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

Kaicheng Chen

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

Dec 29, 2024

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. For estimation, I propose a weighted LASSO with the two-way cluster-robust penalty weights serving as a self-normalization. Due to the cluster dependence driven by the underlying components, the rate of convergence is slow even in an oracle case. Nevertheless, by leveraging a cross-fitting approach robust to panel data dependence, the asymptotic normality on low-dimensional treatment parameters can be established using the weighted LASSO for nuisance estimation. Additionally, I address the challenges posed by unobserved heterogeneity, which introduces a subtle issue for cross-fitting. Strategies and implications on the sparsity condition under various scenarios are discussed. 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. For example, approximating an unknown function $g(X)$ using the 3rd-order polynomial transformation of X involves the variables in X themselves and all quadratic and cubic terms including the interactions. For a vector X with dimension $k = 10$, it involves 285 regressors; and for $k = 20$, it involves 1770 regressors.[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/common time effects and heterogeneous trends/common time effects 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 types of high dimensionality matter potentially:

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

where D_{it} is a vector of low-dimensional treatment or policy variables and U_{it} is an exogenous stochastic error; X_{it} is a vector of potentially high-dimensional control variables; $g(.)$ 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 average partial effect 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. In recent years, to address the high-dimensional issues in the model, regularization approaches, also known as machine learning, have been

[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.

employed for estimation, which trades off bias for smaller variance. However, due to the bias introduced by regularization and overfitting, inference is challenging. Typically, two key components are involved to obtain desirable statistical properties of high-dimensional estimation and inference approaches. The first component typically involves bias correction, accounting for the regularization bias. The second component targets the overfitting issue, caused by potentially too many relevant regressors, and so it often relies on a condition that only a subset of the high-dimensional regressors are relevant, i.e. a sparsity assumption. Additionally, a sample-splitting or cross-fitting procedure is sometimes involved to relax the sparsity assumption so that it allows for more complex/less sparse models.

In a panel data setting, it is soon realized that three challenges would appear if researchers attempt to apply the existing high-dimensional approaches directly. First of all, the statistical properties of many high-dimensional estimators remain unknown with panel data, which is potentially dependent across space and time. Secondly, some procedures for inference such as sample splitting and cross fitting are very specific to the dependence structure of the data and existing approaches are not general enough to deal with two-way dependence in panels. Thirdly, panel data models often take unobserved individual and time effects into consideration, which may lead to another source of high dimensionality and further complicate estimation and inference.

For the first challenge, I proposed a variant of LASSO that uses regressor-specific penalty weights as self-normalization and is robust to two-way cluster dependence and weak temporal dependence across clusters. Such a LASSO approach is named a two-way cluster-LASSO, corresponding to the heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016). By decomposing the moment condition using Hajek projection to unit, time, and a remaining small order term, I am able to leverage moderate deviation theorems for self-normalized sums of independent and weakly dependent random variables to bound the probability of a so-called "regularization event", i.e. $\lambda > 2\max_{1\leq j\leq p}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right|$ where λ is a common penalty level, ω_j are some regressor-specific weights, $f_{it,j}$ are regressors, and V_{it} are stochastic errors. Combining with existing non-asymptotic bounds for the LASSO approach, I derive the rate of convergence for the (post) two-way cluster LASSO.

Due to the two-way dependence driven by underlying unobserved effects, it is revealed that the convergence rate is not as fast as the common rates for LASSO estimation. Indeed, even under a Gaussian condition on the error term, it can be shown that following the analysis in Theorem 29.3 of Hansen (2022) and the results in Theorem 1 in Chiang et al. (2024), the $l2$ convergence rate for the LASSO approach is still $O_P\left(\sqrt{\frac{s\log(p)}{N\wedge T}}\right)$. The problem lies in the underlying factor structure. 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 (throughout the paper, those components do not introduce any endogeneity issue but are only used for characterizing the dependence). This is a common characterization of cluster dependence in cluster-robust inference literature: we notice that α_i introduce cluster/temporal dependence within group i and γ_t introduce cluster/cross-sectional dependence within group t. To estimate the high-dimensional vector θ_0, we consider the sample mean estimator $\hat{\theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$. We can rewrite the estimator through a Hajek projection:

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

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 simplicity, 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 $j = 1, ..., s$, $\hat{\theta}_j = O_P\left(\frac{1}{\sqrt{N \wedge T}}\right)$ and $\|\hat{\theta} - \theta_0\|_2 = \left(\sum_{j=1}^{s}\left(\hat{\theta}_j - \theta_{0j}\right)^2\right)^{1/2} = O_P\left(\sqrt{\frac{s}{N \wedge T}}\right)$. We see that it is due to the underlying factor structure that prevents a faster rate of convergence. This explains why the convergence rate of the proposed two-way cluster-LASSO is shown to be $O_P\left(\sqrt{\frac{s\log(p \vee NT)}{N \wedge T}}\right)$. Unless s is a finite number, it may not be a very helpful rate of convergence for inference purposes.

The question is what rate of convergence is necessary for valid inference in a high-dimensional model with two-way dependence. I will answer it by studying a general inference procedure for high-dimensional panel models, which is also the second challenge. Specifically, I propose an inference procedure for low-dimensional parameters in the presence of high-dimensional nuisance parameters in a semiparametric sense. In the first step, high-dimensional nuisance parameters of an orthogonalized moment function are estimated by some high-dimensional methods. In the second step, low-dimensional parameters of interest are estimated by a parametric specification and plug-ins of nuisance estimates. The innovation comes from a cross-fitting algorithm that constructs main and auxiliary samples "approximately" and asymptotically independent of each other. Accordingly, the dependence between the two-step estimations is eliminated so that a potentially over-fitted nuisance estimate from the first step won't pollute the second-step estimator as much as it would otherwise do. Effectively, this inferential procedure extends the double/debiased machine learning (DML, hereafter) approach by Chernozhukov et al. (2018a) to panel data models, and so it is labeled as 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 makes the first-step estimation through the two-way cluster LASSO feasible under a suitable sparsity condition.

Another idea is to deal with the unobserved factors before considering estimation and inference in a high-dimensional setting. To be concrete, again I consider a component structure representation of the panel data:

$$(Y_{it}, X_{it}, U_{it}) = f(\alpha_i, \gamma_t, \epsilon_{it}), \tag{1.4}$$

where α_i, γ_t, and ε_{it} are unobserved random elements (throughout the paper, those components do not introduce any endogeneity issue but are only used for characterizing the dependence). If f is a linear function, e.g. $X = \alpha_i^x + \gamma_t^x + \varepsilon_{it}^x$, then a two-way within transformation eliminates α_i and γ_t. In a linear regression model with unobserved heterogeneous effects,

$$Y_{it} = X_{it}\beta_0 + c_t + d_t + U_{it},$$

it means that a two-way within-estimator not only deals with the endogeneity issue caused by (c_i, d_t) but also removes the two-way cluster dependence due to the underlying factors. Given that a two-way fixed-effect estimator and a two-way Mundlak approach are algebraically equivalent to the two-way within-estimator (Wooldridge, 2021), the cluster-dependence can also be dealt with by these other two methods. That is good news because the within-transformation methods can only remove the underlying components under very specific function forms but fixed-effect and Mundlak device approaches are much more flexible. If f is nonlinear in its components, e.g. $X_{it} = \alpha_{i1}\gamma_{2t} + \alpha_{i2}\gamma_{1t} + \varepsilon_{it}^x$ and $U_{it} = \alpha_{i1}\gamma_{3t} + \alpha_{i3}\gamma_{1t} + \varepsilon_{it}^u$ where all components are assumed to be i.i.d with mean 0 and variance 1, then the moment function $X_{it}U_{it}$ possesses a non-degenerate component structure: $X_{it}U_{it} = a_i + g_t + \varepsilon_{it}$ where $a_i = \mathrm{E}[X_{it}U_{it}|\alpha_i] = \alpha_{2i}\alpha_{3i}$ and $g_t = \mathrm{E}[X_{it}U_{it}|\gamma_t] = \gamma_{2t}\gamma_{3t}$ (Chiang et al., 2024). While a within-transformation cannot remove the underlying factors, an interactive fixed-effect projection on X and U enables the complete removal of the components. Although it is not clear if the fixed-effect and Mundlak device approaches enable the complete removal of the underlying factor and the resulting cluster dependence due to the unknown function form of f, it does suggest that we would like a fixed-effect or Mundlak device approach as flexible as possible to do the job.

In this model, though the unobserved heterogeneous effects (c_i, d_t) are in general different objects from the underlying cross-sectional and time components (α_i, γ_t), they are also closely related: while (c_i, d_t) cause an identification problem and (α_i, γ_t) bring cluster dependence, flexible modeling of (c_i, d_t) takes care of both issues. This is the reason why researchers care about unobserved heterogeneity and this is also the third challenge in this paper. Consider the model 1.1, a common approach for dealing with the unknown function is to use a series approximation, except that, in this case, c_i and d_t are unobservable. A fixed effect approach takes c_i and d_t as random initially and the analysis is based on conditioning on their realized values, and they will be estimated along with the slope coefficients associated with the observables and their transformations. However, it is well known that this approach would lead to an incidental parameter problem either when N or T diverges. With both N and T diverging, which is the main focus of this paper, the bias caused by the incidental parameter problem can still persist (see, for example, Hahn and Newey, 2004). Alternatively, in high-dimensional literature, regularization approaches are used to estimate fixed-effect parameters (Kock and Tang, 2019; Semenova et al., 2023a) by imposing sparsity directly on the fixed-effect parameters. In this paper, I take a correlated random effect approach as a more tractable way to avoid the incidental parameter problem. Specifically, I model the unobserved heterogeneity through a generalized Mundlak device, i.e. a nonparametric function of the cross-sectional and temporal sample averages and an independent error term. Instead of imposing sparsity on the unobserved effects themselves, this approach imposes the sparsity assumption on the slope coefficient of the proxy of those unobserved effects instead of assuming some of those effects themselves are zero.

It sounds like a good solution where researchers proxy the unobserved effects by observable random variables, which avoids the incidental parameter problem, resolves the endogeneity issue, and potentially removes the cluster dependence; and researchers perform the panel DML estimation and inference procedures with the first-step estimates obtained by the two-way cluster LASSO, assuming those first-step estimators possess desirable rate of convergence. Indeed, if the underlying unit and time components are completely removed, then the dependence only comes from the remainder term e_{it}. Under the assumption, e_{it} can be only weakly dependent over time or simply independent. Then, the proposed toolkit for the two-way clustered panel is still valid with a faster convergence rate. Except, there is one more subtle issue: with unobserved heterogeneous effects c_i and d_t, all relevant approaches mentioned above (within-transformation, fixed-effect, and Mundlak device approaches) inevitably introduce some functions of sample averages into the regressors, which may not be compatible with the cross-fitting scheme. For example, a two-way within-transformation brings full-sample cross-sectional and temporal averages into all regressors in a linear regression model and then the observations are dependent across cross-fitting sub-samples. One alternative way is to conduct within transformation in each sub-sample, but it relies on the linear function form of c_i and d_t. As is further shown in Section 5, in a similar sense, we can impose a stronger Mundlak device condition to get around this non-compatibility, except that this condition may be too strong to be plausible. On the other hand, without cross-fitting, it is unclear whether the panel DML inference is still valid with the growing dimension of the nuisance parameters in a general semiparametric moment restriction model. Therefore, as a special case, I demonstrate in a partial linear model similar to 1.1 but allowing for endogenous treatments, that it is possible to establish an asymptotic normality result without cross-fitting, given a high-level assumption on the first-step estimator. The implication of the high-level assumption on the sparsity condition is also discussed.

In the empirical application, 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 and my question is whether there is improvement can be made in estimation and inference. Transitionally studied in a time series context, it is not considered a high-dimensional problem. While they study it using a panel data framework and, because of that, the dimension of the nuisance parameters does increase with the sample size, it is not considered as a high-dimensional problem in the baseline setting: the identification is through the instrumental variable and at most one control variable is considered. With the inclusion of the fixed-effect parameters and unit-specific slope coefficient, the number of nuisance parameters goes above 100 but is still well below the sample size of about 2000. However, as I demonstrate in Section 7, even in a conventionally low-dimensional setting, there is hidden high dimensionality. With the approaches proposed in this paper, I extend their analysis by permitting flexible modeling and robust inference, which can be regarded as a robustness check. It is shown that even with more complex models, the estimates are very consistent compared to the original set of results while keeping the estimates the same or even less noisy.

The rest of the paper is outlined as follows: The next sub-section reviews (extra) 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 asymptotic properties under two-way dependence. Section 3 introduces a sub-sampling scheme designed for cross-fitting that allows within-cluster dependence and weak dependence across clusters. Section 4 studies the high-dimensional inference problem in a general semiparametric moment restriction model using panel data, which gives a rate requirement of the first-step estimator for obtaining valid inference on the low-dimensional parameter of interest. In Section 5, the high-dimensional partial linear panel model defined at the beginning of the paper is revisited, under which I study the problem of unobserved heterogeneity in detail and illustrate that asymptotic normality can be established with or without cross-fitting. Simulation evidence is given in Section 6 where the toolkit proposed in this paper is competed with each other as well as existing approaches. In Section 7, the empirical estimation of the government spending multiplier is used as an illustration of hidden high dimensionality and the application of the proposed toolkit. Section 8 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) first derive the convergence rate of the prediction risk in terms of the empirical norm under homogeneous Gaussian error, restricted eigenvalue, and sparsity assumption. 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. By utilizing self-normalizing penalty weights and leveraging on a moderate deviation theorem from Jing et al. (2003) and Peña et al. (2009), Belloni et al. (2012) first show the convergence rates of LASSO estimator can be derived under non-Gaussian errors and approximate sparsity. Belloni et al. (2016) extend the analysis to a panel data model with arbitrary within-cluster dependence and cross-sectional independence. Other literature on LASSO-based estimators with dependent data includes Basu and Michailidis (2015); Kock and Callot (2015); Lin and Michailidis (2017), assuming either Gaussian or sub-Gaussian errors. Using the functional dependence measure of Wu (2005) to characterize the dependency, Wu and Wu (2016) relax the Gaussian assumption by deriving Nagaev-type inequalities using only moment conditions, and, recently, Gao et al. (2024) establish Nagaev-types inequalities in a panel data setting with two-way dependence. In the setting of a system of equations, Chernozhukov et al. (2021a) provide a method of choosing the common penalty level through block bootstrap and establishing performance bound for the LASSO estimator under dependence in both space and time characterized by the functional dependence measure. Both the functional dependence characterization and the component structure characterization used in this paper feature two-way dependence in panel data settings, but they are not nested within each other. Thus, the method presented in this paper complements the existing literature.

The inferential theory in high-dimensional regression models typically relies on some bias-correction methods to account for the regularization bias. It takes 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 the debiased machine learning (DML) approach, 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 also related to the doubly-robust literature. 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.

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. It is an important ingredient when the dimension of nuisance parameters increases as the sample size diverges and when the model is less sparse. To relax the sparsity assumption and circumvent the stochastic equicontinuity condition when establishing the asymptotic normality, Belloni et al. (2014) propose a sample-splitting procedure that removes the dependence between the first-step estimates and the data used for the second step. 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. However, the sample splitting or cross-fitting is very specific about the sampling assumption and dependence structure of the data. Chiang et al. (2021, 2022) propose a cross-fitting scheme robust to separately and jointly exchangeable arrays. Semenova et al. (2023a) propose a cross-fitting scheme robust to weak dependence and introduce a coupling approach (due to Strassen, 1965 and Berbee, 1987) to show their cross-fitting sub-samples are independent with the probability approaching one as the sample size grows. 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.

As for the unobserved heterogeneity issue in high-dimensional panel models, it has been considered in Belloni et al. (2016); Kock and Tang (2019); Vogt et al. (2022); Gao et al. (2024) among others. In Belloni et al. (2016), the unobserved heterogeneity can be eliminated by the two-way within-transformation due to the linear additivity. Kock and Tang (2019) instead takes a fixed-effect approach by estimating and penalizing the realized values of the unobserved heterogeneity while imposing sparsity assumption on these fixed-effect parameters. Vogt et al. (2022) and Gao et al. (2024) model the unobserved heterogeneity as interactive fixed-effects. In this paper, I take the stand that the additive form ignores the possibility of interactive effects between the controls and the unobserved heterogeneous effects, and imposing sparsity on the fixed effects might be not very natural in certain applications because it implies that certain units have nonzero heterogeneous effects while not others. Instead, I view unobserved heterogeneous effects are correlated random effects and I generalize the Mundlak device approach due to Mundlak (1978) and two-way Mundlak approach in Wooldridge (2021) in the sense that it allows for nonlinearity in the heterogeneous effects and arbitrary interaction with the covariates. A similar idea has been implemented in Wooldridge and Zhu (2020). Furthermore, it is the first paper that addresses the subtle issue of unobserved heterogeneity when the cross-fitting approach is involved.

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, strictly speaking, it is not a representation anymore but more of an assumption and a general characterization of cluster dependence. Such characterization of the dependence structure is common in economics 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). In this paper, AHK representation introduces dependence both within and across clusters, and the asymptotic variance of the DML estimator has two components: one is due to within-unit dependence and one is due to both within-time dependence and across-time dependence. Therefore, the usual one-way or two-way cluster variance estimator is not valid. I propose variance estimators similar to Chiang et al. (2024) and Chen and Vogelsang (2024), with careful adjustment due to cross-fitting procedures.

1.2. Notation.

Here is a collection of the most frequently used notations in this paper. Some extra notations are defined along with the context. I use E and P as generic expectation and probability operators. I denote $\mathcal{P}_{NT}$ as an expanding collection of all data-generating processes P that satisfy certain conditions. I denote P_{NT} as a sequence of probability laws such that $P_{NT} \in \mathcal{P}_{NT}$. for each (N,T) We will suppress the dependence on (N,T) and P_{NT} whenever clear in its setting. We will use the following vector and matrix norms: we denote $\|.\|$ as 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 $1 \le q < \infty$, and $\|\mathbf{x}\|_\infty := \max_{1\le j\le k} |x_j|$. 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 $\mathbb{E}_{kl}[f_{it}] = \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l} f_{it}$, 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 literature, little is known in terms of statistical properties for high-dimensional methods under cluster dependence in both space and time. In this section, a variant of the $l1$-regularization methods, also known as the LASSO, will be constructed and examined. Besides a common penalty level that is theoretically driven, regressor-specific penalty weights robust to two-way cluster dependence and temporal weak dependence across clusters will be utilized. Such a LASSO approach is named two-way cluster LASSO, corresponding to the existing heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016). The intuition is to replace the regressor-specific penalty weights employed in the aforementioned papers with some alternative choices that are robust to two-way cluster dependence in the panel. However, the valid choices are not immediately clear and the theory is highly non-trivial due to the dependence in both cross-sectional and temporal dimensions.

