Inference in High-Dimensional Panel Models:
Two-Way Dependence and Unobserved Heterogeneity

Kaicheng Chen

Department of Economics, Michigan State University. Email: chenka19@msu.edu.

Mar 19, 2025

Link to the latest manuscript

Abstract

Panel data allows for the modeling of unobserved heterogeneity, which significantly increases the number of nuisance parameters, making high dimensionality a practical issue rather than just a theoretical concern. However, unobserved heterogeneity, along with potential temporal and cross-sectional dependence in panel data, further complicates estimation and inference for high-dimensional models. This paper proposes a toolkit for robust estimation and inference in high-dimensional panel models with large cross-sectional and time sample sizes. To reduce the dimensionality, I propose a weighted LASSO using two-way cluster-robust penalty weights. Due to the cluster dependence, the rate of convergence is slow even in an oracle case. Nevertheless, by leveraging a clustered-panel cross-fitting approach for bias correction, asymptotic normality can be established for the low-dimensional vector of the estimated parameters. As a special case, inferential theories are also established using the full sample in a partial linear model with unobserved time and unit effects. In a panel estimation of the government spending multiplier, I demonstrate how high dimensionality can be hidden and how the proposed toolkit enables flexible modeling and robust inference.

Keywords: high-dimensional regression, two-way cluster dependence, correlated time effects, unobservable heterogeneity, LASSO, Post-LASSO, double/debiased machine learning, cross-fitting.
JEL Classification: C01, C14, C23, C33

1. Introduction

In economic research, high dimensionality typically refers to the large number of unknown parameters relative to the sample size, under which traditional estimations are either infeasible or tend to yield estimates too noisy to be informative. The issue of high dimensionality becomes more relevant as data availability grows and economic modeling involves more flexibility. Commonly, the problem of high dimensionality appears in at least the following three scenarios:

- The dimension of observable and potentially relevant variables can be large relative to the sample. In trade literature, preferential trade agreements (PTAs) usually involve a large number of provisions even though most policy analysis only focuses on the effect of a small subset of the provisions [1]. In demand analysis, even if the focus is on the own-price elasticity, the prices of relevant goods should also be included, unless strong assumptions for aggregation are assumed (see Chernozhukov et al., 2019).

- With nonparametric or semiparametric modeling, the unknown functions are viewed as infinite-dimensional parameters regardless of the dimension of observable variables. If the unknown function $g(X)$ is approximately sparse and can be well-approximated by a linear combination of the 3rd-order polynomial transformation of X, then it would involve 285 transformed regressors when the dimension of X is 10 and 1770 when we start with a dimension of 20. [2]

- The modeling of heterogeneity can raise the number of nuisance parameters drastically. In demand analysis, income effects are specific to products if the homothetic preference assumption fails. For difference-in-difference analysis, allowing unit-specific trends and heterogeneous trends across the covariates can relax/test the parallel trend assumption. For models with unobserved heterogeneity that appears in a nonlinear way, either treating them as parameters to be estimated (fixed effects) or modeling them in a flexible way (correlated random effects) contributes to high dimensionality. [3].

Particularly, the modeling of heterogeneity in panel models makes high dimensionality more of a practical issue rather than just a theoretical concern. As a concrete example, let's consider a panel model where all three sources of high dimensionality are involved:

$$Y_{it} = D_{it}\theta_0 + g_0(X_{it}, c_i, d_t) + U_{it}, \tag{1.1}$$

where D_{it} is a vector of low-dimensional treatment or policy variables. X_{it} is a vector of potentially high-dimensional control variables. D_{it} can also contain some higher-order effects and interactive effects with a subset of the controls to allow for nonlinear and heterogeneous effects in a parametric way. When D_{it} is not conditionally uncounfounded, instrumental variables may be used for identification of θ_0. ; $g(.)$ is an unknown function, e.g. an infinite dimensional parameter; c_i and d_t are unobserved heterogeneous effects, either as fixed-effect parameters or correlated random variables. The interest lies in the inference on the low-dimensional parameters θ_0.

Without considering the features of panel data and the unobserved heterogeneity, it is a classic partial linear model that has been well-studied in previous semiparametric literature. To reduce the dimensionality, sparse approximation and regularization approaches have been widely employed. Essentially, regularization,

[1] Based on data from Mattoo et al. (2020), 282 PTAs were signed and notified to the WTO between 1958 and 2017, encompassing 937 provisions across 17 policy areas. See Breinlich et al. (2022).

[2] For a vector X with dimension k, it is easy to show that the 2nd-order polynomial transformation generates $\frac{k^2}{2} + \frac{3}{2}k$ terms and the 3rd-order polynomial transformation generates $k + \frac{1}{2}k(k+1) + \frac{1}{2}\sum_{l=1}^{k} l(l+1) = \frac{1}{6}k^3 + k^2 + \frac{11}{6}k$ terms.

[3] This is particularly relevant in trade literature where the unobserved heterogeneity derived from the gravity model takes a pairwise form among the importers, exporters, and the time. As each of these three dimensions expands, the number of nuisance parameters explodes quickly. See Correia et al. (2020), Chiang et al. (2021), and Chiang et al. (2023b), for example.

also known as the machine learning approach, trades off bias for smaller variance to achieve desirable rates of convergence. However, due to the bias introduced by regularization and overfitting, inference can be challenging. Typically, bias correction is involved to obtain estimators with better statistical properties and to conduct valid inferences.

In the case of panel data, it is soon realized that at least three challenges would appear if researchers attempt to apply the existing high-dimensional approaches directly. First of all, the statistical properties of many regularized estimators remain unknown with panel data where the observations are potentially dependent across space/unit and time. Secondly, some bias-correction procedures for inference such as sample-splitting/cross-fitting are very particular about the sampling assumption and existing approaches are not valid under two-way dependence in panels. Thirdly, panel data is often leveraged to model unobserved individual and time effects, which may lead to another source of high dimensionality and further complicate estimation and inference.

To reduce the dimensionality, as the first challenge, I proposed a variant of LASSO that uses regressor-specific penalty weights robust to two-way cluster dependence and weak temporal dependence across clusters. Such a LASSO approach is labeled as the two-way cluster-LASSO, corresponding to the heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016). This approach theoretically derives the common penalty level λ up to a constant and a small-order sequence that do not vary across different data-generating processes. Therefore, data-driven tuning, such as cross-validation, is not needed, which makes it more computationally efficient and avoids non-trivial theories that take data-driven tuning into account.

A common and important condition for obtaining the desirable statistical properties of LASSO selection/estimation is the so-called "regularization event", which states that the overall penalty level is sufficiently large to dominate the "noise" in the high-dimensional estimation (but not too large at the same time to avoid under-selection and slow rate of convergence). However, existing approaches for ensuring such an event with probability approaching one are not applicable in this case due to the two-way cluster dependence. Instead, by considering the component structure characterization of the two-way dependence and decomposing the error terms using Hajek projections, I am able to leverage the moderate deviation theorems by Peña et al., 2009 and Gao et al., 2022 and the concentration inequality by Fuk and Nagaev (1971) for bounding the tail probability of the "noise" term. Combining with existing non-asymptotic bounds for the LASSO approach in Belloni et al. (2012), I derive the rate of convergence for the (post) two-way cluster-LASSO.

According to the rate of convergence results, the proposed (post) LASSO estimator is consistent for the slope coefficients of the sparse model. However, it is also revealed that the convergence rate is not as fast as the common rates for LASSO estimation due to the two-way cluster dependence. The problem lies in the underlying component structure. To illustrate, consider the simplest multivariate mean model through a component structure representation:

$$Y_{it} = \theta_0 + f(\alpha_i, \gamma_t, \epsilon_{it}) \tag{1.2}$$

where Y_{it} is a high-dimensional vector with dimension $s = o(NT)$ and $\theta_0 = \mathrm{E}[Y_{it}]$; α_i, γ_t, and ε_{it} are unobserved random elements. This is a common characterization of cluster dependence in the literature on cluster-robust inference. We notice that α_i introduces cluster/temporal dependence within group i and γ_t introduces cluster/cross-sectional dependence within group t. To estimate the high-dimensional vector θ_0, we consider the sample mean estimator $\widehat{\theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$. We can rewrite the estimator through a Hajek projection:

$$\widehat{\theta} - \theta_0 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(a_i + g_t + e_{it}) = \frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{T}\sum_{t=1}^{T} g_t + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}, \tag{1.3}$$

where $a_i := \mathrm{E}[Y_{it} - \theta_0 | \alpha_i], g_t := \mathrm{E}[Y_{it} - \gamma_t]$, and $e_{it} := Y_{it} - \theta_0 - a_i - g_t$. For illustration purposes, suppose those components are i.i.d sequences and independent of each other. Then it can be shown that, under some regularity conditions, for each $j = 1, ..., s$, $\widehat{\theta}_j - \theta_0 = O_P\left(\frac{1}{\sqrt{N \wedge T}}\right)$ and $\|\widehat{\theta} - \theta_0\|_2 = \left(\sum_{j=1}^{s}\left(\widehat{\theta}_j - \theta_{0j}\right)^2\right)^{1/2} = O_P\left(\sqrt{\frac{s}{N \wedge T}}\right)$. While $\widehat{\theta}$ is still consistent when N, T diverge at the same rate, $\|\widehat{\theta} - \theta_0\|_2$ converges slower than $o_P\left(\sqrt{\frac{1}{N \wedge T}}\right) = o_P\left((NT)^{-1/4}\right)$, which is a common rate requirement for inferential theory.

This is where the second challenge arises: if a faster rate of convergence is not achievable due to the two-way cluster dependence, some bias-correction approaches are needed to relax the rate requirement for valid inference. One common approach in semiparametric literature is to add a correction term to the original identifying moment function. It results in an orthogonalized moment condition which often features multiplicative error terms, closely related to the doubly robust estimators. Although the orthogonality property allows the nuisance estimations to be noisier, it generally does not ensure valid inference when there is growing dimensionality in the unknown parameters. An extra bias-correction approach, sample splitting or its generalization cross-fitting, has been proposed for inference in the high-dimensional regression models. The idea of sample splitting in a two-step estimation is to split the sample in a proper way and use the sub-samples separately for each step. If the sub-samples are independent of each other, then the first-step estimates will be independent of the sample used for the second-step estimation. With this property, the error term that causes the bias can vanish with a less stringent rate requirement on the first step. Intuitively, the dependence between the two steps is eliminated so that a potentially over-fitted nuisance estimate from the first step does not pollute the second-step estimator as much as it would otherwise do. However, sample-splitting as well as cross-fitting is very sensitive to the sampling assumption. Building upon recent development of cross-fitting approaches for dependent data (Chiang et al. (2022); Semenova et al. (2023a)), I propose a clustered-panel cross-fitting scheme and I show the constructed main and auxiliary samples are "approximately" and independent of each other. Effectively, this inferential procedure extends the double/debiased machine learning (DML, hereafter) approach by Chernozhukov et al. (2018a) to panel data models, and it is labeled as the panel DML. Asymptotic normality for the panel DML estimator is established given high-level assumptions on the convergence rates regarding the first-step estimator. It is shown that the crude requirement on the rate of convergence can be relaxed to $o\left((N \wedge T)^{-1/4}\right)$ in L^2 norm, which admits the first-step estimation through the two-way cluster LASSO.

For the third challenge caused by the unobserved heterogeneity, existing literature proposes to use the fixed-effect approach assuming the unknown function g_0 in 1.1 is linear in (c_i, d_t) (Belloni et al., 2016; Kock and Tang, 2019), or to use the high-dimensional common correlated effect approaches assuming g_0 is linear in the interactive fixed effects (Vogt et al., 2022). To allow for flexible function forms while remaining tractable, I propose to model (c_i, d_t) as correlated random effects through a generalized Mundlak device while assuming the unknown function to be approximately sparse. In that way, a very rich form of heterogeneity is permitted. While not all of those nonlinear and heterogeneous effects are relevant and the identities of the truly relevant effects are unknown to the researcher, suitable machine learning approaches, e.g. the two-way cluster-LASSO, can be used to select the relevant effects. However, there is one more subtle issue: common approaches including Mundlak device that deal with the unobserved heterogeneous effects introduce cross-sectional and temporal sample averages which in turn bring dependence across cross-fitting sub-samples. Furthermore, even if it remains valid under extra conditions, cross-fitting often causes a loss of efficiency due to the exclusion of observations. On the other hand, without cross-fitting, valid inference remains challenging for high-dimensional panel models in general. Nevertheless, in the case of the partial linear panel model, I further show that inferential theory can be established using the full sample.

In the empirical application, I re-examine the effect of government spending on the output of an open economy following the framework of Nakamura and Steinsson (2014), a well-cited empirical-macro paper. While they study it using a panel data approach considering unobserved heterogeneous effects that raise the nuisance parameters as the sample size grows, it is not considered a high-dimensional problem in the baseline setting: a linear panel model with only a few covariates and additive unobserved heterogeneous effects; the identification is through the instrumental variable. However, even in a conventionally low-dimensional setting, high dimensionality can be hidden because the true model can be highly nonlinear in the covariates and the unobserved heterogeneous effects can enter the model in a flexible way. To avoid the endogeneity caused by potential misspecification in the function form, I consider extending the baseline model in a more flexible way as in 1.1. Due to potential two-way cluster dependence, existing high-dimensional methods designed for independent or weakly dependent data may not be valid. This is where the proposed dependence-robust estimation and inference for high-dimensional models can be leveraged and the results can be used for a robustness check. It is shown that the estimates are consistent with the baseline results, which indicates the nonlinear and interactive effects may not be very relevant in this model. Existing estimation and inference methods that are not robust to either high-dimensionality or two-way cluster dependence tend to over-fit and result in noisy estimates and inaccurate inference results.

The rest of the paper is outlined as follows: The next sub-section reviews relevant literature and summarizes the differences and contributions of this paper relative to the existing ones. Section 2 presents the two-way cluster-LASSO estimator and the investigation of its statistical properties under two-way cluster dependence. Section 3 introduces a sub-sampling scheme designed for cross-fitting that allows within-cluster dependence and weak dependence across clusters. It is then used as a bias-correction approach for valid inference on the low-dimensional parameter considering the effect of the high-dimensional nuisance estimation. In Section 4, the partial linear model with unobserved heterogeneity is studied in detail as a leading example. Simulation evidence is given in Section 5 where the proposed approaches compete with existing ones. In Section 6, the empirical estimation of the government spending multiplier is used as the illustration of hidden high dimensionality and the application of the proposed toolkit. Section 7 concludes the paper with a discussion of limitations and detailed empirical recommendations.

1.1. ***Relation to the Literature***

This paper builds upon literature on $l1$ regularization methods in high-dimensional regression. Bickel et al. (2009) derive the convergence rate of the prediction error in terms of the empirical norm under homogeneous Gaussian error, restricted eigenvalue, and sparsity conditions. Bühlmann and Van De Geer (2011) instead assumes a sub-Gaussian tail property to derive similar results of convergence rates. See Section 29.11 of Hansen (2022) for an illustration and extension of Bickel et al. (2009)'s analysis under heteroskedasticity. Under Gaussian or sub-Gaussian errors, Basu and Michailidis (2015); Kock and Callot (2015); Lin and Michailidis (2017) study LASSO-based approaches for dependent data. To allow for both non-Gaussian errors and dependent data, Wu and Wu (2016), Chernozhukov et al. (2021a), Babii et al. (2022, 2023) Gao et al. (2024) derive Nagaev-type concentration inequalities to bound the tail probability assuming a proper order of the penalty level. However, all aforementioned LASSO-based approaches require delicate tuning of the penalty level to ensure a desirable finite sample performance. The common cross-validation approaches and bootstrap in Chernozhukov et al. (2021a) for choosing the penalty level are computationally costly and are very sensitive to the sampling assumption. Plus, the statistical analysis accounting for the data-driven penalty level is highly non-trivial (see Chetverikov et al., 2021 for validity on cross-validation LASSO under random sampling). As another strand, Belloni et al. (2011, 2012, 2016) propose other variants of LASSO approaches and leverage (self-normalized) moderate deviation theorems to derive theoretically-driven penalty levels. However, their methodologies cannot be easily extended to settings with two-way dependence. The proposed variant of LASSO is built upon the aforementioned literature and employs both Nagaev-type inequalities (Fuk and Nagaev, 1971) and moderate deviation theorem for self-normalized sums (Peña et al., 2009; Gao et al., 2022). To my knowledge, it is the first LASSO and high-dimensional estimator that is robust to the two-way cluster dependence.

The inferential theory in high-dimensional regression models typically relies on some bias-correction methods and they are particularly important here due to the two-way cluster dependence that results in a slow rate of convergence. Bias-correction approaches for inference purposes take various forms in the literature: for example, the low-dimensional projection adjustment in Zhang and Zhang (2014), the de-sparsification procedure in Van de Geer et al. (2014), the decorrelating matrix adjustment in Javanmard and Montanari (2014), the double selection approach in Belloni et al. (2014), the decorrelated score construction in Ning and Liu (2017), the Neyman orthogonal moment construction in Chernozhukov et al. (2018a, 2022a). The last strand of the literature is often labeled as DML, which is closely related to previous semiparametric literature including Ichimura (1987), Robinson (1988), Powell et al. (1989), Newey (1994), and Andrews (1994). The idea of the orthogonalization is to add a correction term to the original identifying moment function so that the second-step estimator is less sensitive to the plug-in of noisy first steps. Due to the resulting multiplicative error term in the orthogonal moment condition, it is closely related to the doubly-robust methods. Newey (1994) provides a general construction of the orthogonal moment condition through the influence functions. It is further facilitated by Ichimura and Newey (2022) for identifying moment conditions satisfying certain restrictions. See Chernozhukov et al. (2018a) and Chernozhukov et al. (2022a) for a summary of such constructions and known orthogonal moment functions. More recently, Chernozhukov et al. (2018b, 2021b, 2022b,c); Jordan et al. (2023) provide an alternative approach by estimating the correction term without knowing its analytical form. For the inferential theory in high-dimensional panel models, this paper takes the orthogonalization step as given and focuses on nuisance estimation and cross-fitting.

Sample-splitting or cross-fitting, serving as another bias-correction approach, has been widely employed in other two-step estimations. The role of cross-fitting in high-dimensional inferential theory is to remove the dependence between the nuisance estimation and the second-step estimation so that the over-fitting bias from the first step has less impact on the second step. Technically, it allows for a slower rate of convergence in the first step and it in turn relaxes the sparsity condition (e.g., Belloni et al., 2014). Chernozhukov et al. (2018a) generalize the sample-splitting procedure as a cross-fitting scheme which further improves finite sample performance by reducing the noise due to arbitrary splitting of the sample. Chiang et al. (2021, 2022) propose a cross-fitting scheme robust to separately and jointly exchangeable arrays. Semenova et al. (2023a) propose a leave-one-neighborhood-out cross-fitting and introduce a coupling approach (due to Strassen, 1965 and Berbee, 1987) to prove the validity of cross-fitting under temporal dependence. The idea of leave-one-neighborhood-out sub-sampling scheme is also shared by $h-$block cross-validation (Burman et al., 1994; Racine, 2000) and big-block-small-block technique in time series literature (e.g., Gao et al., 2022). Built upon previous literature, I propose a more robust cross-fitting scheme that is valid under not only cluster dependence but also weak temporal dependence across clusters.

This paper also belongs to the cluster-robust inference literature. The characterization of the two-way cluster dependence is based on the Aldous-Huber-Kallenberg (AHK) type representation, which is common is this literature (e.g., Djogbenou et al., 2019, Roodman et al., 2019, Davezies et al., 2019, and Menzel, 2021). This original representation only works for exchangeable arrays, which is violated in panel data settings with autocorrelation over time. Chiang et al. (2024) generalizes this representation by allowing the time factor to be correlated over time and Chen and Vogelsang (2024) also considers this representation when deriving fixed-b asymptotic results for inference. Differing from the original representation theorem, it is a fairly general characterization of two-way dependence in the panel. Such characterization of the dependence structure is common in economic studies (e.g., Rajan and Zingales, 1998, Fama and French, 2000, Li et al., 2004, Larrain, 2006, Thompson, 2011, Nakamura and Steinsson, 2014, Guvenen et al., 2017, Ellison et al., 2024, and Nakamura and Steinsson, 2014 among many others).

1.2. Notation.

Here is a collection of frequently used notations in this paper. Some extra notations are defined along with the context. E and P are as generic expectation and probability operators. $\mathcal{P}_{NT}$ is an expanding collection of all data-generating processes P that satisfy certain conditions. P_{NT} is a sequence of probability laws such that $P_{NT} \in \mathcal{P}_{NT}$ for each (N, T). The dependence on (N, T) and P_{NT} will be suppressed whenever clear in the context. $\|.\|$ is the Euclidean (Frobenius) norm for a matrix. Let $\mathbf{x}$ be a generic $k \times 1$ real vector, then the l^q norm is denoted as $\|\mathbf{x}\|_q := \left(\sum_{j=1}^k x_j^q\right)^{1/q}$ for $1 \leq q < \infty$; $\|\mathbf{x}\|_\infty := \max_{1\leq j\leq k} |x_j|$. The $L^q(P)$ norm is denoted as $\|f\|_{P,q} := \left(\int \|f(\omega)\|^q dP(\omega)\right)^{1/q}$ where f is a random element with probability law P. I denote the empirical average of f_{it} over $i = 1, ..., N$ and $t = 1, ..., T$ as $\mathbb{E}_{NT}[f_{it}] = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T f_{it}$ and the empirical L^2 norm as $\|f_{it}\|_{NT,2} = \left(\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \|f_{it}\|^2\right)^{1/2}$. Correspondingly, I denote the empirical average of f_{it} over the sub-sample $i \in I_k$ and $t \in S_l$ as $\mathbb{E}_{kl}[f_{it}] = \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l} f_{it}$ and the empirical L^2 norm over the subsample as $\|f_{it}\|_{kl,2} = \left(\frac{1}{N_k T_l}\sum_{i\in N_l}\sum_{t\in T_l}\|f_{it}\|^2\right)^{1/2}$, where I_k, S_l are sub-sample index sets and N_k, T_l are sub-sample sizes that will be introduced next section.

2. Two-Way Cluster LASSO

In the existing literature, not much is known in terms of statistical properties for high-dimensional methods under cluster dependence in both cross-section and time. In this section, a variant of the $l1$-regularization methods, also known as the LASSO, will be proposed and examined.

To focus on the LASSO approach under two-way dependence, I consider a simple conditional expectation model of a scalar outcome given a potentially high-dimensional vector of covariates. Let (Y_{it}, X_{it}) be a sample with $i = 1, ..., N$ and $t = 1, ..., T$. The conditional expectation model can be expressed as follows:

$$Y_{it} = f(X_{it}) + V_{it},\ \mathrm{E}[V_{it}|X_{it}] = 0 \tag{2.1}$$

where $f(X_{it}) := E[Y_{it}|X_{it}]$ is an unknown conditional expectation function of potentially high-dimensional covariates X_{it}; V_{it} is the associated stochastic error.

To characterize the two-way cluster dependence in the panel, I assume the random elements $W_{it} := (Y_{it}, X_{it}, V_{it})$ are generated by the following process:

Assumption AHK (Aldous-Hoover-Kallenberg Component Structure Characterization)**.**

$$W_{it} = \mu + f\left(\alpha_i, \gamma_t, \varepsilon_{it}\right),\ \ \forall i \geq 1, t \geq 1, \tag{2.2}$$

where $\mu = \mathrm{E}_P[W_{it}]$, f is some unknown measurable function; $(\alpha_i)_{i\geq 1}$, $(\gamma_t)_{t\geq 1}$, and $(\varepsilon_{it})_{i\geq 1, t\geq 1}$ are mutually independent sequences, α_i is i.i.d across i, ε_{it} is i.i.d across i and t, and γ_t is strictly stationary.

Assumption AHK is motivated by a representation theorem for an exchangeable array, named after Aldous-Hoover-Kallenberg (AHK, hereafter), which states that if an array of random variables $(X_{ij})_{i\geq 1, j\geq 1}$ is separately or jointly exchangeable[4], then $X_{ij} = f(\xi_i, \zeta_j, \iota_{ij})$ where $(\xi_i)_{i\geq 1}, (\zeta_j)_{j\geq 1}, (\iota_{ij})_{i\geq 1, j\geq 1}$ are mutually independent, uniformly distributed i.i.d. random variables[5]. However, the exchangeability is not likely to hold for arrays with the presence of a temporal dimension since it is naturally ordered. In macroeconomics, for instance, we can interpret the time components $(\gamma_t)_{t\geq 1}$ as unobserved common time shocks, which are naturally correlated over time, implying the exchangeability violated. Therefore, by allowing γ_t to be correlated, it introduces temporal dependence across all clusters, making the characterization more sensible. The relaxation of the independence condition on $(\gamma_t)_{t\geq 1}$ can be viewed as a generalization of the component structure representation, as argued by Chiang et al. (2024). It is clear that under Assumption AHK, W_{it} and W_{is} are correlated for any i, t, s due to sharing the same cross-sectional cluster. Similarly, due to sharing the same temporal cluster, W_{it} and W_{jt} are dependent for any t, i, j. Furthermore, even if sharing neither the cross-sectional or temporal dimensions, observations can still be dependent due to correlated time effects γ_t. It is important to notice that the components in 2.2 simply characterize the dependence in panel data. Differing from factor models or models with unobserved heterogeneity, they do not affect the identification of the regression model in any way.

Throughout the paper, time effects γ_t are weakly dependent with some regularity condition introduced as follows. Before that, a few more concepts and notations are needed. Let (X, Y) be random elements taking values in Euclidean space $S = (S_1 \times S_2)$ with probability laws P_X and P_Y, respectively. Let $\|\nu\|_{TV}$ denote the total variation norm of a signed measure ν on a measurable space (S, Σ) where Σ is a σ-algebra on S:

$$\|\nu\|_{TV} = \sup_{A\in\Sigma} \nu(A) - \nu(A^c).$$

Define the dependence coefficient of X and Y as:

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

The next assumption regulates the dependence of γ_t using the beta-mixing coefficient:

Assumption AR (Absolute Regularity)**.** *The sequence $\{\gamma_t\}_{t\geq 1}$ is beta-mixing at a geometric rate:*

$$\beta_\gamma(m) = \sup_{s\leq T} \beta\left(\{\gamma_t\}_{t\leq s}, \{\gamma_t\}_{t\geq s+m}\right) \leq c_\kappa exp(-\kappa m), \forall m \in \mathbb{Z}^+, \tag{2.3}$$

for some constants $\kappa > 0$ and $c_\kappa \geq 0$.

---

[4]An array $(X_{ij})_{i\geq 1, j\geq 1}$ is separately exchangeable if $\left(X_{\pi(i),\pi'(j)}\right) \overset{d}{=} \left(X_{ij}\right)$, and jointly exchangeable if the same condition holds with $\pi = \pi'$.

[5]This is first proved in Aldous (1981) and independently proved and generalized to higher dimensional arrays in Hoover (1979). It is then further studied in Kallenberg (1989). For a formal statement of the theorem, see, for example, Theorem 7.22 in Kallenberg (2005).

Condition AR, also known as the beta-mixing condition, restricts the temporal dependence of the common time effects to decay at an exponential rate that is common in literature (for example, see Hahn and Kuersteiner (2011); Fernández-Val and Lee (2013), and can be generated by common autoregressive models as in Baraud et al. (2001).

Due to the potential high dimensionality in X, traditional nonparametric methods are not feasible for estimating the unknown function f. To reduce the dimensionality, a common approach is to consider the sparsity in the model and reduce the dimension through regularization. However, the unknown function f is an infinite dimensional parameter, which is not exactly sparse. Therefore, I take a sparse approximation approach following Belloni et al. (2012):

Assumption ASM (Approximate Sparse Model)**.** *The unknown function f can be well-approximated by a dictionary of transformations $f_{it} = F(X_{it})$ where f_{it} is a $p \times 1$ vector and F is a measurable map, such that*

$$f(X_{it}) = f_{it}\zeta_0 + r_{it}$$

where the coefficients ζ_0 and the approximation error r_{it} satisfy

$$\|\zeta_0\|_0 \leq s = o(N \wedge T),\ \|r_{it}\|_{NT,2} = O_P\left(\sqrt{\frac{s}{N \wedge T}}\right).$$

Assumption ASM views the high-dimensional linear regression as an approximation. It requires a subset of the parameters ζ_0 to be zero while controlling the size of the approximation error. Compared to the sparsity condition in previous literature, here it imposes a slower rate of growth restriction on the non-zero slope coefficients. For example, $s = o(NT)$ corresponds to the case of heteroskedasticity-robust LASSO under i.i.d data in Belloni et al. (2012); $s = (Nl_T)$ corresponds to the cluster-robust LASSO under temporal dependence panel data in Belloni et al. (2016) where $l_T \in [1, T]$ is an information index that equals T when there is no temporal dependence and equals 1 when there is cross-sectional independence and perfect temporal dependence. In other words, the underlying component structure restricts the growth of nonzero slope coefficients of the model in a similar way to the perfect temporal dependence case.

Under Assumption ASM, we can rewrite the model 2.1 as

$$Y_{it} = f_{it}\zeta_0 + r_{it} + V_{it},\ \mathrm{E}[V_{it}|X_{it}] = 0. \tag{2.4}$$

Using 2.4, we can apply $l1$ regularization in the least squared error problem. Let λ be some non-negative common penalty level and ω be some non-negative $p \times p$ diagonal matrix of regressor-specific penalty weights. Consider the following generic weighted LASSO estimator:

$$\hat{\zeta} = \arg\min_{\zeta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - f_{it}\zeta)^2 + \frac{\lambda}{NT} \|\omega^{1/2}\zeta\|_1. \tag{2.5}$$

To obtain the desirable property of LASSO estimation, one needs to choose λ and ω in an optimal way so that the penalty level is large enough to avoid noisy estimation due to overfitting but also the smallest possible since the size of the penalty determines the performance bound of LASSO estimation and too large a penalty level can cause missing variable bias. In other words, the overall penalty level given by both λ and ω decides the trade-off between overfitting variance and regularization bias. For example, let $\ddot{f}_{it}$ be the demeaned f_{it} using the sample mean[6] and one common choice of ω is the empirical Gram matrix $\mathrm{E}[\ddot{f}_{it}'\ddot{f}_{it}]$ that is used to standardize the regressors and so the model selection is not affected by the scale of the regressors. The common penalty level λ is often chosen by some cross-validation algorithms. If the chosen λ satisfies a certain asymptotic order, then the key condition that regularizes the tail behavior of the error term can be established under the conditional Gaussian error or sub-Gaussian error conditions (see Bickel et al., 2009, Bühlmann and Van De Geer, 2011, and Theorem 29.3 of Hansen, 2022):

$$\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right|\leq\frac{\lambda}{2c_1NT}. \tag{2.6}$$

Condition 2.6 is referred to as the "regularization event" in the literature. Combining with some non-asymptotic bounds for LASSO, the rate of convergence for $\widehat{\zeta}$ can be derived. This approach, however, is not applicable when the error terms are considered to exhibit heavy tails. Alternatively, Fuk-Nagaev types of concentration inequality are established to verify Condition 2.6 without relying on the Gaussian or sub-Gaussian assumption (e.g. Babii et al. (2024, 2023); Gao et al. (2024)). These alternative approaches, again, rely on cross-validation for choosing penalty levels, which is computationally costly and sensitive to the tuning of the cross-validation. It is further complicated when the cross-validation needs to be adjusted for dependent data.

Belloni et al. (2012) propose to self-normalize the regressors through regressor-specific penalty weights and leverage moderate deviation theorems (see Jing et al., 2003 and Peña et al., 2009) for the self-normalized sums to verify Condition 2.6. This common penalty level λ of this approach is theoretically derived and only determined by the sample size, the number of regressors, a small-order sequence, and some constants that do not vary across data generate processes. When the dependence is considered only in the temporal dimension, then the existing approach for independent data can be extended by clustering over the cross-sectional dimension (see Belloni et al. (2016)). However, there is no simple extension if the dependence is present in both temporal and cross-sectional dimensions. Instead, I utilize the component structure characterization of the dependence and decompose the high-dimensional error term $f_{it}V_{it}$ using Hajek projection into three parts: $a_i = \mathrm{E}[f_{it}V_{it}|\alpha_i]$, $g_t = \mathrm{E}[f_{it}V_{it}|\gamma_t]$, $e_{it} = f_{it}V_{it} - a_i - g_t$, where a_i are i.i.d over i, g_t are weakly dependent over t, and the remainder can be shown as small order and is well-behaved too. With this observation, appropriate regressor-specific penalty weights can be constructed, and existing moderate deviation theorems

[6]The demeaning is done because of the inclusion of the intercept term. To avoid it to be penalized, it is usually projected out first.

and concentration inequalities can be leveraged.

