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

Kaicheng Chen

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

Oct 22, 2024

Link to the latest manuscript

Abstract

Panel data allows for the modeling of unobserved heterogeneity, which increases the number of nuisance parameters drastically, making high dimensionality a relevant practical issue rather than just a theoretical concern. However, unobserved heterogeneity, along with potential two-way 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 a theoretically driven penalty level and weights. Alternatively, a cross-validation scheme robust to panel data dependence is also proposed. Due to the underlying factor structure that drives the cluster dependence, a slow rate of convergence is revealed. I further investigate the rate requirement for valid inference with the help of a cross-fitting scheme robust to two-way cluster dependence. Moreover, I consider the complication due to unobserved heterogeneity and how it helps eliminate the underlying factors for faster convergence. 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 allows for more flexible modeling and more robust estimation and inference.

Keywords: high-dimensional regression, two-way cluster dependence, correlated time effects, unobservable heterogeneity, cluster-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 in order 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 and so 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, statistical properties of many high-dimensional estimators remain unknown with panel data, which is potentially dependent over both 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 complicates estimation and inference.

For the first challenge, I proposed two LASSO-based approaches. The first one is a weighted-LASSO, whose regressor-specific penalty weights are based on some self-normalization scheme robust to two-way dependence and heteroskedasticity and whose common penalty level is theoretically driven. Such an LASSO approach is named as two-way cluster-LASSO, corresponding to the heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016a). By decomposing the score 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. $\lambda > C \max_{1\leq j\leq p} \left| \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_j^{-1/2} f_{it,j} V_{it} \right|$ where λ is a common penalty level, C is some constant, ω_j are some regressor-specific weights, $f_{it,j}$ are regressors, and V_{it} are stochastic errors. The second LASSO approach chooses the common penalty level by a cross-validation algorithm. Built upon previous model selection literature that utilize cross validation (Shao, 1993; Burman et al., 1994; Racine, 2000), I propose a cross-validation algorithm robust to two-way cluster dependence and serial dependence across clusters, a common feature in panel data. The idea is to construct the training and testing sub-samples by grouping observations by clusters and excluding an neighborhood in temporal dimension that grows with sample sizes. As is shown in Section 4, the training and testing sub-samples generated this way are “approximately” independent, asymptotically. The theoretical challenge of cross-validation LASSO is that the cross-validated λ is often quite small and it results in non-trivial probability of not realizing the regularization event. Taking a different approach, Chetverikov et al. (2021) establish convergence rates results for LASSO that uses cross-validated $\hat{\lambda}$, and the convergence rates are as fast as those utilize the bound under a regularization event up to a small factor. As is shown in simulation, cross-validation LASSO often works better than LASSO using theoretically driven penalty terms. The proposed

panel cross validation is not an exception either. However, to establish such a result in panel data setting is challenging and is not considered in this paper.

Even if similar results are established for the proposed panel cross-validation method, the convergence rate would not be as fast as desired under the two-way dependence driven by underlying factors. Indeed, even under a Gaussian condition on the error term, 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 introduces cluster/temporal dependence within group i and γ_t introduced cluster/cross-sectional dependence within group t. To estimate the high-dimensional vector θ_0, we consider the sample mean estimator $\widehat{\theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$. We can rewrite the estimator through a Hajek projection:

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

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 $j = 1, ..., s$, $\widehat{\theta}_j = O_P\left(\frac{1}{\sqrt{N\wedge T}}\right)$ and $\|\widehat{\theta}-\theta_0\|_2 = \left(\sum_{j=1}^{s}\left(\widehat{\theta}_j - \theta_{0j}\right)^2\right)^{1/2} = O_P\left(\sqrt{\frac{s}{N\wedge T}}\right)$. We see that it is due to the underlying factor structure that prevents the a faster convergence rate. 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 purpose.

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 a orthogonalized score 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 are 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. Note that cross-fitting algorithm is similar to the cross-validation algorithm and they share the same "approximate" independence results. However, they serve for different purposes and the usage of the sub-samples are used differently.

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 is $o\left(\sqrt{N \wedge T}\right)$ in terms of $L2$ norm under the set of main assumptions. Commonly, this is not a demanding requirement, but when the two-way dependence is driven by underlying factors, even the oracle rate of convergence, $O_P\left(\sqrt{\frac{s}{N \wedge T}}\right)$, would not suffice. Alternatively, with an extra m-dependence condition where m is a finite number, I show that the rate requirement can be relaxed to $o\left((N \wedge T)^{-1/4}\right)$, which makes the first-step estimation through the two-way cluster LASSO feasible under a proper sparsity assumption.

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 an 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 a 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 score $X_{it}U_{it}$ 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 resulted cluster dependence due to the unknown function form of f, it does suggest that we would like an 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, a 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 conduct the analysis by conditioning their realized values, and they will be estimated as 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 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 generalize 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 those effects themselves are zero.

It sounds like a good solution where researcher proxy the unobserved effects by observable random variables, which avoids 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 made by the two-way cluster LASSO or panel cross-validation 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 two-way clustered panel should be 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 paper 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 as 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 still well below the sample size 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 in its asymptotic properties under two-way dependence. Section 3 introduces a sub-sampling scheme that allows within-cluster dependence and weak dependence across clusters. This sub-sampling scheme will be used for both cross fitting and cross validation. 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 in 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 are competed with each other as well as existing approaches. In Section 7, the empirical estimation of 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. (2016a) 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 model typically relies on some bias-correction to account for the regularization bias. It take various forms in the literature: for example, the low-dimensional projection adjustment in Zhang and Zhang (2014), the de-sparsification procedure in Van de Geer et al. (2014), the decorrelating matrix adjustment in Javanmard and Montanari (2014), the double selection approach in Belloni et al. (2014) ,the decorrelated score construction in Ning and Liu (2017), the Neyman orthogonal score construction in Chernozhukov et al. (2018a, 2022a). The last strand of the literature is often labelled as the debiased machine learning (DML) approach, which is closely related to previous semi-parametric literature including Ichimura (1987), Robinson (1988), Powell et al. (1989), Newey (1994), and Andrews (1994). The idea of the Neyman orthogonal score 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 using the orthogonal score, 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 scores. 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. In order 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. The validity has been shown borrowing the technical tools introduced in Semenova et al. (2023a). Such sub-sampling schemes are also broadly used for other purposes. For example, a similar sub-sampling scheme is also used for cross validation in choosing penalty level λ for LASSO approach.

As for the unobserved heterogeneity issue in high-dimensional panel models, it has been considered in Belloni et al. (2016a); Kock and Tang (2019); Vogt et al. (2022); Gao et al. (2024) among others. In Belloni et al. (2016a), 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 assuming sparsity on the fixed effects might be not very natural in certain applications since it means that only 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 the 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 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 $1 \leq q < \infty$, and $\|\mathbf{x}\|_\infty := \max_{1\leq j\leq k} |x_j|$. The $L^q(P)$ norm is denoted as $\|f\|_{P,q} := \left(\int \|f(\omega)\|^q dP(\omega)\right)^{1/q}$ where f is a random element with probability law P. I denote the empirical average of f_{it} over $i = 1, ..., N$ and $t = 1, ..., T$ as $\mathbb{E}_{NT}[f_{it}] = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T f_{it}$ and the empirical L^2 norm as $\|f_{it}\|_{NT,2} = \left(\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \|f_{it}\|^2\right)^{1/2}$. Correspondingly, I denote the empirical average of f_{it} over the sub-sample $i \in I_k$ and $t \in S_l$ as $\mathbb{E}_{kl}[f_{it}] = \frac{1}{N_k T_l}\sum_{i\in I_k, t\in S_l} f_{it}$, 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 and LASSO via Panel Cross Validation

Little is known in terms of statistical properties for high-dimensional methods under dependence in both space and time. In this section, a candidate based on $l1$-regularization methods, also known as the LASSO, will be examined. Particularly, I will propose a weighted LASSO where the regressor-specific penalty weights are based on some self-normalization scheme robust to two-way dependence and the common penalty level is theoretically driven. Such an LASSO approach is named as two-way cluster-LASSO, corresponding to the heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016a).

To study this LASSO approach, I consider a high-dimensional regression model of panel data $\{Y_{it}\}$ and $\{X_{it}\}$ with a stochastic error $\{V_{it}\}$. Before I specify the model, I will first introduce the underlying data

generating process of $W_{it} := (Y_{it}, X_{it}, V_{it})$ that characterize the two-way cluster dependence:

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

$$W_{it} = \mu + f\left(\alpha_i, \gamma_t, \varepsilon_{it}\right), \ \ \forall i \geq 1, t \geq 1, \tag{2.1}$$

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, due to sharing the same cross-sectional cluster, W_{it} and W_{is} are dependent for any i, t, s. Similarly, due to sharing the same temporal cluster, W_{jt} are dependent for any t, i, j. Furthermore, even if sharing neither the cross-sectional or temporal cluster, observations can also be correlated due to correlated time effects γ_t. It is important to notice that the components in 2.1 simply characterize the dependence in panel data in a fairly general. Differing from factor models or unobserved heterogeneity, they do not affect identification of the regression model in anyway.

Now let's consider the high-dimensional regression model as follows: for $i = 1, ..., N$ and $t = 1, ..., T$,

$$Y_{it} = f(X_{it}) + V_{it}, \ \mathrm{E}[V_{it}|X_{it}] = 0 \tag{2.2}$$

where f is an unknown function of potentially high-dimensional covariates X_{it}. Since f is an infinite dimensional parameter, the regression model is not exactly sparse. I take a sparse approximation approach as in Belloni et al. (2012):

Assumption ASM (Approximate Sparse Model)**.** *The unknown function f can be well-approximated by a*

[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).

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. Comparing 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. (2016a) 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 restrict the growth of nonzero slope coefficients of the model in a similar way to the perfect temporal dependence case in Belloni et al. (2016a).

