and justifies the conventional inference based on (3.30) in situations where the correlation between the unobserved shocks is strong. This comes at a price: we need to believe that the variance of the idiosyncratic shocks is small. Recall that the variables $\epsilon_{it}^{(w)}, \epsilon_{it}^{(y)}$ represent measurement errors, and thus the small-variance regime can be interpreted as saying that the outcomes are measured well. This assumption is natural if the unit-level outcomes themselves are aggregates (e.g., averages of the individual-level data). We believe that in such situations Theorem 3 can be practically useful.

We also investigate the properties of the test (3.30) using the simulations of Section 2.3. Results are summarized in Table 2 and show the rejection rates for the four different designs described in Section 2.3. We see that while the test based on $(\hat{\delta}, \hat{\pi})$ is not perfect, its size

	Design 1	Design 2	Design 3	Design 4
Size	0.013	0.008	0.045	0.081

Table 2: Rejection rates for $\mathbf{H}_0 : \tau = 1.43$ using (3.30) with a nominal size 0.05, for the simulation designs described in Section 2.3. Results are based on 2000 simulations.

distortions are relatively small. Reasonable performance of our estimator in the last design might be attributed to the fact that in this case the variance of the idiosyncratic noise is much smaller than the size of the fixed effects.

4 Discussion

4.1 Constructed exposures

In applications D_i often is not an observed fixed characteristic of a unit, but rather a data-dependent quantity that is constructed to approximate π_i. For each unit we can construct such proxy by running a regression of W_{it} on Z_t and ψ_t separately for every i using the first part of the data ($t \in [1, T_0]$):

$$W_{it} = \alpha_i + \eta_i^\top \psi_t + \pi_i Z_t + \varepsilon_{it}. \tag{4.1}$$

Let $\hat{\pi}_i$ be the OLS estimator in this regression. Researchers frequently use $D_i = \hat{\pi}_i$ together with the conventional TSLS weights described in Section 2.1. For example, Nakamura and Steinsson [2014] do exactly that, albeit estimating $\hat{\pi}_i$ using the data from all periods. In Nunn and Qian [2014] the authors use a similar algorithm, but instead of running the regression (4.1) they compute the average W_{it} over the first T_0 periods.

Under Assumptions 3.3 and 3.4 $\hat{\pi}_i$ has the following representation:

$$\hat{\pi}_i = \pi_i + \hat{\eta}_{H|Z}\theta_i^{(w)} + u_i, \tag{4.2}$$

where u_i is a mean-zero error (correlated across i). This immediately shows the potential problem of using the $D_i = \hat{\pi}_i$ together with the conventional weights. If the unobserved aggregate shocks are present and have heterogeneous exposures, then $\hat{\pi}_i$ is generically correlated with them. At the same time, using $\hat{\pi}_i$ together with our weights ω is completely natural, because they essentially balance $\theta_i^{(w)}$ away. We do not provide formal results for this case, but we follow this strategy in the simulation described in Section 2.3 and it performs well.

In applications, researchers can go beyond (4.1) and use more elaborate procedures. For example, in Duflo and Pande [2007] the authors project $\hat{\pi}_i$ on the set of available unit characteristics X_i and argue that the resulting D_i are as good as randomly assigned. The validity of this approach depends on the nature of X_i. In Guren et al. [2020b] the authors propose an alternative procedure for constructing the weights in the context of the local equilibrium model described in Section 3. In particular, they suggest estimation of the panel regression by OLS treating $\{(\alpha_i, \pi_i)\}_{i=1}^n$ as fixed effects (using the notation from (3.8)):

$$W_{it} = \alpha_i + \mu_t + \gamma Y_{it} - \pi_i W_t - \gamma \pi_i Y_t + \epsilon_{it}^{(w)}, \tag{4.3}$$

and using $\hat{\pi}_i$ to construct the instrument $\hat{\pi}_i W_t$. Under additional assumptions $\hat{\pi}_i$ converges to π_i as defined in (3.10). The authors emphasize that for this procedure to produce a valid estimator, $\pi_i W_t$ should be uncorrelated with the unobserved term $\theta_i^{(y)} H_t$ in the outcome equation. This does not hold in our setup, because H_t and W_t are correlated by construction, and $\theta_i^{(y)}$ is allowed to be correlated with π_i. Combination of this procedure with Algorithm 1 is promising and we leave its formal analysis to future research.

4.2 Time heterogeneity in exposures

One of the restrictions of the causal model described by Assumption 3.1 is that both Z_t and H_t affect outcomes in a time-invariant way. Formally this means that π_i and $\theta_i^{(w)}, \theta_i^{(y)}$ do not vary over t. The statistical analysis of Section 3 relies on this assumption in an important way – it guarantees that if we find the weights that eliminate $\theta_i^{(w)}, \theta_i^{(y)}$ using the first part of the data, these weights "work" for the second part of the data. As a result, it cannot be completely eliminated, but it can be relaxed. To understand why this is possible, consider a generalization of the time-invariant exposures:

$$\begin{aligned}
&\pi_{it} = \pi_i + \gamma_i^{(\pi)} \phi_t^{(\pi)}, \\
&\theta_{it}^{(w)} = \theta_i^{(w)} + \gamma_i^{(w)} \phi_t^{(w)}, \\
&\theta_{it}^{(y)} = \theta_i^{(y)} + \gamma_i^{(w)} \phi_t^{(w)}.
\end{aligned} \tag{4.4}$$

One can transform this setup to the one described by Assumption 3.1 at the expense of the increased dimension of H_t. While our formal results are derived for one-dimensional H_t, they can be adapted to the multi-dimensional case if the dimension is modest. We expect that as long as $\phi_t^{(k)}$ do not concentrate on particular periods (e.g., second part of the data) the conclusions of Theorems 2,3 would hold.

4.3 Prior knowledge

Algorithm 1 can be extended to accommodate additional knowledge that researchers might have in a given application. It can be done by introducing additional constraints to the optimization problem. A natural constraint can be defined in terms of covariates. For example, if we believe that $\theta_i^{(w)}$ or $\theta_i^{(y)}$ are correlated with observed characteristics X_i, then we can incorporate the following constraint:

$$\frac{1}{n}\sum_{i=1}^{n} w_i X_i = 0. \tag{4.5}$$

Depending on the application researchers might want to control the sign of ω_i by imposing the following constraint for all units i:

$$\omega_i(D_i - \overline{D}) \geq 0 \tag{4.6}$$

This is similar to the standard non-negativity constraint used in the synthetic control algorithm. In Appendix E we discuss how this constraint can help in applications with heterogeneous treatment effects. More generally, any prior information about the complexity and structure of weights can be incorporated into our algorithm. As long as the resulting problem is convex, it can be solved efficiently, delivering alternative unit weights.

There are different possible benefits from introducing restrictions, and in general, the estimator's behavior depends on the nature of the constraints. The key part that is affected is the number of unobservables n and periods T_0 we can allow. In Theorem 2 we require $n \sim T_0$, which can be demanding in applications. If additional, informative constraints hold, T_0 can be much smaller. Precise results of this nature can be derived using the general bound introduced in Hirshberg [2021].

4.4 Shift-share Designs

In this section, we discuss the relationship between our model and models from the shift-share, or "Bartik" instruments, literature (Adao et al., 2019, Borusyak et al., 2018, Goldsmith-Pinkham et al., 2020). We start by considering an extension of our original framework. Assume that instead of a single aggregate shock, we have $|S|$ of them. In a typical application, these will correspond to industry-level shocks. Potential outcomes are now determined by the following equations:

$$\begin{aligned} Y_{it} &= \alpha_{it}^{(y)} + \tau W_{it} + \sum_{s \in S} \theta_{its}^{(y)} H_{ts}, \\ W_{it} &= \alpha_{it}^{(w)} + \sum_{s \in S} \pi_{its} \gamma_{its} Z_{ts} + \sum_{s \in S} \theta_{its}^{(w)} H_{ts}, \end{aligned} \tag{4.7}$$

where s is a generic industry, and we observe $\{\gamma_{its}\}_{i,t,s}$, $\{W_{it}, Y_{it}\}_{it}$, $\{Z_{ts}\}_{t,s}$, and $\sum_i \gamma_{its} = 1$. It is straightforward to see that our model is a special case of this with $|S| = 1$.

The model considered in the shift-share literature is a special case of (4.7) with $T = 1$, and two additional assumptions: (a) for every s, $\{H_{ts}\}_{s \in S}$ is uncorrelated with $\{Z_{ts}\}_{s \in S}$, and (b) $\mathbb{E}[Z_{ts}] = \mu$ and Z_{ts} are uncorrelated over s. Identification is now achieved exploiting variation over industries (see Borusyak et al., 2018). In applications, T is usually not equal to 1, and often the model in differences is considered. At the same time, the identification argument does not exploit the time dimension and focuses on the variation over industries.

One can immediately see that these two models are non-nested, both formally and conceptually: we are focusing on the case, with a single aggregate shock, motivated by the applications in development and macroeconomics. In these applications, correlation between observed and unobserved aggregate shocks is the key problem one has to deal with to make causal claims. Shift-share literature, on the other hand, focuses on the case where the main source of endogeneity is the cross-sectional correlation between $\alpha_{it}^{(y)}$ and $\alpha_{it}^{(w)}$ that typically arises because of simultaneity issues (e.g., when Y_{it} is wage and W_{it} is a labor supply).

We believe that models of the type (4.7) can be promising, because they allow for a combination of two identification arguments: one that is based on the variation over time, and one that is based on the variation over s. In applications, both $|S|$ and T can be modest (especially, if we want shocks to be independent over s), and thus it is natural to use both sources of variation. Also, using a time-series dimension, one can estimate correlations between Z_s and adapt inference to this case.

5 Conclusion

Aggregate shocks provide a natural source of exogenous variation for unit-level outcomes. As a result, they are frequently used to evaluate local-level policies. We argue that this exercise has two conceptual steps: aggregation of unit-level data into a time series and analysis of the aggregated data. We propose a new algorithm for constructing unit weights that are then used to produce aggregate outcomes. In a rich statistical model, we show that our weights approximately eliminate potential unobserved aggregate shocks, leading to a consistent and asymptotically normal estimator. After aggregation, we suggest that researchers use OLS regressions to estimate first stage and reduced-form coefficients. Importantly these regressions should include other variables that capture the underlying trend in the aggregate instrument. We illustrate the performance of the resulting estimator in data-driven simulations that demonstrate its superiority to the conventional algorithm in a variety of practically relevant situations. We also provide conditions under which one can use design-based techniques to conduct valid inference.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. American Economic Review, 93(-):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490):493–505, 2010.

Rodrigo Adao, Michal Kolesár, and Eduardo Morales. Shift-share designs: Theory and inference. The Quarterly Journal of Economics, 134(4):1949–2010, 2019.

Isaiah Andrews, James H Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. Annual Review of Economics, 11:727–753, 2019.

Manuel Arellano. Panel data econometrics. Oxford university press, 2003.

