cemmap
centre for microdata methods and practice

Arellano-bond lasso estimator for dynamic linear panel models

Victor Chernozhukov
Iván Fernández-Val
Chen Huang
Weining Wang

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP09/24

Economic and Social Research Council

ARELLANO-BOND LASSO ESTIMATOR FOR DYNAMIC LINEAR PANEL MODELS*

VICTOR CHERNOZHUKOV, IVÁN FERNÁNDEZ-VAL, CHEN HUANG, AND WEINING WANG†

Abstract. The Arellano-Bond estimator is a fundamental method for dynamic panel data models, widely used in practice. However, the estimator is severely biased when the data's time series dimension T is long due to the large degree of overidentification. We show that weak dependence along the panel's time series dimension naturally implies approximate sparsity of the most informative moment conditions, motivating the following approach to remove the bias: First, apply LASSO to the cross-section data at each time period to construct most informative (and cross-fitted) instruments, using lagged values of suitable covariates. This step relies on approximate sparsity to select the most informative instruments. Second, apply a linear instrumental variable estimator after first differencing the dynamic structural equation using the constructed instruments. Under weak time series dependence, we show the new estimator is consistent and asymptotically normal under much weaker conditions on T's growth than the Arellano-Bond estimator. Our theory covers models with high dimensional covariates, including multiple lags of the dependent variable, common in modern applications. We illustrate our approach by applying it to weekly county-level panel data from the United States to study opening K-12 schools and other mitigation policies' short and long-term effects on COVID-19's spread.

Keywords: Dynamic panel model, Arellano-Bond Estimator, GMM, LASSO, Debiasing

1. Introduction

Panel data involve observations collected for cross-sectional units ($i = 1, \dots, N$) over multiple time periods ($t = 1, \dots, T$). Models for panel data are commonly used in economics and other social sciences because they allow researchers to control for unobserved unit and time heterogeneity and account for unit-level dynamics. These models have multiple applications, including evaluating job training and minimum wage regulations in labor economics,

* We thank Marine Carrasco, Matt Hong and Wendun Wang for helpful comments. Wang gratefully acknowledges financial support from the ESRC (Grant Reference: ES/T01573X/1). The codes to implement the algorithms are publicly accessible via the GitHub repository: Arellano-Bond-LASSO.

† Chernozhukov: Department of Economics and Center for Statistics and Data Science, MIT; Fernández-Val: Department of Economics, Boston University; Huang: Department of Economics and Business Economics, Aarhus University; Wang: Falculty of Economics and Business, University of Groningen.

studying household consumption and economic growth in macroeconomics, estimating demand models for products in microeconomics, and analyzing payout policies and investment decisions in corporate finance. See Bond (2002) for a review of methods and applications of dynamic panel data models.

The Arellano-Bond estimator (AB) is the most widely used fundamental method for panel models (Arellano and Bond, 1991). It applies to dynamic linear models that include lagged dependent variables and predetermined covariates as explanatory variables and unobserved unit and time fixed effects. After taking first differences to remove the unit fixed effects, AB constructs moment conditions using sufficiently lagged dependent variables and covariates as instruments and applies the General Method Moment (GMM) to estimate the model parameters. However, AB might be severely biased in long panels when T is large. The problem arises because the number of moment conditions grows with the square of T, T^2, leading to many instrument bias caused by the large degree of overidentification in the GMM problem (e.g., Newey and Smith, 2004). More precisely, AB has an asymptotic bias of order T/N, which might not be negligible compared to $1/\sqrt{NT}$, the size of the stochastic error, when the time dimension T is sufficiently large relative to N (Alvarez and Arellano, 2003). This problem causes bias in estimates and undercoverage of confidence intervals.

We address the bias issue in AB estimators within long panels through a two-step method. Initially, we select the most informative moment conditions, followed by applying a linear instrumental variable estimation using instruments derived from these conditions. Specifically, as the number of AB's moment conditions varies across time periods, we perform a moment selection procedure on cross-section data for each time individually. We utilize the least absolute shrinkage and selection operator (LASSO) Tibshirani (1996) as our selector, given the naturally sparse structure of the moment conditions under appropriate weak temporal dependence conditions.

Several moment selection methods have been previously established in other contexts. For instance, Donald et al. (2009) introduced an alternative method for selecting instruments based on asymptotic mean squared error calculations. Belloni et al. (2012) described a similar approach to select optimal instruments using LASSO in cross-section instrumental variable (IV) models. Other methods for constructing optimal instruments through model averaging have been proposed by Kuersteiner and Okui (2010) and Okui (2011). Luo (2016) expanded the LASSO selector to nonlinear GMM settings with many potential moments, noting its computational advantages over Donald et al. (2009). Although some studies employ the AB estimator to underpin their analyses or conduct simulations (e.g., Newey and Windmeijer, 2009), none address the AB estimator directly, mainly because they assume temporal independence – an assumption invalidated in the AB context by the introduction of temporal dependencies through first differences to remove unit fixed effects.

LASSO utilizes the ℓ_1-norm to select the moment conditions. To show the validity of the selector, we build on theoretical results achieving near-oracle rates for LASSO and related estimators in Bickel et al. (2009), Belloni et al. (2011) and Belloni and Chernozhukov (2013). An additional complication comes from the high dimensionality of the problem. As we mentioned above, the number of moment conditions grows with T^2. Moreover, we allow for high dimensional covariates including multiple lags of the dependent variable and strictly exogenous covariates, which are common in modern applications, and unit and time fixed effects, whose number grows with the two dimensions of the panel. While our method does not suffer from shrinkage and model selection biases because the moment conditions of the second step are Neyman-orthogonal with respect to the parameters estimated in the first step, it is still subject to over-fitting bias. We deal with this problem by combining the two steps of our procedure using cross-fitting (Chernozhukov et al., 2018). Thus, we partition the panel in two parts. We select the moment conditions in the first part and estimate the parameters in the second part. Then, we repeat the procedure reversing the roles of the two parts and aggregate the results by averaging the estimates of the model parameters from the two orderings. We show theoretically that this procedure removes asymptotic over-fitting bias and improves small sample properties over the method that does not use cross-fitting in various simulation settings.

There is an extensive recent literature on panel data with large T, including dynamic linear models. Alvarez and Arellano (2003) studied the properties of AB and other estimators in long panels. They showed that AB exhibits asymptotic bias when T/N tends to a constant. Okui (2009) proposed a method to select instruments by characterizing the mean squared error of the one-step AB estimator that uses the matrix of second moments of the instruments as weighting matrix in models with homoskedastic errors and strictly exogenous regressors. In the same setting, Carrasco and Nayihouba (2024) developed a version of the one-step AB estimator that regularizes the weighting matrix building on Carrasco (2012). Our instrument selection method is different and applies to both one-step and two-step AB estimators. It also does not rely on homeskedastic errors and allows for weakly exogenous covariates. Chen et al. (2019) developed a debiasing method for AB based on applying the split-panel idea of Dhaene and Jochmans (2015) to the cross-section dimension of the panel. An alternative to debiasing to deal with the many moments problem of GMM is the use of generalized empirical likelihood (GEL) estimators (Newey and Smith, 2004; Newey and Windmeijer, 2009). While GEL estimators have theoretical advantages over GMM, they are not commonly used in practice due to their computational complexity. In that sense, our primary objective is to develop an easily implementable approach for AB in moderately long panels. LASSO methods have also been combined with GMM and GEL in high dimensional models, but such combinations have not been explored in panel settings. Examples include Chang et al. (2015); Shi (2016); Chang et al. (2021), which considered relaxed/penalized empirical likelihood methods to address the issue of many moment conditions with a fixed or diverging number of model parameters. We refer to Belloni et al. (2018) for a recent review on high dimensional methods in GMM settings. Finally, Kock and Tang (2019) and Semenova et al. (2023) applied LASSO methods to dynamic panel models with sparsity in the unit fixed effects.

We make three main theoretical contributions. First, we demonstrate that the moment conditions of AB have an approximately sparse structure under suitable temporal dependence conditions and propose a LASSO version of AB that fruitfully exploits this structure. In particular, the effective dimension of the non-zero coefficients in the first step estimation of the selected instruments at each time period t is the minimum of $\log N$ and t, which is very "low" relative to the cross-sectional size N. Second, we propose a cross-fitting procedure based on sample-splitting, which removes the dependence between the generated errors of the selected instruments and the errors in the main regression by conditioning on the sub-sample. This enables us to remove the over-fitting bias and achieve a more favorable convergence rate of the estimator under simple regularity conditions. It also opens the door to the use of machine learning methods other than LASSO to select the most informative moment conditions. Third, we consider models with high dimensional covariates. Here, we achieve selection of moment conditions and covariates simultaneously by constructing orthogonal conditions and employing regularized GMM with a sparse weighting matrix via Dantzig selector, in order to partial out the effect of high dimensional nuisance parameters.

We apply our proposed method to study the effect of the opening of K-12 schools and other policies on the spread of COVID-19 using a panel of $2,510$ US counties over 32 weeks, extracted from the dataset used in Chernozhukov et al. (2021b). We estimate a panel regression model with rich dynamics, incorporating four lags of the dependent variable and several predetermined covariates. Due to the large number of instruments ($m = 3,402$), the small bias condition for AB of Chen et al. (2019), $m^2/(NT) \to 0$, fails for the AB estimator ($m^2/(NT) \approx 170$).[1] Using our proposed method, we find that policies such as the opening of K-12 schools, college visits, mask mandates, and stay at home orders have larger short-run effects but smaller long-run effects than those estimated using AB. The source of the difference in this case is bias in the estimation of the autoregressive coefficients.

[1]In this calculation $T = 27$ because the first 5 observations are used as initial conditions of the dynamic model.

Notation. For a vector $v = (v_1, ..., v_d)^\top \in \mathbb{R}^d$ and a constant $r \geq 1$, we denote $|v|_r = (\sum_{i=1}^d |v_i|^r)^{1/r}$ and $|v|_\infty = \max_{1\leq i\leq d} |v_i|$. Define $|v|_0$ as the zero norm, i.e. the number of non-zero coordinates. For a matrix $A = (a_{ij})_{1\leq i\leq m, 1\leq j\leq n}$, we define $|A|_{\max} = \max_{1\leq i\leq m, 1\leq j\leq n} |a_{ij}|$, $|A|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|$, $|A|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}|$, and $|A|_{1,1} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$. For a random variable X_{it}, we say $X_{it} \in \mathcal{L}^r$ if $\|X_{it}\|_r \stackrel{\text{def}}{=} (\mathsf{E}\,|X_{it}|^r)^{1/r} < \infty$ for some $r > 0$, and define the sub-Gaussian norm as $\|X_{it}\|_{\psi_{1/2}} = \inf\{s > 0 : \mathsf{E}\exp(X_{it}^2/s^2) \leq 2\}$, where E denotes the expectation conditional on the unit and time effects. We denote cross-section averages of X_{it} by $\bar{\mathsf{E}}$, that is $\bar{\mathsf{E}}X_{it} = \lim_{N\to\infty} N^{-1} \sum_{i=1}^N \mathsf{E}\, X_{it}$. Given two sequences of positive numbers a_n and b_n, we write $a_n \lesssim b_n$ (resp. $a_n \asymp b_n$) if there exists $C > 0$, which does not depend on n, such that $a_n/b_n \leq C$ (resp. $1/C \leq a_n/b_n \leq C$) for all large n. For a sequence of random variables X_n, we use the notation $X_n \lesssim_{\mathrm{P}} b_n$ to denote $X_n = \mathcal{O}_{\mathrm{P}}(b_n)$. For two real numbers, set $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$.

Outline. The rest of the paper is organized as follows. Section 2 introduces the model and estimators. Section 3 presents the main theoretical results. Sections 4 and 5 report the results of the simulation study and empirical application, respectively. Section 6 contains some concluding remarks. Appendix collects the deferred proofs of the theoretical results and supplementary tables with simulation results.

2. Model and Estimators

2.1. **Basic Model.** Let $\{(Y_{it}, D_{it}, C_{it}) : 1 \leq i \leq N, 1 \leq t \leq T\}$ be a panel dataset, where i and t index unit and time period, respectively. Y_{it} is a scalar outcome or response variable, D_{it} is the policy variable or treatment of interest, and C_{it} is a vector of covariates of fixed dimension including, for example, $Y_{i,t-1}$ and other treatments. To measure the effect of D_{it} on Y_{it}, we consider a dynamic linear panel model:

$$Y_{it} = \alpha_i + \gamma_t + X_{it}^\top \theta^0 + \varepsilon_{it}, \quad X_{it} := (D_{it}, C_{it}^\top)^\top, \quad 1 \leq t \leq T, \tag{1}$$

where θ^0 is the parameter of interest, α_i is an unobserved unit effect, γ_t is an unobserved time effect, and ε_{it} is a zero-mean idiosyncratic error. We might also be interested in functions of θ^0 such as long-run effects in dynamic models that include lags of the dependent variable as covariates. We refer to the empirical application in Section 5 for an example.

In the theoretical analysis, we shall treat the unobserved unit and time effects as fixed parameters. This is equivalent to conditioning on the realization of all these effects.[2] We

[2]Due to this conditioning, all probability statements should be qualified with almost surely. We shall omit this qualifier for notational convenience.

assume that $\{(D_{it}, C_{it}, \varepsilon_{it}) : 1 \leq t \leq T\}$ are independent over i, and ε_{it} is an uncorrelated sequence over t. In addition, we assume that the treatment and covariates in X_{it} are predetermined with respect to ε_{it} in the sense that

$$\mathsf{E}(X_{is}\varepsilon_{it}) = 0, \text{ for all } 1 \leq s \leq t,\ 1 \leq t \leq T.$$

We remove the unobserved effects by taking first differences over time and demeaning all the variables at the unit level, namely

$$\Delta\widetilde{Y}_{it} = \Delta\widetilde{X}_{it}^{\top}\theta^0 + \Delta\widetilde{\varepsilon}_{it}, \quad 2 \leq t \leq T, \tag{2}$$

where $\Delta\widetilde{Z}_{it} = \Delta Z_{it} - \sum_{j=1}^{N}\Delta Z_{jt}/N$ and $\Delta Z_{it} = Z_{it} - Z_{i,t-1}$, for $Z_{it} \in \{Y_{it}, X_{it}, \varepsilon_{it}\}$. The transformed error, $\Delta\widetilde{\varepsilon}_{it}$, satisfies the moment conditions

$$\mathsf{E}(X_{is}\Delta\widetilde{\varepsilon}_{it}) = 0, \text{ for all } 1 \leq s \leq t-1,\ 2 \leq t \leq T.$$

AB uses these moment conditions to construct a GMM estimator of θ^0. It should be noted that AB is biased when T is large due to the large number of moment conditions, i.e. $m = (T-2)(T-1)/2 = \mathcal{O}(T^2)$; see, e.g., Newey and Smith (2004) for more discussion. We propose an alternative estimator that is computationally simple and has lower bias when T is large. It is based on the application of LASSO to select the most informative moment conditions to estimate the parameters. Thus, the estimator has two stages. It first selects moment conditions using LASSO, and then estimates the parameters of interest by instrumental variables, with the predicted values of the endogenous regressors obtained from the selected moment conditions serving as instruments. We name the new estimator as AB-LASSO as a shortcut for Arellano-Bond LASSO estimator.

Definition 2.1 (AB-LASSO)**.** *The AB-LASSO estimator consists of two steps:*

1 For $t = 2, \ldots, T$ and $W_{it} \in \Delta\widetilde{X}_{it}$, run the LASSO regressions:

$$\widehat{\Pi}_t \stackrel{\text{def}}{=} (\widehat{\pi}_{t0}, \widehat{\pi}_{t1}^{\top}, \ldots, \widehat{\pi}_{t,t-1}^{\top})^{\top} \in \arg\min_{\pi_{t0},\ldots,\pi_{t,t-1}} \left\{ \sum_{i=1}^{N}\left(W_{it} - \pi_{t0} - \sum_{s=1}^{t-1} X_{is}^{\top}\pi_{ts}\right)^2 + \lambda_t \sum_{s=1}^{t-1}\omega_{ts}|\pi_{ts}|_1 \right\}, \tag{3}$$

where λ_t is a penalty tuning parameter and ω_{ts} is a non-negative weight that is a non-increasing function of $t-s$, e.g., $\omega_{ts} = 1$. Obtain the predicted values of the previous regression, $\widehat{W}_{it} \in \widehat{\Delta\widetilde{X}}_{it}$,

$$\widehat{W}_{it} = \widehat{\pi}_{t0} + \sum_{s=1}^{t-1} X_{is}^{\top}\widehat{\pi}_{ts}.$$

2 Estimate (2) by instrumental variable regression using $\widehat{\Delta \widetilde{X}_{it}}$ as the instrument for $\Delta \widetilde{X}_{it}$, that is

$$\widehat{\theta} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T} \widehat{\Delta \widetilde{X}_{it}} \Delta \widetilde{X}_{it}^{\top} \right)^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} \widehat{\Delta \widetilde{X}_{it}} \Delta \widetilde{Y}_{it}. \tag{4}$$

Remark 2.1 (Initial Conditions)**.** We have implicitly assumed so far that the initial conditions of Y_{it} are observed in models that include lags of the dependent variable as covariates. For example, we have assumed that Y_{i0} is observed when C_{it} includes $Y_{i,t-1}$. If Y_{it} is first observed at $t = 1$, then the vector X_{is} in (3) needs to be modified to include only the observed values of C_{is}. In models where $C_{it} = Y_{i,t-1}$, for example, $X_{i1} = D_{i1}$ instead of $X_{i1} = (D_{i1}, Y_{i0})$.

Remark 2.2 (Neyman-Orthogonality)**.** Let $V_{it} = (1, X_{i1}^{\top}, \dots, X_{i,t-1}^{\top})^{\top}$ and $\Pi_t = (\pi_{t0}, \pi_{t1}^{\top}, \dots, \pi_{t,t-1}^{\top})^{\top}$. The estimator given in (4) is a moment estimator with moment function:

$$g_i(\theta, \Pi_2, \dots, \Pi_T) = \sum_{t=2}^{T} \Pi_t^{\top} V_{it} (\Delta \widetilde{Y}_{it} - \Delta \widetilde{X}_{it}^{\top} \theta).$$

This moment function is Neyman-orthogonal with respect to each of the first stage parameters Π_t, $t = 2, \dots, T$, because

$$\left. \frac{\partial \, \mathsf{E}[g_i(\theta, \Pi_2, \dots, \Pi_T)]}{\partial \Pi_t} \right|_{\theta=\theta^0} = \mathsf{E}[V_{it}(\Delta \widetilde{Y}_{it} - \Delta \widetilde{X}_{it}^{\top} \theta^0)] = 0, \quad t = 2, \dots, T.$$

Note that the 2SLS (Two-Stage Least Squares) version of the second stage that replaces $\Delta \widetilde{X}_{it}$ by $\widehat{\Delta \widetilde{X}_{it}}$ in (4) does not satisfy this condition.

AB is an instrumental variable estimator. Its bias comes from overfitting because the same observations are used to project the endogenous regressors on the instruments and to estimate the parameters (Phillips and Hale, 1977; Angrist and Krueger, 1995; Angrist et al., 1999). The order of the bias is m/n, where m is the number of moment conditions and n is the sample size. In the case of AB, $m = \mathcal{O}(T^2)$ and $n = NT$, so that the order of the bias is T/N. The order of the sampling noise is $n^{-1/2} = (NT)^{-1/2}$, so that the small bias condition of Chen et al. (2019) is $m/n^{1/2} \to 0$ or equivalently $m^2/n = T^3/N \to 0$. AB-LASSO reduces the overfitting bias by selecting moment conditions. Up to logarithmic terms, the small bias condition for AB-LASSO becomes $\max_{2 \le t \le T} s_t^* \sqrt{T/N} \to 0$ (s_t^* is the dimension of effective instruments for each t), see Remark 3.2. When T is moderately large relative to N, AB-LASSO might still exhibit bias. The temporal dependence of the data can further exacerbate the problem in finite samples. Indeed, we observe a more severe bias, compared to the i.i.d. case studied in Belloni et al. (2012), in numerical simulations. To reduce this bias, we develop a sample-splitting procedure over the cross-section dimension following the idea of the split-sample IV estimator of Angrist and Krueger (1995). We name the version of AB-LASSO with sample splitting and cross-fitting as AB-LASSO-SS.