With the observation above, I propose the following common penalty level λ and (infeasible) penalty weights:

$$\lambda = \frac{C_\lambda NT}{(N \wedge T)^{1/2}} \Phi^{-1}\left(1 - \frac{\gamma}{2p}\right), \tag{2.7}$$

$$\omega_j = \frac{N \wedge T}{N^2} \sum_{i=1}^{N} a_{i,j}^2 + \frac{N \wedge T}{T^2} \sum_{b=1}^{B} \left(\sum_{t \in H_b} g_{t,j} \right)^2. \tag{2.8}$$

where $a_{i,j} = \mathrm{E}[f_{it,j} V_{it} | \alpha_i]$, $g_{t,j} = \mathrm{E}[f_{it,j} V_{it} | \gamma_t]$ for $j = 1, ..., p$. C_λ is some sufficiently large constant and γ is a small order sequence. The convergence rate of γ affects the convergence rate of the LASSO estimator: as is revealed later, γ should be $o(1)$ for LASSO to be consistent while a larger γ is necessary for a faster convergence rate. Both C_λ and γ do not vary across different data-generating processes. While there is some guidance about choosing C_λ and γ, the choice is not given exactly in the theoretical results. In practice, it is found that $C_\lambda = 2$ and $\lambda = 0.1/\log(p \vee N \vee T)$ delivers desirable finite sample performance. Looking at the definition of ω in 2.8, we notice that the first term in 2.8 is a variance estimator for i.i.d random variables and the second term is a cluster variance estimator as in Bester et al., 2008 where B is the number of clusters/blocks, h is the block length and H_b is the associated index set. Technically, they are chosen as $B = \mathrm{round}(T/h)$, $h = \mathrm{round}(T^{1/5}) + 1$, and, for $b = 1, .., B$, $H_b = \{t : h(b-1) + 1 \leq t \leq hb\}$.

To implement the penalty weights in 2.8, however, we need to estimate $a_{i,j} = \mathrm{E}[f_{it,j} V_{it} | \alpha_i]$ and $g_{t,j} = \mathrm{E}[f_{it,j} V_{it} | \gamma_t]$ with two challenges. With some initial estimation, we can replace V_{it} with the initial residual $\tilde{V}_{it}$. $\tilde{V}_{it}$ then can be updated iteratively by the residuals from the estimation in 2.5 until it converges, meaning that $\tilde{V}_{it}$ does not update anymore up to a small difference. A common estimator for $a_{i,j}$ is then $\hat{a}_{i,j} = \frac{1}{N} \sum_{t=1}^{T} f_{it,j} \tilde{V}_{it}$. Similarly, we use $\hat{g}_{t,j} = \frac{1}{N} \sum_{i=1}^{N} f_{it,j} \tilde{V}_{it}$ for estimating $g_{t,j}$. Observe that this choice of implementing $\sum_{i=1}^{N} a_{i,j}^2$ is equivalent to the feasible penalty weights of cluster-LASSO in Belloni et al. (2016). It shows that the first term of ω clusters over time so it adjusts for the temporal dependence within each unit cluster. The second term clusters over individuals first and then clusters over time within each block, so it adjusts for cross-sectional dependence and weak temporal dependence across time. The validity of estimating those components through cross-sectional and temporal averages is given in Menzel (2021) for exchangeable arrays. Extending the consistency results for non-exchangeable arrays is not trivial and establishing the uniform convergence result, required due to the high dimensionality, is rather challenging and not the focus of this paper. Following Belloni et al. (2012) and Belloni et al. (2016), the statistical analysis of the weighted LASSO approach is based on high-level assumptions on the feasible penalty weights: Let $\hat{\omega}$ be the feasible diagonal weights and suppose there exists $0 < 1/c_1 < l \leq 1$ and $1 \leq u < \infty$ such that $l \to 1$ and

$$l\omega_j^{1/2} \leq \hat{\omega}_j^{1/2} \leq u\omega_j^{1/2}, \text{ uniformly over j} = 1, ..., \text{p}, \tag{2.9}$$

where $\{\omega_j\}$ and $\{\hat{\omega}_j\}$ are diagonal entries of ω and $\hat{\omega}$, respectively.

Algorithm: Implementation of the Two-Way Cluster-LASSO

i Obtain the initial residuals $\tilde{V}$: estimate a model with a certain (user-specified) number of the most correlated regressors. [7]

ii Set λ according to 2.7 with $C_\lambda = 2$ and $\gamma = 0.1/\log(p \vee N \vee T)$. Calculate $\tilde{\omega}$ according to 2.8.

iii Using $\tilde{\omega}$ for LASSO estimation as in 2.5 and update the residual $\tilde{V}$ using the (post) LASSO estimation.[8]

iv Repeat steps ii-iii until it converges. Obtains the (post) LASSO estimates from the last iteration.

In the low-dimensional case, a key identifying condition is that the population Gram matrix $\mathrm{E}_P[f_{it}'f_{it}]$ is non-singular so that the empirical Gram matrix is also non-singular with high probability. However, as we allow the dimension of f_{it} to be larger than the sample size, the empirical Gram matrix $\mathbb{E}_{NT}[f_{it}'f_{it}]$ is singular. Fortunately, it turns out that we only need certain sub-matrices to be well-behaved for identification. Define

$$\phi_{\min}(m)(M_f) := \min_{\delta \in \Delta(m)} \delta' M_f \delta \text{ and } \phi_{\max}(Cs)(M_f) := \max_{\delta \in \Delta(m)} \delta' M_f \delta,$$

where $\Delta(m) = \{\delta : \|\delta\|_0 = m, \|\delta\|_2 = 1\}$ and $M_f = \mathbb{E}_{NT}[f_{it}'f_{it}]$.

Assumption SE (Sparse Eigenvalues)**.** *For any $C > 0$, there exists constants $0 < \kappa_1 < \kappa_2 < \infty$ such that with probability approaching one, as $(N,T) \to \infty$ jointly, $\kappa_1 \leq \phi_{\min}(Cs)(M_f) < \phi_{\max}(Cs)(M_f) \leq \kappa_2$.*

The sparse eigenvalue assumption follows from Belloni et al. (2012). It implies a restricted eigenvalue condition, which represents a modulus of continuity between the prediction norm and the norm of δ within a restricted set. More primitive sufficient conditions are discussed in Bickel et al. (2009) and Belloni et al. (2012).

Assumption REG (Regularity Conditions)**.** *(i) $\log(p/\gamma) = o\left(T^{1/6}/(\log T)^2\right)$ and $p = o\left(T^{7/6}/(\log T)^2\right)$. (ii) For some $\mu > 1, \delta > 0$, $\max_{j\leq p} \mathrm{E}[|f_{it,j}|^{8(\mu+\delta)}] < \infty$, $\mathrm{E}[|V_{it}|^{8\mu+\delta)}] < \infty$. (iii) $\min_{j\leq p} \mathrm{E}(a_{i,j}^2) > 0$ and $\min_{j\leq p} \mathrm{E}(g_{t,j}^2) > 0$.*

Assumption REG(i) restricts the dimension of f_{it} relative to the sample size. Although the number of regressors is constrained to be of a small order relative to the overall sample size NT as $N,T \to \infty$ jointly, it is still allowed to be greater than the sample size in the finite sample. Note that this requirement is more of a technical constraint and may be further relaxed with refined concentration inequalities for two-way dependent arrays. The moment conditions in Assumption REG(ii) are common in the literature. REG(iii) is a non-degeneracy condition, which is the main case of interest.

[7] This step is for better convergence of the iterative estimation of the penalty weights. A small number of initially included regressors can cause failure to converge.

[8] While they are asymptotically equivalent, post-LASSO suffers from less shrinkage bias in the finite sample.

A common way to mitigate the shrinkage bias of LASSO is to apply least square estimation based on the selected model by LASSO, which is named Post-LASSO. The next theorem delivers a similar result. Let $\widehat{\Gamma} = \{j \in 1, ..., p : |\widehat{\zeta}_j| > 0\}$ where $\widehat{\zeta}_j$ are two-way LASSO estimates. The next theorem gives convergence rates for both two-way cluster-LASSO and its associated Post-LASSO.

Theorem 2.1. *Suppose Assumptions AHK, ASM, AR, REG hold for model 2.1 as $N, T \to \infty$ jointly with $N/T \to c$. Then, by setting λ as 2.7 with some sufficiently large C_λ, we have (i) the event 2.6 happens with probability approaching one. Additionally, suppose that Assumption SE holds and $\widehat{\omega}$ satisfies condition 2.9. Let $\widehat{\zeta}$ be the two-way cluster-LASSO estimator or the post-LASSO estimator based on the two-way cluster-LASSO selection. Then, (ii) $\|\widehat{\zeta}\|_0 = O_P(s)$, and (iii)*

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(f_{it}\widehat{\zeta} - f_{it}\zeta_0\right)^2 =& O_P\left(\frac{s\log(p/\gamma)}{N \wedge T}\right), \\
\left\|\widehat{\zeta} - \zeta_0\right\|_1 =& O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N \wedge T}}\right), \\
\left\|\widehat{\zeta} - \zeta_0\right\|_2 =& O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}}\right).
\end{aligned}$$

Theorem 2.1 establishes convergence rates in terms of the prediction, $l1$, and $l2$ norms for the (post) two-way cluster-LASSO estimator in an approximately sparse model. These results are the first that give convergence rates for a LASSO-based estimator allowing for two-way cluster dependence. It is shown that under the two-way cluster dependence, the two-way cluster-LASSO is consistent but has a convergence rate slower than those of LASSO-based methods under the random sampling condition or weak dependence. Without loss of generality, let $N = N \wedge T$, then by choosing γ according to $\log(1/\gamma) \simeq \log(p \vee NT)$, we have $\left\|\widehat{\zeta} - \zeta_0\right\|_2 = O_P\left(\sqrt{\frac{s\log(p\vee NT)}{N}}\right)$. As a comparison, the rate of convergence in terms of the $l2$ norm is $O_P\left(\sqrt{\frac{s\log p}{NT}}\right)$ under the random sampling and the homoskedasticity Gaussian error assumptions in Bickel et al. (2009) or the heteroskedasticity Gaussian error in Theorem 19.3 of Hansen (2022), $O_P\left(\sqrt{\frac{s\log(p\vee NT)}{NT}}\right)$ under random sampling in Belloni et al. (2012), and $O_P\left(\sqrt{\frac{s\log(p\vee NT)}{Nl_T}}\right)$ under cross-sectional independence in Belloni et al. (2016) where the information index $l_T = 1$ when there is perfect dependence within the cross-sectional cluster.

As illustrated in the Introduction, the slow rate of convergence is due to the underlying factor structure. It is unclear if valid inference is possible under the rate of convergence results in Theorem 2.1 or if it is possible to relax the requirement through a cross-fitting procedure. These questions are addressed in the next section.

3. Clustered-Panel Cross-Fitting and Inference

In this section, I will first propose a sub-sampling scheme for cross-fitting in a two-way clustered panel and then propose a general inference procedure using cross-fitting for a high-dimensional panel model. The idea of the sub-sampling scheme is to split the sample in a proper way so that two resulting sub-samples are independent or, at least, "approximately" independent. With such properties, the sub-sampling scheme can be used for various purposes. For example, it can be used for cross-fitting in a two-step estimation since it effectively eliminates the dependence between the two steps, which in turn relaxes the rate of convergence requirement for the first step for valid inference. It can also be used for cross-validation when choosing tuning parameters in panel data models. In this paper, we will focus on its application in cross-fitting.

Let $\{W_{it} : i = 1, ..., N \textit{ and } t = 1, ..., T\}$ denote a sample of sizes (N, T) from a sequence of random elements $(W_{it})_{i\geq 1, t\geq 1}$ defined on a common measurable space $(\Omega, \mathcal{F})$ and taking values in Euclidean spaces. To allow the dimension of W_{it} to grow with N, T, we denote $(\mathcal{P}_{NT})_{N\geq 1, T\geq 1}$ as an expanding class of probability laws of $\{W_{it} : i = 1, ..., N \textit{ and } t = 1, ..., T\}$ and denote $P \in \mathcal{P}_{NT}$ as a generic probability law for the sample with sizes (N, T).

Under the AHK characterization in Assumption AHK, W_{it} are cluster-dependent over both cross-section and time. Importantly, the cluster dependence does not vanish as the distance between observations (if there is any ordering) increases. If γ_t is weakly dependent, which is the focus of this paper, then the dependence between observations that don't share the same cluster in either dimension dies out as the temporal distance grows. In that case, intuitively, one can split the sample so that the sub-samples do not share the same cluster and are away from each other in temporal distance. This is exactly how this scheme works:

Definition 3.1 (Two-Way Clustered-Panel Cross-Fitting)**.**

(i) Select some positive integers (K, L). Randomly partition the cross-sectional index set $\{1, 2, ..., N\}$ into K folds $\{I_1, I_2, ..., I_K\}$ and partition the temporal index set $\{1, 2, ..., T\}$ into L adjacent folds $\{S_1, S_2, ..., S_L\}$ so that $\bigcup_{k=1}^K I_k = \{1, ..., N\}$, $\bigcup_{l=1}^L S_l = \{1, ..., T\}$[9].

(ii) For each $k = 1, ..., K$ and $l = 1, .., L$, construct the main sample

$$W(k, l) = \{W_{it} : i \in I_k, t \in S_l\},$$

and the auxiliary sample (typically larger)

$$W(-k, -l) = \left\{ W_{it} : i \in \bigcup_{k'\neq k} I_{k'}, t \in \bigcup_{l'\neq l, l\pm 1} S_{l'} \right\},$$

[9] For simplicity, I assume N and T are divisible by K and L, respectively. In practice, if N is not divisible by K, the size for each cross-sectional block can be chosen differently with some length equal to $floor(N/K)$ and others equal to $\text{ceil}(N/K)$. and the same applies to the temporal dimension.

Later on, we also use I_{-k} and S_{-l} to denote the index sets for the auxiliary sample $W(-k,-l)$. Similarly, we denote N_{-k} and T_{-l} as the cross-sectional and temporal sample sizes for the auxiliary sample $W(-k,-l)$. Figure 1 illustrates the cross-fitting with $K = 4$ and $L = 8$.

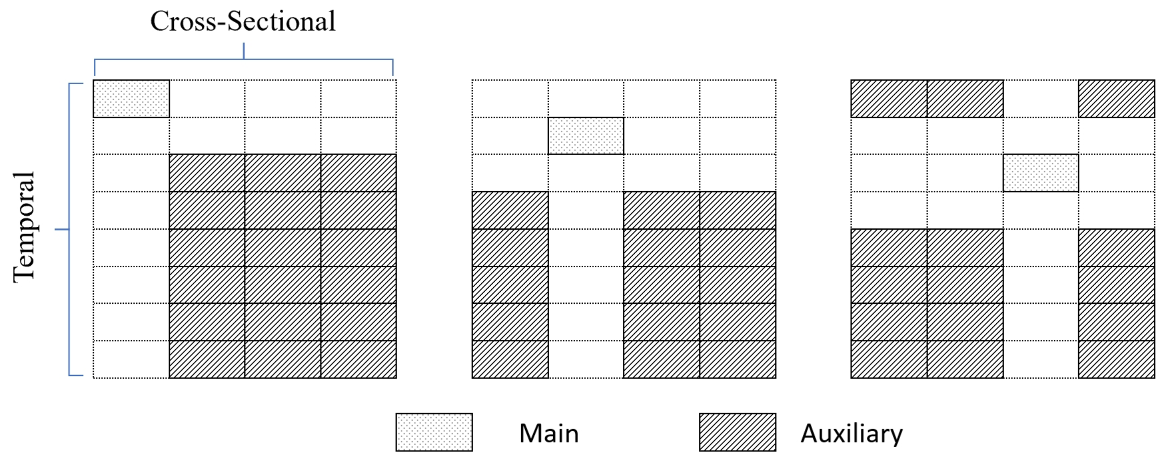

Figure 1: Clustered-Panel cross-fitting with $K = 4$ and $L = 8$. Three graphs from left to right correspond to the main and auxiliary sample constructions with $(k,l) = (1,1)$, $(k,l) = (2,2)$, $(k,l) = (3,3)$. For a simple illustration, observations in the main sample are all adjacent in the cross-sectional dimension but it is not necessary in practice; the same applies to the auxiliary sample.

Since the sub-samples $W(k,l)$ and $W(-k,-l)$ do not share any cluster, they are free from cluster dependence and what's left is the weak dependence over time. Unless imposing $m-$dependence, the sub-samples above will not be independent. However, under certain regularity conditions regarding the weak dependence, it can be shown through the coupling technique that as long as the temporal distance between the sub-samples diverges at a certain rate, there exist coupling sub-samples that are independent of each other while having the same marginal distributions as the constructed sub-samples with probability converging to 1. Such coupling technique is common in the time series context. The following lemma delivers such a result:

Lemma 3.1 (Independent Coupling)**.** *Consider the sub-samples $W(k,l)$ and $W(-k,-l)$ for $k = 1,...,K$ and $l = 1,...,L$. Suppose Assumptions AHK, AR hold and $log(N)/T = o(1)$ as $T \to \infty$. Then, we can construct $\tilde{W}(k,l)$ and $\tilde{W}(-k,-l)$ such that: (i) they are independent of each other; (ii) have the same marginal distribution as $W(k,l)$ and $W(-k,-l)$, respectively; (iii)*

$$\mathrm{P}\left\{(W(k,l), W(-k,-l)) \neq \left(\tilde{W}(k,l), \tilde{W}(-k,-l)\right), \text{ for some } (k,l)\right\} = o(1).$$

Lemma 3.1 shows that the main and auxiliary samples from the proposed clustered-panel cross-fitting scheme are approximately independent as N, T diverge. Note that the hypothetical sample $\tilde{W}(k,l)$ and $\tilde{W}(-k,-l)$ do not matter in practice, but they allow us to treat $W(k,l)$ and $W(-k,-l)$ as $\tilde{W}(k,l)$ and $\tilde{W}(-k,-l)$ with probability approaching 1. The proof of Lemma 3.1 is based on independence coupling results (Strassen, 1965, Dudley and Philipp, 1983, and Berbee, 1987) introduced in Semenova et al. (2023a).

As mentioned at the beginning of the section, one of the primary uses of the sub-sampling scheme is cross-fitting in a two-step estimation. To be concrete, I will define a two-step estimator using the cross-fitting algorithm in the context of a semi-parametric moment restriction model. The theoretical properties of the estimator will be studied in Section 3.1.

Let $\varphi(W_{it}; \theta, \eta)$ denote some identifying moment functions where θ is a low-dimensional vector of parameters of interest and η are nuisance functions. For example, $\eta = g_0$ in 1.1. Let $\psi(W_{it}; \theta, \eta)$ denote some orthogonalized moment function based on $\varphi(W_{it}; \theta, \eta)$. The formal definition of the orthogonality will be delivered in the next subsection. For now, it suffices to be aware that both functions are mean zero while $\psi(W_{it}; \theta, \eta)$ is adjusted for the fact that η_0 needs to be estimated. In model 1.1, $\varphi(W_{it}; \theta, \eta) = D_{it}U_{it}$ and $\psi(W_{it}; \theta, \eta) = \left(D_{it} - E[D_{it}|X_{it}, c_i, d_t]\right)\left(Y_{it} - D_{it}\theta - g(X_{it}, c_i, d_t)\right)$. In the treatment effect model with unconfoundedness conditional on covariates and unobserved heterogeneous effects, $\varphi(W_{it}; \theta, \eta) = \mathrm{E}[Y_{it}|D_{it} = 1, X_{it}, c_i, d_t] - \mathrm{E}[Y_{it}|D_{it} = 0, X_{it}, c_i, d_t] - \theta^{\mathrm{ATE}}$ and $\psi(W_{it}; \theta, \eta)$ is the moment function corresponding to the well-known augmented inverse probability weighting estimator, which is doubly robust.

The panel cross-fitting procedure goes as follows. For each k and l, we use the sub-sample $W(-k, -l)$ to estimate η with the estimator denoted as $\widehat{\eta}_{kl}$. For each $i \in I_k$ and $t \in S_l$, we plug-in $\widehat{\eta}_{kl}$ to the orthogonal moment function, $\psi(W_{it}; \theta, \widehat{\eta}_{kl})$. By averaging $\psi(W_{it}; \theta, \widehat{\eta}_{kl})$ across all $k = 1, ..., K$ and $l = 1, ..., L$, we obtain

$$\overline{\psi}_{kl} := \mathbb{E}_{kl}\left[\psi(W_{it}; \theta, \widehat{\eta}_{kl})\right],$$

which is a sample analog of the population orthogonal moment condition used for estimation. Note that the larger sub-sample $W(-k, -l)$, instead of the smaller sub-sample $W(k, l)$, is used for first-step nuisance estimation because it involves the estimation of high-dimensional unknown parameters. For reference, $W(k, l)$ is referred to as the main sample, and $W(-k, -l)$ is referred to as the auxiliary sample. The next definition summarizes the panel DML estimation and inference procedures for a semiparametric moment restriction model:

Definition 3.2 (Panel DML Algorithm)**.**

(i) Given the identifying moment functions $\varphi(W; \theta, \eta)$ such that $\mathrm{E}_P[\varphi(W; \theta_0, \eta_0)] = 0$, find the orthogonalized moment function $\psi(W, \theta, \eta)$.

(ii) Obtain cross-fitting sub-samples $W(k, l)$ and $W(-k, -l)$ as in Definition 3.1.

(iii) For each k and l, use the sample $W(-k, -l)$ for the first-step estimation and obtain $\widehat{\eta}_{kl}$, then construct $\overline{\psi}_{kl}(\theta) = \mathbb{E}_{kl}[\psi(W_{it}; \theta, \widehat{\eta}_{kl})]$ for each (k, l). Finally, obtain the DML estimator $\widehat{\theta}$ as the solution to

$$\frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\overline{\psi}_{kl}(\theta) = 0. \tag{3.1}$$

Remark 3.1 (The Choice of K and L)**.** *Notice there is a trade-off in setting (K, L) between the first step and second step accuracy: the bigger values of (K, L), the bigger sample size of the auxiliary sample $W(-k, -l)$,* *which is beneficial for high-dimensional first-steps but at the cost of a noisier parametric second step. Due to leaving out the temporal neighborhood, it necessitates an $L \geq 4$ for feasible implementation (if $L = 3$, for example, any main sample $W(k, l)$ with $l = 2$ does not have a well-defined auxiliary sample). On the other hand, it is computationally costly to set the values of (K, L) too large. In practice, $K = 2$ to 4 and $L = 4$ to 8 work well in simulations.*

*3.1. **Panel DML: Inferential Theory***

To investigate the required convergence rate of a high-dimensional estimator for valid inference, I will study a general inference procedure for a high-dimensional panel model characterized by a semiparametric moment restriction. Such an inference procedure is based on the panel cross-fitting approach proposed in Section 3 and the prototypical DML approach proposed in Chernozhukov et al. (2018a).

With the same notations as above, the model is characterized by a semiparametric moment condition $\mathrm{E}[\varphi(W_{it}; \theta_0, \eta_0)] = 0$ where W_{it} are again characterized by an underlying component structure as in Assumption AHK. Let $\psi(W; \theta, \eta)$ be the orthogonalized moment function. Formally, the orthogonality means that it is mean zero and its pathwise or Gateaux derivative with respect to the nuisance parameter is 0 when evaluated at the true values:

$$\mathrm{E}_P[\psi(W_{it}; \theta_0, \eta_0)] = 0, \tag{3.2}$$

$$\partial_r \mathrm{E}_P\left[\psi\left(W_{it}; \theta_0, \eta_0 + r(\eta - \eta_0)\right)\right]|_{r=0} = 0. \tag{3.3}$$

In other words, the nuisance functions have no first-order effect locally on the orthogonalized moment conditions, based on which the estimation of θ_0 is therefore robust to the plug-in of noisy estimates of γ_0. In contrast, the original identifying moment conditions do not possess such a property.

Differing from the existing literature, the approach in this paper focuses on estimation and inference robust to two-way cluster dependence characterized by Assumption AHK. Note that Assumption AHK also includes i.i.d data as a special case. Although the panel DML procedure is also robust to the i.i.d case or, more generally, the case of the degeneracy in components, the theoretical properties are not formally given in this paper. The rates of convergence for both the nuisance estimator and the second-step estimator are different and faster for the i.i.d case but that's not surprising and is not the focus of this paper. To restrict the focus, I will assume a non-degeneracy condition in terms of Hajek projection components. First, I define the Hajek components and their corresponding (long-run) variance-covariance matrices as follows:

$$\begin{aligned}
a_i &:= \mathrm{E}_P\left[\psi\left(W_{it}; \theta_0, \eta_0\right)|\alpha_i\right], \ \ \Sigma_a := \mathrm{E}_P[a_i a_i'], \\
g_t &:= \mathrm{E}_P\left[\psi\left(W_{it}; \theta_0, \eta_0\right)|\gamma_t\right], \ \ \Sigma_g := \sum_{l=-\infty}^{\infty} \mathrm{E}_P[g_t g_{t+l}'], \\
e_{it} &:= \psi\left(W_{it}; \theta_0, \eta_0\right) - a_i - g_t, \ \ \Sigma_e := \sum_{l=-\infty}^{\infty} \mathrm{E}_P[e_{it} e_{i,t+l}'].
\end{aligned}$$

Let $\lambda_{min}[.]$ denote the smallest eigenvalue of a square matrix. The next assumption specifies the non-degeneracy condition and it implies that at least one of the components drives the cluster dependence.

Assumption ND (Non-Degeneracy)**.** *Either $\lambda_{min}[\Sigma_a] > 0$ or $\lambda_{min}[\Sigma_g] > 0$.*

The next two assumptions follow the same format as Chernozhukov et al. (2018a) but, importantly, they characterize some different rates of convergence required for inferential theory. Let (δ_{NT}) and (Δ_{NT}) be some sequence of positive constants converging to 0 as $N, T \to \infty$. Let $\mathcal{T}_{NT}$ be a nuisance realization set such that it contains η_0 and that $\hat{\eta}_{kl}$ belongs to $\mathcal{T}_{NT}$ with probability $1 - \Delta_{NT}$ for each (k, l).

Assumption DML1 (Linear Moment Conditions, Smoothness, and Identification)**.**

(i) *$\psi(W; \theta, \eta)$ is linear in θ:*

$$\psi(w; \theta, \eta) = \psi^a(W, \eta)\theta + \psi^b(W, \eta),\ \forall\ w \in \mathcal{W}, \theta \in \Theta, \eta \in \mathcal{T}.$$

(ii) *$\psi(W; \theta, \eta)$ satisfy the Neyman orthogonality conditions 3.2 and 3.3 with respect to the probability measure P, or, more generally, 3.3 can be replaced by a λ_{NT} near-orthogonality condition*

$$\lambda_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left\| \partial_r \mathrm{E}_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]|_{r=0} \right\| \leq \delta_{NT}/\sqrt{N}.$$

(iii) *The map $\eta \to \mathrm{E}_P[\psi(W_{it}; \theta, \eta)]$ is twice continuously Gateaux-differentiable on $\mathcal{T}$.*

(iv) *The singular values of the matrix $A_0 := \mathrm{E}_P[\psi^a(W_{it}; \eta_0)]$ are bounded below by $c_a > 0$.*

Assumption DML1(i) restricts the focus of this paper to models with linear orthogonal moment conditions, which covers the model in Section 4. For nonlinear orthogonal moment conditions, Chernozhukov et al. (2018a) has shown that the DML estimator has the same desirable properties under more complicated regularity conditions. Focusing on the linear cases allows us to pay more attention to issues specifically attributed to panel data. Assumption DML1(ii) slightly relaxes the orthogonality condition 3.3 by a near-orthogonality condition, which is useful for the approximate sparse model with approximation errors. Assumption DML1(iii) imposes a mild smoothness assumption on the orthogonal moment condition and Assumption DML1(iv) is a common condition for identification.

Assumption DML2 (Moment Regularity and First-Steps)**.**

(i) *For all $i \geq 1$, $t \geq 1$, and some $q > 2$, $c_m < \infty$, the following moment conditions hold:*

$$m_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E}_P \left\| \psi(W_{it}; \theta_0, \eta) \right\|^q \right)^{1/q} \leq c_m,$$
$$m'_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E}_P \left\| \psi^a(W_{it}; \eta) \right\|^q \right)^{1/q} \leq c_m.$$

(ii) The following conditions on the statistical rates r_{NT}, r'_{NT}, λ'_{NT} hold for all $i \geq 1$, $t \geq 1$:

$$
\begin{aligned}
r_{NT} &:= \sup_{\eta \in \mathcal{T}_{NT}} \left\| \mathrm{E}_P[\psi^a(W_{it};\eta) - \psi^a(W_{it};\eta_0)] \right\| \leq \delta_{NT}, \\
r'_{NT} &:= \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E}_P \left\| \psi(W_{it};\theta_0,\eta) - \psi(W_{it};\theta_0,\eta_0) \right\|^2 \right)^{1/2} \leq \delta_{NT}, \\
\lambda'_{NT} &:= \sup_{r \in (0,1), \eta \in \mathcal{T}_{NT}} \left\| \partial_r^2 \mathrm{E}_P[\psi(W_{it};\theta_0,\eta_0 + r(\eta - \eta_0))] \right\| \leq \delta_{NT}/\sqrt{N}.
\end{aligned}
$$

Assumption DML2 regulates the quality of the first-step nuisance estimators. It follows from Chernozhukov et al. (2018a) and it can be verified under primitive conditions in the next section. Observe that, if the orthogonal moment function $\psi(W;\theta,\eta)$ is smooth in η, then λ'_{NT} is the dominant rate and it imposes a crude rate requirement of order $\varepsilon_{NT} = o(N^{-1/4})$ on the first-step nuisance parameter in $L^2(P)$ norm, which is possible for the two-way cluster LASSO estimator to achieve under proper sparsity assumption. Furthermore, in some models including the partial linear model, λ'_{NT} can be exactly 0, then it is possible to obtain the weakest possible rate requirement for the first-step estimator, i.e. $\varepsilon_{NT} = o(1)$.

Theorem 3.1 (Asymptotic Normality and Variance)**.** *Suppose Assumptions AHK, AR, ND, DML1, DML2 hold for any $P \in \mathcal{P}_{NT}$, then for some $\delta_{NT} \geq N^{-1/2}$, as $(N,T) \to \infty$ jointly,*

$$
\sqrt{N}\left(\widehat{\theta} - \theta_0\right) = -\sqrt{N} A_0^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \psi(W_{it};\theta_0,\eta_0) + o_P(1) \Rightarrow \mathcal{N}(0,V),
$$

where

$$
\begin{aligned}
V &:= A_0^{-1} \Omega A_0^{-1\prime}, \\
\Omega &:= \Sigma_a + c\Sigma_g.
\end{aligned}
$$

We observe that the convergence rate of the two-step estimator $\widehat{\theta}$ resulting from the panel DML procedure is non-standard. It is $\sqrt{N}$-consistent instead of $\sqrt{NT}$-consistent. This is because the cluster dependence introduced by the unit and time components does not decay over time or space. Intuitively, with more persistence, the information carried by data is accumulated more slowly. It is a common feature in the literature of robust inference with cluster dependence[10] and it is also related to inferential theory under strong cross-sectional dependence (e.g., Gonçalves, 2011).