Dmitry Arkhangelsky and Guido W Imbens. Double-robust identification for causal panel data models. Technical report, National Bureau of Economic Research, 2021.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Orley C Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs, 1984.

David H Autor, David Dorn, Gordon H Hanson, and Jae Song. Trade adjustment: Worker-level evidence. The Quarterly Journal of Economics, 129(4):1799–1860, 2014.

Jushan Bai. Panel data models with interactive fixed effects. Econometrica, 77(4):1229–1279, 2009.

Kyle Barron, Edward Kung, and Davide Proserpio. The effect of home-sharing on house prices and rents: Evidence from airbnb. Marketing Science, 40(1):23–47, 2021.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. arXiv preprint arXiv:1811.04170, 2018.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? The Quarterly journal of economics, 119(1):249–275, 2004.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research, 2020.

Kirill Borusyak, Peter Hull, and Xavier Jaravel. Quasi-experimental shift-share research designs. Technical report, National Bureau of Economic Research, 2018.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Technical report, National Bureau of Economic Research, 1993.

Gary Chamberlain. Efficiency bounds for semiparametric regression. Econometrica: Journal of the Econometric Society, pages 567–596, 1992.

Andrew Chesher. The effect of measurement error. Biometrika, 78(3):451–462, 1991.

Gabriel Chodorow-Reich, Plamen T Nenov, and Alp Simsek. Stock market wealth and the real economy: A local labor market approach. American Economic Review, 2021.

Marco Del Negro. Aggregate risk sharing across us states and across european countries. Yale University, 1998.

Christian Dippel, Avner Greif, and Daniel Trefler. Outside options, coercion, and wages: Removing the sugar coating. The Economic Journal, 130(630):1678–1714, 2020.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

Oeindrila Dube and Juan F Vargas. Commodity price shocks and civil conflict: Evidence from colombia. The Review of Economic Studies, 80(4):1384–1421, 2013.

Esther Duflo and Rohini Pande. Dams. The Quarterly Journal of Economics, 122(2):601–646, 2007.

KS Evdokimov and Andrei Zeleneev. Simple estimation of semiparametric models with measurement errors. Technical report, Working Paper, Princeton University, 2016.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. American Economic Review, 110(8):2586–2624, 2020.

Adam Guren, Alisdair McKay, Emi Nakamura, and Jón Steinsson. What do we learn from cross-regional empirical estimates in macroeconomics? Technical report, National Bureau of Economic Research, 2020a.

Adam M Guren, Alisdair McKay, Emi Nakamura, and Jón Steinsson. Housing wealth effects: The long view. The Review of Economic Studies, 2020b.

James D Hamilton. Why you should never use the hodrick-prescott filter. Review of Economics and Statistics, 100(5):831–843, 2018.

David A Hirshberg. Least squares with error in variables. Technical report, Stanford University, 2021.

Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen. Estimating vector autoregressions with panel data. Econometrica: Journal of the econometric society, pages 1371–1395, 1988.

Hyungsik Roger Moon and Martin Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica, 83(4):1543–1579, 2015.

Hyungsik Roger Moon and Martin Weidner. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory, 33(1):158–195, 2017.

Emi Nakamura and Jon Steinsson. Fiscal stimulus in a monetary union: Evidence from us regions. American Economic Review, 104(3):753–92, 2014.

Nathan Nunn and Nancy Qian. Us food aid and civil conflict. American Economic Review, 104 (6):1630–66, 2014.

Susanne M Schennach. Recent advances in the measurement error literature. Annual Review of Economics, 8:341–377, 2016.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018.

Appendices

A Preparations

A.1 Notation and definitions

We use $\|\cdot\|_2$ to denote euclidean norm, $\|\cdot\|_{HS}$ to denote Hilbert-Schmidt norm, and $\|\cdot\|_{op}$ – the operator norm. For deterministic sequences we say $x_n \sim y_n$ if $\lim_n \frac{x_n}{y_n}$ exists and is not equal to 0 or infinity. The same applies for random sequences that converge in probability to a deterministic limit.

We use superscript (0) and (1) to distinguish data that belongs to periods $[1, T_0]$ and $(T_0+1, T]$, respectively. For example, $Y^{(0)}$ corresponds to $n \times T_0$ matrix of outcomes from periods $[1, T_0]$ and $L^{(w),(1)}$ corresponds to a sub-matrix with last $(T_0, T]$ columns of matrix $L^{(w)}$. This convention works for any $n \times T$ matrix. For a $T \times T$ matrices $\Lambda^{(z)}, \Lambda^{(h)}$ we use $\Lambda^{(k),(0)}$ and $\Lambda^{(k),(1)}$ $(k \in \{z, h\})$ to denote sub-matrices with rows that correspond to $[1, T_0]$ and $(T_0, T]$, respectively. We also separate each matrix $\Lambda^{(k),(j)}$ into two parts: $\Lambda_0^{(k),(j)}$ and $\Lambda_1^{(k),(j)}$ such that $\Lambda^{(k),(j)}\nu^{(k)} = \Lambda_0^{(k),(j)}\nu^{(k),(0)} + \Lambda_1^{(k),(j)}\nu^{(k),(1)}$.

$$\tilde{\sigma}_y^2 := \sigma_y^2 + \tau\sigma_w^2 + 2\tau\rho_{id}\sigma_y\sigma_w, \tag{1.1}$$

Define the following projection matrices:

$$\begin{aligned}
\Pi_f^{(k)} &= \mathcal{I} - \psi^{(k)}((\psi^{(k)})^\top \psi^{(k)})^{-1}(\psi^{(k)})^\top \\
\Pi_r^{(0)} &= \mathcal{I} - \epsilon^{(z),(0)}(\epsilon^{(z),(0)})^\top \epsilon^{(z),(0)})^{-1}(\epsilon^{(z),(0)})^\top \\
\Pi^{(0)} &= \Pi_f^{(0)}\Pi_r^{(0)} \\
\Pi_\alpha &= \mathcal{I}_n - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^\top
\end{aligned} \tag{1.2}$$

For $k \in \{0, 1\}$ define the regression and correlation coefficients:

$$\begin{aligned}
\eta_{H|Z}^{(k)} &:= \frac{\rho_{ag}\text{trace}\left(\left(\Lambda^{(h),(k)}\right)^\top \Pi_f^{(k)}\Lambda^{(z),(k)}\right)}{\|\Pi_f^{(k)}\Lambda^{(z),(k)}\|_{HS}^2} \\
\rho_{H|Z}^{(k)} &= \frac{\rho_{ag}\text{trace}\left(\left(\Lambda^{(h),(k)}\right)^\top \Pi_f^{(k)}\Lambda^{(z),(k)}\right)}{\|\Pi_f^{(k)}\Lambda^{(z),(k)}\|_{HS}\|\Pi_f^{(k)}\Lambda^{(h),(k)}\|_{HS}}
\end{aligned} \tag{1.3}$$

Define the following symmetric matrix that later plays a crucial role for the analysis of the bias:

$$\begin{aligned}\Gamma := \frac{1}{\hat{\sigma}_W^2} L^{(w),(0)} \Pi_f^{(0)} (L^{(w),(0)})^\top + \frac{1}{\hat{\sigma}_Y^2} \tilde{L}^{(y),(0)} \Pi_f^{(0)} (\tilde{L}^{(y),(0)})^\top + \\ \frac{(1-(\rho_{H|Z}^{(0)})^2)\|\Lambda^{(h),(0)}\|_{HS}^2}{\hat{\sigma}_W^2} (\theta^{(w)})(\theta^{(w)})^\top + \frac{(1-(\rho_{H|Z}^{(0)})^2)\|\Lambda^{(h),(0)}\|_{HS}^2}{\hat{\sigma}_Y^2} (\tilde{\theta}^{(y)})(\tilde{\theta}^{(y)})^\top.\end{aligned} \tag{1.4}$$

Define $D - \overline{D} \in \mathbb{R}^n$ such that $(D - \overline{D})_i = D_i - \frac{1}{n}\sum_{i=1}^n D_i$ and similarly define $\pi - \overline{\pi}$. For $\zeta > 0$ define the following quantity that measures the correlation between D_i and π_i:

$$s(\zeta) := \left| \frac{(D - \overline{D})^\top \left(\zeta^2 \mathcal{I}_n + \frac{1}{T_0}\Pi_\alpha \Gamma \Pi_\alpha\right)^{-1} (\pi - \overline{\pi})}{(D - \overline{D})^\top \left(\zeta^2 \mathcal{I}_n + \frac{1}{T_0}\Pi_\alpha \Gamma \Pi_\alpha\right)^{-1} (D - \overline{D})} \right| \tag{1.5}$$

Finally, for arbitrary weights $\tilde{\omega}$ define the following objects:

$$\begin{aligned}
&\beta^{(y)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n (\beta_i^{(y)} + \tau\beta_i^{(w)})\tilde{\omega}_i, && \beta^{(w)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n \beta_i^{(w)}\tilde{\omega}_i, && \theta^{(y)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n (\theta_i^{(y)} + \tau\theta_i^{(w)})\tilde{\omega}_i, \\
&\theta^{(w)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n \beta_i^{(w)}\tilde{\omega}_i, && \pi^{(w)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n \pi_i^{(w)}\tilde{\omega}_i, && \\
&L_t^{(y)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n (L_{it}^{(y)} + \tau L_{it}^{(w)})\tilde{\omega}_i, && L_t^{(w)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n L_{it}^{(w)}\tilde{\omega}_i, && \\
&\epsilon_t^{(w)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n \epsilon_{it}^{(w)}\tilde{\omega}_i, && \epsilon_t^{(y)}(\tilde{\omega}) := \frac{1}{n}\sum_{i=1}^n (\epsilon_{it}^{(y)} + \tau\epsilon_{it}^{(w)})\tilde{\omega}_i. &&
\end{aligned} \tag{1.6}$$

A.2 Assumptions

Assumption A.1. (Universal constants) *The following restrictions are satisfied for all n, T_0, T_1 for some fixed universal constant $c_p, c_{ag}, c_{h|z}, c_\sigma, c_\tau$:*

$$
\begin{aligned}
&dim\{\psi_t\} < c_p, \quad |\rho_{ag}| < c_{ag} < 1, \quad \max_{k\in\{0,1\}} \left\{|\eta^{(k)}_{H|Z}|\right\} < c_{h|z}, \quad \max\{\sigma_w, \sigma_y\} < c_\sigma, \\
&|\tau| < c_\tau.
\end{aligned}
\tag{1.7}
$$

Assumption A.2. (Alignment and size of times shocks) *As T_1, T_0 go to infinity, the following restrictions are satisfied for $k \in \{0,1\}$:*

$$
\begin{aligned}
&\max_{j\in\{z,h\}} \left\{\frac{\|\Lambda^{(j),(k)}\|_{op}}{\|\Lambda^{(j),(k)}\|_{HS}}\right\} \le \frac{c_{op}}{\sqrt{T_k}}, \qquad \frac{\|(\Lambda^{(h),(k)})^\top \Lambda^{(z),(k)}\|_{op}}{\|(\Lambda^{(h),(k)})^\top \Lambda^{(z),(k)}\|_{HS}} = o(1), \\
&\|\Lambda^{(z),(k)}\|_{op} \sim \|\Lambda^{(h),(k)}\|_{op}, \qquad\qquad \|\Lambda^{(z),(k)}\|_{HS} \sim \|\Lambda^{(h),(k)}\|_{HS}
\end{aligned}
\tag{1.8}
$$

For a generic matrix A quantity $\frac{\|A\|^2_{HS}}{\|A\|^2_{op}}$ generalizes the concept of rank, and thus the first part of this assumption says that after projection most shocks are not aligned with respect to a small number of directions. The second part simply says that distributions of $\epsilon^{(h)}$ and $\epsilon^{(z)}$ are not very different.