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, and the regression model throughout this section is free from the endogeneity issue caused by unobserved heterogeneous effects. 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 $f(X) := E[Y|X]$ is an unknown conditional expectation function of potentially high-dimensional covariates X and 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 from 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 $X_{ij} = f(\nu, \xi_i, \iota_j, \zeta_{ij})$ where $\nu, (\xi_i)_{i\geq 1}, (\iota_j)_{j\geq 1}, (\zeta_{ij})_{i\geq 1, j\geq 1}$ 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 correlated 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 in a fairly general. 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, I consider the time effects γ_t to be weakly dependent, and some regularity condition is introduced for tractability. Before that, a few more concepts and notations need to be introduced. Let (X, Y) be random elements taking values in Euclidean space $S = (S_1 \times S_2)$ with 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 a certain 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 assume sparsity 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 assumption in previous literature in high-dimensional regression, it requires a relatively slow 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 factor structure restricts the growth of nonzero slope coefficients of the model in a similar way to the perfect temporal dependence case in Belloni et al. (2016).

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 estimate ζ_0 allowing its dimension to be greater than the sample size by applying $l1$ 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:

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

To obtain the desirable property of LASSO estimation or variable selection, one needs to choose λ and ω in an optimal way that the penalty level is large enough to avoid noisy estimation due to overfitting but also the smallest possible since the size of 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 $\dot{f}_{it}$ be the demeaned f_{it} by the sample averages[6] and one common choice of ω is the empirical Gram matrix $\mathrm{E}[\dot{f}_{it}' \dot{f}_{it}]$ 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 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):

$$\max_{j=1,\dots,p} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_j^{-1/2} f_{it,j} V_{it} \right| \leq \frac{\lambda}{2c_1 NT}. \tag{2.6}$$

Condition 2.6 is referred to as the "regularization event" in the literature. Combining with some algebraic results, the rate of convergence for LASSO estimation 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 and are 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 and the validity of cross-validation is highly non-trivial.

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 though 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

[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.

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 components: $a_i = \mathrm{E}[f_{it}V_{it}|\alpha_i]$, $g_t = \mathrm{E}[f_{it}V_{it}|\gamma_t]$, $e_{it} = f_{it}V_{it} - a_i - g_t$, 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, I can construct the regressor-specific penalty weights in a way that existing moderate deviation theorems for i.i.d and weakly dependent sums can be leveraged for the first two Hajek components separately while some concentration inequality can be applied to the third component.

For that purpose, 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 guarantees a faster convergence rate of LASSO. Again, both C_λ and γ do not vary across different DGPs. While there is some guidance about choosing C_λ and γ through the asymptotic theory, the choice is not given exactly. 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 some initial residual $\tilde{V}_{it}$ and then $\tilde{V}_{it}$ 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 (unscaled) cluster variance estimator for $\mathrm{Var}(\mathrm{E}_{NT}[f_{it,j}V_{it}])$, which is also used as the regressor-specific penalty weights for 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, while the second term clusters over individual first and then clusters over time so it adjusts for cross-sectional dependence within each time cluster and temporal dependence across time clusters. The validity of estimating those components through cross-sectional and temporal averages is given in Menzel (2021) for exchangeable arrays. Extending the consistency results of those sample averages 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 the practice in Belloni et al. (2012) and

Belloni et al. (2016), the asymptotic analysis in this paper is based on high-level assumptions on the feasible penalty weights: Let $\hat{\omega}$ be the feasible diagonal weights and suppose there exists $0 < l \leq 1$ 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.

Observing that the regressor-specific penalty weights can also be interpreted as estimators for $\text{Var}(\mathrm{E}_{NT}[f_{it,j}V_{it}])$, the diagonal terms of two-way cluster-robust variance estimators may also be used as the penalty weights. For example, consider the variance estimator proposed in Chiang et al., 2024[7]:

$$\omega_j^{\text{CHS}} = \omega_j^{\text{A}} + \omega_j^{\text{DK}} - \omega_j^{\text{NW}}, \tag{2.10}$$

where, with $v_{it,j} \equiv f_{it,j}V_{it}$,

$$\begin{aligned}
\omega_j^{\text{A}} =& \frac{N \wedge T}{N^2T^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} v_{it,j}v_{is,j}, \\
\omega_j^{\text{DK}} =& \frac{N \wedge T}{N^2T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right)\left(\sum_{i=1}^{N} v_{it,j}\right)\left(\sum_{l=1}^{N} v_{ls,j}\right), \\
\omega_j^{\text{NW}} =& \frac{N \wedge T}{N^2T^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right) v_{it,j}v_{is,j}.
\end{aligned}$$

Chen and Vogelsang (2024) propose a two-term version of this estimator by omitting the term ω_j^{NW} due to the observation that the third term is of small order and can introduce finite sample bias. In the extreme case, the three-term estimator ω_j^{CHS} can cause non-trivial encountering of negative variance estimates in the finite sample and it can be alleviated by the two-term estimator

$$\omega_j^{\text{DKA}} := \omega_j^{\text{A}} + \omega_j^{\text{DK}} \tag{2.11}$$

Given consistent residual estimation, Theorem 2 of Chiang et al. (2024) and Proposition 2 of Bester et al. (2008) can be applied to show the point-wise equivalent of the weights in 2.8, 2.10, and 2.11. However, due to the iterative estimation and high dimensionality, showing the uniform asymptotic equivalence of the feasible weights requires extra conditions and is left for future research. All three sets of feasible weights are utilized and compared in the empirical example and the simulation and they do exhibit similar and desirable finite

[7] As it is shown in Chen and Vogelsang (2024), the CHS estimator is algebraically equivalent to an affine combination of a cluster estimator ω_j^{A} (Arellano, 1987), a "HAC of Averages" estimator ω_j^{DK} (Driscoll and Kraay, 1998), and a "Averages of HACs" estimator ω_j^{NW} (Newey, 1994).

sample behaviors. The implementation of the two-way cluster-LASSO using the feasible penalty weights is given as follows:

Algorithm: Implementation of the Two-Way Cluster-LASSO

i Obtain the initial residuals $\tilde{V}$: estimate the model without penalization if feasible; if not feasible, include a certain (user-specified) number of the most correlated regressors. [8]

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, 2.10, or 2.11 using $\tilde{V}$.

iii Using $\tilde{\omega}$ for LASSO estimation as in 2.5 and update the residual $\tilde{V}$ using the (post) LASSO estimation.[9]

iv Repeat steps ii-iii until it converges or up to a pre-set number of iterations. Obtains the (post) LASSO estimates from the last iteration.

Before delivering the main results of the weighted LASSO estimator above, two more sets of regularity conditions are needed. 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, $\kappa_1 \le \phi_{\min}(Cs)(M_f) < \phi_{\max}(Cs)(M_f) \le \kappa_2$.*

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 conditions for both types of assumptions are given in Belloni et al. (2012).

Assumption REG (Regularity Conditions). *(i) $\left[\mathrm{E}(a_{i,j})^2\right]^{1/2} / \left[\mathrm{E}(a_{i,j})^3\right]^{1/3} = O(1)$ where $a_{i,j} := \mathrm{E}[f_{it,j}V_{it}|\alpha_i]$ for $j = 1, ..., p$. (ii) $\log(p \vee NT) = o\left(T^{1/6}/\log T\right)$, $p = o\left(T^{7/6}/\log T\right)$. (iii) For some $s > 1, \delta > 0, \mu > 0$, $\mathrm{E}[\|f_{it,j}\|^{8(s+\delta)}] < \infty$, $\mathrm{E}[\|V_{it}\|^{8(s+\delta)}] < \infty$ and $\mathrm{E}\left[\sum_{t=r}^{r+m} f_{it,j}V_{it}\right]^2 \ge \mu^2 m$ for all j and $r \ge 0, m \ge 1$. (iv) Either $\lambda_{a,j} := \left[\mathrm{E}(a_{i,j}^2)\right]^{1/2} > 0$ or $\lambda_{g,j} := \left[\sum_{\ell=-\infty}^{\infty} \mathrm{E}[g_{t,j}g_{t+\ell,j}]\right]^{1/2} > 0$.*

Assumption REG(i) is needed for applying the moderate deviation theorem from Peña et al. (2009). Assumption REG(ii) restricts the dimension of f_{it}. The first condition is binding when the sample size is

[8]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.

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

relatively small and the second condition is binding when the sample size is large. While the number of regressors is allowed to grow with the sample, the second requirement restricts the dimension of regressors to be larger than the sample size. Note that it is not an intrinsic nature of the proposed approach but more of a technical constraint that could be relaxed and is a focus of ongoing research. Assumption REG(iii) is used to apply the moderate deviation theorem from Gao et al. (2022) and to show the remainder in the decomposed $f_{it,j}V_{it}$ is of small order. (iv) is again a non-degeneracy condition, which is the main case of interest. The penalty weights given in 2.10, which is used for feasible estimation, are actually robust to both degeneracy and non-degeneracy. However, the theory leverages the moderate deviation theorems based on the self-normalization weights given by 2.8.

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. In general, $\widehat{\Gamma}$ also allows the inclusion of additional variables chosen by the researcher. However, such generalization is not considered in this paper to avoid further complications. As is shown in Belloni et al. (2012), the Post-LASSO is able to achieve rates of convergence no worse than LASSO, and under certain conditions, it improves upon LASSO. This finding is also supported by our simulation. 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, SE, REG hold for model 2.1 as $N, T \to \infty$ jointly with $N/T \to c$. Additionally, suppose $\widehat{\omega}$ 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, by setting λ as 2.7 with some sufficiently large C_λ, we have the number of selected regressors bounded above by $O(s)$, and*

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

Theorem 2.1 establishes convergence rates in terms of the prediction, $l1$, and $l2$ norms for the 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, driven by an underlying factor structure, the two-way cluster-LASSO is consistent but, unfortunately, has a convergence rate slower than those of LASSO-based methods under the random sampling condition or the cross-section independence. 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 is revealed in the proof of Theorem 2.1 in the Appendix and briefly illustrated in the Introduction, the slow rate of convergence is due to the underlying factor structure. It is unclear if valid inference is still possible under the rate of convergence results in Theorem 2.1. Or, put in another way, what is the minimum requirement of the convergence rates for valid inference in a high-dimensional panel model under two-way dependence? Is it possible to relax the requirement through a cross-fitting procedure? These questions are addressed in the next two sections.

3. Sub-Sampling Scheme for Panel Cross Fitting

To propose a general inference procedure for a high-dimensional panel model in the next section, I will first introduce a new sub-sampling scheme for a two-way clustered panel. 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. Building upon the cross-fitting algorithm for exchangeable arrays in Chiang et al. (2022) and for weakly dependent panels in Semenova et al. (2023a), I propose a cross-fitting scheme robust to two-way cluster dependence and weak dependence over time.

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 with both W_{is} and W_{jt}. Importantly, these types of cluster dependence do 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 $\bigcup_{k=1}^{K} I_k = \{1, ..., N\}$, $\bigcup_{l=1}^{L} S_l = \{1, ..., T\}$[10].

(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)

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

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 $L = 8$.

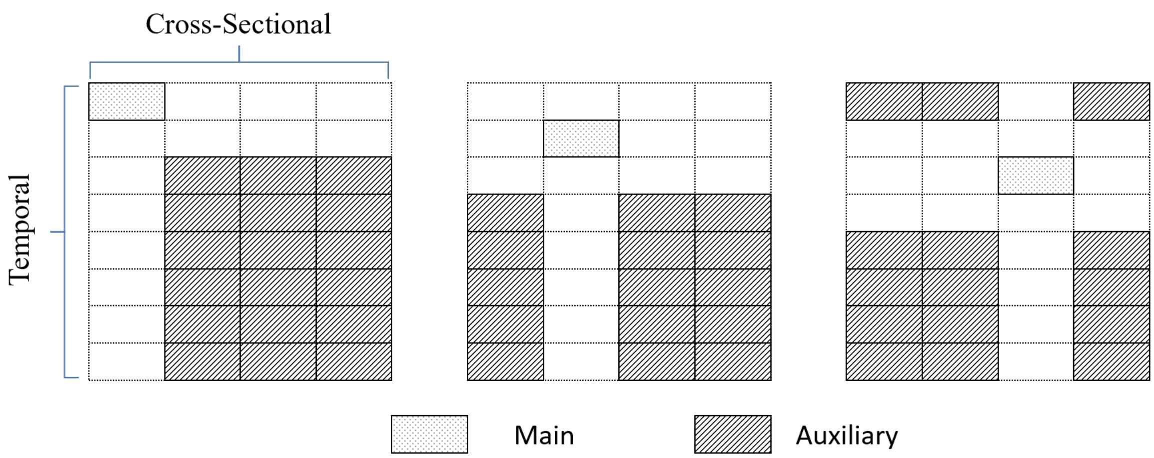

Figure 1: Panel cross fitting with $K = 4$ and $L = 8$. Three graphs from left to right correspond to the main and auxiliary sample constructions with $(k, l) = (1, 1)$, $(k, l) = (2, 2)$, $(k, l) = (3, 3)$. For a simple illustration, observations in the main sample are all adjacent in the cross-sectional 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,

[10] 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 ceil(N/K). and the same applies to the temporal dimension.

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 exists coupling sub-samples that are independent of each other while having the same marginal distributions as the our constructed sub-samples with probability coverging to 1. Such coupling tecnique is common in time series context and the idea in cross-fitting in time dimension follows from Semenova et al. (2023a). The following lemma delievers 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 4.

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 but $\psi(W_{it};\theta,\eta)$ 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 analogue 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 usually involves 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) Select (K,L) and then randomly partition $\{1,2,...,N\}$ into K folds $\{I_1,I_2,...,I_K\}$ and partition $\{1,2,...,T\}$ into L adjacent folds $\{S_1,S_2,...,S_L\}$. 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

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

(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. In our case, 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.*

4. 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 notation from Section 3, 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:

$$\mathrm{E}_P[\psi(W_{it};\theta_0,\eta_0)] = 0, \tag{4.1}$$

$$\partial_r \mathrm{E}_P[\psi(W_{it};\theta_0,\eta_0 + r(\eta-\eta_0))]|_{r=0} = 0. \tag{4.2}$$

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.

Again, the orthogonal moment construction is taken as given. The panel DML procedure is defined in Definition 3.2. Differing from the existing literature, the approach in this paper focuses on estimation and inference robust to two-way cluster dependence and weak dependence across clusters, 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:

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

Let $\lambda_{min}[.]$ denote the smallest eigenvalue of a square matrix. The next assumption specifies the non-degenerate condition.

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

Assumption ND implies that at least one of the components drives the cluster dependence.

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 a_0 and a_1 be some positive and finite constants such that $a_0 < a_1$. 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 $\widehat{\eta}_{kl}$

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 4.1 and 4.2 with respect to the probability measure* P*, or, more generally, 4.2 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 between* a_0 *and* a_1*.*

Assumption DML1(i) restricts the focus of this paper to models with linear orthogonal moment conditions, which covers many applications and the model in Section 5. 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 4.2 by a near-orthogonality condition, which is useful for the approximate sparse model considered in Section 5 because the corresponding orthogonal moment condition does not satisfy 4.2 exactly due to 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* $q > 2$*, the following moment conditions hold:*

$$m_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E}_P \left\| \psi(W_{it}; \theta_0, \eta) \right\|^q \right)^{1/q} \leq a_1,$$
$$m'_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E}_P \left\| \psi^a(W_{it}; \eta) \right\|^q \right)^{1/q} \leq a_1.$$

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

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

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 4.1 (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,*

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

where

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

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[11] and it is also related to inferential theory under strong cross-sectional dependence as in 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 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

[11] 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.

define the CHS-type variance estimator as follows:

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

where, 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,

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

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 4.2 (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 4.2 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 $\widehat{\Omega}_{\text{NW}}$:

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

The estimator is referred to as the DKA variance estimator because it is a sum of Driscoll-Krray and Arellano variance estimators.[12] 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 4.3 (Alternative Consistent Variance Estimator)**.** *Under the same conditions as Theorem 4.2, we have, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

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

Theorem 4.3 formally shows that the double-counting term is of small order under two-way clustering and it implies that the $\widehat{V}_{\text{DKA}}$ is also consistent for Ω under two-way clustering.

To conclude, in this section, I establish the inferential theory for the panel DML estimator, under high-level assumptions on the first-step estimator. It reveals that under the main set of assumptions considered in previous sections, the rate requirement is likely too stringent for any high-dimensional estimator under two-way dependence. However, with an extra m-dependence condition, I show that the crude rate requirement is less stringent and can be achieved by the two-way cluster estimator proposed in Section 2. In the next section, I will study a special case of the semiparametric restriction model but consider the complication due to unobserved heterogeneity. I will demonstrate how to apply the results from this section and previous sections and discuss an extra subtle issue.

5. Partial Linear Model with Unobserved Heterogeneity

Now I am going back to the partial linear model featuring three sources of high dimensionality at the beginning of the paper. We seem to have all toolkits ready for application, but, as is revealed soon, there is one subtle issue caused by the unobserved heterogeneity that has not been addressed. To cover more applications, I will consider a partial linear model with excludable instrument variables: for $i = 1, ..., N$ and $t = 1, ..., T$,

$$Y_{it} = D_{it}\theta_0 + g(X_{it}, c_i, d_t) + U_{it}, \quad \mathrm{E}[U_{it}|X_{it}, c_i, d_t] = \mathrm{E}[Z_{it}U_{it}] = 0, \tag{5.1}$$

[12]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.

where D_{it} is a low-dimensional vector of endogenous treatment variables and we will treat it as a scalar variable for clearer presentation throughout the paper; Z_{it} is an instrumental variable excluded from the outcome equation, and Z_{it} is of the same dimension of D_{it}; g is some unknown measurable functions; X_{it} is a $k \times 1$ vector of control variables[13] and k is allow to grow with the sample size. $\mathbf{D}$ is a $1 \times NT$ vector that stacks D_{it} over $i = 1, ..., N$ and $t = 1, ..., T$; $\mathbf{X}$ is defined similarly. c_i and d_t are treated as random variables. Apparently, model 5.1 includes a partial linear model as a special case by taking $Z_{it} = D_{it}$.