Due to the presence of unit and time components, the asymptotic variance is made of (long-run) variance-covariance matrices of both factors. I consider a two-way cluster robust variance estimator similar to Chiang et al. (2024) (CHS estimator) with adjustment due to cross-fitting. The variance estimator is motivated under

[10]For example, see Hansen, 2007, MacKinnon et al., 2021,Menzel, 2021, Chiang et al., 2022,Chiang et al., 2023a,Chiang et al., 2024,Chen and Vogelsang, 2024 among many others.

arbitrary dependence in panel data and is shown to be robust to two-way clustering with correlated time effects in linear panel models. As is shown in Chen and Vogelsang (2024), such variance estimator can be written as an affine combination of three well-known robust variance estimators: Liang-Zeger-Arellano estimator, Driscoll-Kraay estimator, and the "average of HACs" estimator. Applying this result, we can define the CHS-type variance estimator as follows:

$$\widehat{V}_{\text{CHS}} = \widehat{\bar{A}}^{-1}\widehat{\Omega}_{\text{CHS}}\widehat{\bar{A}}^{-1'},$$
$$\widehat{\Omega}_{\text{CHS}} = \widehat{\Omega}_{\text{A}} + \widehat{\Omega}_{\text{DK}} - \widehat{\Omega}_{\text{NW}},$$

where $\widehat{\bar{A}} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{N_kT_l}\sum_{i\in I_k, s\in S_l}\psi^a(W_{it};\widehat{\eta}_{kl})$ and, with $k\left(\frac{m}{M}\right) := 1-\frac{m}{M}$ for $m=0,1,...,M-1$ and 0 otherwise (i.e., Bartlett kernel) and the bandwidth parameter M chosen from 1 to T_l,

$$\widehat{\Omega}_{\text{A}} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{N_kT_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})',$$
$$\widehat{\Omega}_{\text{DK}} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{K/L}{N_kT_l^2}\sum_{t\in S_l, r\in S_l}k\left(\frac{|t-r|}{M}\right)\sum_{i\in I_k, j\in I_k}\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi_{jr}(\widehat{\theta},\widehat{\eta}_{kl})',$$
$$\widehat{\Omega}_{\text{NW}} := \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{K/L}{N_kT_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}k\left(\frac{|t-r|}{M}\right)\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})'.$$

It is noted that the variance estimator under the cross-fitting is equivalent to estimating the variance in each sub-sample and then averaging across all sub-samples. Since K, L are fixed, the asymptotic analysis is done at the sub-sample level. The next theorem establishes the consistency of this variance estimator under the conventional small-bandwidth assumption.

Theorem 3.2 (Consistent Variance Estimator)**.** *Assumptions AHK, AR, ND, DML1, DML2 hold for any* $P \in \mathcal{P}_{NT}$*, and some* $q > 4$ *(defined in Assumption DML2), and* $M/T^{1/2} = o(1)$*. Then, as* $N, T \to \infty$ *and* $N/T \to c$ *where* $0 < c < \infty$*,*

$$\widehat{V}_{\text{CHS}} = V + o_P(1).$$

Theorem 3.2 can be seen as a generalization of the consistency result for the CHS variance estimator in Chiang et al. (2024) by allowing for the estimated nuisance parameters in the moment functions. A remaining practical issue is that $\widehat{V}$ is not ensured to be positive semi-definite. It has been shown in Chen and Vogelsang (2024) that negative variance estimates happen with a non-trivial number of times under certain data-generating processes. Accordingly, an alternative two-term variance estimator was proposed in Chen and Vogelsang (2024). Following the same idea, I propose an alternative variance estimator by dropping the double-counting term $\widehat{\Omega}_{\mathrm{NW}}$:

$$\widehat{V}_{\mathrm{DKA}} = \widehat{A}^{-1}\widehat{\Omega}_{\mathrm{DKA}}\widehat{A}^{-1\prime},$$
$$\widehat{\Omega}_{\mathrm{DKA}} = \widehat{\Omega}_{\mathrm{A}} + \widehat{\Omega}_{\mathrm{DK}}.$$

The estimator is referred to as the DKA variance estimator because it is a sum of Driscoll-Krray and Arellano variance estimators.[11] Similar approaches can be found in MacKinnon et al. (2021). It relies on the fact that the double-counting term is of small order asymptotically when the panel is two-way clustering. Similar to other two-term cluster-robust variance estimators, it has the computational advantage of guaranteeing positive semi-definiteness but at the cost of inconsistency in the case of no clustering or clustering at the intersection. For theoretical results and more detailed discussions on the trade-off between the ensured positive-definiteness and the risk of being too conservative/losing power, readers are referred to MacKinnon et al. (2021) and Chen and Vogelsang (2024).

Theorem 3.3 (Alternative Consistent Variance Estimator)**.** *Under the same conditions as Theorem 3.2, we have, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

$$\widehat{V}_{\mathrm{DKA}} = \widehat{V}_{\mathrm{CHS}} + o_P(1).$$

Theorem 3.3 formally shows that the double-counting term is of small order under two-way clustering and it implies that the $\widehat{V}_{\mathrm{DKA}}$ is also consistent for Ω under two-way clustering.

To conclude, in this section, the inferential theory is established for the panel DML estimator, under high-level assumptions on the first-step estimator. Even though the rate of convergence can be slow for the nuisance estimations due to the two-way cluster dependence, the cross-fitting approach for panel models allows for valid inference in a general moment restriction model with growing dimensions in the nuisance parameters. In the next section, I will study a special case of the semiparametric restriction model but consider the complication due to unobserved heterogeneity.

4. Partial Linear Model with Unobserved Heterogeneity

In this section, a partial linear model with non-additive unobserved heterogeneous effects is considered. The proposed toolkit is flexible enough to allow for models with instrumental variables used for identification,

[11]Note that, the DKA estimator defined in Chen and Vogelsang (2024) differs from the DKA estimator here by a constant term based on fixed-b asymptotic analysis. Such bias correction is not considered here since the fixed-b properties are not directly applicable in this setting. The conjecture is that the same form of bias correction can be applied here but formally establishing the fixed-b asymptotic results with the presence of estimated nuisance parameters is challenging and out of the scope of this paper, and so is left for future research.

so I consider the following model: for $i = 1, ..., N$ and $t = 1, ..., T$,

$$Y_{it} = D_{it}\theta_0 + g(X_{it}, c_i, d_t) + U_{it}, \ \ \mathrm{E}[U_{it}|X_{it}, c_i, d_t] = 0, \tag{4.1}$$

where D_{it} is a low-dimensional vector of endogenous variables; g is an unknown function of potentially high-dimensional control variables X_{it} and unobserved heterogeneous effects (c_i, d_t). For clearer presentation, D_{it} is treated as a scalar variable. In practice, D_{it} can contain some high-order terms and interactions with a low-dimensional vector of controls. If the lags or leads of D_{it} are considered to be exogenous, they can also included in X_{it}. Doing so would not change the theory for estimation and inference but could change the interpretation of θ_0. Consider an excludable instrumental variable Z_{it} such that $\mathrm{E}[Z_{it}U_{it}] = 0$, which gives the identifying moment condition.

To apply the estimation and inference methods proposed in previous sections, g is again assumed to be approximately sparse. However, it does not suffice since (c_i, d_t) are not observed. To deal with the unobserved heterogeneous effects that cause the endogeneity, I take a correlated random-effects approach through the generalized Mundlak device:

Assumption GMD (Generalized Mundlak Device)**.** *For each $i = 1, ..., N$ and $t = 1, ..., T$,*

$$c_i = h_c(\bar{F}_i, \epsilon_i^c), \tag{4.2}$$

$$d_t = h_d(\bar{F}_t, \epsilon_t^d), \tag{4.3}$$

where $\bar{F}_i = \frac{1}{T}\sum_{t=1}^T F_{it}$, $\bar{F}_t = \frac{1}{N}\sum_{i=1}^N F_{it}$, $F_{it} := (D_{it}, X_{it}')'$; h_c and h_d are some unknown measurable functions; the stochastic errors $(\epsilon_i^c, \epsilon_t^d)$ are independent of $(\bar{F}_i, \bar{F}_t, X_{it}, Z_{it}, U_{it}^D, U_{it}^Y)$; and $(c_i, d_t, \epsilon_i, \epsilon_t)$ are independent of U_{it}.

To justify its use, we shall recall the idea of the conventional Mundlak device. Due to the correlation between (c_i, d_t) and the covariates, the endogeneity issue arises if we don't control for the unobserved heterogeneity. To explicitly model the correlation between the random effects and the covariates, Mundlak (1978) proposes an auxiliary regression between the random effects and the cross-sectional sample average and shows that if the random effects enter the model linearly then the resulting estimator GLS estimator is equivalent to the common within-estimator. Wooldridge (2021) further shows that the equivalence relations exist among the POLS estimators resulting from the Mundlak device, within-transformation, and the fixed-effects dummies. Therefore, if the within-transformation and including fixed-effects dummies are sensible ways of dealing with unobserved heterogeneity, then allowing the Mundlak device to have a more flexible function form should also be reasonable and more robust. A similar assumption is considered in Wooldridge and Zhu (2020).

It seems like one can simply apply the panel DML approach from Section 3.1 with the two-way cluster LASSO estimator employed as the first-step estimator except that there is a subtle issue: the Mundlak device uses the full history of the covariates which potentially generates dependence across the cross-fitting sub-samples. Similar issues also appear in a simple linear panel model with additive unobserved effects where within-transformation also introduces sample-averages. Therefore, the cross-fitting may not be compatible with approaches dealing with unobserved heterogeneity, including the proposed generalized Mundlak device. However, without cross-fitting, it is challenging to establish an inferential theory with growing dimensionality in unknown parameters in general. Nevertheless, as is shown below, it is possible to establishing the asymptotic normality of the panel DML estimator using the full sample in both the first and the second steps for the partial linear model with strengthened sparsity condition. This is helpful not only due to the presence of unobserved heterogeneous effects but also because cross-fitting can be computationally costly and it works in a cost of efficiency loss.

Under model 4.1, $g(X_{it}, c_i, d_t) = \mathrm{E}[Y_{it} - D_{it}\theta_0 | X_{it}, c_i, d_t]$. We can rewrite 4.1 as follows:

$$Y_{it} = \left(D_{it} - g_D(X_{it}, c_i, d_t)\right)\theta_0 + g_Y(X_{it}, c_i, d_t) + U_{it}.$$

where $g_D(X_{it}, c_i, d_t) := \mathrm{E}[D_{it}|X_{it}, c_i, d_t]$ and $g_Y(X_{it}, c_i, d_t) := \mathrm{E}[Y_{it}|X_{it}, c_i, d_t]$. Under Assumption GMD, $g_D(X_{it}, c_i, d_t)$ and $g_Y(X_{it}, c_i, d_t)$ can be rewritten as compound functions, which are assumed to be well-approximated by a linear combination of a $\tau-$th order polynomial transformation L^τ as follows:

$$g_D^*(X_{it}, \bar{F}_i, \epsilon_i^c, \bar{F}_t, \epsilon_t^d) := g_D(X_{it}, h_c(\bar{F}_i, \epsilon_i^c), h_d(\bar{F}_t, \epsilon_t^d)) = L^\tau\left(X_{it}, \bar{F}_i, \bar{F}_t, \epsilon_i^c, \epsilon_t^d\right)\eta_D + r_{it}^D \tag{4.4}$$

$$g_Y^*(X_{it}, \bar{F}_i, \epsilon_i^c, \bar{F}_t, \epsilon_t^d) := g_Y(X_{it}, h_c(\bar{F}_i, \epsilon_i^c), h_d(\bar{F}_t, \epsilon_t^d)) = L^\tau\left(X_{it}, \bar{F}_i, \bar{F}_t, \epsilon_i^c, \epsilon_t^d\right)\eta_Y + r_{it}^Y \tag{4.5}$$

where (η_D, η_Y) are slope coefficients and (r_{it}^D, r_{it}^Y) are the approximation errors. Furthermore, we can define a vector of transformed regressors as $L_{1,it} = L^\tau(X_{it}, \bar{F}_i, \bar{F}_t)$ and a vector of unobserved regressors as $L_{2,it} = L^\tau(X_{it}, \bar{F}_i, \bar{F}_t, \epsilon_i^c, \epsilon_t^d) \backslash L^\tau(X_{it}, \bar{F}_i, \bar{F}_t)$. Let $(\eta_{D,1}, \eta_{D,2})$ be such that $\eta_D = \eta_{D,1} \bigcup \eta_{D,2}$ and

$$L^\tau\left(X_{it}, \bar{F}_i, \bar{F}_t, \epsilon_i^c, \epsilon_t^d\right)\eta_D = L_{1,it}\eta_{D,1} + L_{2,it}\eta_{D,2}.$$

$(\eta_{Y,1}, \eta_{Y,2})$ are defined in the same way. Under the sparse approximation and Assumption GMD, we can rewrite model 4.1 as follows:

$$Y_{it} = \left(D_{it} - L_{1,it}\eta_{D,1} - L_{2,it}\eta_{D,2} - r_{it}^D\right)\theta_0 + L_{1,it}\eta_{Y,1} + L_{2,it}\eta_{Y,2} + r_{it}^Y + U_{it},$$

By defining a new error term $V_{it}^g := \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\left(\eta_{Y,2} - \eta_{D,2}\theta_0\right) + U_{it}$, a new approximation error $r_{it} = r_{it}^Y + r_{it}^D\theta_0$, the vector of observables $f_{it} := \left(L_{1,it}, 1\right)$ with dimension denoted by p, and the nuisance vectors $\beta_0 := \left(\eta_{Y,1}, \mathrm{E}[L_{2,it}]\eta_{Y,2}\right)$, $\pi_0 := \left(\eta_{D,1}, \mathrm{E}[L_{2,it}]\eta_{D,2}\right)$, we can rewrite the model above as

$$Y_{it} = \left(D_{it} - f_{it}\pi_0\right)\theta_0 + f_{it}\beta_0 + r_{it} + V_{it}^g. \tag{4.6}$$

Noticeably, in this case, the parameters associated with the unobservables $L_{2,it}$ can be arbitrarily non-sparse.

Given $\mathrm{E}[Z_{it}U_{it}]$ and the independence between Z_{it} and $(\epsilon_i^c, \epsilon_t^d)$, we have the identifying moment condition $\mathrm{E}[Z_{it}V_{it}^g]=0$. Let ζ_0 be the linear projection parameter of Z_{it} onto f_{it} and let V_{it}^Z be the corresponding linear projection errors. By Chernozhukov et al., 2018a, (2.18), the near-Neyman orthogonal moment function is given by:

$$\psi_{it}(\theta_0,\eta_0) := \left(Z_{it} - f_{it}\zeta_0\right)\left(Y_{it} - f_{it}\beta_0 - \left(D_{it} - f_{it}\pi_0\right)\theta_0\right). \tag{4.7}$$

where we denote $\eta_0 = (\zeta_0, \beta_0, \pi_0)$. Under the sparse approximation, we can also rewrite the conditional expectation models for Y and D as

$$\begin{aligned} Y_{it} &= E[Y_{it}|X_{it}, c_i, d_t] + U_{it}^Y = f_{it}\beta_0 + r_{it}^Y + V_{it}^Y \\ D_{it} &= E[Y_{it}|X_{it}, c_i, d_t] + U_{it}^D = f_{it}\pi_0 + r_{it}^D + V_{it}^D. \end{aligned}$$

where $V_{it}^Y = \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\eta_{Y,2} + U_{it}^Y$ and $V_{it}^D = \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\eta_{D,2} + U_{it}^D$. For $l = Z, Y, D$, let ω_l, as defined in 2.8 with V_{it} replaced by V_{it}^l be the infeasible penalty weights for the two-way cluster LASSO estimations of $(\zeta_0, \beta_0, \pi_0)$. Correspondingly, let $\widehat{V}^l$ be the residuals and let $\widehat{\omega}_l$ be the feasible penalty weights. The two-step debiased estimator $\widehat{\theta}$ for θ_0 using the full-sample is defined as the solution of $\mathbb{E}_{NT}[\psi_{it}(\theta,\widehat{\eta})] = 0$ where $\widehat{\eta}$ are the (post) two-way cluster LASSO estimators for η_0 obtained in the first step using the full-sample.

The additional notations introduced below are used in the statistical analysis and delivering the main results:

$$\begin{aligned} a_i &= E[V_{it}^Z V_{it}^g|\alpha_i], \quad g_t = E[V_{it}^Z V_{it}^g|\gamma_t], \quad \Sigma_a = E[a_i a_i'], \quad \Sigma_g = \sum_{l=-\infty}^{\infty} E[g_t g_{t+l}'] \\ A_0 &= E_P[V_{it}^Z V_{it}^D], \quad \Omega_0 = \Sigma_a + c\Sigma_g, \\ a_{i,j,l} &= E[f_{it,j} V_{it}^l|\alpha_i], \quad g_{t,j,l} = E[f_{it,j} V_{it}^l|\gamma_t], \quad l = Z, Y, D \end{aligned}$$

Assumption REG-P (Regularity Conditions for the Partial Linear Model)**.**

(i) *A_0 is non-singular.*

(ii) *For any ϵ, $h_c(F,\epsilon)$ and $h_d(F,\epsilon)$ are invertible in F.*

(iii) *For some $\mu > 1, \delta > 0$, $\max_{j\le p}\mathrm{E}[|f_{it,j}|^{8(\mu+\delta)}] < \infty$ and $\mathrm{E}[|V_{it}^l|^{8(\mu+\delta)}] < \infty$ for $l = g, D, Y, Z$.*

(iv) *Either $\lambda_{min}[\Sigma_a] > 0$ or $\lambda_{min}[\Sigma_g] > 0$, and $\min_{j\le p}\mathrm{E}[a_{i,j}^l]^2 > 0$ and $\min_{j\le p}\mathrm{E}[g_{t,j}^l]^2 > 0$, $l = D, Y, Z$.*

(v) $\log(p/\gamma) = o\left(T^{1/6}/(\log T)^2\right)$ *and* $p = o(T^{7/6}/(\log T)^2)$.

(vi) *The feasible penalty weights $\widehat{\omega}_l$ satisfy condition 2.9 for $l = D, Y, Z$.*

This set of regularity conditions follows from the assumptions for two-way cluster-LASSO and the panel-DML inference. The only extra condition is Assumption REG-P(ii) which is a smoothness condition that ensures the exogeneity properties of $\bar{F}_i$ and $\bar{F}_t$ inherited from (c_i, ϵ_i) and (d_t, ϵ_t).

Theorem 4.1. *Suppose, for $P = P_{NT}$ for each (N,T), the following conditions hold for model 4.1 and $W_{it} = \left(Y_{it}, D_{it}, X_{it}, Z_{it}, U_{it}, c_i, d_t, \epsilon_i, \epsilon_t\right)$: (i) Assumptions AHK, AR, SE, GMD, REG-P; (ii) sparse approximation in 4.4 and 4.5 with $s = o\left(\frac{\sqrt{N \wedge T}}{\log(p/\gamma)}\right)$, $\|r^l_{it}\|_{NT,2} = o_P\left(\sqrt{\frac{1}{N \wedge T}}\right)$ for $l = Y, D$. Then, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V := A_0^{-1}\Omega_0 A_0^{-1}$.

Theorem 4.1 establishes the validity of the proposed inference procedure using the full sample. Note that the sparsity condition and the condition of the approximation errors are stronger than the ones needed for two-way LASSO estimation itself. To estimate the asymptotic variance, the following variance estimators are adapted from Chiang et al. (2024) and Chen and Vogelsang (2024) using the full sample:

$$\hat{V}_{\text{CHS}} = \hat{A}_{NT}^{-1}\hat{\Omega}_{\text{CHS}}\hat{A}_{NT}^{-1\prime}, \quad \hat{\Omega}_{\text{CHS}} = \hat{\Omega}_{\text{A}} + \hat{\Omega}_{\text{DK}} - \hat{\Omega}_{\text{NW}}, \tag{4.8}$$

$$\hat{V}_{\text{DKA}} = \hat{A}_{NT}^{-1}\hat{\Omega}_{\text{DKA}}\hat{A}_{NT}^{-1\prime}, \quad \hat{\Omega}_{\text{DKA}} = \hat{\Omega}_{\text{A}} + \hat{\Omega}_{\text{DK}}, \tag{4.9}$$

where

$$\begin{aligned}
\hat{A}_{NT} &:= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\hat{\zeta})(D_{it} - f_{it}\hat{\pi}), \\
\hat{\Omega}_{\text{A}} &:= \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\psi_{it}(\hat{\theta},\hat{\eta})\psi_{ir}(\hat{\theta},\hat{\eta})', \\
\hat{\Omega}_{\text{DK}} &:= \frac{1}{NT^2}\sum_{t=1}^{T}\sum_{r=1}^{T}k\left(\frac{|t-r|}{M}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\psi_{it}(\hat{\theta},\hat{\eta})\psi_{jr}(\hat{\theta},\hat{\eta})', \\
\hat{\Omega}_{\text{NW}} &:= \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}k\left(\frac{|t-r|}{M}\right)\psi_{it}(\hat{\theta},\hat{\eta})\psi_{ir}(\hat{\theta},\hat{\eta})'.
\end{aligned}$$

For simplicity, we deliver the consistency results of variance estimators assuming the approximation is exact. Allowing for approximation errors does not change the main idea but only requires more regularity conditions on the approximation error and lengthier derivations.

Theorem 4.2. *Suppose assumptions for Theorem 4.1 holds for $P = P_{NT}$ for each (N,T) with $r_{it}^D = r_{it}^Y = 0$ a.s., and $M/T^{1/2} = o(1)$. Then, $(N,T) \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

$$\begin{aligned}
\hat{V}_{\text{CHS}} &= V + o_P(1), \\
\hat{V}_{\text{DKA}} &= \hat{V}_{\text{CHS}} + o_P(1).
\end{aligned}$$

5. Monte Carlo Simulation

In this section, the finite sample performance of the panel DML estimation and inference procedure are examined in a Monte Carlo simulation study. We will start with an exactly sparse linear model without considering approximation errors and unobserved heterogeneous effects, and then we will further consider the partial linear model with correlated random effects.

Firstly, the linear model with high-dimensional covariates and exact sparsity is specified as follows:

$$\begin{aligned}
&\textbf{DGP(i)} - \textbf{Linear Model}:\\
&Y_{it} = \; D_{it}\theta_0 + X_{it}\beta_0 + U_{it},\\
&D_{it} = \; X_{it}\pi_0 + V_{it},
\end{aligned}$$

where $\theta_0 = 1/2$ is the true parameter of interest, and $\beta_0 = c_\beta \times (1, 1, ..., 1, 0, ..., 0)'$, $\pi_0 = c_\pi \times (1, 1, ..., 1, 0, ..., 0)'$ are p-dimensional nuisance parameters where the first s entries are 1 and the rest of the elements are 0; c_β and c_π are constants that control the relevance of the covariates.

Secondly, the partial linear model with correlated random effects is specified as follows:

$$\begin{aligned}
&\textbf{DGP(ii)} - \textbf{Partial Linear Model}:\\
&Y_{it} = \; D_{it}\theta_0 + \left(X_{it}\beta_0 + c_i + d_t\right)^2 + U_{it},\\
&D_{it} = \; \frac{\exp(X_{it}\pi_0)}{1 + \exp(X_{it}\pi_0)} + V_{it},\\
&c_i = \bar{D}_i + \bar{X}_i\xi_0 + \epsilon_i^c, \;\; d_t = \bar{D}_t + \bar{X}_t\zeta_0 + \epsilon_t^d,
\end{aligned}$$

where $\beta_0 = c_\beta \times \left(1/2^2, 1/2^3, ..., 1/2^{(p+1)}\right)'$, $\pi_0 = c_\pi \times \left(1/2^2, 1/2^3, ..., 1/2^{(p+1)}\right)'$, $\xi_0 = c_\xi \left(1/2^2, 1/2^3, ..., 1/2^{(p+1)}\right)$, and $\zeta_0 = c_\zeta \left(1/2^2, 1/2^3, ..., 1/2^{(p+1)}\right)$; ϵ_i^c and ϵ_t^d are each a random draw from the uniform distribution $U(0, 1)$. The nuisance functions in both Y and D are taken as unknown. Although these nuisance functions are not exactly sparse, they are smooth enough and can be well-approximated by a polynomial series. The correlated random effects are generated by the Mundlak device which is taken as known and will be used for estimation.

For the linear model, to feature in the two-way dependence in $V_{it}U_{it}$ as well as $X_{it}U_{it}$ and $X_{it}V_{it}$, (X_{it}, U_{it}, V_{it}) are generated by the underlying components as follows: for each $j = 1, ..., p$,

$$\begin{aligned}
&\textbf{DGP(i)} - \textbf{Additive Components}:\\
&X_{it,j} = \; w_1\alpha_{i,j} + w_2\gamma_{t,j} + w_3\varepsilon_{it,j},\\
&U_{it} = \; w_1\alpha_i^u + w_2\gamma_t^u + w_3\varepsilon_{it}^u,\\
&V_{it} = w_1\alpha_i^v + w_2\gamma_t^v + w_3\varepsilon_{it}^v,
\end{aligned}$$

where the components $\alpha_i^u, \alpha_i^v, \varepsilon_{it}^u, \varepsilon_{it}^v, \alpha_{i,j}, \gamma_{t,j}$ are each random draws from a uniform distribution $U(-\sqrt{3}, \sqrt{3})$

for each j; $\varepsilon_{it} = (\varepsilon_{it,1}, ..., \varepsilon_{it,p})'$ is a random draw from a joint normal distribution with mean 1 and variance-covariance matrix equal to $\imath^{|j-k|}$, $\imath \in [0, 1)$, in the (j, k)'s entry; The components γ_t^u, γ_t^v each follows an AR(1) process with the coefficient equal to ρ and the initial values randomly drawn from the normal distribution with mean 0 and variance $1 - \rho^2$ for some $\rho \in [0, 1)$. The weights (w_1, w_2, w_3) are non-negative with $w_1^2 + w_2^2 + w_3^2 = 1$. The default weights are $w_1 = w_2 = w_3 = 1/\sqrt{3}$.

For the partial linear model, the Mundlak device will be used for estimation. It is well-known that the Mundlak device is mechanically equivalent to within-transformation in a linear panel model, in which the within-transformation would also remove the additive components in DGP(i) and eliminate the two-way dependence in the within-transformed random variables. When the true model is partially linear in the covariates, the Mundlak device also projects out many underlying components and removes most of the dependence driven by the additive components. To illustrate it is not necessarily the case in general, a multiplicative component structure is considered as follows:

$$\begin{aligned}
&\textbf{DGP(ii) – Multiplicative Components}: \\
&X_{it,j} = w_1 \alpha_{i,j} + w_2 \gamma_{t,j} + w_3 \varepsilon_{it,j}, \\
&U_{it} = \frac{w_4}{c_p} \sum_{j=1}^{p} \left[\alpha_i^u \gamma_{t,j} + \alpha_{i,j} \gamma_t^u \right] + w_5 \varepsilon_{it}^u, \\
&V_{it} = \frac{w_4}{c_p} \sum_{j=1}^{p} \left[\alpha_i^v \gamma_{t,j} + \alpha_{i,j} \gamma_t^v \right] + w_5 \varepsilon_{it}^v,
\end{aligned}$$

where the components are generated the same way as in DGP(i) - Linear Components. The weights are non-negative with $w_1^2 + w_2^2 + w_3^2 = 1$ and $w_4^2 + w_5^2 = 1$. The default weights are $w_1 = w_2 = (2/5)^{0.5}$, $w_3 = w_5 = (1/5)^{0.5}$, and $w_4 = (4/5)^{0.5}$. c_p is a scaling factor that ensures the sums of multiplicative components in both U_{it} and V_{it} have variance around 1. With the default weights, c_p is set as $3/2$. The multiplicative components construction here is a generalization of the example in Chiang et al. (2024). To see why $U_{it}V_{it}$ features a component structure, we can expand the product and observe that it includes terms such as $\alpha_i^u \alpha_i^v \gamma_{t,j}^2$ for $j = 1, ..., p$ whose conditional expectations given $\alpha = (\alpha_i^u, \alpha_i^v, \alpha_{i,1}, ..., \alpha_{i,p})$ are $\alpha_i^u \alpha_i^v$ since $\gamma_{t,j}$ has variance 1 and is independent of α. Likewise, the product also includes terms like $\gamma_t^u \gamma_t^v \alpha_{i,j}^2$ whose conditional expectations given $\gamma = (\gamma_t^u, \gamma_t^v, \gamma_{t,1}, ..., \gamma_{t,p})$ are $\gamma_t^u \gamma_t^v$. Importantly, these underlying common factors do not introduce endogeneity as they may seem to.

The simulation study examines the Monte Carlo bias (Bias), standard deviation (SD), mean square error (MSE), and coverage probability of estimators for θ_0. All estimations are based on the orthogonal moment condition given by 4.7 with $Z_{it} = D_{it}$ ($f_{it} = X_{it}$ in DGP(i)). The comparison will be among procedures with and without cross-fitting. The first-step estimations will be based on the POLS estimator (if feasible), the post heteroskedasticity-robust LASSO from Belloni et al. (2012), the post square-root LASSO from Belloni et al. (2011), the post cluster-robust LASSO from Belloni et al. (2016), and the post two-way cluster-LASSO. The CHS-type and DKA-type variance estimators (different formulas for estimations with and without cross-fitting) will be used to obtain sample coverage probabilities. In some unreported simulations, I also compare CHS/DKA type variance estimators with Eicker-Huber-White type estimators in Chernozhukov et al. (2018a) for random sampling data and Cameron-Galbach-Miller type estimator from Chiang et al. (2022) for multiway clustered data. Since it is well-known that inference based on variance estimators not sufficiently accounting for the dependence would cause over-rejection, it is omitted here.

Table 5.1: DGP(i) with $N = T = 25$, $s = 5$, $p = 200$, $\iota = 0.5$, $\rho = 0.5$, $c_\beta = c_\pi = 0.5$

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	First-Step Ave. Sel. D	Second-Step Bias	Second-Step SD	Second-Step RMSE	Coverage (%) CHS	Coverage (%) DKA
No	POLS	200	200	0.003	0.053	0.053	78.9	95.1
	H LASSO	26.0	26.0	0.062	0.065	0.090	58.5	78.7
	R LASSO	17.6	17.6	0.070	0.067	0.097	65.2	79.5
	C LASSO	8.6	8.9	0.036	0.095	0.101	80.0	87.5
	TW LASSO	6.7	6.9	0.023	0.096	0.099	84.3	90.4
Yes	POLS	200	200	0.006	0.113	0.113	98.2	99.4
	H LASSO	16.9	16.6	0.053	0.131	0.141	96.0	97.6
	R LASSO	9.5	9.5	0.054	0.130	0.141	96.0	98.2
	C LASSO	8.0	8.1	0.041	0.130	0.136	96.2	97.4
	TW LASSO	6.7	6.4	0.057	0.126	0.138	95.8	97.2

Note: Simulation results are based on 1000 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. 10 most relevant regressors (based on the sample correlation with the outcome) are used for initial estimation and at most 10 iterations are used in calculating the penalty weights. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO. Post-LASSO POLS is performed in all first steps. Nominal coverage probability: 0.95.

The simulation results are based on 1000 Monte Carlo replications. It is a relatively small number of replications but it is necessitated by the high computational cost of multiple high-dimensional estimation and inference procedures, particularly with cross-fitting. Results are obtained across DGPs varied by the sample sizes (N, T), the dimensions of covariates p, the number of non-zero slope coefficients s, the other sparsity parameter b, the common coefficient a, the multicollinearity parameter ι and the temporal correlation parameter ρ. For the panel DML inferential procedure with cross-fitting, the tuning parameters (K, L), the number of cross-fitting blocks, needs to be chosen. For variance estimation, bandwidth parameters M of the Bartlett kernel are required. I use the min-MSE rule from Andrews (1991) for both purposes. For a generic scalar score v_{it}, the formula is given as follows:

$$\hat{M} = 1.8171 \left(\frac{\widehat{\rho}^2}{\left(1 - \widehat{\rho}^2\right)^2} \right)^{1/3} T^{1/3} + 1,$$

where $\hat{\rho}$ is the OLS estimator from the regression $\bar{\hat{v}}_t = \rho\bar{\hat{v}}_{t-1} + \eta_t$ where $\bar{\hat{v}}_t = \frac{1}{N}\sum_{i=1}^{N}\hat{v}_{it}$ and $\hat{v}_{it} = \hat{U}_{it}\hat{V}_{it}$.