Assumption A.3. (Size of deterministic trends) *As T_1 go to infinity, there exists a sequence $s_{T_1} \to \infty$ such that the following restrictions hold:*

$$
\sup_{x\neq 0, x^\top \mathbf{1}_n = 0, \alpha, \beta} \left\{\frac{\|x\left(\alpha L^{(w),(1)} + \beta \tilde{L}^{(y),(1)}\right)^\top \Pi_f^{(1)} \Lambda_1^{(z),(1)}\|_\infty}{\|x\left(\alpha L^{(w),(1)} + \beta \tilde{L}^{(y),(1)}\right)^\top \Pi_f^{(1)} \Lambda_1^{(z),(1)}\|_2}\right\} \le \frac{1}{s_{T_1}}
\tag{1.9}
$$

This assumption requires deterministic trends in $L^{(w),(1)}$ and $L^{(y),(1)}$ to be "well-spread" over time even after projection and integration with $\Lambda_1^{(z),(1)}$. For bounded deterministic trends and $\Lambda_1^{(z),(1)}$ generated by a stationary process we would expect s_{T_1} to behave as $\sqrt{T_1}$.

Assumption A.4. (Size of predictable part) *As T_0, T_1 go to infinity the following holds for s_{T_1} from Assumption A.3:*

$$
\|(\Lambda_1^{(z),(1)})^{-1}(\Lambda_0^{(z),(1)})\nu^{(z),(0)}\|_1 = o_p(s_{T_1}).
\tag{1.10}
$$

This assumption restricts the size of the part of $\epsilon^{(z),(1)}$ that is predictable using the past. In a stationary auto-regressive model we have $\|(\Lambda_1^{(z),(1)})^{-1}(\Lambda_0^{(z),(1)})\nu^{(z),(0)}\|_1 = O_p(1)$ and thus the assumption is satisfied for any diverging sequence s_{T_1}. If $\epsilon^{(z)}$ follow a random walk, then $\|(\Lambda_1^{(z),(1)})^{-1}(\Lambda_0^{(z),(1)})\nu^{(z),(0)}\|_1 \sim \sqrt{T_0}$. Since s_{T_1} cannot be larger than $\sqrt{T_1}$, in this case, Assumption A.4 fails for the regime that we are mainly interested in ($T_0 \sim T_1$).

Assumption A.5. (Order of the variances) *The following holds as n, T_0 approach infinity for a universal constant c_{var} and deterministic κ^2:*

$$\max\left\{\hat{\sigma}_Y^2, \hat{\sigma}_W^2\right\} \leq \kappa^2(1 + o_p(1)), \quad \frac{\kappa^2}{\min\{\hat{\sigma}_Y^2, \hat{\sigma}_W^2\}} \leq c_{var}(1 + o_p(1)). \tag{1.11}$$

Assumption A.6. (Quality and size of D_i) *Let $(D - \overline{D})_i := D_i - \frac{1}{n}\sum_{i=1}^n D_i$ and assume that the following holds:*

$$\|D - \overline{D}\|_2 \sim \sqrt{n} \tag{1.12}$$

For some universal constant $c_s > 0$ the following holds:

$$\inf_{\kappa\zeta > \max\{\sigma_w, \sqrt{\sigma_y^2 + \tau\sigma_w^2 + 2\tau\rho_{id}\sigma_y\sigma_w}\}} |s(\zeta)| > c_s \tag{1.13}$$

A.3 Technical lemmas

Lemma A.1. *Let Π be an orthogonal projector on p-dimensional subspace or $\mathbb{R}^T$ and consider a $T \times n$ matrix A such that $\frac{\|A\|_{op}}{\|A\|_{HS}} = o\left(\frac{1}{\sqrt{p}}\right)$. Then the following holds:*

$$\frac{\|(\mathcal{I}_T - \Pi)A\|_{HS}}{\|A\|_{HS}} = 1 + o(1) \tag{1.14}$$

Proof. The result follows from the chain of inequalities:

$$\left|\frac{\|(\mathcal{I}_T - \Pi)A\|_{HS}}{\|A\|_{HS}} - 1\right| \leq \frac{\|\Pi A\|_{HS}}{\|A\|_{HS}} \leq \frac{\|\Pi\|_{HS}\|A\|_{op}}{\|A\|_{HS}} = \sqrt{p} \times o\left(\frac{1}{\sqrt{p}}\right) = o(1) \tag{1.15}$$

□

Lemma A.2. *Suppose Assumptions 3.4,A.1,A.2 hold, then the following is true as T_0, T_1 goes to infinity for $k \in \{0, 1\}$:*

$$\begin{aligned}
&\frac{\|\Pi_f^{(k)}\epsilon^{(z),(k)}\|_2}{\|\Lambda^{(z),(k)}\|_{HS}} = 1 + o_p(1), \\
&\frac{(\epsilon^{(h),(k)})^\top \Pi_f^{(k)}\epsilon^{(z),(k)}}{\|\Pi_f^{(k)}\epsilon^{(z),(k)}\|_2^2} = \eta_{H|Z}^{(k)} + \mathcal{O}_p\left(\frac{\|(\Lambda^{(h),(k)})^\top \Lambda^{(z),(k)}\|_{HS}}{\|\Lambda^{(z),(k)}\|_{HS}^2}\right), \\
&\frac{\left\|\left(\mathcal{I}_{T_0} - \frac{1}{\|\epsilon^{(z),(0)}\|_2^2}\epsilon^{(z),(0)}(\epsilon^{(z),(0)})^\top\right)\Pi_f^{(0)}\epsilon^{(h),(0)}\right\|_2^2}{\|\Lambda^{(h),(0)}\|_{HS}^2} = 1 - \rho_{ag}^2 \frac{trace^2\left(\Lambda^{(z),(0)}(\Lambda^{(h),(0)})^\top\right)}{\|\Lambda^{(z),(0)}\|_{HS}^2\|\Lambda^{(h),(0)}\|_{HS}^2} + o_p(1)
\end{aligned} \tag{1.16}$$

Proof. We prove the first two claims for $k = 1$, the result for $k = 0$ follows in exactly the same way. Theorem 6.3.2 in Vershynin [2018] implies the following:

$$\left|\|\Pi_f^{(1)}\epsilon^{(z),(1)}\|_2 - \|\Pi_f^{(1)}\Lambda^{(z),(1)}\|_{HS}\right| = \mathcal{O}_p\left(\|\Pi_f^{(1)}\Lambda^{(z),(1)}\|_{op}\right) \tag{1.17}$$

which together with Assumptions A.1, A.2 and Lemma A.1 implies the first claim. We also have the following decomposition:

$$(\epsilon^{(h),(1)})^\top \Pi_f^{(1)}\epsilon^{(z),(1)} = \rho_{ag}(\Lambda^{(h),(1)}\nu^{(z)})^\top \Pi_f^{(1)}\Lambda^{(z),(1)}\nu^{(z)} + \sqrt{1-\rho_{ag}^2}(\Lambda^{(h),(1)}\nu^{(h)})^\top \Pi_f^{(1)}\Lambda^{(z),(1)}\nu^{(z)} \tag{1.18}$$

Applying Hanson-Wright inequality to the first part, Lemma 6.2.3 from Vershynin [2018], Assumption A.1, Asssumption A.2, and Lemma A.1 we get he following:

$$\left|(\epsilon^{(h),(1)})^\top \Pi_f^{(1)}\epsilon^{(z),(1)} - \eta_{H|Z}\|\Pi_f^{(1)}\Lambda^{(z),(1)}\|_{HS}^2\right| = \mathcal{O}_p\left(\|(\Lambda^{(z),(1)})^\top \Lambda^{(h),(1)}\|_{HS}\right) \tag{1.19}$$

proving the second claim. The third follows in the same way. □

B General results

B.1 General version of Theorem 1

The results in this subsection apply to a generic weights $\tilde{\omega}$ that depend on the first part of the sample.

B.1.1 Expansion of the estimator

We focus on first stage and reduced form coefficients separately. We start with the decomposition of the first stage:

$$
\begin{aligned}
\hat{\pi} = \frac{1}{n}\frac{\tilde{\omega}^\top W^{(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}} &= \frac{1}{n}\tilde{\omega}^\top\pi + \text{bias}^{(w)} + \text{time noise}^{(w)} + \text{cross noise}^{(w)} \\
\text{bias}^{(w)} &:= \frac{1}{n}\tilde{\omega}^\top\theta^{(w)}\frac{(H^{(1)})^\top\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi^{(1)}Z^{(1)}} \\
\text{time noise}^{(w)} &:= \frac{1}{n}\frac{\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}} \\
\text{cross noise}^{(w)} &:= \frac{1}{n}\frac{\tilde{\omega}^\top E^{(w),(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}}
\end{aligned}
\tag{2.1}
$$

Similar decomposition holds for the reduced form:

$$
\begin{aligned}
\hat{\delta} = \frac{1}{n}\frac{\tilde{\omega}^\top Y^{(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}} &= \frac{\tau}{n}\tilde{\omega}^\top\pi + \text{bias}^{(y)} + \text{time noise}^{(y)} + \text{cross noise}^{(y)} \\
\text{bias}^{(y)} &:= \frac{1}{n}\tilde{\omega}^\top(\tilde{\theta}^{(y)})\frac{(H^{(1)})^\top\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}} \\
\text{time noise}^{(y)} &:= \frac{1}{n}\frac{\tilde{\omega}^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}} \\
\text{cross noise}^{(y)} &:= \frac{1}{n}\frac{\tilde{\omega}^\top \tilde{E}^{(y),(1)}\Pi_f^{(1)}Z^{(1)}}{(Z^{(1)})^\top\Pi_f^{(1)}Z^{(1)}}
\end{aligned}
\tag{2.2}
$$

B.1.2 Analysis of the error

Lemma B.1. *Suppose Assumptions 3.3, 3.4, A.1, A.2 hold, then the following is true:*

$$
\begin{pmatrix} \textit{cross noise}^{(y)} \\ \textit{cross noise}^{(w)} \end{pmatrix} = \frac{\|\tilde{\omega}\|_2}{n\|\Lambda^{(z),(1)}\|_{HS}}(\xi_{cr} + o_p(1))
\tag{2.3}
$$

where $\xi_{cr} \sim \mathcal{N}(0, \Sigma)$.