Definition 2.2 (AB-LASSO-SS)**.** *The AB-LASSO-SS estimator consists of the following steps:*

1 *Partition the sample $\{(Y_{it}, D_{it}, C_{it}) : 1 \leq i \leq N, 1 \leq t \leq T\}$ along the cross-section dimension into two parts or sub-samples A and B, corresponding to the indexes $i \in \{1, \dots, \lfloor N/2 \rfloor\} =: \mathbb{I}_A$ and $i \in \{\lfloor N/2 \rfloor + 1, \dots, N\} =: \mathbb{I}_B$, where $\lfloor \cdot \rfloor$ denotes the integer part.*

2 *Take first differences over time and demean at the unit level all the variables in each sub-sample, namely $\Delta \widetilde{Z}_{it,s} = \Delta Z_{it} - \sum_{j \in \mathbb{I}_s} \Delta Z_{jt}/|\mathbb{I}_s|$, $i \in \mathbb{I}_s$, $s \in \{A, B\}$, and $\Delta Z_{it} = Z_{it} - Z_{i,t-1}$, for $Z_{it} \in \{Y_{it}, X_{it}\}$.*

3 *For $t = 2, \dots, T$ and $W_{it} \in \Delta \widetilde{X}_{it,A}$, run step 1 of AB-LASSO in sub-sample A, that is estimate the LASSO regressions:*

$$\widehat{\Pi}_{t,A} \stackrel{\text{def}}{=} (\widehat{\pi}_{t0,A}, \widehat{\pi}_{t1,A}^\top, \dots, \widehat{\pi}_{t,t-1,A}^\top)^\top \in \arg \min_{\pi_{t0}, \dots, \pi_{t,t-1}} \left\{ \sum_{i \in \mathbb{I}_A} \left(W_{it} - \pi_{t0} - \sum_{s=1}^{t-1} X_{is}^\top \pi_{ts} \right)^2 + \lambda_t \sum_{s=1}^{t-1} \omega_{ts} |\pi_{ts}|_1 \right\}, \tag{5}$$

where λ_t is a penalty tuning parameter and ω_{ts} is a non-negative weight that is a non-increasing function of $t - s$, e.g., $\omega_{ts} = 1$. Obtain the predicted values in sub-sample B using the previous estimates from sub-sample A, $\widehat{W}_{it,BA} \in \widehat{\Delta \widetilde{X}_{it,BA}}$,

$$\widehat{W}_{it,BA} = \widehat{\pi}_{t0,A} + \sum_{s=1}^{t-1} X_{is}^\top \widehat{\pi}_{ts,A}, \quad i \in \mathbb{I}_B.$$

Run the second step of AB-LASSO in sub-sample B using the instruments $\widehat{\Delta \widetilde{X}_{it,BA}}$,

$$\widehat{\theta}_{B,A} = \left(\sum_{i \in \mathbb{I}_B} \sum_{t=2}^{T} \widehat{\Delta \widetilde{X}_{it,BA}} \Delta \widetilde{X}_{it,B}^\top \right)^{-1} \sum_{i \in \mathbb{I}_B} \sum_{t=2}^{T} \widehat{\Delta \widetilde{X}_{it,BA}} \Delta \widetilde{Y}_{it,B}. \tag{6}$$

4 *Run step 3 reversing the roles of sub-samples A and B to obtain $\widehat{\theta}_{A,B}$.*

5 *Compute the cross-fitting estimator of θ^0 as the average of the estimators in the two orderings*

$$\widehat{\theta}_{SS} = (\widehat{\theta}_{A,B} + \widehat{\theta}_{B,A})/2. \tag{7}$$

Remark 2.3 (K-Fold and Multiple Splitting)**.** The above cross-fitting procedure can be further generalized with K-fold sample splitting (e.g. $K = 5$). Each of the K sub-samples is used as the main sample for estimating (6) while the rest form the auxiliary sample to fit the LASSO estimate in (5). The resulting K estimates corresponding to the different partitions are averaged. The cross-section mean-difference is taken within the main and auxiliary samples. Moreover, since the ordering of the cross-section units is arbitrary by the independence assumption, we repeat the procedure for multiple splits by randomly permuting the index i across units and aggregate the estimates by averaging across permutations. The use of multiple sample splits makes the estimator invariant to the ordering of the cross-section units.

Remark 2.4 (Comparison with SSIV)**.** AB-LASSO-SS has two main differences with respect to the split-sample IV (SSIV) estimator of Angrist and Krueger (1995) applied to a dynamic panel model. First, we use LASSO instead of ordinary least squares (OLS) in the first step to project the endogenous regressors on the instruments. In numerical simulations, we find that using OLS in the first stage produces biased estimators of the parameters of interest, even when we combine it with sample splitting, see Tables B.3 and B.4 in the Appendix. Second, we use cross-fitting to improve efficiency and multiple sample splits to avoid the temptation of data mining.

2.2. **General Model with Many Exogenous Covariates.** The basic model in (1) can be extended by including additional covariates:

$$Y_{it} = \alpha_i + \gamma_t + X_{it}^\top \theta^0 + \varepsilon_{it}, \quad X_{it} := (D_{it}, C_{it}^\top, X_{2,it}^\top)^\top, \tag{8}$$

where $X_{2,it}$ is a possibly high-dimensional vector of covariates that is independent over i and satisfies

$$\mathsf{E}(\Delta X_{2,it} \Delta \varepsilon_{it}) = 0.$$

Examples of such covariates include strictly exogenous variables with respect to ε_{it} and lagged predetermined covariates such as second and higher lags of the dependent variable. Denote the dimension of $X_{2,it}$ by d_2. The high-dimensional case arises when d_2 is large relative to the sample size $n = NT$ such that it is more appropriate to treat d_2 as increasing in the asymptotic analysis. For this case, we propose a debiasing procedure to partial out the effect of $X_{2,it}$.

Remark 2.5 (Strictly Exogenous Covariates)**.** If $X_{2,it}$ includes strictly exogenous covariates, then there are additional moment conditions that can be used to estimate θ^0. In particular,

$$\mathsf{E}(X_{2,is}^{se} \Delta \widetilde{\varepsilon}_{it}) = 0, \text{ for all } 1 \leq s \leq T, \quad 2 \leq t \leq T,$$

where $X_{2,is}^{se} \subseteq X_{2,it}$ is the subset of strictly exogenous covariates of $X_{2,it}$. These additional moment conditions can be incorporated to step 1 of AB-LASSO.

To explain the partialling-out procedure, it is convenient to rewrite the extended model (8) as:

$$Y_{it} = \alpha_i + \gamma_t + X_{1,it}^\top \theta_1^0 + X_{2,it}^\top \theta_2^0 + \varepsilon_{it}, \quad X_{1,it} := (D_{it}, C_{it}^\top)^\top,$$

where $\theta_1^0 \in \mathbb{R}^{d_1}$ and $\theta_2^0 \in \mathbb{R}^{d_2}$. Denote $d = d_1 + d_2$, where d_1 is fixed and d_2 is growing with n. Assume the sparsity assumption $|\theta_2^0|_0 = o(n)$. The moment functions are given by

$$g_{it}(\theta_1, \theta_2) = \mathsf{E}\left\{(\Delta\widetilde{Y}_{it} - \Delta\widetilde{X}_{1,it}^\top \theta_1 - \Delta\widetilde{X}_{2,it}^\top \theta_2)U_{it}\right\},$$

where $U_{it} = (U_{it}^{0\top}, \Delta\widetilde{X}_{2,it}^\top)^\top$, U_{it}^0 $(d_1 \times 1)$ contains the most informative IVs for $\Delta\widetilde{X}_{1,it}$.

For each t, we wish to construct instruments for $\Delta\widetilde{X}_{1,it}$ from U_{it} that are orthogonal to $\Delta\widetilde{X}_{2,it}$. We can achieve this goal by finding a weighting matrix $\mathcal{W}_t$ $(d \times d_1)$ such that

$$\mathsf{E}(\Delta\widetilde{X}_{2,it}U_{it}^\top)\mathcal{W}_t = 0,$$

and $\mathsf{E}(\Delta\widetilde{X}_{1,it}U_{it}^\top)\mathcal{W}_t$ is of the rank d_1. This problem can be solved by Dantzig selector:

$$\begin{aligned} &\widehat{\mathcal{W}}_t = \arg\min_{\mathcal{W}_t} |\mathcal{W}_t|_{1,1} \quad \text{subject to} \\ &\left| N^{-1}\sum_{i=1}^N \left\{ \begin{pmatrix} \Delta\widetilde{X}_{1,it} \\ \Delta\widetilde{X}_{2,it} \end{pmatrix} \widehat{U}_{it}^\top \right\} \mathcal{W}_t - \mathbf{I}_{d\times d_1} \right|_{\max} \le \ell_t, \end{aligned} \tag{9}$$

where we have replaced U_{it}^0 by the LASSO predictions, i.e. $\widehat{U}_{it} = (\widehat{\Delta\widetilde{X}_{1,it}}^\top, \Delta\widetilde{X}_{2,it}^\top)^\top$, and $\mathbf{I}_{d\times d_1}$ represents the $d \times d_1$ sub-matrix of the $d \times d$ identity matrix. Then, the instrument for $\Delta\widetilde{X}_{1,it}$ is $\widehat{\mathcal{W}}_t^\top \widehat{U}_{it}$ and the estimator of the parameters of interest becomes:[3]

$$\widehat{\theta}_1 = \left(\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta\widetilde{X}_{1,it}^\top\right)^{-1}\left(\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta\widetilde{Y}_{it}\right). \tag{10}$$

The low-dimensional case arises when $X_{2,it}$ has few components such that we can treat d_2 as fixed. In this scenario, regularization is not necessary when solving the weighting matrix $\mathcal{W}_t$, and the constraint in (9) becomes binding with $\ell_t = 0$. Specifically, to partial out the effect of the nuisance parameters that are not of interest, we can choose $\mathcal{W}_t$ in the form of a co-projection matrix to ensure the orthogonality holds. We will not delve further into discussing this case because it can be treated by partialling out $X_{2,it}$ together with the unit and time effects in (2) using linear projections. In particular, we can define $\Delta\widetilde{Z}_{it}$ as the residuals of the linear regression of ΔZ_{it} on $X_{2,it}$ and a set of indicators for the time effects.

3. Main Theorems

In this section, we present the theoretical foundation of the proposed estimator. As a special case of the general model in (8), we gain insights into the extended model by exploiting the specific structure within the basic model (1). We will start by demonstrating some results related to the basic model. In particular, we will first address the model without the time

[3]When X_{2it} includes second or higher lags of the dependent variable the summation over t in (10) needs to be modified to include only the observed values of $\widehat{U}_{it}$. See Remark 2.1 for a related discussion.

effect γ_t and then we discuss how the theory adapts in presence of γ_t. Throughout this section, we impose the following conditions on the data generating processes.

Assumption 3.1 (Data Generating Processes). *The processes $X_{it} \in \mathbb{R}^d$ and ε_{it} are stationary over t and i.i.d. over i, conditional on unobserved individual and time effects, and admit the representations: $X_{it} = F_{it}(\dots, \xi_{i,t-1}, \xi_{it})$ and $\varepsilon_{it} = g_{it}(\dots, \zeta_{i,t-1}, \zeta_{it})$, where $F_{it}(\cdot) = (f_{it,1}(\cdot), \dots, f_{it,d}(\cdot))^\top$, F_{it} and g_{it} are measurable functions, and ξ_{it}, ζ_{it} for $t \in \mathbb{Z}, i \in \mathbb{N}$, are i.i.d. random elements.*[4]

We allow for overlap in the innovations ξ_{it} and ζ_{it}, as long as the exogeneity conditions specified in Section 2 are satisfied, i.e., $\mathsf{E}(X_{is}\varepsilon_{it}) = 0$, for all $1 \leq s \leq t$. The following definition, along with Assumptions 3.1 and 3.2(i) below, adapt the functional dependence measure proposed by Wu (2005) for stationary time series processes to heterogeneous panel data processes.

Definition 3.1 (Dependence Adjusted Norm). *For each $k = 1, \dots, d$, let*

$$X^*_{it,k}(\ell) = f_{it,k}(\dots, \xi^*_{i,t-\ell}, \dots, \xi_{it}),$$

where $\xi_{i,t-\ell}$ is replaced by an i.i.d. copy $\xi^*_{i,t-\ell}$. For $r \geq 1$, define the functional dependence measure $\delta_{it,k,r}(\ell) \stackrel{\text{def}}{=} \|X^*_{it,k}(\ell) - X_{it,k}\|_r$, which measures the dependency of $\xi_{i,t-\ell}$ on $X_{it,k}$. Additionally, define $\Delta_{k,r,m} \stackrel{\text{def}}{=} \max\limits_{1\leq i\leq N, 1\leq t\leq T} \sum_{\ell=m}^{\infty} \delta_{it,k,r}(\ell)$, which measures the cumulative effects for all $\ell \geq m$ and is uniform over i and t. Moreover, the dependence adjusted norm of $X_{it,k}$ is introduced by $\|X_{\cdot,k}\|_{r,\varsigma} \stackrel{\text{def}}{=} \sup_{m\geq 0}(m+1)^{\varsigma}\Delta_{k,r,m}$, where $\varsigma > 0$.

Assumption 3.2 (Data Generating Processes, Continued). *(i) For each $k = 1, \dots, d$, assume that $\|X_{\cdot,k}\|_{r,\varsigma} < \infty$ for some $r \geq 4$, $\varsigma > 0$, and*

$$\|X_{\cdot,k}\|_{\psi_\nu,\varsigma} \stackrel{\text{def}}{=} \sup_{r\geq 2} r^{-\nu}\|X_{\cdot,k}\|_{r,\varsigma} < \infty, \text{ for some } \nu \geq 0, \varsigma > 0.$$

Specifically, $\|X_{\cdot,k}\|_{\psi_\nu,\varsigma}$ is the dependence adjusted sub-Gaussian or sub-exponential norm, with ν taking values of 1/2 or 1, respectively.

(ii) ε_{it} is a martingale difference sequence (m.d.s.) over t, with respect to the filtration $\mathcal{F}_{it} = \{(X_{is})_{s=1}^t, (Y_{is})_{s=1}^{t-1}\}$, i.e. $\mathsf{E}(\varepsilon_{it} \mid \mathcal{F}_{it}) = 0$. Analogous assumption to part (i) for $X_{it,k}$ holds for ε_{it}.

(iii) The sub-Gaussian norms $\max\limits_{1\leq k\leq d}\|X_{it,k}\|_{\psi_{1/2}} < \infty$, and $\|\varepsilon_{it}\|_{\psi_{1/2}} < \infty$, for all $i = 1, \dots, N, t = 1, \dots, T$.

[4]It is worth noting that in Assumption 3.1, we assume that conditional on the individual and time effects $\{\alpha_1, \dots, \alpha_N, \gamma_1, \dots, \gamma_T\}$ is equivalent to conditional on $\{\alpha_i, \gamma_t\}$. As a result, in Definition 3.1, the functional dependence measure $\delta_{it,k,r}(\ell)$ is a random function of α_i and γ_t, and we shall define $\Delta_{k,r,m}$ (as well as the dependence adjusted norm) as a uniform measure over i and t.

Example 3.1 provides an heterogeneous linear process that satisfies Assumptions 3.2(i).

Example 3.1 (Heterogeneous Linear Process)**.** Assume that X_{it} is univariate. For each $i = 1, \ldots, N$, consider the linear process:

$$X_{it} = \sum_{\ell \geq 0} a_{i\ell} \xi_{i,t-\ell}, \quad 1 \leq t \leq T,$$

where the coefficients $a_{i\ell}$ can be heterogeneous over i and ℓ, and satisfy $|a_{i\ell}| \leq |c|^{\ell}$, for some $|c| < 1$ and for all i and ℓ. The unobservable ξ_{it}'s are sub-Gaussian random variables that are i.i.d. over i and t, and have finite rth-moment for some $r \geq 4$. It follows that

$$\begin{aligned}
\delta_{it,r}(\ell) &= \|X^*_{it}(\ell) - X_{it}\|_r = \|a_{i\ell}\xi^*_{i,t-\ell} - a_{i\ell}\xi_{i,t-\ell}\|_r = |a_{i\ell}|\|\xi^*_{i,t-\ell} - \xi_{i,t-\ell}\|_r, \\
\Delta_{r,m} &= \max_{1 \leq i \leq N, 1 \leq t \leq T} \sum_{\ell \geq m} \delta_{it,r}(\ell) \leq \max_{1 \leq i \leq N, 1 \leq t \leq T} \sum_{\ell \geq m} |c|^{\ell} \|\xi^*_{i,t-\ell} - \xi_{i,t-\ell}\|_r \propto |c|^m, \\
\|X_{\cdot}\|_{r,\varsigma} &= \sup_{m \geq 0} (m+1)^{\varsigma} \Delta_{r,m} < \infty.
\end{aligned}$$

In the special case of a heterogenous over i and stationary over t AR(1) process:

$$X_{it} = \beta_i X_{i,t-1} + \xi_{it}, \quad |\beta_i| < 1,$$

we have $a_{it} = \beta_i^t$, and $\Delta_{r,m} \propto \max_{1 \leq i \leq N} |\beta_i|^m$.

The m.d.s. condition in Assumption 3.2(ii) aligns with the standard large T panel literature; see, for example, Alvarez and Arellano (2003) and Arellano (2003, page 145). However, it's noteworthy that our theoretical analysis does not heavily rely on this assumption. With some tedious but standard techniques, such as those demonstrated in Chen et al. (2022), we could generalize the setting and achieve the same convergence rate without requiring any additional theoretical insights.

Additionally, we note that for practitioners, the sub-Gaussian conditions in Assumption 3.2(iii) rule out heavy-tail distributions for X_{it} and ε_{it}. However, this assumption is not critical for our analysis. They can be relaxed to a polynomial tail conditions with more demanding rate assumptions.

3.1. **Basic Model: Consistency of Step 1.** We will first demonstrate the consistency property of the LASSO estimator $\widehat{\Pi}_t$, which is obtained in step 1 of AB-LASSO by (3). For this purpose, a few definitions and assumptions are introduced as follows.

For each $W_{it} \in \Delta X_{it}$, denote the $N \times 1$ vector of $(W_{it})_{i=1}^N$ by $\boldsymbol{W}_t$. Recall that $V_{it} = (1, X_{i1}^{\top}, \ldots, X_{i,t-1}^{\top})^{\top}$. Denote the dimension of V_{it} as m_t, which is the number of instruments for each time period t. We further stack $V_{it}^{\top}$ by rows for all $i = 1, \ldots, N$, to create the

$N \times m_t$ matrix $\boldsymbol{V}_t$. For each $t = 2, \ldots, T$, define the true model for step 1:

$$\boldsymbol{W}_t = \boldsymbol{V}_t \Pi_t^0 + \boldsymbol{\eta}_t,$$

with $\Pi_t^0 \stackrel{\text{def}}{=} \bar{\mathsf{E}}(V_{it}V_{it}^\top)^{-1}\bar{\mathsf{E}}(V_{it}W_{it})$, and $\boldsymbol{\eta}_t$ is an $N \times 1$ vector of the i.i.d. errors $(\eta_{it})_{i=1}^N$. Assumptions 3.1-3.2 guarantee that each component in V_{it} satisfies the finite moment conditions $\mathsf{E}\,|V_{it,k}|^{2r} < \infty$ and $\mathsf{E}\,|V_{it,k}\eta_{it}|^r < \infty$, for some $r \geq 2$, where $k = 1, \ldots, m_t$, $i = 1, \ldots, N$, $t = 2, \ldots, T$.

In addition to the true model, given $0 < s_t^* \leq m_t$, we consider a sparse approximation $\Pi_t^* = \arg\min_{|\Pi_t|_0 \leq s_t^*} \bar{\mathsf{E}}|V_{it}^\top(\Pi_t - \Pi_t^0)|^2$. The oracle order of s_t^* is determined by the degree of temporal dependency in the data, which we will discuss in specific cases later. To quantify the approximation error of Π_t^* with respect to Π_t^0 empirically, we consider the prediction norm defined by

$$|\Pi_t^* - \Pi_t^0|_{2,N} \stackrel{\text{def}}{=} N^{-1/2}|\boldsymbol{V}_t(\Pi_t^* - \Pi_t^0)|_2 = \Big[\frac{1}{N}\sum_{i=1}^N \{V_{it}^\top(\Pi_t^* - \Pi_t^0)\}^2\Big]^{1/2} =: C_{s_t^*}. \tag{11}$$

We shall express all the general rate of convergence results in terms of $C_{s_t^*}$ and provide bounds for $C_{s_t^*}$ in specific examples (see Appendix A.2.2).

Let $\Pi_{t,k}^*$ be the k-th element of Π_t^*, $k = 1, \ldots, m_t$. Define the indices sets $J_t \stackrel{\text{def}}{=} \{k \in \{1, \ldots, m_t\} : \Pi_{t,k}^* \neq 0\}$ and $J_t^c \stackrel{\text{def}}{=} \{k \in \{1, \ldots, m_t\} : \Pi_{t,k}^* = 0\}$. For any $\delta_t \in \mathbb{R}^{m_t}$, let $J_{t,0} \subseteq \{1, \ldots, m_t\}$ be a set of indices with cardinality $|J_{t,0}| \leq s_t^*$, and let $J_{t,1} \subseteq \{1, \ldots, m_t\}$ be the set of indices corresponding to the s_t^* largest in absolute value coordinates of δ_t outside of $J_{t,0}$. In the case of $s_t^* > m_t/2$, it corresponds to the $m_t - s_t^*$ largest absolute values. Define $J_{t,01} \stackrel{\text{def}}{=} J_{t,0} \cup J_{t,1}$. Let δ_{t,J_t} be the sub-vector of δ_t corresponding to J_t, similarly for δ_{t,J_t^c} and $\delta_{t,J_{t,01}}$.

To show identification of Π_t^*, we consider two events (for each t) associated with the restricted eigenvalue (RE) conditions, as outlined in Section 3 of Bickel et al. (2009). For $c_0 > 0$, define

$$\begin{aligned}
\mathcal{A}_{1t} &\stackrel{\text{def}}{=} \Big\{\min_{\delta_t \neq 0, |\delta_t|_0 \leq s_t^*, |\delta_{t,J_t^c}|_1 \leq c_0|\delta_{t,J_t}|_1} \frac{|\boldsymbol{V}_t\delta_t|_2}{\sqrt{N}|\delta_{t,J_t}|_2} \geq \kappa_t(c_0, s_t^*)\Big\},\\
\mathcal{A}_{1t}' &\stackrel{\text{def}}{=} \Big\{\min_{\delta_t \neq 0, |\delta_t|_0 \leq s_t^*, |\delta_{t,J_t^c}|_1 \leq c_0|\delta_{t,J_t}|_1} \frac{|\boldsymbol{V}_t\delta_t|_2}{\sqrt{N}|\delta_{t,J_{t,01}}|_2} \geq \kappa_t(c_0, s_t^*, s_t^*)\Big\},
\end{aligned}$$

where $\kappa_t(c_0, s_t^*)$ and $\kappa_t(c_0, s_t^*, s_t^*)$ are positive constants that depend on s_t^* and c_0. In Lemma 3.1, we will prove that these events occur with probabilities approaching to 1 as $N \to \infty$, for some $\kappa_t(\cdot) > 0$ related to the RE condition in population, as per Assumption 3.3.