Table 5.1 presents a set of baseline results that are obtained for a decent number of regressors ($p = 200$) among which 5 are associated with non-zero slope coefficients. The number of covariates is much larger than either cross-sectional or temporal dimensions. On the other hand, the number of non-zero coefficients can be regarded as a small order of the sample sizes, approximately satisfying the sparsity condition. In the first step, model selections are done using different LASSO approaches reported in the second column. The number of selected regressors for both Y and D are reported in the third and fourth columns. First, comparing the results obtained without using cross-fitting, it is shown that when the number of regressors is not extremely large relative to the sample size, the POLS estimator dominates the sparse methods through different LASSOs in terms of Monte Carlo bias, standard deviation, and coverage probability obtained using DKA standard error, even though the true model is sparse. Among the sparse methods, the proposed two-way cluster-LASSO exhibits the smallest bias and best coverage, while its standard deviation is slightly larger than the heteroskedastic-robust LASSO and root-LASSO. In terms of selection, the proposed method selects the number of regressors closest to the true number of relevant regressors while other sparse methods over-select to different extents.

Table 5.2: DGP(i) with $N = T = 25$, $s = 5$, $p = 600$, $\iota = 0.5$, $\rho = 0.5$, $c_\beta = c_\pi = 0.5$

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	Sel. D	Second-Step Bias	SD	RMSE	Coverage (%) CHS	DKA
No	POLS	600	600	0.008	0.221	0.221	26.6	38.6
	H LASSO	39.5	39.8	0.073	0.049	0.087	51.2	78.9
	R LASSO	25.1	25.3	0.079	0.055	0.097	52.4	79.1
	C LASSO	14.0	15.2	0.058	0.096	0.112	68.8	78.4
	TW LASSO	6.9	7.5	0.033	0.098	0.103	81.6	88.1
Yes	H LASSO	24.8	24.7	0.056	0.134	0.146	94.5	98.4
	R LASSO	12.1	12.1	0.054	0.137	0.147	94.5	96.1
	C LASSO	10.7	11.6	0.043	0.139	0.145	95.1	96.1
	TW LASSO	6.8	7.6	0.065	0.140	0.154	90.7	95.1

Note: Simulation results are based on 1000 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. 10 most relevant regressors (based on the sample correlation with the outcome) are used for initial estimation and at most 10 iterations are used in calculating the penalty weights. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO. Post-LASSO POLS is performed in all first steps. Nominal coverage probability: 0.95.

When cross-fitting is employed, all methods have witnessed a significant improvement in terms of sample coverage. This is particularly true for LASSO-based methods that are not designed for dependent data. This is not too surprising because those non-robust sparse methods tend to over-select and the cross-fitting is designed to remove the overfitting bias and to restore asymptotic normality. As a cost of cross-fitting, the Monte Carlo standard deviation increased, indicating the efficient loss due to the exclusion of sub-samples in the first-step estimation. It is also worth emphasizing that the CHS- and DKA-type variance estimators designed for cross-fitting approaches play an important role in the desirable sample coverage. In some unreported simulations, it is shown that inference based on the cross-fitting variance estimators proposed in Chernozhukov et al. (2018a) and Chiang et al. (2022) suffer from severe under-coverage. This is not surprising but the implication is more subtle: while two-way dependence potentially affects both estimation and inference, its negative impact on the inference is more salient.

As the dimension of the covariates significantly increases and becomes as large as the overall sample size (so that POLS remains in the competition), a different pattern is revealed. Table 5.2 also reports simulation results under the DGP(i) except that the dimension p now increases to 600, slightly smaller than the overall sample size 625. First, we compare the results obtained without cross-fitting. The simulation results demonstrate that the methods based on the POLS with no selection and those based on the existing LASSO approaches with over-selection all suffer from severe under-coverage. The proposed methods, in contrast, continue to select the number of relevant regressors closest to the true number regardless of the increased number of irrelevant regressors. When cross-fitting is performed, there is again a significant improvement across all approaches in terms of the sample coverage but it is also in the cost of efficiency loss measured by the increase in SD.

Table 5.3: DGP(ii) with $N = T = 25$, $s = p = 10$, $\iota = 0.5$, $\rho = 0.5$, $c_\beta = 1, c_\pi = 4, c_\xi = c_\zeta = 1/4$; 2nd-order polynomial series are used for approximation

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	First-Step Ave. Sel. D	Second-Step Bias	Second-Step SD	Second-Step RMSE	Coverage (%) CHS	Coverage (%) DKA
No	POLS	560	560	0.012	0.173	0.173	54.4	67.4
	H LASSO	12.2	3.4	0.032	0.126	0.130	87.2	90.8
	R LASSO	11.0	3.3	0.030	0.127	0.130	86.2	91.0
	C LASSO	12.3	24.7	0.030	0.127	0.130	87.8	91.8
	TW LASSO	9.3	3.1	0.023	0.127	0.129	87.8	93.6
Yes	H LASSO	9.0	2.6	0.015	0.156	0.157	95.6	98.8
	R LASSO	6.9	2.0	0.010	0.157	0.158	95.8	98.8
	C LASSO	9.1	3.1	0.003	0.153	0.153	96.6	99.0
	TW LASSO	6.8	1.2	0.020	0.151	0.152	97.2	98.8

Note: Simulation results are based on 1000 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. 10 most relevant regressors (based on the sample correlation with the outcome) are used for initial estimation and at most 10 iterations are used in calculating the penalty weights. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO. Post-LASSO POLS is performed in all first steps. Nominal coverage probability: 0.95.

We have seen the case with exact sparsity in Tables 5.1 and 5.2. As claimed in the theory, the proposed estimation and inference procedures are also valid under approximate sparsity. Table 5.3 reports the simulation results under DGP(ii) where the true model is nonlinear in the the control variables and correlated random effects. The functional form of the nonlinearity is not given and is approximated by the second-order polynomial series. While only 10 observable covariates are considered, the Mundlak device and the polynomial transformation generate 560 regressors that will be included in the approximately sparse linear model. Due to the large number of regressors relative to the overall sample size, the approach based on the POLS estimation has the largest Monte Carlo standard deviation and root mean square error, and it suffers from severe under-coverage. Compared to the POLS and other sparse methods, the proposed two-way cluster-LASSO method selects the most sparse model while having the smallest bias and root mean square error, and it also achieves the best coverage. When the clustered-panel cross-fitting is employed in the inference procedure, we find that the Monte Carlo coverage probability for confidence intervals based on CHS-type standard errors improves significantly, and the confidence intervals based on DKA-type standard errors switch from slight under-coverage to over-coverage. As the correlated random effects are used for estimation, they project out most of the components that drive the two-way cluster dependence under DGP(ii). In that case, the adjustment in the standard error formulas due to cross-fitting can be conservative.

6. Empirical Application

In this section, I re-examine the effects of government spending on the output of an open economy following the framework of Nakamura and Steinsson (2014). It is one of the most cited empirical-macro papers on the American Economic Review and it investigates one classic quantity of interest in economics: the government spending multiplier. The question is can we improve on the estimation and inference through more robust and flexible methods? As I will show, it is made possible by the proposed toolkit in this paper.

This framework utilizes the regional variation in military spending in the US to estimate the percentage increase in output that results from the increase of government spending by 1 percent of GDP, i.e. government spending multiplier. It is referred to as the "open economy relative multiplier" because this framework takes advantage of uniform monetary and tax policies across the regions in the US to difference-out their effects on government spending and output. The parameter of interest is a scalar and the baseline model is identified without considering control variables, so why is the high dimensionality relevant here? As it will be revealed very soon, indeed, the high dimensionality from heterogeneity and flexible modeling can be hidden.

Due to the endogeneity in the variation of the regional military procurement, Nakamura and Steinsson (2014) achieves identification through an instrumental variable (IV) approach. As argued by the authors, the national military spending is largely determined by geopolitical events so it is likely exogenous to the unobserved factors of regional military spending and it affects the regional military spending disproportionally. In other words, the identifying assumption is that the buildups and drawdowns in national military spending are not due to unbalanced military development across regions. Based on this observation, a share-shift type IV is considered and the share is estimated by regressing the regional military spending on the national military spending allowing for region-specific constant slope coefficients.[12] To focus on the main idea, the shares are

[12]All quantities, unless specifically defined, are in terms of two-year growth rate of the real per capita values. Per capita is in

taken as given and the resulting instrument variable is treated as observable instead of generated regressors to avoid further complication.

In this paper, to avoid the endogeneity caused by the misspecification of the function form, I extend the linear model with additive unobserved heterogeneous effects to a partial linear model with non-additive unobserved heterogeneous effects. Let D_{it} be the percentage change in per capita regional military spending in state i and time t and Z_{it} be the IV. Specifically, the baseline model from the original study and the one from this paper differ as follows:

$$\begin{aligned}
&\textbf{Baseline model}:\\
&Y_{it} = \theta_0 D_{it} + \pi_i W_t + c_i + d_t + U_{it}.\\
&\textbf{Partial linear model}:\\
&Y_{it} = \theta_0 D_{it} + g(X_{it}, W_t, c_i, d_t) + U_{it}.
\end{aligned}$$

where θ_0 is the parameter of interest, i.e. the true multiplier; X_{it} and W_t are exogenous control variables with the latter being only time-varying; π_i are non-random unit specific slope coefficients of W_t; (c_i, d_t) are unobserved heterogeneous effects. In the original study, the linear model is estimated by the two-stage least square (2SLS) with two-way fixed effects. In the extended model, I model the unobserved heterogeneous effects as correlated random effects and take a sparse approximation approach for the infinite-dimensional nuisance parameters as in Section 4. Specifically, c_i is assumed to be a function of $(\bar{D}_i, \bar{X}_i)$ and d_t is assumed to be a function of $(\bar{D}_t, \bar{X}_t, W_t)$. Then, through sparse approximation, the feasible (near) Neyman-orthogonal moment function is given by 4.7 with $f_{it} = (L^{\tau}(X_{it}, W_t, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t), 1)$.

In the baseline specification of Nakamura and Steinsson (2014), W_t are not included in the baseline model. In their alternative specifications, W_t is chosen as the real interest rate or the change in national oil price. These two variables are never included together in the original study. Note that allowing the unit-specific slope coefficients for controls generates many nuisance parameters: with 51 state groups[13], one control would increase 51 parameters and two controls would generate 102 parameters, without considering interactions or higher order terms. With a sample size of less than 2000, the high dimensionality in nuisance parameters could result in a noisy estimate of θ_0. In this paper, to obtain a more precise estimate and make the excludability assumption of the IV more plausible, besides the controls from the original study, I also consider additional controls. As is shown in Table 3 of Nakamura and Steinsson (2014), the change in state population is likely not affected by the treatment (the regional military spending), so it is immune to the "bad control" problem[14]; But it could affect the treatment and the outcome, so it is included in X_{it}. By considering

terms of total population. Nakamura and Steinsson (2014) also presents results when per capita is calculated using the working age population as a robustness check.

[13]The regions in this analysis are defined by the states. Nakamura and Steinsson (2014) also presents results on regions as clusters of states.

[14]Angrist and Pischke, 2009 and Chen and Kim, 2024 provide detailed discussions on how endogenous control can pollute the

more flexible function forms and additional exogenous control variables, the excludability condition of the instruments is more plausible. On the other hand, the high-dimensionality arose from the flexible function form and the unobserved heterogeneity necessitates the use of high-dimensional methods. Moreover, state-level yearly variables of those macroeconomic characteristics are often considered to be cluster-dependent in both cross-sectional and time groups due to correlated time shocks and state-unobserved factors. These concerns justify the use of robust estimation and inference methods proposed in this paper.

Table 6.1: Multiplier estimates from the original model

(1) Unobs. Heterog.	(2) Oil Price	(3) Real Int.	(4) Pop.	(5) First Stage	(6) IV 1 $\widehat{\theta}$	(7) CHS s.e.	(8) DKA s.e.
	No	No	No	POLS	1.43	0.68	0.81
	Yes	No	No	POLS	1.30	0.56	0.72
Fixed Effects	No	Yes	No	POLS	1.40	0.57	0.70
	Yes	Yes	No	POLS	1.27	0.45	0.71
	Yes	Yes	Yes	POLS	1.36	0.43	0.56

Note: Standard errors are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 5.

The data is available through Nakamura and Steinsson (2014). It is a balanced (after trimming) state-level yearly panel data with 51 states from 1971-2005 years. The military spending data is collected from the electronic database of DD-350 military procurement forms of the US Department of Defense. The state output is measured by state DGP collected from the US Bureau of Economics Analysis (BEA). The state population data is from the Census Bureau. Data on oil prices is from West Texas Intermediate. The Federal Funds rate is from the FRED database of the St. Louis Federal Reserve. The state inflation measures are constructed from several sources. For more details on data construction, readers are referred to Nakamura and Steinsson (2014).

Table 6.1 provides benchmark results for the original model with different choices of control variables. All estimates (columns 6) of are given by 2SLS with two-way fixed effects and the standard errors (s.e.) are calculated using CHS and DKA formulas given in Section 4. The estimates of the multiplier replicate those given in Nakamura and Steinsson (2014) with significant differences in the standard errors. It is because the variance estimates here account for the potential two-way dependence while the variance estimator used in Nakamura and Steinsson (2014) assumes cross-sectional independence.

The main comparisons are done in Tables 6.2 and 6.3. In Table 6.2, no cross-fitting is performed in the first stage. The number of parameters associated with regressors generated by the polynomials transformations are reported in column (4) and the number of selected parameters associated with Z are reported in column (6)[15]. Overall, with more controls and the polynomial transformation of the observables, the stan-

identification/estimation.

[15]Across all first-step LASSO approaches, more parameters associated with Z are selected compared to those associated with Y

Table 6.2: Estimates of the open economy relative multiplier from the extended model.

(1) Cross-Fitting	(2) Unobs. Heterog.	(3) Poly. Trans.	(4) Param. Gen.	(5) First Stage	(6) Z: Param. Sel.	(7) $\widehat{\theta}$	(8) CHS s.e.	(9) DKA s.e.
No	Mundlak	None	7	POLS	7	1.51	0.66	0.82
				H LASSO	2	1.43	0.66	0.81
				C LASSO	4	1.43	0.66	0.81
				TW LASSO	2	1.43	0.70	0.84
No	Mundlak	2nd	35	POLS	35	1.73	0.99	1.15
				H LASSO	6	1.73	1.01	1.17
				CR LASSO	5	1.75	1.02	1.19
				TW LASSO	3	1.47	0.62	0.77
No	Mundlak	3rd	119	POLS	119	2.20	1.19	1.37
				H LASSO	10	1.97	1.16	1.38
				CR LASSO	6	0.98	0.66	0.82
				TW LASSO	5	1.47	0.61	0.76

Note: Tuning parameters are chosen as $C_\lambda = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. 7 most relevant regressors (based on the sample correlation with the outcome) are used for initial estimation and at most 10 iterations are used in calculating the penalty weights. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO. The number of predictors generated by the polynomial transformation and the number of selected predictors for Z are reported in columns (4) and (6). Standard errors are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 5.

dard errors are generally larger than those in 6.1. With no transformations of the original regressors, the estimates obtained by four different methods are similar and they are consistent with the baseline results. It is noticeable that the proposed approach TW LASSO using the DKA-type penalty weights achieves an estimate that is consistent with the baseline results and has the least variability. As the flexibility and number of nuisance parameters increase with the higher-order polynomial transformations, the number of selected regressors increases across all methods. While the standard errors of most approaches climb become larger and the estimates deviate from the baseline results, the proposed approach remains less noisy. This indicates that many higher-order polynomials included in the extended model for robustness in the function form may not matter that much but sorely contribute to the noise; while the existing approaches tend to over-select those terms under potential two-way dependence, the proposed method is robust against over-selection.

As in the Monte Carlo simulation, the results obtained with cross-fitting are also examined. Although the theoretical results for inference procedure based on cross-fitting methods with the presence of the Mundlak device are not formally given in this paper, the conjecture is that it is still valid under the same set of conditions given in Section 4. Table 6.3 demonstrates the comparison between various sparse methods with

and X. The difference in the LASSO selection is less evident for Y and X while the pattern is similar.

Table 6.3: Estimates of the open economy relative multiplier from the extended model.

(1) Cross-Fitting	(2) Unobs. Heterog.	(3) Poly. Trans.	(4) Param. Gen.	(5) First Stage	(6) Z: Param. Ave. Sel.	(7) $\widehat{\theta}$	(8) CHS s.e.	(9) DKA s.e.
				H LASSO	2.0	1.28	1.73	2.00
Yes	Mundlak	None	7	C LASSO	2.0	1.32	1.75	2.03
				TW LASSO	2.6	1.18	1.77	2.05
				H LASSO	5.2	1.12	2.18	2.52
Yes	Mundlak	2nd	35	C LASSO	5.8	1.46	1.95	2.24
				TW LASSO	4.1	1.20	1.42	1.70
				H LASSO	8.3	1.81	3.17	3.47
Yes	Mundlak	3rd	119	C LASSO	6.5	1.25	1.59	1.91
				TW LASSO	5.3	1.50	1.18	1.44

Note: The tuning parameters are chosen as $(K, L) = (4, 8)$, $C_\lambda = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. 7 most relevant regressors (based on the sample correlation with the outcome) are used for initial estimation and at most 10 iterations are used in calculating the penalty weights. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO. The number of predictors generated by the polynomial transformation and the number of selected predictors for Z are reported in columns (4) and (6). Standard errors are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 5.

the clustered-panel cross-fitting [16]. It reveals a similar pattern as in Table 6.2: The variability of different methods increases as the model approximated by higher-order polynomial series, except for the proposed approach which witnesses more accuracy as the approximation is made more flexible.

To conclude, the empirical study of the government spending multiplier using a flexible model and sparse methods illustrates the issue of hidden dimensionality. In the current example, the estimates obtained through the high-dimensional methods do not deviate much from the baseline results, so it implies the nonlinear effects omitted from the original model may not be very relevant. While the proposed two-way cluster-LASSO and the inference procedure with or without cross-fitting remain relatively accurate and provide results as a robustness check, other sparse methods tend to over-select and become too noisy to be interpretable.

7. Conclusion and Discussion

The inferential theory for high-dimensional models is particularly relevant in panel data settings where the modeling of unobserved heterogeneity commonly leads to high-dimensional nuisance parameters. This paper enriches the toolbox of researchers in dealing with high-dimensional panel models. Particularly, I propose a package of tools that deal with the estimation and inference in high-dimensional panel models that feature two-way cluster dependence and unobserved heterogeneity. I first develop a weighted LASSO

[16] Due to a smaller sample used in the first-step estimation and multicollinearity among the polynomial terms, methods based on the POLS first-step is too noisy and so they are omitted for comparison here.

approach that is robust to two-way cluster dependence in the panel data. As is shown in the statistical analysis of the two-way cluster LASSO, the convergence rates are slow due to the cluster dependence, making it challenging for inference purposes. However, by utilizing a cross-fitting method designed for a two-way clustered panel, the rate requirement for the first step can be substantially relaxed, making the proposed two-way cluster-LASSO a feasible first-step estimator for the panel-DML inference procedure in a high-dimensional semiparametric model. Individually, both the two-way cluster-LASSO and the clustered-panel cross-fitting can be of independent interest; Together, they extend the DML approach to panel data settings. I further consider the unobserved heterogeneity in panel models. Due to the potential non-compatibility of cross-fitting with common fixed-effect and random-effect methods, I study the statistical properties of the proposed estimation and inference procedures using the full sample in both the first and the second steps. The validity is established under a slightly stronger sparsity condition in a partial linear panel model, as a special case.

The estimation and inferential theory are empirically relevant. I illustrate the proposed approaches in an empirical example and exemplify that high-dimensionality can be hidden in questions not traditionally considered high-dimensional. In practice, when the question is naturally high-dimensional and answered by panel data, then the proposed approaches are natural solutions. When the questions are originally not high-dimensional, it is reasonable to start with a simple model as a baseline and then extend it to a more general and flexible model for a robustness check.

While both theoretical and simulation results support the proposed approaches, some limitations remain in certain scenarios. The feasible penalty weight estimation is highly non-trivial due to two-way cluster dependence and high dimensionality. The statistical analysis of the two-way cluster LASSO relies on high-level assumptions on the feasible penalty weights. Even though the iterative feasible weights estimation possesses desirable finite sample properties among the scenarios considered in the Monte Carlo simulation, many subtle issues lack of theoretical guarantee. A devoted exploration of such issues requires a more comprehensive treatment and is an important direction of future research.

References

Aldous, D.J., 1981. Representations for partially exchangeable arrays of random variables. Journal of Multivariate Analysis 11, 581–598.

Andrews, D.W., 1994. Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica , 43–72.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817.

Angrist, J.D., Pischke, J.S., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Babii, A., Ball, R.T., Ghysels, E., Striaukas, J., 2023. Machine learning panel data regressions with heavy-tailed dependent data: Theory and application. Journal of Econometrics 237, 105315.

Babii, A., Ghysels, E., Striaukas, J., 2022. Machine learning time series regressions with an application to nowcasting. Journal of Business & Economic Statistics 40, 1094–1106.

Babii, A., Ghysels, E., Striaukas, J., 2024. High-dimensional granger causality tests with an application to vix and news. Journal of Financial Econometrics 22, 605–635.

Baraud, Y., Comte, F., Viennet, G., 2001. Adaptive estimation in autoregression or-mixing regression via model selection. The Annals of Statistics 29, 839–875.

Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models .

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

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies 81, 608–650.

Belloni, A., Chernozhukov, V., Hansen, C., Kozbur, D., 2016. Inference in high-dimensional panel models with an application to gun control. Journal of Business & Economic Statistics 34, 590–605.

Belloni, A., Chernozhukov, V., Wang, L., 2011. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98, 791–806.

Berbee, H., 1987. Convergence rates in the strong law for bounded mixing sequences. Probability theory and related fields 74, 255–270.

Bester, C.A., Conley, T.G., Hansen, C.B., 2008. Inference with dependent data using cluster covariance estimators .

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and dantzig selector .

Breinlich, H., Corradi, V., Rocha, N., Ruta, M., Santos Silva, J., Zylkin, T., 2022. Machine learning in international trade research-evaluating the impact of trade agreements .

Bühlmann, P., Van De Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. Biometrika 81, 351–358.

Chen, K., Kim, K.i., 2024. Identification of nonseparable models with endogenous control variables. arXiv preprint arXiv:2401.14395 .

Chen, K., Vogelsang, T.J., 2024. Fixed-b asymptotics for panel models with two-way clustering. Journal of Econometrics 244, 105831.

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

Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W.K., Robins, J.M., 2022a. Locally robust semiparametric estimation. Econometrica 90, 1501–1535.

Chernozhukov, V., Hausman, J.A., Newey, W.K., 2019. Demand analysis with many prices. Technical Report. National Bureau of Economic Research.

Chernozhukov, V., Karl Härdle, W., Huang, C., Wang, W., 2021a. Lasso-driven inference in time and space. The Annals of Statistics 49, 1702–1735.

Chernozhukov, V., Newey, W., Quintas-Martınez, V.M., Syrgkanis, V., 2022b. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests, in: International Conference on Machine Learning, PMLR. pp. 3901–3914.

Chernozhukov, V., Newey, W.K., Quintas-Martinez, V., Syrgkanis, V., 2021b. Automatic debiased machine learning via neural nets for generalized linear regression. arXiv preprint arXiv:2104.14737 .

Chernozhukov, V., Newey, W.K., Robins, J., 2018b. Double/de-biased machine learning using regularized Riesz representers. Technical Report. cemmap working paper.

Chernozhukov, V., Newey, W.K., Singh, R., 2022c. Automatic debiased machine learning of causal and structural effects. Econometrica 90, 967–1027.

Chetverikov, D., Liao, Z., Chernozhukov, V., 2021. On cross-validated lasso in high dimensions. The Annals of Statistics 49, 1300–1317.

Chiang, H.D., Hansen, B.E., Sasaki, Y., 2024. Standard errors for two-way clustering with serially correlated time effects. Review of Economics and Statistics , 1–40.

Chiang, H.D., Kato, K., Ma, Y., Sasaki, Y., 2022. Multiway cluster robust double/debiased machine learning. Journal of Business and Economic Statistics 40, 1046–1056.

Chiang, H.D., Kato, K., Sasaki, Y., 2023a. Inference for high-dimensional exchangeable arrays. Journal of the American Statistical Association 118, 1595–1605.

Chiang, H.D., Ma, Y., Rodrigue, J., Sasaki, Y., 2021. Dyadic double/debiased machine learning for analyzing determinants of free trade agreements. arXiv preprint arXiv:2110.04365 .

Chiang, H.D., Rodrigue, J., Sasaki, Y., 2023b. Post-selection inference in three-dimensional panel data. Econometric Theory 39, 623–658.

Correia, S., Guimarães, P., Zylkin, T., 2020. Fast poisson estimation with high-dimensional fixed effects. The Stata Journal 20, 95–115.

Davezies, L., D'Haultfœuille, X., Guyonvarch, Y., 2019. Empirical process results for exchangeable arrays. arXiv preprint arXiv:1906.11293 .

Davidson, J., 1994. Stochastic limit theory: An introduction for econometricians. OUP Oxford.

Dehling, H., Wendler, M., 2010. Central limit theorem and the bootstrap for u-statistics of strongly mixing data. Journal of Multivariate Analysis 101, 126–137.

Djogbenou, A.A., MacKinnon, J.G., Nielsen, M.Ø., 2019. Asymptotic theory and wild bootstrap inference with clustered errors. Journal of Econometrics 212, 393–412.

Dudley, R.M., Philipp, W., 1983. Invariance principles for sums of banach space valued random elements and empirical processes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 62, 509–552.

Ellison, M., Lee, S.S., O'Rourke, K.H., 2024. The ends of 27 big depressions. American Economic Review 114, 134–168.

Fama, E.F., French, K.R., 2000. Forecasting profitability and earnings. The journal of business 73, 161–175.

Fernández-Val, I., Lee, J., 2013. Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. Quantitative Economics 4, 453–481.

Fuk, D.K., Nagaev, S.V., 1971. Probability inequalities for sums of independent random variables. Theory of Probability & Its Applications 16, 643–660.

Gao, J., Peng, B., Yan, Y., 2024. Robust inference for high-dimensional panel data models. Available at SSRN 4825772 .

Gao, L., Shao, Q.M., Shi, J., 2022. Refined cramér-type moderate deviation theorems for general self-normalized sums with applications to dependent random variables and winsorized mean. The Annals of Statistics 50, 673–697.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models .

Gonçalves, S., 2011. The moving blocks bootstrap for panel linear regression models with individual fixed effects. Econometric Theory 27, 1048–1082.

Guvenen, F., Schulhofer-Wohl, S., Song, J., Yogo, M., 2017. Worker betas: Five facts about systematic earnings risk. American Economic Review 107, 398–403.

Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27, 1152–1191.

Hansen, B., 2022. Econometrics. Princeton University Press.

Hansen, B.E., 1992. Consistent covariance matrix estimation for dependent heterogeneous processes. Econometrica , 967–972.

Hansen, C.B., 2007. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. Journal of Econometrics 141, 597–620.

Hoover, D.N., 1979. Relations on probability spaces and arrays of random variables. t, Institute for Advanced Study .

Ichimura, H., 1987. Estimation of single index models. Ph.D. thesis. Massachusetts Institute of Technology.

Ichimura, H., Newey, W.K., 2022. The influence function of semiparametric estimators. Quantitative Economics 13, 29–61. `arXiv:1508.01378`.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research 15, 2869–2909.

Jing, B.Y., Shao, Q.M., Wang, Q., 2003. Self-normalized cramér-type large deviations for independent random variables. The Annals of probability 31, 2167–2215.

Jordan, M.I., Wang, Y., Zhou, A., 2023. Data-driven influence functions for optimization-based causal inference. `arXiv:2208.13701`.

Kallenberg, O., 1989. On the representation theorem for exchangeable arrays. Journal of Multivariate Analysis 30, 137–154.

Kallenberg, O., 2005. Probabilistic symmetries and invariance principles. volume 9. Springer.

Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. Journal of Econometrics 186, 325–344.

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

Larrain, B., 2006. Do banks affect the level and composition of industrial volatility? The Journal of Finance 61, 1897–1925.

Li, K., Morck, R., Yang, F., Yeung, B., 2004. Firm-specific variation and openness in emerging markets. Review of Economics and Statistics 86, 658–669.

Lin, J., Michailidis, G., 2017. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. Journal of Machine Learning Research 18, 1–49.

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2021. Wild bootstrap and asymptotic inference with multiway clustering. Journal of Business & Economic Statistics 39, 505–519.

Mattoo, A., Rocha, N., Ruta, M., 2020. Handbook of deep trade agreements. World Bank Publications.

Menzel, K., 2021. Bootstrap with cluster-dependence in two or more dimensions. Econometrica 89, 2143–2188.

Mundlak, Y., 1978. On the pooling of cross-section and time-series data. Econometrica 46, X6.

Nakamura, E., Steinsson, J., 2014. Fiscal stimulus in a monetary union: Evidence from us regions. American Economic Review 104, 753–792.

Newey, W.K.., 1994. The Asymptotic Variance of Semiparametric Estimators. Econometrica 62, 1349–1382.

Ning, Y., Liu, H., 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. The Annals of Statistics 45, 158 – 195.

Peña, V.H., Lai, T.L., Shao, Q.M., 2009. Self-normalized processes: Limit theory and Statistical Applications. Springer.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica , 1403–1430.

Racine, J., 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of econometrics 99, 39–61.

Rajan, R.G., Zingales, L., 1998. Financial dependence and growth. The American Economic Review 88, 559.

Robinson, P.M., 1988. Root-n-consistent semiparametric regression. Econometrica , 931–954.

Roodman, D., Nielsen, M.Ø., MacKinnon, J.G., Webb, M.D., 2019. Fast and wild: Bootstrap inference in stata using boottest. The Stata Journal 19, 4–60.

Semenova, V., Goldman, M., Chernozhukov, V., Taddy, M., 2023a. Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. Quantitative Economics 14, 471–510.

Semenova, V., Goldman, M., Chernozhukov, V., Taddy, M., 2023b. Supplement to "Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence". Quantitative Economics 14, 471–510.

Strassen, V., 1965. The existence of probability measures with given marginals. The Annals of Mathematical Statistics 36, 423–439.

Thompson, S.B., 2011. Simple formulas for standard errors that cluster by both firm and time. Journal of financial Economics 99, 1–10.

Vogt, M., Walsh, C., Linton, O., 2022. Cce estimation of high-dimensional panel data models with interactive fixed effects. arXiv preprint arXiv:2206.12152 .

Wooldridge, J.M., 2021. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Available at SSRN 3906345 .

Wooldridge, J.M., Zhu, Y., 2020. Inference in approximately sparse correlated random effects probit models with panel data. Journal of Business & Economic Statistics 38, 1–18.

Wu, W.B., Wu, Y.N., 2016. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors .

Zhang, C.H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society Series B: Statistical Methodology 76, 217–242.

Appendix A.

We will first introduce two lemmas regarding the law of large number (LLN) and the central limit theorem (CLT) for two-way clustered arrays with correlated time effects. They are restated and generalized from Theorems 1 and 2 in Chiang et al. (2024). The following notations will also be used frequently throughout the appendices: Let $\{W_{it} : i = 1,..,N; t = 1,...,T\}$ be an array of random vectors taking values in $\mathbb{R}^p$. Let $F : \mathbb{R}^p \to \mathbb{R}^k$ be a measurable function where k is a constant. We define the Hajek projection terms $a_i = E[F(W_{it}) - E[F(W_{it})]|\alpha_i]$, $g_t = E[F(W_{it}) - E[F(W_{it})]|\gamma_t]$, and $e_{it} = W_{it} - E[F(W_{it})] - a_i - g_t$ and their corresponding (long-run) variance covariance matrices:

$$\Sigma_a = E[a_i a_i'], \quad \Sigma_g = \sum_{l=-\infty}^{\infty} E[g_t g_{t+l}'], \quad \Sigma_e = \sum_{l=-\infty}^{\infty} E[e_{it} e_{i,t+l}'].$$

We can rewrite $F(W_{it}) = a_i + g_t + e_{it}$. Suppose that W_{it} satisfy Assumptions AHK and AR, then the decomposition has the following properties:

(i) $\{a_i\}_{i\geq 1}$ is a sequence of i.i.d random vectors, $\{g_t\}_{t\geq 1}$ are strictly stationary and β-mixing with the the mixing coefficient $\beta_g(m) \leq \beta_\gamma(m)$ for all $m \geq 1$; for each i, $\{e_{it}\}_{t\geq 1}$ is also strictly stationary; and a_i is independent of g_t.