Proof. First, we condition on $\tilde{\omega}, Z$ and use the fact that idiosyncratic errors are independent over time and of

$\epsilon^{(z)}, \epsilon^{(h)}$ and have a normal distribution. Then we use Lemma A.2 to go from $\|\Pi_f^{(1)}\epsilon^{(z),(1)}\|_2$ to $\|\Lambda^{(z),(1)}\|_{HS}$ at the expense of $(1 + o_p(1))$ factor. □

Define the following objects:

$$\begin{aligned} \rho(\tilde{\omega}) &:= \frac{\tilde{\omega}^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\left(\tilde{\omega}^\top L^{(w,(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\right)^\top}{\|\tilde{\omega}^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2\|\tilde{\omega}^\top L^{(w,(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2}, \\ \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega}) &:= \begin{pmatrix} \frac{\|\tilde{\omega}^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2}{n\|\Lambda^{(z),(1)}\|_{HS}^2} & 0 \\ 0 & \frac{\|\tilde{\omega}^\top L^{(w,(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2}{n\|\Lambda^{(z),(1)}\|_{HS}^2} \end{pmatrix} \begin{pmatrix} \sqrt{1-\rho^2(\tilde{\omega})} & \rho(\tilde{\omega}) \\ 0 & 1 \end{pmatrix} \end{aligned} \tag{2.4}$$

Lemma B.2. *Suppose Assumptions 3.3, 3.4, A.1, A.2, A.3, A.4 hold, then we have the following:*

$$\begin{pmatrix} \textit{time noise}^{(y)} \\ \textit{time noise}^{(w)} \end{pmatrix} = \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega})(\xi_{z,T_1} + o_p(1)) \tag{2.5}$$

where $\mathbb{E}[\xi_{z,T_1}] = 0$, $\mathbb{V}[\xi_{z,T_1}] = \mathcal{I}_2$, *and* ξ_{z,T_1} *is independent of* ξ_{cr}. *As* T_1 *increases* ξ_{z,T_1} *converges in distribution to* $\mathcal{N}(0, \mathcal{I}_2)$.

Proof. To prove this we first separate $\epsilon^{(z),(1)}$ into two parts: predictable from $\epsilon^{(z),(0)}$ and unpredictable one. By Assumption 3.4 we get the unpredictable part is equal to $\Lambda_1^{(z),(1)}\nu^{(z),(1)}$ and predictable is equal to $\mu_{1|0} := \Lambda_0^{(z),(1)}\nu^{(z),(0)}$. The unpredictable part delivers the result (using Lemma A.2) distribution described in the statement. Asymptotic normality follows from Assumption A.3 and multivariate Lindeberg's CLT. To finish the proof we have to prove that the bias from $\mu_{1|0}$ is of the smaller order. It from the chain of inequalities that follow from Assumptions A.3, A.4:

$$\begin{aligned} &\frac{|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_0^{(z),(1)}\nu^{(z),(0)}|}{\|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2} = \frac{|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}(\Lambda_1^{(z),(1)})^{-1}\Lambda_0^{(z),(1)}\nu^{(z),(0)}|}{\|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2} \leq \\ &\quad \frac{\|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_\infty\|(\Lambda_1^{(z),(1)})^{-1}\Lambda_0^{(z),(1)}\nu^{(z),(0)}\|_1}{\|\tilde{\omega}^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2} \leq \frac{\|(\Lambda_1^{(z),(1)})^{-1}\Lambda_0^{(z),(1)}\nu_0^{(z)}\|_1}{s_{T_1}} = o_p(1). \end{aligned} \tag{2.6}$$

The same holds for $\tilde{L}^{(y),(1)}$ instead of $L^{(w),(1)}$ concluding the proof. □

Corollary B.1. *Suppose Assumptions 3.3, 3.4, A.1, A.2 hold, then the following is true:*

$$\begin{pmatrix} \textit{bias}^{(y)} \\ \textit{bias}^{(w)} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\tilde{\omega}^\top \tilde{\theta}^{(y)} \\ \frac{1}{n}\tilde{\omega}^\top \theta^{(w)} \end{pmatrix} \left(\eta_{H|Z} + \mathcal{O}_p\left(\frac{\|(\Lambda^{(h),(1)})^\top \Lambda^{(z),(1)}\|_{HS}}{\|\Lambda^{(z),(1)}\|_{HS}^2}\right)\right) \tag{2.7}$$

Proof. The result is a direct consequence of Lemma A.2 and definition of the bias. □

Next theorem follows from the lemmas above.

Theorem B.1. *Suppose Assumption 3.3, 3.4, A.1, A.2, A.3, A.4 hold. Then the following is true:*

$$\begin{pmatrix} \hat{\delta} - \frac{\tau}{n}\tilde{\omega}^\top \pi \\ \hat{\pi} - \frac{1}{n}\tilde{\omega}^\top \pi \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\tilde{\omega}^\top \tilde{\theta}^{(y)} \\ \frac{1}{n}\tilde{\omega}^\top \theta^{(w)} \end{pmatrix} \left(\eta_{H|Z} + \mathcal{O}_p\left(\frac{\|(\Lambda^{(h),(1)})^\top \Lambda^{(z),(1)}\|_{HS}}{\|\Lambda^{(z),(1)}\|_{HS}^2} \right) \right) + \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega})(\xi_{z,T_1} + o_p(1)) + \frac{\|\tilde{\omega}\|_2}{n\|\Lambda^{(z),(1)}\|_{HS}}(\xi_{cr} + o_p(1)) \quad (2.8)$$

B.2 General version of Theorem 2

Our previous analysis indicates that the key component of bias of an arbitrary weighted estimator depends on the covariance of the weights with $\theta^{(w)}$ and $\tilde{\theta}^{(y)}$. In this section we bound this covariance using results from Hirshberg [2021]. In this section ω refers to the weights described in Section 2.2.

B.2.1 Random oracle weights

Let $\mathbb{E}_{cs}[\cdot]$ be the expectation with respect to noise $E^{(w)}, E^{(y)}$, conditional on $\nu^{(z),(0)}, \nu^{(h),(0)}$. We define random oracle weights $\omega^\star$ as a solution to the following problem:

$$\omega^\star = \arg\min_{\{w\}} \left\{ \zeta^2 \frac{T_0}{n^2} \|w\|_2^2 + \frac{\mathbb{E}_{cs}\|\frac{1}{n} w^\top Y^{(0)} \Pi^{(0)}\|_2^2}{\hat{\sigma}_Y^2} + \frac{\mathbb{E}_{cs}\|\frac{1}{n} w^\top W^{(0)} \Pi^{(0)}\|_2^2}{\hat{\sigma}_W^2} \right\}$$

$$\text{subject to:} \qquad \frac{1}{n}\sum_{i=1}^n w_i D_i = 1, \qquad \frac{1}{n}\sum_{i=1}^n w_i = 0, \tag{2.9}$$

Define $r^\star := \omega - \omega^\star$ – the deviation of the empirical weights from the oracle weights. Define the implicit regularization parameter

$$\tilde{\zeta}^2 := \kappa^2 \zeta^2 + \frac{\kappa^2 \tilde{\sigma}_y^2 (T_0 - p - 1)}{T_0 \hat{\sigma}_Y^2} + \frac{\kappa^2 \sigma_w^2 (T_0 - p - 1)}{T_0 \hat{\sigma}_W^2}, \tag{2.10}$$

and the following symmetric matrix:

$$\begin{aligned}\hat{\Gamma} := {} & \frac{1}{\hat{\sigma}_W^2} L^{(w),(0)} \Pi^{(0)} (L^{(w),(0)})^\top + \frac{1}{\hat{\sigma}_Y^2} \tilde{L}^{(y),(0)} \Pi^{(0)} (\tilde{L}^{(y),(0)})^\top + \\ & \frac{\|\Pi_0 \epsilon^{(h),(0)}\|_2^2}{\hat{\sigma}_W^2} (\theta^{(w)})(\theta^{(w)})^\top + \frac{\|\Pi_0 \epsilon^{(h),(0)}\|_2^2}{\hat{\sigma}_Y^2} (\tilde{\theta}^{(y)})(\tilde{\theta}^{(y)})^\top + \\ & \frac{1}{\hat{\sigma}_W^2} L^{(w),(0)} \Pi^{(0)} \epsilon^{(h),(0)} (\theta^{(w)})^\top + \frac{1}{\hat{\sigma}_W^2} \theta^{(w)} (\epsilon^{(h),(0)})^\top \Pi^{(0)} (L^{(w),(0)})^\top + \\ & \qquad \frac{1}{\hat{\sigma}_Y^2} \tilde{L}^{(y),(0)} \Pi^{(0)} \epsilon^{(h),(0)} (\tilde{\theta}^{(y)})^\top + \frac{1}{\hat{\sigma}_Y^2} \tilde{\theta}^{(y)} (\epsilon^{(h),(0)})^\top \Pi^{(0)} (L^{(y),(0)})^\top\end{aligned} \tag{2.11}$$

Using these definitions and computing expectations in (2.9) we get another representation for $\omega^\star$:

$$\begin{aligned}
&\omega^\star = \arg\min_{\{w\}} \left\{ \tilde{\zeta}^2 \frac{T_0}{n^2}\|w\|_2^2 + \kappa^2 w^\top \hat{\Gamma} w \right\} \\
&\quad \text{subject to:} \\
&\qquad \frac{1}{n}\sum_{i=1}^n w_i D_i = 1, \\
&\qquad \frac{1}{n}\sum_{i=1}^n w_i = 0,
\end{aligned} \tag{2.12}$$

B.2.2 Deterministic oracle weights

We define additional, "deterministic" oracle weights ω_{det} as a solution to the following problem:

$$\begin{aligned}
&\omega_{det} = \arg\min_{\{w\}} \left\{ \tilde{\zeta}^2 \frac{T_0}{n^2}\|w\|_2^2 + \frac{\kappa^2}{n^2} w^\top \Gamma w \right\} \\
&\quad \text{subject to:} \\
&\qquad \frac{1}{n}\sum_{i=1}^n w_i D_i = 1, \\
&\qquad \frac{1}{n}\sum_{i=1}^n w_i = 0,
\end{aligned} \tag{2.13}$$

Define $r_{det} := \omega^\star - \omega_{det}$ – the deviation of the random oracle weights from the deterministic ones.

B.2.3 Technical lemmas

The first lemma describes the key properties of ω_{det}. Define $\chi := \max\{\sigma_w, \tilde{\sigma}_y, c_{fe}\}$.