Assumption 3.3 (RE Condition). *For any constant $c_0 > 0$, define the subspace*

$$\Omega_t(c_0, s_t^*) \stackrel{\text{def}}{=} \{\delta_t/|\delta_t|_2 : \delta_t \in \mathbb{R}^{m_t}, \delta_t \neq 0, |\delta_t|_0 \leq s_t^*, |\delta_{t,J_t^c}|_1 \leq c_0|\delta_{t,J_t}|_1\}.$$

Assume that there exist some positive constants $C_{\min}$ and $C_{\max}$ such that

$$C_{\min} \leq \min_{2\leq t\leq T} \min_{\delta_t \in \Omega_t(c_0, s_t^*)} \delta_t^\top \bar{\mathsf{E}}(V_{it}V_{it}^\top)\delta_t \leq \max_{2\leq t\leq T} \max_{\delta_t \in \Omega_t(c_0, s_t^*)} \delta_t^\top \bar{\mathsf{E}}(V_{it}V_{it}^\top)\delta_t \leq C_{\max}.$$

Lemma 3.1 (Identification). *Under Assumptions 3.1-3.2, and Assumption 3.3 holding with $C_{\min} = \min_{2\leq t\leq T} \kappa_t^2(c_0, s_t^*) - \Delta_{N,T}$ and $C_{\max} = \max_{2\leq t\leq T} \kappa_t^2(c_0, s_t^*) + \Delta_{N,T}$, where $\min_{2\leq t\leq T} \kappa_t(c_0, s_t^*) > 0$ and $\Delta_{N,T} \overset{\text{def}}{=} \max_{2\leq t\leq T} \sqrt{s_t^*/N}\log m_t \to 0$ as $N, T \to \infty$, then for each t,*

$$\min_{\delta_t\neq 0, |\delta_t|_0 \leq s_t^*, |\delta_{t,J_t^c}|_1 \leq c_0|\delta_{t,J_t}|_1} \frac{|\boldsymbol{V}_t\delta_t|_2}{\sqrt{N}|\delta_{t,J_t}|_2} \geq \kappa_t(c_0, s_t^*),$$

holds with probability $1 - \mathcal{o}(1)$, as $N \to \infty$.

Lemma 3.1 shows that $\mathrm{P}(\mathcal{A}_{1t}) \to 1$ as $N \to \infty$ for each t, which is in line with Lemma 1 of Belloni and Chernozhukov (2013). We can similarly verify that $\mathrm{P}(\mathcal{A}_{1t}') \to 1$ as $N \to \infty$. These results establish the identification of the sparse solution Π_t^* within the subspace. In Appendix A.2.2, we show that for a specific example, the oracle order of sparsity is bounded as $s_t^* \asymp \log N \wedge t$.

Recall the LASSO estimator $\widehat{\Pi}_t$ obtained by (3). To achieve good prediction performance of the estimator, properly chosen penalty tuning parameters and weights are necessary. For each $t = 2, \ldots, T$, let $\boldsymbol{\omega}_t$ be an $m_t \times 1$ vector, with the first element being 1 and the remaining elements collecting the non-negative penalty weights $(\omega_{ts}\mathbf{1}_s)_{s=1}^{t-1}$, where $\mathbf{1}_s$ represents a vector of ones with the same dimension as X_{is}.

Assumption 3.4 (Penalty Weights). *Assume that $|\boldsymbol{\omega}_t|_\infty$ is bounded by constant, $|\boldsymbol{\omega}_{t,J_t}|_2 \leq \sqrt{s_t^*}$, where $\boldsymbol{\omega}_{t,J_t}$ is the sub-vector of $\boldsymbol{\omega}_t$ corresponding to J_t.*

According to the Karush-Kuhn-Tucker conditions of LASSO, the solution $\widehat{\Pi}_t$ satisfies

$$|\{\boldsymbol{V}_t^\top(\boldsymbol{W}_t - \boldsymbol{V}_t\widehat{\Pi}_t)\} \oslash \boldsymbol{\omega}_t|_\infty \leq \lambda_t,$$

where $\oslash$ represents the Hadamard division, i.e. element-wise division. On the other hand, to ensure the true Π_t^0 is feasible for the LASSO problem with high probability, we need the event

$$\mathcal{A}_{2t} \overset{\text{def}}{=} \{c|\boldsymbol{V}_t^\top\boldsymbol{\eta}_t \oslash \boldsymbol{\omega}_t|_\infty \leq \lambda_t\}$$

to occur with high probability, where $c > 2$ is a constant. This suggests an ideal choice of the tuning parameter λ_t is given by the tail quantile of the random variable $c|\boldsymbol{V}_t^\top\boldsymbol{\eta}_t \oslash \boldsymbol{\omega}_t|_\infty$. Furthermore, by relying on Assumptions 3.1-3.2 and 3.4, we can apply Lemma A.2, which provides the maximal tail probability for the partial sum of the m_t-dimensional process $\varpi_{it} \overset{\text{def}}{=} V_{it}\eta_{it} \oslash \boldsymbol{\omega}_t$, to derive an upper bound for λ_t of the order $\sqrt{N\log m_t}$.

Lastly, to conclude the consistency of the LASSO estimators, we present the prediction performance bounds for $\delta_{\Pi,t} \overset{\text{def}}{=} \widehat{\Pi}_t - \Pi_t^*$ in the following theorem. This will be combined with the prediction norm of the approximation error in (11) to derive a performance bound for $\widehat{\Pi}_t - \Pi_t^0$ using the triangle inequality.

Theorem 3.1 (Prediction Performance Bounds of LASSO)**.** *Under the same assumptions as in Lemma 3.1 and Assumption 3.4, on the event $\mathcal{A}_{2t}$ for each $t = 2, \ldots, T$, we can conclude, with probability $1 - o(1)$,*

$$|\delta_{\Pi,t}|_{2,N} \leq 2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa_t(3, s_t^*),$$

$$|\delta_{\Pi,t}|_1 \leq 7\sqrt{s_t^*}\{2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa_t(3, s_t^*)\}/\kappa_t(3, s_t^*) + 56NC_{s_t^*}^2/(3\lambda_t).$$

Based on these findings, in Corollary 3.1, we provide the joint prediction performance bounds, where the ℓ_2-norm bound is derived by following Theorem 7.2 of Bickel et al. (2009).

Corollary 3.1 (Joint Prediction Performance Bounds of LASSO)**.** *Under the same assumptions as in Theorem 3.1, if $\mathrm{P}\big(\bigcap_{t=2}^T(\mathcal{A}_{1t} \cap \mathcal{A}_{2t})\big) \to 1$ as $N, T \to \infty$, then with probability $1 - o(1)$,*

$$\max_{2\leq t\leq T} |\delta_{\Pi,t}|_1 \leq 7 \max_{2\leq t\leq T} \sqrt{s_t^*}\{2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa_t(3, s_t^*)\}\Big/\Big(\min_{2\leq t\leq T} \kappa_t(3, s_t^*)\Big) + 56N \max_{2\leq t\leq T} C_{s_t^*}^2\Big/\Big(3 \min_{2\leq t\leq T} \lambda_t\Big).$$

In addition, if $\mathrm{P}\big(\bigcap_{t=2}^T(\mathcal{A}'_{1t} \cap \mathcal{A}_{2t})\big) \to 1$ as $N, T \to \infty$, then with probability $1 - o(1)$,

$$\max_{2\leq t\leq T} |\delta_{\Pi,t}|_2 \leq 4 \max_{2\leq t\leq T}\{2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa_t(3, s_t^*, s_t^*)\}\Big/\Big(\min_{2\leq t\leq T} \kappa_t(3, s_t^*, s_t^*)\Big).$$

3.2. **Basic Model: Inference Theory for Step 2.** In this subsection, we establish the asymptotic normality of the final estimator for both AB-LASSO and AB-LASSO-SS, which will allow us to perform large sample inference on the parameter of interest and functions of it.

Define $\mathbf{\Theta}_t^0$ (resp. $\widehat{\mathbf{\Theta}}_t$) by stacking Π_t^0 (resp. $\widehat{\Pi}_t$) by rows for all $W_{it} \in \Delta X_{it}$. Specifically, when the number of components in X_{it} is d, we have $\mathbf{\Theta}_t^0$ and $\widehat{\mathbf{\Theta}}_t$ with dimensions $d \times m_t$ for each $t = 2, \ldots T$. Recall the definition $V_{it} = (1, X_{i1}^\top, \ldots, X_{i,t-1}^\top)^\top$. It follows that $\widehat{\Delta X_{it}} = \widehat{\mathbf{\Theta}}_t V_{it}$, and the AB-LASSO estimator obtained in (4) (irrespective of the demeaning transformation to remove the time effects) can be expressed by

$$\widehat{\theta} - \theta^0 = \left(\sum_{i=1}^N \sum_{t=2}^T \widehat{\mathbf{\Theta}}_t V_{it} \Delta X_{it}^\top\right)^{-1} \left(\sum_{i=1}^N \sum_{t=2}^T \widehat{\mathbf{\Theta}}_t V_{it} \Delta\varepsilon_{it}\right).$$

The asymptotic variance of $\widehat{\theta}$ has the sandwich form. We impose the following assumptions to derive the specific formula for it.

Assumption 3.5 (Nonsingularity). *Assume that as $N, T \to \infty$, the limit matrix $Q = \lim_{N,T\to\infty} (NT)^{-1} \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 \mathsf{E}(V_{it} \Delta X_{it}^\top)$ is nonsingular.*

Assumption 3.2(ii) implies that $\mathsf{E}(\varepsilon_{i,t-1}\varepsilon_{is} \mid V_{it}) = 0$, for $1 \leq s < t-1$, hence $\mathsf{E}(\Delta\varepsilon_{it}\Delta\varepsilon_{i,t-\ell} \mid V_{it}) = 0$, for $\ell > 1$. By defining $\Sigma_{0,t} \stackrel{\text{def}}{=} \lim_{N\to\infty} N^{-1} \sum_{i=1}^N \mathsf{E}(V_{it}V_{it}^\top (\Delta\varepsilon_{it})^2)$ and $\Sigma_{1,t} \stackrel{\text{def}}{=} \lim_{N\to\infty} N^{-1} \sum_{i=1}^N \mathsf{E}(V_{it}V_{i,t-1}^\top \Delta\varepsilon_{it}\Delta\varepsilon_{i,t-1})$, we can express the asymptotic variance of $\widehat{\theta}$ in the form of

$$\Omega = Q^{-1}\mathbf{\Sigma}(Q^{-1})^\top, \tag{12}$$

where

$$\mathbf{\Sigma} = \lim_{T\to\infty} \Big(\frac{1}{T} \sum_{t=2}^T \mathbf{\Theta}_t^0 \Sigma_{0,t} \mathbf{\Theta}_t^{0\top} + \frac{1}{T} \sum_{t=3}^T \mathbf{\Theta}_t^0 \Sigma_{1,t} \mathbf{\Theta}_{t-1}^{0\top} + \frac{1}{T} \sum_{t=3}^T \mathbf{\Theta}_{t-1}^0 \Sigma_{1,t}^\top \mathbf{\Theta}_t^{0\top} \Big).$$

Accordingly, the empirical analog of Ω is

$$\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\mathbf{\Sigma}}(\widehat{Q}^{-1})^\top, \tag{13}$$

where $\widehat{Q} = (NT)^{-1} \sum_{i=1}^N \sum_{t=2}^T \widehat{\mathbf{\Theta}}_t V_{it} \Delta X_{it}^\top$, and

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T} \sum_{t=2}^T \widehat{\mathbf{\Theta}}_t \widehat{\Sigma}_{0,t} \widehat{\mathbf{\Theta}}_t^\top + \frac{1}{T} \sum_{t=3}^T \widehat{\mathbf{\Theta}}_t \widehat{\Sigma}_{1,t} \widehat{\mathbf{\Theta}}_{t-1}^\top + \frac{1}{T} \sum_{t=3}^T \widehat{\mathbf{\Theta}}_{t-1} \widehat{\Sigma}_{1,t}^\top \widehat{\mathbf{\Theta}}_t^\top,$$

with $\widehat{\Sigma}_{0,t} = N^{-1} \sum_{i=1}^N V_{it}V_{it}^\top (\Delta\widehat{\varepsilon}_{it})^2$ and $\widehat{\Sigma}_{1,t} = N^{-1} \sum_{i=1}^N V_{it}V_{i,t-1}^\top \Delta\widehat{\varepsilon}_{it}\Delta\widehat{\varepsilon}_{i,t-1}$.

Remark 3.1 (Consistency of $\widehat{\mathbf{\Sigma}}$ and $\widehat{Q}$). The consistency property for $\widehat{\mathbf{\Sigma}}$ and $\widehat{Q}$ is essential for forming feasible confidence intervals for θ^0 based on $\widehat{\Omega}$. Such a requirement for $\widehat{\mathbf{\Sigma}}$ can be inferred from $\max_{1\leq i\leq N, 2\leq t\leq T} |\widehat{\varepsilon}_{it} - \varepsilon_{it}| = o_{\mathrm{P}}(1)$ and $\max_{2\leq t\leq T} |\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^0|_{1,1} = o_{\mathrm{P}}(1)$.

To verify these conditions, consider $\Delta\widehat{\varepsilon}_{it} = \Delta Y_{it} - \Delta X_{it}^\top \widehat{\theta}$:

$$\max_{1\leq i\leq N, 2\leq t\leq T} |\Delta\widehat{\varepsilon}_{it} - \Delta\varepsilon_{it}| = \max_{1\leq i\leq N, 2\leq t\leq T} |\Delta X_{it}^\top (\widehat{\theta} - \theta^0)| \leq \max_{1\leq i\leq N, 2\leq t\leq T} |\Delta X_{it}|_\infty |(\widehat{\theta} - \theta^0)|_1.$$

Under the conditions of Theorem 3.2, we have the above term of order $\sqrt{\log(NT)}/\sqrt{NT}$. Additionally, we have $\max_{2\leq t\leq T} |\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^0|_{1,1} \lesssim_{\mathrm{P}} \max_{2\leq t\leq T} s_t^* \log m_t / \sqrt{N}$, as shown in Corollary 3.1.

Concerning $\widehat{Q}$, we observe that

$$\begin{aligned}
&\Big|(NT)^{-1} \sum_{i=1}^N \sum_{t=2}^T \{\widehat{\mathbf{\Theta}}_t V_{it} \Delta X_{it}^\top - \mathbf{\Theta}_t^0 \mathsf{E}(V_{it}\Delta X_{it}^\top)\}\Big|_{\max} \\
\leq\ & \Big|(NT)^{-1} \sum_{i=1}^N \sum_{t=2}^T (\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^0) V_{it} \Delta X_{it}^\top\Big|_{\max} + \Big|(NT)^{-1} \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 \{V_{it}\Delta X_{it}^\top - \mathsf{E}(V_{it}\Delta X_{it}^\top)\}\Big|_{\max}.
\end{aligned}$$

As shown in the proof of Theorem 3.2, the first term is bounded by $\max_{2\le t\le T}(s_t^* \log m_t)^{3/2}/N$, with probability tending to 1, and the second term is of order $\mathcal{O}_{\mathrm{P}}(1/\sqrt{NT})$. Hence, the consistency for $\widehat{Q}$ follows.

Theorem 3.2 (Asymptotic Normality of AB-LASSO). *Under Assumptions 3.1-3.5, assuming the asymptotic variance Ω is a positive definite matrix, and $\max_{2\le t\le T} s_t^* \log m_t \sqrt{T/N} \to 0$, the AB-LASSO $\widehat{\theta}$ obtained by* (4) *is a consistent estimator of θ^0, and*

$$\sqrt{NT}(\widehat{\theta}-\theta^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0,\Omega). \tag{14}$$

Theorem 3.3 (Asymptotic Normality of AB-LASSO-SS). *Under Assumptions 3.1-3.5, assuming the asymptotic variance Ω is a positive definite matrix, and $\max_{2\le t\le T} \sqrt{s_t^*} \log m_t/\sqrt{N} \to 0$, the AB-LASSO-SS $\widehat{\theta}_{SS}$ obtained by* (7) *is a consistent estimator of θ^0, and*

$$\sqrt{NT}(\widehat{\theta}_{SS}-\theta^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0,\Omega). \tag{15}$$

Remark 3.2 (Discussion of the Rates). It is important to note that the small bias condition $\max_{2\le t\le T} s_t^* \log m_t \sqrt{T/N} \to 0$, as stated in Theorem 3.2, is relaxed in Theorem 3.3. This relaxation results in a more favorable convergence rate for the estimator when using AB-LASSO-SS. This observation certifies that sample-splitting effectively mitigates the overfitting bias by employing different sub-samples for instruments selection and parameter estimation.

More fundamentally, to prove the two theorems above, we require the term

$$(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=2}^{T} Q^{-1}(\widehat{\mathbf{\Theta}}_t-\mathbf{\Theta}_t^0)V_{it}\Delta\varepsilon_{it} \to 0,$$

as $N,T\to\infty$. Without sample-splitting, the generated errors $(\widehat{\mathbf{\Theta}}_t-\mathbf{\Theta}_t^0)$ might be correlated with the ordinary errors ε_{it}, and the order of this term is given by $\max_{2\le t\le T} s_t^* \log m_t \sqrt{T/N}$. While using sample-splitting, we achieve a smaller order of this term: $\max_{2\le t\le T} \sqrt{s_t^*} \log m_t/\sqrt{N}$.

Remark 3.3 (Time Effects). When time effects are included, the cross-sectional demeaning of the variables introduces weak cross-sectional dependence to the data of order $1/N$. This dependence makes the asymptotic analysis much more cumbersome, but does not significantly affect the results provided that $T/N\to 0$, as $N,T\to\infty$. Thus, the expressions of the asymptotic variance only need to be adjusted by substituting the variables with their demeaned counterparts. In Appendix A.3.3, we illustrate how the theorems outlined in this section adapt in the presence of time effects in an example with a dynamic panel model with one lagged dependent variable as a covariate, that is, a panel AR(1) model.

3.3. **General Model.** Before concluding this section, we present the limiting distribution of the estimator $\widehat{\theta}_1$ under the general model with many exogenous covariates, obtained by (10), irrespective of the demeaning transformation to remove the time effects. Recall the definition $U_{it} = (U_{it}^{0\top}, \Delta X_{2,it}^\top)^\top$, where U_{it}^0 contains the ideal IVs for $\Delta X_{1,it}$[5]. The asymptotic variance of $\widehat{\theta}_1$ takes the form of

$$\Omega_1 = Q_1^{-1}\mathbf{\Sigma}_1(Q_1^{-1})^\top,$$

where $Q_1 = \lim_{N,T\to\infty}(NT)^{-1}\sum_{i=1}^N\sum_{t=2}^T \mathcal{W}_t^\top\,\mathsf{E}(U_{it}\Delta X_{1,it}^\top)$ is assumed to be a nonsingular matrix, and

$$\begin{aligned}\mathbf{\Sigma}_1 = \lim_{N,T\to\infty}\Big\{&(NT)^{-1}\sum_{i=1}^N\sum_{t=2}^T \mathcal{W}_t^\top\,\mathsf{E}(U_{it}U_{it}^\top(\Delta\varepsilon_{it})^2)\mathcal{W}_t\\ &+(NT)^{-1}\sum_{i=1}^N\sum_{t=3}^T \mathcal{W}_t^\top\,\mathsf{E}(U_{it}U_{i,t-1}^\top\Delta\varepsilon_{it}\Delta\varepsilon_{i,t-1})\mathcal{W}_{t-1}\\ &+(NT)^{-1}\sum_{i=1}^N\sum_{t=3}^T \mathcal{W}_{t-1}^\top\,\mathsf{E}(U_{i,t-1}U_{it}^\top\Delta\varepsilon_{i,t-1}\Delta\varepsilon_{it})\mathcal{W}_t\Big\}.\end{aligned}$$

Define the moments $M_t \overset{\text{def}}{=} \bar{\mathsf{E}}\left\{\begin{pmatrix}\Delta X_{1,it}\\ \Delta X_{2,it}\end{pmatrix}U_{it}^\top\right\}$, and the empirical counterpart $\widehat{M}_t \overset{\text{def}}{=} N^{-1}\sum_{i=1}^N\left\{\begin{pmatrix}\Delta X_{1,it}\\ \Delta X_{2,it}\end{pmatrix}\widehat{U}_{it}^\top\right\}$, where U_{it}^0 in U_{it} is replaced by the LASSO predictions, i.e. $\widehat{U}_{it} = (\widehat{\Delta X_{1,it}}^\top, \Delta X_{2,it}^\top)^\top$. Some additional assumptions on the moments M_t and the weighting matrix $\mathcal{W}_t$ under the general model are made as follows.

Assumption 3.6 (Additional Assumptions on the General Model)**.** *(i) There exist sequences of constants $c_n, w_n \ge 0$ (where $n = NT$), such that $\max_{2\le t\le T}(|\mathcal{W}_t|_{1,1}\vee|M_t^{-1}|_\infty) \le c_n$, and*

$$\max_{2\le t\le T}\sum_{i=1}^d\sum_{j=1}^{d_1}|\mathcal{W}_{t,ij}|^r \lesssim w_n,\ \ \textit{for some } 0\le r<1,$$

where $\mathcal{W}_{t,ij}$ is the element in the i-th row and j-th column of $\mathcal{W}_t$.

(ii) For each $t = 2,\dots,T$, assume that $|M_t - \widehat{M}_t|_{\max} \lesssim_{\mathrm{P}} \rho_{N,t}$, with $\rho_{N,t}\to 0$ as $N\to\infty$.

(iii) Let $v_n \overset{\text{def}}{=} \log(N\vee T\vee d)$, and assume that

$$c_n^2 w_n (v_n/N)^{\frac{1-r}{2}}\sqrt{T} + c_n\max_{2\le t\le T}s_t^*\log m_t\sqrt{T}/\sqrt{N}\to 0,\ \ \textit{as } N,T\to\infty,$$

with the same r that makes part (i) hold.

[5]The ideal IVs for $\Delta X_{1,it}$ are structured similarly to those for ΔX_{it} in the basic model, expressed as $\mathbf{\Theta}_t^0 V_{it}$. The covariates V_{it} are expanded by the additional moments arising from the strictly exogenous covariates in $X_{2,it}$, as commented in Remark 2.5.