Under Assumption ASM, we can rewrite the model 2.2 as

$$Y_{it} = f_{it}\zeta_0 + r_{it} + V_{it},\ \mathrm{E}[V_{it}|X_{it}] = 0 \tag{2.3}$$

Using , we can estimate ζ_0 allowing its dimension to be greater than the sample size by applying $l1$ normalization in the least squared error problem. Let λ be some common penalty level and ω be some desired $p \times p$ diagonal matrix of regressor-specific penalty weights which may not be feasible. Let ω be some diagonal matrix of penalty weights. Consider the following generic weighted LASSO estimator:

$$\widehat{\zeta} = \arg\min_{\zeta} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - f_{it}\zeta)^2 + \frac{\lambda}{NT}\|\omega^{1/2}\zeta\|_1. \tag{2.4}$$

To reserve the desirable properties of such estimators under two-way dependence, we need to construct a weight matrix ω robust to the dependence and derive a common penalty level λ that is large enough but smallest possible trade-off regularization bias for smaller variance. To obtain the rate of convergence of such estimators, one common approach is through verifying the following condition:

$$\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2} f_{it,j}V_{it}\right| \leq \frac{\lambda}{2c_1 NT}. \tag{2.5}$$

Condition 2.5 is referred to as the "regularization event" in the literature. It is a common condition that

appears in asymptotic analysis of LASSO approaches. Intuitively, it rules that the penalty level λ should be as large as the noise due to the stochastic error and the large number of regressors. Under the event 2.5, finite sample bounds in terms for the LASSO estimators are available in the literature. See for example, Lemma 6 of Belloni et al. (2012).

However, verifying event 2.5 happening with high probability is challenging. Due to the high dimensionality of f_{it}, conventional central limit theorem approximation is unable to provide a fast enough convergence rate for the term on the left-hand side. In earlier literature, Gaussian or sub-Gaussian errors are often assumed for using Gaussian tail inequality to show condition 2.5 (see Bickel et al., 2009, Bühlmann and Van De Geer, 2011, and Theorem 29.3 of Hansen, 2022). In Belloni et al. (2012, 2014, 2016b), the regularization event is shown by utilizing a moderate deviation theorem (see Jing et al., 2003 and Peña et al., 2009) for self-normalized sums without relying on Gaussian or sub-Gaussian properties. However, this approach is not feasible under two-way dependence. Instead, I decompose $f_{it,j}V_{it}$ using Hajek projection components and utilize moderate deviation theorems for both i.i.d and dependent sums for each Hajek projection components separately.

For that purpose, I propose the following common penalty level λ and (infeasible) penalty weights:

$$\lambda = 6c_1 \frac{NT}{(N \wedge T)^{1/2}} \Phi^{-1}\left(1 - \frac{\gamma}{2p}\right), \tag{2.6}$$

$$\omega_j = \frac{N \wedge T}{N^2} \sum_{i=1}^{N} a_{i,j}^2 + \frac{N \wedge T}{T^2} \sum_{b=1}^{B} \left(\sum_{t \in H_b} g_{t,j} \right)^2. \tag{2.7}$$

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 $j = 1, ..., p$; $B = \mathrm{round}(T/h)$, $h = \mathrm{round}(T^{1/5}) + 1$, and, for $b = 1, .., B$, $H_b = \{t : h(b-1) + 1 \leq t \leq hb\}$. γ is a tuning parameter which should be sufficiently small. Technically, it is chosen according to

$$\log(1/\gamma) \simeq \log(p \vee NT). \tag{2.8}$$

To utilize the moderate deviation theorem for weakly dependent sums due to Gao et al. (2022), a regularity condition is needed to restrict the weak dependence of $\{\gamma_t\}$. To be more specific, we need to introduce a few more concepts and notations. Let (X, Y) be random element 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:

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

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

$$\beta_\gamma(q) = \sup_{s\leq T} \beta\left(\{\gamma_t\}_{t\leq s}, \{\gamma_t\}_{t\geq s+q}\right) \leq c_\kappa exp(-\kappa q), \forall q \in \mathbb{Z}^+, \tag{2.9}$$

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

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).

To implement the penalty weights in 2.7, 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. Firstly, V_{it} is unknown and so we will need some initial estimation to replace V_{it} with some residual $\tilde{V}_{it}$. Belloni et al. (2012) propose an iteration process for constructing the penalty weights. Secondly, by replacing V_{it} with $\tilde{V}_{it}$ from some initial estimation or iteration process, we still do not observe $\mathrm{E}[f_{it,j}\tilde{V}_{it}|\alpha_i]$ or $\mathrm{E}[f_{it,j}\tilde{V}_{it}|\gamma_t]$. Common estimators for these quantities are the temporal sample mean and cross-sectional sample mean of $f_{it,j}\tilde{V}_{it}$, respectively. The validity of those estimators is only established for exchangeable arrays (see, for example, Menzel (2021) and Chiang et al. (2023a)). In our panel data setting, the exchangeability fails. Moreover, the second term in 2.7, the cluster covariance estimator for the long-run variance, doesn't perform very well when the temporal sample sizes are small, which is likely to be the case with the cross-fitting sub-samples. As shown in some unreported simulation studies, the performance of the two-way cluster-LASSO estimator and the corresponding panel DML estimator is indeed not satisfying. Therefore, I propose an alternative construction of infeasible penalty weights which takes the form of a two-way cluster-robust standard error, named in this paper as CHS estimator (Chiang et al., 2024)[6]:

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

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

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

[6] 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).

This set of infeasible penalty weights has several advantages over the one defined in 2.7: (1) Given V_{it}, it avoids the estimation of the unobserved components so it tends to have better finite sample performance when the sample sizes are reasonably small. (2) Applying the results given by Bester et al. (2011) and Chiang et al. (2024), it is straightforward to show the two sets of infeasible penalty weights have the same probability limit, i.e. $\omega_j = \omega_j^{\text{CHS}} + o_P(1)$ for $j = 1, ..., p$, under regularity conditions given in this paper and a non-degeneracy condition. (3) When the non-degeneracy condition is violated, e.g. i.i.d data, the two term penalty weights given by 2.7 would be too conservative, but the three term penalty weights given by 2.10 is still robust because using the results in Chiang et al. (2024) one can show 2.10 has the same probability limit as the infeasible penalty weights given in Belloni et al. (2012) under the i.i.d case.

Again, V_{it} is unknown, so to make it feasible, V_{it} needs to be replaced by some initial estimate $\tilde{V}_{it}$ and estimate ω^{CHS} iteratively. Following Belloni et al. (2012), I set the initial residual as $\tilde{V}_{it} = Y_{it} - \mathbb{E}_{NT}[Y_{it}]$ and $\tilde{v}_{it} = f_{it}\tilde{V}_{it} - \mathbb{E}_{NT}[f_{it}\tilde{V}_{it}]$. Then, $\tilde{V}_{it}$ will be updated iteratively by the residuals from the estimation in 2.4 and . The validity of iterative estimation of the penalty weights is given in Belloni et al. (2012) under high-level assumptions. To avoid extra complexity of the presentation, I follow previous literature (Belloni et al., 2012, 2016a) and impose a high-level assumption on the feasible penalty weights: Let $\widehat{\omega}$ be the feasible diagonal weights. There exists some constants c_1 and u such that $c_1 > 1$ and $1 \leq u < \infty$ and

$$\frac{1}{c_1}\omega_j^{1/2} \leq \widehat{\omega}_j^{1/2} \leq u\omega_j^{1/2}, \text{ uniformly over j} = 1, ..., \text{p}, \tag{2.11}$$

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

Before I deliver 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 $\text{E}_P[f_{it}f_{it}']$ 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.

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

$$\kappa_1 \leq \min_{\delta \in \Delta(Cs)} \delta' M_f \delta < \max_{\delta \in \Delta(Cs)} \delta' M_f \delta \leq \kappa_2,$$

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

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, and it is also needed in the proof. Here we assume the sparse eigenvalue condition because we need it to show $l2$ convergence rate for the LASSO estimator and it implies the restricted eigenvalue condition as shown in Bickel et al. (2009). 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)* $\log p = o\left((T)^{1/6}/\log T\right)$, $p = o\left((T)^{13/12}/(\log T)^{1/2}\right)$. *(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* $\mathrm{E}\left[\sum_{t=r}^{r+m} f_{it,j}V_{it}\right]^2 \geq \mu^2 m$ *for all* j *and* $r \geq 0, m \geq 1$. *(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}: Although it still allows the regressor to be larger than the sample size, it requires the number of included regressors to grow slower than that in Belloni et al. (2016a) with $\log^3(p) = o(NT)$. This is because the weak dependence in our setting makes the tail probability decay at a slower rate, as shown in the results of Gao et al. (2022). The first condition is binding when the sample size is relatively small and the second condition is binding when the sample size is large. 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, is actually robust to both degeneracy and non-degeneracy. However, the theory leverages the moderate deviation theorems based on the self-normalization weights given by 2.7.

Theorem 2.1. *Suppose Assumptions AHK, ASM, AR, SE, REG hold for model 2.2. Additionally,* λ *is set as 2.6,* $\widehat{\omega}$ *satisfies condition 2.11, and* γ *is chosen according to 2.8. Then, as* $N, T \to \infty$ *jointly with* $N/T \to c$*, we have*

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(f_{it}'\widehat{\zeta} - f_{it}'\zeta_0\right)^2 &= O_P\left(\frac{s\log(p \vee NT)}{N \wedge T}\right), \\
\left\|\widehat{\zeta} - \zeta_0\right\|_1 &= O_P\left(s\sqrt{\frac{\log(p \vee NT)}{N \wedge T}}\right), \\
\left\|\widehat{\zeta} - \zeta_0\right\|_2 &= O_P\left(\sqrt{\frac{s\log(p \vee NT)}{N \wedge T}}\right).
\end{aligned}$$

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 a 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. For example, 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. (2016a) 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 it in another way, what is the minimum requirement of the convergence rates for valid inference in 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 and Cross Validation

To propose a general inference procedure for high-dimensional panel model in the next section and also propose an alternative cross-validation LASSO approach, I will first introduce a new sub-sampling scheme that is robust to two-way cluster dependence and weak dependence over time. 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. By exploiting all possibilities of valid splitting, we are able to obtain multiple pairs of sub-samples by which researches can repeat the intended procedures and average the results. In this paper, this algorithm will be used for both cross-fitting and cross-validation purposes. Before diving into the algorithm, I will first introduce the data structure of interest.