Let $W_{it} = (Y_{it}, D_{it}, Z_{it}, X_{it}, U_{it})$. Again, the two-way dependence in W_{it} is characterized by Assumption AHK and the temporal dependence across clusters is regularized by Assumption AR. As is briefly mentioned in the Introduction, (α_i, γ_t) in Assumption AHK are in general very different objects from (c_i, d_t). Firstly, (α_i, γ_t) are each a vector of arbitrary dimensions while the latter ones are each scalar random elements. Secondly, we are not trying to model (α_i, γ_t) like factor models but, instead, we use those as a dependence measure.

Due to 5.1, an identifying moment condition for θ_0 is given by $\mathrm{E}[Z_{it}U_{it}] = 0$. By the influence function adjustment approach due to Newey (1994), we obtain the following (infeasible) orthogonality moment condition:

$$\mathrm{E}\left[Z_{it} - \mathrm{E}\left[Z_{it}|X_{it}, c_i, d_t\right]\right]\left[Y_{it} - \mathrm{E}\left[Y_{it}|X_{it}, c_i, d_t\right] - \theta_0\left(D_{it} - \mathrm{E}\left[D_{it}|X_{it}, c_i, d_t\right]\right)\right] = 0. \tag{5.2}$$

However, we are not ready to approximate the conditional expectation functions in 5.2 yet since the random effects c_i, d_t are not observed. To proxy the unobserved heterogeneous effects, I take a correlated random-effects approach through a generalized Mundlak device:

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

$$c_i = h_c(\bar{D}_i, \bar{X}_i) + \epsilon_i^c, \tag{5.3}$$

$$d_t = h_d(\bar{D}_t, \bar{X}_t) + \epsilon_t^d, \tag{5.4}$$

where h_c and h_d are some unknown measurable functions; $\bar{D}_i = \frac{1}{T}\sum_{t=1}^{T} D_{it}$, $\bar{D}_t = \frac{1}{N}\sum_{i=1}^{N} D_{it}$, and $(\bar{X}_i, \bar{X}_t)$ are defined similarly; the stochastic errors $(\epsilon_i^c, \epsilon_t^d)$ are each i.i.d random variables, independent of $(\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{Z})$.

A similar assumption is considered in Wooldridge and Zhu (2020). 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 fixed-effect approach, i.e.

[13]If D_{it} is exogenous, then it is allowed to include lags or leads of D_{it} in X_{it}. It would not change the theory for estimation and inference but doing so would change the interpretation of θ_0.

the 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 and commonly accepted ways of dealing with unobserved heterogeneity, then allowing the Mundlak regression to have a more flexible function form should also be sensible and more robust.

Under Assumption GMD, the conditional expectations in 5.1 are functions of $(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)$. I now take a sparse approximation approach as in Section 2. Specifically, the unknown conditional expectations are approximated by a linear combination of a τ-th order polynomial transformation in the sense of Assumption ASM:

$$\mathrm{E}\left[Y_{it}|X_{it}, c_i, d_t\right] = L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\eta_Y + r_{it}^Y, \tag{5.5}$$

$$\mathrm{E}\left[D_{it}|X_{it}, c_i, d_t\right] = L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\eta_D + r_{it}^D, \tag{5.6}$$

$$\mathrm{E}\left[Z_{it}|X_{it}, c_i, d_t\right] = L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\eta_Z + r_{it}^Z, \tag{5.7}$$

where η_Y and η_D are slope parameters; r_{it}^Y, r_{it}^D and r_{it}^Z are the remainder terms or approximation errors. Furthermore, we can define a vector of transformed regressors as

$$L_{1,it} = L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t),$$

for which we denote the dimension as p, and a vector of unobserved regressors as

$$L_{2,it} = L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\backslash L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t)$$

For equation 5.5, we denote the parameters associated with $L_{1,it}$ and $L_{2,it}$ as η_{Y1} and η_{Y2}, then we have $L^{\tau}(X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\eta_Y = L_{1,it}\eta_{Y1} + L_{2,it}\eta_{Y2}$. We can define (η_{D1}, η_{D2}) and (η_{Z1}, η_{Z2}) in the same way. Let $U_{it}^Y := Y_{it} - \mathrm{E}\left[Y_{it}|X_{it}, c_i, d_t\right]$, $U_{it}^D := D_{it} - \mathrm{E}\left[D_{it}|X_{it}, c_i, d_t\right]$, and $U_{it}^Z := Z_{it} - \mathrm{E}\left[Z_{it}|X_{it}, c_i, d_t\right]$. Due to the independence of between $(\epsilon_i^c, \epsilon_t^d)$ and $\mathbf{X}$, we have $\mathrm{E}[L_{2,it}|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = \mathrm{E}[L_{2,it}]$. By defining the following stochastic error terms

$$V_{it}^Y = \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\eta_{Y2} + U_{it}^Y, \quad V_{it}^D = \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\eta_{D2} + U_{it}^D, \quad V_{it}^Z = \left(L_{2,it} - \mathrm{E}[L_{2,it}]\right)\eta_{Z2} + U_{it}^Z,$$

we have $\mathrm{E}[V_{it}^Y|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = \mathrm{E}[V_{it}^D|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = \mathrm{E}[V_{it}^z|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = 0$ and

$$Y_{it} = \mathrm{E}[L_{2,it}]\eta_{Y2} + L_{1,it}\eta_{Y1} + r_{it}^Y + V_{it}^Y, \tag{5.8}$$

$$D_{it} = \mathrm{E}[L_{2,it}]\eta_{D2} + L_{1,it}\eta_{D1} + r_{it}^D + V_{it}^D, \tag{5.9}$$

$$Z_{it} = \mathrm{E}[L_{2,it}]\eta_{Z2} + L_{1,it}\eta_{Z1} + r_{it}^Z + V_{it}^Z. \tag{5.10}$$

Again, the high-dimensional linear regression model defined by 5.8 -5.10 is viewed as an approximation.

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

It seems like one can simply apply the panel DML approach from Section 4 with the two-way cluster LASSO estimator employed as the first-step machine learner 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. Alternatively, one could assume the generalized Mundlak device holds in each sub-sample:

Assumption GMD' (Generalized Mundlak Device in Sub-Samples)**.** *For each $i \in I_k$ and $t \in S_l$ where $k = 1, ..., K$ and $l = 1, ..., L$,*

$$c_i = h_c(\bar{D}_{i,l}, \bar{X}_{i,l}) + \epsilon^c_{i,l},$$
$$d_t = h_d(\bar{D}_{t,k}, \bar{X}_{t,k}) + \epsilon^d_{t,k},$$

where $\bar{D}_{i,l} = 1/T_l \sum_{t\in S_l} D_{it}$, $\bar{D}_{t,k} = 1/N_k \sum_{i\in I_k} D_{it}$, and $(\bar{X}_{i,l}, \bar{X}_{t,k})$ are defined simiarly; $(\epsilon^c_{i,l}, \epsilon^d_{t,k})$ i.i.d random variables and are independent of $(\mathbf{D}, \mathbf{X}, \mathbf{Z})$.

Indeed, under Assumption GMD', panel-cross fitting and Theorems 4.1 - 4.3 can be used directly. However, it is soon realized that this assumption may not be plausible. In particular, the independence between $(\epsilon^c_{i,l}, \epsilon^d_{t,k})$ and $(\mathbf{D}, \mathbf{X}$ is not likely to hold: For example, for any i, c_i can be modeled by the sub-samples averages with $l = 1, 2$ as $c_i = h_c(\bar{D}_{i,1}, \bar{X}_{i,1}) + \epsilon^c_{i,1}$ and $c_i = h_c(\bar{D}_{i,2}, \bar{X}_{i,2}) + \epsilon^c_{i,2}$, and it implies that $h_c(\bar{D}_{i,1}, \bar{X}_{i,1}) - h_c(\bar{D}_{i,2}, \bar{X}_{i,2}) = \epsilon^c_{i,2} - \epsilon^c_{i,1}$ almost surely. Observe that $\epsilon^c_{i,1}$ is a function of $(\bar{D}_{i,1}, \bar{D}_{i,2}, \bar{X}_{i,1}, \bar{X}_{i,2}, \varepsilon^c_{i,2})$ and so the independence between $\epsilon^c_{i,1}$ and $\mathbf{X}$ is only possible in very special cases. Therefore, strengthening Assumption GMD to Assumption GMD' is not a desirable option.

A similar but feasible idea is to assume the unobserved heterogeneous effects are linearly additive:

$$Y_{it} = D_{it}\theta_0 + g(X_{it}) + c_i + d_t + U_{it}, \quad \mathrm{E}[U_{it}|\mathbf{X}] = \mathrm{E}[Z_{it}U_{it}] = 0. \tag{5.11}$$

In that case, the within transformation can be done in each cross-fitting sub-sample so that the demeaned version of Lemma 3.1 can be established similarly. Again, the asymptotic normality can be obtained by applying Theorem 4.1 but the validity depends on the linear function form in c_i and d_t.

Alternatively, one could maintain Assumption GMD but, for each $(i, t) \in W(-k, -l)$, approximate the full-sample temporal and cross-sectional averages through the sub-sample averages using only observations in $W(-k, -l)$. The sub-sample averages use a big portion of the full-sample so the difference between the sub-sample and full-sample averages should vanish as the sample sizes (N, T) grow. However, the temporal and cross-sectional averages $(\bar{X}_i, \bar{X}_t)$ are also high-dimensional vectors, and so the approximation error for the whole vector may not vanish fast enough, as shown below. For simplicity, we ignore the dependence of (c_i, d_t) on the low-dimensional treatment D. Let N_{-k} and T_{-l} be the cross-sectional and temporal sample sizes of the auxiliary sample $W(-k, -l)$. Define $\bar{X}_{t,-k} := 1/N_{-k} \sum_{i\in I_{-k}} X_{it}$ and $\bar{X}_{i,-l} := 1/T_{-l} \sum_{t\in S_{-l}} X_{it}$

as the sub-sample averages for each $(i,t) \in W(-k,-l)$. We can rewrite, for each $(i,t) \in W(-k,-l)$,

$$L^\tau(X_{it}, \bar{X}_i, \bar{X}_t) = L^\tau(X_{it}, \bar{X}_{i,-l} + v_{i,-l}, \bar{X}_{t,-k} + v_{t,-k})$$

where $v_{i,-l} := \bar{X}_i - \bar{X}_{i,-l}$ and $v_{t,-k} := \bar{X}_t - \bar{X}_{t,-k}$ are the vectors of approximation errors. Let $v_{i,-l,j}$ be the j-th entry of $v_{i,-l}$ for $j = 1, ..., p$, and, similarly, denote $v_{t,-k,j}$ as the j-th entry of $v_{t,-k}$. If X_{it} are independent over (i,t), then the fasted convergence rate one can obtain for those approximation errors is $v_{i,-l,j} = O_P(T^{-1/2})$ and $v_{t,-k,j} = O_P(N^{-1/2})$ for each j. Under two-way dependence, the convergence rates will be even slower than that. Define the approximation error from the sub-sample approximation as

$$\varrho_{it} := L^\tau(X_{it}, \bar{X}_{i,-l} + v_{i,-l}, \bar{X}_{t,-k} + v_{t,-k})\eta_{Y1} - L^\tau(X_{it}, \bar{X}_{i,-l}, \bar{X}_{t,-k})\eta_{Y1}.$$

Under the fastest vanishing rates and the sparsity assumption $\|\eta_{Y1}\|_0 \leq s$, we can use the property of the polynomial transformation to show that

$$\varrho_{it} = O_P\left(\frac{s}{N \wedge T}\right).$$

However, as illustrated in the proof of Theorem 2.1, to ensure the convergence rates of the two-way cluster LASSO, we need the approximation error to vanish at least as fast as $O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right)$. Therefore, the possible fasted convergence rate of the approximation error is still much slower than required.

As demonstrated above, the cross-fitting approach is in general not compatible with approaches dealing with unobserved heterogeneity in the model, including the the Mundlak device. However, without cross-fitting, it is challenging to establish an inferential theory with high dimensionality. Recall that the cross fitting is used, when establishing the asymptotic normality of the panel DML estimator, to relax the sparsity condition. Therefore, intuitively, with a more sparse model, it is possible to establish asymptotic normality in a particular model. In other words, what would be the sufficient rate requirement and its associated sparsity condition that allow for an asymptotic normality result without cross-fitting? For the rest of the section, I will give such a result based on a high-level assumption on the convergence rates of the first step estimators and I will then compare those rates to their counterparts under alternative specification, such as Assumption GMD' and linear function form in (c_i, d_t), with cross-fitting.

To simplify the notation, we rewrite 5.8 - 5.10 as

$$Y_{it} = f_{it}\beta_0 + r^Y_{it} + V^Y_{it}, \ \ \mathrm{E}[V^Y_{it}|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = 0, \tag{5.12}$$
$$D_{it} = f_{it}\pi_0 + r^D_{it} + V^D_{it}, \ \ \mathrm{E}[V^D_{it}|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = 0, \tag{5.13}$$
$$Z_{it} = f_{it}\zeta_0 + r^Z_{it} + V^Z_{it}, \ \ \mathrm{E}[V^Z_{it}|X_{it}, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t] = 0, \tag{5.14}$$

where $f_{it} := (L_{1,it}, 1)$, $\beta_0 := (\eta_{Y1}, \mathrm{E}[L_{2,it}]\eta_{Y2})'$, $\pi_0 := (\eta_{D1}, \mathrm{E}[L_{2,it}]\eta_{D2})'$, and $\zeta_0 := (\eta_{Z1}, \mathrm{E}[L_{2,it}]\eta_{Z2})'$. Recall that f_{it} includes polynomial transformations of the unit and time sample averages. To avoid the extra complexity, the asymptotic analysis will be conditional on the realized values of sample averages, as a standard practice for panel models with fixed effects.

Now we have approximated the unknown functions and unobserved random effects using a linear combination of transformations of observables. The feasible (near) Neyman-orthogonal moment function is then given by

$$\psi(W_{it};\theta,\eta) := \left(Z_{it} - f_{it}\zeta_0\right)\left(Y_{it} - f_{it}\beta_0 - \theta_0\left(D_{it} - f_{it}\pi_0\right)\right). \tag{5.15}$$

where $W_{it} := (Y_{it}, D_{it}, Z_{it}, f_{it})$ and $\eta := (\beta, \pi, \zeta)$. The estimator $\widehat{\theta}$ is then obtained by solving the sample analog of $\mathrm{E}[\psi(W_{it};\theta_0,\eta_0)] = 0$ with the nuisance parameters η_0 replaced by the first-step estimates through some high-dimensional methods. The next theorem,

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

(i) *For some* $s > 1,\ \delta > 0$, $\mathrm{E}\|D_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|Z_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|U_{it}\|^{8(s+\delta)} < \infty$, $\mathrm{E}\|V_{it}^{\iota}\|^{8(s+\delta)} < \infty$ *and* $\max_j \mathrm{E}\|f_{it,j}V_{it}^{\iota}\|^{4(s+\delta)} < \infty$, *for* $\iota = Y, D, Z$. *Moreover, there exist neighborhoods* $\mathcal{N}_m(\xi_0)$ *with* $0 < m < \infty$, *such that* $\mathrm{E}\left[\sup_{\beta\in\mathcal{N}_m(\xi_0)}|f_{it}\xi|\right]^4 < \infty$ *for* $\xi = \beta, \pi,$ and ζ.

(ii) $\lambda_{min}[\Lambda_a\Lambda_a'] > 0$ *or* $\lambda_{min}[\Lambda_g\Lambda_g'] > 0$, *where* $\Lambda_a\Lambda_a' = \mathrm{E}[a_ia_i']$, $\Lambda_b\Lambda_b' = \sum_{l=-\infty}^{\infty}\mathrm{E}[g_tg_{t+l}']$, *and* $a_i = \mathrm{E}[\psi(W_{it};\theta_0,\eta_0)|\alpha_i]$, $g_t = \mathrm{E}[\psi(W_{it};\theta_0,\eta_0)|\gamma_t]$.

(iii) $\mathrm{E}[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi_0)]$ *is non-singular.*

For brevity and simplicity, the approximation error is not considered in delivering the main theorem. Incorporating the approximation error for establishing the asymptotic normality requires extra regularity assumption but it does not substantially affect the argument.

Theorem 5.1. *Suppose, for* $P = P_{NT}$ *for each* (N,T), *(i) Assumptions AHK, AR, GMD, REG-P hold; (ii) the sparse approximation in 5.5-5.7 holds with the approximation error being 0 almost surely; and (iii) the first-step estimators* $\xi_0 = \zeta_0,\ \beta_0,$ and π_0 *obeys* $\|f_{it}(\hat{\xi}-\xi_0)\|_{NT,2} = o_P\left((N\wedge T)^{-1/2}\right)$ *and* $\|\hat{\xi}-\xi_0\|_2 = o_P(1)$. *Then, as* $N,T\to\infty$ *and* $N/T\to c$ *where* $0 < c < \infty$,

$$\sqrt{N\wedge T}V^{-1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0,1)$$

where $V := A_0^{-1}\Omega A_0^{-1}$, $A_0 := \mathrm{E}_P[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi)]$ *and* $\Omega := \Lambda_a\Lambda_a' + c\Lambda_g\Lambda_g'$ *with* Λ_a, Λ_g *defined in Assumption REG-P(ii).*

The consistency result is given as follows:

$$\widehat{V}_{\mathrm{CHS}} = \widehat{A}_{NT}^{-1}\widehat{\Omega}_{\mathrm{CHS}}\widehat{A}_{NT}^{-1\prime}, \quad \widehat{\Omega}_{\mathrm{CHS}} = \widehat{\Omega}_{\mathrm{A}} + \widehat{\Omega}_{\mathrm{DK}} - \widehat{\Omega}_{\mathrm{NW}}, \tag{5.16}$$
$$\widehat{V}_{\mathrm{DKA}} = \widehat{A}_{NT}^{-1}\widehat{\Omega}_{\mathrm{DKA}}\widehat{A}_{NT}^{-1\prime}, \quad \widehat{\Omega}_{\mathrm{DKA}} = \widehat{\Omega}_{\mathrm{A}} + \widehat{\Omega}_{\mathrm{DK}}, \tag{5.17}$$

where

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

where $\psi(W_{it}; \widehat{\theta}, \tilde{\eta}) = \left(Z_{it} - f_{it}\tilde{\zeta}\right)\left(Y_{it} - f_{it}\tilde{\beta} - (D_{it} - f_{it}\tilde{\pi})\widehat{\theta}\right)$.