(iv) The tuning parameter $\ell_t \geq 0$ *used in* (9) *satisfies* $\max_{2\leq t\leq T} \ell_t \vartheta_n = o(1/\sqrt{NT})$, *and* $\max_{2\leq t\leq T}(\ell_t + \rho_{N,t}c_n) \lesssim c_n\sqrt{v_n/N}$, *where* $\vartheta_n \overset{\text{def}}{=} |\theta_2^0|_1$.

The consistency condition stated in Assumption 3.6(ii) can be achieved by bounding $|M_t - \widehat{M}_t|_{\max}$ using the concentration inequality provided in Lemma A.2, relying on Assumptions 3.1-3.2.

Moreover, the crucial rate condition in Assumption 3.6(iii) can be further improved to $c_n^2 w_n(v_n/N)^{\frac{1-r}{2}} + c_n \max_{2\leq t\leq T} \sqrt{s_t^*}\log m_t/\sqrt{N} = o(1)$, if a sample-splitting procedure is employed. Specifically, $\widehat{\mathcal{W}}_t$ by the Dantzig selector and instruments selection by LASSO are obtained from a sub-sample that is cross-sectionally independent of the sub-sample used for computing the final estimator in (10). More detailed explanations regarding this improvement can be found in the proof of Theorem 3.4.

Theorem 3.4 (Asymptotic Normality for the General Model Estimator)**.** *Under Assumptions 3.1-3.6, assuming the asymptotic variance* Ω_1 *is a positive definite matrix, we have the general model estimator* $\widehat{\theta}_1$ *obtained by* (10) *is consistent, and*

$$\sqrt{NT}(\widehat{\theta}_1 - \theta_1^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \Omega_1). \tag{16}$$

4. Simulation Study

We illustrate the finite sample properties of the proposed AB-LASSO and AB-LASSO-SS estimators, comparing them with other alternative methods based on AB. We consider the following data generating process: for $i = 1, \ldots, N$, $t = 1, \ldots, T$,

$$\begin{aligned} Y_{it} &= \alpha_i + \gamma_t + \theta_1 Y_{i,t-1} + \theta_2 D_{it} + \varepsilon_{it}, \\ D_{it} &= \rho D_{i,t-1} + v_{it}, \end{aligned}$$

where $\alpha_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\alpha^2)$ and $\gamma_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\gamma^2)$ are mutually independent generated. For each i,

$$\begin{pmatrix} \varepsilon_{i,t-1} \\ v_{it} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right),$$

such that D_{it} is predetermined with respect to ε_{it}, but it is not strictly exogenous. We set $\rho = 0.5$, $\theta_1 = 0.8$, $\theta_2 = 1$, and $\sigma_\alpha = \sigma_\gamma = 1$. To start the process, we set the initial values of Y and D to zero for all the units and use the first 10 periods as burn-in sample.

As in the empirical application in Section 5, we assume that Y_{it} is first observed at $t = 1$ and use all the available lags of Y_{it} and D_{it} to construct the moment conditions, that is,

$$\mathsf{E}(Z_{it}\Delta\widetilde{\varepsilon}_{it}) = 0, \quad Z_{it} = (Y_{i,t-2}, \ldots, Y_{i1}, D_{i,t-1}, \ldots, D_{i1})^\top.$$

See Remark 2.1 for a description of how the AB-LASSO and AB-LASSO-SS are modified when we do not observe the initial condition Y_{i0}. The LASSO fitting is carried out with post-LASSO using a penalty (λ) that is independent of the design matrix. This approach is chosen for its conservatism to prevent overfitting (Chernozhukov et al., 2021a). We set the penalty weights equal to one, i.e., $\omega_{ts} = 1$ for all t, s. For the AB-LASSO-SS estimator, we implement a K-fold sample-splitting and cross-fitting procedure with $K \in \{2, 5\}$. According to the formula for the asymptotic variance estimation shown in (13), we calculate the standard error using the residuals based on the final estimate obtained after aggregation. We also repeat the estimators 100 times using random sample splits and aggregate the results through the medians. In addition to the conventional two-step AB, we compare the results with the debiased AB of Chen et al. (2019) using one random split (i.e. 2 folds) along the cross-section dimension. We refer this estimator as DAB-SS. For both comparison estimators, we use the analytical standard error clustered at the individual level for AB, which are also asymptotically valid for DAB-SS.

For each estimator, we report the root mean square error (RMSE), standard deviation and bias in percent of the true parameter value, together with the length and empirical coverage of confidence intervals (CI) with a nominal confidence level of 95%. Tables 4.1 and 4.2 display the results for the parameters θ_1 and θ_2, respectively, based on 500 simulations for $N = 200$ and $T \in \{30; 40; 50\}$. These sample sizes lead to numbers of moment conditions $m \in \{784; 1,444; 2,304\}$ that are large relative to the corresponding sample sizes $n = NT \in \{6,000; 8,000; 10,000\}$. The resulting orders of the small bias condition are $m^2/n \in \{102; 261; 531\}$, which are not negligible.

We find that the performance of AB-LASSO-SS in terms of RMSE and CI coverage is comparable to AB when $T = 30$. Thus, the bias reduction of AB-LASSO-SS relative to AB is exactly compensated by an increase of dispersion in the RMSE. However, as T increases, AB-LASSO-SS outperforms AB in all dimensions. For all the sample sizes considered, AB has a bias of similar order of magnitude as the standard deviation and that grows with T as the asymptotic theory predicts. It is evident that using LASSO to select the most relevant moments followed by a sample-splitting procedure, significantly reduces the bias and results in more accurate coverage rates. The confidence interval (CI) length of AB-LASSO-SS is generally shorter than that for AB in most cases, suggesting that the bias reduction does not come at the expense of more dispersion for sufficiently large T.

We also find that the AB-LASSO suffers from severe overfitting bias for both coefficients. This bias can be even greater than the original bias of AB and highlights the necessity of carrying out sample splitting. DAB-SS improves over AB in terms of the treatment coefficient θ_2 (when T is large) but not the autoregressive coefficient θ_1. Overall, we find that the results for AB-LASSO-SS remain robust across different choices of K in terms of estimation, and that using $K = 5$ split folds improves the inference accuracy with lower standard error and coverage closer to the nominal level.

Table 4.1. Results for $\theta_1 = 0.8$ with $N = 200$

	AB	AB-LASSO	AB-LASSO-SS ($K = 2$)	AB-LASSO-SS ($K = 5$)	DAB-SS
			$T = 30$		
RMSE	**0.04**	0.16	**0.04**	**0.04**	0.15
std. dev.	**0.02**	**0.02**	0.04	0.04	0.07
bias	-0.03	-0.16	**0.00**	**0.00**	0.13
CI length	0.14	0.09	0.23	0.14	0.14
coverage	0.97	0.00	0.99	**0.95**	0.21
			$T = 40$		
RMSE	0.10	0.16	**0.04**	**0.04**	0.33
std. dev.	0.05	**0.02**	0.04	0.04	0.12
bias	-0.09	-0.16	**0.00**	**0.00**	0.31
CI length	0.25	0.07	0.20	0.13	0.25
coverage	0.81	0.00	0.99	**0.93**	0.06
			$T = 50$		
RMSE	0.26	0.16	0.04	**0.03**	0.36
std. dev.	0.08	**0.02**	0.04	0.03	0.17
bias	-0.25	-0.16	**0.00**	**0.00**	0.32
CI length	0.30	0.05	0.15	0.11	0.30
coverage	0.22	0.00	0.99	**0.93**	0.21

Notes: The numbers in the table are divided by 0.8 for RMSE, standard deviation (std. dev.), bias, and CI length. Superior results are indicated in bold.

We present some supplementary simulation results in Appendix B. In Tables B.1 and B.2, the results for $N = 400$ are provided. With a relatively large N, the bias in AB estimation is partially mitigated. However, when $T \in \{40; 50\}$, AB still produces considerably wider confidence intervals compared to AB-LASSO-SS, which affects the credibility of the coverage rate for inference purposes. Additionally, implementing AB demands a greater amount of computational memory. For example, with $N = 400$ and $T = 30$, running AB for one sample requires approximately 3.5 Gb of memory usage, whereas only 15 Mb are needed for AB-LASSO-SS with a single partition of 5 folds. We track the RAM usage in R using the package `peakRAM`.

Table 4.2. Results for $\theta_2 = 1$ with $N = 200$

	AB	AB-LASSO	AB-LASSO-SS ($K = 2$)	AB-LASSO-SS ($K = 5$)	DAB-SS
			$T = 30$		
RMSE	0.07	0.18	**0.05**	**0.05**	0.12
std. dev.	0.04	**0.03**	0.05	0.05	0.10
bias	-0.06	-0.18	**0.00**	0.01	0.08
CI length	0.28	0.11	0.28	0.18	0.28
coverage	0.97	0.00	0.99	**0.95**	0.76
			$T = 40$		
RMSE	0.16	0.18	0.05	**0.04**	0.15
std. dev.	0.07	**0.03**	0.05	0.04	0.14
bias	-0.14	-0.18	**0.00**	**0.00**	0.04
CI length	0.45	0.10	0.25	0.15	0.45
coverage	0.88	0.00	0.99	**0.94**	0.83
			$T = 50$		
RMSE	0.24	0.19	**0.04**	**0.04**	0.18
std. dev.	0.09	**0.02**	0.04	0.04	0.18
bias	-0.23	-0.18	**0.00**	**0.00**	**0.00**
CI length	0.53	0.09	0.23	0.17	0.53
coverage	0.68	0.00	1.00	**0.95**	0.86

Notes: Superior results are indicated in bold.

In Tables B.3 and B.4, we compare AB-LASSO and AB-LASSO-SS (5 folds) with other naive approaches that fit the instruments through OLS regression rather than using LASSO, for cases with $N = 200$ and $T \in \{30; 40; 50\}$. The results highlight the crucial role of the moment selection in reducing bias. Utilizing LASSO to select the most informative moments followed by a sample-splitting procedure, leads to narrower confidence intervals and contributes to more efficient inference compared to AB-OLS-SS, which tends to overfit.

5. School Opening and COVID-19 Spread

We apply AB-LASSO-SS to study the effect of K-12 schools opening and other policies on the spread of COVID-19 in the U.S. We use a balanced panel of 2,510 counties over 32 weeks between April 1st and December 2nd, 2020. This panel was extracted from Chernozhukov et al. (2021b), which constructed an unbalanced panel of U.S. counties including 7-day moving averages of daily observations for the same period. We aggregate the observations at the week level to avoid spurious serial correlation coming from the moving averages.

In this application, Y_{it} is the logarithm of the number of reported COVID-19 cases in county i at week t, D_{it} is a measure of visits to K-12 schools from SafeGraph foot traffic data, C_{it} contains other treatments and control variables, α_i is a county fixed effect and γ_t is a week fixed effect. We estimate the model:

$$Y_{it} = \alpha_i + \gamma_t + \theta_0 D_{i,t-1} + \beta_1 Y_{i,t-1} + \beta_2 Y_{i,t-2} + \beta_3 Y_{i,t-3} + \beta_4 Y_{i,t-4} + \theta_1^\top C_{1i,t-1} + \theta_2 C_{2it} + \varepsilon_{it},$$

where C_{1it} includes measure of visits to colleges and policy indicators on mask mandates, stay-at-home orders and the ban on gatherings of more than 50 persons, and C_{2it} includes a measure of the weekly growth rate in the number of tests. We assume that there is no serial correlation in ε_{it} over t. The variables in D_{it} and C_{1it} enter the model lagged one week to account for the time lag between infection and case confirmation. Additionally, we assume that D_{it}, C_{1it} and C_{2it} are predetermined with respect to ε_{it} and therefore use $Y_{i,t-2}, \dots, Y_{i1}$, $D_{i,t-1}, \dots, D_{i1}$, $C_{1i,t-1}, \dots, C_{1i1}$ and $C_{2i,t-1}, \dots, C_{2i1}$ to construct moment conditions at each t. This yields $m = 3,402$ moment conditions and $n = NT = 2,510 \times 27 = 67,770$ observations.[6] AB is likely to be biased in this case because $m^2/n \approx 170$.

Table 5.1 presents estimates and standard errors for AB, DAB-SS and AB-LASSO-SS with $K = 2$ and $K = 5$. For the AB-LASSO-SS estimators, the penalty tuning parameter λ and penalty weights ω are chosen in the same way as in the simulation study of Section 4. The coefficients of the model measure the short run effects of the covariates. In addition to these coefficients, we report results for the long-run effects obtained as $\theta_k/(1 - \sum_{j=1}^4 \beta_j)$ where θ_k is the coefficient of the covariate of interest and $\beta_1, \dots, \beta_4$ are the coefficients of $Y_{i,t-1}, \dots, Y_{i,t-4}$, respectively. All the methods reveal positive and significant effects of K-12 school and college visits on the spread of COVID-19. In particular, an increase in visits to K-12 schools and colleges is associated with a higher number of cases in both the short and long run. The estimated effects of mask mandates and stay-at-home orders are negative and significant both in the short and long runs. The effect of banning gathering is found to be negative and significant under AB and DAB-SS, but positive and barely significant under AB-LASSO-SS.

Comparing AB-LASSO-SS with AB, AB-LASSO-SS produces similar estimates of short-run effects but significantly smaller long-run effects in absolute value. AB-LASSO-SS also produces more precise estimates of both effects than AB. The difference might be attributed to the bias in the autoregressive coefficients in AB. In particular, AB-LASSO-SS gives significantly smaller estimates of the coefficient of $Y_{i,t-1}$. In general, the results of AB-LASSO-SS

[6]The first 5 weeks are used as initial conditions.

are not sensitive to the number of folds K. Debiasing the standard AB through half-splitting the panel does not result in significant changes in the estimates.

According to AB-LASSO-SS, we conclude that the opening of K-12 schools one week is associated with an increase in the number of covid-19 cases of about 50% the week after and has a compounded long run increase of more than 100%. Mask mandates and stay-at-home orders are associated with more modest effects. The reduction in the number of cases is about 8% and 12% after a week and about 20% and 30% in the long run, respectively. These effects are all statistically and economically relevant.

6. Concluding Remarks

We propose a LASSO and cross-fitting based estimator of dynamic linear panel models. This estimator shows better large sample properties and finite-sample performance in simulations than the classical AB estimator in long panels. In an empirical application, our estimator finds that policies such as the closure of K-12 schools, mask mandates and stay-at-home orders reduced the spread of COVID-19 both in the short and long run. Our estimates of the long run effects, however, are less optimistic than the estimates obtained with AB.

A potential avenue for future research is to analyze the performance of our method under weak identification. This situation arises, for example, when the process of the outcome Y_{it} is very persistent such that lagged values of Y_{it} might not strongly correlated with the outcome in differences ΔY_{it}, see, for example, Blundell and Bond (1998). Here, we conjecture that standard methods for dealing with weak instruments for cross-section data, such as the use of the Anderson-Rubin statistics, can be fruitfully applied to our setting (Anderson and Rubin, 1949). We leave this analysis to future research.

Table 5.1. Short-run and long-run effects on COVID-19 Cases

	AB-LASSO-SS ($K = 2$)	AB-LASSO-SS ($K = 5$)	AB	DAB-SS
$Y_{i,t-1}$	0.63***	0.58***	0.78***	0.79***
	(0.02)	(0.02)	(0.01)	(0.01)
$Y_{i,t-2}$	-0.02**	-0.02***	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
$Y_{i,t-3}$	0.04***	0.04***	0.06***	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)
$Y_{i,t-4}$	-0.01*	-0.01**	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
K-12 school opening: Short-run	0.50***	0.48***	0.47***	0.44***
	(0.15)	(0.14)	(0.10)	(0.10)
Long-run	1.44***	1.15***	3.17***	3.19***
	(0.42)	(0.35)	(0.69)	(0.73)
College visits: Short-run	1.73***	1.85***	1.14***	1.03***
	(0.33)	(0.33)	(0.34)	(0.34)
Long-run	5.02***	4.41***	7.69***	7.41***
	(1.01)	(0.84)	(2.32)	(2.47)
Mask mandates: Short-run	-0.09***	-0.08***	-0.10***	-0.10***
	(0.02)	(0.02)	(0.01)	(0.01)
Long-run	-0.24***	-0.19***	-0.66***	-0.69***
	(0.06)	(0.05)	(0.10)	(0.10)
Stay-at-home orders: Short-run	-0.12***	-0.12***	-0.08***	-0.09***
	(0.03)	(0.03)	(0.02)	(0.02)
Long-run	-0.33***	-0.28***	-0.54***	-0.62***
	(0.10)	(0.08)	(0.16)	(0.17)
Banning gatherings: Short-run	0.06	0.06*	-0.08***	-0.08***
	(0.03)	(0.03)	(0.02)	(0.02)
Long-run	0.16	0.14*	-0.52***	-0.58***
	(0.10)	(0.08)	(0.16)	(0.17)
Tests weekly growth	0.003	0.003	0.01	0.01
	(0.004)	(0.004)	(0.01)	(0.01)

Notes: Analytical standard errors in parentheses. Significant codes: 0.01***, 0.05**, 0.1*.

References

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

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

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint arXiv:1806.01888*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Bond, S. R. (2002). Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Economic Journal*, 1(2):141–162.

Burkholder, D. L. (1988). Sharp inequalities for martingales and stochastic integrals. In *Colloque Paul Lévy sur les processus stochastiques (22-26 juin 1987. École Polytechnique, Palaiseau)*, number 157-158 in Astérisque, pages 75–94. Société mathématique de France.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Carrasco, M. and Nayihouba, A. (2024). Regularized estimation of dynamic panel models. *Econometric Theory*, 40(2):360–418.

Chang, J., Chen, S. X., and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1):283–304.

Chang, J., Chen, S. X., Tang, C. Y., and Wu, T. T. (2021). High-dimensional empirical likelihood inference. *Biometrika*, 108(1):127–147.

Chen, L., Wang, W., and Wu, W. B. (2022). Inference of breakpoints in high-dimensional time series. *Journal of the American Statistical Association*, 117(540):1951–1963.

Chen, S., Chernozhukov, V., and Fernández-Val, I. (2019). Mastering panel metrics: Causal impact of democracy on growth. *AEA Papers and Proceedings*, 109:77–82.

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

Chernozhukov, V., Härdle, W., Huang, C., and Wang, W. (2021a). LASSO-driven inference in time and space. *The Annals of Statistics*, 49(3):1702–1735.

Chernozhukov, V., Kasahara, H., and Schrimpf, P. (2021b). The association of opening K-12 schools with the spread of COVID-19 in the United States: County-level panel data analysis. *Proceedings of the National Academy of Sciences*, 118(42):e2103420118.

Dhaene, G. and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030.

Dirksen, S. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36.

El Machkouri, M., Volnỳ, D., and Wu, W. B. (2013). A central limit theorem for stationary random fields. *Stochastic Processes and their Applications*, 123(1):1–14.

Kock, A. B. and Tang, H. (2019). Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory*, 35(2):295–359.

Kuersteiner, G. and Okui, R. (2010). Constructing optimal instruments by first-stage prediction averaging. *Econometrica*, 78(2):697–718.

Luo, Y. (2016). Selecting informative moments via LASSO. *Unpublished manuscript.*

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Newey, W. K. and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719.

Okui, R. (2009). The optimal choice of moments in dynamic panel data models. *Journal of Econometrics*, 151(1):1–16.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70–86.

Phillips, G. D. A. and Hale, C. (1977). The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review*, 18(1):219–228.

Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163.

Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Proceedings of the 25th Annual Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 10.1–10.24.

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

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, 195(1):104–119.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Zhang, D. and Wu, W. B. (2017). Supplement to "Gaussian approximation for high dimensional time series". DOI: 10.1214/16-AOS1512SUPP.

Appendix

A. Technical Proofs

A.1. **Some Useful Lemmas and Auxiliary Results.** Define $\boldsymbol{x} = (x_1, \ldots, x_n)^\top, \boldsymbol{y} = (y_1, \ldots, y_n)^\top$, where $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are sequences of independent, mean-zero, unit variance, sub-Gaussian random variables. Let $\mathcal{A}$ be a class of $n \times m$ matrices. For $A_1, A_2 \in \mathcal{A}$, define the d_q-metric as $d_q(A_1, A_2) \stackrel{\text{def}}{=} \|A_1 - A_2\|_q$, where $\|A\|_q$ are the Schatten norms of matrix A: $\|A\|_q \stackrel{\text{def}}{=} \Big(\sum_{i=1}^{\min(m,n)} \sigma_i^q(A)\Big)^{1/q}$ for $1 \leq q < \infty$, and $\|A\|_q \stackrel{\text{def}}{=} \sigma_1$ for $q = \infty$, if A has the singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$. For $\epsilon > 0$, the ϵ-covering number of $\mathcal{A}$ with respect to the d_q-metric is denoted by $\mathcal{N}(\epsilon, \mathcal{A}, d_q)$.

LEMMA A.1. *Define* $E = \sup_{A \in \mathcal{A}} \max(\mathsf{E}\,|A^\top \boldsymbol{x}|_2^2, \mathsf{E}\,|A^\top \boldsymbol{y}|_2^2) + \gamma_2^2(\mathcal{A}, d_\infty) + \bar{\Delta}_2(\mathcal{A})\gamma_2(\mathcal{A}, d_\infty)$, $V = \bar{\Delta}_\infty(\mathcal{A})\big(\bar{\Delta}_2(\mathcal{A}) + \gamma_2(\mathcal{A}, d_\infty)\big)$, $U = \bar{\Delta}_\infty^2(\mathcal{A})$, *where* $\gamma_p(\mathcal{A}, d_q) \lesssim \int_0^\infty \big(\log \mathcal{N}(\epsilon, \mathcal{A}, d_q)\big)^{1/p} d\epsilon$, $\bar{\Delta}_q(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_q$. *For any* $u \geq 0$, *there exists* $c_1, c_2 > 0$ *such that*

$$\mathrm{P}\Big(\sup_{A \in \mathcal{A}} |\boldsymbol{x}^\top A A^\top \boldsymbol{y}| \geq c_1 E + u\Big) \leq 2 \exp\big(-c_2 \min(u^2/V, u/U)\big).$$

A special case of Lemma A.1, with $\boldsymbol{x} = \boldsymbol{y}$, is stated in Theorem 6.2 of Dirksen (2015). The applicability of the rate in a general context is evidently clear.