(ii) a_i, b_t, e_{it} are mean zero.

(iii) Conditional on (γ_t, γ_r), e_{it} and e_{jr} are independent for $j \neq i$.

(iv) The sequences $\{a_i\}$, $\{g_t\}$, $\{e_{it}\}$ are mutually uncorrelated.

Properties (i) and (ii) are straightforward. Property (iii) is due to the assumption that $\{\alpha_i\}$ and $\{\varepsilon_{it}\}$ are each i.i.d sequence and independent of each other. Property (iv) is less obvious. One can show $\mathrm{E}_P[e_{it}|\gamma_r] = 0$ and $\mathrm{E}_P[e_{it}|\alpha_j]$ for any i, t, j, r. It is less obvious to see $\mathrm{E}_P[e_{it}|\gamma_r] = 0$ for some $r \neq t$:

$$\begin{aligned}
\mathrm{E}_P[e_{it}|\gamma_r] &= \mathrm{E}_P[\psi\left(W_{it}; \theta_0, \eta_0\right)|\gamma_r] - \mathrm{E}_P[a_i|\gamma_r] - \mathrm{E}_P[g_s|\gamma_t] \\
&= \mathrm{E}_P[\mathrm{E}_P[\psi\left(f(\alpha_i, \gamma_t, \varepsilon_{it}); \theta_0, \eta_0\right)|\gamma_t, \gamma_r]|\gamma_r] - \mathrm{E}_P[a_i] - \mathrm{E}_P[g_t|\gamma_r] \\
&= \mathrm{E}_P[\mathrm{E}_P[\psi\left(f(\alpha_i, \gamma_t, \varepsilon_{it}); \theta_0, \eta_0\right)|\gamma_t]|\gamma_r] - \mathrm{E}_P[a_i] - \mathrm{E}_P[g_t|\gamma_r] \\
&= \mathrm{E}_P[g_t|\gamma_r] - \mathrm{E}_P[g_t|\gamma_r] = 0
\end{aligned}$$

where the second equality follows from the iterated expectation and the independence of α_i and γ_r and the third equality follows from that given γ_t, γ_r is independent of $(\alpha_i, \gamma_t, \varepsilon_{it})$.

Using the properties above, one can derive the LLN and CLT for two-way clustered panel data. The following lemma is regarding the LLN.

Lemma A.1 *Suppose that W_{it} satisfy Assumptions AHK and AR and $E\left[\|F(W_{it})\|^{4(r+\delta)}\right] < \infty$. Then,*

i $\|\Sigma_a\| < \infty$, $\|\Sigma_g\| < \infty$, *and* $\|\Sigma_e\| < \infty$ *where*

ii $\mathrm{Var}\left(\mathrm{E}_{NT}[F(W_{it})]\right) = \frac{1}{N}\Sigma_a + \frac{1}{T}\Sigma_g(1 + o(1)) + \frac{1}{NT}\Sigma_e(1 + o(1))$ *as* $N, T \to \infty$.

iii $\mathrm{E}_{NT}[F(W_{it})] \xrightarrow{p} E[F(W_{it})]$ *as* $N, T \to \infty$.

Lemma A.2 *With the same setting as in Lemma A.1, further assume that either* $\lambda_{min}\left[\Sigma_a\right] > 0$ *or* $\lambda_{min}\left[\Sigma_g\right] > 0$. *Then, as* $N, T \to \infty$ *and* $N/T \to c$, $\sqrt{N}\left(\mathrm{E}_{NT}[F(W_{it})] - E[F(W_{it})]\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_a + c\Sigma_g)$

Lemmas A.1 and A.2 are the same as those for Theorems 1 and 2 in Chiang et al. (2024) except that W_{it} are replaced by $F(W_{it})$ and we don't consider the i.i.d case here. The proofs with W_{it} replaced by $F(W_{it})$ still go through so they are not repeated here.

The following lemma provides a probability limit of the infeasible penalty weights.

Lemma A.3 *Let* ω_j *be as defined in 2.8 with the bandwidth* M *such that* $M/T^{0.5} = o(1)$. *With the same setting as in Lemma A.2 for* $F(W_{it}) = f_{it,j}V_{it}$, $\omega_j \xrightarrow{p} \frac{N\wedge T}{N}\Sigma_a + \frac{N\wedge T}{T}\Sigma_g$ *as* $N, T \to \infty$ *and* $N/T \to c$.

Proof of Lemma A.3. Since $a_{i,j}$ is independent over i, we can apply the weak law of large number and obtain

$$\frac{N\wedge T}{N^2}\sum_{i=1}^{N} a_{i,j}^2 = \frac{N\wedge T}{N}\Sigma_a + o_P(1)$$

To show the convergence of the second term, we can apply Proposition 2 of Bester et al. (2008) by verifying its Assumption 7. Since the block size here $h = round(T^{1/5}) + 1$, it diverges with the time sample size and $h/T \to 0$ as $T \to \infty$. and Assumption 7(i) follows. Note that the β-mixing property of $g_{t,j}$ implies that it is also α-mixing with the mixing coefficient $\alpha_g(q) \le \beta_g(q) \le \beta_\gamma(q) = c_\kappa \exp(-\kappa q)$ for all $q \ge 1$. Let ζ be some positive constant, then we have

$$\sum_{q=1}^{\infty} q^2 \alpha_g(q)^{\zeta/(4+\zeta)} \le c_\kappa^{\zeta/(4+\zeta)} \sum_{q=1}^{\infty} q^2 \exp(-\kappa\zeta q/(4+\zeta)) = c_\kappa^{\zeta/(4+\zeta)} \sum_{q=1}^{\infty} q^2 \exp(-aq)$$

where $a := \frac{\kappa\zeta}{4+\zeta}$. We can use the ratio test to examine the convergence of sum:

$$\lim_{q\to\infty} \frac{(q+1)^2 \exp(-a(q+1))}{q^2 \exp(-aq)} = \lim_{q\to\infty}\left(\frac{q+1}{q}\right)\exp(-a) = \exp(-a)$$

Since $\kappa > 0$ and $\zeta > 0$, we have $a > 0$ and so $\exp(-a) < 1$. Thus we conclude the infinite sum does not diverge to infinity. The third condition is ensured by our assumptions directly. Thus, by Proposition 2 of Bester et al. (2008), we have

$$\frac{N\wedge T}{T^2}\sum_{b=1}^{B}\left(\sum_{t\in H_b} g_{t,j}\right)^2 = \frac{N\wedge T}{T}\Sigma_g + o_P(1).$$

The conclusion follows. □

The following notations and the lemma is used for deriving the performance bounds for post-LASSO.

Corresponding to $\widehat{\Gamma}$ defined above Theorem 2.1, here we define Γ_0 as the support of ζ_0. Define $\widehat{m} = \|\widehat{\Gamma}\backslash\Gamma_0\|_0$. Define $\mathcal{P}_\Gamma$ as the projection matrix such that it projects an $NT \times 1$ vector onto the linear span of $NT \times 1$ vector f_j with $j \in \Gamma$. The post-LASSO estimator $\widehat{\zeta}_{PL}$ is defined as the OLS estimator of the linear projection of Y_{it} onto $\{f_{it,j} \ : \ j \in \widehat{\Gamma}\}$.

Lemma A.4 *Under Assumption ASM, if* $S_{\max} := \max_{1\le j\le p} |\mathbb{E}_{NT}[\omega^{-1/2} f_{it,j} V_{it}]| \le \frac{\lambda}{2c_1 NT}$, $0 < a = \min_j \omega^{1/2} \le \max_j \omega^{1/2} = b < \infty$, *and* $u \ge 1 \ge l \ge 1/c_1$, *then*

$$
\begin{aligned}
&\|f(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} \\
=&\left(\sqrt{\frac{s}{\phi_{min}(s)(M_f)}} + \sqrt{\frac{\widehat{m}}{\phi_{min}(\widehat{m})(M_f)}}\right) O_P\left(\frac{\lambda}{NT}\right) + O_P\left(\|f(X_{it}) - \left(\mathcal{P}_{\widehat{\Gamma}} f\right)_{it}\|_{NT,2}\right).
\end{aligned}
$$

Proof of Lemma A.4. We can decompose $f(X_{it}) - f_{it}\widehat{\zeta}_{PL}$ as follows:

$$
\begin{aligned}
&f(X_{it}) - f_{it}\widehat{\zeta}_{PL} = f(X_{it}) - (\mathcal{P}_{\widehat{\Gamma}} Y)_{it} = \left((I_{NT} - \mathcal{P}_{\widehat{\Gamma}}) f(X) - \mathcal{P}_{\widehat{\Gamma}} V\right)_{it} \\
=&\left((I_{NT} - \mathcal{P}_{\widehat{\Gamma}}) f - (\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0} + \mathcal{P}_{\Gamma_0}) V\right)_{it} \le \left\|\left(I_{NT} - \mathcal{P}_{\widehat{\Gamma}}) f\right)_{it}\right\|_{NT,2} + \left\|\left(\mathcal{P}_{\Gamma_0} V\right)_{it}\right\|_{NT,2} + \left\|\left(\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0} V\right)_{it}\right\|_{NT,2}.
\end{aligned}
$$

where the last equality follows from the property of the linear projection and the inequality follows from Minkowski's inequality. By Hölder's inequality and the property of spectral norm, we have

$$
\begin{aligned}
&\left\|\left(\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0} V\right)_{it}\right\|_{NT,2} = \frac{1}{\sqrt{NT}}\left\|\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0} V\right\|_2 \le \frac{1}{\sqrt{NT}}\left\|f_{\widehat{\Gamma}\backslash\Gamma_0}(f'_{\widehat{\Gamma}\backslash\Gamma_0} f_{\widehat{\Gamma}\backslash\Gamma_0})^{-1}\right\|_\infty \left\|f'_{\widehat{\Gamma}\backslash\Gamma_0} V\right\|_2 \\
\le&\frac{1}{\sqrt{NT}}\sqrt{\frac{1}{NT\phi_{min}(\widehat{m})(M_f)}}\left(\sum_{j\in\widehat{\Gamma}\backslash\Gamma_0}\left(\sum_{i=1}^N\sum_{t=1}^T f_{it,j}V_{it}\right)^2\right)^{1/2} \le \sqrt{\frac{\widehat{m}}{\phi_{min}(\widehat{m})(M_f)}} S_{\max} \\
=&\sqrt{\frac{\widehat{m}}{\phi_{min}(\widehat{m})(M_f)}} O_P\left(\frac{\lambda}{NT}\right)
\end{aligned}
$$

where the last line follows from $\min_j \omega_j^{1/2} = a > 0$ and $S_{\max} \le \frac{\lambda}{2c_1 NT}$. By similar arguments, we have

$$
\begin{aligned}
&\left\|\left(\mathcal{P}_{\Gamma_0} V\right)_{it}\right\|_{NT,2} = \frac{1}{\sqrt{NT}}\left\|\mathcal{P}_{\Gamma_0} V\right\|_2 \le \frac{1}{\sqrt{NT}}\left\|f_{\Gamma_0}(f'_{\Gamma_0} f_{\Gamma_0})^{-1}\right\|_\infty \left\|f'_{\Gamma_0} V\right\|_2 \\
\le&\frac{1}{\sqrt{NT}}\sqrt{\frac{1}{NT\phi_{min}(s)(M_f)}}\left(\sum_{j\in\Gamma_0}\left(\sum_{i=1}^N\sum_{t=1}^T f_{it,j}V_{it}\right)^2\right)^{1/2} \le \sqrt{\frac{s}{\phi_{min}(s)(M_f)}} O_P\left(\frac{\lambda}{NT}\right).
\end{aligned}
$$

□

Proof of Theorem 2.1. In the proof, we will show L1 and L2 convergence rates for $\widehat{\zeta}$. We will first show the regularization event in terms of the infeasible penalty weights ω as defined in 2.8. Due to the AHK representation as in Assumption AHK, we can decompose $f_{it,j}V_{it}$ as $f_{it,j}V_{it} = a_{i,j} + g_{t,j} + e_{it,j}$ where $a_{i,j} := \mathrm{E}[f_{it,j}V_{it}|\alpha_i]$, $g_{t,j} = \mathrm{E}[f_{it,j}V_{it}|\gamma_t]$, and $e_{it,j} = f_{it,j}V_{it} - a_{i,j} - g_{t,j}$, for $j = 1, .., p$.

To show the regularization event holds with probability approaching one, we bound the probability of the following event for each $j = 1, ..., p$:

$$
\begin{aligned}
&\mathrm{P}\left(\frac{1}{NT}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1 NT}\right)\\
=&\mathrm{P}\left(\omega_j^{-1/2}\left|\frac{1}{N}\sum_{i=1}^{N} a_{i,j} + \frac{1}{T}\sum_{t=1}^{T} g_{t,j} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it,j}\right| > \frac{\lambda}{2c_1 NT}\right)\\
\leq&\mathrm{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\omega_{a,j}^{-1/2} a_{i,j}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}\omega_{g,j}^{-1/2} g_{t,j}\right| + \left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} e_{it,j}\right| > \frac{\lambda}{2c_1 NT}\right)\\
\leq&\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{a,j}^{-1/2} a_{i,j}\right| > \frac{\sqrt{N}\lambda}{6c_1 NT}\right) + \mathrm{P}\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\omega_{g,j}^{-1/2} g_{t,j}\right| > \frac{\sqrt{T}\lambda}{6c_1 NT}\right) + \mathrm{P}\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} e_{it,j}\right| > \frac{\lambda}{6c_1}\right)\\
:=&p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)
\end{aligned}
$$

where $\omega_{a,j} := \frac{N\wedge T}{N^2}\sum_{i=1}^{N} a_{i,j}^2$ and $\omega_{g,j} := \frac{N\wedge T}{T^2}\sum_{b=1}^{B}\left(\sum_{t\in H_b} g_{t,j}\right)^2$. The first inequality follows from the triangle inequality and the fact that $\omega_j^{1/2} = \left(\omega_{a,j} + \omega_{g,j}\right)^{1/2} \geq \max\{\omega_{a,j}^{1/2}, \omega_{g,j}^{1/2}\}$. The second inequality follows from a union-bound inequality. Applying union-bound inequality again, we obtain

$$
\mathrm{P}\left(\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1 NT}\right) \leq \sum_{j=1}^{p}\left[p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)\right]
$$

To bound $p_{1,j}(\lambda)$, we will apply a moderate deviation theorem for self-normalized sums of independent random variables. For $j = 1, ..., p$, define $\Xi_{a,j} = \frac{\left[\mathrm{E}(a_{i,j})^2\right]^{1/2}}{\left[\mathrm{E}(a_{i,j})^3\right]^{1/3}}$. Under Assumption REG(i), $\max_{j\leq p} E|a_{i,j}|^3 < \infty$ by Holder's inequality and Jensen's inequality. By Assumption REG(ii), $\min_{j\leq p} E|a_{i,j}|^2 > 0$. Therefore, $\min_j \Xi_{a,j} > 0$. By Theorem 7.4 of Peña et al. (2009) with $\delta = 1$, we have for any $x \in [0, N^{1/6}\Xi_{a,j}]$ that

$$
\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{a,j}^{-1/2} a_{i,j}\right| > x\right) \leq 2\left(1 - \Phi(x)\right)\left[1 + O(1)\left(\frac{1+x}{N^{1/6}\Xi_{a,j}}\right)^3\right]
$$

Let $l_{a,N}$ be some positive increasing sequence. If $N^{1/6}\Xi_{a,j}/l_{a,N} - 1 > 0$ and $x \in [0, N^{1/6}\Xi_{a,j}/l_{a,N} - 1]$, then

$$\mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{a,j}^{-1/2}a_{i,j}\right|>x\right)\leq 2\left(1-\Phi(x)\right)\left[1+O(1)\left(\frac{1}{l_{a,N}}\right)^{3}\right]$$

Then, setting $\lambda = 6c_1\frac{NT}{\sqrt{N}}\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)$ gives

$$\sum_{j=1}^{p}p_{1,j}(\lambda)\leq 2p(1-\Phi(\Phi^{-1}(1-\gamma/2p)))\leq\gamma[1+O(1)(1/l_{a,N})^{3}]$$

given that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)\in[0,N^{1/6}\min_j \Xi_{a,j}/l_{a,N}-1]$ and $N^{1/6}\min_j \Xi_{a,j}/l_{a,N}-1>0$. Note that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)\lesssim\sqrt{\log(p/\gamma)}=o\left(N^{1/12}/\log N\right)$ under Assumption REG(i) and $N/T\to c$ as $N,T\to\infty$. Therefore, it suffices to take $l_{a,N}=O(\log N)$, and it follows that $\sum_{j=1}^{p}p_{1,kl}(\lambda)\to 0$ as $\gamma\to 0$ and $(N,T)\to\infty$.

To bound $p_{2,j}(\lambda)$, we utilize a moderate deviation theorem for self-normalized sums of weakly dependent random variables. Observe that $g_{t,j}=\mathrm{E}[f_{it,j}V_{it}|\gamma_t]$ is beta-mixing with coefficient $\beta_g(q)$ satisfying

$$\beta_g(q)\leq\beta_\gamma(q)\leq c_\kappa exp(-\kappa q)\;\forall q\in\mathbb{Z}^{+}$$

Furthermore, by the strict stationarity and the non-degeneracy condition in Assumption REG(iii), we can verify that for some $\nu>0$, $\mathrm{E}\left[\sum_{t=r}^{r+m}g_{t,j}\right]^{2}\geq\nu^{2}m$ for all $t\geq 1, r\geq 0, m\geq 1$. By Assumption REG(ii) and Holder's inequality, we have $\mathrm{E}|f_{it,j}V_{it}|^{4(\mu+\delta)}<\infty$ for some $\mu>1,\ \delta>0$. Then, by Theorem 3.2 of Gao et al. (2022) with $\tau=1$ and $\alpha=\frac{1}{1+2\tau}$, we have

$$\sum_{j=1}^{p}\mathrm{P}\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\omega_{g,j}^{-1/2}g_{t,j}\right|>x\right)\leq 2p\left(1-\Phi(x)\right)\left[1+O(1)\left(\frac{1}{l_{g,T}}\right)^{2}\right]$$

uniformly for $x\in\left(0,d_0(\log T)^{-1/2}T^{1/12}/l_{g,T}\right)$ where d_0 is some positive constant and $l_{g,T}$ is some positive increasing sequence. Then, setting $\lambda=6c_1\frac{NT}{\sqrt{T}}\Phi^{-1}(1-\frac{\gamma}{2p})$ gives, for all $j=1,...,p$,

$$\sum_{j=1}^{p}p_{2,j}(\lambda)\leq\gamma\left[1+O(1)\left(\frac{1}{l_{g,T}}\right)^{2}\right]$$

given that $\Phi^{-1}(1-\frac{\gamma}{2p})\in\left(0,d_0(\log T)^{-1/2}T^{1/12}/l_{g,T}\right)$. Under Assumption REG(i), we have $\log(p/\gamma)=o\left(T^{1/6}/(\log T)^2)\right)$ and so $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)\lesssim\sqrt{\log(p/\gamma)}=o\left(T^{1/12}/(\log T\right)$. Therefore, by taking $l_{g,T}=O\left((\log T)^{1/2}\right)$, it follows that $\sum_{j=1}^{p}p_{2,j}(\lambda)\to 0$ as $\gamma\to 0$ and $(N,T)\to\infty$.

Consider $p_{3,j}(\lambda)$. Define $\bar{e}_{i,j}:=\frac{1}{T}\sum_{t=1}^{T}e_{it,j}$. Observe that $\mathrm{E}[\bar{e}_{i,j}]=0$ by iterated expectation and conditional on $\{\gamma_t\}_{t=1}^T$, $\bar{e}_{i,j}$ are independent over i . We have shown previously that $\mathrm{E}|f_{it,j}V_{it}|^{4(\mu+\delta)} < \infty$ for some $\mu > 1,\ \delta > 0$. Given that $e_{it,j} = f_{it,j}V_{it} - a_{i,j} - g_{t,j}$ and $\mathrm{E}|a_{i,j}|^{4(\mu+\delta)} < \infty$, $\mathrm{E}|g_{t,j}|^{4(\mu+\delta)} < \infty$ due to Jansen's inequality and iterated expectation, we have $\mathrm{E}|e_{it,j}|^{4(\mu+\delta)} < \infty$ and so $\mathrm{E}|\bar{e}_{i,j}|^{4(\mu+\delta)} < \infty$ due to Minkowski's inequality. Note that

$$\mathrm{Var}\left(\bar{e}_{i,j}\right) = \frac{1}{T}\sum_{l=-(T-1)}^{T-1}\left(1-\frac{|l|}{T}\right)\mathrm{E}(e_{it,j}e_{i,t+l,j}) = \frac{1}{T}\Sigma_e(1+o(1)).$$

where Σ_e is defined in the beginning Appendix A with $k=1$ in this case. By Lemma A.1, $|\Sigma_{e,j}| < \infty$. Furthermore, as is shown below, $\omega_j^{1/2}$ is bounded from below by some constant $a > 0$. Now, by the conditional version of Corollary 4 from Fuk and Nagaev (1971), there exists some constant a_1 and a_2 such that

$$\begin{aligned}
&\mathrm{P}\left(\left|\sum_{i=1}^N \omega_j^{-1/2}\bar{e}_{i,j}\right| > \frac{\lambda}{6c_1T}|\{\gamma_t\}_{t=1}^T\right) \leq \mathrm{P}\left(\left|\sum_{i=1}^N \bar{e}_{i,j}\right| > \frac{a\lambda}{6c_1T}|\{\gamma_t\}_{t=1}^T\right)\\
&\leq a_1(\lambda/T)^{-4}\sum_{i=1}^N E\left(|\bar{e}_{i,j}|^4|\{\gamma_t\}_{t=1}^T\right) + \exp\left(-\frac{a_2(\lambda/T)^2}{\sum_{i=1}^N \mathrm{Var}\left(\bar{e}_{i,j}|\{\gamma_t\}_{t=1}^T\right)}\right)
\end{aligned}$$

Note that $\exp(-1/z)$ is not globally concave but it is concave for $z > 1/2$ and is bounded by z/e^2 for $z \in (0, 1/2)$ where e is the Euler's number. Denote $z = \frac{(T/\lambda)^2}{a_2}\sum_{i=1}^N \mathrm{Var}(\bar{e}_{i,j}|\{\gamma_t\}_{t=1}^T)$. Then, we have

$$\exp\left(-\frac{a_2(\lambda/T)^2}{\sum_{i=1}^N \mathrm{Var}(\bar{e}_{i,j}|\{\gamma_t\}_{t=1}^T)}\right) = \exp\left(-1/z\right) \leq z/e^2 1\{z\in(0,1/2)\} + \exp(-1/z)1\{z > 1/2\}.$$

By Fubini theorem, Jensen's inequality, and the bounded moments, we have

$$\begin{aligned}
p_{3,j}(\lambda) =&\mathrm{P}\left(\left|\sum_{i=1}^N \omega_j^{-1/2}\bar{e}_{i,j}\right| > \frac{\lambda}{6c_1T}\right)\\
\leq& a_1(\lambda/T)^{-4}\sum_{i=1}^N E\left(|\bar{e}_{i,j}|^4\right) + \frac{(T/\lambda)^2}{a_2}\sum_{i=1}^N \mathrm{Var}(\bar{e}_{i,j})/e^2 + \exp\left(-\frac{a_2(\lambda/T)^2}{\sum_{i=1}^N \mathrm{Var}(\bar{e}_{i,j})}\right)\\
=&O\left((\lambda/T)^{-4}N\right) + O\left(\frac{TN}{\lambda^2}\right) + \exp\left(-\frac{a_2(\lambda/T)^2}{N/T\Sigma_e}\right).
\end{aligned}$$

Therefore, we have $\sum_{j=1}^p p_{3,j}(\lambda) = O\left(p(\lambda/T)^{-4}N\right) + O\left(\frac{pTN}{\lambda^2}\right) + p\exp\left(-\frac{a_2(\lambda/T)^2}{N/T\Sigma_e}\right)$. Given that $N/T \to c$ where $0 < c < \infty$, by taking $\lambda_e = \frac{p^{1/4}TN^{1/4}}{\varepsilon^{1/4}} \vee \frac{\sqrt{pNT}}{\varepsilon^{1/2}} \vee \sqrt{\frac{NT\log(p/\gamma)}{a_2/\Sigma_e}}$ for some $\varepsilon = o(1)$, $\sum_{j=1}^p p_{3,j}(\lambda_e) = o(1)$. Under REG(i), $\log(p/\gamma) = o\left(T^{1/6}/(\log T)^2\right)$ and $p = o(T^{7/6}/(\log T)^2)$, then $\lambda_e = O(\lambda)$ where $\lambda = 6c_1\frac{NT}{\sqrt{N\wedge T}}\Phi^{-1}(1-\frac{\gamma}{2p})$. Therefore, we have shown $\sum_{j=1}^p p_{3,j}(\lambda) \to 0$ for $\lambda = 6c_1\frac{NT}{\sqrt{N\wedge T}}\Phi^{-1}(1-\frac{\gamma}{2p})$.

Put together, we have shown

$$\mathrm{P}\left(\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right|\leq\frac{\lambda}{2c_1NT}\right)\to 1. \tag{A.1}$$

Now, we can apply Lemma 6 of Belloni et al. (2012) to obtain the finite sample bounds on $\left\|f_{it}\left(\widehat{\zeta}-\zeta_0\right)\right\|_{NT,2}$ and $\left|\omega^{1/2}\left(\widehat{\zeta}-\zeta_0\right)\right|_1$. Let δ be some generic vector of nuisance parameters and let J_p^1 be a subset of an index set $J_p=1,\ldots,p$ and $J_p^0=J_p\backslash J_p^1$. Let δ^1 be a copy of δ with its j-th element replaced by 0 for all $j\in J_p^0$ and similarly let δ^0 be a copy of δ with its j-th element replaced by 0 for all $j\in J_p^1$. Define the restricted eigenvalues and Gram matrix as follows:

$$K_C(M_f)=\min_{\delta:\|\delta^0\|_1\leq C\|\delta^1\|_1,\ \|\delta\|\neq 0,\ |J_p^1|\leq s}\frac{\sqrt{s\delta' M_f\delta}}{\left\|\delta^1\right\|_1},\quad M_f=\mathrm{E}_{NT}[f_{it}'f_{it}].$$

Define the weighted restricted eigenvalues as follows:

$$K_C^{\omega}(M_f)=\min_{\delta:\|\omega^{1/2}\delta^0\|_1\leq C\|\omega^{1/2}\delta^1\|_1,\ \|\delta\|\neq 0,\ |J_p^1|\leq s}\frac{\sqrt{s\delta' M_f\delta}}{\left\|\omega^{1/2}\delta^1\right\|_1}.$$

Let $a:=\min_{j=1,\ldots,p}\omega_j^{1/2}$, $b:=\max_{j=1,\ldots,p}\omega_j^{1/2}$. As is shown in Belloni et al. (2016),

$$K_C^{\omega}(M_f)\geq\frac{1}{b}K_{bC/a}(M_f). \tag{A.2}$$

Denote $\Sigma_{a,j}=\left(\mathrm{E}[a_{i,j}^2]\right)^{1/2}$ and $\Sigma_{g,j}=\left(\sum_{l=-\infty}^{\infty}\mathrm{E}[g_{t,j}g_{t+l,j}]\right)^{1/2}$. By Lemma A.2 above, we have $\omega_j\xrightarrow{p}\frac{N\wedge T}{N}\Sigma_{a,j}+\frac{N\wedge T}{T}\Sigma_{g,j}$. By Lemma A.1, $|\Sigma_{a,j}|<\infty$ and $|\Sigma_{g,j}|<\infty$. Assumption REG(iii) implies that $\min_{j\leq p}\Sigma_{a,j}^2>0$. Therefore, we have ω_j bounded below by zero and bounded above for each $j=1,..,p$ with probability approaching one as $N,T\to\infty$. Assumption (ASM), the condition 2.9, and A.1, Lemma 6 of Belloni et al. (2012) implies that

$$\begin{aligned}
\left\|f_{it}\left(\widehat{\zeta}-\zeta_0\right)\right\|_{NT,2}&\leq\left(u+\frac{1}{c_1}\right)\frac{\sqrt{s}\lambda}{NTK_{c_0}^{\omega}(M_f)}+2\|r\|_{NT,2}=O_P\left(\frac{1}{K_{c_0}^{\omega}(M_f)}\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}+\sqrt{\frac{s}{N\wedge T}}\right),\\
\left\|\omega^{1/2}\left(\widehat{\zeta}-\zeta_0\right)\right\|_1&\leq\frac{3c_0\sqrt{s}}{K_{2c_0}^{\omega}(M_f)}\left[\left(u+\frac{1}{c_1}\right)\frac{\sqrt{s}\lambda}{NTK_{c_0}^{\omega}(M_f)}+2\|r\|_{NT,2}\right]+3c_0\frac{NT}{\lambda}\|r\|_{NT,2}^2,\\
&=O_P\left(\frac{s}{K_{2c_0}^{\omega}(M_f)K_{c_0}^{\omega}(M_f)}\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}+\sqrt{\frac{s}{N\wedge T}}+\frac{s/\sqrt{N\wedge T}}{\log(p/\gamma)}\right)
\end{aligned}$$

where $c_0 := \frac{uc+1}{lc-1} > 1$. By A.2, we have $1/K^{\omega}_{c_0}(M_f) \leq b/K_{\bar{C}}(M_f)$ where $\bar{C} := bc_0/a$. By arguments given in Bickel et al. (2009), Assumption SE implies that $1/K_C(M_f) = O_P(1)$ for any $C > 0$. Therefore,

$$\left\| f_{it}\left(\hat{\zeta} - \zeta_0\right)\right\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}}\right), \quad \left|\omega^{1/2}\left(\hat{\zeta} - \zeta_0\right)\right|_1 = O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N \wedge T}}\right).$$

By Holder's inequality and that $\min_j \omega_j^{1/2} \geq a > 0$

$$\|\hat{\zeta} - \zeta_0\|_1 \leq \|\omega^{-1/2}\|_\infty \left|\omega^{1/2}\left(\hat{\zeta} - \zeta_0\right)\right|_1 = O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N \wedge T}}\right) = O_P\left(s\sqrt{\frac{\log(p \vee NT)}{N \wedge T}}\right)$$

where the first inequality follows from the .

The $l2$ rate of convergence will be derived after the sparsity bounds. We now switch the focus to the Post-LASSO. By the finite sample bounds of Lemma A.4, we have

$$\begin{aligned}\|f(X_{it}) - f_{it}\hat{\zeta}_{PL}\|_{NT,2} =& \left(\sqrt{\frac{s}{\phi_{min}(s)(M_f)}} + \sqrt{\frac{\hat{m}}{\phi_{min}(\hat{m})(M_f)}}\right) O_P\left(\frac{\lambda}{NT}\right) \\ &+ O_P\left(\|f(X_{it}) - \left(\mathcal{P}_{\hat{\Gamma}}f\right)_{it}\|_{NT,2}\right),\end{aligned} \tag{A.3}$$

By the finite sample bounds of Lemma 7 from Belloni et al. (2012), we have

$$\|f_{it}(\hat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \leq \|f_{it}(X_{it}) - f_{it}\hat{\zeta}_{PL}\|_{NT,2} + \|r_{it}\|_{NT,2}, \tag{A.4}$$

$$\|\omega^{1/2}(\hat{\zeta}_{PL} - \zeta_0)\|_1 \leq \frac{b\sqrt{\hat{m}+s}}{\sqrt{\phi_{min}(\hat{m}+s)(M_f)}} \times \|f_{it}(\hat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \tag{A.5}$$

$$\|f(X_{it}) - \mathcal{P}_{\hat{\Gamma}}f(X_{it})\|_{NT,2} \leq \left(u + \frac{1}{c_1}\right)\frac{\lambda\sqrt{s}}{NTK^{\omega}_{c_0}(M_f)} + 3\|r_{it}\|_{NT,2}. \tag{A.6}$$

The finite sample bound of Lemma 8 from Belloni et al. (2012) gives

$$\hat{m} \leq \phi_{max}(\hat{m})(M_f)a^{-2}\left(\frac{2c_0\sqrt{s}}{K^{\omega}_{c_0}(M_f)} + \frac{6c_0NT\|r_{it}\|_{NT,2}}{\lambda}\right)^2.$$

where $a > 0$ has been shown previously.