Lemma B.3. *Suppose Assumptions 3.5, A.1, A.5, A.6 hold, in addition, as n and T_0 approach infinity $\frac{\kappa\zeta}{\chi} = c_\zeta + o(1)$, where $c_\zeta > 2$ and $\frac{T_0}{n} = c_{asp} + o(1)$, where $c_{asp} > 0$. Then the following is true for n and T_0 large enough:*

$$\begin{aligned}
&\|D - \overline{D}\|_2^2 \le \|\omega_{det}\|_2^2 \le cn(1 + c_{var})c_{\breve{\omega}}^2 \left(1 + \frac{c_{fe}^2}{c_\zeta^2 \chi^2 c_{asp}}\right)(1 + o_p(1)) \\
&\kappa^2 \omega_{det}^\top \Gamma \omega \le cn^2 \left(c_{\breve{\omega}}^2\right)(c_{asp} c_\zeta^2 \chi^2 + c_{fe}^2)(1 + c_{var})(1 + o_p(1))
\end{aligned} \tag{2.14}$$

Proof. To prove the result we define yet another weights:

$$\tilde{\omega}_{det} = \arg\min_{\{w\}} \left\{ \tilde{\zeta}^2 \frac{T_0}{n^2} \|w\|_2^2 + \frac{\kappa^2}{n^2} w^\top \Gamma w \right\}$$

subject to:

$$\frac{1}{n}\sum_{i=1}^{n} w_i \pi_i = 1, \qquad (2.15)$$

$$\frac{1}{n}\sum_{i=1}^{n} w_i = 0,$$

The difference between these weights and ω_{det} is that they average up to 1 with respect to π_i, not D_i. It is straightforward to see that the solution has the "ridge" form:

$$\frac{1}{n}\omega_{det} = \frac{(\tilde{\zeta}^2 \mathcal{I}_{T_0} + \frac{\kappa^2}{T_0}\Pi_\alpha \Gamma \Pi_\alpha)^{-1}(\pi - \overline{\pi})}{(\pi - \overline{\pi})^\top (\tilde{\zeta}^2 \mathcal{I}_{T_0} + \frac{\kappa^2}{T_0}\Pi_\alpha \Gamma \Pi_\alpha)^{-1}(\pi - \overline{\pi})} \qquad (2.16)$$

and thus we get the following

$$\frac{1}{n}(\tilde{\omega}_{det})^\top D = \frac{(D - \overline{D})^\top (\tilde{\zeta}^2 \mathcal{I}_{T_0} + \frac{\kappa^2}{T_0}\Pi_\alpha \Gamma \Pi_\alpha)^{-1}(\pi - \overline{\pi})}{(\pi - \overline{\pi})^\top (\tilde{\zeta}^2 \mathcal{I}_{T_0} + \frac{\kappa^2}{T_0}\Pi_\alpha \Gamma \Pi_\alpha)^{-1}(\pi - \overline{\pi})} = s\left(\frac{\tilde{\zeta}}{\kappa}\right) \qquad (2.17)$$

By construction we have for n, T_0 large enough $\frac{\tilde{\zeta}}{\kappa} > \zeta > \frac{c_\zeta \max\{\sigma_w, \tilde{\sigma}_y, c_{fe}\}}{2\kappa} > \frac{\max\{\sigma_w, \tilde{\sigma}_y\}}{\kappa}$ which by Assumption A.6 implies $s\left(\frac{\tilde{\zeta}}{\kappa}\right)$ is bounded away from zero by c_s and thus we can define the following weights:

$$\breve{\omega}_{new} := \frac{1}{s\left(\frac{\tilde{\zeta}}{\kappa}\right)}\tilde{\omega}_{det} \qquad (2.18)$$

that now average to 1 once multiplied by D_i. With these weights we get the following inequalities:

$$\tilde{\zeta}^2 \frac{T_0}{n^2}\|\omega_{det}\|_2^2 + \frac{\kappa^2}{n^2}(\omega_{det})^\top \Gamma(\omega_{det}) \le \tilde{\zeta}^2 \frac{T_0}{n^2}\|\breve{\omega}_{new}\|_2^2 + \frac{\kappa^2}{n^2}(\breve{\omega}_{new})^\top \Gamma(\breve{\omega}_{new}) = \frac{1}{s^2\left(\frac{\tilde{\zeta}}{\kappa}\right)}\left(\tilde{\zeta}^2 \frac{T_0}{n^2}\|\breve{\omega}_{det}\|_2^2 + \frac{\kappa^2}{n^2}(\breve{\omega}_{det})^\top \Gamma(\breve{\omega}_{det})\right) \qquad (2.19)$$

As a result, we need to bound only the last part. Here we use Assumption 3.5 to get the following for n, T_0 large enough:

$$\tilde{\zeta}^2 \frac{T_0}{n^2}\|\breve{\omega}_{det}\|_2^2 + \frac{\kappa^2}{n^2}(\breve{\omega}_{det})^\top \Gamma(\breve{\omega}_{det}) \le c(1 + c_{var})(c_\zeta^2 c_{asp}\chi^2 + c_{fe}^2)(1 + o_p(1)) \qquad (2.20)$$

Applying this bound separately to $\tilde{\zeta}^2 \frac{T_0}{n^2}\|\omega_{det}\|_2^2$ and $\frac{\kappa^2}{n^2}(\omega_{det})^\top \Gamma(\omega_{det})$, and using the fact that $\|\omega_{det}\|$ cannot be smaller that $\|D - \overline{D}\|$ (solution for $\tilde{\zeta}$ equal to infinity) we get the result. □

Our next lemma provides connection between $\omega^\star$ and ω_{det}.

Lemma B.4. *Suppose Assumptions 3.4, 3.5, A.1, A.2, A.5 hold, and* $\frac{1}{\|\Lambda^{(h),(0)}\|_{HS}} = o(1)$, $\frac{\kappa\zeta}{\chi} = c_\zeta + o(1)$ *and* $c_\zeta > 2$ *as* n, T_0 *approach infinity. Then the following is true as* n, T_0 *approach infinity:*

$$
\begin{aligned}
&|r_{det}^\top \theta^{(w)}| = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^\top \Gamma \omega_{det}}}{\|\Lambda^{(h),(0)}\|_{HS}}\right)\\
&|r_{det}^\top \tilde{\theta}^{(y)}| = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^\top \Gamma \omega_{det}}}{\|\Lambda^{(h),(0)}\|_{HS}}\right)\\
&\|r_{det}\|_2 = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^\top \Gamma \omega_{det}}}{\sqrt{T_0} c_\zeta \chi}\right) \qquad (\omega^\star)^\top \hat{\Gamma}\omega^\star \leq c(\omega_{det})^\top \Gamma \omega_{det}(1 + o_p(1))
\end{aligned}
\tag{2.21}
$$

Proof. Observe that in the described regime we have the following $\tilde{\zeta}^2 \geq c_\zeta^2 \max\{\sigma_w^2, \tilde{\sigma}_y^2\}$. Consider the following chain of inequalities:

$$
\begin{aligned}
0 \geq T_0\tilde{\zeta}^2\|\omega^\star\|_2^2 + \kappa^2(\omega^\star)^\top\hat{\Gamma}\omega^\star - T_0\tilde{s}^2\|\omega_{det}\|_2^2 - \kappa^2(\omega_{det})^\top\hat{\Gamma}\omega_{det} \geq\\
2\kappa^2 r_{det}^\top(\hat{\Gamma} - \Gamma)\omega_{det} + \kappa^2 r_{det}^\top(\hat{\Gamma} - \Gamma)r_{det} + \kappa^2 r_{det}^\top \Gamma r_{det} + T_0\tilde{\zeta}^2\|r_{det}\|_2^2
\end{aligned}
\tag{2.22}
$$

Here the first follows by definition of $\omega^\star$, the second follows by definition of ω_{det}. Define the following variables:

$$
\begin{aligned}
&x_1^2 := T_0\tilde{\zeta}^2\|r_{det}\|_2^2\\
&x_2^2 := \kappa^2 r_{det}^\top \Gamma r_{det}\\
&x_3^2 := \kappa^2(\omega_{det})^\top \Gamma \omega_{det}
\end{aligned}
\tag{2.23}
$$

We will bound the first two terms in the sum above using this qualities. We have the following expansion:

$$
\begin{aligned}
&\kappa^2(\hat{\Gamma} - \Gamma) = \frac{\kappa^2}{\hat{\sigma}_W^2} L^{(w),(0)}(\Pi^{(0)} - \Pi_f^{(0)})(L^{(w),(0)})^\top + \frac{\kappa^2}{\hat{\sigma}_Y^2}\tilde{L}^{(y),(0)}(\Pi^{(0)} - \Pi_f^{(0)})(\tilde{L}^{(y),(0)})^\top +\\
&(\|\Pi_0\epsilon^{(h),(0)}\|_2^2 - (1 - (\rho_{H|Z}^{(0)})^2)\|\Lambda^{(h),(0)}\|_{HS}^2)\left(\frac{\kappa^2}{\hat{\sigma}_W^2}(\theta^{(w)})(\theta^{(w)}) + \frac{\kappa^2}{\hat{\sigma}_Y^2}\tilde{\theta}^{(y)}(\tilde{\theta}^{(y)})^\top\right) +\\
&\frac{\kappa^2}{\hat{\sigma}_W^2} L^{(w),(0)}\Pi^{(0)}\epsilon^{(h),(0)} + \frac{\kappa^2}{\hat{\sigma}_W^2}\theta^{(w)}(\epsilon^{(h),(0)})^\top\Pi^{(0)}(L^{(w),(0)})^\top +\\
&\qquad\qquad \frac{\kappa^2}{\hat{\sigma}_Y^2}\tilde{L}^{(y),(0)}\Pi^{(0)}\epsilon^{(h),(0)}(\tilde{\theta}^{(y)})^\top + \frac{\kappa^2}{\hat{\sigma}_Y^2}\tilde{\theta}^{(y)}(\epsilon^{(h),(0)})^\top\Pi^{(0)}(L^{(y),(0)})^\top
\end{aligned}
\tag{2.24}
$$

We bound $\kappa^2 r_{det}^\top(\hat{\Gamma} - \Gamma)r_{det}$ in terms of $|x_2|$ and $|x_3|$. We only need to bound the last six terms, because the first two are positive. We start with the last four, they all behave in the same way, so we focus on the one that involves $\theta^{(w)}$ and $L^{(w)}$:

$$\left|\frac{\kappa^2}{\hat{\sigma}_W^2}(r_{det})^\top \theta^{(w)}(\epsilon^{(h),(0)})^\top \Pi^{(0)}(L^{(w),(0)})^\top r_{det}\right| \leq$$

$$(1+o_p(1))|\frac{|x_2^2|}{\|\Lambda^{(h),(0)}\|_{HS}} \frac{\left|(\epsilon^{(h),(0)})^\top \Pi^{(0)}\Pi_f^{(0)}(L^{(w),(0)})^\top r_{det}\right|}{\|\Pi_f^{(0)}(L^{(w),(0)})^\top r_{det}\|_2} \leq$$

$$(1+o_p(1))|\frac{|x_2^2|}{\|\Lambda^{(h),(0)}\|_{HS}} \sup_{x,\|x\|_2=1, x\in \mathrm{Im}(\Pi_f^{(0)}(L^{(w),(0)})^\top)} |(\epsilon^{(h),(0)})^\top \Pi^{(0)} x| \quad (2.25)$$