Given an $s \in \mathbb{N}$, consider s probability spaces denoted as $(\Omega_1, \mathrm{P}_1), \ldots, (\Omega_s, \mathrm{P}_s)$. Suppose we have a parameter set $\mathcal{T}$ containing s-tuples $\tau = (\tau_1, \ldots, \tau_s)$. For each $\tau \in \mathcal{T}$, we have an s-tuple $X_\tau = (X_{\tau_1}, \ldots, X_{\tau_s})$ of sub-exponential random variables $X_{\tau_i} : \Omega_i \to \mathbb{R}$, and define the sub-exponential norm as $\|X_{\tau_i}\|_{\psi_1} = \inf\{v > 0 : \mathsf{E} \exp(|X_{\tau_i}|/v) \leq 2\}$. Consider the empirical process given by

$$W_\tau = \frac{1}{s}\sum_{i=1}^{s}(X_{\tau_i} - \mathsf{E}\, X_{\tau_i}),$$

Bernstein's inequality (referenced, for instance, as Lemma 5.1 of Dirksen (2015)) implies that the process $(W_\tau)_{\tau \in \mathcal{T}}$ exhibits a mixed tail behavior with respect to the metrics $(\frac{1}{s}d_1, \frac{1}{\sqrt{s}}d_2)$, where

$$d_1(\tau, \tau') = \max_{1 \leq i \leq s} \|X_{\tau_i} - X_{\tau_i'}\|_{\psi_1},\ d_2(\tau, \tau') = \Big(\frac{1}{s}\sum_{i=1}^{s} \|X_{\tau_i} - X_{\tau_i'}\|_{\psi_1}^2\Big)^{1/2}.$$

LEMMA A.2 (Corollary 5.2 of Dirksen (2015), Supremum of Empirical Processes)**.** *Let* $\sigma, K > 0$ *be constants such that*

$$\sup_{\tau \in \mathcal{T}} \frac{1}{s}\sum_{i=1}^{s} \mathsf{E}\,|X_{\tau_i} - \mathsf{E}\, X_{\tau_i}|^r \leq \frac{r!}{2}\sigma^2 K^{r-2}, \quad r = 2, 3, \ldots.$$

Then, for any $1 \le q < \infty$,

$$\Big(\mathsf{E}\sup_{\tau\in\mathcal{T}}|W_\tau|^q\Big)^{1/q} \lesssim \Big(\frac{1}{\sqrt{s}}\gamma_2(\mathcal{T},d_2)+\frac{1}{s}\gamma_1(\mathcal{T},d_1)\Big)+\sqrt{q}\frac{\sigma}{\sqrt{s}}+q\frac{K}{s}.$$

In particular, there exist constants $c, C > 0$ *such that for any* $u \ge 1$,

$$\mathrm{P}\Big(\sup_{\tau\in\mathcal{T}}|W_\tau| \ge C\Big(\frac{1}{\sqrt{s}}\gamma_2(\mathcal{T},d_2)+\frac{1}{s}\gamma_1(\mathcal{T},d_1)\Big)+c\Big(\frac{\sigma}{\sqrt{s}}\sqrt{u}+\frac{K}{s}u\Big)\Big) \le e^{-u}.$$

LEMMA A.3 (Burkholder (1988); Rio (2009))**.** *Let* $q > 1$, $q' = \min(q, 2)$. *Let* $M_n = \sum_{t=1}^n \xi_t$, *where* $\xi_t \in \mathcal{L}^q$ *(i.e.,* $\|\xi_t\|_q < \infty$*) are martingale differences. Then*

$$\|M_n\|_q^{q'} \le K_q^{q'}\sum_{t=1}^n \|\xi_t\|_q^{q'} \quad \textit{where} \quad K_q = \max((q-1)^{-1}, \sqrt{q-1}).$$

LEMMA A.4 (Tail Probabilities for High-dimensional Partial Sums)**.** *For a mean zero* p*-dimensional random variable* $X_t \in \mathbb{R}^p$ $(p > 1)$, *let* $T_n = \sum_{t=1}^n X_t$ *and* $T_{n,m} = \sum_{t=1}^n X_{t,m}$, *where* $X_{t,m} = \mathsf{E}(X_t \mid \varepsilon_{t-m},\dots,\varepsilon_t)$. *Assume that* $\Phi_{\psi_\nu,\varsigma} = \max_{1\le j\le p}\sup_{q\ge 2} q^{-\nu}\|X_{j,\cdot}\|_{q,\varsigma} < \infty$ *for some* $\nu \ge 0$, *and let* $\gamma = 2/(1+2\nu)$. *Then for all* $x > 0$, *we have*

$$\mathrm{P}(|T_n - T_{n,m}|_\infty \ge x) \lesssim p\exp\{-C_\gamma x^\gamma m^{\varsigma\gamma}/(\sqrt{n}\Phi_{\psi_\nu,\varsigma})^\gamma\},$$

where C_γ *is a positive constant only depends on* γ.

Lemma A.4 follows from Lemma C.3 of Zhang and Wu (2017) and applying the Bonferroni inequality. In particular, $\nu = 1$ corresponds to the sub-exponential case, and $\nu = 1/2$ corresponds to the sub-Gaussian case.

A.1.1. *Relaxing the Tail Assumption.* Lemma A.1 primarily focuses on sub-Gaussian random variables. Now, we explore ways to ease this assumption. Specifically, we examine the case of $m = 1$, i.e., A is a vector of dimension $n \times 1$.

THEOREM A.1. *Assume that* $x_i, y_i \in \mathcal{L}^q$ *for some* $q > 2$, $\sup_{A\in\mathcal{A}}|A|_2 \lesssim \sqrt{n}c_n$. *Then, we have with probability* $1 - o(1)$,

$$\sup_{A\in\mathcal{A}}|\boldsymbol{x}^\top AA^\top \boldsymbol{y}| \lesssim (E+\sqrt{V}+U)n^{2r}\gamma_n + (n^{-(q-2)r/2+1}c_n)^2\gamma_n,$$

where E, V, U *are defined in Lemma A.1,* $0 < r \le 1/2$, *and* γ_n *is a slowly growing sequence of positive constants.*

Proof. We proceed without loss of generality by assuming $\boldsymbol{x} = \boldsymbol{y}$. Let $\boldsymbol{z} = (z_1,\dots,z_n)^\top$, where $z_i = x_i\mathbf{1}(|x_i| \le M)$ represents the truncated random variables, with $M = cn^r$ for constants $c > 0$ and $0 < r \le 1/2$. Denote $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$, where $w_i = x_i \mathbf{1}(|x_i| > M) = x_i - z_i$. It follows that

$$\begin{aligned}\sup_{A\in\mathcal{A}} |\boldsymbol{x}^\top AA^\top \boldsymbol{x}| &\le \sup_{A\in\mathcal{A}} |\boldsymbol{z}^\top AA^\top \boldsymbol{z}| + \sup_{A\in\mathcal{A}} |\boldsymbol{w}^\top AA^\top \boldsymbol{z}| + \sup_{A\in\mathcal{A}} |\boldsymbol{x}^\top AA^\top \boldsymbol{w}| \\ &\le \sup_{A\in\mathcal{A}} |\boldsymbol{z}^\top AA^\top \boldsymbol{z}| + 2|\boldsymbol{w}|_2 \sup_{A\in\mathcal{A}} |AA^\top \boldsymbol{x}|_2 \\ &\le \sup_{A\in\mathcal{A}} |\boldsymbol{z}^\top AA^\top \boldsymbol{z}| + 2|\boldsymbol{w}|_2 \sup_{A\in\mathcal{A}} |A|_2 \sup_{A\in\mathcal{A}} |A^\top \boldsymbol{x}|.\end{aligned}$$

Utilizing Lemma A.1, we bound the first term as $\sup_{A\in\mathcal{A}} |\boldsymbol{z}^\top AA^\top \boldsymbol{z}| \lesssim_{\mathrm{P}} (E + \sqrt{V} + U) n^{2r}\gamma_n$, where γ_n is a sequence of positive numbers growing slowly. Moreover, by Markov inequality, we have

$$\mathrm{P}(|\boldsymbol{w}|_2^2 > s^2) \le n\, \mathsf{E}[x_i^2 \mathbf{1}(|x_i| > M)]/s^2 \le n\, \mathsf{E}\, |x_i|^q/(s^2 M^{q-2}),$$

Note that $\mathsf{E}\,|x_i|^q$ is bounded for some $q > 2$. By letting $s^2 = n^{-(q-2)r+1}\gamma_n$, we have the tail probability tends to zero as $n \to \infty$, that is $|\boldsymbol{w}|_2^2 \lesssim_{\mathrm{P}} n^{-(q-2)r+1}\gamma_n$. Given $\sup_{A\in\mathcal{A}} |A|_2 \lesssim \sqrt{n} c_n$, it follows that

$$\sup_{A\in\mathcal{A}} |\boldsymbol{x}^\top AA^\top \boldsymbol{x}| \lesssim_{\mathrm{P}} (E + \sqrt{V} + U) n^{2r}\gamma_n + (n^{-(q-2)r/2+1} c_n)^2 \gamma_n.$$

□

When $r = \log\log n / \log n$ (implying $n^r = \log n$), the second term in the bound becomes $(\log n)^{-(q-2)+2/r} c_n^2 \gamma_n$. For sufficiently large q, this term would be dominated by the first term, resulting in a similar rate as in the sub-Gaussian case, albeit subject to a scaling factor up to a logarithmic order.

A.1.2. *Relaxing the Independence Assumption.* Consider two processes $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ (with $x_i, y_i \in \mathbb{R}$), both of which are stationary with zero mean and unit variance, and admit the representations: $x_i = f(\mathcal{F}_i)$ and $y_i = g(\mathcal{F}_i)$, where f, g are measurable functions, and $\mathcal{F}_i \overset{\text{def}}{=} \{\ldots, \eta_{i-1}, \eta_i\}$, with η_i for $i \in \mathbb{Z}$ being i.i.d. random elements.

Define the projector operator $\mathcal{P}_\ell(x_i y_j) \overset{\text{def}}{=} \mathsf{E}(x_i y_j \mid \mathcal{F}_{s-\ell}) - \mathsf{E}(x_i y_j \mid \mathcal{F}_{s-\ell-1})$, where $s = \min(i,j)$. Note that $\mathcal{P}_\ell(x_i y_j)$ is m.d.s. with respect to $\mathcal{F}_{s-1}$. For $q \ge 1$, $\varsigma > 0$, we introduce the norm

$$\Theta_{q,\varsigma} \overset{\text{def}}{=} \sup_{d\ge 0} (d+1)^\varsigma \sum_{\ell \ge d} \max_{1\le i,j\le n} \|\mathcal{P}_\ell(x_i y_j)\|_q,$$

to measure the degree of dependence. This norm is directly linked to the dependence adjusted norm for $x_i y_j$. Additionally, we denote a truncation argument as $x_i^m \overset{\text{def}}{=} \mathsf{E}(x_i \mid \eta_{i-m}, \ldots, \eta_i)$.

Assumption A.1. *(i) Assume that* $\sup_{q\ge 2} q^{-\nu}\Theta_{q,\varsigma} < \infty$ *for some* $\nu \ge 0$, $\varsigma > 0$.

(ii) The sub-Gaussian norm $\|x_i y_j\|_{\psi_{1/2}} < \infty$, *for all* $i, j = 1, \ldots, n$.

(ii') The sub-exponential norm $\|x_i y_j\|_{\psi_1} < \infty$, *for all* $i, j = 1, \ldots, n$.

(iii) There exists a finite set $\mathcal{A} = \{\delta \in \mathbb{R}^n : |\delta|_\infty \le c_{\max}\}$, for some $c_{\max} > 0$, such that

$$\int_0^\infty \big(\log \mathcal{N}(\epsilon, \mathcal{A}, d_\infty)\big)^{1/2} d\epsilon \lesssim (\log N_n)^{1/2},$$

where N_n is a sequence of positive constants greater than 1.

THEOREM A.2. *Under Assumption A.1 with (ii), we have*

$$\sup_{\delta \in \mathcal{A}} \Big| \sum_{i=1}^n \sum_{j=1}^n \delta_i x_i \delta_j y_j \Big| / c_{\max}^2 \lesssim_{\mathrm{P}} n + \log N_n + \sqrt{n} (\log N_n)^{1/2} + n m^{-\varsigma} \log N_n.$$

Under Assumption A.1 with (ii'), we have

$$\sup_{\delta \in \mathcal{A}} \Big| \sum_{i=1}^n \sum_{j=1}^n \delta_i x_i \delta_j y_j \Big| / c_{max}^2 \lesssim_{\mathrm{P}} \{n + \log N_n + \sqrt{n} (\log N_n)^{1/2}\} (\log n)^2 \gamma_n + n m^{-\varsigma} (\log N_n)^{3/2},$$

where γ_n is a slowly growing sequence of positive constants, see the proof of Theorem A.1.

Proof. Step 1: First, we need to prove that the deviation $\sup_{\delta \in \mathcal{A}} \big| \sum_{i=1}^n \sum_{j=1}^n (\delta_i x_i \delta_j y_j - \delta_i x_i^m \delta_j y_j^m) \big|$, subject to the truncation error, is sufficiently small.

For each $i = 1, \ldots, n$, define $z_i(\delta) \overset{\text{def}}{=} \sum_{j=1}^n (\delta_i x_i \delta_j y_j - \delta_i x_i^m \delta_j y_j^m)$. On Assumption A.1(i), by applying Lemma A.4 on the summation $\sum_{i=1}^n z_i(\delta)$, we obtain that

$$\sup_{\delta \in \mathcal{A}} \Big| \sum_{i=1}^n \sum_{j=1}^n (\delta_i x_i \delta_j y_j - \delta_i x_i^m \delta_j y_j^m) \Big| \lesssim_{\mathrm{P}} n c_{\max}^2 m^{-\varsigma} (\log N_n)^{1/\gamma},$$

where $\gamma = 1$ for the sub-Gaussian case, and $\gamma = 2/3$ for the sub-exponential case.

Step 2: Next, we shall bound the sum of the truncated terms. Divide the sample $\{1, \ldots, n\}$ into $L = \lfloor n/m \rfloor$ blocks: A_l, $l = 1, \ldots, L$, each of size m. Without loss of generality, assume that L is an even number and let $B_o, B_e \subseteq \{1, \ldots, L\}$ be the indices sets for the odd and even blocks, respectively.

For each block $l = 1, \ldots, L$, define $\tilde{x}_l^m(\delta) \overset{\text{def}}{=} \sum_{i \in A_l} \delta_i x_i^m$, $\tilde{y}_l^m(\delta) \overset{\text{def}}{=} \sum_{i \in A_l} \delta_i y_i^m$. It follows that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \delta_i x_i^m \delta_j y_j^m &= \sum_{l=1}^L \sum_{l'=1}^L \sum_{i \in A_l} \sum_{j \in A_{l'}} \delta_i x_i^m \delta_j y_j^m \\ &= \sum_{l \in B_o} \sum_{l' \in B_o} \tilde{x}_l^m(\delta) \tilde{y}_{l'}^m(\delta) + \sum_{l \in B_o} \sum_{l' \in B_e} \tilde{x}_l^m(\delta) \tilde{y}_{l'}^m(\delta) + \sum_{l \in B_e} \sum_{l' \in B_e} \tilde{x}_l^m(\delta) \tilde{y}_{l'}^m(\delta). \end{aligned}$$

It is worth noting that $\{\tilde{x}_l^m(\delta)\}_{l \in B_o}$ and $\{\tilde{x}_l^m(\delta)\}_{l \in B_e}$ are sequences of independent random variables, similarly for $\{\tilde{y}_l^m(\delta)\}_{l \in B_o}$ and $\{\tilde{y}_l^m(\delta)\}_{l \in B_e}$. We shall apply Lemma A.1 to bound each of the terms above. Taking the first term as an example, we denote a scaling constant $c_l(\delta) > 0$ such that $\mathrm{Var}(\tilde{x}_l^m(\delta)/c_l(\delta)) = \mathrm{Var}(\tilde{y}_l^m(\delta)/c_l(\delta)) = 1$, and stack them into a vector

$\boldsymbol{c}(\delta) = (c_l(\delta))_{l\in B_o}$. Using $\boldsymbol{c}(\delta)$ as the weights in the quadratic form, it follows that for any $u \geq 0$, there exist $c_1, c_2 > 0$ such that

$$\mathrm{P}\Big(\sup_{\delta\in\mathcal{A}}\Big|\sum_{l\in B_o}\sum_{l'\in B_o}\tilde{x}_l^m(\delta)\tilde{y}_{l'}^m(\delta)\Big|/c_{\max}^2 \geq c_1 E + u\Big) \leq 2\exp\big(-c_2\min(u^2/V, u/U)\big),$$

where $E = n + \log N_n + \sqrt{n}(\log N_n)^{1/2}$, $V = n + \sqrt{n}(\log N_n)^{1/2}$, $U = n$.

To prove the case of (ii'), we just need to replace Lemma A.1 with Theorem A.1. In particular, we choose r such that $n^r = \log n$ and assume the moments condition holds with a sufficiently high order to absorb the second term in the bound.

By combining the results of Step 1 and Step 2, we can conclude the proof. $\square$

For a class of measurable functions $\mathcal{G}$ mapping to the real space $\mathbb{R}$. let the d_1-metric and d_2-metric be denoted as $d_1(f,g) = \max_{1\leq i\leq n}\|f(x_i) - g(x_i)\|_{\psi_1}$ and $d_2(f,g) = \{\mathsf{E}\,\|f(x_i) - g(x_i)\|_{\psi_1}^2\}^{1/2}$, $\forall f, g \in \mathcal{G}$. In the following theorem, we will derive a bound for the empirical process $\sup_{f\in\mathcal{G}}|n^{-1}\sum_{i=1}^n\{f(x_i) - \mathsf{E}\,f(x_i)\}|$ under certain conditions.

Consider analogous definitions as mentioned earlier:

$$\mathcal{P}_\ell(f(x_i)) \overset{\text{def}}{=} \mathsf{E}(f(x_i) \mid \mathcal{F}_{i-\ell}) - \mathsf{E}(f(x_i) \mid \mathcal{F}_{i-\ell-1}),\ \Theta_{f,q,\varsigma} \overset{\text{def}}{=} \sup_{d\geq 0}(d+1)^\varsigma\sum_{\ell\geq d}\max_{1\leq i\leq n}\|\mathcal{P}_\ell(f(x_i))\|_q,$$

$$f^m(x_i) \overset{\text{def}}{=} \mathsf{E}(f(x_i) \mid \eta_{i-m},\ldots,\eta_i).$$

Assumption A.2. *(i) The function class $\mathcal{G}$ is enveloped with $F = \sup_{f\in\mathcal{G}}|f|$, with $\max_{1\leq i\leq n}(\mathsf{E}\,|F(x_i)|^2)^{1/2} < c_n$. Additionally, assume that there exists a sequence of positive constants (greater than 1) N_n, such that*

$$\int_0^\infty \log\mathcal{N}(\epsilon,\mathcal{G},d_1)d\epsilon \lesssim \log N_n, \quad \int_0^\infty \big(\log\mathcal{N}(\epsilon,\mathcal{G},d_2)\big)^{1/2}d\epsilon \lesssim (\log N_n)^{1/2}.$$

(ii) For any $f\in\mathcal{G}$, assume that $\sup_{q\geq 1} q^{-\nu}\Theta_{f,q,\varsigma} < \infty$ for some $\nu \geq 0$, $\varsigma > 0$.

(iii) There exist constants $\sigma, K > 0$ such that

$$\sup_{f\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \mathsf{E}\,|f(x_i) - \mathsf{E}\,f(x_i)|^q \leq \frac{q!}{2}\sigma^2K^{q-2}, \quad (q = 2,3,\ldots).$$

THEOREM A.3. *Under Assumption A.2, we have*

$$\sup_{f\in\mathcal{G}}\Big|\frac{1}{n}\sum_{i=1}^n\{f(x_i) - \mathsf{E}\,f(x_i)\}\Big|/c_n \lesssim_{\mathrm{P}} m^{-\varsigma}(\log N_n)^{1/\gamma}/\sqrt{n} + \sqrt{(\log N_n)/n} + \sqrt{m}(\log N_n)/n,$$

where $\gamma = 1$ for the sub-Gaussian case, and $\gamma = 2/3$ for the sub-exponential case.

Proof. To begin, we decompose the process as:

$$n^{-1}\sum_{i=1}^n\{f(x_i) - \mathsf{E}\,f(x_i)\} = n^{-1}\sum_{i=1}^n\{f(x_i) - f^m(x_i) + f^m(x_i) - \mathsf{E}\,f^m(x_i)\}.$$

In the subsequent two steps, we will analyze each part of the deviations.

Step 1: Given Assumptions A.2(i)-(ii), applying Lemma A.4 yields:

$$\sup_{f\in\mathcal{G}} |n^{-1}\sum_{i=1}^{n}\{f(x_i)-f^m(x_i)\}| \lesssim_{\mathrm{P}} n^{-1/2}m^{-\varsigma}(\log N_n)^{1/\gamma}c_n,$$

where $\gamma=1$ for the sub-Gaussian case, and $\gamma=2/3$ for the sub-exponential case.

Step 2: We partition the sample into blocks. The definitions of L, A_l, B_o, and B_e remain consistent with those in the proof of Theorem A.2. For each block $l=1,\dots,L$, define $z_{l,f}^m \overset{\text{def}}{=} \sum_{i\in A_l}\{f^m(x_i)-\mathsf{E}\, f^m(x_i)\}$. It follows that

$$\sum_{i=1}^{n}\{f^m(x_i)-\mathsf{E}\, f^m(x_i)\} = \sum_{l=1}^{L} z_{l,f}^m = \sum_{l\in B_o} z_{l,f}^m + \sum_{l\in B_e} z_{l,f}^m.$$

Note that $\{z_{l,f}^m\}_{l\in B_o}$ and $\{z_{l,f}^m\}_{l\in B_o}$ are sequences of independent random variables. We shall apply Lemma A.2 to bound each of them.