Let $\mathcal{M} = \left\{ m \in \mathbb{N} : m > 2\phi_{max}(m)(M_f)a^{-2}\left(\frac{2c_0\sqrt{s}}{K^{\omega}_{c_0}(M_f)} + \frac{6c_0NT\|r_{it}\|_{NT,2}}{\lambda}\right)^2\right\}$. Lemma 10 of Belloni et al. (2012) gives

$$\widehat{m} \leq \min_{m \in \mathcal{M}} \phi_{max}(m \wedge NT)(M_f) a^{-2} \left(\frac{2c_0\sqrt{s}}{K^{\omega}_{c_0}(M_f)} + \frac{6c_0 NT \|r_{it}\|_{NT,2}}{\lambda} \right)^2 . \tag{A.7}$$

Note that $\frac{6c_0 NT\|r_{it}\|_{NT,2}}{\lambda\sqrt{s}} = O_P(1/\log(p \wedge NT)) \xrightarrow{p} 0$. Recall that $1/K^{\omega}_{c_0}(M_f) \leq b/K_{\bar{C}}(M_f) < \infty$. Let $\mu := \min_m \left\{ \sqrt{\phi_{max}(m)(M_f)/\phi_{min}(m)(M_f)} : m > 18\bar{C}^2 s\phi_{max}(m)(M_f)/K^2_{\bar{C}}(M_f) \right\}$, and let $\bar{m}$ be the integer associated with μ. By the definition of $\mathcal{M}$, it implies that $\bar{m} \in \mathcal{M}$ with probability approaching one, which implies $\bar{m} > \widehat{m}$ due to A.7. By Lemma 9 (the sub-linearity of sparse eigenvalues) from Belloni et al. (2012) and A.7, we have

$$\widehat{m} \lesssim_P s\mu^2 \phi_{min}(\bar{m}+s)/K^2_{\bar{C}} \lesssim s\mu^2\phi_{min}(\hat{m}+s)/K^2_{\bar{C}}.$$

Combining the results above with A.3 and A.6 to gives

$$\|f(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} = O_P\left(\sqrt{\frac{s\mu^2 \log(p/\gamma)}{(N \wedge T)K^2_{\bar{C}}}} + \|r_{it}\|_{NT,2} + \frac{\lambda\sqrt{s}}{NTK^{\omega}_{c_0}(M_f)} \right).$$

Recall that $b < \infty$ and Condition SE imply $1/K^{\omega}_{c_0}(M_f) \leq 1/K_{\bar{C}}(M_f) < \infty$. Then, Condition SE, Condition ASM and the choice of λ together imply

$$\|f(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}} \right).$$

For the $l1$ convergence rate, note that $\|\widehat{\zeta}_{PL} - \zeta_0\|_0 \leq \widehat{m} + s$. Then, applying Cauchy-Schwarz inequality to $\|\widehat{\zeta}_{PL} - \zeta_0\|_1 = \sum_{j=1}^p |\widehat{\zeta}_{PL} - \zeta_0| = \sum_{j \in \{\widehat{\Gamma} \bigcup \Gamma_0\}} |\widehat{\zeta}_{PL} - \zeta_0|$ gives

$$\|\widehat{\zeta}_{PL} - \zeta_0\|_1 \leq \sqrt{\widehat{m}+s}\|\widehat{\zeta}_{PL} - \zeta_0\|_2$$

To derive the convergence rates in $l2$-norm of the Post-LASSO estimator (the $l2$ rate for the LASSO estimator is obtained similarly), we will utilize the sparse eigenvalue condition and the prediction norm. If $\widehat{\zeta}_{PL} - \zeta_0 = 0$, then the conclusion holds trivially. Otherwise, define $b = \left(\widehat{\zeta}_{PL} - \zeta_0\right)/\|\widehat{\zeta}_{PL} - \zeta_0\|_2^{-1}$. Then, we have $\|b\|_2 = 1$ and so $b \in \Delta(\widehat{m}+s) = \{\delta : \|\delta\|_0 = \widehat{m}+s, \|\delta\|_2 = 1\}$. By Assumption SE, we have

$$0 < \kappa_1 \leq \phi_{min}(\widehat{m}+s)(M_f) \leq \frac{(b'M_f b)^{1/2}}{\|b\|_2} = \frac{\left\| f_{it}\left(\widehat{\zeta}_{PL} - \zeta_0\right)\right\|_{NT,2}}{\|\widehat{\zeta}_{PL} - \zeta_0\|_2},$$

Therefore, using the bound on the prediction norm above, we conclude that

$$\|\widehat{\zeta}_{PL} - \zeta_0\|_2 \le \frac{\left\| f_{it}\left(\widehat{\zeta}_{PL} - \zeta_0\right)\right\|_{NT,2}}{\kappa_1} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}}\right).$$

It implies that $\|\widehat{\zeta}_{PL} - \zeta_0\|_1 = \sqrt{\widehat{m} + s}\, O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right) = O_P\left(\sqrt{\frac{s^2\log(p/\gamma)}{N\wedge T}}\right)$.

□

Appendix B.

The following lemma, quoted from Semenova et al. (2023a)(Lemma A.3), is a result follows from the weak form of Strassen's coupling Strassen (1965) and the strong form of Strassen's coupling via Lemma 2.11 of Dudley and Philipp (1983):

Lemma B.1 *Let (X, Y) be random element taking values in Polish space $S = (S_1 \times S_2)$ with laws P_X and P_Y, respectively. Then, we can construct $(\tilde{X}, \tilde{Y})$ taking values in (S_1, S_2) such that (i) they are independent of each other; (ii) their laws $\mathcal{L}(\tilde{X}) = P_X$ and $\mathcal{L}(\tilde{Y}) = P_Y$; (iii)*

$$\mathrm{P}\left\{(X, Y) \neq (\tilde{X}, \tilde{Y})\right\} = \frac{1}{2}\|P_{X,Y} - P_X \times P_Y\|_{TV}$$

The proof is provided in Semenova et al. (2023b). To apply the independence coupling result for cross-fitting in the panel data, we need to introduce another lemma:

Lemma B.2 *Let* $X_1, ..., X_q$ and Y be random elements taking values in Polish space $S = (S_1 \times ... \times S_m \times S_y)$.

$$\beta((X_1, ..., X_m), Y) \le \sum_{i=1}^{q} \beta(X_i, Y).$$

Proof of Lemma B.2. By Lemma B.1, we have

$$\begin{aligned}\beta((X_1, ..., X_m), Y) &= \frac{1}{2}\left\| P_{(X_1,...,X_q),Y} - P_{(X_1,...,X_m)} \times P_Y \right\|_{TV} \\ &= \mathrm{P}((X_1, ..., X_m, Y) \neq (\tilde{X}_1, ..., \tilde{X}_m, \tilde{Y})) \le \sum_{i=1}^{m} \mathrm{P}((X_i, Y) \neq (\tilde{X}_i, \tilde{Y})) = \sum_{i=1}^{m} \beta(X_i, Y),\end{aligned}$$

where the inequality follows from the union bound. □

Now we can prove Lemma 3.1 from the main body of the paper:

Proof of Lemma 3.1. By Lemma B.1, for each (k, l) we have

$$
\begin{aligned}
&\mathrm{P}\{(W(k,l), W(-k,-l)) \neq (\tilde{W}(k,l), \tilde{W}(-k,-l))\} \\
=&\beta(W(k,l), W(-k,-l)) = \beta\left(\{W_{it}\}_{i\in I_k, t\in S_l}, \bigcup_{k'\neq k, l'\neq l, l\pm 1} \{W_{it}\}_{i\in I_{k'}, t\in S_{l'}}\right) \\
\leq& \sum_{i\in I_k} \beta\left(\{W_{it}\}_{t\in S_l}, \bigcup_{k'\neq k, l'\neq l, l\pm 1} \{W_{it}\}_{i\in I_{k'}, t\in S_{l'}}\right) \leq \sum_{k'\neq k, l'\neq l, l\pm 1} \sum_{j\in I_{k'}} \sum_{i\in I_k} \beta\left(\{W_{it}\}_{t\in S_l}, \{W_{jt}\}_{t\in S_{l'}}\right)
\end{aligned}
$$

where the last two inequalities follow from Lemma B.2. Note that for $s, m \geq 1$, we have

$$
\begin{aligned}
&\beta(\{W_{it}\}_{t\leq s}, \{W_{jt}\}_{t\geq s+m}) \\
=&\left\| P_{\{W_{it}\}_{t\leq s}, \{W_{jt}\}_{t\geq s+m}} - P_{\{W_{it}\}_{t\leq s}} \times P_{\{W_{jt}\}_{t\geq s+m}} \right\|_{TV} \leq \sup_{A\in\sigma(\{W_{jt}\}_{t\geq s+m})} \mathrm{E}_P \left|\mathrm{P}(A|\sigma(\{W_{it}\}_{t\leq s})) - \mathrm{P}(A)\right| \\
=&\sup_{A\in\sigma(\{W_{jt}\}_{t\geq s+m})} \mathrm{E}_P \left|\mathrm{P}(\mathrm{P}(A|\sigma(\alpha_i, \{\gamma_t\}_{t\leq s}, \{\varepsilon_{it}\}_{t\leq s}))|\sigma(\{W_{it}\}_{t\leq s})) - \mathrm{P}(A)\right| \\
=&\sup_{A\in\sigma(\{W_{jt}\}_{t\geq s+m})} \mathrm{E}_P \left|\mathrm{P}(A|\sigma(\{\gamma_t\}_{t\leq s}) - \mathrm{P}(A)\right| = \sup_{A\in\sigma(\{\gamma_t\}_{t\geq s+m})} \mathrm{E}_P \left|\mathrm{P}(A|\sigma(\{\gamma_t\}_{t\leq s}) - \mathrm{P}(A)\right| \leq c_\kappa exp(-\kappa m),
\end{aligned}
$$

where the last inequality follows from Assumption 2. Therefore,

$$
\mathrm{P}\{(W(k,l), W(-k,-l)) \neq (\tilde{W}(k,l), \tilde{W}(-k,-l))\} \leq KLN^2 c_\kappa exp(-\kappa T_l),
$$

which in turn gives

$$
\mathrm{P}\{(W(k,l), W(-k,-l)) \neq (\tilde{W}(k,l), \tilde{W}(-k,-l)),\ for\ some\ (k,l)\} \leq K^2 L^2 N^2 c_\kappa exp(-\kappa T_l),
$$

where $T_l = T/L$. Given that $\log(N)/T = o(1)$ and (K, L) are finite, it follows that

$$
\mathrm{P}\{(W(k,l), W(-k,-l)) \neq (\tilde{W}(k,l), \tilde{W}(-k,-l)),\ for\ some\ (k,l)\} = o(1)
$$

□

Proof of Theorem 3.1. By Assumption DML2(i), with probability $1 - \Delta_{NT}$, $\widehat{\eta}_{kl} \in \mathcal{T}_{NT}$. So, $\mathrm{P}(\widehat{\eta}_{kl} \in \mathcal{T}_{NT},\ \forall(k,l)) \geq 1 - KL\Delta_{NT} = 1 - o(1)$. Let's denote the event $\mathrm{P}(\widehat{\eta}_{kl} \in \mathcal{T}_\eta,\ \forall(k,l))$ as $\mathcal{E}_\eta$ and the event $\{(W(k,l), W(-k,-l)) = (\tilde{W}(k,l), \tilde{W}(-k,-l)),\ for\ some\ (k,l)\}$ as $\mathcal{E}_{cp}$. By Lemma 3.1, we have $\mathrm{P}(\mathcal{E}_{cp}) = 1 - o(1)$. By union bound inequality, we have $\mathrm{P}(\mathcal{E}_\eta^c \cup \mathcal{E}_{cp}^c) \leq \mathrm{P}(\mathcal{E}_\eta^c) + \mathrm{P}(\mathcal{E}_{cp}^c) = o(1)$. So, $\mathrm{P}(\mathcal{E}_\eta \cap \mathcal{E}_{cp}) = 1 - \mathrm{P}(\mathcal{E}_\eta^c \cup \mathcal{E}_{cp}^c) \geq 1 - o(1)$.

Let $\widehat{\theta}$ be a solution from equation 3.1. To simplify the notation, we denote

$$
\begin{aligned}
&\widehat{A}_{kl} = \mathbb{E}_{kl}[\psi^a(W_{it}, \widehat{\eta}_{kl})],\ \widehat{\widehat{A}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\widehat{A}_{kl},\ A_0 = \mathrm{E}_P[\psi^a(W_{it};\eta_0)],\\
&\widehat{B}_{kl} = \mathbb{E}_{kl}[\psi^b(W_{it}, \widehat{\eta}_{kl})],\ \widehat{\widehat{B}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\widehat{B}_{kl},\ B_0 = \mathrm{E}_P[\psi^b(W_{it};\eta_0)],\\
&\widehat{\widehat{\psi}}(\theta) = \widehat{\widehat{A}}\theta + \widehat{\widehat{B}},\ \bar{\psi}(\theta,\eta) = \mathbb{E}_{NT}\psi(W_{it};\theta,\eta).
\end{aligned}
$$

Claim B.1. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\|\widehat{\widehat{A}} - A_0\| = O_P(N^{-1/2} + r_{NT})$.

By Claim 1 and Assumption 3(iii) that all singular values of A_0 are bounded below by zero, it follows that all singular values of $\widehat{\widehat{A}}$ are also bounded below from zero, on event $\mathcal{E}_\eta$. Then, by the linearity in Assumption 3(i), we can write $\widehat{\theta} = -\widehat{\widehat{A}}^{-1}\widehat{\widehat{B}},\ \theta_0 = -A_0^{-1}B_0$. Then, by basic algebra, we have

$$
\begin{aligned}
\sqrt{N}(\widehat{\theta} - \theta_0) =& \sqrt{N}(-\widehat{\widehat{A}}^{-1}\widehat{\widehat{B}} - \theta_0) = -\sqrt{N}\widehat{\widehat{A}}^{-1}(\widehat{\widehat{B}} + \widehat{\widehat{A}}\theta_0) = -\sqrt{N}\widehat{\widehat{A}}^{-1}\widehat{\widehat{\psi}}(\theta_0)\\
=& \sqrt{N}A_0^{-1}\bar{\psi}(\theta_0,\eta_0) + \sqrt{N}A_0^{-1}\left(\widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\right)\\
&+ \sqrt{N}\left[\left(A_0 + \widehat{\widehat{A}} - A_0\right)^{-1} - A_0^{-1}\right]\left(\bar{\psi}(\theta_0,\eta_0) + \widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\right)
\end{aligned}
$$

Claim B.2. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\|\widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\| = O_P(r'_{NT}/\sqrt{N} + \lambda_{NT} + \lambda'_{NT})$.

By Assumption DML2(i) and Jensen's inequality, we have $\|A_0\| \leq m'_{NT} \leq c_m$. Then, Claim B.2 implies that

$$
\begin{aligned}
\|\sqrt{N}A_0^{-1}\left(\widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\right)\| =& O_P(1)O_P(\sqrt{N}r'_{NT} + \sqrt{N}\lambda_{NT} + \sqrt{N}\lambda'_{NT})\\
=& O_P(r'_{NT} + \sqrt{N}\lambda_{NT} + \sqrt{N}\lambda'_{NT}),
\end{aligned}
$$

Since $\mathrm{E}[\bar{\psi}(\theta_0,\eta_0)] = 0$, by Lemma A.2, we have $\sqrt{N}\bar{\psi}(\theta_0,\eta_0) \xrightarrow{d} \mathcal{N}(0,\Omega)$ where $\Omega = \Sigma_a + c\Sigma_g$ and $\|\Omega\| < \infty$. By Claims B.1, B.2, and the asymptotic normality of $\sqrt{N}\bar{\psi}(\theta_0,\eta_0)$, we have

$$
\begin{aligned}
&\left\|\sqrt{N}\left[\left(A_0 + \widehat{\widehat{A}} - A_0\right)^{-1} - A_0^{-1}\right]\left(\bar{\psi}(\theta_0,\eta_0) + \widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\right)\right\|\\
\leq& \left\|\widehat{\widehat{A}}^{-1}\right\|\left\|\widehat{\widehat{A}} - A_0\right\|\left\|A_0^{-1}\right\|\left\|\sqrt{N}\left(\bar{\psi}(\theta_0,\eta_0) + \widehat{\widehat{\psi}}(\theta_0) - \bar{\psi}(\theta_0,\eta_0)\right)\right\|\\
=& O_P(1)O_P\left(N^{-1/2} + r_{NT}\right)O_P(1)\left(O_P(1) + O_P\left(r'_{NT} + \sqrt{N}\lambda_{NT} + \sqrt{N}\lambda'_{NT}\right)\right) \quad = O_P\left(N^{-1/2} + r_{NT}\right),
\end{aligned}
$$

and $\sqrt{N}\left(\widehat{\theta}-\theta_0\right) = A_0^{-1}\mathcal{N}(0,\Omega) + O_P\left(N^{-1/2} + r_{NT} + r'_{NT} + \sqrt{N}\lambda_{NT} + \sqrt{N}\lambda'_{NT}\right) \xrightarrow{d} A_0^{-1}\mathcal{N}(0,\Omega)$.

Proof of Claim B.1. Fix any (k,l), we have

$$\left\|\widehat{A}_{kl} - A_0\right\| \leq \left\|\widehat{A}_{kl} - \mathrm{E}_P[\widehat{A}_{kl}|W(-k,-l)]\right\| + \left\|\mathrm{E}_P[\widehat{A}_{kl}|W(-k,-l)] - A_0\right\| =: \|\Delta_{A,1}\| + \|\Delta_{A,2}\|.$$

On the event $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, we have $\widehat{\eta}_{kl} \in \mathcal{T}_{NT}$ and the independence between $W(-k,-l)$ and $W(k,l)$. So, due to Assumption DML2, we have $\left\|\Delta_{A,2}\right\| \leq r_{NT}$. By iterated expectation,- $\mathrm{E}_P[\Delta_{A,1}] = 0$. To simplify the notation, we denote $\ddot{\psi}_{it}^{a,kl} := \psi^a(W_{it},\widehat{\eta}_{kl}) - \mathrm{E}_P[\psi^a(W_{it},\widehat{\eta}_{kl})|W(-k,-l)]$. Consider $\left\|\Delta_{A,1}\right\|$:

$$\begin{aligned}
&\mathrm{E}\left(\left\|\Delta_{A,1}\right\|^2 |W(-k,-l)\right) = \left(\frac{1}{N_kT_l}\right)^2 \mathrm{E}_P\left[\left\|\sum_{i\in I_k,t\in S_l}\ddot{\psi}_{it}^{a,kl}\right\|^2 |W(-k,-l)\right] \\
&\leq \left(\frac{1}{N_kT_l}\right)^2 \sum_{i\in I_k,t\in S_l,r\in S_l}\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{is}^{a,kl}\rangle|W(-k,-l)\right]\right| + \sum_{t\in S_l,i\in I_k,j\in I_k}\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{jt}^{a,kl}\rangle|W(-k,-l)\right]\right| \\
&\quad + \sum_{t\in S_l,i\in I_k}\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{it}^{a,kl}\rangle|W(-k,-l)\right]\right| + 2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k}\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{j,t+m}^{a}\rangle|W(-k,-l)\right]\right| \\
&\quad + 2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i\in I_k}\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{i,t+m}^{a}\rangle|W(-k,-l)\right]\right| =: \left(\frac{1}{N_kT_l}\right)^2\left(a(1)+a(2)+a(3)+2a(4)+2a(5)\right).
\end{aligned}$$

By conditional Cauchy-Schwarz inequality, for any i,t,j,s, we have

$$\begin{aligned}
\left|\mathrm{E}_P\left[\langle\ddot{\psi}_{it}^{a,kl},\ddot{\psi}_{js}^{a,kl}\rangle|W(-k,-l)\right]\right| &\leq \left(\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right]\mathrm{E}_P\left[\|\ddot{\psi}_{js}^{a,kl}\|^2|W(-k,-l)\right]\right)^{1/2} \\
&= \mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&a(1) \leq N_kT_l^2\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right], \quad a(2) \leq N_k^2T_l\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right], \\
&a(3) \leq N_kT_l\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right], \quad a(5) \leq N_kT_l^2\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right].
\end{aligned}$$

On the event $\mathcal{E}_\eta \cap \mathcal{E}_{cp}$, we have, for $i \in I_k, t \in S_l$,

$$\left(\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2|W(-k,-l)\right]\right)^{1/2} \lesssim \left(\mathrm{E}_P\left[\|\psi^a(W_{it},\widehat{\eta}_{kl})\|^2|W(-k,-l)\right]\right)^{1/2} < \infty,$$

where the first inequality follows from expanding the term and applying Jensen's inequality and the second inequality follows from Assumption DML2(i). Let D denote the dimension of $\psi^a(W,\eta)$, then we have

$$a(4) = a(5) + \sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j\in I_k, i\neq j} \sum_{d=1}^{D} \mathrm{E}_P\left[\ddot{\psi}^{a,kl}_{d,i,t}\ddot{\psi}^{a,kl}_{d,j,t+m}|W(-k,-l)\right]$$

For each $i \in I_k, t\in S_l$, we can decompose $\ddot{\psi}^{a,kl}_{d,i,t} = a^{kl}_i + g^{kl}_t + e^{kl}_{it}$ where $a_i = \mathrm{E}[\ddot{\psi}^{a,kl}_{d,i,t}|\alpha_i]$, $g_i = \mathrm{E}[\ddot{\psi}^{a,kl}_{d,i,t}|\gamma_t]$, and $e_{it} = \ddot{\psi}^{a,kl}_{d,i,t} - a_i - g_t$. Conditional on $W(-k,-l)$, $(a^{kl}_i, g^{kl}_t, e^{kl}_{it})$ are mutually uncorrelated, $a_i \perp a_j$ for $i\neq j$, and g^{kl}_t is also beta-mixing with $\beta_g(m) \leq \beta_\gamma(m)$. Therefore, we have

$$\begin{aligned}
&\mathrm{E}_P\left[\ddot{\psi}^{a,kl}_{d,i,t}\ddot{\psi}^{a,kl}_{d,j,t+m}|W(-k,-l)\right] = \mathrm{E}_P\left[g^{kl}_t g^{kl}_{t+m} + e^{kl}_{it}e^{kl}_{j,t+m}|W(-k,-l)\right]\\
=&\mathrm{E}_P\left[g^{kl}_t g^{kl}_{t+m}|W(-k,-l)\right] + \mathrm{E}_P\left[\mathrm{E}_P\left[e^{kl}_{it}e^{kl}_{j,t+m}|\alpha_i,\alpha_j,W(-k,-l)\right]|W(-k,-l)\right]
\end{aligned}$$

Note that β-mixing of γ_t implies α-mixing with the mixing coefficient $\alpha_\gamma(m) \leq \beta_\gamma(m)$ for all $m\in\mathbb{Z}^+$, and conditional on $W(-k,-l)$ and α_i, e^{kl}_{it} is also α-mixing with the mixing coefficient not larger than $\alpha_\gamma(m)$ by Theorem 14.12 of Hansen (2022). Then, we have

$$\begin{aligned}
&\left|\mathrm{E}_P\left[\mathrm{E}_P\left[e^{kl}_{it}e^{kl}_{j,t+m}|\alpha_i,\alpha_j,W(-k,-l)\right]|W(-k,-l)\right]\right| \leq \mathrm{E}_P\left[\left|\mathrm{E}_P\left[e^{kl}_{it}e^{kl}_{j,t+m}|\alpha_i,\alpha_j,W(-k,-l)\right]\right||W(-k,-l)\right]\\
\lesssim&8\alpha_\gamma(m)^{1-2/q}\left(\mathrm{E}_P[|\ddot{\psi}^{a,kl}_{d,i,t}|^q|W(-k,-l)]\right)^{1/q}\left(\mathrm{E}_P[|\ddot{\psi}^{a,kl}_{d,j,t+m}|^q|W(-k,-l)]\right)^{1/q} \lesssim 32\alpha_\gamma(m)^{1-2/q}c_m^2,
\end{aligned}$$

where the first inequality follows from the Jensen's inequality; the second inequality follows from the fact that $\mathrm{E}[e^{kl}_{it}|\alpha_i,W(-k,-l)] = 0$, and Theorem 14.13(ii) of Hansen (2022); the last inequality follows from the moment conditions in Assumption DML2 and that $W(-k,-l)$ is independent of $W(k,l)$ on $\mathcal{E}_{cp}$. Similarly,

$$\left|\mathrm{E}_P\left[g^{kl}_t g^{kl}_{t+m}|W(-k,-l)\right]\right| \lesssim \alpha_\gamma(m)^{1-2/q}c_m^2,$$

Then, we have

$$\begin{aligned}
&\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j\in I_k, i\neq j} \sum_{d=1}^{D} \mathrm{E}_P\left[\ddot{\psi}^{a,kl}_{d,i,t}\ddot{\psi}^{a,kl}_{d,j,t+m}|W(-k,-l)\right]\\
\lesssim& c_m^2\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j\in I_k, i\neq j} \sum_{d=1}^{D}\alpha_\gamma(m)^{1-2/q} \leq c_m^2 D\sum_{m=1}^{\infty} c_\kappa exp(-\kappa m)^{1-2/q} \leq \frac{c_m^2 D c_\kappa}{exp(\kappa(1-2/q))-1} < \infty,
\end{aligned}$$

where the last inequality follows from the geometric sum. Thus, as $(N_k, T_l)\to\infty$ we have

$$\mathrm{E}\left(\left\|\Delta_{A,1}\right\|^2|W(-k,-l)\right) = \left(\frac{1}{N_k T_l}\right)^2[a(1)+a(2)+(3)+2a(4)+2a(5)] = O_P(1/T_l) = O_P(1/N).$$

where the last step follows from that L is constant and $N/T \to c$ as $N,T\to\infty$. By Markov's inequality, we conclude that conditional on $W(-k,-l)$, $\|\Delta_{A,1}\| = O_P(1/\sqrt{N})$. By Lemma 6.1 that conditional convergence implies unconditional convergence, we have $\|\Delta_{A,1}\| = O_P(1/\sqrt{N})$. To summarize, we have $\left\|\widehat{A}_{kl} - A_0\right\| = O_P(N^{-1/2} + \delta_{NT})$, which implies $\left\|\widehat{\bar{A}} - A_0\right\| = O_P(N^{-1/2} + r_{NT})$.

Proof of Claim B.2: Since K and L are finite, it suffices to show for any k, l,

$$\left\|\mathbb{E}_{kl}\left[\psi(W_{it};\theta_0,\widehat{\eta}_{kl}) - \psi(W_{it};\theta_0,\eta_0)\right]\right\| = O_P(r'_{NT}/\sqrt{N_k} + \lambda_{NT} + \lambda'_{NT}).$$

To simplify the notation, we denote

$$\begin{aligned}
\ddot{\psi}_{it}^{kl} &= \psi(W_{it};\theta_0,\widehat{\eta}_{kl}) - \psi(W_{it};\theta_0,\eta_0),\\
\tilde{\ddot{\psi}}_{it}^{kl} &= \ddot{\psi}_{it}^{kl} - \mathrm{E}_P[\ddot{\psi}_{it}^{kl}|W(-k,-l)],\\
b(1) &= \left\|\frac{\sqrt{N_k}}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left[\ddot{\psi}_{it}^{kl} - \mathrm{E}_P[\ddot{\psi}_{it}^{kl}|W(-k,-l)]\right]\right\|\\
b(2) &= \left\|\frac{1}{N_k T_l}\mathrm{E}_P\left[\psi(W_{it};\theta_0,\widehat{\eta}_{kl})|W(-k,-l)\right] - \mathrm{E}_P\left[\psi(W_{it};\theta_0,\eta_0)\right]\right\|.
\end{aligned}$$

We also denote $\tilde{\ddot{\psi}}_{d,it}$ as each element in the vector $\tilde{\ddot{\psi}}_{it}^{kl}$ for $d = 1,...,D$, while suppressing the subscripts k, l for convenience. By triangle inequality, we have

$$\left\|\mathbb{E}_{kl}\left[\psi(W_{it};\theta_0,\widehat{\eta}_{kl}) - \psi(W_{it};\theta_0,\eta_0)\right]\right\| \le b(1)/\sqrt{N_k} + b(2).$$

To bound $b(1)$, first note that it is mean zero by the iterated expectation argument. On the event $\mathcal{E}_\eta \cap \mathcal{E}_{cp}$, we have

$$\begin{aligned}
&\mathrm{E}_P[b(1)^2|W(-k,-l)] \le \frac{1}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{is}^{kl}\rangle|W(-k,-l)\right]\right|\\
&+\sum_{t\in S_l, i\in I_k, j\in I_k}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{jt}^{kl}\rangle|W(-k,-l)\right]\right| + \sum_{t\in S_l, i\in I_k}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{it}^{kl}\rangle|W(-k,-l)\right]\right|\\
&+2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{j,t+m}^{kl}\rangle|W(-k,-l)\right]\right| + 2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i\in I_k}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{i,t+m}^{kl}\rangle|W(-k,-l)\right]\right|\\
&=:\ c(1) + c(2) + c(3) + 2c(4) + 2c(5).
\end{aligned}$$

By conditional Cauchy-Schwarz inequality, for any i, t, j, s, we have

$$\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl},\tilde{\ddot{\psi}}_{js}^{kl}\rangle|W(-k,-l)\right]\right| \le \left(\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{js}^{kl}\|^2|W(-k,-l)\right]\right)^{1/2}.$$

Applying Minkowski's inequality, Jensen's inequality on the event $\mathcal{E}_\eta \cap \mathcal{E}_{cp}$, we have, for $i \in I_k, t \in S_l$,

$$
\begin{aligned}
&\left(\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]\right)^{1/2}\\
\leq&\left(\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{kl}\|^2|W(-k,-l)\right]\right)^{1/2}+\left(\mathrm{E}_P\left[\|\mathrm{E}_P[\ddot{\psi}_{it}^{kl}|W(-k,-l)]\|^2|W(-k,-l)\right]\right)^{1/2}\\
\leq&2\left(\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{kl}\|^2|W(-k,-l)\right]\right)^{1/2}\leq 2r'_{NT}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&c(1)\leq \mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]=O({r'_{NT}}^2), && c(2)\leq c\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]=O({r'_{NT}}^2),\\
&c(3)\leq \frac{1}{N_k}\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]=O({r'_{NT}}^2/N), && c(5)\leq \mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]]=O({r'_{NT}}^2).
\end{aligned}
$$

Following similar arguments as for bounding $a(4)$, $c(4)$ is of order $O({r'_{NT}}^2)$. So, we have shown

$$
\mathrm{E}_P[b(1)^2|W(-k,-l)]=O_P\left({r'_{NT}}^2\right),
$$

which implies $b(1)=O_P(r'_{NT})$ by Markov inequality and Lemma 6.1 of Chernozhukov et al. (2018a).