Let Π_w be the projector on $\mathrm{Im}((L^{(w),(0)})\Pi_f^{(0)})$, by Assumption 3.5 we know that the dimension of this subspace is $o(\min\{n, T_0\})$. Using this and Lemmas A.1, A.2 we get the following:

$$\sup_{x,\|x\|_2=1, x\in \mathrm{Im}(\Pi_f^{(0)}(L^{(w),(0)})^\top)} |(\epsilon^{(h),(0)})^\top \Pi^{(0)} x| = \|\Pi_w^{(0)}\Pi^{(0)}\epsilon^{(h),(0)}\|_2 \leq$$

$$\|\Pi_w^{(0)}\epsilon^{(h),(0)}\|_2 + |\hat{\eta}_{H|Z}^{(0)}|\|\Pi_w^{(0)}\epsilon^{(z),(0)}\|_2 = o_p\left(\|\Lambda^{(h),(0)}\|_{HS}\right) \quad (2.26)$$

This implies the following bound:

$$\left|\frac{\kappa^2}{\hat{\sigma}_W^2}(r_{det})^\top \theta^{(w)}(\epsilon^{(h),(0)})^\top \Pi^{(0)}(L^{(w),(0)})^\top r_{det}\right| \leq o_p\left(x_2^2\right) \quad (2.27)$$

Using Lemma A.2 we get the following:

$$\left|\|\Pi_0\epsilon^{(h),(0)}\|_2^2 - (1-(\rho_{H|Z}^{(0)})^2)\|\Lambda^{(h),(0)}\|_{HS}^2\right| \left(\frac{\kappa^2}{\hat{\sigma}_W^2}(r_{det})^\top(\theta^{(w)})(\theta^{(w)})^\top r_{det} + \frac{\kappa^2}{\hat{\sigma}_Y^2}(r_{det})^\top \tilde{\theta}^{(y)}(\tilde{\theta}^{(y)})^\top r_{det}\right) =$$

$$o_p(x_2^2) \quad (2.28)$$

Next we bound $\left|\kappa^2\omega_{det}^\top(\hat{\Gamma}-\Gamma)r_{\mathrm{det}}\right|$ in terms of x_2, x_3. Applying exactly the same argument as above we get the following:

$$\left|\kappa^2\omega_{det}^\top(\hat{\Gamma}-\Gamma)r_{\mathrm{det}}\right| = o_p\left(|x_2x_3|\right) + |x_2x_3|\mathcal{O}_p\left(\frac{1}{\|\Lambda^{(z),(0)}\|_{HS}}\right) = o_p\left(|x_2x_3|\right) \quad (2.29)$$

Combining this with previously derived bound we get the following equation:

$$x_1^2 + x_2^2(1+o_p(1)) + o_p\left(|x_2x_3|\right) \leq 0 \quad (2.30)$$

This implies that $|x_2| = o_p(|x_3|)$ and the same holds for $|x_1|$. Going back to the original notation this gives us the following bounds:

$$
\begin{aligned}
&|r_{det}^{\top}\theta^{(w)}| = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^{\top}\Gamma\omega_{det}}}{\|\Lambda^{(h),(0)}\|_{HS}}\right)\\
&|r_{det}^{\top}\tilde{\theta}^{(y)}| = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^{\top}\Gamma\omega_{det}}}{\|\Lambda^{(h),(0)}\|_{HS}}\right)\\
&\|r_{det}\|_2 = o_p\left(\frac{\sqrt{\kappa^2(\omega_{det})^{\top}\Gamma\omega_{det}}}{\sqrt{T_0}c_{\zeta}\chi}\right)
\end{aligned}
\tag{2.31}
$$

By combining the bounds above and using a straightforward implication:

$$
\left|(\omega_{det})^{\top}(\hat{\Gamma}-\Gamma)\omega_{det}\right| \leq o_p\left((\omega_{det})^{\top}\Gamma\omega_{det}\right) \tag{2.32}
$$

we get the following bound, thus finishing the proof:

$$
(\omega^{\star})^{\top}\hat{\Gamma}\omega^{\star} \leq 2(\omega_{det})^{\top}\hat{\Gamma}\omega_{det} + 2(r_{det})^{\top}\hat{\Gamma}r_{det} \leq c(\omega_{det})^{\top}\Gamma\omega_{det}(1+o_p(1)). \tag{2.33}
$$

□

The next lemma provides a technical tool that connects the bound on the covariance we are interested to the bound on $(\omega^{\star})^{\top}\hat{\Gamma}\omega^{\star}$.

Lemma B.5. *Suppose Assumptions 3.4, 3.5, A.1, A.2, A.5 hold, $\frac{T_0}{n} = c_{asp} + o(1)$, where $c_{asp} \in (0,1)$; also suppose that for some (possibly random) x the following holds with probability at least $1-\alpha$:*

$$
\begin{aligned}
&\|x\|_2 \leq s,\\
&\|x^{\top}(\theta^{(w)}(\epsilon^{(h),(0)})^{\top} + L^{(w)})\Pi^{(0)}\|_2^2 \leq cs^2T_0\\
&\|x^{\top}(\tilde{\theta}^{(y)}(\epsilon^{(h),(0)})^{\top} + \tilde{L}^{(y)})\Pi^{(0)}\|_2^2 \leq cs^2T_0
\end{aligned}
\tag{2.34}
$$

Then the following holds with probability at least $1-\alpha$

$$
\begin{aligned}
&|x^{\top}\theta^{(w)}| \leq c\frac{s\sqrt{T_0}}{\|\Lambda^{(h),(0)}\|_{HS}}(1+o_p(1))\\
&|x^{\top}\tilde{\theta}^{(y)}| \leq c\frac{s\sqrt{T_0}}{\|\Lambda^{(h),(0)}\|_{HS}}(1+o_p(1))
\end{aligned}
\tag{2.35}
$$

Proof. In the course of the previous lemma we showed the following:

$$
\kappa^2 r_{det}^{\top}\hat{\Gamma}r_{det} \geq \kappa^2 r_{det}^{\top}\Gamma r_{det}(1+o_p(1)) \tag{2.36}
$$

It is clear from the proof that the same result hold for arbitrary vector x, which implies the result in a straightforward way. □

The next lemma establishes the rate for $(r^\star)^\top \hat{\Gamma} r^\star$ and $\|r^\star\|_2$.

Lemma B.6. *Suppose Assumptions 3.1,3.2,3.3,3.4,3.5, A.1,A.2 hold; also suppose that as n, T_0 approach infinity we have the following:*

$$\frac{T_0}{n} = c_{asp} + o(1), \quad \frac{\kappa\zeta}{\chi} = c_\zeta + o(1), \quad c_{fe} < c\sigma. \tag{2.37}$$

where $c_{asp} \in (0,1)$ and $c_\zeta > c \max\left\{1, \frac{\max\{\sigma_w, \tilde{\sigma}_y\}}{\sqrt{c_{asp}}}\right\}$. Then we have the following with probability at least $1 - c\exp\left(-c(\min\left\{\sigma_w^2, \tilde{\sigma}_y^2\right\} n\right)$:

$$\begin{aligned}
&\|r^\star\|_2 \le c\sqrt{n}\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1)) \\
&\|(r^\star)^\top(\theta^{(w)}(\epsilon^{(h),(0)})^\top + L^{(w)})\Pi^{(0)}\|_2 \le cn\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1)) \\
&\|(r^\star)^\top(\tilde{\theta}^{(y)}(\epsilon^{(h),(0)})^\top + \tilde{L}^{(y)})\Pi^{(0)}\|_2 \le cn\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1))
\end{aligned} \tag{2.38}$$

Proof. We have the following implication from Lemmas B.3, B.4:

$$\begin{aligned}
&\frac{1}{n}\sqrt{\kappa^2(\omega^\star)^\top \hat{\Gamma}\omega^\star} \le c c_\zeta \chi(1 + o_p(1)), \\
&\frac{1}{n}\|\omega^\star\|_2 \le \frac{c}{\sqrt{n}}(1 + o_p(1))
\end{aligned} \tag{2.39}$$

Define the following objects:

$$\begin{aligned}
X_1 &:= \left(Y^{(0)}\Pi^{(0)}\right)^\top, & X_2 &:= \left(W^{(0)}\Pi^{(0)}\right)^\top, \\
\varepsilon_1 &:= \left((E^{(y),(0)} + \tau E^{(w),(0)})\Pi^{(0)}\right)^\top, & \varepsilon_1 &:= \left((E^{(w),(0)})\Pi^{(0)}\right)^\top, \\
A_1 &:= X_1 - \varepsilon_1, & A_2 &:= X_2 - \varepsilon_2,
\end{aligned} \tag{2.40}$$

and consider the optimization problem:

$$\hat{\theta} := \arg\min_{\theta \in \Theta}\left\{T_0\zeta^2\kappa^2\||\theta\|_2^2 + \frac{\kappa^2\|X_1\theta\|_2^2}{\hat{\sigma}_Y^2} + \frac{\kappa^2\|X_2\theta\|_2^2}{\hat{\sigma}_W^2}\right\} \tag{2.41}$$

It is immediate for the appropriate Θ we get $\hat{\theta} = \frac{\omega}{n}$. Also, if we take expectations conditionally on $\{Z_t, H_t\}_{t=1}^{T_0}$ and then optimize, then the solution would be equal to $\theta^\star := \frac{\omega^\star}{n}$. As a result, to control $r^\star$ we need to control $\theta - \theta^\star$.

Define η^2 as a solution to the following equation:

$$\zeta^2 = (\eta^2 - 1)\left(\frac{1}{\hat{\sigma}_Y^2}\tilde{\sigma}_y^2 + \frac{1}{\hat{\sigma}_W^2}\sigma_w^2\right)\frac{T_0 - p - 1}{T_0}. \tag{2.42}$$

By assumption for n, T_0 large enough we have that $\eta^2 > 1 + \frac{b}{2}$. With this parameter the minimized function in (2.41) has the following form:

$$\left(T_0(\eta^2 - 1)\frac{\kappa^2}{\hat{\sigma}_Y^2}\tilde{\sigma}_y^2\frac{T_0 - p - 1}{T_0}\||\theta\|_2^2 + \frac{\kappa^2\|X_1\theta\|_2^2}{\hat{\sigma}_Y^2}\right) + \left(T_0(\eta^2 - 1)\frac{\kappa^2}{\hat{\sigma}_W^2}\sigma_w^2\frac{T_0 - p - 1}{T_0}\||\theta\|_2^2 + \frac{\kappa^2\|X_2\theta\|_2^2}{\hat{\sigma}_W^2}\right) \tag{2.43}$$

A more general version of such problem, but with a single X, instead of X_1, X_2 has been considered in Hirshberg [2021]. His analysis trivially extends to this version, by simply considering two separate bounds for each term. We consider one such bound, the second is analogous.