Recalling the definition of the projector operation and utilizing Lemma A.3, we observe that

$$\begin{aligned}
\|z_{l,f}^m\|_q &= \Big\|\sum_{\ell\geq 0}\mathcal{P}_\ell(z_{l,f}^m)\Big\|_q\\
&\leq \sum_{\ell\geq 0}\Big\|\sum_{i\in A_l}\mathcal{P}_\ell(f^m(x_i)-\mathsf{E}\, f^m(x_i))\Big\|_q\\
&\lesssim \sqrt{m}\sum_{\ell\geq 0}\max_{1\leq i\leq n}\|\mathcal{P}_\ell(f^m(x_i))\|_q \leq \sqrt{m}\Theta_{f,q,\varsigma}.
\end{aligned}$$

Then, after proper scaling on the block sums, we can evoke Lemma A.2 to obtain that

$$\sup_{f\in\mathcal{G}}\Big|n^{-1}\sum_{l=1}^{L} z_{l,f}^m\Big|/c_n \lesssim_{\mathrm{P}} \sqrt{\log N_n}/\sqrt{n}+\sqrt{m}\log N_n/n.$$

By combining the results of Step 1 and Step 2, we can conclude the proof. □

A.2. **Proofs of Section 3.1.**

A.2.1. *The RE Condition.*

Proof of Lemma 3.1. Recall that δ_{t,J_t} is a sub-vector of δ_t corresponding to the indices set J_t. It suffices to show that the event

$$\min_{\delta_t\neq 0, |\delta_t|_0\leq s_t^*, |\delta_{t,J_t^c}|_1\leq c_0|\delta_{t,J_t}|_1}\frac{|\boldsymbol{V}_t\delta_t|_2}{\sqrt{N}|\delta_t|_2}\geq \kappa_t(c_0,s_t^*)$$

occurs with probability $1-\mathcal{o}(1)$ as it is less likely to hold than the original event required for identification.

Consider the sphere $\Omega_t(c_0, s_t^*)$ defined in Assumption 3.3. To simplify notation, we will henceforth refer to it as Ω_t. Let Ω_t^ϵ be the ϵ-net of Ω_t with respect to $|\cdot|_2$. According to Rudelson and Zhou (2012), the cardinality of Ω_t^ϵ is bounded as $|\Omega_t^\epsilon| \lesssim \binom{m_t}{s_t^*}(c/\epsilon)^{s_t^*} \le \{cem_t/(s_t^*\epsilon)\}^{s_t^*}$, where m_t is the dimension of V_{it} (i.e. the number of instruments for each period t), and $c > 0$ is an absolute constant. Moreover, for any point $\delta \in \Omega_t$, let $\pi(\delta)$ denote the closest point to δ within Ω_t^ϵ.

We first observe that

$$\begin{aligned}\min_{\delta_t\in\Omega_t} |\boldsymbol{V}_t\delta_t|_2^2 &\ge \min_{\delta_t\in\Omega_t} |N\delta_t^\top \bar{\mathsf{E}}(V_{it}V_{it}^\top)\delta_t| - \max_{\delta_t\in\Omega_t} |\delta_t^\top\{\boldsymbol{V}_t^\top\boldsymbol{V}_t - N\bar{\mathsf{E}}(V_{it}V_{it}^\top)\}\delta_t| \\ &\ge \min_{\delta_t\in\Omega_t} |N\delta_t^\top \bar{\mathsf{E}}(V_{it}V_{it}^\top)\delta_t| - \max_{\delta_t\in\Omega_t} |\boldsymbol{V}_t\delta_t|_2^2 - \max_{\delta_t\in\Omega_t} |N\delta_t^\top \bar{\mathsf{E}}(V_{it}V_{it}^\top)\delta_t|.\end{aligned}$$

Next, we shall derive the upper bound for the second term. By defining $C_t \stackrel{\text{def}}{=} \max_{\delta_t\in\Omega_t} |\boldsymbol{V}_t(\delta_t - \pi(\delta_t))|_2$, and $D_t \stackrel{\text{def}}{=} \max_{\pi(\delta_t)\in\Omega_t^\epsilon} |\boldsymbol{V}_t\pi(\delta_t)|_2$, we have the following inequalities

$$D_t - C_t \le \max_{\delta_t\in\Omega_t} |\boldsymbol{V}_t\delta_t|_2 \le D_t + C_t.$$

Since $C_t \le \epsilon \max_{\delta_t\in\Omega_t} |\boldsymbol{V}_t\delta_t|_2$, we can further obtain that

$$D_t/(1+\epsilon) \le \max_{\delta_t\in\Omega_t} |\boldsymbol{V}_t\delta_t|_2 \le D_t/(1-\epsilon).$$

To bound D_t, we apply the tail probability inequality in Lemma A.2 on the mean zero random variable $z_{it}(\delta_t) \stackrel{\text{def}}{=} \{\pi(\delta_t)^\top V_{it}\}^2 - \bar{\mathsf{E}}\{\pi(\delta_t)^\top V_{it}\}^2$, over all $\pi(\delta_t) \in \Omega_t^\epsilon$. It follows that $D_t^2 \lesssim_{\mathrm{P}} \sqrt{N}\{s_t^* \log(cem_t/(s_t^*\epsilon))\}^{1/2}$.

By choosing a sufficiently small ϵ, e.g., $\epsilon = 1/(\log s_t^*)^{1/4}$, and given Assumption 3.3 with the particular $\kappa_t(\cdot)$ specified in the conditions of the lemma, we have

$$\min_{\delta_t\in\Omega_t} N^{-1/2}|\boldsymbol{V}_t\delta_t|_2 = \min_{\delta_t\neq 0, |\delta_t|_0\le s_t^*, |\delta_{t,J_t^c}|_1\le c_0|\delta_{t,J_t}|_1} \frac{|\boldsymbol{V}_t\delta_t|_2}{\sqrt{N}|\delta_t|_2} \ge \kappa_t(c_0, s_t^*),$$

which holds with probability $1 - \mathcal{o}(1)$, as $N \to \infty$. □

A.2.2. *Oracle Order of s_t^*.* Recall that V_{it} is a vector of length m_t that gathers the instruments for each $t = 2, \dots, T$. Let $J_t \subseteq \{1, \dots, m_t\}$ be a set of indices with cardinality $|J_t| \le s_t \le m_t$, and let $J_t^c = \{k \in \{1, \dots, m_t\} : k \notin J_t\}$ be the complement set. Denote V_{it,J_t} and V_{it,J_t^c} as the sub-vectors of V_{it} corresponding to J_t and J_t^c, respectively. Consider the true model

$$W_{it} = V_{it}^\top\Pi_t^0 + \eta_{it} = V_{it,J_t}^\top\Pi_{t,J_t}^0 + \underbrace{V_{it,J_t^c}^\top\Pi_{t,J_t^c}^0 + \eta_{it}}_{=:\check{\eta}_{it}},$$

where $\Pi_t^0 \stackrel{\text{def}}{=} \bar{\mathsf{E}}(V_{it}V_{it}^\top)^{-1}\bar{\mathsf{E}}(V_{it}W_{it})$, Π_{t,J_t}^0 and $\Pi_{t,J_t^c}^0$ are the sub-vectors of Π_t^0 corresponding to J_t and J_t^c, respectively. Let $\delta_{-J_t}^0 \stackrel{\text{def}}{=} \Pi_t^0 - \Pi_{J_t}^0$, where $\Pi_{J_t}^0$ is a vector of length m_t with elements corresponding to J_t being Π^0_{t,J_t} and zeros elsewhere. Moreover, define $\Pi^\dagger_{t,J_t} \overset{\text{def}}{=} \bar{\mathsf{E}}(V_{it,J_t}V_{it,J_t}^\top)^{-1}\bar{\mathsf{E}}(V_{it,J_t}W_{it})$, and $\widehat{\Pi}_{t,J_t} \overset{\text{def}}{=} \big(\sum_{i=1}^N V_{it,J_t}V_{it,J_t}^\top\big)^{-1}\big(\sum_{i=1}^N V_{it,J_t}W_{it}\big)$.

To choose the optimal value of s_t, we consider the oracle risk minimization problem:

$$s_t^* = \arg\min_{s_t}\Big\{\min_{J_t:|J_t|\le s_t} N^{-1}\sum_{i=1}^N (V_{it}^\top\Pi^0_t - V_{it,J_t}^\top\Pi^\dagger_{t,J_t})^2 + s_t\check{\sigma}_t^2/N\Big\}, \tag{A.1}$$

where $\check{\sigma}_t^2 \overset{\text{def}}{=} \bar{\mathsf{E}}(\check{\eta}_{it}^2 \mid V_{it,J_t})$. In the following theorem, we shall illustrate the oracle order of s_t^* for a specific case.

THEOREM A.4 (Oracle Order of s_t^*)**.** *Under Assumptions 3.1-3.2, and assuming that* $\bar{\mathsf{E}}(V_{it}^\top\delta^0_{-J_t}) \lesssim c^{-s_t}$ *for some constant* $c>0$*, we can conclude that the optimal* s_t^* *defined in* (A.1) *is bounded as* $s_t^* \asymp \log N \wedge t$.

Proof. Let $e_{it} \overset{\text{def}}{=} V_{it,J_t}^\top\big(\sum_{i=1}^N V_{it,J_t}V_{it,J_t}^\top\big)^{-1}\big(\sum_{i=1}^N V_{it,J_t}\check{\eta}_{it}\big)$. Observe that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \mathsf{E}\{(V_{it}^\top\Pi^0_t - V_{it,J_t}^\top\widehat{\Pi}_{t,J_t})^2 \mid V_{it,J_t}\}\\
=\;& \frac{1}{N}\sum_{i=1}^N \mathsf{E}\bigg(\bigg[V_{it}^\top\Pi^0_t - V_{it,J_t}^\top\Big(\sum_{i=1}^N V_{it,J_t}V_{it,J_t}^\top\Big)^{-1}\Big(\sum_{i=1}^N V_{it,J_t}(V_{it,J_t}^\top\Pi^0_{t,J_t} + \check{\eta}_{it})\Big)\bigg]^2 \mid V_{it,J_t}\bigg)\\
=\;& \frac{1}{N}\sum_{i=1}^N \delta^{0\top}_{-J_t}V_{it}V_{it}^\top\delta^0_{-J_t} - \frac{2}{N}\sum_{i=1}^N \mathsf{E}(e_{it}V_{it}^\top\delta^0_{-J_t} \mid V_{it,J_t}) + \frac{1}{N}\sum_{i=1}^N \mathsf{E}(e_{it}^2 \mid V_{it,J_t}).
\end{aligned}$$

By applying the concentration inequality, we can bound the first term in probability by c^{-2s_t}, and the second term as $\frac{1}{N}\sum_{i=1}^N \mathsf{E}(e_{it}V_{it}^\top\delta^0_{-J_t} \mid V_{it,J_t}) \lesssim_{\mathrm{P}} c^{-2s_t} \vee c^{-s_t}\sqrt{\log m_t/N}$. As for the last term, under the cross-sectional independence assumption, we obtain that

$$\frac{1}{N}\sum_{i=1}^N \mathsf{E}(e_{it}^2 \mid V_{it,J_t}) = \frac{1}{N^2}\sum_{i=1}^N \check{\sigma}_t^2\,\mathrm{tr}\bigg(\Big(\sum_{i=1}^N V_{it,J_t}V_{it,J_t}^\top\Big)^{-1}V_{it,J_t}V_{it,J_t}^\top\bigg) + o_{\mathrm{P}}(1) \lesssim_{\mathrm{P}} N^{-1}s_t.$$

Finally, to minimize the order of $c^{-2s_t} + c^{-s_t}\sqrt{\log m_t/N} + N^{-1}s_t$, we find that the oracle order of the optimal s_t^* is approximately given by $s_t^* \asymp \log N \wedge t$. □

It is worth noting that the approximate sparse error, as quantified by (11), is bounded as $C_{s_t^*} \lesssim_{\mathrm{P}} (c^{-2s_t^*} + c^{-s_t^*}\sqrt{\log m_t/N} + N^{-1}s_t^*)^{1/2}$ is this particular case. Therefore, we can further deduce the oracle bound for $C_{s_t^*}$ based on the optimal s_t^* found above.

A.2.3. *Prediction Performance of* $\widehat{\Pi}_t$.

Proof of Theorem 3.1. First observe that

$$|\boldsymbol{W}_t - \boldsymbol{V}_t\widehat{\Pi}_t|_2^2 = |\boldsymbol{V}_t(\Pi_t^0 - \widehat{\Pi}_t) + \boldsymbol{\eta}_t|_2^2 = N|\Pi_t^0 - \widehat{\Pi}_t|_{2,N}^2 + |\boldsymbol{\eta}_t|_2^2 + 2\langle \boldsymbol{V}_t(\Pi_t^0 - \widehat{\Pi}_t), \boldsymbol{\eta}_t\rangle,$$
$$|\boldsymbol{W}_t - \boldsymbol{V}_t\Pi_t^*|_2^2 = |\boldsymbol{V}_t(\Pi_t^0 - \Pi_t^*) + \boldsymbol{\eta}_t|_2^2 = N|\Pi_t^0 - \Pi_t^*|_{2,N}^2 + |\boldsymbol{\eta}_t|_2^2 + 2\langle \boldsymbol{V}_t(\Pi_t^0 - \Pi_t^*), \boldsymbol{\eta}_t\rangle.$$

By the definition of the LASSO estimator, we have

$$|\boldsymbol{W}_t - \boldsymbol{V}_t\widehat{\Pi}_t|_2^2 + \lambda|\boldsymbol{\omega}_t \circ \widehat{\Pi}_t|_1 \le |\boldsymbol{W}_t - \boldsymbol{V}_t\Pi_t^*|_2^2 + \lambda|\boldsymbol{\omega}_t \circ \Pi_t^*|_1,$$

where $\circ$ denotes the Hadamard product. It follows that

$$\begin{aligned}
|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} &\le |\Pi_t^0 - \widehat{\Pi}_t|_{2,N}^2 + |\Pi_t^0 - \Pi_t^*|_{2,N}^2 \\
&= N^{-1}|\boldsymbol{W}_t - \boldsymbol{V}_t\widehat{\Pi}_t|_2^2 - N^{-1}|\boldsymbol{W}_t - \boldsymbol{V}_t\Pi_t^*|_2^2 + 2N^{-1}\langle \boldsymbol{V}_t\delta_{\Pi,t}, \boldsymbol{\eta}_t\rangle \\
&\le 2N^{-1}\langle \boldsymbol{V}_t\delta_{\Pi,t}, \boldsymbol{\eta}_t\rangle + N^{-1}\lambda_t(|\boldsymbol{\omega}_t \circ \Pi_t^*|_1 - |\boldsymbol{\omega}_t \circ \widehat{\Pi}_t|_1).
\end{aligned}$$

If $|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} \le 0$, then we have $|\delta_{\Pi,t}|_{2,N} \le 2C_{s_t^*}$ and the desired bound in the conclusion holds. In the following, we shall derive the bound for the case of $|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} > 0$.

On the event $\mathcal{A}_{2t}$ (with $c = 4$), we have

$$\langle \boldsymbol{V}_t\delta_{\Pi,t}, \boldsymbol{\eta}_t\rangle = \langle \boldsymbol{\eta}_t^\top \boldsymbol{V}_t \oslash \boldsymbol{\omega}_t, \boldsymbol{\omega}_t \circ \delta_{\Pi,t}\rangle \le \lambda_t|\boldsymbol{\omega}_t \circ \delta_{\Pi,t}|_1/4,$$

which implies that

$$\begin{aligned}
&|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} + (2N)^{-1}\lambda_t|\boldsymbol{\omega}_t \circ \delta_{\Pi,t}|_1 \\
\le\;& N^{-1}\lambda_t|\boldsymbol{\omega}_t \circ \delta_{\Pi,t}|_1 + N^{-1}\lambda_t(|\boldsymbol{\omega}_t \circ \Pi_t^*|_1 - |\boldsymbol{\omega}_t \circ \widehat{\Pi}_t|_1).
\end{aligned} \tag{A.2}$$

Recall that J_t is the indices set for nonzero elements in Π_t^*, and J_t^c is the complement set for zero ones. Let δ_{Π,t,J_t} and δ_{Π,t,J_t^c} denote the sub-vectors of $\delta_{\Pi,t}$ corresponding to J_t and J_t^c, similarly for Π_t^*, $\widehat{\Pi}_t$, and $\boldsymbol{\omega}_t$. Then, by (A.2) we obtain that

$$\begin{aligned}
&|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} + (2N)^{-1}\lambda_t|\boldsymbol{\omega}_t \circ \delta_{\Pi,t}|_1 \\
\le\;& N^{-1}\lambda_t|\boldsymbol{\omega}_{t,J_t} \circ \delta_{\Pi,t,J_t}|_1 + N^{-1}\lambda_t|\boldsymbol{\omega}_{t,J_t} \circ \Pi_{t,J_t}^*|_1 + N^{-1}\lambda_t|\boldsymbol{\omega}_{t,J_t} \circ \widehat{\Pi}_{t,J_t}|_1 \\
\le\;& 2N^{-1}\lambda_t|\boldsymbol{\omega}_{t,J_t} \circ \delta_{\Pi,t,J_t}|_1.
\end{aligned}$$

Without loss of generality, we set the penalty wights to be 1. Given $|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} \ge 0$, (A.2) also implies that $\delta_{\Pi,t}$ satisfies $|\delta_{\Pi,t,J_t^c}|_1 \le 3|\delta_{\Pi,t,J_t}|_1$. On the event $\mathcal{A}_{1t}$, we get $|\delta_{\Pi,t}|_{2,N} \ge \kappa_t(3, s_t^*)|\delta_{\Pi,t,J_t}|_2$. Therefore, on the events $\mathcal{A}_{1t}$ and $\mathcal{A}_{2t}$, we have

$$|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} \le 2N^{-1}\sqrt{s_t^*}\lambda_t|\delta_{\Pi,t,J_t}|_2 \le 2N^{-1}\sqrt{s_t^*}\lambda_t|\delta_{\Pi,t}|_{2,N}/\kappa(3, s_t^*),$$

which gives the bound for prediction norm:

$$|\delta_{\Pi,t}|_{2,N} \le 2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa(3, s_t^*).$$

Next, we shall derive the bound for $|\delta_{\Pi,t}|_1$. When $|\delta_{\Pi,t,J_t^c}|_1 \leq 6|\delta_{\Pi,t,J_t}|_1$ is satisfied, we have

$$|\delta_{\Pi,t}|_1 \leq 7|\delta_{\Pi,t,J_t}|_1 \leq 7\sqrt{s_t^*}|\delta_{\Pi,t,J_t}|_2 \leq 7\sqrt{s_t^*}|\delta_{\Pi,t}|_{2,N}/\kappa(3, s_t^*).$$

When $|\delta_{\Pi,t,J_t^c}|_1 > 6|\delta_{\Pi,t,J_t}|_1$, we have $|\delta_{\Pi,t}|_{2,N}^2 - 2C_{s_t^*}|\delta_{\Pi,t}|_{2,N} < 0$. Then, by (A.2) we can find that $C_{s_t^*}^2 \geq \frac{1}{16}N^{-1}\lambda_t|\delta_{\Pi,t,J_t^c}|_1$, and thus

$$|\delta_{\Pi,t}|_1 < \frac{7}{6}|\delta_{\Pi,t,J_t^c}|_1 \leq \frac{56}{3}NC_{s_t^*}^2/\lambda_t.$$

Overall, $|\delta_{\Pi,t}|_1$ is bounded as

$$|\delta_{\Pi,t}|_1 \leq 7\sqrt{s_t^*}\{2C_{s_t^*} + 2N^{-1}\sqrt{s_t^*}\lambda_t/\kappa_t(3, s_t^*)\}/\kappa_t(3, s_t^*) + 56NC_{s_t^*}^2/(3\lambda_t).$$

$\square$

A.3. **Proofs of Section 3.2.**

A.3.1. *Asymptotic Normality of AB-LASSO.*

Proof of Theorem 3.2. Recall the expression

$$\widehat{\theta} - \theta^0 = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathbf{\Theta}}_t V_{it}\Delta X_{it}^\top\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathbf{\Theta}}_t V_{it}\Delta\varepsilon_{it}\right),$$

and rewrite the first summation as

$$\begin{aligned}
& (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathbf{\Theta}}_t V_{it}\Delta X_{it}^\top \\
= \quad & (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*)V_{it}\Delta X_{it}^\top + (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{\Theta}_t^*\{V_{it}\Delta X_{it}^\top - \mathsf{E}(V_{it}\Delta X_{it}^\top)\} \\
& + (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{\Theta}_t^*\,\mathsf{E}(V_{it}\Delta X_{it}^\top) \\
=: \quad & I_1 + I_2 + I_3
\end{aligned}$$

To deal with the inverse of the sum, we consider the following expansion:

$$\begin{aligned}
(I_1 + I_2 + I_3)^{-1} &= [I_3\{\mathbf{I}_{d\times d} + I_3^{-1}(I_1 + I_2)\}]^{-1} \\
&= \left[\sum_{k=0}^{\infty}\big\{-I_3^{-1}(I_1 + I_2)\big\}^k\right]I_3^{-1} = \{\mathbf{I}_{d\times d} - I_3^{-1}(I_1 + I_2) + R_n\}I_3^{-1},
\end{aligned}$$

where $\mathbf{I}_{d\times d}$ represents the $d \times d$ identity matrix and R_n denotes the remainder of order $\mathcal{O}_{\mathrm{P}}((NT)^{-1/2})$. We observe an approximate sparse error between $\mathbf{\Theta}_t^0$ and $\mathbf{\Theta}_t^*$, with the average error rate being of a small order $T^{-1}\sum_{t=2}^{T}|\mathbf{\Theta}_t^* - \mathbf{\Theta}_t^0|_\infty = \mathcal{O}(1)$. This generally holds true under a decaying temporal dependence structure. The error is deemed negligible in the subsequent asymptotic analysis. Base on Assumption 3.5, we assert that I_3 is inverterable in the limit, with the presence of only a negligible error.