Without the cross-fitting procedure, the consistency results in Theorems 4.2 and 4.3 are not applicable anymore. Therefore, we need another consistency result delivered by the following theorem:

Theorem 5.2. *Suppose assumptions for Theorem 5.1 holds for $P = P_{NT}$ for each (N, T) and $M/T^{1/2} = o(1)$. Then, $(N, T) \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

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

Theorems 5.1 and 5.2 together validate the panel DML estimation and inference procedure without cross-fitting in this partial linear model.

To summarize what we find in this section, first, with the inclusion of non-additive unobserved heterogeneity, the cross-fitting approach is in general not valid anymore. There are alternative conditions, either through strengthening the Mundlak device or restricting the function form of the unobserved heterogeneous effects. Under the alternative conditions, results for panel-DML are directly applicable, and the two-way cluster LASSO serves as the first-step estimator with desirable convergence rates. Recall that the rate requirement imposed by DML2 in Theorem 4.1 imposes a rate requirement for the first-step estimator to be $o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm, which translates to $\left\|\widehat{\xi} - \xi_0\right\| = o_P\left((N \wedge T)^{-1/4}\right)$ for $\xi = \zeta,\ \beta,\ \pi$ in this case, under regularity conditions. For the two-way cluster LASSO estimator, it imposes a sparsity condition that $s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$, which is a fairly strong sparsity condition. Without the alternative conditions regarding the unobserved heterogeneity, it is shown in Theorem 5.1 that a sufficient condition regarding the first-step estimator is $o_P\left((N \wedge T)^{1/2}\right)$ in terms of l^2 and prediction norms. As discussed in the Introduction, it is generally not achievable with the underlying component structure. The hope is that, with the help of the generalized Mundlak device, it is possible the underlying component structure could be removed. We have discussed the cases where the components are completely removed but it is unclear under an unknown function of the components. These different approaches under various scenarios are also summarized in Table 5.1. In the next section, these different methods will be put under examination through a simulation study.

Table: Summary of Results

Conditions on (c_i, d_t)	Approaches for (c_i, d_t)	First-Step Estimation	First-Step Sufficient Rates	Sparsity TW CL LASSO
Not Present	NA	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Linear	Within-Transformation	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Non-Additive	Sub-sample Mundlak	Panel CF	$o\left((N \wedge T)^{-1/4}\right)$ in $L^2(P)$ norm	$s = o\left(\frac{(N \wedge T)^{1/2}}{\log(p \vee NT)}\right)$
Non-Additive	Full-sample Mundlak	Panel CF	Not Valid	NA
Non-Additive	Full-sample Mundlak	Full Sample	$o_P\left((N \wedge T)^{-1/2}\right)$ in l^2 norm	Not Achievable

6. Monte Carlo Simulation

In this section, we examine the performance of the panel DML estimation and inference procedure in a Monte Carlo simulation study. To focus on the performance of the panel DML procedure and the LASSO estimator, the DGPs considered in this section are free of correlated random effects. Firstly, I consider the following triangular model without approximation error, simplified from the partial linear model in Section 5:

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

where $\theta_0 = 1/2$ and $\beta_0 = \pi_0 = a\,(1, 1, ..., 1, 0, ..., 0)'$ are p-dimensional parameter vectors where the first s entries are 1 and the rest of the elements are 0; a is a constant that controls the relevance of the covariates.

Secondly, I also consider the case where the true nuisance function form is unknown but it is smooth enough so it can be well-approximated by a linear combination of polynomially transformed variables:

$$\begin{aligned} \textbf{Nonlinear Model}: \ Y_{it} &= D_{it}\theta_0 + \frac{X_{it}\beta_0}{1 + \left(X_{it}\beta_0\right)^2} + U_{it}, \\ D_{it} &= \frac{X_{it}\pi_0}{1 + [\exp(X_{it}\pi_0)]^{-1}} + V_{it}, \end{aligned}$$

where the parametrization is the same as DGP(i). For DGP(ii), we pretend the nuisance functions of X_{it} are unknown, but, since they are sufficiently smooth functions of X_{it} with bounded derivatives, it can be shown by Taylor expansion that they can be well-approximated (in the sense of Assumption ASM) by a polynomial series of sufficiently large order.

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 $j = 1, ..., p$,

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

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 with support $(-\sqrt{3}, \sqrt{3})$ 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 $\iota^{|j-k|}$, $\iota \in [0, 1)$, in the (j, k)'s entry; The components γ_t^u, γ_t^v each follow a 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 $w_1 + w_2 + w_3 = 1$.

In practice, it is common to apply a within-transformation for a linear model with additive unobserved heterogeneity. In that case, if the component structure is exactly as described above, then the within-transformation not only deals with the endogeneity due to the unobserved heterogeneity but also eliminates the two-way dependence introduced by the linear components. To illustrate it is not necessarily the case in general and verify the theory under a different scenario, I also consider a multiplicative component structure DGP that features two-way dependence in the moment function even after the within-transformation:

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

where the components are generated the same way as the Linear Components. The weights $(w_1, w_2, w_3, w_4, w_5)$ are non-negative with $w_1^2 + w_2^2 + w_3^2 = 1$ and $w_4^2 + w_5^2 = 1$. c_p is a scaling factor that ensures the sums of multiplicative components in both U_{it} and V_{it} are variance 1.

The multiplicative components construction here is a generalization of the example in Chiang et al. (2024) where it is used to illustrate that the two-way within-transformation of the original linear panel model may not eliminate the underlying components. To see why $U_{it}V_{it}$ features a component structure, we can expand the product and observe that it includes terms like $\alpha_i^u\alpha_i^v\gamma_{t,j}^2$ 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 $\gamma_t^u\gamma_t^v$. We can also show that $X_{it,j}U_{it}$ and $X_{it,j}V_{it}$ possess a components structure in a similar way. 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. 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 heteroskedasticity-robust LASSO from Belloni et al. (2012), the square-root LASSO from Belloni et al. (2011), the cluster-robust LASSO from Belloni et al. (2016), and the two-way cluster-LASSO with CHS-type and DKA-type feasible penalty weights. 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 from the current version. Theoretically, among the candidates, only the approaches using cross-fitting and with the first step through the two-way cluster-LASSO estimator are guaranteed to produce asymptotically normal estimators and have correct size when CHS/DKA type variance estimators are used for inference. As we will see, cross-fitting indeed improves the sample coverage probability across all approaches with different first-step estimators but approaches without using cross-fitting also perform reasonably well when the model is less dense.

The simulation results are based on 500 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. The estimation is based on the orthogonal moment condition given as follows by 5.15 with $Z_{it} = D_{it}, f_{it} = X_{it}, \zeta_0 = \pi_0$. 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 the two-way cluster-LASSO estimation, the sufficiently large constant C_λ, the $o(1)$ term γ, and the number of iterations m for the feasible weights are needed. For both variance estimation and feasible weight 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 $\widehat{\rho}$ is the OLS estimator from the regression $\bar{\hat{v}}_t = \rho\bar{\hat{v}}_{t-1} + \eta_t$ where $\bar{\hat{v}}_t = \frac{1}{N}\sum_{i=1}^{N}\hat{v}_{it}$. For variance estimation, $\hat{v}_{it} = \hat{U}_{it}\hat{V}_{it}$; For feasible weights estimation, $\hat{v}_{it} = x_{it}\hat{U}_{it}$ or $\hat{v}_{it} = x_{it}\hat{V}_{it}$ where x_{it} is a generic scalar regressor.

Table 6.1: Linear Model with Additive Components
with $N = T = 30$, $s = 5$, $p = 200$, $\iota = 0.5$, $\rho = 0.75$, $a = 0.5$

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	Sel. D	Second-Step Bias	SD	RMSE	Coverage (%) CHS	DKA
No	POLS	200	200	0.001	0.047	0.047	78.6	94.4
	H LASSO	24.3	24.7	0.056	0.058	0.080	63.4	82.0
	R LASSO	20.1	20.6	0.057	0.060	0.082	68.0	83.0
	C LASSO	8.8	8.7	0.042	0.085	0.095	80.0	86.0
	TW LASSO CHS	5.0	4.9	0.005	0.107	0.108	82.8	87.4
	TW LASSO DKA	5.0	4.7	0.009	0.108	0.108	81.4	87.0
Yes	POLS	200	200	0.001	0.128	0.128	97.6	98.4
	H LASSO	15.5	15.7	0.057	0.139	0.150	93.6	95.8
	R LASSO	11.7	11.9	0.054	0.141	0.151	94.6	96.4
	C LASSO	5.5	5.7	0.008	0.152	0.152	93.4	95.0
	TW LASSO CHS	5.0	4.7	0.012	0.159	0.159	90.4	93.6
	TW LASSO DKA	4.9	4.4	0.026	0.158	0.160	89.6	92.8

Note: Simulation results are based on 500 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2.1$, $m = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO with CHS and DKA labeled for the feasible penalty weights. Post-LASSO POLS is performed in all first steps. Based on the sample correlation with the outcome, the 5 most relevant regressors are used to obtain the initial feasible penalty weights for all three LASSO approaches. Nominal coverage probability: 0.95.

Table 6.1 presents a set of baseline results that are obtained for a reasonably large 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. Still, the number of non-zero coefficients can be regarded as a small order of the sample sizes, 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, we take a look at the rows without using cross-fitting. It is shown that the proposed two-way cluster-LASS, either performed with CHS or DKA feasible weights, selects almost perfectly while other LASSO approaches all over-select to a different degree. Among the LASSO-based approaches, the proposed methods suffer from the least finite sample bias. However, the proposed methods have slightly larger variability in terms of Monte Carlo SD and RMSE. For sample coverage, both CHS- and DKA-type variance estimators are used for all methods. In terms of variance estimation and inference, all high-dimensional methods suffer from under-coverage slightly while the proposed methods have the least size distortion. This can result from a strong serial correlation in the time effects ($\rho = 0.75$) and can also be explained by the fact that under high-dimensionality and two-way cluster dependence, the asymptotic normality is not guaranteed without cross-fitting, as discussed in Section 2. Surprisingly, POLS performs slightly better than high-dimensional methods for a reasonably high-dimensional model. However, as we will see later, it is not true anymore when the dimension of covariates grows further.

Table 6.2: Linear Model with Additive Components
with $N = T = 30$, $s = 5$, $p = 800$, $\iota = 0.5$, $\rho = 0.75$, $a = 0.5$

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	Sel. D	Second-Step Bias	SD	RMSE	Coverage (%) CHS	DKA
No	POLS	800	800	0.004	0.097	0.097	47.6	65.8
	H LASSO	39.1	39.6	0.066	0.043	0.078	47.8	74.4
	R LASSO	31.9	32.2	0.068	0.048	0.083	49.2	70.2
	C LASSO	13.4	13.8	0.058	0.078	0.097	66.6	76.2
	TW LASSO CHS	5.0	4.8	0.011	0.108	0.108	79.8	86.6
	TW LASSO DKA	5.0	4.6	0.017	0.113	0.114	78.6	85.4
Yes	H LASSO	24.6	25.7	0.029	0.151	0.154	91.8	94.2
	R LASSO	18.1	18.3	0.060	0.133	0.146	94.0	97.0
	C LASSO	6.5	6.5	0.011	0.151	0.151	93.2	95.6
	TW LASSO CHS	5.1	4.7	0.021	0.157	0.158	89.0	94.0
	TW LASSO DKA	4.9	4.2	0.039	0.154	0.159	88.4	92.0

Note: Simulation results are based on 500 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2.1$, $m = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO with CHS and DKA labeled for the feasible penalty weights. Post-LASSO POLS is performed in all first steps. Based on the sample correlation with the outcome, the 5 most relevant regressors are used to obtain the initial feasible penalty weights for all three LASSO approaches. 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, which is not predicted by the theory. As a cost, we can see from the increased Monte Carlo SD that there is a loss of efficiency by doing cross-fitting, as expected. 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 is as large as the overall sample size, a different pattern is revealed. Table 6.2 considers the same DGP as Table 6.1 except that the dimension p now increases to 800, slightly smaller than the sample size 900. In principle, we can compare a case with the dimension of covariates larger than the sample size, but then it is not possible to compare the results based on the POLS first-steps. First, we compare the results without cross-fitting. The simulation results demonstrate that the methods based on the POLS first-steps 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 almost perfectly regardless of the increased number of irrelevant regressors. Again, when cross-fitting is performed, there is a significant improvement across all approaches in terms of the sample coverage. This is expected: as shown in the proof of Theorem 5.2, the error term R^2_{NT} contains terms such as $V^D_{it}(\beta_0 - \tilde{\beta})$ that are not mean 0 and may not vanish fast enough; this is more severe when no selection or inconsistent selection is performed in the first stage. Through the proper cross-fitting scheme, such terms are conditionally mean-zero even with over-selection. Such property of cross-fitting methods is thus translated into better coverages across all approaches considered here.

Table 6.3: Nonlinear Model with Additive Components with $N = T = 30$, $s = 5$, $p = 10$, $\iota = 0.5$, $\rho = 0.75$, $a = 0.33$

Cross Fitting	First-Step Estimator	First-Step Ave. Sel. Y	Sel. D	Second-Step Bias	SD	RMSE	Coverage (%) CHS	DKA
No	POLS	494	494	0.005	0.100	0.100	71.8	79.8
	H LASSO	9.5	17.5	0.01	0.100	0.101	83.0	87.6
	R LASSO	8.2	15.3	0.011	0.100	0.101	81.8	88.0
	C LASSO	6.3	13.0	0.001	0.101	0.101	83.8	89.6
	TW LASSO CHS	4.9	10.2	0.006	0.105	0.106	81.6	87.2
	TW LASSO DKA	2.2	6.7	0.013	0.105	0.106	82.6	87.6
Yes	POLS	494	494	0.011	0.166	0.166	99.2	99.8
	H LASSO	7.3	13.5	0.037	0.159	0.163	88.4	93.0
	R LASSO	5.2	10.4	0.037	0.157	0.162	87.6	92.8
	C LASSO	4.9	9.2	0.027	0.148	0.150	91.4	93.6
	TW LASSO CHS	7.3	12.4	0.028	0.150	0.152	91.2	93.6
	TW LASSO DKA	3.6	8.0	0.027	0.149	0.151	91.0	93.6

Note: Simulation results are based on 500 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2.1$, $m = 5$, and $\gamma = 0.1/\log(p \vee N \vee T)$. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO with CHS and DKA labeled for the feasible penalty weights. Post-LASSO POLS is performed in all first steps. Based on the sample correlation with the outcome, the 5 most relevant regressors are used to obtain the initial feasible penalty weights for all three LASSO approaches. Nominal coverage probability: 0.95.

We have seen the case with exact sparsity in Tables 6.1 and 6.2. As argued in the theory, the proposed estimation and inference procedures are also valid under approximate sparsity. In Table 6.3, we consider a case where 10 observable control variables are considered by the researcher while only 5 of those are relevant.

Those 5 relevant controls enter the model through unknown but smooth nuisance functions as described above. The researchers are not aware of the irrelevance of the other five controls and neither do researchers have knowledge of the nuisance function forms. However, the smoothness allows for the approximation of the nuisance function using polynomials series. In particular, we use the 3rd-order polynomial transformation of the 10 observable controls for approximation. Due to bounded derivatives of the smooth nuisance functions and the irrelevance of some of the controls, the sparse approximation is valid. As is shown in Table 6.3, compared to previous sets of results, the advantage of the proposed methods is not obvious in this case. From the number of selected regressors and the Monte Carlo SD across all methods, it seems the dependence is not as strong as in previous DGPs. Given the small number of Monte Carlo replications, the differences in sample coverage probabilities across different approaches are not really distinguishable. On the other hand, the improvement made by the use of the cross-fitting remains significant.

Table 6.4: Linear Model with Multiplicative Components
with $N = T = 30$, $s = 5$, $p = 200$, $\iota = 0.5$, $\rho = 0.75$, $a = 0.5$

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	0.001	0.139	0.139	77.0	81.6
	H LASSO	5.8	5.9	0.002	0.135	0.135	83.8	89.4
	R LASSO	5.6	5.6	0.001	0.135	0.135	84.2	89.2
	C LASSO	5.9	6.1	0.001	0.137	0.137	84.2	89.6
	TW LASSO CHS	12.2	12.6	0.001	0.137	0.137	83.6	89.2
	TW LASSO DKA	5.1	5.0	0.001	0.138	0.138	83.4	88.6
Yes	POLS	200	200	0.004	0.148	0.148	91.8	94.6
	H LASSO	6.0	6.0	0.003	0.149	0.149	95.0	97.2
	R LASSO	5.4	5.4	0.002	0.149	0.149	94.8	97.6
	C LASSO	6.2	6.1	0.013	0.152	0.153	94.0	96.6
	TW LASSO CHS	12.7	12.2	0.024	0.152	0.153	92.6	96.2
	TW LASSO DKA	5.1	4.8	0.023	0.154	0.155	92.4	95.4

Note: Simulation results are based on 500 replications. Tuning parameters: $(K, L) = (4, 8)$, $C_\lambda = 2.1$, $m = 2$, and $\gamma = 0.1/\log(p \vee N \vee T)$. H: heteroskedastic-LASSO; R: square-root-LASSO; C: cluster-LASSO; TW: two-way cluster-LASSO with CHS and DKA labeled for the feasible penalty weights. Post-LASSO POLS is performed in all first steps. Based on the sample correlation with the outcome, the 5 most relevant regressors are used to obtain the initial feasible penalty weights for all three LASSO approaches. Nominal coverage probability: 0.95.

Lastly, I use the multiplicative components to generate two-way cluster dependence and compare the results in Table 6.4. There are two main differences compared to previous results. First, the two-way dependence is present but relatively weak. In some unreported simulation results using variance estimators not accounting for two-way dependence, all methods suffer from under-coverage issues, implying the presence of two-way dependence. Additionally, we can see the over-selection problem is almost not present for heteroskedasticity-LASSO, square-root-LASSO, and cluster-LASSO, also indicating the dependence is quite weak. Secondly, notice that the two-way cluster LASSO with CHS-type penalty weights over-selects, in a more severe way than existing approaches not accounting for two-way dependence. Further examination of the issue reveals that it is due to the finite-sample bias of the HAC-type formula: it is well known that HAC-type variance estimators are not guaranteed to be positive-semi definite. Under certain DGPs, the HAC-type estimator is more likely to be negative or close to 0, causing large finite sample bias in variance estimation. This is likely to be the case here. Indeed, some penalty weights estimated using the CHS formula are negative and replaced by 0. There are other finite sample adjustments available but in general, it is a disadvantage for the CHS-type penalty weights. The DKA-type feasible weights are guaranteed to be positive and it has been shown in the literature that the two-term formulas for variance estimation tend to have better finite sample performance (Chen and Vogelsang (2024)). However, it is also well-known that the two-term formulas for variance estimation suffer from an overestimation problem when the true DGP is i.i.d over both clusters or is only clustered at the intersection of the two clusters (MacKinnon et al. (2021)). In practice, it is very rare to encounter panel data not featuring any dependence across space or time. As a result, it is in general recommended to use the DKA-type formula for both penalty weights and variance estimation.