Let $\{W_{it} : i = 1, ..., N \text{ and } t = 1, ..., T\}$ 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 $\{W_{it} : i = 1, ..., N \text{ and } t = 1, ..., T\}$ and denote $P \in \mathcal{P}_{NT}$ as a generic probability law for the sample with sizes (N, T).

Again, 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 dimensions 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. I will first give the general scheme and then combine with the specific purposes, cross fitting and cross validation, to present their corresponding algorithms.

First, let K and L be some positive integers chosen by the researcher (tuning parameters). 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 $\text{floor}(N/K)$ and others equal to $\text{ceil}(N/K)$. and the same applies to the temporal dimension. Then, partition the cross-sectional indices $\{1, 2, ..., N\}$ into K equal-size folds $\{I_1, I_2, ..., I_K\}$ and partition the temporal indices $\{1, 2, ..., T\}$ into L adjacent equal-size folds $\{S_1, S_2, ..., S_L\}$ so that $\bigcup_{k=1}^{K} I_k = \{1, ..., N\}$, $\bigcup_{l=1}^{L} S_l = \{1, ..., T\}$, and the

sub-sample sizes are $N_k = N/K$ and $T_l = T/L$. Let $W(k,l) = \{W_{it} : i \in I_k, t \in S_l\}$ denote one part of sub-sample, which is typically smaller, and the set $W(-k,-l) = \{W_{it} : i \in \bigcup_{k' \neq k} I_{k'}, t \in \bigcup_{l' \neq l, l \pm 1} S_{l'}\}$ as the other sub-sample. 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 $L = 8$.

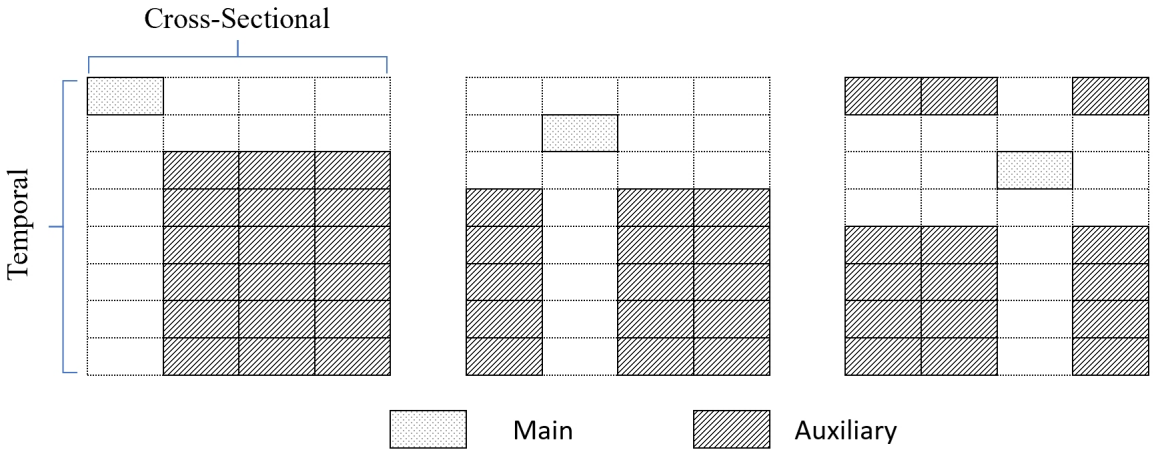

Figure 1: Panel cross fitting with $K = 4$ and $L = 8$. Three graphs from left to right correspond to the main and auxiliary sample constructions with $(k,l) = (1,1)$, $(k,l) = (2,2)$, $(k,l) = (3,3)$. For a simple illustration, observations in the main sample are all adjacent in the cross-sectional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

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. 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).

For the rest of the section, I will illustrate how this sub-sampling scheme is used in cross fitting and cross validation.

3.1. ***Panel Cross-Fitting and DML Algorithm***

One of the primary use 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 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 score of 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 score, $\psi(W_{it};\theta,\widehat{\eta}_{kl})$. By averaging the scores 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 3.1 (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 first step estimation and obtain $\widehat{\eta}_{kl}$, then construct the averages of scores $\overline{\psi}_{kl}(\theta) = \mathbb{E}_{kl}[\psi(W_{it};\theta,\widehat{\eta}_{kl})]$ 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 a high-dimensional nonparametric 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.*

3.2. ***Panel Cross Validation for LASSO***

The other use of resampling scheme is cross validation with panel data. Cross validation is commonly used for model selection, bandwidth choice in nonparametric estimation, and penalty selection in high-dimensional regression. The general idea of the cross validation is to evaluate the model fit based on hold-out estimation. The basic cross validation is based on leave-one-out estimation, but it is computationally costly when the sample size is large. Alternatively, based on leave-one-fold-out estimation, K-fold cross validation has lower computational cost and is less noisy. For both algorithms, the validity of cross validation, i.e. the unbiasedness or consistency of the cross-validation criterion for the expected prediction error, often depends on the independence between the hold-out testing sample and the training sample (see, for example, Shao, 1993 for model selection in linear regression model and Theorem 19.7 of Hansen, 2022 for nonparametric bandwidth selection). For time series data, Burman et al. (1994) and Racine (2000) propose h-block and hv-block cross-validation, respectively. They are based on the same idea that by excluding a growing temporal neighborhood of the testing sample, the training and the testing samples are approximately independent when the data is weakly dependent. This property is exactly provided by the sub-sampling scheme proposed here.

In the context of LASSO approaches, K-fold cross validation is a popular criterion for choosing the penalty term λ in practice due to its computation efficiency and good finite sample performance. The theoretical properties of cross-validated LASSO has been studied in Chetverikov et al. (2021) in a random sampling context. The purpose of this section is to provide a more robust cross-validation algorithm in the panel data context. However, a rigorous theoretical examination of the panel cross-validation LASSO is not pursued here.

We now introduce the the panel cross validation algorithm in the context of the LASSO estimation. Consider a high-dimensional linear regression model $Y = X\beta_0 + U$ where β_0 is a high-dimensional parameter

vector of dimension p. We define the LASSO estimators as a function of λ:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - X_{it}\beta)^2 + \frac{\lambda}{NT} \|\beta\|_1,$$
$$\hat{\beta}_{kl}(\lambda) = \arg\min_{\beta} \frac{1}{N_{-k}T_{-l}} \sum_{i \in I_{-k}} \sum_{t \in S_{-l}} (Y_{it} - X_{it}\beta)^2 + \frac{\lambda}{N_{-k}T_{-l}} \|\beta\|_1,$$

The next definition presents the cross-validation algorithm.

Definition 3.2 (Panel Cross-Validation LASSO)**.**

(i) 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 testing sample

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

and the training sample

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

(ii) Pre-select a list of λ ranging from 0 to $C\sqrt{\frac{NT}{\log(p)}}$ with a large enough C.

(iii) Take an λ in ascending order from the preset list. For each (k, l), fit the penalized regression model using the training sample $W(-k, -l)$ and obtain the average prediction error

$$CV_{kl}(\lambda) = \frac{1}{N_k T_l} \sum_{i \in I_k, t \in S_l} \left(Y_{it} - X_{it}\hat{\beta}_{kl}(\lambda)\right)^2.$$

(iv) Calculate $CV(\lambda) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} CV_{kl}(\lambda)$ and the standard error of $CV(\lambda)$

$$se(\lambda) = \left(\frac{1}{KL(KL-1)} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(CV_{kl}(\lambda) - CV(\lambda)\right)^2 \right)^{1/2}$$

(v) Repeat steps (ii)-(iv) and find the $\hat{\lambda} = \arg\min_{\lambda} CV(\lambda)$ or $\tilde{\lambda} = \max\{\lambda : CV(\lambda) < CV(\hat{\lambda}) + se(\hat{\lambda})\}$.

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 3.1 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 score. 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 3.1. 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:

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

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 Orthogonal Score, 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* $\boldsymbol{P}$*, 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 scores, which covers many applications and the model in Section 5. For nonlinear orthogonal scores, 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 score does not satisfy condition 4.2 exactly due to approximation errors. Assumption DML1(iii) imposes a mild smoothness assumption on the orthogonal score and Assumption DML1(iv) is a common condition for identification.

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

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

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

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

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

Assumption DML2 regulates the quality of the first-step nuisance estimators. Compared to Assumption 3.2 of Chernozhukov et al. (2018a), here r'_{NT} is required to converge at a faster rate $\delta_{NT}/\sqrt{N}$ instead of δ_{NT}. As is shown in the proof of Lemma A.2 in the Appendix, it is due to the across-cluster dependence between the data W_{it} and W_{jr} for $j \neq i$, $r \neq t$ makes the estimation error more persistent, and the mixing condition itself does not allow the error term to decay at a fast enough rate. We seem to compensate for this by requiring a more stringent convergence rate for the nuisance estimator, but this is not the case. To see how it changes the rate requirement of the nuisance estimators, we bound the nuisance estimator by $\|\widehat{\eta} - \eta_0\|_{P,2} \lesssim \varepsilon_{NT}$ where $\varepsilon_{NT} = o(1)$ as $N, T \to \infty$. If $(\theta, \eta) \mapsto \psi(W; \theta, \eta)$ is smooth, we can bound the rates in Assumption DML2(ii) by $r_{NT} \lesssim \varepsilon_{NT}$, $r'_{NT} \lesssim \varepsilon_{NT}$, $\lambda'_{NT} \lesssim \varepsilon_{NT}^2$. Then, $r_{NT} = o(1)$, $r'_{NT} = o\left(N^{-1/2}\right)$, and $\lambda'_{NT} = o\left(N^{-1/2}\right)$ together impose the crude rate requirement for $\widehat{\eta}$, $\varepsilon_{NT} = o\left(N^{-1/2}\right)$. In our setting, since N and T grow jointly at the same rate, we have $N \approx (NT)^{1/2}$, which means that $\varepsilon_{NT} = o\left((NT)^{-1/4}\right)$ which is the same crude requirement made in Chernozhukov et al. (2018a). This is because the main theorem below on the asymptotic normality of the second-step estimator also has a slower convergence rate, which translates into a less stringent requirement for the first-step estimator.