The second summation in the expression of $(\widehat{\theta} - \theta^0)$ is rewritten as

$$\begin{aligned} &(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathbf{\Theta}}_t V_{it}\Delta\varepsilon_{it} \\ = \;& (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*)V_{it}\Delta\varepsilon_{it} + (NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{\Theta}_t^* V_{it}\Delta\varepsilon_{it} =: P_1 + P_2 \end{aligned}$$

It follows that

$$\widehat{\theta} - \theta^0 = I_3^{-1}P_2 + I_3^{-1}P_1 - I_3^{-1}(I_1 + I_2)I_3^{-1}(P_1 + P_2) + o_{\mathrm{P}}((NT)^{-1/2}). \tag{A.3}$$

We will prove that $I_3^{-1}P_2$ is the leading term, with $\sqrt{NT}I_3^{-1}P_2$ exhibiting asymptotic Gaussianity and analyze the orders of the other terms.

We refer to a central limit theorem for stationary random field (Theorem 1 of El Machkouri et al. (2013)) to establish asymptotic normality. To achieve this, we must verify the necessary conditions outlined below.

Define an index set $\mathcal{J}_{N,T} \overset{\text{def}}{=} \{(i,t) : 1 \le i \le N, 2 \le t \le T\}$. As $N, T \to \infty$, it follows that the cardinality $|\mathcal{J}_{N,T}| \to \infty$, while the ratio $|\partial\mathcal{J}_{N,T}|/|\mathcal{J}_{N,T}| \to 0$, where $\partial\mathcal{J}_{N,T}$ contains the boundary points of $\mathcal{J}_{N,T}$. Under Assumption 3.1, the d-dimensional process $Z_{(i,t)} \overset{\text{def}}{=} \mathbf{\Theta}_t^* V_{it}\Delta\varepsilon_{it}$ is stationary over t and i.i.d. across i. For $k = 1, \dots, d$, $Z_{(i,t),k}$ can be represented as $Z_{(i,t),k} = h_{(i,t),k}(\dots, \eta_{(i,t-1)}, \eta_{(i,t)})$, where $h_{(i,t),k}$ are measurable functions, and $\eta_{(i,t)}$ for $i \in \mathbb{N}$, $t \in \mathbb{Z}$, are i.i.d. random elements. By Definition 3.1 and Assumption 3.2(i), it follows that

$$\sum_{(i,t)\in\mathcal{J}_{N,T}} \|Z^*_{(i,t),k} - Z_{(i,t),k}\|_2 < \infty.$$

Moreover, Assumption 3.2(ii), along with the cross-sectional independence assumption, implies that for $k, k' = 1, \dots, d$, the variance

$$\begin{aligned} &\mathsf{E}\Big[\Big(\sum_{(i,t)\in\mathcal{J}_{N,T}} Z_{(i,t),k)}\Big)\Big(\sum_{(i,t)\in\mathcal{J}_{N,T}} Z_{(i,t),k'}\Big)\Big] \\ = &\sum_{(i,t)\in\mathcal{J}_{N,T}} \mathsf{E}(Z_{(i,t),k}Z_{(i,t),k'}) \quad + \sum_{(i,t)\in\mathcal{J}'_{N,T}} \mathsf{E}(Z_{(i,t),k}Z_{(i,t-1),k'}) + \sum_{(i,t)\in\mathcal{J}'_{N,T}} \mathsf{E}(Z_{(i,t-1),k}Z_{(i,t),k'}) \end{aligned}$$

is of order NT, where $\mathcal{J}'_{N,T} \overset{\text{def}}{=} \{(i,t) : 1 \le i \le N, 3 \le t \le T\}$. Therefore, based on Assumption 3.5, by applying Theorem 1 of El Machkouri et al. (2013) and Slutsky's theorem, we deduce that

$$\sqrt{NT}I_3^{-1}P_2 \xrightarrow{\mathcal{L}} \mathrm{N}(0, Q^{-1}\mathbf{\Sigma}(Q^{-1})^\top).$$

Recalling the subspace $\Omega_t(c_0, s_t^*)$ defined in Assumption 3.3, and considering the entropy condition (with respect to the d_2-metric):

$$\text{ent}\left(\epsilon, \bigcup_{2\le t\le T} \Omega_t(c_0, s_t^*)\right) \lesssim T \max_{2\le t\le T} s_t^* \log(m_t/\epsilon), \text{ for all } 0<\epsilon<1,$$

by employing Theorem A.3 based on Assumptions 3.1-3.2, we obtain:

$$\sup_{\delta_t\in\Omega_t(c_0,s_t^*),\, t=2,\dots,T} \left|(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\delta_t^\top V_{it}\Delta\varepsilon_{it}\right| \lesssim_{\mathrm{P}} \sqrt{\max_{2\le t\le T} s_t^*\log m_t}/\sqrt{N}. \tag{A.4}$$

Consequently, we bound $|I_3^{-1}P_1|_\infty$ as:

$$|I_3^{-1}P_1|_\infty \le |I_3^{-1}|_\infty \left|(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*)V_{it}\Delta\varepsilon_{it}\right|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T} s_t^*\log m_t/N.$$

Next, we proceed to bound $|I_3^{-1}(I_1+I_2)I_3^{-1}(P_1+P_2)|_\infty$. Observe that

$$\begin{aligned}|I_3^{-1}(I_1+I_2)I_3^{-1}(P_1+P_2)|_\infty &\le |I_3^{-1}I_1I_3^{-1}P_1|_\infty + |I_3^{-1}I_2I_3^{-1}P_1|_\infty \\ &\quad + |I_3^{-1}I_1I_3^{-1}P_2|_\infty + |I_3^{-1}I_2I_3^{-1}P_2|_\infty. \end{aligned} \tag{A.5}$$

We first look at the rate of $|I_3^{-1}I_1I_3^{-1}P_1|_\infty$. By letting $\boldsymbol{D}_t \overset{\text{def}}{=} \widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*$, we have

$$I_3^{-1}I_1I_3^{-1}P_1 = (NT)^{-2}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T} I_3^{-1}\mathbf{\Delta}_t V_{it}\Delta X_{it}^\top I_3^{-1}\boldsymbol{D}_{t'}V_{i't'}\Delta\varepsilon_{i't'}.$$

In particular, for $k=1,\dots,d$, the kth element of $I_3^{-1}I_1I_3^{-1}P_1$ is given by

$$\begin{aligned}
& (NT)^{-2}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[I_3^{-1}]_{k\cdot}\boldsymbol{D}_t V_{it}\Delta X_{it}^\top I_3^{-1}\boldsymbol{D}_{t'}V_{i't'}\Delta\varepsilon_{i't'} \\
= \; & (NT)^{-2}\sum_{k'=1}^{d}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[I_3^{-1}\boldsymbol{D}_t]_{k\cdot}[V_{it}\Delta X_{it}^\top]_{\cdot k'}[I_3^{-1}\boldsymbol{D}_{t'}]_{k'\cdot}V_{i't'}\Delta\varepsilon_{i't'} \\
= \; & (NT)^{-2}\sum_{k'=1}^{d}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[\Delta X_{it}V_{it}^\top]_{k'\cdot}[I_3^{-1}\boldsymbol{D}_t]_{k\cdot}^\top[I_3^{-1}\boldsymbol{D}_{t'}]_{k'\cdot}V_{i't'}\Delta\varepsilon_{i't'} \\
= \; & (NT)^{-2}\sum_{k'=1}^{d}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[\Delta X_{it}V_{it}^\top - \mathsf{E}(\Delta X_{it}V_{it}^\top)]_{k'\cdot}[I_3^{-1}\boldsymbol{D}_t]_{k\cdot}^\top[I_3^{-1}\boldsymbol{D}_{t'}]_{k'\cdot}V_{i't'}\Delta\varepsilon_{i't'} \\
& + (NT)^{-2}\sum_{k'=1}^{d}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[\mathsf{E}(\Delta X_{it}V_{it}^\top)]_{k'\cdot}[I_3^{-1}\boldsymbol{D}_t]_{k\cdot}^\top[I_3^{-1}\boldsymbol{D}_{t'}]_{k'\cdot}V_{i't'}\Delta\varepsilon_{i't'},
\end{aligned}$$

where $[\cdot]_{k\cdot}$ denotes the kth row and $[\cdot]_{\cdot k}$ denotes the kth column of the matrix, respectively.

Consider the class of functions

$$\mathcal{A}_{t,t'} \overset{\text{def}}{=} \{A_{t,t'} = a_t a_{t'}^\top : a_t \in \Omega(c_0, s_t^*), a_{t'} \in \Omega(c_0, s_{t'}^*)\},$$

with the entropy condition (with respect to the d_∞-metric):

$$\mathrm{ent}\left(\epsilon, \bigcup_{2\le t,t'\le T} \mathcal{A}_{t,t'}\right) \lesssim T^2\Big\{\max_{2\le t\le T} s_t^* \log(m_t/\epsilon)\Big\}^2, \text{ for all } 0<\epsilon<1.$$

Applying Theorem A.2 based on Assumptions 3.1-3.2, we find that

$$\begin{aligned}&\sup_{A_{t,t'}\in\mathcal{A}_{t,t'},\, t,t'=2,\dots,T}\left|(NT)^{-2}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[\Delta X_{it}V_{it}^\top - \mathsf{E}(\Delta X_{it}V_{it}^\top)]_{k'\cdot}A_{t,t'}V_{i't'}\Delta\varepsilon_{i't'}\right|\\ \lesssim_{\mathrm{P}}\ & (NT)^{-1}\log^3(NT).\end{aligned}$$

Moreover, combining (A.4) with the fact that

$$\sup_{a_t\in\Omega_t(c_0,s_t^*),\, t=2,\dots,T}\left|(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}[\mathsf{E}(\Delta X_{it}V_{it}^\top)]_{k'\cdot}a_t\right| \lesssim \max_{2\le t\le T} s_t^*\log m_t/\sqrt{N},$$

we achieve:

$$\begin{aligned}&\sup_{A_{t,t'}\in\mathcal{A}_{t,t'},\, t,t'=2,\dots,T}\left|(NT)^{-2}\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{i'=1}^{N}\sum_{t'=2}^{T}[\mathsf{E}(\Delta X_{it}V_{it}^\top)]_{k'\cdot}A_{t,t'}V_{i't'}\Delta\varepsilon_{i't'}\right|\\ \lesssim_{\mathrm{P}}\ & \max_{2\le t\le T}(s_t^*\log m_t)^{3/2}/N.\end{aligned}$$

As a result, we bound $|I_3^{-1}I_1I_3^{-1}P_1|_\infty$ as

$$|I_3^{-1}I_1I_3^{-1}P_1|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T}(s_t^*\log m_t)^{5/2}/N^2.$$

Regarding the other terms on the right-hand side of (A.5), note that $|I_3^{-1}I_2|_\infty = \mathcal{O}_{\mathrm{P}}(1/\sqrt{NT})$ and $|I_3^{-1}P_2|_\infty = \mathcal{O}_{\mathrm{P}}(1/\sqrt{NT})$, which imply that $|I_3^{-1}I_2I_3^{-1}P_2|_\infty = \mathcal{O}_{\mathrm{P}}(1/(NT))$. Combining the bound for $|I_3^{-1}P_1|_\infty$ we have found above, we also deduce that $|I_3^{-1}I_2I_3^{-1}P_1|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T} s_t^*\log m_t/\sqrt{N^3T}$. Lastly, as for $|I_3^{-1}I_1|_\infty$, applying Theorem A.3 gives that

$$\sup_{\delta_t\in\Omega_t(c_0,s_t^*),\, t=2,\dots,T}\left|(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\delta_t^\top\{V_{it}\Delta X_{it}^\top - \mathsf{E}(V_{it}\Delta X_{it}^\top)\}\right|_\infty \lesssim_{\mathrm{P}} \sqrt{\max_{2\le t\le T} s_t^*\log m_t}/\sqrt{N}.$$

Combining this with the fact that

$$\sup_{\delta_t\in\Omega_t(c_0,s_t^*),\, t=2,\dots,T}\left|(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\delta_t^\top\,\mathsf{E}(V_{it}\Delta X_{it}^\top)\right|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T} s_t^*\log m_t/\sqrt{N},$$

we conclude that $|I_3^{-1}I_1|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T}(s_t^*\log m_t)^{3/2}/N$, which implies that $|I_3^{-1}I_1I_3^{-1}P_2|_\infty \lesssim_{\mathrm{P}} \max_{2\le t\le T}(s_t^*\log m_t)^{3/2}/\sqrt{N^3T}$.

Under the condition $\max_{2\le t\le T} s_t^*\log m_t\sqrt{T}/\sqrt{N}\to 0$, we have $\sqrt{NT}|I_3^{-1}P_1|_\infty = o_{\mathrm{P}}(1)$, as well as $\sqrt{NT}|I_3^{-1}(I_1+I_2)I_3^{-1}(P_1+P_2)|_\infty = o_{\mathrm{P}}(1)$. Thus, the proof is concluded. □

A.3.2. *Asymptotic Normality of AB-LASSO-SS.*

Proof of Theorem 3.3. Analogous to (A.3), we derive expansions for $\widehat{\theta}_{A,B} - \theta^0$ and $\widehat{\theta}_{B,A} - \theta^0$, respectively. With sample-splitting, a significant difference arises in the convergence rate of $(NT)^{-1/2} \sum_{i \in \mathbb{I}_s} \sum_{t=2}^{T} (\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*) V_{it} \Delta\varepsilon_{it}$, where $s \in \{A, B\}$. As $\widehat{\mathbf{\Theta}}_t$ is obtained from a sub-sample uncorrelated with the one considered in the summation, we have

$$(NT)^{-1/2} \left| \sum_{i \in \mathbb{I}_s} \sum_{t=2}^{T} (\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^*) V_{it} \Delta\varepsilon_{it} \right|_\infty \lesssim_{\mathrm{P}} \max_{2 \le t \le T} \sqrt{s_t^*} \log m_t / \sqrt{N}.$$

Thus, the required condition reduces to $\max_{2 \le t \le T} \sqrt{s_t^*} \log m_t / \sqrt{N} \to 0$, as $N, T \to \infty$. The remainder of the proof follows that of Theorem 3.2 in a similar manner. □

A.3.3. *Time Effects.* In this subsection, we discuss how the inference theory, such as Theorem 3.2, adapts in the presence of time effects γ_t. We illustrate this by considering an example with a panel AR(1) model:

$$Y_{it} = \alpha_i + \gamma_t + \theta_1 Y_{i,t-1} + \theta_2 D_{it} + \varepsilon_{it}, \quad |\theta_1| < 1. \tag{A.6}$$

We assume that $\mathsf{E}(Y_{is}\varepsilon_{it}) = 0$ for all $1 \le s < t$, and $\mathsf{E}(\Delta D_{it} \Delta\varepsilon_{it}) = 0$. In addition to Assumptions 3.1-3.2, for simplicity, in this particular example, we assume that α_i, γ_t, and D_{it} have zero mean, D_{it} has no serial correlation over t. Let σ_D^2 denote the variance of D_{it}. We also assume that the fourth moment of D_{it} is bounded by a constant.

By recursively substituting the lagged values, the model in (A.6) can be rewritten as:

$$Y_{it} = \alpha_i \sum_{\ell \ge 0} \theta_1^\ell + \sum_{\ell \ge 0} \theta_1^\ell \gamma_{t-\ell} + \theta_2 \sum_{\ell \ge 0} \theta_1^\ell D_{i,t-\ell} + \sum_{\ell \ge 0} \theta_1^\ell \varepsilon_{i,t-\ell}. \tag{A.7}$$

For $Z_{it} \in \{Y_{it}, D_{it}, \varepsilon_{it}\}$, let $\bar{Z}_{\cdot t} = \sum_{i=1}^{N} Z_{it}/N$ (similarly, $\bar{\alpha}_{\cdot} = \sum_{i=1}^{N} \alpha_i / N$), $\widetilde{Z}_{it} = Z_{it} - \bar{Z}_{\cdot t}$, and $\Delta\widetilde{Z}_{it} = \Delta Z_{it} - \sum_{i=1}^{N} \Delta Z_{it}/N$. It follows that

$$\bar{Y}_{\cdot t} = \bar{\alpha}_{\cdot} \sum_{\ell \ge 0} \theta_1^\ell + \sum_{\ell \ge 0} \theta_1^\ell \gamma_{t-\ell} + \theta_2 \sum_{\ell \ge 0} \theta_1^\ell \bar{D}_{\cdot, t-\ell} + \sum_{\ell \ge 0} \theta_1^\ell \bar{\varepsilon}_{\cdot, t-\ell}, \tag{A.8}$$

$$\widetilde{Y}_{it} = (\alpha_i - \bar{\alpha}_{\cdot}) \sum_{\ell \ge 0} \theta_1^\ell + \theta_2 \sum_{\ell \ge 0} \theta_1^\ell \widetilde{D}_{i,t-\ell} + \sum_{\ell \ge 0} \theta_1^\ell \widetilde{\varepsilon}_{i,t-\ell}. \tag{A.9}$$

Recall the definitions of V_{it} and $\mathbf{\Theta}_t^0$ as provided in Section 3. In this example, we have $X_{it} = (D_{it}, Y_{i,t-1})^\top$. Given the assumption $\mathsf{E}(\Delta D_{it} \Delta\varepsilon_{it}) = 0$, we do not need to project ΔD_{it} onto the instruments. Hence, $\mathbf{\Theta}_t^0$, which collects the coefficients in the reduced form, is simply a vector of dimension $1 \times m_t$.

Similar to the proof of Theorem 3.2, we can show that terms involving $\widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}_t^0$ are of smaller order. The cross-sectional demeaning would similarly affect the rate of these terms, akin to the dominant ones. Therefore, our focus will be on examining the orders of the terms involving $\mathbf{\Theta}_t^0$. Specifically, we will analyze the orders of $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{\varepsilon}_{it}$ and $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{X}_{it}$, which are pivotal in the rate analysis for proving asymptotic normality.

Regarding $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{\varepsilon}_{it}$, we observe that, provided $T/N \to 0$, we have:

$$\begin{aligned}\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{\varepsilon}_{it} &= \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \varepsilon_{it} - \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \bar{\varepsilon}_{it} \\ &= \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \varepsilon_{it} - N^{-1} \sum_{i=1}^N \sum_{j=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \varepsilon_{jt} = \mathcal{O}_{\mathrm{P}}(\sqrt{NT}).\end{aligned}$$

Note that the covariance of the two components is of order T, which is negligible compared to $\sqrt{NT}$ if $T/N \to 0$, as $N, T \to \infty$.

Concerning $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{X}_{it}$, it consists of two components related to $\Delta \widetilde{Y}_{i,t-1}$ and $\Delta \widetilde{D}_{it}$. We will now elaborate on the order of $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{Y}_{i,t-1}$. A similar analysis applies to $\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{D}_{it}$.