7. 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 does not even need a control for identification, 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 in settings to which researchers don't usually relate high dimensionality.

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.[14] To focus on the main idea, the shares are taken as given and the resulting instrument variable is treated as observable instead of generated regressors to avoid further complication.

In this paper, 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:

Original Linear Model	Partial Linear Model
$Y_{it} = \theta_0 D_{it} + \pi_i W_t + c_i + d_t + U_{it}$	$Y_{it} = \theta_0 D_{it} + g(X_{it}, W_t, c_i, d_t) + U_{it}$
$D_{it} = \alpha_0 Z_{it} + \beta_i W_t + c_i + d_t + V_{it}$	

where θ_0 is the parameter of interest, i.e. the true multiplier, and α_0 is the scalar parameter associated with Z_{it}.; X_{it} and W_t are exogenous control variables with the latter being only time-varying; π_i and β_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 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 5. Specifically, c_i is assumed to be a function of $\bar{X}_i$ and d_t is assumed to be a function of $(\bar{X}_t, W_t)$. Then, through sparse approximation, the feasible (near) Neyman-orthogonal moment function is given by

$$\left(Z_{it} - f_{it}\zeta_0\right)\left(Y_{it} - f_{it}\beta_0 - \theta_0\left(D_{it} - f_{it}\pi_0\right)\right)$$

where $f_{it} = (L^{\tau}(X_{it}, W_t, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t), 1)$ and (β, π, ζ) are associated slope coefficients defined the same way as in 5.12-5.14.

In the original study, W_t are not included in the baseline model. In the 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[15], one control would increase 51 parameters and two controls would generate 102 parameters, given no interactions or higher order terms. With a sample size less than 2000, the high dimensionality in nuisance parameters could result in a noisy estimate of θ_0. In this

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

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

paper, to obtain a more precise estimate and make the excludability assumption of the IV to be more plausible, besides the controls from the original study, I also consider additional controls. For X_{it}, I include the change in state population. As is shown in Table 3 of Nakamura and Steinsson (2014), the state population is likely not affected by the treatment (the regional military spending), so it is immune to the "bad control" problem[16]; But it could affect the treatment and the outcome. By considering 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 selection methods. Moreover, state-level yearly variables of those macroeconomic characteristics are often considered to be cluster-dependent in both space 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.

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 7.1: 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: The data is a balance panel with $(N, T) = (51, 39)$. Standard errors and feasible penalty weights are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 6.

Table 7.1 provides benchmark results for the original model with different choices of control variables. All estimates (columns 6 and 9) 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 5. The estimates of the multiplier are matched with those given in Nakamura and Steinsson (2014) with significant differences in the standard

[16] Angrist and Pischke, 2009, Frölich, 2008, and Chen et al., 2024 provide detailed discussions on when endogenous control pollute the identification/estimation and when they are innocuous.

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.

Table 7.2: 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	0.70	0.84
				C LASSO	4	1.43	0.66	0.81
				TW LASSO DKA	1	1.47	0.61	0.77
No	Mundlak	2nd	35	POLS	35	1.73	0.99	1.15
				H LASSO	6	1.73	1.03	1.19
				CR LASSO	5	1.40	0.70	0.86
				TW LASSO DKA	3	1.47	0.61	0.77
No	Mundlak	3rd	119	POLS	119	2.20	1.19	1.37
				H LASSO	9	2.00	1.18	1.39
				CR LASSO	6	0.97	0.65	0.80
				TW LASSO DKA	5	1.47	0.61	0.77

Note: The data is a balance panel with $(N, T) = (51, 39)$. The tuning parameters are chosen as $m = 10$, $C_\lambda = 2.1$, $\gamma = 0.1/\log(p \vee N)$. The control variables and the Mundlak sample averages $(X_{it}, W_t, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t)$ are used to obtain the initial feasible penalty weights for all three LASSO approaches. 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 and feasible penalty weights are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 6.

The main comparisons are done in Tables 7.2 and 7.3. In Table 7.2, no cross-fitting is performed in the first stage. The number of parameters associated with regressors generated by polynomials transformations are reported in column (4) and the number of selected parameters associated with Z are reported in column (6)[17]. Overall, with more controls and the polynomial transformation of the observables, the standard errors are generally larger than those in 7.1. 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

[17] Across all first-step LASSO approaches, more parameters associated with Z are selected compared to those associated with Y and X. The difference in the LASSO selection is less evident for Y and X while the pattern is similar.

terms under potential two-way dependence, the proposed method is robust against over-selection.

Table 7.3: 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.
Yes	Mundlak	None	7	H LASSO	2.0	1.45	1.66	1.93
				C LASSO	2.7	1.32	1.75	2.03
				TW LASSO DKA	2.0	1.51	1.29	1.56
Yes	Mundlak	2nd	35	H LASSO	5.1	0.98	2.10	2.43
				C LASSO	5.5	1.12	1.90	2.18
				TW LASSO DKA	4.6	1.01	1.22	1.47
Yes	Mundlak	3rd	119	H LASSO	8.2	2.05	2.92	3.46
				C LASSO	6.5	1.23	1.57	1.88
				TW LASSO DKA	6.0	1.09	1.18	1.46

Note: The data is a balance panel with $(N, T) = (51, 39)$. The tuning parameters are chosen as $(K, L) = (4, 8)$, $m = 10$, $C_\lambda = 2.1$, $\gamma = 0.1/\log(p \vee N \vee T)$. The control variables and the Mundlak sample averages $(X_{it}, W_t, \bar{D}_i, \bar{D}_t, \bar{X}_i, \bar{X}_t)$ are used to obtain the initial feasible penalty weights for all three LASSO approaches. Number of predictors generated by the polynomial transformation and the average (over cross-fitting sub-samples) number of selected predictors for Z are reported in columns (4) and (6). Standard errors and feasible penalty weights are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 6.

For more robustness in inference, as is shown in both theoretical and simulation results, I further consider different methods implemented with cross-fitting in Table 7.3[18]. It shows a similar pattern as in Table 7.2: 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. What's going on here could be potential nonlinearity in the true model unknown to the researcher. When no transformation or the 2nd-order polynomial transformation is used, a limited amount of nonlinearity is captured in the misspecified model. With the 3rd-order polynomial approximation, while the nonlinearity might be captured by all three methods in comparison, as revealed by the increased number of selected regressors, the proposed two-way cluster LASSO achieves that with fewer and likely more accurate selections. Unfortunately, although the cross-fitting approach is asymptotically valid and has better size control as shown in the simulation, it is in the cost of efficient loss, as told by the larger standard errors.

8. 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

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

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 in two-way cluster dependence and unobserved heterogeneity. I first develop a weighted LASSO approach that is robust to two-way cluster dependence in the panel data. As is shown in the asymptotic analysis of the two-way cluster LASSO, the convergence rates are quite 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 valid first-step estimator for inference purposes in a high-dimensional semiparametric model. Individually, both the two-way cluster-LASSO and the cross-fitting can be of independent interest; Together, they extend the DML approach to panel data settings. Furthermore, I study the subtle issues of cross-fitting and unobserved heterogeneity in panel models and provide strategies under various scenarios.

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 there is no reason not to apply the approaches in this paper. 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. First, the Mundlak device and, in general, many other approaches for dealing with unobserved heterogeneity are not compatible with the cross-fitting schemes due to the dependence introduced by the full history of the data that is used for modeling the unobserved heterogeneous effects. On the other hand, it is generally not feasible to establish the inferential theory without cross-fitting in high-dimensional semi-parametric models. In that sense, the cross-fitting approach is naturally limited in use for panel data models. Secondly, the feasible penalty weight estimation is highly non-trivial due to two-way cluster dependence and high dimensionality. The asymptotic 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, there are many subtle issues that are lack of theoretical guarantee. A devoted exploration of such issues requires a more comprehensive treatment and is another 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.

Arellano, M., 1987. Computing robust standard errors for within-groups estimators. Oxford Bulletin of Economics and Statistics 49, 431–434.

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., 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.

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

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

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.

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.

Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. Review of Economics and Statistics 80, 549–559.

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.

Frölich, M., 2008. Parametric and nonparametric regression in the presence of endogenous control variables. International Statistical Review 76, 214–227.

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.

Hahn, J., Newey, W., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72, 1295–1319.

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. URL: `https://arxiv.org/abs/2208.13701`, `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.

Newey, W.K., McFadden, D., 1994. Large sample estimation testing. Handbook of Econometrics 4, 2113–2245.

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. URL: https://doi.org/10.1214/16-AOS1448, doi:10.1214/16-AOS1448.

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.

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., 2005. Nonlinear system theory: Another look at dependence. Proceedings of the National Academy of Sciences 102, 14150–14154.

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

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 $j = 1, ..., p$:

$$
\begin{aligned}
&\mathrm{P}\left(\frac{1}{NT}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1NT}\right) = \mathrm{P}\left(\omega_j^{-1/2}\left|\frac{1}{N}\sum_{i=1}^{N}a_{i,j} + \frac{1}{T}\sum_{t=1}^{T}g_{t,j} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it,j}\right| > \frac{\lambda}{2c_1NT}\right)\\
\leq &\mathrm{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\omega_{a,j}^{-1/2}a_{i,j}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}\omega_{g,j}^{-1/2}g_{t,j}\right| + \left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}e_{it,j}\right| > \frac{\lambda}{2c_1NT}\right)\\
\leq &\mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{N}\sum_{i=1}^{N}\omega_{a,j}^{-1/2}a_{i,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right) + \mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{T}\sum_{t=1}^{T}\omega_{g,j}^{-1/2}g_{t,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right)\\
&+\mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}e_{it,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right) := p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)
\end{aligned}
$$

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 equality follows from the Hajek projection and it is definitional. 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 third inequality follows from a union-bound inequality. Then, by the union-bound inequality again, we have

$$
\mathrm{P}\left(\max_{j=1,...,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1NT}\right) \leq \sum_{j=1}^{p}\left[p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)\right]
$$

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

$$
\Xi_{a,j} = \frac{\left[\mathrm{E}(a_{i,j})^2\right]^{1/2}}{\left[\mathrm{E}(a_{i,j})^3\right]^{1/3}}.
$$

Let $l_{a,NT}$ be some positive increasing sequence. Without loss of generality, we assume $N \wedge T = N$ from now on. By Theorem 7.4 of Peña et al. (2009) with $\delta = 1$, we have for any $x \in [0, N^{1/6}\Xi_{a,j}/l_{a,NT}]$ that

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

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

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

given that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in [0, N^{1/6}\Xi_{a,j}/l_{a,NT}]$ for all $j=1,...,p$. And so we have

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

To show the right-hand-side converges to 0 as $\gamma \to 0$ and $(N,T) \to \infty$, there should exist an increasing sequence $l_{a,NT}$ such that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in [0, N_k^{1/6}\min_j\{\Xi_{a,j}\}/l_{a,NT}]$. Under Assumption REG and the assumption that $N/T \to c$, we have $\log(p \vee NT) = o(N^{1/3})$ as $(N,T) \to \infty$. With the choice $\log(1/\gamma) \lesssim \log(p \vee NT)$, we have

$$\begin{aligned} N^{1/6}\min_j\{\Xi_{a,j}\}/l_{a,NT} &= O(N^{1/6}/l_{a,NT}) \\ \Phi^{-1}\left(1-\frac{\gamma}{2p}\right) &\lesssim \sqrt{\log(p/\gamma)} \lesssim \sqrt{\log p + \log(p \vee NT)} = o(N^{1/6}) \end{aligned}$$

Therefore, such positive sequence $l_{a,NT} \to \infty$ indeed exists and so we conclude $\sum_{j=1}^{p} p_{1,kl}(\lambda) \to 0$ 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

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

Then, by Theorem 3.2 of Gao et al. (2022) with $\tau = 1$ and $\alpha = \frac{1}{1+2\tau}$, we have

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

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

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

given that $\Phi^{-1}(1-\frac{\gamma}{2p}) \in \big(0, d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT}\big)$. And so we have

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

To show the right-hand-side converge to 0 as $\gamma \to 0$ and $(N,T) \to 0$, there should exists an increasing sequence $l_{g,NT}$ such that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in (0, d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT})$. Under Assumption REG, we have $\log(p \vee NT) = o(T^{1/6}/\log T)$. Under the choice $\log(1/\gamma) \lesssim \log(p \vee NT)$, we have

$$\begin{aligned} d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT} &= O(T^{1/12}/(\log T)^{1/2}/l_{g,NT}) \\ \Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \simeq \sqrt{\log(p/\gamma)} \simeq \sqrt{\log p + \log(p \vee NT)} &= o(T^{1/12}/(\log T)^{1/2}). \end{aligned}$$

Therefore, such positive sequence $l_{g,NT} \to \infty$ indeed exists and so we conclude $\sum_{j=1}^{p} p_{2,j}(\lambda) \to 0$ as $\gamma \to 0$ and $(N,T) \to \infty$.

To bound $p_{3,j}(\lambda)$, first notice that by the same arguments made in the Proof of Claim C.3 that shows $\|\Lambda_e\Lambda_e'\| < \infty$ and $\mathrm{Var}\left(\frac{1}{N^{1/2}T}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\right) = O(1/T)$, we can show, under Assumption AHK, AR, ASM, REG(iii), that

$$\mathrm{Var}\left(\frac{1}{N^{1/2}T}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it,j}\right) = O(1/T)$$

Therefore, by Chebyshev's inequality and that $p = o(T^{7/6}/\log T)$, we have

$$\begin{aligned} p_{3,j}(\lambda) =& \mathrm{P}\left(\left|\frac{1}{N^{1/2}T}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}e_{it,j}\right| > \frac{\lambda}{6c_1N^{1/2}T}\right) \\ \leq& \mathrm{P}\left(\left|\frac{\underline{\omega}^{-1/2}}{N^{1/2}T}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it,j}\right| > \Phi^{-1}\left(1-\frac{\gamma}{2p}\right)\right) \\ \leq& \underline{\omega}^{-1}\mathrm{Var}\left(\frac{1}{N^{1/2}T}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it,j}\right)/\Phi^{-2}\left(1-\frac{\gamma}{2p}\right) \\ =& O(1/T)/\Phi^{-2}\left(1-\frac{\gamma}{2p}\right) \simeq O(1/T)/(T^{1/6}/\log T), \\ \sum_{j=1}^{p} p_{3,j}(\lambda) \simeq& O(p\log T/T^{7/6}) = o(1). \end{aligned}$$

Put together, we have shown, for all (k, l),

$$\mathrm{P}\left(\max_{j=1,\dots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} f_{it,j} V_{it}\right| \leq \frac{\lambda}{2c_1 NT}\right) \to 1. \tag{8.1}$$

Secondly, we will apply Lemma 6 of Belloni et al. (2012) to obtain the finite sample bounds on

$$\begin{aligned}\left\|\mathbf{f}\left(\hat{\zeta}-\zeta_0\right)\right\|_{NT,2} &= \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(f_{it}\hat{\zeta}-f_{it}\zeta_0\right)^2\right)^{1/2}, \\ \left|\omega^{1/2}\left(\hat{\zeta}-\zeta_0\right)\right|_1 &= \left|\sum_{j=1}^{p}\omega_j^{1/2}\left(\hat{\zeta}_j-\zeta_{0,j}\right)\right|.\end{aligned}$$

Let δ be some generic vector of nuisance parameters and let J_p^1 be a subset of an index set $J_p = 1, \dots, p$ 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:

$$\begin{aligned}K_C(M_f) &= \min_{\delta:\|\delta^0\|_1 \leq C\|\delta^1\|_1,\ \|\delta\|\neq 0,\ |J_p^1|\leq s} \frac{\sqrt{s\delta' M_f \delta}}{\left\|\delta^1\right\|_1}, \\ M_f &= \frac{1}{NT}\sum_{i\in I_k, t\in T_l} f_{it} f_{it}'.\end{aligned}$$

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

$$K_C^{\omega}(M_f) \geq \frac{1}{b} K_{bC/a}(M_f). \tag{8.2}$$

By results in Chiang et al. (2024), we have, as $N, T \to \infty$ jointly,

$$\omega_j \xrightarrow{p} \frac{N \wedge T}{N}\lambda_{a,j}^2 + \frac{N \wedge T}{T}\lambda_{g,j}^2 \quad \text{for each } j = 1, \dots, p.$$

By Theorem 2 of Chiang et al. (2024), we have $|\lambda_{a,j}^2| < \infty$ and $|\lambda_{g,j}^2| < \infty$. By the non-degeneracy assumption, we have either $\lambda_{a,j}^2 > 0$ or $\lambda_{g,j}^2 > 0$. Therefore, we have ω_j bounded below by zero and bounded above for each $j = 1, .., p$ with probability approaching one as $N, T \to \infty$.

Under Assumption (ASM), the condition 2.9, and 8.1, Lemma 6 of Belloni et al. (2012) implies that

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

where $c_0 := \frac{uc+1}{lc-1} > 1$. By 8.2, 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,

$$
\begin{aligned}
\left\|\mathbf{f}\left(\widehat{\zeta}-\zeta_0\right)\right\|_{NT,2} =& O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right),\\
\left|\omega^{1/2}\left(\widehat{\zeta}-\zeta_0\right)\right|_1 =& O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}\right),
\end{aligned}
$$

Then, applying the results above gives

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

where the first inequality follows from the Holder's.