As briefly discussed in the introduction and explicitly shown in Section 2, however, this rate requirement, $\varepsilon_{NT} = o\left((NT)^{-1/4}\right)$, is not satisfied for common estimators for a high-dimensional regression under two-way cluster dependence. To obtain a less stringent rate requirement, one proposal would be add an extra m-dependent condition beside the beta-mixing condition on γ_t with a finite m, which states that the sequence γ_t is still beta-mixing as in Assumption AR but the dependence is strictly 0 for observations that are more than m periods apart in temporal dimension. Indeed, it is a stronger assumption on the unobserved time effects but the m-dependence is also common in literature and probable in practice. Under this extra regularity condition, I can replace the rate requirement for r'_{NT} by δ_{NT}. In that case, $\lambda'_{NT} = o(N^{1/2})$ will be the demanding rate requirement and it imposes that $\varepsilon_{NT} = o(N^{-1/4})$, 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 a weakest possible rate requirement for the first-step estimator, i.e. $\varepsilon_{NT} = o(1)$. In the Appendix, I define the Assumption DML2', which basically replace the rate requirement for r'_{NT}, as a weaker alternative assumption of Assumption DML2. The next main theorem presents the asymptotic normality under these two set of assumptions discussed above.

Theorem 4.1 (Asymptotic Normality and Variance)**.** *Suppose Assumptions AHK, AR, ND, DML1 hold for*

any $P \in \mathcal{P}_{NT}$, and, either Assumption DML2 or DML2' holds, 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

$$V := A_0^{-1} \Omega A_0^{-1'},$$
$$\Omega := \Lambda_a \Lambda_a' + c\Lambda_g \Lambda_g'.$$

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 of 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 slower. It is a common feature in the literature of robust inference with cluster dependence[7] 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 define the CHS-type variance estimator as follows:

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

where, 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

[7] 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.

parameter M chosen from 1 to T_l,

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

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 hold for any $P \in \mathcal{P}_{NT}$, and, either Assumption DML2 or DML2' holds with some $p > 4$ (defined in Assumption DML2 or DML2'), and $M/T^{1/2} = o(1)$. Then, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$,*

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

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 score.

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 non-trivial number of times under certain data generating processes. Accordingly, an alternative two-term variance estimator was proposed in Chen and Vogelsang (2024). Following the same idea, I propose an alternative variance estimator by dropping the double-counting term $\widehat{\Omega}_{\mathrm{NW}}$:

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

The estimator is referred to as the DKA variance estimator because it is a sum of Driscoll-Krray and Arellano variance estimators.[8] Similar approaches can be found in MacKinnon et al. (2021). It relies on the

[8]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

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 toolkit 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 $t = 1, ..., T$,

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

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}

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.

is a $k \times 1$ vector of control variables[9] and $k >> NT$ is allowed. $\mathbf{X}$ is a $k \times NT$ matrix that stacks X_{it} over $i = 1, ..., N$ and $t = 1, ..., T$. c_i and d_t are treated as random variables. Apparently, model 5.1 includes 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 5.15 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 $t = 1, ..., T$,*

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

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

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

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. 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

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

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 $(X_{it}, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)$. 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:

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

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

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

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 $L_{1,it} = L^\tau(X_{it}, \bar{X}_i, \bar{X}_t)$, for which we denote the dimension as p, and a vector of unobserved regressors as $L_{2,it} = L^\tau(X_{it}, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d) \backslash L^\tau(X_{it}, \bar{X}_i, \bar{X}_t)$. 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 $L^\tau(X_{it}, \bar{X}_i, \bar{X}_t, \epsilon_i^c, \epsilon_t^d)\eta_Y = L_{1,it}\eta_{Y1} + L_{2,it}\eta_{Y2}$. 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 $\mathrm{E}[L_{2,it}|X_{it}, \bar{X}_i, \bar{X}_t] = \mathrm{E}[L_{2,it}]$. 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 $\mathrm{E}[V_{it}^Y|X_{it}, \bar{X}_i, \bar{X}_t] = \mathrm{E}[V_{it}^D|X_{it}, \bar{X}_i, \bar{X}_t] = \mathrm{E}[V_{it}^z|X_{it}, \bar{X}_i, \bar{X}_t] = 0$ 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 are 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* $l = 1, ..., L$,

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

where $(\epsilon^c_{i,l}, \epsilon^d_{t,k})$ *are independent of* $\mathbf{X}$*;* $\bar{X}_{t,k} = 1/N_k \sum_{i \in I_k} X_{it}$ *and* $\bar{X}_{i,l} = 1/T_l \sum_{t \in S_l} X_{it}$.

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 $\mathbf{X}$ 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 $c_i = h_c(\bar{X}_{i,1}) + \epsilon^c_{i,1}$ and $c_i = h_c(\bar{X}_{i,2}) + \epsilon^c_{i,2}$, and it implies that $h_c(\bar{X}_{i,1}) - h_c(\bar{X}_{i,2}) = \epsilon^c_{i,2} - \epsilon^c_{i,1}$ almost surely. Observe that $\epsilon^c_{i,1}$ is a function of $(\bar{X}_{i,1}, \bar{X}_{i,2}, \epsilon^c_{i,2})$ and so the independence between $\epsilon^c_{i,1}$ and $\mathbf{X}$ is only possible in very unique 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:

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

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. 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 $(i, t) \in W(-k, -l)$,

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

where $\nu_{i,-l} := \bar{X}_i - \bar{X}_{i,-l}$ and $\nu_{t,-k} := \bar{X}_t - \bar{X}_{t,-k}$ are the vectors of approximation errors. Let $\nu_{i,-l,j}$ be the j-th entry of $\nu_{i,-l}$ for $j = 1, ..., p$, and, similarly, denote $\nu_{t,-k,j}$ as the j-th entry of $\nu_{t,-k}$. If X_{it} are independent over (i, t), then the fasted convergence rate one can obtain for those approximation errors is $\nu_{i,-l,j} = O_P(T^{-1/2})$ and $\nu_{t,-k,j} = O_P(N^{-1/2})$ 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

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

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 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 results 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

$$Y_{it} = f_{it}\beta_0 + r^Y_{it} + V^Y_{it}, \ \mathrm{E}[V^Y_{it}|X_{it}, \bar{X}_i, \bar{X}_t] = 0, \tag{5.12}$$

$$D_{it} = f_{it}\pi_0 + r^D_{it} + V^D_{it}, \ \mathrm{E}[V^D_{it}|X_{it}, \bar{X}_i, \bar{X}_t] = 0, \tag{5.13}$$

$$Z_{it} = f_{it}\zeta_0 + r^Z_{it} + V^Z_{it}, \ \mathrm{E}[V^Z_{it}|X_{it}, \bar{X}_i, \bar{X}_t] = 0, \tag{5.14}$$

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 $\zeta_0 := (\eta_{Z1}, \mathrm{E}[L_{2,it}]\eta_{Z2})'$. Remind 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 analogue 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* $\Lambda_a \Lambda_a' = \mathrm{E}[a_i a_i']$, $\Lambda_b \Lambda_b' = \sum_{l=-\infty}^{\infty} \mathrm{E}[g_t g_{t+l}']$, *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 are 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* $\xi_0 = \zeta_0$, β_0, and π_0 *obeys* $\|f_{it}(\hat{\xi} - \xi_0)\|_{NT,2} = o_P\left((N \wedge T)^{-1/2}\right)$ *and* $\|\hat{\xi} - \xi_0\|_2 = o_P\left((N \wedge T)^{-1/2}\right)$. *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

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

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 sectio, 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 may produces first-step estimator with desirable convergence rates. Remind that the rate requirement imposed by DML2 in Theorem 4.1 is not achievable but, under Assumption DML2', it 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 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.

Table 5.1: Summary of Findings from Sections 2 - 5

Panel Cross Fitting	GMD	Not valid	
	GMD' or Linearity	DML2	Not achievable rates
		DML2'	$s = o\left(\frac{(N\wedge T)^{1/2}}{\log(p\vee NT)}\right)$
No Cross Fitting		First-step: $o_P\left((N \wedge T)^{-1/2}\right)$	

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.

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 and approximation errors. Specifically, I consider the following triangular model, simplified from the partial linear model in Section 5:

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

where $\theta_0 = 1$ and $\beta_0 = \pi_0 = (2^{-q}, 3^{-q}, ..., (s+1)^{-q}, 0, ..., 0)'$ are p-dimensional parameter vectors where the first s entries are non-zero; q is another sparsity parameter that rules the rates of polynomial decay.

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

$$\begin{aligned} X_{it,j} &= \alpha_{i,j} + \gamma_{t,j} + \varepsilon_{it,j}, \\ U_{it} &= \sum_{j=1}^{p} \left[\alpha_i^U \gamma_{t,j} + \alpha_{i,j}\gamma_t^U\right] + \varepsilon_{it}^U, \\ V_{it} &= \sum_{j=1}^{p} \left[\alpha_i^V \gamma_{t,j} + \alpha_{i,j}\gamma_t^V\right] + \varepsilon_{it}^V, \end{aligned}$$

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 the standard normal distribution 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 $\iota^{|j-k|}$ in the (j,k)'s entry for $j, k = 1, ..., p$; 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 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 or, equivalently (see Wooldridge (2021)), the Mundlak approach in a linear panel model may not eliminate the underlying components. To see why the score $U_{it}V_{it}$ 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 score. The comparison will be among procedures with and without cross fitting. The first-step estimations will be based on the two-way cluster-LASSO estimator, a non-weighted LASSO estimator with cross validation, and the POLS estimator (when feasible). The CHS-type and DKA-type variance estimators (different formulas for estimations with and without cross fitting) will be used for obtaining sample coverage probabilities. I also compare CHS/DKA-type estimators with Eicker-Huber-White (EHW) type estimator from Chernozhukov et al. (2018a) for random sampling data and Cameron-Galbach-Miller (CGM) type estimator from Chiang et al. (2022) for multiway clustered data. Theoretically, only the approaches with the first step through the two-way cluster-LASSO estimator and variance estimation using CHS/DKA have been shown valid (with or without cross fitting).