All the assumptions of Hirshberg [2021] are satisfied by Assumption 3.3 and 3.4 once we condition on $\{Z_t, H_t\}_{t=1}^{T_0}$, and thus we can use its Theorem 1. It implies the following for $\eta^2 > 1 + \frac{cR}{n}$:

$$\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \leq s \\
&\|A_2(\hat{\theta} - \theta^\star)\| \leq s\sqrt{T_0}\eta
\end{aligned} \tag{2.44}$$

with probability at least $1 - c\exp\left(-c\min\left\{\sigma_w^2\frac{T_0-p-1}{T_0}n, R, T_0\right\}\right)$, where s satisfies the following constraint:

$$\begin{aligned}
s^2 \geq \frac{c\sigma_w^2\frac{T_0-p-1}{T_0}s^2 n}{T_0\eta^2}(1 + o_p(1)) + \frac{c(\sigma_w^2\frac{T_0-p-1}{T_0}(R\|\theta^\star\|_2^2)^{1,1/2}}{\eta^2 T_0}(1 + o_p(1)) +& \\
\frac{c\sqrt{\sigma_w^2\frac{T_0-p-1}{T_0}}\|A_2(\hat{\theta} - \theta^\star)\|_2 s\sqrt{n} + c\sigma_w^2\frac{T_0-p-1}{T_0}(n\|\theta^\star\|_2^2)^{1/2}s\sqrt{n}}{\eta^2 T_0}&(1 + o_p(1)),
\end{aligned} \tag{2.45}$$

and R satisfies the following constraint:

$$R \geq c\sigma_{R+1}(A_2)\frac{s\sqrt{n}}{s + \frac{\sqrt{\sigma_w^2\frac{T_0-p-1}{T_0}}}{\sqrt{n}}(1 + o_p(1))} \tag{2.46}$$

By choosing $R = \min\{n, T_0\}$ we can simplify this to the following condition:

$$s^2 \geq c\left(\frac{\sigma_w^2 s^2}{c_{asp}\eta^2} + \frac{\sigma_w^2 + \sigma_w}{\eta^2 T_0} + \frac{\sigma_w c_\zeta \chi + \sigma_w^2}{\eta^2 T_0}s\sqrt{n}\right)(1 + o_p(1))) \tag{2.47}$$

which delivers the following condition with probability at least $1 - c\exp\left(-c\sigma_w^2 n\right)$ as long as $\eta^2 \geq c\frac{\sigma_w^2}{c_{asp}}$ which is guaranteed by assumptions on c_ζ:

$$\|\hat{\theta} - \theta^\star\|_2 \leq c\left(\frac{\chi^2 c_\zeta}{\sqrt{c_{asp}}\sqrt{n}} + \frac{\sqrt{\chi} + \chi}{\sqrt{c_{asp}}\sqrt{n}}\right)(1 + o_p(1)) \tag{2.48}$$

The same bound is true if we use A_1 and thus with probability at least $1 - c\exp\left(-c\min\{\sigma_w^2, \tilde{\sigma}_y^2\}n\right)$:

$$\begin{aligned}
&\|\hat{\theta} - \theta^\star\|_2 \le c\left(\frac{\chi^2 c_\zeta}{\sqrt{c_{asp}}\sqrt{n}} + \frac{\sqrt{\chi} + \chi}{\sqrt{c_{asp}}\sqrt{n}}\right)(1 + o_p(1)) \\
&\|A_1(\hat{\theta} - \theta^\star)\|_2 \le c\left(\chi 2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1)) \\
&\|A_2(\hat{\theta} - \theta^\star)\|_2 \le c\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1))
\end{aligned} \tag{2.49}$$

Translating the bound into our original notation and simplifying we get the following bounds thus concluding the proof:

$$\begin{aligned}
&\|r^\star\|_2 \le c\sqrt{n}\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1)) \\
&\|(r^\star)^\top(\theta^{(w)}(\epsilon^{(h),(0)})^\top + L^{(w)})\Pi^{(0)}\|_2 \le cn\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1)) \\
&\|(r^\star)^\top(\tilde{\theta}^{(y)}(\epsilon^{(h),(0)})^\top + \tilde{L}^{(y)})\Pi^{(0)}\|_2 \le cn\left(\chi^2 c_\zeta + \sqrt{\chi} + \chi\right)(1 + o_p(1))
\end{aligned} \tag{2.50}$$

□

Lemma B.7. *Suppose Assumptions 3.1,3.2,3.3,3.4,3.5, A.1,A.2 hold; also suppose that as n, T_0 approach infinity we have the following:*

$$\frac{T_0}{n} = c_{asp} + o(1), \quad \frac{\kappa\zeta}{\chi} = c_\zeta + o(1), \quad c_{fe} < c\sigma. \tag{2.51}$$

where $c_{asp} \in (0,1)$ and $c_\zeta > c\max\left\{1, \frac{\max\{\sigma_w, \tilde{\sigma}_y\}}{\sqrt{c_{asp}}}\right\}$. Then we have the following with probability at least $1 - c\exp\left(-c(\min\left\{\sigma_w^2, \tilde{\sigma}_y^2\right\}n\right)$:

$$\max\left\{\frac{1}{n}|\omega^\top\tilde{\theta}^{(y)}|, \frac{1}{n}|\omega^\top\theta^{(w)}|\right\} \le c\frac{\chi^2 c_\zeta + \sqrt{\chi}}{\|\Lambda^{(h),(0)}\|_{HS}}(1 + o_p(1)) \tag{2.52}$$

Proof. We have the following (the same for the second quantity):

$$|\omega^\top\theta^{(w)}| \le |\omega_{det}^\top\theta^{(w)}| + |r_{det}^\top\theta^{(w)}| + |(r^\star)^\top\theta^{(w)}| \tag{2.53}$$

Define $\sigma := \max\{\sigma_w, \tilde{\sigma}_y\}$; applying Lemmas B.3,B.4,B.5,B.6 we get the following result with probability at least $1 - c\exp\left(-c(\max\left\{\sigma_w^2, \tilde{\sigma}_y^2\right\}n\right)$:

$$\frac{1}{n}|\omega^\top\theta^{(w)}| \le c\frac{\chi^2 c_\zeta + \sqrt{\chi}}{\|\Lambda^{(h),(0)}\|_{HS}}(1 + o_p(1)) \tag{2.54}$$

□

Theorem B.2. *Suppose Assumptions 3.1,3.2,3.3,3.4,3.5,A.1,A.2, A.5,A.6 hold; also suppose that as n, T_0, T_1*

approach infinity in such a way that the following relationships hold:

$$\frac{T_0}{n} = c_{asp} + o(1), \quad \frac{\kappa\zeta}{\chi} = c_\zeta + o(1), \quad \frac{1}{n\min\{\sigma_w^2, \tilde{\sigma}_y^2\}} = o(1), \tag{2.55}$$

where $c_{asp} \in (0,1)$, *and* $\frac{c}{\chi} > c_\zeta > c\max\left\{1, \frac{\max\{\sigma_w, \tilde{\sigma}_y\}}{\sqrt{c_{asp}}}\right\}$. *Then the following is true:*

$$\begin{pmatrix} \hat{\delta} - \frac{\tau}{n}\tilde{\omega}^\top \pi \\ \hat{\pi} - \frac{1}{n}\tilde{\omega}^\top \pi \end{pmatrix} = \frac{\sqrt{\chi}\xi_{bias}}{\|\Lambda^{(h),(0)}\|_{HS}} \left(\eta_{H|Z} + \mathcal{O}_p\left(\frac{\|(\Lambda^{(h),(1)})^\top \Lambda^{(z),(1)}\|_{HS}}{\|\Lambda^{(z),(1)}\|_{HS}^2}\right)\right)(1 + o_p(1)) + \\ \Sigma_{ag}^{\frac{1}{2}}(\tilde{\omega})(\xi_{z,T_1} + o_p(1)) + \frac{\|\tilde{\omega}\|_2}{n\|\Lambda^{(z),(1)}\|_{HS}}(\xi_{cr} + o_p(1)) \tag{2.56}$$

where ξ_{bias} *is a bounded random variable independent of* ξ_{z,T_1}, ξ_{cr}.

Proof. Proof follows immediately from the previous results. □

B.3 General version of Theorem 3

Theorem B.3. *Suppose conditions of Theorem B.2 hold, and* $\max\{\sigma_w, \tilde{\sigma}_y\} = o(1)$, $\frac{T_0}{T_1} = c_{int} + o(1)$, *for* $c_{int} \in (0,1)$. *In addition suppose that the following is true for* $c_{\rho_l} < 1$, *and* $c_{int} \in (0,1)$:

$$\begin{aligned}
&\frac{1}{n}\|\omega^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2 = c\|\Lambda^{(z),(1)}\|_{HS}(1+o_p(1)),\\
&\frac{1}{n}\|\omega^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2 = c\|\Lambda^{(z),(1)}\|_{HS}(1+o_p(1)),\\
&\left|\frac{\omega^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\left(\omega^\top L^{(w),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\right)^\top}{\|\omega^\top \tilde{L}^{(y),(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2\|\omega^\top L^{(w,(1)}\Pi_f^{(1)}\Lambda_1^{(z),(1)}\|_2}\right| < c_{\rho_l}(1+o(1)),\\
&\frac{\|\Lambda^{(k),(0)}\|_{HS}}{\|\Lambda^{(k),(1)}\|_{HS}} = c_{int} + o(1),\\
&\|\hat{\Lambda}^{(z),(1)} - \Lambda^{(z),(1)}\|_{HS} = o_p\left(\|\Lambda^{(z),(1)}\|_{HS}\right),\\
&\|\hat{\Lambda}_1^{(z)} - \Lambda_1^{(z)}\|_{op} = o_p\left(\|\Lambda^{(z)}\|_{op}\right).
\end{aligned} \tag{2.57}$$

Then under $\mathbf{H}_0 : \tau = \tau_0$ *we get the following result:*

$$\mathbb{E}\left[\left\{|\hat{\delta} - \tau_0\hat{\pi}| \leq z_{\alpha/2}\hat{\sigma}(\tau_0)\right\}\right] \to 1-\alpha \tag{2.58}$$

Proof. From Theorem B.2 and assumptions we get the following result:

$$\begin{pmatrix} \hat{\delta} - \frac{\tau}{n}\omega^\top\pi \\ \hat{\pi} - \frac{1}{n}\omega^\top\pi \end{pmatrix} = \Sigma_{ag}^{\frac{1}{2}}(\omega)\xi_{z,T_1}(1+o_p(1)). \tag{2.59}$$