The $l2$ rate of convergence will be derived after the sparsity bounds. We now switch the focus to the Post-LASSO. The proof follows closely Steps 5-8 of Theorem 1 in Belloni et al. (2016). 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

$j \in \Gamma$. We want to show, first, under Assumptions AHK, ASM, AR, REG,

$$\begin{aligned}\|f(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} =& O_P\left(\sqrt{\frac{s\log(p\vee NT)}{(N\wedge T)\phi_{min}(s)(M_f)}} + \sqrt{\frac{\widehat{m}\log(p\vee NT)}{(N\wedge T)\phi_{min}(\widehat{m})(M_f)}}\right) \quad (8.3)\\ &+ O_P\left(\|f(X_{it}) - \left(\mathcal{P}_{\widehat{\Gamma}}f\right)_{it}\|_{NT,2}\right),\\ \|f_{it}(\widehat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \leq& \|f_{it}(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} + \|r_{it}\|_{NT,2}, \quad (8.4)\\ \|\omega^{1/2}(\widehat{\zeta}_{PL} - \zeta_0)\|_1 \leq& \frac{b\sqrt{\widehat{m}+s}}{\sqrt{\phi_{min}(\widehat{m}+s)(M_f)}} \times \|f_{it}(\widehat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \quad (8.5)\end{aligned}$$

Let $S_j := 2\mathbb{E}_{NT}[\omega^{-1/2}f_{it,j}V_{it}]$. Furthermore, under $c_1 \max_{1\leq j\leq p}|S_j| \leq \frac{\lambda}{NT}$ and $u \geq 1 \geq l \geq 1/c_1$, we want to show

$$\|f(X_{it}) - \mathcal{P}_{\widehat{\Gamma}}f(X_{it})\|_{NT,2} \leq \left(u + \frac{1}{c_1}\right)\frac{\lambda\sqrt{s}}{NTK^{\omega}_{c_0}(M_f)} + 3\|r_{it}\|_{NT,2}. \quad (8.6)$$

First, note that

$$f(X_{it}) - f_{it}\widehat{\zeta}_{PL} = f(X_{it}) - (\mathcal{P}_{\widehat{\Gamma}}Y)_{it} = \left((I_{NT} - \mathcal{P}_{\widehat{\Gamma}})f(X) - \mathcal{P}_{\widehat{\Gamma}}V\right)_{it} = \left((I_{NT} - \mathcal{P}_{\widehat{\Gamma}})f - (\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0} + \mathcal{P}_{\Gamma_0})V\right)_{it},$$

where the last inequality follows from the property of the linear projection. Then, by Minkowski inequality, we have

$$\|f(X_{it}) - f_{it}\widehat{\zeta}_{PL}\|_{NT,2} \leq \left\|\left(I_{NT} - \mathcal{P}_{\widehat{\Gamma}})f\right)_{it}\right\|_{NT,2} + \left\|\left(\mathcal{P}_{\Gamma_0}V\right)_{it}\right\|_{NT,2} + \left\|\left(\mathcal{P}_{\widehat{\Gamma}\backslash\Gamma_0}V\right)_{it}\right\|_{NT,2}. \quad (8.7)$$

We will proceed by bounding the last two terms above. First, consider the last term. By Hölder's inequality and the property of spectral norm, we have

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

We have shown previously that $b = \max_{j=1,\dots,p}\omega_j^{1/2} < \infty$, so

$$\frac{1}{NT}\max_{j=1,\dots,p}\left|\sum_{i=1}^N\sum_{t=1}^T f_{it,j}V_{it}\right| \leq b/2\max_{j=1,\dots,p}\left|S_j\right| = O_P(\lambda/NT) = O_P\left(\frac{\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)}{\sqrt{N\wedge T}}\right) = O_P\left(\sqrt{\frac{\log(p\vee NT)}{N\wedge T}}\right).$$

Next, consider the middle term.

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

Combining the bounds for the last two terms and 8.7, we obtain the inequality 8.3. The inequality 8.4 follows directly from the triangle inequality. The inequalities 8.5 and 8.5 follow from Lemma 7 of Belloni et al. (2012) since there are finite sample bounds, not affected by the panel data dependence.

Next, to obtain the sparsity bound for the two-way cluster LASSO estimator, we first apply Lemma 8 of Belloni et al. (2012) to obtain that

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

where $a > 0$ has been shown previously.

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

$$
\widehat{m} \leq \min_{m\in\mathcal{M}}\phi_{max}(m\wedge NT)(M_f)a^{-2}\left(\frac{2c_0\sqrt{s}}{K_{c_0}^{\omega}(M_f)} + \frac{6c_0NT\|r_{it}\|_{NT,2}}{\lambda}\right)^2. \tag{8.8}
$$

Note that $\frac{6c_0NT\|r_{it}\|_{NT,2}}{\lambda\sqrt{s}} = O_P(1/\log(p\wedge NT)) \xrightarrow{p} 0$. Recall that $1/K_{c_0}^{\omega}(M_f) \leq b/K_{\bar{C}}(M_f) < \infty$. Let $\mu := \min_m\left\{\sqrt{\phi_{max}(m)(M_f)/\phi_{min}(m)(M_f)} : m > 18\bar{C}^2 s\phi_{max}(m)(M_f)/K_{\bar{C}}^2(M_f)\right\}$, 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 8.8. By Lemma 9 (the sub-linearity of sparse eigenvalues) from Belloni et al. (2012) and 8.8, we have

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

Recall that $c_1\max_{1\leq j\leq p}|S_j| \leq \frac{\lambda}{NT}$ with probability approaching one, so we can combine the results above with 8.3 and 8.6 to obtain

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

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

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

For the $l1$ convergence rate, note that $\|\widehat{\zeta}_{PL} - \zeta_0\|_0 \le \widehat{m} + s$. 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

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

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

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

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

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

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

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 $S = (S_1 \times ... \times S_m \times S_y)$.

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

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

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

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

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

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

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

□

Appendix C

For proving Theorem 4.1, we need another lemma, which is quoted from Chernozhukov et al. (2018a) and restated as follows:

Lemma C.1 (Conditional convergence implies unconditional) *Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors. (i) If, for $\epsilon_m \to 0$, $\mathrm{P}\left(\|X_m\| > \epsilon_m | Y_m\right) \xrightarrow{p} 0$, then $\mathrm{P}\left(\|X_m\| > \epsilon_m\right) \to 0$. In particular, this occurs if $\mathrm{E}_P[\|X_m\|^q/\epsilon_m^q \xrightarrow{p} 0]$ for some $q \geq 1$, by Markov inequality. (ii) Let $\{A_m\}$ be a sequence of positive constants. If $\|X_m\| = O_P(A_m)$ conditional on Y_m, namely, that for any $l_m \to \infty$, $\mathrm{P}(\|X_m\| > l_m A_m | Y_m) \xrightarrow{p} 0$, then $\|X_m\| = O_P(A_m)$ unconditionally, namely, that for any $l_m \to \infty$, $\mathrm{P}(\|X_m\| > l_m A_m) \to 0$.*

Proof of Theorem 4.1. 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. Denote

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

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

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 $\widehat{\bar{A}}$ are also bounded below from zero, on event $\mathcal{E}_\eta$. Then, by the linearity in Assumption 3(i), we can write

$$
\widehat{\theta} = -\widehat{\bar{A}}^{-1}\widehat{\bar{B}},\ \ \theta_0 = -A_0^{-1}B_0.
$$

Then, we have

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

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

By Claim C.2, we have

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

Claim C.3. $\sqrt{N}\bar{\psi}(\theta_0, \eta_0) \xrightarrow{d} N(0, \Omega)$ *where* $\Omega = \Lambda_a\Lambda_a' + c\Lambda_g\Lambda_g$ *and* $\|\Omega\| < \infty$.

By Claims A.1, A.2, and A.3, we have

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

Now, we can combine the results and obtain

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

Proof of Claim C.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 $\|\Delta_{A,2}\|\leq r_{NT}$. By iterated expectation, we have $\mathrm{E}_P[\Delta_{A,1}]=0$. To simplify the notation, we define

$$\ddot{\psi}_{it}^{a,kl}:=\psi^a(W_{it},\widehat{\eta}_{kl})-\mathrm{E}_P[\psi^a(W_{it},\widehat{\eta}_{kl})|W(-k,-l)]$$

Then, we have

$$\mathrm{Var}\left(\left\|\widehat{A}_{kl}-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]\right\||W(-k,-l)\right)=\left(\frac{1}{N_kT_l}\right)^2\mathrm{E}_P\left[\left\|\sum_{i\in I_k,t\in S_l}\ddot{\psi}_{it}^{a,kl}\right\|^2|W(-k,-l)\right]$$

Expanding $\mathrm{E}_P\left[\left\|\sum_{i\in I_k,t\in S_l}\ddot{\psi}_{it}^{a,kl}\right\|^2|W(-k,-l)\right]$ and using triangle inequality gives

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

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

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

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

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

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 $\beta_g(m) \leq \beta_\gamma(m)$, as is shown in the proof of Claim C.3 below. 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

$$
\begin{aligned}
&\left|\mathrm{E}_P\left[\mathrm{E}_P\left[e^{kl}_{it}e^{kl}_{j,t+m}|\alpha_i,\alpha_j,W(-k,-l)\right]|W(-k,-l)\right]\right| \leq \mathrm{E}_P\left[\left|\mathrm{E}_P\left[e^{kl}_{it}e^{kl}_{j,t+m}|\alpha_i,\alpha_j,W(-k,-l)\right]\right||W(-k,-l)\right] \\
\lesssim&8\alpha_\gamma(m)^{1-2/q}\left(\mathrm{E}_P[|\ddot{\psi}^{a,kl}_{d,i,t}|^q|W(-k,-l)]\right)^{1/q}\left(\mathrm{E}_P[|\ddot{\psi}^{a,kl}_{d,j,t+m}|^q|W(-k,-l)]\right)^{1/q} \\
\lesssim&32\alpha_\gamma(m)^{1-2/q}a_1^2,
\end{aligned}
$$

where the first inequality follows from the Jensen's inequality; the second inequality follows from the fact that $\mathrm{E}[e^{k,l}_{it}\alpha_i, W(-k,-l)] = 0$, 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,

$$
\left|\mathrm{E}_P\left[g^{kl}_t g^{kl}_{t+m}|W(-k,-l)\right]\right| \lesssim \alpha_\gamma(m)^{1-2/q}a_1^2,
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k, i\neq j}\sum_{d=1}^{D}\mathrm{E}_P\left[\ddot{\psi}^{a,kl}_{d,i,t}\ddot{\psi}^{a,kl}_{d,j,t+m}|W(-k,-l)\right] \\
\lesssim&a_1^2\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k, i\neq j}\sum_{d=1}^{D}\alpha_\gamma(m)^{1-2/q} \\
\leq&a_1^2 D\sum_{m=1}^{\infty}c_\kappa exp(-\kappa m)^{1-2/q} \leq \frac{a_1^2 D c_\kappa}{exp(\kappa(1-2/q))-1} < \infty,
\end{aligned}
\tag{8.9}
$$

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

$$
\mathrm{Var}\left(\left\|\widehat{A}_{kl} - \mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]\right\|\right) = \left(\frac{1}{N_k T_l}\right)^2[a(1)+a(2)+(3)+2a(4)+2a(5)] \to 0.
$$

By Chebyshev's inequality, we conclude

$$\left\|\widehat{A}_{kl}-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]\right\|=o_P(1).$$

Also note that if we scale $\left\|\widehat{A}_{kl}-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]\right\|$ by $N_k^{-1/2}$, it is bounded in probability. So, we can be more specific about the convergence rate:

$$\left\|\widehat{A}_{kl}-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]\right\|=O_P(N_k^{-1/2})=O_P(N^{-1/2}),$$

where the last equality is due to $K<\infty$. To summarize, we have $\left\|\widehat{A}_{kl}-A_0\right\|=O_P(N^{-1/2}+\delta_{NT})$, which implies $\left\|\widehat{\bar{A}}-A_0\right\|=O_P(N^{-1/2}+r_{NT})$. So, Claim C.1 is proved.

Proof of Claim C.2.

$$\begin{aligned}
&\left\|\widehat{\bar{\psi}}(\theta_0)-\bar{\psi}(\theta_0,\eta_0)\right\|\\
=&\left\|\frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{E}_{kl}[\psi(W_{it};\theta_0,\widehat{\eta}_{kl})]-\mathbb{E}_{NT}[\psi(W_{it},\theta_0,\eta_0)]\right\|\\
=&\frac{1}{KL}\left\|\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{E}_{kl}\left[\psi(W_{it};\theta_0,\widehat{\eta}_{kl})-\psi(W_{it};\theta_0,\eta_0)\right]\right\|
\end{aligned}$$

Since K and L are finite, it suffices to show

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

We also define

$$\ddot{\psi}_{it}^{kl}:=\psi(W_{it};\theta_0,\widehat{\eta}_{kl})-\psi(W_{it};\theta_0,\eta_0),$$

and define

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

Then, 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. We further define

$$\tilde{\ddot{\psi}}_{it}^{kl} := \ddot{\psi}_{it}^{kl} - \mathrm{E}_P[\ddot{\psi}_{it}^{kl}|W(-k,-l)],$$

and denote $\tilde{\ddot{\psi}}_{d,it}$ as each element in the vector $\tilde{\ddot{\psi}}_{it}^{kl}$ for $d = 1, ..., D$, while still suppressing the subscripts k, l for convenience. Similar to what we have shown in the proof of Claim C.1, on the event $\mathcal{E}_\eta \cap \mathcal{E}_{cp}$, we have

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

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

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

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 8.9, $c(4)$ is of order $O({r'_{NT}}^2)$. So, we have shown

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

which implies $b(1) = O_P(r'_{NT})$.

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 $\mathcal{E}_\eta \cap \mathcal{E}cp$, we have

$$\left\|f'_{kl}(0)\right\| \leq \lambda_{NT},$$

under Assumption DML1(ii)(near-orthogonality), 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 C.2.

Proof of Claim C.3. Following the decomposition approach taken in Chiang et al. (2024), we define the following terms

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

Then, we can rewrite $\psi\left(W_{it};\theta_0,\eta_0\right) = a_i + g_t + e_{it}$. Under Assumptions 2 and 2, 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 $(\alpha_i, \gamma_t, \varepsilon_{it})$.

With the decomposition, we can re-express $\sqrt{N}\bar{\psi}(\theta_0,\eta_0)$ as follows:

$$\sqrt{N}\bar{\psi}(\theta_0,\eta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i + \frac{\sqrt{c}}{\sqrt{T}}\sum_{t=1}^{T} g_t + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}$$

Since the summations on the RHS are mutually uncorrelated, we have

$$\begin{aligned}
\mathrm{Var}\left(\sqrt{N}\bar{\psi}(\theta_0,\eta_0)\right) &= \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i\right) + \mathrm{Var}\left(\frac{\sqrt{c}}{\sqrt{T}}\sum_{t=1}^{T} g_t\right) + \mathrm{Var}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\right) \\
&= \Lambda_a\Lambda_a' + c\left(\Lambda_g\Lambda_g' + o(1)\right) + \frac{c}{N}\left(\Lambda_e\Lambda_e' + o(1)\right),
\end{aligned}$$

where the second equality follows from the stationarity of g_t and e_{it}, the uncorrelatedness of e_{it} over i, and the moment condition on $\psi(W_{it};\theta_0,\eta_0)$ (Assumption DML2(i)).

Next, we will show $\|\Lambda_a\Lambda_a'\| < \infty$, $\|\Lambda_g\Lambda_g'\| < \infty$, $\|\Lambda_e\Lambda_e'\| < \infty$. With an application of (conditional) Jensen's inequality and under Assumption DML2, we have

$$\begin{aligned}
\|\Lambda_a\Lambda_a'\| = \|\mathrm{E}_P[a_i'a_i]\| \le \mathrm{E}_P\|a_ia_i'\| = \mathrm{E}_P\|a_i\|^2 = \mathrm{E}_P\|\mathrm{E}_P[\psi(W_{it};\theta_0,\eta_0)|\alpha_i]\|^2 \\
\le \mathrm{E}_P[\mathrm{E}_P\|\psi(W_{it};\theta_0,\eta_0)\|^2||\alpha_i] = \mathrm{E}_P\|\psi(W_{it};\theta_0,\eta_0)\|^2 \le m_{NT} < \infty.
\end{aligned}$$

Due to the β-mixing property of g_t, we have

$$\begin{aligned}
\left\|\Lambda_g\Lambda_g'\right\| = \left\|\sum_{q=-\infty}^{\infty}\mathrm{E}_P[g_tg_{t+q}']\right\| \le \|\mathrm{E}_P[g_tg_t']\| + 2\sum_{q=1}^{\infty}\|\mathrm{E}_P[g_tg_{t+q}']\| \\
\le \mathrm{E}_P\|\psi(W_{it};\theta_0,\eta_0)\|^2 + 16\left(\mathrm{E}_P\|g_t\|^q\right)^{2/q}\sum_{q=1}^{\infty}\beta_g(q)^{1-2/q} \le \infty
\end{aligned}$$

where the second inequality follows Theorem 14.13(ii) of Hansen (2022) (with the α-mixing coefficient replaced by the β-mixing coefficient) and the third inequality follows from Assumptions 2 and DML2 with $p > 2$.

Note that, conditional on α_i, e_{it} is also β-mixing with the same mixing coefficient as γ_t. So, by Jensen's inequality, Theorem 14.13(ii), and again Jensen's inequality, we have

$$\left\| \mathrm{E}_P[e_{it} e_{i,t+q}] \right\| \leq \mathrm{E}_P \left\| \mathrm{E}_P[e_{it} e_{i,t+q} | \alpha_i] \right\| \leq 8 \left(\mathrm{E}_P \| e_{it} \|^q \right)^{2/q} \alpha_g(q)^{1-2/q}. \tag{8.10}$$

Then, similarly, we have

$$\| \Lambda_e \Lambda_e' \| = \left\| \sum_{q=-\infty}^{\infty} \mathrm{E}_P[e_{it} e_{i,t+q}'] \right\| < \infty.$$

Then, as $N, T \to \infty$, we have

$$\mathrm{Var}\left(\sqrt{N} \bar{\psi}(\theta_0, \eta_0) \right) \to \Lambda_a \Lambda_a' + c \Lambda_g \Lambda_g' = \Omega.$$

We have shown $\{a_i\}_{i \geq 1}$ is a sequence of i.i.d random vector with mean zero and finite variance $\Lambda_a \Lambda_a$. Then, Lindeberg-Lévy CLT applies:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \xrightarrow{d} N(0, \Lambda_a \Lambda_a').$$

Note that g_t is mean zero, $\mathrm{E}_P \| g_t \|^2 < \infty$, strictly stationary and β-mixing. Previously, we have shown $\sum_{q=1}^{\infty} \beta_g(q)^{1-2/q} < \infty$ for some $p > 2$. Then, the central limit theorem for mixing sequences applies here (see Theorem 14.15 of Hansen (2022)):

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{t=1}^{T} g_t \xrightarrow{d} N(0, \Lambda_g \Lambda_g')$$

Lastly, $\| \Lambda_e \Lambda_e' \| < \infty$ implies $\mathrm{Var}\left(\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it} \right) \to 0$. By Chebyshev's inequality, we have each component of the vector $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it} \xrightarrow{p} 0$, and so

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it} \xrightarrow{p} 0.$$

Since $\{a_i\}_{i \geq 1}$ and $\{g_t\}_{t \geq 1}$ are independent, we have $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i + \frac{\sqrt{c}}{\sqrt{T}} \sum_{t=1}^{T} g_t \xrightarrow{d} N(0, \Lambda_a \Lambda_a' + c \Lambda_g \Lambda_g') =$

$N(0, \Omega)$ Therefore, as, $N, T \to \infty$, we have

$$\sqrt{N}\bar{\psi}(\theta_0, \eta_0) \xrightarrow{d} N(0, \Omega),$$

as claimed.