Results are obtained across DGPs varied by the sample sizes (N, T), the dimensions of covariates p, the number of non-zero slope coefficients s (or the ratio $\frac{s}{NT}$), the other sparsity parameter q, 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, inputs for the feasible weights estimation parameter c, the conservative level parameter 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 is 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:

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

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

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.

Table 6.1: Simulation results for DGP with $N = T = 25$, $p = s = 500$, $q = 2$, $\rho = 0.75$, and $\iota = 0.5$

Cross Fitting	First-Step Estimator	Bias	SD	RMSE	Coverage (%) EHW	CGM	CHS	DKA
No	POLS	0.011	0.209	0.209	33.0	46.0	35.8	45.2
	10-fold CV LASSO	0.003	0.195	0.195	37.3	78.6	78.0	81.6
	two-way cluster-LASSO	0.009	0.196	0.196	40.1	80.8	80.0	83.4
Yes	10-fold CV LASSO	0.002	0.218	0.218	35.2	82.0	86.2	89.6
	two-way cluster-LASSO	0.002	0.225	0.225	37.6	82.2	84.8	89.2

Note: Simulation results are based on 500 Monte Carlo replications. The tuning parameters are $(K, L) = (2, 4)$, $m = 5$, $c = 1.1$ and $\gamma = 0.1$. The nominal coverage probability is 0.95.

Table 6.1 presents the first set of results that compares different estimation and inference procedures under a small sample size with a large number of regressors and relatively strong dependence. The estimation is based on the orthogonal moment condition given by 5.15 with $Z_{it} = D_{it}$ and $f_{it} = X_{it}$. In this exact linear model, the orthogonal moment condition using POLS first step produces estimator algebraically equivalent to POLS estimation of the outcome equation directly by Frisch–Waugh–Lovell Theorem. Note that POLS estimator is not feasible with cross fitting since the sub-sample size is smaller than the number of nuisance parameters. For the rest of the approaches, we find that the sample biases and standard deviations are not very distinguishable across different approaches but the estimator with POLS first step has a deficient sample coverage no matter what variance estimators are used. This can be seen in the proof of Theorem Theorem 5.2 that the error term R^2_{NT} contains terms such as $V^D_{it}(\beta_0 - \tilde{\beta})$ which is not mean 0 and may not vanish fast enough with POLS estimates $\tilde{\beta}$. With LASSO based methods, the overfitting bias is less severe; With the help of cross-fitting algorithm, it shows the performance is further improved in terms of sample bias and, especially, the sample coverage, although in the cost of slight efficiency loss. Table 6.2 presents results under the same DGP except that the temporal dependence of the common time effects is much weaker. Correspondingly, we observe that the sample coverages are closer to the nominal rates for all methods; the bias and standard deviations are also dropped noticeably.

Table 6.2: Simulation results for DGP with $N = T = 25$, $p = s = 500$, $q = 2$, $\rho = 0.25$, and $\iota = 0.5$

Cross Fitting	First-Step Estimator	Bias	SD	RMSE	Coverage (%) EHW	CGM	CHS	DKA
No	POLS	0.01	0.209	0.209	47.1	49.9	48.0	49.8
	10-fold CV LASSO	0.007	0.154	0.154	49.1	91.6	89.0	91.4
	two-way cluster-LASSO	0.005	0.150	0.151	49.4	92.0	89.0	91.6

Note: Simulation results are based on 500 Monte Carlo replications. The tuning parameters are $(K, L) = (2, 4)$, $m = 5$, $c = 1.1$ and $\gamma = 0.1$. The nominal coverage probability is 0.95.

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 paper on American Economic Review and it investigates one classic quantity of interest in economics: the

Table 6.3: Simulation results for i.i.d data with $N = T = 25$, $p = s = 500$, $q = 2$, and $\iota = 0.5$

Cross Fitting	First-Step Estimator	Bias	SD	RMSE	Coverage (%) EHW	CGM	CHS	DKA
No	POLS	0.01	0.209	0.209	47.1	49.9	48.0	49.8
	10-fold CV LASSO	0.007	0.154	0.154	49.1	91.6	89.0	91.4
	two-way cluster-LASSO	0.005	0.150	0.151	49.4	92.0	89.0	91.6

Note: Simulation results are based on 500 Monte Carlo replications. The tuning parameters are $(K, L) = (2, 4)$, $m = 5$, $c = 1.1$ and $\gamma = 0.1$. The nominal coverage probability is 0.95.

government spending multiplier. The question is that 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 the government spending and the output. The parameter of interest is a scalar and the baseline model does not even need a control for identification, so why it is relevant and where the high dimensionality comes from? As it will be revealed very soon, indeed, the high dimensionality from heterogeneity and flexible modelling can be hidden in settings to which researcher 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 a instrumental variable (IV) approach. As argued by the authors, the national military spending is largely determined by geopolitical event 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, two types of share-shift/Bartik IV methods are considered: the first one estimates the share directly by regressing the regional military spending on the national military spending allowing for region-specific constant slope coefficients.[10] The second IV takes a simpler approach that uses the fraction of five-year average military spending in the region relative to the five-year average national military spending in the first five years of the sample as the share. To fucos on the main idea, we take both shares as given and the resulting instruments as observable regressors instead of generated regressors to avoid further complication.

In this paper, I extend the linear model with fixed 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 Bartik IV. Specifically, the baseline model from the original study and the

[10] 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 working age population as a robustness check.

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 effect. 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 $f_{it} = (L^\tau(X_{it}, W_t, \bar{X}_i, \bar{X}_t), 1)$ and (β, π, ζ) are associated slope coefficients defined the same as 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. One reason could be that allowing the unit specific slope coefficients for controls generates too many nuisance parameters: With 51 state groups[11], one control would increases 51 parameters and two controls would generate 102 parameters, given no interactions or higher order terms. With the sample size less than 2000, it could substantially affect the variance of the estimator and cause the estimates of θ_0 to be too noisy to be meaningful. In this paper, to obtain a more precise estimate and make the excludability assumption of the IV to be more plausible, beside 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[12]; But it could affect the treatment and the outcome. I also consider military spending in other regions of the world as additional controls in W_t. Remind that the identification relies on the assumption that the military spending in the US is mostly driven by the geopolitical events. If we control the military spending as a fraction of the GDP which is a proxy for those

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

[12] 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.

geopolitical events, then we would expect the estimate to be more precise.

By considering more flexible function forms and additional exogenous control variables, the excludability condition of the instruments is more plausible. However, we don't know which regions triggers the most response of the US military procurement and the inclusion of additional irrelevant variables would simply increase the noise in estimation. Moreover, with the sparse approximation through a polynomial transformation of the original observables, the number of nuisance parameters increases in a polynomial rate. These concerns necessitates the use of high-dimensional selection methods. On the other hand, 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, which calls for 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 are from the FRED database of 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). The additional military spending data for other regions of the world is from the Yearbook: Armaments, Disarmament and International Security by Stockholm International Peace Research Institute (SIPRI).

Table 7.1: Estimates of the open economy relative multiplier from the original model.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Unobs.	Oil	Real		First	IV 1	CHS	DKA	IV 2	CHS	DKA
Heterog.	Price	Int.	Pop.	Stage	$\widehat{\theta}$	s.e.	s.e.	$\widehat{\theta}$	s.e.	s.e.
	No	No	No	POLS	1.43	0.68	0.81	2.48	1.23	1.56
	Yes	No	No	POLS	1.30	0.56	0.72	2.34	1.15	1.43
Fixed Effects	No	Yes	No	POLS	1.40	0.57	0.71	2.41	0.97	1.33
	Yes	Yes	No	POLS	1.27	0.45	0.71	2.26	0.93	1.21
	Yes	Yes	Yes	POLS	1.36	0.43	0.56	2.14	0.77	1.03

Note: The data is a balance panel with $(N, T) = (51, 39)$. Full sample is used for all steps. 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 choice of IVs and 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 estimate results are matched with those given in Nakamura and Steinsson (2014) with significant difference in the standard errors. It is because the variance estimates here accounts for the potential two-way dependence while the variance estimator used in Nakamura and Steinsson (2014) assumes cross-sectional independence. The estimates given by the second IV appear uniformly larger and more noisy than those produced by the first IV. It is

argued in Nakamura and Steinsson (2014) that the second IV uses an imperfect proxy for the responsiveness of states to the national military spending shocks. So the analysis in this paper will also focus on the results due to the first IV.

Table 7.2: Estimates of the open economy relative multiplier from the extended model with .

(1) Unobs. Heterog.	(2) Poly. Transform.	(3) Param. Gen.	(4) First Stage	(5) Param. Sel. for Y	(6) IV 1 $\widehat{\theta}$	(7) CHS s.e.	(8) DKA s.e.	(9) IV2 $\widehat{\theta}$	(10) CHS s.e.	(11) DKA s.e.
			POLS	8	2.27	0.87	0.95	3.09	0.97	1.14
Mundlak	None	8	10-Fold LASSO	8	2.27	0.87	0.96	3.09	0.98	1.15
			TW CL LASSO	0	0.57	0.80	0.85	0.65	0.82	0.89
			POLS	44	1.47	0.41	0.59	2.03	1.02	1.27
Mundlak	2nd	44	10-Fold LASSO	44	1.47	0.45	0.61	2.03	0.96	1.21
			TW CL LASSO	6	0.75	0.85	0.90	0.80	0.87	0.94
			POLS	164	1.45	0.71	0.83	2.41	1.38	1.72
Mundlak	3rd	164	10-Fold LASSO	76	1.57	0.69	0.81	2.64	1.52	1.82
			TW CL LASSO	33	1.44	0.66	0.71	2.14	0.73	0.84

Note: Note: The data is a balance panel with $(N, T) = (51, 39)$. Full sample is used for all steps. Standard errors and feasible penalty weights are calculated with the truncation parameter M chosen by the min-MSE rule given in Section 6. Number of parameters generated by the polynomial transformation and selected by the LASSO-based methods are reported in columns (3) and (5). The tuning parameters for the two-way cluster-LASSO are chosen as $m = 5$, $c = 1.1$ and $\gamma = 0.1$.