Define the following object (residuals):

$$\begin{aligned}
e_y &:= \frac{1}{n}\omega^\top Y^{(1)}\Pi_f^{(1)}\left(\mathcal{I}_{T_1} - \frac{1}{\|\epsilon^{(z),(1)}\|_2^2}\epsilon^{(z),(1)}(\epsilon^{(z),(1)})^\top\right)\\
e_w &:= \frac{1}{n}\omega^\top W^{(1)}\Pi_f^{(1)}\left(\mathcal{I}_{T_1} - \frac{1}{\|\epsilon^{(z),(1)}\|_2^2}\epsilon^{(z),(1)}(\epsilon^{(z),(1)})^\top\right)
\end{aligned} \tag{2.60}$$

and observe that the following is true for e_y (expansion for e_w is the same):

$$\begin{aligned}
e_y = \varepsilon_y + \frac{1}{n}\omega^\top\tilde{\theta}^{(y)}(\epsilon^{(h),(1)})^\top\left(\mathcal{I}_{T_1} - \frac{1}{\|\epsilon^{(z),(1)}\|_2^2}\epsilon^{(z),(1)}(\epsilon^{(z),(1)})^\top\right) - \\
\frac{1}{\|\epsilon^{(z),(1)}\|_2^2}\frac{1}{n}\omega^\top\tilde{L}^{(y),(1)}\Pi_f^{(1)}\epsilon^{(z),(1)}(\epsilon^{(z),(1)})^\top + \frac{1}{n}\omega^\top\tilde{E}^{(y),(1)}\Pi_f^{(1)}\left(\mathcal{I}_{T_1} - \frac{1}{\|\epsilon^{(z),(1)}\|_2^2}\epsilon^{(z),(1)}(\epsilon^{(z),(1)})^\top\right) = \\
\varepsilon_y + o_p(1) + \mathcal{O}_p(1) + o_p(1) = \varepsilon_y + r_y
\end{aligned} \tag{2.61}$$

By assumptions we have the following two bounds:

$$\left|\|\varepsilon_k\hat{\Lambda}_1^{(z),(1)}\|_2 - \|\varepsilon_k\Lambda_1^{(z),(1)}\|_2\right| \leq \|\varepsilon_k\|_2\|\hat{\Lambda}_1^{(z),(1)} - \Lambda_1^{(z),(1)}\|_{op} = o_p(\|\Lambda^{(z),(1)}\|_{HS}), \tag{2.62}$$

and

$$\begin{aligned}\left|\|e_k\hat{\Lambda}_1^{(z),(1)}\|_2^2 - \|\varepsilon_k\hat{\Lambda}_1^{(z),(1)}\|_2^2\right| \leq 2\|r_k\|_2\|\hat{\Lambda}^{(z),(1)}\|_{op}\|\varepsilon_k\hat{\Lambda}_1^{(z),(1)}\|_2 + \|r_k\|_2^2\|\hat{\Lambda}_1^{(z),(1)}\|_{op}^2 = \\ \mathcal{O}_p\left(\|\Lambda^{(z),(1)}\|_{HS}\|\Lambda^{(z),(1)}\|_{op} + \|\Lambda^{(z),(1)}\|_{op}\right) = o_p\left(\|\Lambda^{(z),(1)}\|_{HS}\right).\end{aligned} \tag{2.63}$$

Using this we arrive to the following:

$$\begin{aligned}\hat{\Sigma}(\omega) := \frac{1}{\|\hat{\Lambda}^{(z),(1)}\|_{HS}^4}\begin{pmatrix}\|e_y^\top\hat{\Lambda}_1^{(z),(1)}\|_2^2 & e_y^\top\hat{\Lambda}_1^{(z),(1)}(\hat{\Lambda}_1^{(z),(1)})^\top e_w \\ e_y^\top\hat{\Lambda}_1^{(z),(1)}(\hat{\Lambda}_1^{(z),(1)})^\top e_w & \|e_w^\top\hat{\Lambda}_1^{(z),(1)}\|_2^2\end{pmatrix} = \\ \frac{1}{\|\Lambda^{(z),(1)}\|_{HS}^4}\begin{pmatrix}\|\varepsilon_y^\top\Lambda_1^{(z),(1)}\|_2^2 & \varepsilon_y^\top\Lambda_1^{(z),(1)}(\Lambda_1^{(z),(1)})^\top\varepsilon_w \\ \varepsilon_y^\top\Lambda_1^{(z),(1)}(\Lambda_1^{(z),(1)})^\top\varepsilon_w & \|\varepsilon_w^\top\Lambda_1^{(z),(1)}\|_2^2\end{pmatrix}(1 + o_p(1)) = \Sigma(\omega)(1 + o_p(1))\end{aligned} \tag{2.64}$$

For arbitrary fixed τ_0 the following objects:

$$\begin{aligned}&\hat{\sigma}(\tau_0) = \sqrt{(1, -\tau_0)\hat{\Sigma}(\omega)(1, -\tau_0)^\top}, \\ &\sigma(\tau_0) = \sqrt{(1, -\tau_0)\Sigma(\omega)(1, -\tau_0)^\top}\end{aligned} \tag{2.65}$$

The result above tells us that $\hat{\sigma}(\tau_0) = \sigma(\tau_0)(1 + o_p(1))$. Under $\mathbf{H}_0 : \tau = \tau_0$ we get the following for the test:

$$\mathbb{E}\left[\left\{|\hat{\delta} - \tau_0\hat{\pi}| \leq z_{\alpha/2}\hat{\sigma}(\tau_0)\right\}\right] = \mathbb{E}[\{|\xi_{z,T_1}(1 + o_p(1))| \leq z_{\alpha/2}(1 + o_p(1))\}] \to 1 - \alpha, \tag{2.66}$$

and thus concluding the proof. □

C Verification

For AR(1) with coefficient ρ process we have the following:

$$\begin{aligned}&\|\Lambda^{(j),(k)}\|_{HS} = \sqrt{\frac{T_k}{(1 - \rho^2)}}(1 + o(1)) \\ &\|\Lambda^{(h),(1)}\|_{op} < \frac{\rho}{1 - \rho^2} \\ &\|(\Lambda_1^{(z),(1)})^{-1}(\Lambda_0^{(z),(1)})\nu^{(z),(0)}\|_1 = \mathcal{O}_p(1)\end{aligned} \tag{3.1}$$

Assumption A.3 does not have to hold, because the statement of all the theorems guarantees its version that is sufficient for all the results. Assumption A.1 holds as a result of all other assumptions. Assumption A.6 holds as a result of Assumption 3.7.

D Parameters of the simulation

The variance matrix of $(\epsilon_{it}^{(y)}, \epsilon_{it}^{(w)})$:

$$\Sigma = \begin{pmatrix} 0.001 & 0.000 \\ 0.000 & 0.003 \end{pmatrix} \tag{4.1}$$

The model for Z_t:

$$\begin{aligned} Z_t &= \nu_t^{(z)} + 1.14\nu_{t-1}^{(z)} + 0.52\nu_{t-2}^{(z)} \\ \nu_t^{(z)} &\sim \mathcal{N}(0, 0.43) \end{aligned} \tag{4.2}$$

The model for H_t:

$$H_t = 0.5Z_t + 0.25\tilde{Z}_t \tag{4.3}$$

where $\tilde{Z}_t$ has the same distribution as Z_t and is independent of it. Exposures $\theta_i^{(w)}$ and $\theta_i^{(y)}$ have the following form:

$$\begin{aligned} \theta_i^{(w)} &= 0.3\pi_i + \sqrt{1 - 0.3^2}\xi_i^{(w)} \\ \theta_i^{(y)} &= 0.6\pi_i + \sqrt{1 - 0.6^2}\xi_i^{(y)} \end{aligned} \tag{4.4}$$

where $\xi_i^{(w)}, \xi_i^{(y)}$ are indpendent realizations of standard normal random variables.

E Heterogeneous treatment effects

While formally our results (in particular Theorem B.2) apply for the case of constant τ they can be generalized to allow for unit-specific effects τ_i. In this case, interpretation of the resulting estimand becomes essential. We want to stress that the same question applies to the conventional TSLS algorithm from Section 2.1. While there are available results in the literature that provide interpretation of an IV-like ratio in similar setups (e.g., see appendix of Borusyak and Hull [2020]), they do not directly apply to our setting. Below we sketch an informal argument that suggests that in the low-noise regime (as in Theorem 3), our estimator converges to a convex combination of τ_i as long as some additional assumptions hold.

For arbitrary weights $\tilde{\omega}$ we have the same expansion as in Section B.1:

$$\hat{\tau} = \frac{\hat{\delta}}{\hat{\pi}} = \frac{\frac{1}{n}\tilde{\omega}^\top(\tau \circ \pi) + \text{bias}^{(y)} + \text{noise}^{(y)}}{\frac{1}{n}\tilde{\omega}^\top \pi + \text{bias}^{(y)} + \text{noise}^{(w)}} \tag{5.1}$$

Generalizing the arguments of Section B.2 we can conclude the following for the weights ω:

$$\hat{\tau} = \frac{\frac{1}{n}\omega^\top(\tau \circ \pi) + o_p(1)}{\frac{1}{n}\omega^\top \pi + o_p(1)}$$

We can further split this sum in the following way:

$$\hat{\tau} = \frac{\frac{1}{n}\omega_{det}^\top(\tau \circ \pi) + \frac{1}{n}(\omega - \omega_{det})^\top(\tau \circ \pi) + o_p(1)}{\frac{1}{n}\omega_{det}^\top \pi + \frac{1}{n}(\omega - \omega_{det})^\top \pi + o_p(1)} \tag{5.2}$$

Results of Theorem B.2 imply that $\|\omega - \omega_{det}\|_2 = o_p(\sqrt{n})$ and thus we get the following result:

$$\hat{\tau} = \frac{\frac{1}{n}\omega_{det}^\top(\tau \circ \pi) + +o_p(1)}{\frac{1}{n}\omega_{det}^\top \pi + o_p(1)}$$

Under Assumption 3.7 we get that the denominator is equal to $\frac{1}{c_\pi} \neq 0$ and thus we get the following:

$$\hat{\tau} = \tau_{det} + o_p(1) \tag{5.3}$$

where $\tau_{det} = \frac{\frac{1}{n}\omega_{det}^\top(\tau\circ\pi)}{\frac{1}{n}\omega_{det}^\top \pi}$. This implies that we need to provide the interpretation for τ_{det}.

Suppose there exist a sequence of numbers $\{r_i\}_{i=1}^n$ such that the following is true:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \omega_{i,det} r_i = o(1) \\
&\frac{1}{n}\sum_{i=1}^n \omega_{i,det} \tau_i r_i = o(1) \\
&\gamma_i := \frac{\omega_i(\pi_i - r_i)}{\frac{1}{n}\sum_{i=1}^n \omega_i(\pi_i - r_i)} \geq 0
\end{aligned} \tag{5.4}$$

Then the following holds:

$$\tau_{det} = \frac{1}{n}\sum_{i=1}^n \gamma_i \tau_i + o(1) \tag{5.5}$$

which implies that $\hat{\tau}$ converges to a convex combination of unit-specific treatment effects.

One way to guarantee that such sequence $\{r_i\}_{i=1}^n$ exists is to use constant $r_i = \overline{\pi}$ and impose an additional convex constraint in the optimization problem:

$$\omega_i(D_i - \overline{D}) \geq 0 \text{ for all } i \tag{5.6}$$

This constraint is similar in spirit to the non-negativity constraint imposed in synthetic control literature. Of course, then to guarantee a good performance of the oracle one would need to assume that there exists good balancing weights with such properties. In this case γ_i is non-negative by construction as long as Assumption 3.7 holds, the first restriction in (5.4) is satisfied by definition, and the second restriction is likely to be satisfied, because the weights $\omega_{i,det}$ have to balance $\tau_i \mu_t^{(w)}$.