□

Proof of Theorem 4.2. By the same arguments as at the beginning of the proof of Theorem 4.1, 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 C.1, we have $\|\widehat{\bar{A}} - A_0\| = O_P(N^{-1/2} + r_{NT})$ 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 C.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 $\|\widehat{\Omega}_{\mathrm{CHS},kl} - \Omega\| = o_P(1)$ where

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

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 define

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

Claim C.4. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{a,kl} - \Lambda_a \Lambda_a\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$.
Claim C.5. *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 C.6. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{c,kl}\right| = O_P\left(T^{-1}\right)$.
Claim C.7. *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 4.2.

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

$$\left|\widehat{\Omega}_{a,kl} - \Lambda_a \Lambda_a'\right| \leq \left|I_{a,1}^{(kl)}\right| + \left|I_{a,2}^{(kl)}\right| + \left|I_{a,3}^{(kl)}\right|,$$

where

$$\begin{aligned}
I_{a,1}^{(kl)} &:= \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \left\{ \widehat{\psi}_{it}^{(kl)} \widehat{\psi}_{ir}^{(kl)} - \psi_{it}^{(0)} \psi_{ir}^{(0)} \right\}, \\
I_{a,2}^{(kl)} &:= \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \left\{ \psi_{it}^{(0)} \psi_{ir}^{(0)} - \mathrm{E}_P[\psi_{it}^{(0)} \psi_{ir}^{(0)}] \right\}, \\
I_{a,3}^{(kl)} &:= \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \mathrm{E}_P[\psi_{it}^{(0)} \psi_{ir}^{(0)}] - \mathrm{E}_P[a_i a_i].
\end{aligned}$$

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, $I_{a,3}^{(kl)} = O(1/T_l^2) = O(T^{-2})$.

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

$$\left| I_{a,2}^{(kl)} \right| \leq \left| \frac{1}{N_k} \sum_{i \in I_k} I_{a,2,i}^{(kl)} \right| + \left| \frac{1}{N_k} \sum_{i \in I_k} \tilde{I}_{a,2,i}^{(kl)} \right|,$$

where

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

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

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

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

$$\mathrm{E}_P \left| h_i(\gamma_t, \gamma_r) \right|^{2+\delta} < \infty, \tag{8.11}$$

$$\int \int \left| h_i(u, v) \right|^{2+\delta} dF(u) dF(v) < \infty, \tag{8.12}$$

for some $\delta > 0$ and $F(.)$ is the common CDF of γ_t.

Consider condition 8.11. By Minkowski's inequality, Jensen's inequality, and the law of iterated expectation, we have

$$\begin{aligned}\left(\mathrm{E}_P \left| h_i(\gamma_t, \gamma_r) \right|^{2+\delta} \right)^{\frac{1}{2+\delta}} &\leq \left(\mathrm{E}_P \left| \psi^{(0)}_{it} \psi^{(0)}_{ir} \right|^{2+\delta} \right)^{\frac{1}{2+\delta}} + \mathrm{E}_P \left| \psi^{(0)}_{it} \psi^{(0)}_{ir} \right| \\ &\leq \left(\mathrm{E}_P \left| \psi^{(0)}_{it} \right|^{4+2\delta} \right)^{\frac{1}{2+\delta}} + \mathrm{E}_P \left| \psi^{(0)}_{it} \right|^2\end{aligned}$$

where the second inequality follows from Hölder's inequality and the identical distribution of γ_t. Let $\delta = \frac{p-4}{2}$, then $\left(\mathrm{E}_P \left| \psi^{(0)}_{it} \right|^{4+2\delta} \right)^{\frac{1}{2+\delta}} < a_1$ and $\mathrm{E}_P \left| \psi^{(0)}_{it} \right|^2 \leq a_1^2$ follows from Assumption DML2(i). Therefore, condition 8.11 is satisfied.

Consider condition 8.12. 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 8.12 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 $\tilde{I}_{a,2,i}^{(kl)}=o_P(T^{-1/2})$.

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

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

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

$$\begin{aligned}\left(\mathrm{E}_P\left[\left|I_{a,2,i}^{(kl)}\right|^2\right]\right)^{1/2} \lesssim & \frac{1}{T_l^2}\sum_{t,r\in S_l}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{ir}^{(0)}\right]^2\right)^{1/2}\\ \leq & \frac{1}{T_l^2}\sum_{t,r\in S_l}\left(\mathrm{E}_P\left[\psi_{it}^{(0)}\right]^4\right)^{1/2} \leq a_1^2,\end{aligned}$$

where the last inequality follows from Assumption 3.5(ii). 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 $\left|I_{a,2}^{kl}\right| = O_P\left(N^{-1/2}\right) + o_P\left(T^{-1/2}\right)$.

Now, consider $I_{a,1}^{kl}$.

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

where

$$R_{kl} = \left(\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right)^2\right)^{1/2}.$$

By Markov inequality and under Assumption DML2(i), we have

$$\mathrm{E}_P\left[\frac{1}{N_kT_l}\sum_{i\in I_k,t\in S_l}\left(\psi_{it}^{(0)}\right)^2\right]=\mathrm{E}_P\left|\psi(W_{it};\theta_0,\eta_0)\right|^2\leq a_1^2.$$

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

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

By Markov inequality and Assumption DML2(i), we have $\frac{1}{N_kT_l}\sum_{i\in I_k,t\in S_l}\left|\psi^a(W_{it};\widehat{\eta}_{kl})\right|^2=O_P(1)$. By Theorem 4.1, $\left|\widehat{\theta}-\theta_0\right|^2=O_p(N^{-1})$. 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

$$\begin{aligned}&\mathrm{E}_P\left[\frac{1}{N_kT_l}\sum_{i\in I_k,t\in S_l}\left|\widehat{\psi}_{it}^{(kl)}-\psi_{it}^{(0)}\right|^2|W(-k,-l)\right]\\ =&\mathrm{E}_P\left[\left|\psi(W_{it};\theta_0,\widehat{\eta}_{kl})-\psi(W_{it};\theta_0,\eta_0)\right|^2|W(-k,-l)\right]\leq(\delta_{NT}')^2,\end{aligned}$$

where the last inequality follows from Assumption DML2(ii). Then, by Markov inequality, 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

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

Proof of Claim C.5. By triangle inequality, we have

$$\left|\widehat{\Omega}_{b,kl}-c\mathrm{E}_P[g_tg_t']\right|\leq\left|I_{b,1}^{(kl)}\right|+\left|I_{b,2}^{(kl)}\right|+\left|I_{b,3}^{(kl)}\right|,$$

where

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

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

Consider $I_{b,3}^{(kl)}$. Note that

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

where the second equality follows from the law of total covariance. 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

$$
\frac{1}{c} \left| I_{b,2}^{kl} \right| \leq \left| \frac{1}{T_l} \sum_{t\in S_l} I_{b,2,t}^{(kl)} \right| + \left| \frac{1}{T_l} \sum_{t\in S_l} \tilde{I}_{b,2,t}^{(kl)} \right|,
$$

where

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

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 3.5(ii) with some $p > 4$. Therefore, by Markov inequality, we have

$\left|\frac{1}{T_l}\sum_{t\in S_l}\tilde{I}^{(kl)}_{b,2,t}\right| = O_P(N_k^{-1/2}) = O_P(N^{-1/2})$.

Now consider $\left|\frac{1}{T_l}\sum_{t\in S_l} I^{(kl)}_{b,2,t}\right|$. Note that conditional on $\{\alpha_i\}$, $I^{(kl)}_{b,2,t}$ 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

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

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

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

By conditional Markov inequality, we have

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

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

$$\begin{aligned}\left(\mathrm{E}_P\left[\left|I^{(kl)}_{b,2,t}\right|^{\frac{q}{2}}\right]\right)^{\frac{2}{q}} \lesssim &\frac{1}{N_k^2}\sum_{i,j\in I_k}\left(\mathrm{E}_P\left[\psi^{(0)}_{it}\psi^{(0)}_{jt}\right]^{\frac{q}{2}}\right)^{\frac{2}{q}}\\ \leq &\frac{1}{N_k^2}\sum_{i,j\in I_k}\left(\mathrm{E}_P\left[\psi^{(0)}_{it}\right]^{q}\right)^{\frac{2}{q}} \leq a_1^2,\end{aligned}$$

where the last inequality follows from Assumption 3.5(ii). Then, by the law of iterated expectation, we have

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

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

The argument to bound $I^{kl}_{b,1}$ is similar to the that for $I^{kl}_{a,1}$. By the similar inequality for $\left|I^{kl}_{a,1}\right|$, we have

$$\frac{1}{c}\left|I^{kl}_{b,1}\right| \lesssim R_{kl}\left\{\left(\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi^{(0)}_{it}\right)^2\right)^{1/2}+R_{kl}\right\},$$

where

$$R_{kl}=\left(\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\widehat{\psi}^{(kl)}_{it}-\psi^{(0)}_{it}\right)^2\right)^{1/2}.$$

We have shown in the proof of Claim C.4 that $\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi^{(0)}_{it}\right)^2 = O_P(1)$ and $R_{kl}=O_P\left(N^{-1}+(r'_{NT})^2\right)$. So $\left|I^{kl}_{b,1}\right| = O_P\left(N^{-1/2}+r'_{NT}\right)$. To summarize

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

which completes the proof of Claim C.5.

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

$$\left|\widehat{\Omega}_{c,kl}\right| \le \left|I^{(kl)}_{c,1}\right|+\left|I^{(kl)}_{c,2}\right|+\left|I^{(kl)}_{c,3}\right|$$

where

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

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

$$\mathrm{E}_P[\psi^{(0)}_{it}\psi^{(0)}_{it}] \le a_1^2.$$

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

To bound $I^{kl}_{c,2}$, consider the variance of $I^{kl}_{c,2}$. Denote $\xi_{it} = \psi^{(0)}_{it}\psi^{(0)}_{it} - \mathrm{E}_P[\psi^{(0)}_{it}\psi^{(0)}_{it}]$.

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

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

$$
\mathrm{E}_P\left|\xi_{it}\right|^2 \lesssim \mathrm{E}_P\left[\psi(W_{it};\theta_0,\eta_0)^4\right] \leq a_1^4.
$$

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

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

$$
\begin{aligned}
\left|I^{kl}_{c,1}\right| &\leq \frac{K/L}{N_k T_l^2} \sum_{i\in I_k, t\in S_l} \left|\widehat{\psi}^{(kl)}_{it}\widehat{\psi}^{(kl)}_{it} - \psi^{(0)}_{it}\psi^{(0)}_{it}\right| \\
&\leq \frac{K/L}{N_k T_l^2} \sum_{i\in I_k, t\in S_l} \left\{ \left|\widehat{\psi}^{(kl)}_{it} - \psi^{(0)}_{it}\right|^2 + 2\left|\psi^{(0)}_{it}\right|\left|\widehat{\psi}^{(kl)}_{it} - \psi^{(0)}_{it}\right| \right\} \\
&\lesssim \frac{K/L}{T_l} R_{kl} \left\{ \left(\frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l}\left(\psi^{(0)}_{it}\right)^2\right)^{1/2} + R_{kl} \right\},
\end{aligned}
$$

where

$$R_{kl} = \left(\frac{1}{N_k T_l} \sum_{i \in I_k, t \in S_l} \left(\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)} \right)^2 \right)^{1/2}.$$

We have shown in the proof of Claim C.4 that $\frac{1}{N_k T_l} \sum_{i \in I_k, t \in S_l} \left(\psi_{it}^{(0)} \right)^2 = O_P(1)$ and $R_{kl} = O_P\left(N^{-1} + (r'_{NT})^2 \right)$. So, $\left| I_{c,1}^{kl} \right| = O_P\left((NT)^{-1} + (r'_{NT})^2/T \right)$. To summarize

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

which completes the proof of Claim C.6.

Proof of Claim C.7. 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 component representation and its properties (iii) and (iv) shown in the proof of Claim C.3. 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), $\left(\mathrm{E}_P|\psi_{it}^{(0)}|^q\right)^{1/q} \leq a_1$ for $p > 4$. 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

$$\begin{aligned}\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \leq& \mathrm{E}_P\left[\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}|\{\alpha_i\}_{i\in I_k}\right]\right|\right] \\ \leq& 8\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{it}^{(0)}|^q|\alpha_i\right]\right)^{1/q}\left(\mathrm{E}_P\left[|\psi_{j,t+m}^{(0)}|^q|\alpha_j\right]\right)^{1/q}\alpha_\gamma(m)^{1-2/q}\right] \\ \leq& 8\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{it}^{(0)}|^q|\alpha_i\right]\right)^{1/q}\right]\mathrm{E}_P\left[\left(\mathrm{E}_P\left[|\psi_{j,t+m}^{(0)}|^q|\alpha_j\right]\right)^{1/q}\right]\alpha_\gamma(m)^{1-2/q} \\ \lesssim& a_1^2\alpha_\gamma(m)^{1-2/q}\end{aligned}$$

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

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

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

$$\begin{aligned}\left|I_{d,3}^{(kl)}\right| \leq & \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} \left|k\left(\frac{m}{M}\right)-1\right| \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j \in I_k, j\neq i} \left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \\ \leq & c a_1^2 \sum_{m=1}^{M-1} \left|k\left(\frac{m}{M}\right)-1\right| \alpha_\gamma(m)^{1-2/q}.\end{aligned}$$

Note that for each m, $\left|k\left(\frac{m}{M}\right)-1\right| \to 0$ as $M \to \infty$. Since $\left|k\left(\frac{m}{M}\right)-1\right|\alpha_\gamma(m)^{1-2/q} \leq 1$, 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

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

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

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

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

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

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 3.5(ii) with some $p>4$. Therefore, we have $\left(\mathrm{E}_P\left|\sum_{m=1}^{M-1} k\left(\frac{m}{M}\right)\tilde{I}^{(kl)}_{d,2,tm}\right|^2\right)^{1/2} \leq O_P\left(\frac{M}{N^{1/2}}\right) = O_P\left(\frac{M}{T^{1/2}}\right)$. 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 $I^{kl}_{d,1}$. Denote

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

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

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

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

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

which completes the proof of Claim C.7.

□

Proof of Theorem 4.3. Since (K, L) are fixed constants, it suffices to show for each (k, l) that $\widehat{\Omega}_{\mathrm{NW},kl} := \frac{K/L}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} k\left(\frac{|t-r|}{M}\right) \psi(W_{it}; \widehat{\theta}, \widehat{\eta}_{kl})\psi(W_{ir}; \widehat{\theta}, \widehat{\eta}_{kl})' = o_P(1)$. 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 4.2, and $\widehat{\Omega}_{e,kl}$ is defined as follows:

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

Observe that by replacing $\widehat{\Omega}_{d,kl}$ by $\widehat{\Omega}_{e,kl}$, each step in the proof of Claim C.7 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 D

Proof of Theorem 5.1. Let $P \in \mathcal{P}_{NT}$ for each (N, T). WLOG, we assume $N \wedge T = N$. First, we can rewrite $\widehat{\theta}$ as

$$\widehat{\theta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\tilde{\zeta})(D_{it} - f_{it}\tilde{\pi})\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\tilde{\zeta})(Y_{it} - f_{it}\tilde{\beta})$$

By rescaling $\widehat{\theta}$ and plugging in the reduced-form model of Y_{it} in 5.12, we have

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\tilde{\zeta})(D_{it} - f_{it}\tilde{\pi})\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it} - f_{it}\tilde{\zeta})(V_{it}^{Y} + f_{it}(\beta - \tilde{\beta}))$$

Define

$$\begin{aligned}
A_0 &:= -\mathrm{E}_P[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi_0)], & B_0 &:= \mathrm{E}_P[(Z_{it} - f_{it}\zeta_0)(Y_{it} - f_{it}\beta_0)],\\
A_{NT} &:= -\mathbb{E}_{NT}[(Z_{it} - f_{it}\zeta_0)(D_{it} - f_{it}\pi_0)], & B_{NT} &:= \mathbb{E}_{NT}[(Z_{it} - f_{it}\zeta_0)(Y_{it} - f_{it}\beta_0)],\\
\hat{A}_{NT} &:= -\mathbb{E}_{NT}[(Z_{it} - f_{it}\tilde{\zeta})(D_{it} - f_{it}\tilde{\pi})], & \hat{B}_{NT} &:= \mathbb{E}_{NT}[(Z_{it} - f_{it}\tilde{\zeta})(Y_{it} - f_{it}\tilde{\beta})]\\
\psi_{NT} &:= \mathbb{E}_{NT}[(Z_{it} - f_{it}\zeta_0)(Y_{it} - f_{it}\beta_0 - \theta_0(D_{it} - f_{it}\pi_0))], & &\\
\widehat{\psi}_{NT}(\theta) &:= \mathbb{E}_{NT}[(Z_{it} - f_{it}\tilde{\zeta})(Y_{it} - f_{it}\tilde{\beta} - \theta(D_{it} - f_{it}\tilde{\pi}))]. & &
\end{aligned}$$

Following the similar algebra as in the beginning of the Proof of Theorem 4.1, we have

$$\begin{aligned}\sqrt{N}V^{-1/2}(\widehat{\theta}-\theta_0) =& \sqrt{N}V^{-1/2}A_0^{-1}\psi_{NT} + \sqrt{N}V^{-1/2}A_0^{-1}(\widehat{\psi}_{NT}(\theta_0)-\psi_{NT}) \\ &+ \sqrt{N}V^{-1/2}[(A_0+\widehat{A}_{NT}-A_0)^{-1}-A_0^{-1}](\psi_{NT}+\widehat{\psi}_{NT}(\theta_0)-\psi_{NT})\end{aligned}$$

We will show (1) $\|A_0-\widehat{A}_{NT}\| = o_P(1)$, (2) $\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0)-\psi_{NT}\| = o_P(1)$, and (3) $\sqrt{N}\Omega^{-1/2}\psi_{NT} \xrightarrow{d} N(0,1)$ with $\|\Omega\| \leq \infty$. With the identification condition in Assumption REG-P(iii), we have $A_0^{-1} > 0$ and so $V^{-1/2} > 0$. The conclusion of the theorem follows.