The main comparison is done in Table 7.2. Besides the three controls in the original model, an additional set of military spendings in other regions/countries are added. With more controls and the polynomial transformation of the observables, the standard errors are generally larger than those in 7.1. Surprisingly, the estimates under the 3rd order polynomial transformation are quite consistent across different approaches and and consistent with the results in Table 7.1, with estimates using the two-way cluster-LASSO first step being slightly less noisy. This could be because both the 10-fold cross validation LASSO and POLS potentially include many irrelevant regressors and result in noisy estimates in the first stage which then translate into overfitting bias and larger variance in the second stage.

The estimated under the 1st or 2nd order polynomial transformations are not very consistent across different methods. When no transformation (or, equivalently, the 1st order polynomials transformation) is used, POLS and unweighted LASSO produce estimates much larger than the baseline results in Table 7.1 while two-way cluster-LASSO produces an estimate much smaller. When the 2nd order polynomial transformation is used, unweighted LASSO continues to select all regressors so its behavior is very similar to POLS, and they all produce estimates consistent with the baseline results. The estimates from the two-way cluster-LASSO, on the other hand, seems to under-select and remains smaller than the baseline results. The discrepancy between different methods disappear when the 3rd order polynomial transformation is used and the estimator for θ_0 using the two-way cluster-LASSO has the least standard error.[13] What's going on here could be the

[13]Recall in the simulation that the test using POLS estimates tend to over-reject due to potentially invalid inference. Therefore, the standard error for estimator using POLS first-step could underestimate the true standard deviation.

rich nonlinearity in the true model unknown to the researcher. When no transformation or the 2nd order polynomial transformation is used, very limited amount of nonlinearity is captured in the misspecified model for estimation. The first order and second order polynomials terms are used as a weak proxy for the nonlinearity in the true model and these weak signals are not well-captured by the two-way cluster-LASSO[14], which is a more conservative selection method relative to unweighted LASSO. However, with 3rd-order polynomial terms considered in the specification, all methods includes enough amount of nonlinearity in the estimation and produces similar estimates consistent with the baseline model, but two-way cluster-LASSO achieves that with much fewer regressors selected and accordingly results in a smaller standard error.

8. Conclusion and Discussion

The inferential theory for high-dimensional model is particularly relevant in panel data setting where the modelling of unobserved heterogeneity commonly leads to high-dimensional nuisance parameters. This paper enriches the toolbox of researchers in dealing with high-dimensional panel models. Particularly, I propose a package of tools that deal with the estimation and inference in high-dimensional panel model that feature in two-way cluster dependence and unobserved heterogeneity. I first develop a weighted LASSO approach that deal with two-way cluster dependence in the panel data and the common penalty level is theoretically driven. Alternatively, a panel cross-validation scheme is also provide for choosing the common penalty level. 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. Indeed, even in an oracle case where the nonzero slope coefficients are known to the researcher, the rates of convergence are still not desirable.

To investigate the minimum rate requirement, I then propose a general inference procedure, panel DML, with a cross-fitting scheme robust to two-way cluster dependence in panel data. It is shown that the rate requirement for the first-step estimator is still not achievable even with the help of cross-fitting. Alternatively, an extra m-dependence condition on the time effects helps relax the rate requirement and makes the two-way cluster LASSO feasible for inference purpose under the two-way cluster dependence. I further consider the complication due to unobserved heterogeneous effects. Due to the potential non-compatibility with the cross-fitting approach, I provide inferential results with or without cross-fitting in a partial linear model that allows for endogenous treatments.

For empirical practice, it is illustrated in the example that dimensionality can be hidden in questions not traditionally considered as high dimensional. I further demonstrate how to apply the toolkit proposed in this paper to allow to a more flexible model and more robust estimation and inference. In practice, when the question is naturally high-dimensional and answered by panel data, then there is no reason not applying the approaches in this paper. When the questions is 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.

[14] It is a common feature for weighted LASSO with self-normalization weights. Also see Belloni et al. (2012).

While both theoretical and simulations results support the proposed approaches, there are some limitations that 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, cross-validation methods are popular in practice and it is shown in the simulation that the proposed panel cross-validation LASSO exhibits desirable and, sometimes, better finite sample properties compared to other approaches. However, the theoretical investigation is not provided in this paper and it is generally limited in the literature. Lastly, many nonlinear high-dimensional methods are critical in empirical studies but their theoretical results of many nonlinear high-dimensional methods remain unknown in panel data settings. For example, the estimation for propensity score can be done in a logistic LASSO to ensure the estimates are bounded in a unit interval but the theoretical results have not been covered in panel data models. Developing high-dimensional methods that incorporates non-linearity in the estimation is also a 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.

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., 2016a. Inference in high-dimensional panel models with an application to gun control. Journal of Business & Economic Statistics 34, 590–605.

Belloni, A., Chernozhukov, V., Wei, Y., 2016b. Post-selection inference for generalized linear models with many controls. Journal of Business & Economic Statistics 34, 606–619.

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., 2011. Inference with dependent data using cluster covariance estimators. Journal of Econometrics 165, 137–151.

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

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

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

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

Chen, K., 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.

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

Chiang, H.D., Hansen, B.E., Sasaki, Y., 2024. Standard errors for two-way clustering with serially correlated time effects Forthcoming Rev. Econ. Stat.

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. The Review of Economics and Statistics 80, 549–560.

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 .

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

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

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

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

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

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

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

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

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American statistical Association 88, 486–494.

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

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

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

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

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

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 A.1, for each (k,l) we have

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

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

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

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

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

□

Lemma A.3 (Minkowski's inequality for infinite sums) *Let* $1\leq p\leq\infty$ *and* $\{f_m\}$ *be a sequence of functions*

in $L^p(\mathbb{R})$, then we have

$$\left\| \sum_{m=1}^{\infty} f_m \right\|_p \leq \sum_{m=1}^{\infty} \|f_m\|_p .$$

Proof of Lemma A.3. By Fatou's lemma, we have

$$\liminf_{n\to\infty} \left\| \sum_{m=1}^{n} f_m \right\|_p^p \geq \left\| \liminf_{n\to\infty} \sum_{m=1}^{n} f_m \right\|_p^p = \left\| \sum_{m=1}^{\infty} f_m \right\|_p^p .$$

Raise both sides to the power of $1/p$ and apply Minkowski's inequality, we have

$$\sum_{m=1}^{\infty} \|f_m\|_p = \liminf_{n\to\infty} \sum_{m=1}^{n} \|f_m\|_p \geq \liminf_{n\to\infty} \left\| \sum_{m=1}^{n} f_m \right\|_p \geq \left\| \sum_{m=1}^{\infty} f_m \right\|_p .$$

□

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

Lemma A.4 (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 $l_m \to \infty$, $\mathrm{P}(\|X_m\| > l_m A_m) \to 0$.*

Appendix B

Proof of Theorem 2.1. In the proof, we will show L1 and L2 convergence rate for $\widehat{\zeta}$. We will first show the regularization event in terms of the infeasible penalty weights ω as defined in 2.7. 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 hold with probability approaching one, we bound the probability of the

following event for each $j = 1, ..., p$:

$$\mathrm{P}\left(\frac{1}{NT}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1NT}\right) = \mathrm{P}\left(\omega_j^{-1/2}\left|\frac{1}{N}\sum_{i=1}^{N}a_{i,j} + \frac{1}{T}\sum_{t=1}^{T}g_{t,j} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it,j}\right| > \frac{\lambda}{2c_1NT}\right)$$

$$\leq \mathrm{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\omega_{a,j}^{-1/2}a_{i,j}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}\omega_{g,j}^{-1/2}g_{t,j}\right| + \left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}e_{it,j}\right| > \frac{\lambda}{2c_1NT}\right)$$

$$\leq \mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{N}\sum_{i=1}^{N}\omega_{a,j}^{-1/2}a_{i,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right) + \mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{T}\sum_{t=1}^{T}\omega_{g,j}^{-1/2}g_{t,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right)$$

$$+ \mathrm{P}\left(\left|\frac{(N\wedge T)^{1/2}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}e_{it,j}\right| > \frac{(N\wedge T)^{1/2}\lambda}{6c_1NT}\right) := p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)$$

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), we have for any $x \in [0, N_k^{1/6}\Xi_{a,j}/l_{a,NT} - 1]$ 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

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

given that $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in [0, N^{1/6}\Xi_{a,j}/l_{a,NT} - 1]$ 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 $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in [0, N_k^{1/6}\min_j\{\Xi_{a,j}\}/l_{a,NT} - 1]$. Under Assumption REG and the assumption that $N/T \to c$, we have $(\log p)^{1/2} = o(N^{1/12}/(\log N)^{1/2})$ as $(N,T) \to \infty$. With the choice $\log(1/\gamma) \lesssim \log(p \vee NT)$, we have

$$N_k^{1/6}\min_j\{\Xi_{a,j}\}/l_{a,NT} - 1 = O(N^{1/6}/l_{a,NT})$$
$$\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \lesssim \sqrt{\log(p/\gamma)} \lesssim \sqrt{\log p + \log(p \vee NT)} = o(N^{1/12}/(\log N)^{1/2})$$

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 $x \in \left(0, d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT} - 1\right)$ 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 $\Phi^{-1}(1-\frac{\gamma}{2p}) \in \left(0, d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT} - 1\right)$. 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 $\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \in (0, d_0(\log T)^{-1/2}T^{1/12}/l_{g,NT} - 1)$. Under Assumption REG and the assumption that $N/T \to c$, we have $(\log p)^{1/2} = o(T^{1/12}/(\log T)^{1/2})$. Under the choice $\log(1/\gamma) \lesssim \log(p \vee NT)$, we have