To bound $b(2)$, we first define

$$
f_{kl}(r):=\mathrm{E}_P\left[\psi(W_{it},\theta_0,\eta_0+r(\widehat{\eta}_{kl}-\eta_0)|W(-k,-l)\right]-\mathrm{E}_P\left[\psi(W_{it};\theta_0,\eta_0)\right],\ r\in[0,1],
$$

for some $i\in I_k, t\in S_l$. So, $b(2)=\|f_{kl}(1)\|$. By expanding $f_{kl}(r)$ around 0 using mean value theorem and evaluating at $r=1$, we have

$$
f_{kl}(r)=f_{kl}(0)+f'_{kl}(0)+f''_{kl}(\tilde{r})/2,
$$

where $\tilde{r}\in(0,1)$. We note that $f_{kl}(0)=0$ on the event $\mathcal{E}_{cp}$. On the event $\mathcal{E}_\eta\cap\mathcal{E}cp$ and under Assumption DML1(ii)(near-orthogonality), we have $\left\|f'_{kl}(0)\right\|\leq\lambda_{NT}$ and $\left\|f''_{kl(0)}\right\|\leq\lambda'_{NT}$. Therefore, we have shown that $b(2)=O_P(\lambda_{NT})+O_P(\lambda'_{NT})$. Combining the bounds for $b(1)$ and $b(2)$ completes the proof of Claim B.2. □

Proof of Theorem 3.2. By the same arguments for Theorem 3.1, we have $\mathrm{P}(\mathcal{E}_\eta\cap\mathcal{E}_{cp})=1-\mathrm{P}(\mathcal{E}_\eta^c\cup\mathcal{E}_{cp}^c)\geq 1-o(1)$. By Claim B.1, we have $\|\hat{\widehat{A}}-A_0\|=O_P(N^{-1/2}+r_{NT})$ on event $\{\mathcal{E}_\eta\cap\mathcal{E}_{cp}\}$. Therefore, due to $\|A_0^{-1}\|\leq a_0^{-1}$ ensured by Assumption DML1(iv) and $\Omega<\infty$ as shown in Claim B.2, it suffices to show $\|\widehat{\Omega}_{\mathrm{CHS}}-\Omega\|=o_P(1)$. Furthermore, since K, L are fixed constants, it suffices to show for each (k,l) that

$\|\widehat{\Omega}_{\text{CHS},kl} - \Omega\| = o_P(1)$ where

$$\begin{aligned}
\widehat{\Omega}_{\text{CHS},kl} :=& \widehat{\Omega}_{a,kl} + \widehat{\Omega}_{b,kl} - \widehat{\Omega}_{c,kl} + \widehat{\Omega}_{d,kl} + \widehat{\Omega}'_{d,kl},\\
\widehat{\Omega}_{a,kl} :=& \frac{1}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} \psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})',\\
\widehat{\Omega}_{b,kl} :=& \frac{K/L}{N_k T_l^2} \sum_{t\in S_l, i\in I_k, j\in I_k} \psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{jt};\widehat{\theta},\widehat{\eta}_{kl})',\\
\widehat{\Omega}_{c,kl} :=& \frac{K/L}{N_k T_l^2} \sum_{i\in I_k, t\in S_l} \psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})',\\
\widehat{\Omega}_{d,kl} :=& \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j\in I_k, j\neq i} \psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{j,t+m};\widehat{\theta},\widehat{\eta}_{kl})'.
\end{aligned}$$

Since a sequence of symmetric matrices Ω_n converges to a symmetric matrix Ω_0 if and only if $e'\Omega_n e \to e'\Omega_0 e$ for all comfortable e, it suffices to assume without loss of generality that the dimension of ψ to be 1. To simplify the expression, we denote

$$\psi_{it}^{(0)} = \psi(W_{it};\theta_0,\eta_0), \quad \widehat{\psi}_{it}^{(kl)} = \psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})$$

Claim B.3. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{a,kl} - \Sigma_a\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$.
Claim B.4. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{b,kl} - c\mathrm{E}_P[g_t g_t]\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$.
Claim B.5. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{c,kl}\right| = O_P\left(T^{-1}\right)$.
Claim B.6. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{d,kl} - c\sum_{m=1}^{\infty} \mathrm{E}_P[g_t g_{t+m}]\right| = o_P(1)$.

The decomposition techniques used in the proofs of Claims A.4, A.5, and A.7 follow the proofs of Lemma 1 and Lemma 2 in Appendix E of Chiang et al. (2024). Combining the Claims A.4-A.7 completes the proof of Theorem 3.2.

Proof of Claim B.3. By triangle inequality, we have

$$\begin{aligned}
\left|\widehat{\Omega}_{a,kl} - \Sigma_a\right| \leq& \left|I_{a,1}^{(kl)}\right| + \left|I_{a,2}^{(kl)}\right| + \left|I_{a,2}^{(kl)}\right|,\\
I_{a,1}^{(kl)} :=& \frac{1}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} \left\{\widehat{\psi}_{it}^{(kl)}\widehat{\psi}_{ir}^{(kl)} - \psi_{it}^{(0)}\psi_{ir}^{(0)}\right\},\\
I_{a,2}^{(kl)} :=& \frac{1}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} \left\{\psi_{it}^{(0)}\psi_{ir}^{(0)} - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}]\right\},\\
I_{a,2}^{(kl)} :=& \frac{1}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}] - \mathrm{E}_P[a_i a_i].
\end{aligned}$$

By law of total covariance and mean-zero property of $\psi_{it}^{(0)}$, we have

$$\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{ir}^{(0)}\right] = \mathrm{E}_P[\mathrm{E}_P(\psi_{it}^{(0)}, \psi_{ir}^{(0)}|\alpha_i)] + \mathrm{E}_P\left(\mathrm{E}_P[\psi_{it}^{(0)}|\alpha_i]\mathrm{E}_P[\psi_{ir}^{(0)}|\alpha_i]\right)$$

Due to the identical distribution of α_i and mean zero, we have

$$\frac{1}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}] = \frac{1}{T_l^2}\sum_{t\in S_l, r\in S_l}\left\{\mathrm{E}_P[\mathrm{E}_P(\psi_{it}^{(0)}\psi_{ir}^{(0)}|\alpha_i)] + \mathrm{E}_P(\mathrm{E}_P[\psi_{it}^{(0)}|\alpha_i]\mathrm{E}_P[\psi_{ir}^{(0)}|\alpha_i])\right\}$$

Conditional on α_i, $\{\psi_{it}^{(0)}\}_{t\geq 1}$ is β-mixing with the mixing coefficient same as γ_t. Therefore, we can apply Theorem 14.13(ii) in Hansen (2022) and Jensen's inequality:

$$\mathrm{E}_P\left|\mathrm{E}_P\left[\psi_{it}^{(0)}, \psi_{ir}^{(0)}|\alpha_i\right]\right| \leq 8\left(\mathrm{E}_P|\psi_{it}^{(0)}|^q\right)^{2/q}\beta_\gamma(|t-r|)^{1-2/q}$$

Note that $\sum_{t\in S_l, r\in S_l}\beta_\gamma(|t-r|)^{1-2/q} \leq \infty$ under Assumption 2. So, $I_{a,2}^{(kl)} = O(1/T_l^2) = O(T^{-2})$.

To bound $I_{a,2}^{(kl)}$, we can rewrite it by triangle inequality as follows:

$$\begin{aligned}
\left|I_{a,2}^{(kl)}\right| &\leq \left|\frac{1}{N_k}\sum_{i\in I_k} I_{a,2,i}^{(kl)}\right| + \left|\frac{1}{N_k}\sum_{i\in I_k}\tilde{I}_{a,2,i}^{(kl)}\right|,\\
I_{a,2,i}^{(kl)} &:= \frac{1}{T_l^2}\sum_{t,r\in S_l}\left\{\psi_{it}^{(0)}\psi_{ir}^{(0)} - \mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\{\gamma_t\}_{t\in S_l}\right]\right\},\\
\tilde{I}_{a,2,i}^{(kl)} &:= \frac{1}{T_l^2}\sum_{t,r\in S_l}\left\{\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\{\gamma_t\}_{t\in S_l}\right] - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}]\right\}.
\end{aligned}$$

Due to identical distribution of α_i, $\tilde{I}_{a,2,i}^{(kl)}$ does not vary over i so that $\mathrm{E}_P\left|\frac{1}{N_k}\sum_{i\in I_k}\tilde{I}_{a,2,i}^{(kl)}\right|^2 = \mathrm{E}_P\left|\tilde{I}_{a,2,i}^{(kl)}\right|^2$. Denote $h_i(\gamma_t, \gamma_r) = \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\gamma_t, \gamma_r] - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}]$. By direct calculation, we have

$$\mathrm{E}_P\left|\tilde{I}_{a,2,i}^{(kl)}\right|^2 = \frac{1}{T_l^4}\sum_{t,r,t',r'\in S_l}\mathrm{E}_P\left[h_i(\gamma_t,\gamma_r)h_i(\gamma_{t'},\gamma_{r'})\right].$$

To bound the RHS above, we can apply Lemma 3.4 in Dehling and Wendler (2010) by verifying the following two conditions:

$$\mathrm{E}_P\left|h_i(\gamma_t,\gamma_r)\right|^{2+\delta} < \infty, \tag{B.1}$$

$$\int\int\left|h_i(u,v)\right|^{2+\delta}dF(u)dF(v) < \infty, \tag{B.2}$$

for some $\delta > 0$ and $F(.)$ is the common CDF of γ_t. Consider condition B.1. By Minkowski's inequality, Jensen's inequality, and the law of iterated expectation, we have

$$\left(\mathrm{E}_P\left|h_i(\gamma_t,\gamma_r)\right|^{2+\delta}\right)^{\frac{1}{2+\delta}} \leq \left(\mathrm{E}_P\left|\psi_{it}^{(0)}\psi_{ir}^{(0)}\right|^{2+\delta}\right)^{\frac{1}{2+\delta}} + \mathrm{E}_P\left|\psi_{it}^{(0)}\psi_{ir}^{(0)}\right| \leq \left(\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^{4+2\delta}\right)^{\frac{1}{2+\delta}} + \mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2$$

where the second inequality follows from Hölder's inequality and the identical distribution of γ_t. Let $\delta = \frac{p-4}{2}$, then $\left(\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^{4+2\delta}\right)^{\frac{1}{2+\delta}} < c_m$ and $\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2 \leq c_m^2$ follows from Assumption DML2(i). Therefore, condition B.1 is satisfied.

Consider condition B.2. By Minkowski's inequality and Jensen's inequality, we have

$$\begin{aligned}
&\left(\int\int\left|\mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\gamma_t=u,\gamma_r=v]-\mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}]\right|^{2+\delta}dF(u)dF(v)\right)^{\frac{1}{2+\delta}}\\
\leq&\left(\int\int\left|\mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\gamma_t=u,\gamma_r=v]\right|^{2+\delta}dF(u)dF(v)\right)^{\frac{1}{2+\delta}}+\mathrm{E}_P\left|\psi_{it}^{(0)}\psi_{ir}^{(0)}\right|\\
\leq&\left(\int\int\left(\mathrm{E}_P\left[\left|\psi_{it}^{(0)}\right|^2|\gamma_t=u\right]\right)^{\frac{2+\delta}{2}}\left(\mathrm{E}_P\left[\left|\psi_{ir}^{(0)}\right|^2|\gamma_r=v\right]\right)^{\frac{2+\delta}{2}}dF(u)dF(v)\right)^{\frac{1}{2+\delta}}+\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2\\
\leq&\left(\int\int\mathrm{E}_P\left[\left|\psi_{it}^{(0)}\right|^{2+\delta}|\gamma_t=u\right]\mathrm{E}_P\left[\left|\psi_{ir}^{(0)}\right|^{2+\delta}|\gamma_r=v\right]dF(u)dF(v)\right)^{\frac{1}{2+\delta}}+\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2\\
=&\left(\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^{4+2\delta}\right)^{\frac{1}{2+\delta}}+\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2
\end{aligned}$$

where the second inequality follows from (conditional) Hölder's inequality and identical distribution of γ_t; the third inequality follows from Jensen's inequality; the last equality follows from the law of iterated expectation and the identical distribution of γ_t. Therefore, condition B.2 is also satisfied with $\delta = \frac{p-4}{2}$. By Lemma 3.4 in Dehling and Wendler (2010), we conclude

$$\mathrm{E}_P\left|\tilde{I}_{a,2,i}^{(kl)}\right|^2 = \frac{1}{T_l^4}\sum_{t,r,t',r'\in S_l}\mathrm{E}_P\left[h_i(\gamma_t,\gamma_r)h_i(\gamma_{t'},\gamma_{r'})\right] = o(T_l^{-1}) = o(T^{-1}).$$

Therefore, by Markov inequality, we have $\tilde{I}_{a,2,i}^{(kl)} = o_P(T^{-1/2})$. Next. consider $\left|\frac{1}{N_k}\sum_{i\in I_k}I_{a,2,i}^{(kl)}\right|$. Note that conditional on $\{\gamma_t\}_{t\in S_l}$, $I_{a,2,i}^{(kl)}$ is i.i.d over i. So, we have

$$\mathrm{E}_P\left[\left|\frac{1}{N_k}\sum_{i\in I_k}I_{a,2,i}^{(kl)}\right|^2|\{\gamma_t\}_{t\in S_l}\right] = \frac{1}{N_k^2}\sum_{i\in I_k}\mathrm{E}_P\left[\left|I_{a,2,i}^{(kl)}\right|^2|\{\gamma_t\}_{t\in S_l}\right] = \frac{1}{N_k}\mathrm{E}_P\left[\left|I_{a,2,i}^{(kl)}\right|^2|\{\gamma_t\}_{t\in S_l}\right]$$

By conditional Markov inequality, we have

$$\mathrm{P}\left(\left|\frac{1}{N_k}\sum_{i\in I_k} I_{a,2,i}^{(kl)}\right| > \varepsilon|\{\gamma_t\}_{t\in S_l}\right) = O\left(\frac{1}{N_k}\mathrm{E}_P\left[\left|I_{a,2,i}^{(kl)}\right|^2 |\{\gamma_t\}_{t\in S_l}\right]\right)$$

By Minkowski's inequality for infinite sums, Jensen's inequality, and Hölder's inequality, we have

$$\left(\mathrm{E}_P\left[\left|I_{a,2,i}^{(kl)}\right|^2\right]\right)^{1/2} \lesssim \frac{1}{T_l^2}\sum_{t,r\in S_l}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{ir}^{(0)}\right]^2\right)^{1/2} \leq \frac{1}{T_l^2}\sum_{t,r\in S_l}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\right]^4\right)^{1/2} \leq c_m^2,$$

where the last inequality follows from Assumption DML2(i) . Then, by law of iterated expectation, we have

$$\mathrm{P}\left(\left|\frac{1}{N_k}\sum_{i\in I_k} I_{a,2,i}^{(kl)}\right| > \varepsilon\right) = O\left(N_k^{-1}\right),$$

and $\left|\frac{1}{N_k}\sum_{i\in I_k} I_{a,2,i}^{(kl)}\right| = O_P\left(N_k^{-1/2}\right) = O_P\left(N^{-1/2}\right)$. Therefore, we have shown $\left|I_{a,2}^{kl}\right| = O_P\left(N^{-1/2}\right) + o_P\left(T^{-1/2}\right)$.

Next, consider $I_{a,1}^{kl}$. By product decomposition, triangle inequality, and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left|I_{a,1}^{kl}\right| \leq & \frac{1}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\left|\widehat{\psi}_{it}^{(kl)}\widehat{\psi}_{ir}^{(kl)\prime} - \psi_{it}^{(0)}\psi_{ir}^{(0)}\right| \\
\leq & \frac{1}{N_k T_l^2}\sum_{i\in I_k, t\in S_l, r\in S_l}\left\{\left|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right|\left|\widehat{\psi}_{ir}^{(kl)} - \psi_{ir}^{(0)}\right| + \left|\psi_{it}^{(0)}\right|\left|\widehat{\psi}_{ir}^{(kl)} - \psi_{ir}^{(0)}\right| + \left|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right|\left|\widehat{\psi}_{ir}^{(kl)\prime}\right|\right\} \\
\lesssim & R_{kl}\left\{\left\|\psi_{it}^{(0)}\right\|_{kl,2} + R_{kl}\right\},
\end{aligned}$$

where $R_{kl} = \left\|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right\|_{kl,2}$. By Markov inequality and under Assumption DML2(i), we have

$$\mathrm{E}_P\left[\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi_{it}^{(0)}\right)^2\right] = \mathrm{E}_P\left|\psi(W_{it};\theta_0,\eta_0)\right|^2 \leq c_m^2.$$

Therefore, $\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi_{it}^{(0)}\right)^2 = O_P(1)$. To bound R_{kl}, note that by Assumption DML1(i) (linearity) we have

$$\begin{aligned}
R_{kl}^2 = & \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi^a(W_{it};\widehat{\eta}_{kl})(\widehat{\theta} - \theta_0) + \psi(W_{it};\theta_0,\widehat{\eta}_{kl}) - \psi(W_{it};\theta_0,\eta_0)\right)^2 \\
\lesssim & \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left|\psi^a(W_{it};\widehat{\eta}_{kl})\right|^2\left|\widehat{\theta} - \theta_0\right|^2 + \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right|^2
\end{aligned}$$

By Markov inequality and Assumption DML2(i), we have $\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left|\psi^a(W_{it};\hat{\eta}_{kl})\right|^2 = O_P(1)$. By Theorem 3.1, $\left|\hat{\theta}-\theta_0\right|^2 = O_p(N^{-1})$. Therefore, the first term on RHS is $O_P(N^{-1})$. For the second term on RHS, consider its conditional expectation given the auxiliary sample $W(-k,-l)$. On the event $\mathcal{E}_\eta \cap \mathcal{E}_{cp}$, we have

$$\mathrm{E}_P\left[\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left|\hat{\psi}_{it}^{(kl)}-\psi_{it}^{(0)}\right|^2 |W(-k,-l)\right] = \mathrm{E}_P\left[\left|\psi(W_{it};\theta_0,\hat{\eta}_{kl})-\psi(W_{it};\theta_0,\eta_0)\right|^2 |W(-k,-l)\right] \le \delta_{NT}^2,$$

where the last inequality follows from Assumption DML2(ii). Then, by Markov inequality and Lemma 6.1 from Chernozhukov et al. (2018a), we have $R_{kl}^2 = O_P\left(N^{-1}+(r'_{NT})^2\right)$ and so $\left|I_{a,1}^{kl}\right| = O_P\left(N^{-1/2}+r'_{NT}\right)$. To summarize, we have shown

$$\left|\hat{\Omega}_{a,kl}-\Sigma_a\right| = O_P\left(N^{-1/2}+r'_{NT}\right)+O_P(N^{-1/2})+o_P(T^{-1/2})+O(T^{-2}) = O_P\left(N^{-1/2}+r'_{NT}\right)$$

Proof of Claim B.4. By triangle inequality, we have

$$\begin{aligned}
\left|\hat{\Omega}_{b,kl}-c\mathrm{E}_P[g_t g_t']\right| &\le \left|I_{b,1}^{(kl)}\right|+\left|I_{b,2}^{(kl)}\right|+\left|I_{b,3}^{(kl)}\right|,\\
I_{b,1}^{(kl)} &:= \frac{K/L}{N_k T_l^2}\sum_{t\in S_l, i\in I_k, j\in I_k}\left\{\hat{\psi}_{it}^{(kl)}\hat{\psi}_{jt}^{(kl)}-\psi_{it}^{(0)}\psi_{jt}^{(0)}\right\},\\
I_{b,2}^{(kl)} &:= \frac{K/L}{N_k T_l^2}\sum_{t\in S_l, i\in I_k, j\in I_k}\left\{\psi_{it}^{(0)}\psi_{jt}^{(0)}-\mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}]\right\},\\
I_{b,3}^{(kl)} &:= \frac{K/L}{N_k T_l^2}\sum_{t\in S_l, i\in I_k, j\in I_k}\mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}]-c\mathrm{E}_P[g_t g_t'],
\end{aligned}$$

and $\frac{K/L}{N_k T_l^2} = \frac{c}{N_k^2 T_l}$.

Consider $I_{b,3}^{(kl)}$. By the the law of total covariance, we have

$$\mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}] = cov(\psi_{it}^{(0)},\psi_{jt}^{(0)}) = \mathrm{E}_P[cov(\psi_{it}^{(0)},\psi_{jt}^{(0)}|\gamma_t)]+cov(\mathrm{E}_P[\psi_{it}^{(0)}|\gamma_t],\mathrm{E}_P[\psi_{jt}^{(0)}|\gamma_t]) = 0+\mathrm{E}_P[g_t g_t'],$$

Due to identical distribution of γ_t, $\mathrm{E}_P[g_t g_t']$ does not vary over t and so $I_{b,3}^{(kl)} = 0$.

To bound $I_{b,2}^{(kl)}$, we can rewrite it by triangle inequality as follows

$$\frac{1}{c}\left|I_{b,2}^{kl}\right| \leq \left|\frac{1}{T_l}\sum_{t\in S_l} I_{b,2,t}^{(kl)}\right| + \left|\frac{1}{T_l}\sum_{t\in S_l}\tilde{I}_{b,2,t}^{(kl)}\right|,$$
$$I_{b,2,t}^{(kl)} := \frac{1}{N_k^2}\sum_{i,j\in I_k}\left\{\psi_{it}^{(0)}\psi_{jt}^{(0)} - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}|\{\alpha_i\}_{i\in I_k}]\right\}$$
$$\tilde{I}_{b,2,t}^{(kl)} := \frac{1}{N_k^2}\sum_{i,j\in I_k}\left\{\mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}|\{\alpha_i\}_{i\in I_k}] - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{jt}^{(0)}]\right\}$$

Due to identical distribution of γ_t, $\tilde{I}_{b,2,t}^{(kl)}$ does not vary over t so that $\mathrm{E}_P\left|\frac{1}{T_l}\sum_{t\in S_l}\tilde{I}_{b,2,t}^{(kl)}\right|^2 = \mathrm{E}_P\left|\tilde{I}_{b,2,t}^{(kl)}\right|^2$. Denote $\zeta_{ij,t} = \psi_{it}^{(0)}\psi_{jt}^{(0)}$. By direct calculation, we have

$$\begin{aligned}\mathrm{E}_P\left|\tilde{I}_{b,2,t}^{(kl)}\right|^2 =& \frac{1}{N_k^4}\sum_{i,j\in I_k}\sum_{i',j'\in I_k}\mathrm{E}_P\left[\left(\mathrm{E}_P[\zeta_{ij,t}|\alpha_i,\alpha_j] - \mathrm{E}_P[\zeta_{ij,t}]\right)\left(\mathrm{E}_P[\zeta_{i'j'}|\alpha_{i'},\alpha_{j'}] - \mathrm{E}_P[\zeta_{i'j'}]\right)\right]\\ \lesssim& \frac{1}{N_k}\mathrm{E}_P[\zeta_{ij,t}]^2 < \frac{1}{N_k}\mathrm{E}_P\left[\psi_{it}^{(0)}\right]^4 = O(1/N_k).\end{aligned}$$

where the first inequality follows from the assumption that α_i is independent over i and an application of Hölder's inequality and Jensen's inequality. The second inequality follows from Hölder's inequality and the last equality follows from Assumption DML2(i) with some $q > 4$. Therefore, by Markov inequality, we have $\left|\frac{1}{T_l}\sum_{t\in S_l}\tilde{I}_{b,2,t}^{(kl)}\right| = O_P(N_k^{-1/2}) = O_P(N^{-1/2})$.

Now consider $\left|\frac{1}{T_l}\sum_{t\in S_l} I_{b,2,t}^{(kl)}\right|$. Note that conditional on $\{\alpha_i\}$, $I_{b,2,t}^{(kl)}$ is also β-mixing with the mixing coefficient same as γ_t. Then, with an application of the conditional version of Theorem 14.2 from Davidson (1994), we have

$$\left(\mathrm{E}_P\left[\left|\mathrm{E}_P[I_{b,2,t}^{(kl)}|\{\alpha_i\}_{i\in I_k},\mathcal{F}_{-\infty}^{t-l}]\right|^2|\{\alpha_i\}_{i\in I_k}\right]\right)^{1/2}] \leq 2(2^{1/2}+1)\beta(l)^{1/2-\frac{2}{q}}\left(\mathrm{E}_P\left[|I_{b,2,t}^{(kl)}|^{\frac{q}{2}}|\{\alpha_i\}_{i\in I_k}\right]\right)^{\frac{2}{q}}.$$

Then, we can apply the conditional version of Lemma A from Hansen (1992) to show that

$$\begin{aligned}\left(\mathrm{E}_P\left[\left|\frac{1}{T_l}\sum_{t\in S_l} I_{b,2,t}^{(kl)}\right|^2|\{\alpha_i\}_{i\in I_k}\right]\right)^{1/2} &\lesssim \frac{1}{T_l}\sum_{l=1}^{\infty}\beta(l)^{1/2-\frac{2}{q}}\left(\sum_{t\in S_l}\left(\mathrm{E}_P\left[|I_{b,2,t}^{(kl)}|^{\frac{q}{2}}|\{\alpha_i\}_{i\in I_k}\right]\right)^{\frac{4}{q}}\right)^{1/2}\\ &\lesssim \frac{1}{\sqrt{T_l}}\left(\mathrm{E}_P\left[|I_{b,2,t}^{(kl)}|^{\frac{q}{2}}|\{\alpha_i\}_{i\in I_k}\right]\right)^{\frac{2}{q}}\end{aligned}$$

By conditional Markov inequality, we have

$$\mathrm{P}\left(\left|\frac{1}{T_l}\sum_{t\in S_l} I_{b,2,t}^{(kl)}\right| > \varepsilon|\{\alpha_i\}_{i\in I_k}\right) = O\left(T_l^{-1}\mathrm{E}_P\left[\left|I_{b,2,t}^{(kl)}\right|^{\frac{q}{2}}|\{\alpha_i\}_{i\in I_k}\right]\right)$$

By Minkowski's inequality for infinite sums, Jensen's inequality, and Hölder's inequality, we have

$$\left(\mathrm{E}_P\left[\left|I_{b,2,t}^{(kl)}\right|^{\frac{q}{2}}\right]\right)^{\frac{2}{q}} \lesssim \frac{1}{N_k^2}\sum_{i,j\in I_k}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{jt}^{(0)}\right]^{\frac{q}{2}}\right)^{\frac{2}{q}} \le \frac{1}{N_k^2}\sum_{i,j\in I_k}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\right]^{q}\right)^{\frac{2}{q}} \le c_m^2,$$

where the last inequality follows from Assumption DML2(i) . Then, by the law of iterated expectation, we have

$$\mathrm{P}\left(\left|\frac{1}{T_l}\sum_{t\in S_l} I_{b,2,t}^{(kl)}\right| > \varepsilon\right) = O\left(T_l^{-1/2}\right).$$

Therefore, we have shown $\left|I_{b,2}^{kl}\right| = O_P(N_k^{-1}) + O_P(T_l^{-1/2}) = O_P(T^{-1/2})$.

Consider $I_{b,1}^{kl}$. By the similar inequality for $\left|I_{a,1}^{kl}\right|$, we have

$$\frac{1}{c}\left|I_{b,1}^{kl}\right| \lesssim R_{kl}\left\{\left(\frac{1}{N_kT_l}\sum_{i\in I_k,t\in S_l}\left(\psi_{it}^{(0)}\right)^2\right)^{1/2} + R_{kl}\right\},$$

where $R_{kl} = \left\|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right\|_{kl,2}$. We have shown in the proof of Claim B.3 that $\left\|\psi_{it}^{(0)}\right\|_{kl,2} = O_P(1)$ and $R_{kl}^2 = O_P\left(N^{-1} + (r'_{NT})^2\right)$. So $\left|I_{b,1}^{kl}\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$. To summarize

$$\left|\widehat{\Omega}_{b,kl} - c\mathrm{E}_P[g_tg_t']\right| = O_P\left(N^{-1/2}\right) + O_P\left(T^{-1/2}\right) + O_P\left(N^{-1/2} + r'_{NT}\right) = O_P\left(N^{-1/2} + r'_{NT}\right),$$

which completes the proof of Claim B.4.

Proof of Claim B.5. By triangle inequality, we have $\left|\widehat{\Omega}_{c,kl}\right| \le \left|I_{c,1}^{(kl)}\right| + \left|I_{c,2}^{(kl)}\right| + \left|I_{c,3}^{(kl)}\right|$ where

$$\begin{aligned}
I_{c,1}^{(kl)} &:= \frac{K/L}{N_kT_l^2}\sum_{i\in I_k,t\in S_l}\left\{\widehat{\psi}_{it}^{(kl)}\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\psi_{it}^{(0)}\right\},\\
I_{c,2}^{(kl)} &:= \frac{K/L}{N_kT_l^2}\sum_{i\in I_k,t\in S_l}\left\{\psi_{it}^{(0)}\psi_{it}^{(0)} - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{it}^{(0)}]\right\},\\
I_{c,3}^{(kl)} &:= \frac{K/L}{N_kT_l^2}\sum_{i\in I_k,t\in S_l}\mathrm{E}_P[\psi_{it}^{(0)}\psi_{it}^{(0)}],
\end{aligned}$$

Consider $I_{c,3}^{(kl)}$. Note that under Assumption DML2(i), we have

$$\mathrm{E}_P[\psi_{it}^{(0)}\psi_{it}^{(0)}] \leq c_m^2.$$

Thus, $I_{c,3}^{(kl)} = O_P(1/T_l) = O_P(T^{-1})$.

Consider $I_{c,2}^{kl}$. We denote $\xi_{it} = \psi_{it}^{(0)}\psi_{it}^{(0)} - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{it}^{(0)}]$. By expanding $\mathrm{E}\left|I_{c,2}^{kl}\right|^2$ and applying Hölder's inequality, we have

$$\begin{aligned}
\mathrm{E}\left|I_{c,2}^{kl}\right|^2 \leq \left(\frac{K/L}{N_k T_l^2}\right)^2 &\left\{ \sum_{i\in I_k, t\in S_l, r\in S_l} \mathrm{E}_P\left|\xi_{it}\right|^2 + \sum_{t\in S_l, i\in I_k, j\in I_k} \mathrm{E}_P\left|\xi_{it}\right|^2 + \sum_{t\in S_l, i\in I_k} \mathrm{E}_P\left|\xi_{it}\right|^2 \right. \\
&\left. +2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k} \mathrm{E}_P\left|\xi_{it}\right|^2 + 2\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i\in I_k} \mathrm{E}_P\left|\xi_{it}\right|^2 \right\}.
\end{aligned}$$

where the last inequality follows from Note that for each i, t, by Hölder's inequality and Assumption DML2(i), we have

$$\mathrm{E}_P\left|\xi_{it}\right|^2 \lesssim \mathrm{E}_P\left[\psi(W_{it};\theta_0,\eta_0)^4\right] \leq c_m^4.$$

Thus, $\mathrm{E}\left|I_{c,2}^{(kl)}\right|^2 = O(T^{-2})$ and so $I_{c,2}^{(kl)} = O_P(T^{-1})$.

Now consider $I_{c,1}^{(kl)}$. Following the same steps for $I_{b,1}^{(kl)}$, we have

$$\left|I_{c,1}^{(kl)}\right| \lesssim \frac{K/L}{T_l} R_{kl}\left\{\left\|\psi_{it}^{(0)}\right\|_{kl,2} + R_{kl}\right\},$$

where $R_{kl} = \left\|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right\|_{kl,2}$. We have shown in the proof of Claim B.3 that $\left\|\psi_{it}^{(0)}\right\|_{kl,2} = O_P(1)$ and $R_{kl}^2 = O_P\left(N^{-1} + (r'_{NT})^2\right)$. So, $\left|I_{c,1}^{(kl)}\right| = O_P\left(N^{-1/2}/T + r'_{NT}/T\right)$. To summarize

$$\left|\widehat{\Omega}_{c,kl}\right| = O_P\left(T^{-1}\right) + O_P\left(N^{-1/2}/T + r'_{NT}/T\right) = O_P\left(T^{-1}\right),$$

which completes the proof of Claim B.5.