First, consider statement (3). By definition, we have

$$\psi_{NT} = \mathbb{E}_{NT}\left(Z_{it}-f_{it}\zeta_0\right)\left(U_{it}-(L_{2,it}-\mathrm{E}[L_{2,it}])\eta_{Y2}+(L_{2,it}-\mathrm{E}[L_{2,it}])\eta_{D2}\theta_0\right)$$

It is clear that Z_{it} is uncorrelated with U_{it} and linear combinations of $L_{2,it}$ due to $\mathrm{E}[Z_{it}U_{it}]=0$ and the independence between Z_{it} and the Mundlak errors. Since $f_{it}\zeta_0$ is part of $\mathrm{E}[Z_{it}|X_{it},c_i,d_t]$, it is also uncorrelated with U_{it} and linear combinations of $L_{2,it}$. Because the term in the second parenthesis is mean zero, we have $\mathrm{E}[\psi_{NT}]=0$. The rest of the statement follows from the same arguments as in the Proof of Claim C.3.

Consider statement (1): $\|A_0-\widehat{A}_{NT}\|$. It is assumed that $\|\tilde{\zeta}-\zeta_0\| = o_P(1)$, $\|\tilde{\beta}-\beta_0\| = o_P(1)$ and $\|\tilde{\pi}-\pi_0\| = o_P(1)$. Then, following the same idea of Lemma 4.3 from Newey and McFadden (1994), there exists a sequence $\delta_{NT} \to 0$ such that $\|\tilde{\pi}-\pi_0\| \leq \delta_{NT}$ and $\|\tilde{\zeta}-\zeta_0\| \leq \delta_{NT}$ with probability approaching one. Then,

$$\begin{aligned}&\|(Z_{it}-f_{it}\tilde{\zeta})(D_{it}-f_{it}\tilde{\pi})-(D_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)\| \\ \leq& \sup_{\|\pi-\pi_0\|\leq\delta_{NT},\|\zeta-\zeta_0\|\leq\delta_{NT}} \|(D_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)-(D_{it}-Z_{it}\pi_0)(D_{it}-f_{it}\pi_0)\| \to 0\end{aligned}$$

Let $\mathcal{N}_m(\xi_0) = \{\xi \in \mathbb{R}^p : \|\xi-\xi_0\| \leq m\}$ for $\xi = \beta,\pi,\zeta$. Then, by Hölder's inequality, we have

$$\begin{aligned}\mathrm{E}_P\left[\sup_{\pi\in\mathcal{N}_m(\pi_0),\zeta\in\mathcal{N}_m(\zeta_0)}\|(Z_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)\|\right] \leq& \|Z_{it}\|_{P,2}\|D_{it}\|_{P,2} + \|Z_{it}\|_{P,2}\left\|\sup_{\pi\in\mathcal{N}_m(\pi_0)}|f_{it}\pi|\right\|_{P,2} \\ + \|D_{it}\|_{P,2}\left\|\sup_{\pi\in\mathcal{N}_m(\zeta_0)}|f_{it}\zeta|\right\|_{P,2} + \left\|\sup_{\pi\in\mathcal{N}_m(\pi_0)}|f_{it}\pi|\right\|_{P,2}&\left\|\sup_{\pi\in\mathcal{N}_m(\zeta_0)}|f_{it}\zeta|\right\|_{P,2} < \infty.\end{aligned}$$

For large enough (N,T), we have

$$\begin{aligned}&\sup_{\|\pi-\pi_0\|\leq\delta_{NT},\|\zeta-\zeta_0\|\leq\delta_{NT}}\|(Z_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)-(D_{it}-Z_{it}\pi_0)(D_{it}-f_{it}\pi_0)\| \\ \leq& 2\sup_{\pi\in\mathcal{N}_m(\pi_0),\zeta\in\mathcal{N}_m(\zeta_0)}\|(Z_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)\|.\end{aligned}$$

So, we can apply the dominated convergence theorem: as $(N,T) \to \infty$,

$$\mathrm{E}\left[\sup_{\|\pi-\pi_0\|\le\delta_{NT},\|\zeta-\zeta_0\|\le\delta_{NT}} \|(Z_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)-(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)\|\right] \to 0.$$

Therefore, by Markov inequality, we have

$$\mathbb{E}_{NT}\left[\sup_{\|\pi-\pi_0\|\le\delta_{NT},\|\zeta-\zeta_0\|\le\delta_{NT}} \|(Z_{it}-f_{it}\zeta)(D_{it}-f_{it}\pi)-(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)\|\right] \xrightarrow{p} 0.$$

By Assumptions AHK, AR, REG-P, Theorem 1 of Chiang et al. (2024) applies, giving the weak law of large number for $\mathbb{E}_{NT}(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)$, i.e.,

$$\mathbb{E}_{NT}[(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)] \xrightarrow{p} \mathrm{E}[(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)].$$

Then, by the triangle inequality, we have

$$\|A_0-\widehat{A}_{NT}\| = \|A_0-\widehat{A}_{NT} \pm \mathbb{E}_{NT}(Z_{it}-f_{it}\zeta_0)(D_{it}-f_{it}\pi_0)\| \le o_P(1).$$

Consider statement (3): $\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0)-\psi_{NT}\|$. Note that $g(X_{it},c_i,d_t) = \mathrm{E}[Y_{it}|X_{it},c_i,d_t]-\theta_0\mathrm{E}[D|X_{it},c_i,d_t]$. Then, by straightforward algebra, we have

$$\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0)-\psi_{NT}\| \le \sum_{k=1}^{6}\Delta_k,$$

where

$$\begin{aligned}
\Delta_1 &:= \left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\zeta_0-\tilde{\zeta})f_{it}f_{it}(\beta_0-\tilde{\beta})\right\|, \quad \Delta_2 := \left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\zeta_0-\tilde{\zeta})f_{it}f_{it}(\pi_0-\tilde{\pi})\right\| \\
\Delta_3 &:= \left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}^{Z}f_{it}(\beta_0-\tilde{\beta})\right\|, \quad \Delta_4 := \theta_0\left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}^{Z}f_{it}(\pi_0-\tilde{\pi})\right\| \\
\Delta_5 &:= \left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}^{Y}f_{it}(\zeta_0-\tilde{\zeta})\right\|, \quad \Delta_6 := \theta_0\left\|\frac{\sqrt{N}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}^{D}f_{it}(\zeta_0-\tilde{\zeta})\right\|
\end{aligned}$$

It is assumed that $\left|f_{it}(\xi-\tilde{\xi})\right|_{NT,2} = o_P\left(N^{-1/2}\right)$ for $\xi=\beta,\pi,\zeta$. Then, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\Delta_1 &\le \sqrt{N}\left\|f_{it}(\zeta_0-\tilde{\zeta})\right\|_{NT,2}\left\|f_{it}(\beta_0-\tilde{\beta})\right\|_{NT,2} = o_P\left(N^{-1/2}\right), \\
\Delta_3 &\le \sqrt{N}\left\|f_{it}(\beta_0-\tilde{\beta})\right\|_{NT,2}\left\|V_{it}^{Z}\right\|_{NT,2} = o_P(1)\left\|V_{it}^{Z}\right\|_{NT,2},
\end{aligned}$$

By Theorem 1 of Chiang et al. (2024), we have $\frac{1}{NT}\left(V_{it}^{Z}\right)^2 \xrightarrow{p} \mathrm{E}\left[V_{it}^{Y}\right]^2$. Therefore,

$$\Delta_3 = o_P(1).$$

Bounds for Δ_2 and $\Delta_4, \Delta_5, \Delta_6$ are obtained similarly: $\Delta_2 = o_P\left(N^{-1/2}\right)$ and $\Delta_4, \Delta_8, \Delta_9 = o_P(1)$. Therefore, $\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0) - \psi_{NT}\| = o_P(1)$ and so

$$\sqrt{N}V^{-1/2}(\widehat{\theta} - \theta_0) = \sqrt{N}V^{-1/2}A_0^{-1}\psi_{NT} + o_P(1) \xrightarrow{d} N(0,1).$$

□

Proof of Theorem 5.2. We have shown in the proof of Theorem 5.1 that $A_0 - \widehat{A}_{NT} = o_P(1)$. Therefore, it suffices to show $\widehat{\Omega}_{\mathrm{CHS}} - \Omega = o_P(1)$. We decompose $\widehat{\Omega}_{\mathrm{CHS}}$ as follows:

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

where $\psi(W_{it};\widehat{\theta},\tilde{\eta}) = (Z_{it} - f_{it}\tilde{\zeta})(Y_{it} - f_{it}\tilde{\beta} - \widehat{\theta}(D_{it} - f_{it}\tilde{\pi}))$. We need to show $\widehat{\Omega}_a \xrightarrow{p} \Lambda_a\Lambda_a = \mathrm{E}_P[a_i^2]$, $\widehat{\Omega}_b \xrightarrow{p} c\mathrm{E}[g_t^2]$, $\widehat{\Omega}_c = o_P(1)$, 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

$$\left|\widehat{\Omega}_a - \mathrm{E}_P[a_i^2]\right| \leq \left|I_{a,1}\right| + \left|I_{a,2}\right| + \left|I_{a,3}\right|,$$

where

$$\begin{aligned}
I_{a,1} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\psi(W_{it};\widehat{\theta},\tilde{\eta})\psi(W_{ir};\widehat{\theta},\tilde{\eta}) - \psi(W_{it};\theta_0,\eta_0)\psi(W_{ir};\theta_0,\eta_0)\right\},\\
I_{a,2} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\psi(W_{it};\theta_0,\eta_0)\psi(W_{ir};\theta_0,\eta_0) - \mathrm{E}[\psi(W_{it};\theta_0,\eta_0)\psi(W_{ir};\theta_0,\eta_0)]\right\},\\
I_{a,3} :=& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\left\{\mathrm{E}[\psi(W_{it};\theta_0,\eta_0)\psi(W_{ir};\theta_0,\eta_0)] - \mathrm{E}[a_i^2]\right\}.
\end{aligned}$$

Note that in proving Claim C.4, the cross-fitting device is only used to show that $I_{a,1}$ is of small order. 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 C.4. So it is omitted here.

Consider $I_{a,1}$. Following the algebra in the proof of Claim C.4, we have

$$\left|I_{a,1}\right| \lesssim R_{NT}\left\{\left|\psi(W_{it};\theta_0,\eta_0)\right|_{NT,2} + R_{NT}\right\}$$

where

$$R_{NT} := \left\|\psi(W_{it};\hat{\theta},\tilde{\eta}) - \psi(W_{it};\theta_0,\eta_0)\right\|_{NT,2}$$

Under Assumption REG-P(i) we have $\mathrm{E}_P[\psi(W_{it};\theta_0,\eta_0)]^2 = O_P(1)$ and so, by Markov inequality, we have $\left|\psi(W_{it};\theta_0,\eta_0)\right|_{NT,2} = O_P(1)$.

Consider R_{NT}^2. By Minkowski's inequality, we have

$$\begin{aligned} R_{NT} &= \left\|\psi^a(W_{it};\tilde{\eta})(\hat{\theta}-\theta_0) + \psi(W_{it};\theta_0,\tilde{\eta}) - \psi(W_{it};\theta_0,\eta_0)\right\|_{NT,2} \\ &\lesssim \left\|\psi^a(W_{it};\tilde{\eta})(\hat{\theta}-\theta_0)\right\|_{NT,2} + \left\|\psi(W_{it};\theta_0,\tilde{\eta}) - \psi(W_{it};\theta_0,\eta_0)\right\|_{NT,2}, \end{aligned}$$

where $\psi^a(W_{it};\tilde{\eta}) := \left(Z_{it} - f_{it}\tilde{\zeta}\right)(D_{it} - f_{it}\tilde{\pi})$. By Minkowski's inequality and Hölder's inequality, we have

$$\|\psi^a(W_{it};\tilde{\eta})\|_{P,2} \leq \left(\|Z_{it}\|_{P,4} + \|f_{it}\tilde{\zeta}\|_{P,4}\right)\left(\|D_{it}\|_{P,4} + \|f_{it}\tilde{\pi}\|_{P,4}\right)$$

It is assumed that $\|\tilde{\pi}-\pi_0\| = o_p(1)$, $\|\tilde{\zeta}-\zeta_0\| = o_p(1)$, $\|\tilde{\beta}-\beta_0\| = o_p(1)$. Then, there exists a sequence $\delta_{NT} \to 0$ such that $\|\tilde{\pi}-\pi_0\| + \|\tilde{\zeta}-\zeta_0\| + \|\tilde{\beta}-\beta_0\| \leq \delta_{NT}$ with probability approaching one. Then, for large enough (N,T), we have

$$\begin{aligned} \mathrm{E}_P[f_{it}\tilde{\pi}]^4 \leq & \mathrm{E}_P\left[\sup_{\|\pi-\pi_0\|\leq\delta_{NT}} |f_{it}\pi|^4\right] = \mathrm{E}_P\left[\sup_{\|\pi-\pi_0\|\leq\delta_{NT}} |f_{it}\pi|\right]^4 \leq \mathrm{E}_P\left[\sup_{\pi\in\mathcal{N}_m(\pi_0)} |f_{it}\pi|\right]^4 < \infty, \\ \mathrm{E}_P[f_{it}\tilde{\zeta}]^4 \leq & \mathrm{E}_P\left[\sup_{\|\zeta-\zeta_0\|\leq\delta_{NT}} |f_{it}\zeta|^4\right] = \mathrm{E}_P\left[\sup_{\|\zeta-\zeta_0\|\leq\delta_{NT}} |f_{it}\zeta|\right]^4 \leq \mathrm{E}_P\left[\sup_{\zeta\in\mathcal{N}_m(\zeta_0)} |f_{it}\zeta|\right]^4 < \infty \end{aligned}$$

So, we have $\mathrm{E}_P[\psi^a(W_{it};\tilde{\eta})]^2 < \infty$. By Markov inequality, we conclude that $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(\psi^a(W_{it};\tilde{\eta})\right)^2 = O_P(1)$. By Theorem 5.1, we have $(\hat{\theta}-\theta_0)^2 = O_P(N^{-1})$. Therefore,

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(\psi^a(W_{it};\tilde{\eta})\right)^2(\hat{\theta}-\theta_0)^2 = O_P(N^{-1}).$$

Consider the second term, $\left\|\psi(W_{it};\theta_0,\tilde{\eta})-\psi(W_{it};\theta_0,\eta_0)\right\|_{NT,2}$:

$$\left|\psi(W_{it};\theta_0,\tilde{\eta})-\psi(W_{it};\theta_0,\eta_0)\right|^2 \leq \sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0+\|\pi-\pi_0\|\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2$$

$$= \sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \Big|(\zeta_0-\tilde{\zeta})'f_{it}'f_{it}(\beta_0-\tilde{\beta})-\theta_0(\zeta_0-\tilde{\zeta})'f_{it}'f_{it}(\pi_0-\tilde{\pi})+(\zeta_0-\tilde{\zeta})'f_{it}'V_{it}^Y$$

$$-\theta_0(\zeta_0-\tilde{\zeta})'f_{it}'V_{it}^D+V_{it}^Z f_{it}(\beta_0-\tilde{\beta})+V_{it}^Z f_{it}(\pi_0-\tilde{\pi})\Big|^2 \to 0.$$

For large enough N,T, we have

$$\sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2 \leq \sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2$$

By Minkowski's inequality and Hölder's inequality, we have

$$\left(\mathrm{E}\left[\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2\right]\right)^{1/2}$$

$$=\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|\right\|_{P,2} = 2\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)} \left|\psi(W_{it};\theta_0,\eta)\right|\right\|_{P,2}$$

$$\leq 2\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0)} \left|Z_{it}-f_{it}\zeta\right|\right\|_{4,P}\left\|\sup_{\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)} \left|Y_{it}-f_{it}\beta-\theta_0(D_{it}-f_{it}\pi)\right|\right\|_{4,P}$$

$$\leq 2\left(\left\|Z_{it}\right\|_{4,P}+\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0)} f_{it}\zeta\right\|_{4,P}\right)\left(\left\|Y_{it}\right\|_{4,P}+\left\|\sup_{\beta\in\mathcal{N}_m(\beta_0)} f_{it}\beta\right\|_{4,P}+\theta_0\left\|D_{it}\right\|_{4,P}+\theta_0\left\|\sup_{\pi\in\mathcal{N}_m(\pi_0)} f_{it}\pi\right\|_{4,P}\right)<\infty.$$

So, we can apply the dominated convergence theorem to obtain

$$\mathrm{E}\left[\sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2\right]\to 0$$

Then, by Markov inequality, we have

$$\mathbb{E}_{NT}\left[\sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2\right]\xrightarrow{p} 0.$$

Therefore, as $(N,T)\to\infty$, we have

$$\left\|\psi(W_{it};\theta_0,\tilde{\eta})-\psi(W_{it};\theta_0,\eta_0)\right\|_{NT,2}=\left(\mathbb{E}_{NT}\left[\psi(W_{it};\theta_0,\tilde{\eta})-\psi(W_{it};\theta_0,\eta_0)\right]^2\right)^{1/2}$$

$$\leq\left(\mathbb{E}_{NT}\left[\sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2\right]\right)^{1/2}\xrightarrow{p} 0.$$

It follows that $R_{NT} = o_P(1)$.

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 4.2 (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. So we do not repeat those here. Combining these results, we obtain $\widehat{\Omega} = \mathrm{E}_P(a_i^2) + c\mathrm{E}_P(g_t 2) + c\sum_{m=1}^{\infty} \mathrm{E}_P(g_t g_{t+m}) = \Lambda_a \Lambda_a + c\Lambda_g \Lambda_g$.

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

$$\Omega_{\mathrm{DKA}} = \widehat{\Omega}_c + \widehat{\Omega}_e - \widehat{\Omega}_d,$$

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 C.7, we have $\widehat{\Omega}_e = \widehat{\Omega}_d + o_P(1)$; From above, we also have $\widehat{\Omega}_c = o_P(1)$. Therefore, we conclude that $\widehat{\Omega}_{NW} = o_P(1)$. So it is proved. $\square$