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

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

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

Therefore, by Chebyshev's inequality we have

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

If $p \leq NT$, then $\sum_{j=1}^{p} p_{3,j}(\lambda) \to 0$. Conciser the case $p > NT$, we have $p_{3,j}(\lambda) \simeq O\left(\frac{1}{T(\log p)^{1/2}}\right)$. Since $p = o\left(T^{13/12}/(\log T)^{1/2}\right)$

$$\sum_{j=1}^{p} p_{3,j}(\lambda) = o(T^{13/12}/(\log T)^{1/2}) \times O\left(\frac{1}{T^{13/12}/(\log T)^{1/2}}\right) = o(1).$$

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

$$\mathrm{P}\left(\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1/2}f_{it,j}V_{it}\right| \leq \frac{\lambda}{2c_1NT}\right) \to 1. \tag{8.1}$$

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

$$\left\|\mathbf{f}\left(\widehat{\zeta}-\zeta_0\right)\right\|_{NT,2}=\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(f_{it}'\widehat{\zeta}-f_{it}'\zeta_0\right)^2\right)^{1/2},$$
$$\left|\omega^{1/2}\left(\widehat{\zeta}-\zeta_0\right)\right|_1=\left|\sum_{j=1}^{p}\omega_j^{1/2}\left(\widehat{\zeta}_j-\zeta_{0,j}\right)\right|.$$

Let δ be some generic vector of nuisance parameters and let J_p^1 be a subset of an index set $J_p=1,...,p$ 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 weighted restricted eigenvalues and the Gram matrix

$$K_C(M_f)=\min_{\delta:\,\|\omega^{1/2}\delta^0\|_1\le C\|\omega^{1/2}\delta^1\|_1,\ \|\delta\|\neq 0,\ |J_p^1|\le s}\frac{\sqrt{s\delta' M_f\delta}}{\left\|\omega^{1/2}\delta^1\right\|_1},$$
$$M_f=\frac{1}{NT}\sum_{i\in I_k,t\in T_l}f_{it}f_{it}'$$

Under Assumption (ASM), the condition 2.11, and 8.1, Lemma 6 of Belloni et al. (2012) implies that

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

where $c_0=\frac{uc+1}{lc-1}>1$. By arguments given in Bickel et al. (2009), Assumption SE implies that $1/K_C(M_f)=O_P(1)$ for any $C>0$. Therefore,

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

To obtain the $l1$ convergence rate, we need to show that $b/a = O(1)$ where

$$a := \min_{j=1,...,p} \omega_j^{1/2},\ b := \max_{j=1,...,p} \omega_j^{1/2}.$$

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

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

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 $N, T \to \infty$.

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.

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

$$0 < \kappa_1 \leq \left(\min_{\delta \in \Delta(m)} (M_f)\right)^{1/2} \leq \frac{(b' M_f b)^{1/2}}{\|b\|_2} = \frac{\left\|\mathbf{f}\left(\widehat{\zeta} - \zeta_0\right)\right\|_{NT,2}}{\|\widehat{\zeta} - \zeta_0\|_2},$$

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

$$\|\widehat{\zeta} - \zeta_0\|_2 \leq \frac{\left\|\mathbf{f}\left(\widehat{\zeta} - \zeta_0\right)\right\|_{NT,2}}{\left(\min_{\delta \in \Delta(m)} (M_f)\right)^{1/2}} = O_P\left(\sqrt{\frac{s \log(p/\gamma)}{N \wedge T}}\right) = O_P\left(\sqrt{\frac{s \log(p \vee NT)}{N \wedge T}}\right)$$

□

Appendix C

Assumption DML2' (Alternative Score Regularity and First-Steps)**.**

(i) For all $i \geq 1$, $t \geq 1$, and some $p > 2$, the following moment conditions hold:

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

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

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

(iii) Additionally, γ_t is independent of γ_{t+m} for any t and $m < \infty$.

Proof of Theorem 4.1. The proof will proceed with Assumption DML2 with the only reference to Assumption DML2' in the proof Lemma A.2 below.

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 $\{(W(k,l), W(-k,-l)) = (\tilde{W}(k,l), \tilde{W}(-k,-l)),\ \textit{for some}\ (k,l)\}$ 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

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

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

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

Claim C.2. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\|\widehat{\bar{\psi}}(\theta_0)-\bar{\psi}(\theta_0,\eta_0)\| = O_P(r'_{NT}+\lambda_{NT}+\lambda'_{NT})$ *under Assumption DML2 and* $\|\widehat{\bar{\psi}}(\theta_0)-\bar{\psi}(\theta_0,\eta_0)\| = O_P(\rho_{NT}/\sqrt{N}+\lambda_{NT}+\lambda'_{NT})$ *under Assumption DML2'.*

By Claim C.2, we have, under Assumption DML2,

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

and under Assumption DML2',

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

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

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

Now, we can combine the results and obtain, under Assumption DML2

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

and under Assumption DML2',

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

as desired.

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

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

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, for the second term on the RHS of the inequality above, we have

$$\left\|\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l))]-A_0\right\|\leq r_{NT}.$$

For the first term, note that

$$\mathrm{E}_P\left[\widehat{A}_{kl}-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]|W(-k,-l)\right]=\mathrm{E}_P\left[\widehat{A}_{kl}|W(-k,-l)\right]-\mathrm{E}_P[\widehat{A}_{kl}|W(-k.-l)]=0.$$

So, it is mean zero by iterated expectation. 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

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

On the event $\mathcal{E}_{cp}$, we have $W(k,l) \perp W(-k,-l)$ and so

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

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

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

Therefore, we have

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

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

$$\begin{aligned}
&\left(\mathrm{E}_P\left[\|\ddot{\psi}_{it}^{a,kl}\|^2 | W(-k,-l)\right]\right)^{1/2} \\
\leq &\left(\mathrm{E}_P\left[\|\psi^a(W_{it}, \widehat{\eta}_{kl})\|^2 | W(-k,-l)\right]\right)^{1/2} + \left(\mathrm{E}_P\left[\|\psi^a(W_{it}, \widehat{\eta}_{kl})\|^2 | W(-k,-l)\right]\right)^{1/2} \\
\leq &2a_1 < \infty,
\end{aligned}$$

where the first inequality follows from conditional Minkowski's inequality and 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

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

Note that β-mixing of γ_t implies α-mixing with the mixing coefficient $\alpha_\gamma(q) \leq \beta_\gamma(q)$, and conditional on $W(-k,-l)$ and $\{a_i\}_{i \in I_k}$, $(\ddot{\psi}_{d,i,t}^{a,kl}, \ddot{\psi}_{d,j,t}^{a})$ is also α-mixing with the mixing coefficient not larger than $\alpha_\gamma(q)$ by Theorem 14.12 of Hansen (2022). Then, we have

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

where the first inequality follows from the law of iterated expectation and the Jensen's inequality; the second inequality follows from Theorem 14.13(ii) of Hansen (2022); the second-to-last inequality follows from the fact that a_i and α_j are independent; the last inequality follows from the moment conditions in Assumption DML2 and that $W(-k,-l)$ is independent of $W(k,l)$ on $\mathcal{E}_{cp}$ with an application of Minkowski's inequality.

Then, we have

$$
\begin{aligned}
(\frac{1}{N_k T_l})^2 a(4) &\leq 8a_1^2(\frac{1}{N_k T_l})^2 \sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j\in I_k} \sum_{d=1}^{D} \alpha_\gamma(m)^{1-2/p} \\
&\leq 8a_1^2 \frac{D}{T_l} \sum_{m=1}^{\infty} c_\kappa exp(-\kappa m)^{1-2/p} \\
&\leq 8a_1^2 c_\kappa \frac{D}{T_l} \frac{1}{exp(\kappa(1-2/p))-1} \to 0, \ as\ T \to \infty,
\end{aligned}
$$

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

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

By Chebyshev's inequality, we conclude

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

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

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

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

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

Then, by triangle inequality we have

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

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

$$\begin{aligned}c(1) &\leq \mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right] = O({r'_{NT}}^2), \\ c(2) &\leq c\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right] = O({r'_{NT}}^2), \\ c(3) &\leq \frac{1}{N_k}\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right] = O({r'_{NT}}^2/N), \\ c(4) &\leq N_k\mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right] = O(N{r'_{NT}}^2) \\ c(5) &\leq \mathrm{E}_P\left[\|\tilde{\ddot{\psi}}_{it}^{kl}\|^2|W(-k,-l)\right]] = O({r'_{NT}}^2).\end{aligned}$$

Note that, under Assumption DML2', $c(4) = \frac{1}{N_k T_l^2}\sum_{l=1}^{m}\sum_{t=\min(S_l)}^{\max(S_l)-l}\sum_{i,j\in I_k}\left|\mathrm{E}_P\left[\langle\tilde{\ddot{\psi}}_{it}^{kl}, \tilde{\ddot{\psi}}_{j,t+l}^{kl}\rangle|W(-k,-l)\right]\right|$ where m is a finite number. Therefore,

$$c(1) = O({\rho_{NT}}^2),\ c(2) = O({\rho_{NT}}^2),\ c(3) = O({\rho_{NT}}^2/N),\ c(4) = O({\rho_{NT}}^2),\ c(5) = O({\rho_{NT}}^2).$$

So, we have shown, under Assumption DML2,

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

which implies $b(1) = O_P(\sqrt{N}r'_{NT})$, and under Assumption DML2', $b(1) = O_P(\rho_{NT})$.

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 $\beta_g(q) \leq \beta_\gamma(q)$ for all $q \geq 1$; 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

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

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

$$\left\|\mathrm{E}_P[e_{it}e_{i,t+q}]\right\| \le \mathrm{E}_P\left\|\mathrm{E}_P[e_{it}e_{i,t+q}|\alpha_i]\right\| \le 8\left(\mathrm{E}_P\|e_{it}\|^p\right)^{2/p}\alpha_g(q)^{1-2/p}. \tag{8.2}$$

Then, similarly, we have

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

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 $\{a_i\}_{i\ge1}$ 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 $\sum_{q=1}^{\infty} \beta_g(q)^{1-2/p} < \infty$ 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. Under either set of the assumptions specified in Theorem 4.2, we have the results in Theorem 4.1. The proof will proceed with Assumption DML2 since the rate requirement r'_{NT} in Assumption DML2 can be replaced with a weaker rate requirement ρ_{NT} in Assumption DML2' without affecting the proof and the rest of the conditions in Assumption DML2' are weaker than Assumption DML2.