By (A.7) and (A.8), we obtain:

$$\begin{aligned}\Delta \widetilde{Y}_{i,t-1} &= Y_{i,t-1} - Y_{i,t-2} - \bar{Y}_{\cdot,t-1} - \bar{Y}_{\cdot,t-2} \\ &= \theta_2 \sum_{\ell \geq 0} \theta_1^\ell (D_{i,t-1-\ell} - D_{i,t-2-\ell}) + \sum_{\ell \geq 0} \theta_1^\ell (\varepsilon_{i,t-1-\ell} - \varepsilon_{i,t-2-\ell}) \\ &\quad - \theta_2 \sum_{\ell \geq 0} \theta_1^\ell (\bar{D}_{\cdot,t-1-\ell} - \bar{D}_{\cdot,t-2-\ell}) - \theta_2 \sum_{\ell \geq 0} \theta_1^\ell (\bar{\varepsilon}_{\cdot,t-1-\ell} - \bar{\varepsilon}_{\cdot,t-2-\ell}) \\ &= \theta_2 \sum_{\ell \geq 0} \theta_1^\ell (\widetilde{D}_{i,t-1-\ell} - \widetilde{D}_{i,t-2-\ell}) + \sum_{\ell \geq 0} \theta_1^\ell (\widetilde{\varepsilon}_{i,t-1-\ell} - \widetilde{\varepsilon}_{i,t-2-\ell}).\end{aligned}$$

By letting $S_{it}^{\widetilde{D}} \stackrel{\text{def}}{=} \sum_{\ell \geq 0} \theta_1^\ell \widetilde{D}_{i,t-\ell}$ and $S_{it}^{\widetilde{\varepsilon}} \stackrel{\text{def}}{=} \sum_{\ell \geq 0} \theta_1^\ell \widetilde{\varepsilon}_{i,t-\ell}$, it follows that

$$\sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} \Delta \widetilde{Y}_{i,t-1} = \theta_2 \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) + \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} (S_{i,t-1}^{\widetilde{\varepsilon}} - S_{i,t-2}^{\widetilde{\varepsilon}}).$$

We will examine the first term more closely. Let $\mathbb{Y}_t, \mathbb{D}_t$ be two sets of indices from $\{1, \ldots, m_t\}$. For $k = 1, \ldots, m_t$, if the k-th element in V_{it}, denoted as $V_{it,k}$, corresponds to a lag of Y_{it}, specifically $Y_{i,t-l(k)}$, where $l(k)$ represents the corresponding lag order, then k is included in $\mathbb{Y}_t$. Similarly, if $V_{it,k}$ corresponds to a lag of D_{it}, namely $D_{i,t-l(k)}$, then k is part of $\mathbb{D}_t$. For simplicity, we exclude the intercept term from V_{it}. Additionally, let $\mathbf{\Theta}_{t,k}^0$ denote the k-th element in the vector $\mathbf{\Theta}_t^0$. Using these notations, the first term can be expressed as:

$$
\begin{aligned}
& \theta_2 \sum_{i=1}^N \sum_{t=2}^T \mathbf{\Theta}_t^0 V_{it} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
= \;& \theta_2 \sum_{i=1}^N \sum_{t=2}^T \Big\{ \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 Y_{i,t-l(k)} + \sum_{k \in \mathbb{D}_t} \mathbf{\Theta}_{t,k}^0 D_{i,t-l(k)} \Big\} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
= \;& \theta_2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \Big(Y_{i,t-l(k)} - \frac{\alpha_i}{1-\theta_1} \Big) (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
& + \theta_2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \frac{\alpha_i}{1-\theta_1} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
& + \theta_2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{D}_t} \mathbf{\Theta}_{t,k}^0 D_{i,t-l(k)} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
=: \;& I + II + III.
\end{aligned}
$$

Next, we will analyze these three parts respectively. By inserting the representation given in (A.7), we obtain that

$$
I = \theta_2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \Big\{ \sum_{\ell \geq 0} \theta_1^\ell \gamma_{t-l(k)-\ell} + \theta_2 \sum_{\ell \geq 0} \theta_1^\ell D_{i,t-l(k)-\ell} + \sum_{\ell \geq 0} \theta_1^\ell \varepsilon_{i,t-l(k)-\ell} \Big\} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}).
$$

We will particularly examine the term involving $D_{i,t-l(k)-\ell}$, as the other terms follow a similar pattern. Specifically, we decompose it into two parts:

$$
\begin{aligned}
& \theta_2^2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \sum_{\ell \geq 0} \theta_1^\ell D_{i,t-l(k)-\ell} (S_{i,t-1}^{\widetilde{D}} - S_{i,t-2}^{\widetilde{D}}) \\
= \;& \theta_2^2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \sum_{\ell \geq 0} \theta_1^\ell D_{i,t-l(k)-\ell} \Big\{ \sum_{\ell \geq 0} \theta_1^\ell (D_{i,t-1-\ell} - D_{i,t-2-\ell}) \Big\} \\
& + \theta_2^2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 \sum_{\ell \geq 0} \theta_1^\ell D_{i,t-l(k)-\ell} \Big\{ \sum_{\ell \geq 0} \theta_1^\ell (\bar{D}_{\cdot,t-1-\ell} - \bar{D}_{\cdot,t-2-\ell}) \Big\} \\
= \;& \theta_2^2 \sum_{i=1}^N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 S_{i,t-l(k)}^D (S_{i,t-1}^D - S_{i,t-2}^D) + \theta_2^2 N \sum_{t=2}^T \sum_{k \in \mathbb{Y}_t} \mathbf{\Theta}_{t,k}^0 S_{\cdot,t-l(k)}^{\bar{D}} (S_{\cdot,t-1}^{\bar{D}} - S_{\cdot,t-2}^{\bar{D}}),
\end{aligned}
$$

where $S_{it}^D \stackrel{\text{def}}{=} \sum_{\ell\geq 0}\theta_1^\ell D_{i,t-\ell}$ and $S_{\cdot t}^{\bar D} \stackrel{\text{def}}{=} \sum_{\ell\geq 0}\theta_1^\ell \bar D_{\cdot,t-\ell}$. Given the assumptions we imposed on the process $\{D_{it}\}$, we can observe that:

$$\sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{Y}_t}\mathbf{\Theta}^0_{t,k}\,\mathsf{E}\{S^D_{i,t-l(k)}(S^D_{i,t-1}-S^D_{i,t-2})\} = \sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{Y}_t}\mathbf{\Theta}^0_{t,k}\sigma_D^2\theta_1^{l(k)-2}/(1-\theta_1)$$
$$N\sum_{t=2}^T\sum_{k\in\mathbb{Y}_t}\mathbf{\Theta}^0_{t,k}\,\mathsf{E}\{S^{\bar D}_{\cdot,t-l(k)}(S^{\bar D}_{\cdot,t-1}-S^{\bar D}_{\cdot,t-2})\} = \sum_{t=2}^T\sum_{k\in\mathbb{Y}_t}\mathbf{\Theta}^0_{t,k}\sigma_D^2\theta_1^{l(k)-2}/(1-\theta_1).$$

Consequently, we can conclude that $I = \mathcal{O}_{\mathrm{P}}(NT)$, if $T/N\to 0$.

Similarly, we express term III as follows:

$$\begin{aligned}
III &= \theta_2\sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}D_{i,t-l(k)}\Big\{\sum_{\ell\geq 0}\theta_1^\ell(D_{i,t-1-\ell}-D_{i,t-2-\ell})\Big\}\\
&\quad+\theta_2\sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}D_{i,t-l(k)}\Big\{\sum_{\ell\geq 0}\theta_1^\ell(\bar D_{\cdot,t-1-\ell}-\bar D_{\cdot,t-2-\ell})\Big\}\\
&=\theta_2\sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}D_{i,t-l(k)}(S^D_{i,t-1}-S^D_{i,t-2})+\theta_2 N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}\bar D_{\cdot,t-l(k)}(S^{\bar D}_{\cdot,t-1}-S^{\bar D}_{\cdot,t-2}).
\end{aligned}$$

Furthermore, we find that:

$$\sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}\,\mathsf{E}\{D_{i,t-l(k)}(S^D_{i,t-1}-S^D_{i,t-2})\} = \sum_{i=1}^N\sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}\sigma_D^2\theta_1^{l(k)-2}(\theta_1+1)$$
$$N\sum_{t=2}^T\sum_{k\in\mathbb{Y}_t}\mathbf{\Theta}^0_{t,k}\,\mathsf{E}\{\bar D_{\cdot,t-l(k)}(S^{\bar D}_{\cdot,t-1}-S^{\bar D}_{\cdot,t-2})\} = \sum_{t=2}^T\sum_{k\in\mathbb{D}_t}\mathbf{\Theta}^0_{t,k}\sigma_D^2\theta_1^{l(k)-2}(\theta_1+1),$$

which implies that $III = \mathcal{O}_{\mathrm{P}}(NT)$, given $T/N^2\to 0$. A similar argument applies to term II.

In summary, in this specific panel AR(1) model, we find that $\sum_{i=1}^N\sum_{t=2}^T\mathbf{\Theta}^0_t V_{it}\Delta\widetilde{\varepsilon}_{it} = \mathcal{O}_{\mathrm{P}}(\sqrt{NT})$ and $\sum_{i=1}^N\sum_{t=2}^T\mathbf{\Theta}^0_t V_{it}\Delta\widetilde{X}_{it} = \mathcal{O}_{\mathrm{P}}(NT)$, if $T/N\to 0$, as $N,T\to\infty$. Therefore, we conclude that when time effects are included, the cross-sectional demeaning of the variables would not significantly affect the inference theory provided in Section 3.2.

A.3.4. *Asymptotic Normality for the General Model Estimator.*

Proof of Theorem 3.4. Observe that the estimator $\widehat{\theta}_1$ obtained by (10), regardless of the demeaning transformation due to the presence of time effects, can be expressed as

$$\begin{aligned}\widehat{\theta}_1 - \theta_1^0 &= \left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta X_{1,it}^\top\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta X_{2,it}^\top \theta_2^0\right)\\ &\quad + \left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta X_{1,it}^\top\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=2}^T \widehat{\mathcal{W}}_t^\top \widehat{U}_{it}\Delta \varepsilon_{it}\right)\\ &=: L_1 + L_2.\end{aligned}$$

We will demonstrate that $|L_1|_\infty$ is negligible asymptotically, and that L_2 approaches $(NT)^{-1}\sum_{i=1}^N\sum_{t=2}^T \mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}$ closely enough. This will lead to the conclusion upon application of the central limit theorem.

Step 1: We first show that the Dantzig estimator $\widehat{\mathcal{W}}_t$ is close to the desired weighting matrix $\mathcal{W}_t$ with respect to various norms. According to the definitions of $\mathcal{W}_t$ and $\widehat{\mathcal{W}}_t$ provided in Section 2.2, we derive the following results:

$$\begin{aligned}|\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{\max} &= |M_t^{-1}|_\infty|\widehat{M}_t\widehat{\mathcal{W}}_t + (M_t - \widehat{M}_t)\widehat{\mathcal{W}}_t - M_t\mathcal{W}_t|_{\max}\\ &\le |M_t^{-1}|_\infty\{|\widehat{M}_t\widehat{\mathcal{W}}_t - \mathbf{I}_{d\times d_1}|_{\max} + |(M_t - \widehat{M}_t)\widehat{\mathcal{W}}_t|_{\max}\}\\ &\le |M_t^{-1}|_\infty(\ell_t + |M_t - \widehat{M}_t|_{\max}|\widehat{\mathcal{W}}_t|_{1,1})\\ &\le |M_t^{-1}|_\infty(\ell_t + |M_t - \widehat{M}_t|_{\max}|\mathcal{W}_t|_{1,1}).\end{aligned}$$

By Assumption 3.6, it follows that

$$\max_{2\le t\le T}|\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{\max} \lesssim_{\mathrm{P}} c_n \max_{2\le t\le T}(\ell_t + \rho_{N,t}c_n) \lesssim c_n^2\sqrt{v_n/N}. \tag{A.10}$$

The last inequality is implied by bounding $|M_t - \widehat{M}_t|_{\max}$ using the tail probability inequality in Lemma A.2 based on Assumptions 3.1-3.2.

To analyze $|\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{1,1}$, we consider a truncation argument with $\tau_n = \sqrt{v_n/N}$. Specifically, we have

$$\begin{aligned}|\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{1,1} &= |\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{1,1}\mathbf{1}(|\mathcal{W}_t|_{\max}\le\tau_n) + |\widehat{\mathcal{W}}_t - \mathcal{W}_t|_{1,1}\mathbf{1}(|\mathcal{W}_t|_{\max}>\tau_n)\\ &\le 2|\mathcal{W}_t|_{1,1}\mathbf{1}(|\mathcal{W}_t|_{\max}\le\tau_n) + \sum_{i=1}^d\sum_{j=1}^{d_1}|\widehat{\mathcal{W}}_{t,ij} - \mathcal{W}_{t,ij}|\mathbf{1}(|\mathcal{W}_t|_{\max}>\tau_n).\end{aligned}$$

Utilizing the bound in (A.10) and Assumption 3.6, it follows that

$$\begin{aligned}
\max_{2\le t\le T}|\widehat{\mathcal{W}}_t-\mathcal{W}_t|_{1,1} &\lesssim_{\mathrm{P}} 2\max_{2\le t\le T}|\mathcal{W}_{t,ij}|^r\mathbf{1}(|\mathcal{W}_t|_{\max}\le\tau_n)/|\mathcal{W}_{t,ij}|^{r-1}\\
&\quad+c_n^2\sqrt{v_n/N}\sum_{i=1}^{d}\sum_{j=1}^{d_1}|\mathcal{W}_{t,ij}/\tau_n|^r\mathbf{1}(|\mathcal{W}_t|_{\max}>\tau_n)\\
&\le 2\tau_n^{1-r}w_n+\tau_n^{-r}w_nc_n^2\sqrt{v_n/N}\\
&\lesssim c_n^2w_n(v_n/N)^{\frac{1-r}{2}},
\end{aligned}$$

where the parameter $0\le r<1$ ensures that Assumption 3.6(i) holds.

Step 2: Next, we analyze the rate of $|L_1|_\infty$. Note that the dimension of $X_{1,it}$, i.e., d_1, is fixed. The max norm, spectral norm (denoted by $|\cdot|_2$), and infinity norm of a $d_1\times d_1$ matrix are equivalent up to a constant factor that depends on d_1. Similarly, the ℓ_1-norm and ℓ_∞-norm of a $d_1\times 1$ vector exhibit the same relationship. Recall the constraints in the Dantzig estimator in (9), we find that

$$\left|N^{-1}\sum_{i=1}^{N}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top-\mathbf{I}_{d_1\times d_1}\right|_2\lesssim\ell_t,\quad\left|N^{-1}\sum_{i=1}^{N}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{2,it}^\top\right|_{\max}\le\ell_t.$$

Denote the smallest and largest singular values of a matrix by $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$, respectively. Applying Weyl's inequality for singular values, $\sigma_{\min}(C+D)\ge\sigma_{min}(C)-\sigma_{\max}(D)$, we can bound $|L_1|_\infty$ as follows:

$$\begin{aligned}
|L_1|_\infty &\le\left|\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top\right)^{-1}\right|_\infty\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{2,it}^\top\theta_2^0\right|_\infty\\
&\lesssim\left|\left(\frac{1}{NT}\sum_{t=2}^{T}\sum_{i=1}^{N}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top-\mathbf{I}_{d_1\times d_1}+\mathbf{I}_{d_1\times d_1}\right)^{-1}\right|_2\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{2,it}^\top\right|_{\max}|\theta_2^0|_1\\
&\le\left\{\sigma_{\min}(\mathbf{I}_{d_1\times d_1})-\frac{1}{T}\sum_{t=2}^{T}\sigma_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top-\mathbf{I}_{d_1\times d_1}\right)\right\}^{-1}\max_{2\le t\le T}\ell_t\vartheta_n\\
&\lesssim\max_{2\le t\le T}\ell_t\vartheta_n=o(1/\sqrt{NT})\quad\text{(by Assumption 3.6(iv))}.
\end{aligned}$$

Step 3: Lastly, we establish the asymptotic normality of the leading term L_2. To achieve this, we verify that L_2 is sufficiently close to $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}$. Based on the findings in Step 2, we observe that

$$\begin{aligned}
&\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}-L_2\right|_\infty\\
\leq\ &\left|\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top\right)^{-1}\right|_{\max}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta\varepsilon_{it}-\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it})\right|_1\\
&+\left|\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta X_{1,it}^\top\right)^{-1}-\mathbf{I}_{d_1\times d_1}\right|_{\max}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}\right|_1\\
\lesssim\ &\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathcal{W}}_t^\top\widehat{U}_{it}\Delta\varepsilon_{it}-\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it})\right|_\infty+o_{\mathrm{P}}(1/\sqrt{NT})\\
\leq\ &\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathcal{W}}_t-\mathcal{W}_t)^\top\widehat{U}_{it}\Delta\varepsilon_{it})\right|_\infty+\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top(\widehat{U}_{it}-U_{it})\Delta\varepsilon_{it}\right|_\infty+o_{\mathrm{P}}(1/\sqrt{NT}).
\end{aligned}$$

Using the results obtained in Step 1, we find that

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}(\widehat{\mathcal{W}}_t-\mathcal{W}_t)^\top\widehat{U}_{it}\Delta\varepsilon_{it})\right|_\infty\lesssim_{\mathrm{P}} c_n^2 w_n(v_n/N)^{\frac{1-r}{2}}/\sqrt{N}.$$

Furthermore, following similar steps as in bounding $|I_3^{-1}P_1|_\infty$ in the proof of Theorem 3.2, we obtain that

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top(\widehat{U}_{it}-U_{it})\Delta\varepsilon_{it}\right|_\infty\lesssim_{\mathrm{P}} c_n\max_{2\leq t\leq T}s_t^*\log m_t/N.$$

Combining these findings with Assumption 3.6(iii), we conclude that

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}-L_2\right|_\infty=o_{\mathrm{P}}(1/\sqrt{NT}).$$

The proof is then completed by applying a central limit theorem to $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathcal{W}_t^\top U_{it}\Delta\varepsilon_{it}$, following a similar approach as in the proof of Theorem 3.2.

It is noteworthy that with the implementation of a sample-splitting procedure, where $\widehat{\mathcal{W}}_t$ and $\widehat{U}_{it}$ are estimated from an auxiliary sub-sample, we could get sharper bounds as follows:

$$\left|\frac{1}{NT}\sum_{i\in\mathbb{I}_s}\sum_{t=2}^{T}(\widehat{\mathcal{W}}_t-\mathcal{W}_t)^\top\widehat{U}_{it}\Delta\varepsilon_{it})\right|_\infty\lesssim_{\mathrm{P}} c_n^2 w_n(v_n/N)^{\frac{1-r}{2}}/\sqrt{NT},$$

$$\left|\frac{1}{NT}\sum_{i\in\mathbb{I}_s}\sum_{t=2}^{T}\mathcal{W}_t^\top(\widehat{U}_{it}-U_{it})\Delta\varepsilon_{it}\right|_\infty\lesssim_{\mathrm{P}} c_n\max_{2\leq t\leq T}\sqrt{s_t^*}\log m_t/(N\sqrt{T}),$$

where $s\in\{A,B\}$. As a result, the required rate condition in Assumption 3.6(iii) can be improved to $c_n^2 w_n(v_n/N)^{\frac{1-r}{2}}+c_n\max_{2\leq t\leq T}\sqrt{s_t^*}\log m_t/\sqrt{N}=o(1)$. □

B. Supplementary Tables for Simulation Study

Table B.1. Results for $\theta_1 = 0.8$ with $N = 400$

	AB	AB-LASSO	AB-LASSO-SS ($K = 2$)	AB-LASSO-SS ($K = 5$)	DAB-SS
			$T = 30$		
RMSE	**0.01**	0.09	0.02	0.02	0.02
std. dev.	**0.01**	**0.01**	0.02	0.02	0.02
bias	-0.01	-0.09	**0.00**	**0.00**	0.01
CI length	0.05	0.05	0.09	0.07	0.05
coverage	**0.96**	0.00	0.98	**0.94**	0.81
			$T = 40$		
RMSE	**0.01**	0.09	**0.01**	**0.01**	0.07
std. dev.	**0.01**	**0.01**	**0.01**	**0.01**	0.03
bias	-0.01	-0.09	**0.00**	**0.00**	0.07
CI length	0.07	0.05	0.08	0.06	0.07
coverage	0.99	0.00	0.99	**0.95**	0.15
			$T = 50$		
RMSE	0.03	0.09	**0.01**	**0.01**	0.20
std. dev.	0.02	**0.01**	**0.01**	**0.01**	0.06
bias	-0.03	-0.09	**0.00**	**0.00**	0.19
CI length	0.11	0.04	0.07	0.05	0.11
coverage	**0.94**	0.00	0.98	**0.94**	0.00

Notes: The numbers in the table are divided by 0.8 for RMSE, standard deviation (std. dev.), bias, and CI length. Superior results are indicated in bold.

Table B.2. Results for $\theta_2 = 1$ with $N = 400$

	AB	AB-LASSO	AB-LASSO-SS ($K = 2$)	AB-LASSO-SS ($K = 5$)	DAB-SS
			$T = 30$		
RMSE	**0.02**	0.11	**0.02**	**0.02**	0.05
std. dev.	**0.01**	0.02	0.02	0.02	0.03
bias	-0.01	-0.10	**0.00**	**0.00**	0.04
CI length	0.08	0.07	0.15	0.09	0.14
coverage	1.00	0.00	0.99	**0.95**	0.47
			$T = 40$		
RMSE	0.04	0.11	**0.02**	**0.02**	0.09
std. dev.	**0.02**	**0.02**	**0.02**	**0.02**	0.05
bias	-0.03	-0.10	**0.00**	**0.00**	0.08
CI length	0.15	0.06	0.09	0.07	0.15
coverage	**0.96**	0.00	0.98	**0.94**	0.47
			$T = 50$		
RMSE	0.09	0.11	**0.02**	**0.02**	0.10
std. dev.	0.03	**0.01**	0.02	0.02	0.07
bias	-0.08	-0.11	**0.00**	**0.00**	0.07
CI length	0.24	0.05	0.08	0.06	0.24
coverage	0.86	0.00	0.98	**0.96**	0.76

Notes: Superior results are indicated in bold.

Table B.3. Results for $\theta_1 = 0.8$ with $N = 200$

	AB-LASSO	AB-LASSO-SS ($K = 5$)	AB-OLS	AB-OLS-SS ($K = 5$)
		$T = 30$		
RMSE	0.16	**0.04**	0.29	0.07
std. dev.	**0.02**	0.04	**0.02**	0.07
bias	-0.16	**0.00**	-0.29	0.01
CI length	0.09	0.14	0.07	0.25
coverage	0.00	**0.95**	0.00	0.94
		$T = 40$		
RMSE	0.16	**0.04**	0.32	0.07
std. dev.	0.02	0.04	**0.01**	0.07
bias	-0.16	**0.00**	-0.32	0.02
CI length	0.07	0.13	0.06	0.25
coverage	0.00	0.93	0.00	**0.96**
		$T = 50$		
RMSE	0.16	**0.03**	0.35	0.07
std. dev.	0.02	0.03	**0.01**	0.07
bias	-0.16	**0.00**	-0.35	0.01
CI length	0.05	0.11	0.05	0.26
coverage	0.00	0.93	0.00	**0.96**

Notes: The numbers in the table are divided by 0.8 for RMSE, standard deviation (std. dev.), bias, and CI length. Superior results are indicated in bold.

Table B.4. Results for $\theta_2 = 1$ with $N = 200$

	AB-LASSO	AB-LASSO-SS ($K = 5$)	AB-OLS	AB-OLS-SS ($K = 5$)
		$T = 30$		
RMSE	0.18	**0.05**	0.35	0.11
std. dev.	**0.03**	0.05	**0.03**	0.10
bias	-0.18	**0.01**	-0.35	0.02
CI length	0.11	0.18	0.10	0.38
coverage	0.00	**0.95**	0.00	**0.95**
		$T = 40$		
RMSE	0.18	**0.04**	0.37	0.11
std. dev.	0.03	0.04	**0.02**	0.10
bias	-0.18	**0.00**	-0.37	0.03
CI length	0.10	0.15	0.08	0.39
coverage	0.00	**0.94**	0.00	**0.96**
		$T = 50$		
RMSE	0.19	**0.04**	0.38	0.11
std. dev.	**0.02**	0.04	**0.02**	0.11
bias	-0.18	**0.00**	-0.38	0.02
CI length	0.09	0.17	0.07	0.42
coverage	0.00	**0.95**	0.00	0.96

Notes: Superior results are indicated in bold.