Proof of Claim B.6. By triangle inequality, we have

$$\left|\widehat{\Omega}_{d,kl} - c\sum_{m=1}^{\infty}\mathrm{E}_P[g_t g_t']\right| \leq \left|I_{d,1}^{(kl)}\right| + \left|I_{d,2}^{(kl)}\right| + \left|I_{d,3}^{(kl)}\right| + \left|I_{d,4}^{(kl)}\right| + \left|I_{d,5}^{(kl)}\right| + \left|I_{d,6}^{(kl)}\right|$$

where

$$
\begin{aligned}
I_{d,1}^{(kl)} &:= \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i \in I_k, j \in I_k, j \neq i} \left\{ \widehat{\psi}_{it}^{(kl)} \widehat{\psi}_{j,t+m}^{(kl)} - \psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right\}, \\
I_{d,2}^{(kl)} &:= \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i \in I_k, j \in I_k, j \neq i} \left\{ \psi_{it}^{(0)} \psi_{j,t+m}^{(0)} - \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right] \right\}, \\
I_{d,3}^{(kl)} &:= \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} \left(k\left(\frac{m}{M}\right) - 1 \right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i \in I_k, j \in I_k, j \neq i} \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right], \\
I_{d,4}^{(kl)} &:= \frac{K/L}{N_k T_l^2} \sum_{m=M}^{\infty} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i \in I_k, j \in I_k, j \neq i} \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right], \\
I_{d,5}^{(kl)} &:= \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i \in I_k, j \in I_k, j \neq i} \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right] - c \sum_{m=1}^{\infty} \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right], \\
I_{d,6}^{(kl)} &:= c \sum_{m=1}^{\infty} \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right] - c \sum_{m=1}^{\infty} \mathrm{E}_P \left[g_t g_{t+m} \right]
\end{aligned}
$$

and $\frac{K/L}{N_k T_l^2} = \frac{c}{N_k^2 T_l}$.

Consider $I_{d,6}^{(kl)}$. By the law of total covariance, we have

$$
\begin{aligned}
\mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right] &= cov(\psi_{it}^{(0)}, \psi_{j,t+m}^{(0)}) \\
&= \mathrm{E}_P[cov(\psi_{it}^{(0)}, \psi_{j,t+m}^{(0)}) | \gamma_t, \gamma_{t+m}] + cov(\mathrm{E}_P[\psi_{it}^{(0)} | \gamma_t], \mathrm{E}_P[\psi_{j,t+m}^{(0)} | \gamma_{t+m}]) \\
&= 0 + \mathrm{E}_P[g_t g_{t+m}'],
\end{aligned}
$$

where the last equality follows from the properties of Hajek projection components, as discussed in the beginning of Appendix A. Therefore, $I_{d,6}^{(kl)} = 0$.

Consider $I_{d,5}^{(kl)}$. The strict stationarity of γ_t implies that $\psi_{it}^{(0)}$ is also strictly stationary over t. And under Assumption 2, there is no heterogeneity across i. Then, as $M, T \to \infty$, we have $I_{d,5}^{(kl)} = o(1)$.

Consider $I_{d,4}^{(kl)}$. Under Assumption DML2(i), $\left(\mathrm{E}_P |\psi_{it}^{(0)}|^q \right)^{1/q} \leq c_m$ for $q > 4$. And conditional on α_i, $\psi_{it}^{(0)}$ is β-mixing with the mixing coefficient not larger than that of γ_t. Then by Theorem 14.13(ii) in Hansen (2022), we have

$$
\left| \mathrm{E}_P \left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} | \{\alpha_i\}_{i \in I_k} \right] \right| \leq 8 \left(\mathrm{E}_P \left[|\psi_{it}^{(0)}|^q | \alpha_i \right] \right)^{1/q} \left(\mathrm{E}_P \left[|\psi_{j,t+m}^{(0)}|^q | \alpha_j \right] \right)^{1/q} \alpha_\gamma(m)^{1-2/q}
$$

By iterated expectation and Jensen's inequality, we have

$$\begin{aligned}\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \le& \mathrm{E}_P\left[\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}|\{\alpha_i\}_{i\in I_k}\right]\right|\right] \\ \le& 8\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{it}^{(0)}|^q|\alpha_i\right]\right)^{1/q}\left(\mathrm{E}_P\left[|\psi_{j,t+m}^{(0)}|^q|\alpha_j\right]\right)^{1/q}\alpha_\gamma(m)^{1-2/q}\right] \\ \le& 8\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{it}^{(0)}|^q|\alpha_i\right]\right)^{1/q}\right]\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{j,t+m}^{(0)}|^q|\alpha_j\right]\right)^{1/q}\right]\alpha_\gamma(m)^{1-2/q} \\ \lesssim& c_m^2\alpha_\gamma(m)^{1-2/q}\end{aligned}$$

where the third inequality follows from that α_i are independent over i. Then, as $M\to\infty$,

$$\begin{aligned}\left|I_{d,4}^{(kl)}\right| \le& \frac{K/L}{N_kT_l^2}\sum_{m=M}^{\infty}\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\sum_{i\in I_k,j\in I_k,j\neq i}\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \lesssim \sum_{m=M}^{\infty}\alpha_\gamma(m)^{1-2/q} \le \sum_{m=M}^{\infty}\beta_\gamma(m)^{1-2/q} \\ \le& c_\kappa\sum_{m=M}^{\infty}exp(-\kappa m) = c_\kappa\left(\frac{1}{1-e^{-k}}-\frac{1-e^{-kM}}{1-e^{-\kappa}}\right) = O\left(e^{-\kappa M}\right).\end{aligned}$$

Consider $I_{d,3}^{(kl)}$.

$$\begin{aligned}\left|I_{d,3}^{(kl)}\right| \le& \frac{K/L}{N_kT_l^2}\sum_{m=1}^{M-1}\left|k\left(\frac{m}{M}\right)-1\right|\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\sum_{i\in I_k,j\in I_k,j\neq i}\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \\ \le& cc_m^2\sum_{m=1}^{M-1}\left|k\left(\frac{m}{M}\right)-1\right|\alpha_\gamma(m)^{1-2/q}.\end{aligned}$$

Note that for each m, $\left|k\left(\frac{m}{M}\right)-1\right|\to 0$ as $M\to\infty$. Since $\left|k\left(\frac{m}{M}\right)-1\right|\alpha_\gamma(m)^{1-2/q}\le 1$, we can apply dominated convergence theorem to conclude that $I_{d,3}^{(kl)}=o(1)$.

To bound $I_{d,2}^{(kl)}$, we can rewrite it by triangle inequality as follows

$$\frac{1}{c}\left|I_{d,2}^{(kl)}\right| \le \left|\sum_{m=1}^{M-1}\frac{k\left(\frac{m}{M}\right)}{T_l}\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}I_{d,2,tm}^{(kl)}\right| + \left|\sum_{m=1}^{M-1}\frac{k\left(\frac{m}{M}\right)}{T_l}\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\tilde{I}_{d,2,tm}^{(kl)}\right|,$$

where

$$\begin{aligned}I_{d,2,tm}^{(kl)} :=& \frac{1}{N_k^2}\sum_{i,j\in I_k,i\neq j}\left\{\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}-\mathrm{E}_P[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}|\{\alpha_i\}_{i\in I_k}]\right\} \\ \tilde{I}_{d,2,tm}^{(kl)} :=& \frac{1}{N_k^2}\sum_{i,j\in I_k,i\neq j}\left\{\mathrm{E}_P[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}|\{\alpha_i\}_{i\in I_k}]-\mathrm{E}_P[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}]\right\}\end{aligned}$$

Due to identical distribution of γ_t, $\tilde{I}^{(kl)}_{d,2,tm}$ does not vary over t so that $\mathrm{E}_P \left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \tilde{I}^{(kl)}_{d,2,tm} \right|^2 \leq \mathrm{E}_P \left| \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \tilde{I}^{(kl)}_{d,2,tm} \right|^2$. And by Minkowski's inequality, we have

$$\left(\mathrm{E}_P \left| \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \tilde{I}^{(kl)}_{d,2,tm} \right|^2 \right)^{1/2} \leq \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \left(\mathrm{E}_P \left[\tilde{I}^{(kl)}_{d,2,tm} \right]^2 \right)^{1/2}$$

Denote $\zeta_{ijm} = \psi^{(0)}_{it} \psi^{(0)}_{j,t+m}$. By direct calculation, we have

$$\begin{aligned}
\mathrm{E}_P \left| \tilde{I}^{(kl)}_{d,2,tm} \right|^2 =& \frac{1}{N_k^4} \sum_{i,j \in I_k} \sum_{i',j' \in I_k} \mathrm{E}_P \left[\left(\mathrm{E}_P[\zeta_{ijm} | \alpha_i, \alpha_j] - \mathrm{E}_P[\zeta_{ij,t}] \right) \left(\mathrm{E}_P[\zeta_{i'j'} | \alpha_{i'}, \alpha_{j'}] - \mathrm{E}_P[\zeta_{i'j'}] \right) \right] \\
\lesssim& \frac{1}{N_k} \mathrm{E}_P[\zeta_{ijm}]^2 < \frac{1}{N_k} \mathrm{E}_P \left[\psi^{(0)}_{it} \right]^4 = O(1/N_k).
\end{aligned}$$

where the first inequality follows from the assumption that α_i is independent over i and an application of Hölder's inequality and Jensen's inequality. The second inequality follows from Hölder's inequality and the last equality follows from Assumption DML2(i) with some $q > 4$. Therefore, we have

$$\left(\mathrm{E}_P \left| \right|^2 \right)^{1/2} \leq O_P\left(\frac{M}{N^{1/2}} \right) = O_P\left(\frac{M}{T^{1/2}} \right).$$

By Markov inequality, we have $\left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \tilde{I}^{(kl)}_{d,2,tm} \right| = O_P\left(\frac{M}{T^{1/2}} \right)$.

Now consider $\left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} I^{(kl)}_{d,2,tm} \right|$. By Minkowski's inequality, we have

$$\left(\mathrm{E}_P \left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} I^{(kl)}_{d,2,tm} \right|^2 \right)^{1/2} \leq \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \left(\mathrm{E}_P \left| \frac{1}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} I^{(kl)}_{d,2,tm} \right|^2 \right)^{1/2}$$

Following the same steps as for $I^{(kl)}_{b,2,tm}$, we can show

$$\mathrm{E}_P \left| \frac{1}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} I^{(kl)}_{d,2,tm} \right|^2 = O\left(T_l^{-1} \right).$$

Therefore, $\left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} I^{(kl)}_{d,2,tm} \right| = O_P\left(\frac{M}{T_l^{-1/2}} \right) = O_P\left(\frac{M}{T^{-1/2}} \right)$. We have shown $\left| I^{(kl)}_{b,2} \right| = O_P(1/N_k) + O_P\left(\frac{M}{T^{-1/2}} \right) = O_P\left(\frac{M}{T^{-1/2}} \right)$.

Consider $I_{d,1}^{(kl)}$. Denote

$$I_{d,1,m}^{(kl)}=\frac{K/L}{N_kT_l^2}\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\sum_{i\in I_k,j\in I_k,j\neq i}\left\{\widehat{\psi}_{it}^{(kl)}\widehat{\psi}_{j,t+m}^{(kl)}-\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right\},$$

for each m. Then, $I_{d,1}^{(kl)}=\sum_{m=1}^{M-1}k\left(\frac{m}{M}\right)I_{d,1,m}^{(kl)}$. Following the same steps as for $I_{a,1}^{(kl)}$, we can show

$$\left|I_{d,1,m}^{(kl)}\right|=O_P(T^{-1/2}+r'_{NT}),$$

for each m. Therefore, $\left|I_{d,1}^{(kl)}\right|=O_P\left(\frac{M}{T^{-1/2}}+Mr'_{NT}\right)$. Note that $Mr'_{NT}\leq M\delta_{NT}N^{-1/2}=\frac{M}{T^{1/2}}\frac{T^{1/2}}{N^{1/2}}\delta_{NT}=o(1)$.

To summarize

$$\begin{aligned}\left|\widehat{\Omega}_{d,kl}-c\sum_{m=1}^{\infty}\mathrm{E}_P[g_tg_t']\right|&=O_P\left(\frac{M}{T^{-1/2}}+Mr'_{NT}\right)+O_P\left(\frac{M}{T^{1/2}}\right)+o(1)+O(e^{-\kappa M})+o(1)+0\\&=o_P(1).\end{aligned}$$

which completes the proof of Claim B.6.

□

Proof of Theorem 3.3. Since (K,L) are fixed constants, it suffices to show for each (k,l) that $\widehat{\Omega}_{\mathrm{NW},kl}:=\frac{K/L}{N_kT_l^2}\sum_{i\in I_k,t\in S_l,r\in S_l}k\left(\frac{|t-r|}{M}\right)\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{ir};\widehat{\theta},\widehat{\eta}_{kl})'=o_P(1)$. Note that we can rewrite $\widehat{\Omega}_{\mathrm{NW},kl}$ as

$$\widehat{\Omega}_{\mathrm{NW},kl}=\widehat{\Omega}_{c,kl}+\widehat{\Omega}_{e,kl}-\widehat{\Omega}_{d,kl}$$

where $\widehat{\Omega}_{c,kl}$ and $\widehat{\Omega}_{d,kl}$ are defined in the beginning of the proof of Theorem 3.2, and $\widehat{\Omega}_{e,kl}$ is defined as follows:

$$\widehat{\Omega}_{e,kl}:=\frac{K/L}{N_kT_l^2}\sum_{m=1}^{M-1}k\left(\frac{m}{M}\right)\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\sum_{i\in I_k,j\in I_k}\psi(W_{it};\widehat{\theta},\widehat{\eta}_{kl})\psi(W_{j,t+m};\widehat{\theta},\widehat{\eta}_{kl})'.$$

Observe that by replacing $\widehat{\Omega}_{d,kl}$ by $\widehat{\Omega}_{e,kl}$, each step in the proof of Claim B.6 also follows. It implies that $\widehat{\Omega}_{e,kl}=\widehat{\Omega}_{d,kl}+o_P(1)$. By Lemma A.6, we have $\widehat{\Omega}_{c,kl}=O_P(T^{-1})$. Therefore, we conclude that $\widehat{\Omega}_{\mathrm{NW},kl}=o_P(1)$. □

Appendix C.

Proof of Theorem 4.1. Let $P \in \mathcal{P}_{NT}$ for each (N,T). We denote

$$A_{NT} = \frac{1}{NT}(V^Z)'V^D,\ \hat{A}_{NT} = \frac{1}{NT}(Z - f\widehat{\zeta}_0)'(D - f\widehat{\pi}_0),$$
$$\psi_{NT} = \frac{1}{NT}(V^Z)'V^g,\ \hat{\psi}_{NT} = \frac{1}{NT}(Z - f\hat{\zeta}_0)'\left(Y - f\hat{\beta} - (D - f\hat{\zeta})'\theta_0\right).$$

We can write $\widehat{\theta} - \theta_0 = \hat{A}_{NT}^{-1}\hat{\psi}_{NT}$. By product decomposition, we have

$$\widehat{\theta} - \theta_0 = A_{NT}^{-1}\psi_{NT} + A_{NT}^{-1}\left[\hat{\psi}_{NT} - \psi_{NT}\right] + \left[\hat{A}_{NT}^{-1} - A_{NT}^{-1}\right]\left[\hat{\psi}_{NT} - \psi_{NT}\right] + \left[\hat{A}_{NT}^{-1} - A_{NT}^{-1}\right]\psi_{NT}$$

For the asymptotic normality of $\sqrt{N \wedge T}\left(\widehat{\theta} - \theta_0\right)$, we need to show the following statements: (i) $A_{NT} \xrightarrow{p} A_0 = \mathrm{E}[V_{it}^Z V_{it}^D]$; (ii) $\sqrt{N \wedge T}\psi_{NT} \xrightarrow{d} \mathcal{N}(0, \Omega_0)$; (iii) $\sqrt{N \wedge T}\left[\hat{\psi}_{NT} - \psi_{NT}\right] = o(1)$; (iv) $\widehat{A}_{NT} - A_{NT} = o_P(1)$. With statements (i) - (iv) and the identification condition in Assumption REG-P(i) such that $\tilde{A}_0$ is non-singular, $\sqrt{N \wedge T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, A_0^{-1}\Omega_0 A_0^{-1'}\right)$. Then, the conclusion of the theorem follows.

Before we show Statement (i) - (iv), we note that Assumptions REG-P(ii) and AHK imply that $(\bar{F}_i, \bar{F}_t)$ are functions of only $(\alpha_i, \gamma_t, \varepsilon_{it})$, and so are f_{it} and V_{it}^l for $l = g, D, Y, Z$. Therefore, the results based on Hajek projection are still applicable. Also, due to Assumptions REG-P(ii), $\bar{F}_i$ is a function of only (c_i, ϵ_i) and $\bar{F}_t$ is a function of only (d_t, ϵ_t), so f_{it} is a function of $(X_{it}, c_i, \epsilon_i, d_t, \epsilon_t)$ which are mean independent of U_{it}^D. Therefore, $\mathrm{E}_P[f_{it}V_{it}^D] = \mathrm{E}_P\left[f_{it}[(L_{2,it} - E[L_{2,it}])\eta_{D,2} + U_{it}^D]\right] = 0$ given that f_{it} is uncorrelated with $L_{2,it}$ as discussed in the main text. Similarly, we have $\mathrm{E}_P[f_{it}V_{it}^D] = 0$.

Statement (i) follows from Lemma A.1 under Assumptions AHK, AR, and REG-P(iii). For Statement (ii), we first observe that $V_{it}^Z = Z_{it}(1 - \zeta_0)$ where $\zeta_0 = \left(E[f_{it}'f_{it}]\right)^{-1} E[f_{it}'Z_{it}]$. Due to the exogeneity condition $\mathrm{E}_P[Z_{it}U^g] = 0$ and the independence between $(\bar{F}_i, \bar{F}_t, Z_{it}, X_{it})$ and (ϵ_i, ϵ_t), we have $\mathrm{E}_P[V_{it}^Z V_{it}^g] = 0$. With the additional Assumption REG-P(iv), Statement (ii) follows from Lemma A.2.

Consider Statement (iii). By product decomposition and triangle inequality, we have

$$\begin{aligned}
NT|\hat{\psi}_{NT} - \psi_{NT}| \leq& |(f(\zeta_0 - \widehat{\zeta}))'(f(\beta_0 - \widehat{\beta}) + V^Y + r^Y - \theta_0(f(\pi_0 - \widehat{\pi}) + V^D + r^D))| \\
&+ |(Z - f\zeta_0)'(\theta_0(f(\widehat{\pi} - \pi_0)) - f(\beta_0 - \widehat{\beta}) + r^g)| \\
\lesssim& |(f(\zeta_0 - \widehat{\zeta}))'f(\beta_0 - \widehat{\beta})| + |(f(\zeta_0 - \widehat{\zeta}))'V^Y| + |(f(\zeta_0 - \widehat{\zeta}))'r^Y| \\
&+ |(f(\zeta_0 - \widehat{\zeta}))'f(\pi_0 - \widehat{\pi})| + |(f(\zeta_0 - \widehat{\zeta}))'V^D| + |(f(\zeta_0 - \widehat{\zeta}))'r^D| \\
&+ |(V^Z)'f(\widehat{\pi} - \pi_0)| + |(V^Z)'f(\beta_0 - \widehat{\beta})| + +|(V^Z)'r^g|
\end{aligned} \tag{C.1}$$

Under Assumptions AHK, AR, the sparse approximation conditions as well as Assumption REG-P(ii) - (vii), we can apply Theorem 2.1 to obtain that $\|f_{it}(\eta_0 - \widehat{\eta})\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}}\right)$, $\|\eta_0 - \widehat{\eta}\|_1 =$

$O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}\right)$ for $\eta = \zeta, \pi, \beta$, and $P\left(\max_{j=1,\dots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{j,l}^{-1/2} f_{it,j} V_{it}^{l}\right| \geq \frac{\lambda}{2c_1 NT}\right) \to 0$ for $l = Z, D, Y$ where $\lambda = \frac{6c_1 NT}{\sqrt{N\wedge T}}\Phi^{-1}(1-\gamma/2p)$. By Lemma A.2, $\omega_{j,l} \xrightarrow{p} \frac{A\wedge T}{N}\Sigma_{a,j,l} + \frac{N\wedge T}{T}\Sigma_{g,j,l}$ where $\min_{j\leq p}\Sigma_{a,j}^{l} > 0$ by Assumption REG-P(iv) and Lemma A.1. Therefore, $\min_j \omega_{j,l}^{-1/2} > 0$, which implies $\|f'V^l\|_\infty = O_P(\Phi^{-1}(1-\gamma/2p)/\sqrt{N\wedge T}) = O_P\left(\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}\right)$ for $l = D, Y, Z$.

Consider the first term in C.1. By Cauchy-Swartz inequality, we have $\frac{\sqrt{N\wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))' f(\beta_0 - \widehat{\beta})| \leq \sqrt{N\wedge T}\|f_{it}(\zeta_0 - \widehat{\zeta})\|_{NT,2}\|f_{it}(\beta_0 - \widehat{\beta})\|_{NT,2} = O_P\left(\frac{s\log(p/\gamma)}{\sqrt{N\wedge T}}\right)$. Consider the second term in C.1. By Holder's inequality, we have $\frac{\sqrt{N\wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))' V^Y| \leq \frac{\sqrt{N\wedge T}}{NT}\|\zeta_0 - \widehat{\zeta}\|_1\|f'V^Y\|_\infty = O_P\left(s\frac{\log(p/\gamma)}{\sqrt{N\wedge T}}\right)$. Consider the third term in C.1. By Cauchy-Swartz inequality, we have $\frac{\sqrt{N\wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))' r^Y| \leq \sqrt{N\wedge T}\|f_{it}(\zeta_0 - \widehat{\zeta})\|_{NT,2}\|r_{it}^Y\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right)$. For the last term of C.1, Cauchy-Swartz inequality implies that $\frac{\sqrt{N\wedge T}}{NT}|(V^Z)'r| \leq \sqrt{N\wedge T}\|V_{it}^Y\|_{NT,2}\|r_{it}^Y\|_{NT,2}$. By Assumption REG-P(ii), we have $|\mathrm{E}[(V_{it}^Y)^2]^{4(\mu+\delta)}| < \infty$. Then we can apply Lemma A.1 and obtain that $\|V_{it}^Y\|_{NT,2} \to (\mathrm{E}[(V_{it}^Y)^2])^{1/2}$. Therefore, we have $\frac{\sqrt{N\wedge T}}{NT}|(V^Z)'r| = o_P(1)$. The arguments for the rest of the terms in C.1 are similar. Under the sparsity condition $s = \frac{\sqrt{N\wedge T}}{\log(p/\gamma)}$, we conclude that $\sqrt{NT}|\hat{\psi}_{NT} - \psi_{NT}| = o_P(1)$.

Consider Statement (vi). By product decomposition, we have

$$
\begin{aligned}
NT\left\|\widehat{A}_{NT} - A_{NT}\right\|_1 &= \left\|\left(f(\zeta_0 - \widehat{\zeta})\right)' f(\pi_0 - \widehat{\pi}) + \left(f(\zeta_0 - \widehat{\zeta})\right)'(D - f\pi_0) + (Z - f\zeta_0)' f(\pi_0 - \widehat{\pi})\right\|_1 \\
&\leq \left\|\left(f(\zeta_0 - \widehat{\zeta})\right)' f(\pi_0 - \widehat{\pi})\right\|_1 + \left\|\left(f(\zeta_0 - \widehat{\zeta})\right)'(r^D + V^D)\right\|_1 + \left\|(V^Z)' f(\pi_0 - \widehat{\pi})\right\|_1
\end{aligned}
$$

We observe that, by similar arguments for Statement (v), $\left\|\widehat{A}_{NT} - A_{NT}\right\|_1 = o_P(1)$. We have shown Statement (i)-(iv), completing the proof. □

Proof of Theorem 4.2. We have shown in the proof of Theorem 4.1 that $\widehat{A}_{NT} - A_{NT} = o_P(1)$ and $A_{NT} - A_0 = o_P(1)$. By triangle inequality, we have $\widehat{A}_{NT} - A_0 = o_P(1)$. Then, it suffices to show $\widehat{\Omega}_{\mathrm{CHS}} - \Omega = o_P(1)$. We decompose $\widehat{\Omega}_{\mathrm{CHS}}$ as follows:

$$
\begin{aligned}
\widehat{\Omega}_{\mathrm{CHS}} &:= \widehat{\Omega}_a + \widehat{\Omega}_b - \widehat{\Omega}_c + \widehat{\Omega}_d + \widehat{\Omega}_d', \\
\widehat{\Omega}_a &:= \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{ir}(\widehat{\theta},\widehat{\eta})', \quad \widehat{\Omega}_b := \frac{1}{NT^2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{jt}(\widehat{\theta},\widehat{\eta})', \\
\widehat{\Omega}_c &:= \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{it}(\widehat{\theta},\widehat{\eta})', \quad \widehat{\Omega}_d := \frac{1}{NT^2}\sum_{m=1}^{M-1}k\left(\frac{m}{M}\right)\sum_{t=1}^{T-m}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{j,t+m}(\widehat{\theta},\widehat{\eta})'.
\end{aligned}
$$

where $\psi_{it}(\theta,\eta) = (Z_{it} - f_{it}\zeta)(Y_{it} - f_{it}\beta - \theta(D_{it} - f_{it}\pi))$ and $\eta = (\zeta, \beta, \pi)$. We need to show $\widehat{\Omega}_a \xrightarrow{p} \Sigma_a = \mathrm{E}_P[a_i^2]$, $\widehat{\Omega}_b \xrightarrow{p} c\mathrm{E}[g_t^2]$, $\widehat{\Omega}_c = o_P(1)$, and $\widehat{\Omega}_d \xrightarrow{p} c\sum_{m=1}^{\infty}\mathrm{E}_P[g_t g_{t+m}]$.

First, consider $\widehat{\Omega}_a - \mathrm{E}_P[a_i^2]$. By triangle inequality, we have

$$
\begin{aligned}
\left|\widehat{\Omega}_a - \mathrm{E}_P[a_i^2]\right| &\leq |I_{a,1}| + |I_{a,2}| + |I_{a,2}|, \\
I_{a,1} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{ir}(\widehat{\theta},\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0)\right\}, \\
I_{a,2} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0) - \mathrm{E}[\psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0)]\right\}, \\
I_{a,2} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\mathrm{E}[\psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0)] - \mathrm{E}[a_i^2]\right\}.
\end{aligned}
$$

Note that in proving Claim B.3, the cross-fitting device is only used to show that $I_{a,1}$ is of small order. Since the arguments for showing $I_{a,2}$ and $I_{a,3}$ to be of small order are basically the same as those in the proof of Claim B.3, they are not repeated here.

Consider $I_{a,1}$. By product decomposition, triangle inequality, and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
|I_{a,1}| \lesssim& R_{NT}\left\{|\psi_{it}(\theta_0,\eta_0)|_{NT,2} + R_{NT}\right\} \\
R_{NT} :=& \left\|\psi_{it}(\widehat{\theta},\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0)\right\|_{NT,2}
\end{aligned}
$$

By Minkowski's inequality, we have

$$
\begin{aligned}
R_{NT} &= \left\|\psi_{it}^a(\eta_0)(\widehat{\theta}-\theta_0) + (\psi_{it}^a(\eta_0) - \psi_{it}^a(\widehat{\eta}))(\widehat{\theta}-\theta_0) + \psi_{it}(\theta_0,\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0)\right\|_{NT,2} \\
&\leq \left\|\psi_{it}^a(\eta_0)(\widehat{\theta}-\theta_0)\right\|_{NT,2} + \left\|(\psi_{it}^a(\eta_0) - \psi_{it}^a(\widehat{\eta}))(\widehat{\theta}-\theta_0)\right\|_{NT,2} + \left\|\psi_{it}(\theta_0,\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0)\right\|_{NT,2} \\
&=: R_{a,1} + R_{a,2} + R_{a,3},
\end{aligned}
$$

where $\psi_{it}^a(\eta) := \left(Z_{it} - f_{it}\zeta\right)(D_{it} - f_{it}\pi)$. Under Assumption REG-P(ii), we have $\mathrm{E}_P[\psi_{it}^a(\eta_0)]^2 = \mathrm{E}_P[V_{it}^Z(V_{it}^D + r_{it}^D)]^2 = O_P(1)$, and Markov inequality implies that $\|\psi_{it}^a(\eta_0)\|_{NT,2} = O_P(1)$. By Theorem 4.1, we have $\widehat{\theta} - \theta_0 = O_P\left(\frac{1}{\sqrt{N\wedge T}}\right)$. Therefore, $R_{a,1} \leq \|\psi_{it}^a(\eta_0)\|_{NT,2}|\widehat{\theta}-\theta_0| = O_P\left(\frac{1}{\sqrt{N\wedge T}}\right)$. To bound $R_{a,2}$, we note

$$
\|\psi_{it}^a(\eta_0) - \psi_{it}^a(\widehat{\eta})\|_{NT,2} = \left\|f_{it}(\widehat{\zeta}-\zeta_0)(D_{it} - f_{it}\pi_0) + f_{it}(\widehat{\zeta}-\zeta_0)f_{it}(\widehat{\pi}-\pi_0) + (Z_{it} - f_{it}\zeta_0)f_{it}(\widehat{\pi}-\pi_0)\right\|_{NT,2}
$$

Under Assumption REG-P(iii), we have $\mathrm{E}_P|V_{it}^D|^{8(\mu+\delta)} < \infty$, which implies $\mathrm{E}_P\left[\max_{i\leq N,t\leq T}|V_{it}^D|^2\right] \lesssim (NT)^{\frac{1}{4(\mu+\delta)}}$. By Markov inequality, we have $\max_{i\leq N,t\leq T}|V_{it}^D|^2 = O_P((NT)^{\frac{1}{4(\mu+\delta)}})$. As in the proof of Theorem 4.1, Theorem 2.1 can be applied to obtain $\left\| f_{it}(\widehat{\zeta} - \zeta_0) \right\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right)$. Then, we have

$$\begin{aligned} R_{a,2} = &\left\| f_{it}(\widehat{\zeta} - \zeta_0) V_{it}^D \right\|_{NT,2} \leq \left(\max_{i\leq N, t\leq T} |V_{it}^D|^2 \right)^{1/2} \left\| f_{it}(\widehat{\zeta} - \zeta_0) \right\|_{NT,2} \\ =&O_P((NT)^{\frac{1}{8(\mu+\delta)}}) O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right) = O_P((NT)^{\frac{1}{8(\mu+\delta)}}) o_P\left(\frac{1}{(N\wedge T)^{1/4}}\right) = o_P(1). \end{aligned}$$

Similar arguments can be made to show $R_{a,3}$. Therefore, we have $R_{NT} = o_P(1)$ and so $\widehat{\Omega}_a \xrightarrow{p} \Sigma_a$

It is left to show that $\widehat{\Omega}_b \xrightarrow{p} c\mathrm{E}[g_t^2]$, $\widehat{\Omega}_c = o_P(1)$, and $\widehat{\Omega}_d \xrightarrow{p} c\sum_{m=1}^{\infty} \mathrm{E}_P[g_t g_{t+m}]$. As is shown in the proof of Theorem 3.2 (Lemmas A.5-A.7), the only step in showing these claims that involve cross-fitting technique is to show the same term R_{NT} to converge to 0 in probability. Otherwise, the arguments are basically the same and not repeated here. Combining these results, we obtain $\widehat{\Omega} \xrightarrow{p} \mathrm{E}_P(a_i^2) + c\mathrm{E}_P(g_t^2) + c\sum_{m=1}^{\infty} \mathrm{E}_P(g_t g_{t+m}) = \Sigma_a + c\Sigma_g$.

To show $\widehat{V}_{\mathrm{DKA}} = \widehat{V}_{\mathrm{CHS}} + o_P(1)$, it suffices to show $\widehat{\Omega}_{\mathrm{NW}} = o_P(1)$. We decompose Ω_{NW} as follows:

$$\widehat{\Omega}_{\mathrm{NW}} = \widehat{\Omega}_c + \widehat{\Omega}_e - \widehat{\Omega}_d,$$

where $\widehat{\Omega}_c$ and $\widehat{\Omega}_d$ are defined as above and $\widehat{\Omega}_e$ is defined as follows:

$$\widehat{\Omega}_e := \frac{1}{NT^2} \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \sum_{t=1}^{T-m} \sum_{i=1}^{N} \sum_{j=1}^{N} \psi(W_{it}; \widehat{\theta}, \tilde{\eta}) \psi(W_{j,t+m}; \widehat{\theta}, \tilde{\eta}).$$

Following the same arguments as in the proof of Claim B.6, we have $\widehat{\Omega}_e = \widehat{\Omega}_d + o_P(1)$. We have shown $\widehat{\Omega}_c = o_P(1)$. Therefore, we conclude that $\widehat{\Omega}_{NW} = o_P(1)$. So it is proved. □