By the same arguments as in the beginning of 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 $\|\widehat{\tilde{A}} - A_0\| = O_P(N^{-1/2} + r_{NT})$ 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

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

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

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

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

Note that $\sum_{t\in S_l, r\in S_l}\beta_\gamma(|t-r|)^{1-2/p} \leq \infty$ 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, $\tilde{I}_{a,2,i}^{(kl)}$ does not vary over i so that $\mathrm{E}_P\left|\frac{1}{N_k}\sum_{i\in I_k}\tilde{I}_{a,2,i}^{(kl)}\right|^2 = \mathrm{E}_P\left|\tilde{I}_{a,2,i}^{(kl)}\right|^2$. Denote $h_i(\gamma_t,\gamma_r) = \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}|\gamma_t,\gamma_r] - \mathrm{E}_P[\psi_{it}^{(0)}\psi_{ir}^{(0)}]$. By direct calculation, we have

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

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

$$\mathrm{E}_P\left|h_i(\gamma_t,\gamma_r)\right|^{2+\delta} < \infty, \tag{8.3}$$

$$\int\int\left|h_i(u,v)\right|^{2+\delta} dF(u)dF(v) < \infty, \tag{8.4}$$

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

Consider condition 8.3. By Minkowski's inequality, Jensen's inequality, and the law of iterated expectation, we have

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

where the second inequality follows from Hölder's inequality and the identical distribution of γ_t. Let $\delta = \frac{p-4}{2}$, then $\left(\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^{4+2\delta}\right)^{\frac{1}{2+\delta}} < a_1$ and $\mathrm{E}_P\left|\psi_{it}^{(0)}\right|^2 \leq a_1^2$ follows from Assumption DML2(i). Therefore, condition 8.3 is satisfied.

Consider condition 8.4. 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 8.4 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 $\left|\frac{1}{N_k}\sum_{i\in I_k} I^{(kl)}_{a,2,i}\right|$. Note that conditional on $\{\gamma_t\}_{t\in S_l}$, $I^{(kl)}_{a,2,i}$ is i.i.d over i. So, we have

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

By conditional Markov inequality, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

where the first inequality follows from the assumption that α_i is independent over i and an application of Hölder's inequality and Jensen's inequality. The second inequality follows from Hölder's inequality and the last equality follows from Assumption 3.5(ii) with some $p > 4$. Therefore, by Markov inequality, we have $\left| \frac{1}{T_l} \sum_{t \in S_l} \tilde{I}_{b,2,t}^{(kl)} \right| = O_P(N_k^{-1/2}) = O_P(N^{-1/2})$.

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

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

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

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

By conditional Markov inequality, we have

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

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

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

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

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

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

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

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

where

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

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

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

which completes the proof of Claim C.5.

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

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

where

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

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

$$\mathrm{E}_P[\psi_{it}^{(0)} \psi_{it}^{(0)}] \leq a_1^2.$$

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

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

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

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

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

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

where

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

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

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

$$I_{d,2}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j\in I_k, j\neq i} \left\{ \psi_{it}^{(0)} \psi_{j,t+m}^{(0)} - \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right] \right\},$$

$$I_{d,3}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} \left(k\left(\frac{m}{M}\right) - 1\right) \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j\in I_k, j\neq i} \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right],$$

$$I_{d,4}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{m=M}^{\infty} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j\in I_k, j\neq i} \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right],$$

$$I_{d,5}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{m=1}^{M-1} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \sum_{i\in I_k, j\in I_k, j\neq i} \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right] - c \sum_{m=1}^{\infty} \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right],$$

$$I_{d,6}^{(kl)} := c \sum_{m=1}^{\infty} \mathrm{E}_P\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)}\right] - c \sum_{m=1}^{\infty} \mathrm{E}_P\left[g_t g_{t+m}\right]$$

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), $\left(\mathrm{E}_P|\psi_{it}^{(0)}|^p\right)^{1/p} \leq a_1$ 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

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

By iterated expectation and Jensen's inequality, we have

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

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

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

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

$$\begin{aligned}\left|I_{d,3}^{(kl)}\right| \leq & \frac{K/L}{N_k T_l^2}\sum_{m=1}^{M-1}\left|k\left(\frac{m}{M}\right)-1\right|\sum_{t=\lfloor S_l\rfloor}^{\lceil S_l\rceil-m}\sum_{i\in I_k, j\in I_k, j\neq i}\left|\mathrm{E}_P\left[\psi_{it}^{(0)}\psi_{j,t+m}^{(0)}\right]\right| \\ \leq & ca_1^2\sum_{m=1}^{M-1}\left|k\left(\frac{m}{M}\right)-1\right|\alpha_\gamma(m)^{1-2/p}.\end{aligned}$$

Note that for each m, $\left|k\left(\frac{m}{M}\right)-1\right| \to 0$ as $M \to \infty$. Since $\left|k\left(\frac{m}{M}\right)-1\right|\alpha_\gamma(m)^{1-2/p} \leq 1$, 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, $\tilde{I}^{(kl)}_{d,2,tm}$ does not vary over t so that $\mathrm{E}_P \left| \sum_{m=1}^{M-1} \frac{k\left(\frac{m}{M}\right)}{T_l} \sum_{t=\lfloor S_l \rfloor}^{\lceil S_l \rceil - m} \tilde{I}^{(kl)}_{d,2,tm} \right|^2 \leq \mathrm{E}_P \left| \sum_{m=1}^{M-1} k\left(\frac{m}{M}\right) \tilde{I}^{(kl)}_{d,2,tm} \right|^2$. And by Minkowski's inequality, we have

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

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

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

where the first inequality follows from the assumption that α_i is independent over i and an application of Hölder's inequality and Jensen's inequality. The second inequality follows from Hölder's inequality and the last equality follows from Assumption 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 $I_{d,1}^{kl}$. Denote

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

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

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

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

To summarize

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

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 rescealing $\widehat{\theta}$ 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 $\|A_0 - \widehat{A}_{NT}\| = o_P(1)$, $\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0) - \psi_{NT}\| = o_P(1)$, and $\sqrt{N}\Omega^{-1/2}\psi_{NT} \xrightarrow{d} N(0,1)$ 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 in the theorem follows.

First, consider the last statement. By definition, we have

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

It is clear that the first term above is mean 0 due to the exogeneity of Z_{it} and $\mathbf{X}$ with respect to U_{it}. Due to the independence of the Mundlak device error from both Z_{it} and $\mathbf{X}$, we have

$$\begin{aligned}
\mathrm{E}[(Z_{it} - f_{it}\zeta_0)(L_{2,it} - \mathrm{E}[L_{2,it}])\eta_{Y2}] &= \mathrm{E}[(Z_{it} - f_{it}\zeta_0)E[(L_{2,it} - \mathrm{E}[L_{2,it}])\eta_{Y2}|Z_{it}, \mathbf{X}]] = 0, \\
\mathrm{E}[(Z_{it} - f_{it}\zeta_0)(L_{2,it} - \mathrm{E}[L_{2,it}])\eta_{D2}\theta_0] &= \mathrm{E}[(Z_{it} - f_{it}\zeta_0)E[(L_{2,it} - \mathrm{E}[L_{2,it}])\eta_{D2}|Z_{it}, \mathbf{X}]\theta_0] = 0
\end{aligned}$$

The rest of the statement follows from the same arguments as in the Proof of Lemma A.3.

Consider $\|A_0 - \widehat{A}_{NT}\|$. It is assumed that $\|\tilde{\xi} - \zeta_0\| = o_P\left(N^{-1/2}\right)$, $\|\tilde{\beta} - \beta_0\| = o_P\left(N^{-1/2}\right)$ and $\|\tilde{\pi} - \pi_0\| = o_P\left(N^{-1/2}\right)$. 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 $(N, T) \to \infty$,

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

Therefore, by Markov inequality, we have

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

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

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

Consider $\sqrt{N}\|\widehat{\psi}_{NT}(\theta_0) - \psi_{NT}\|$. 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 straighforward algebra, we have

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

where

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

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

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

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

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

□

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 $\widehat{\Omega}_{\text{CHS}} - \Omega = o_P(1)$. We decompose $\widehat{\Omega}_{\text{CHS}}$ as follows:

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

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 Lemma A.4, we only use the cross-fitting technique to show $I_{a,1}$ is of small order. The arguments for showing $I_{a,2}$ and $I_{a,3}$ to be of small order is basically same as those in the proof of Lemma A.4. So it is omitted here.

Consider $I_{a,1}$. Following the algebra in the proof of Lemma A.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

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

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

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

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 $\|\tilde{\pi}-\pi_0\| = o_p(N^{-1/2})$, $\|\tilde{\zeta}-\zeta_0\| = o_p(N^{-1/2})$, $\|\tilde{\beta}-\beta_0\| = o_p(N^{-1/2})$. 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 $(\widehat{\theta}-\theta_0)^2 = O_P(N^{-1})$. Therefore,

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

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

$$\begin{aligned} &\left|\psi(W_{it};\theta_0,\tilde{\eta}) - \psi(W_{it};\theta_0,\eta_0)\right|^2 \leq \sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0+\|\pi-\pi_0\|\|\leq\delta_{NT}} \left|\psi(W_{it};\theta_0,\eta) - \psi(W_{it};\theta_0,\eta_0)\right|^2 \\ = &\sup_{\|\zeta-\zeta_0\|+\|\beta-\beta_0\|+\|\pi-\pi_0\|\leq\delta_{NT}} \Big|(\zeta_0-\tilde{\zeta})' f_{it}' f_{it}(\beta_0-\tilde{\beta}) - \theta_0(\zeta_0-\tilde{\zeta})' f_{it}' f_{it}(\pi_0-\tilde{\pi}) + (\zeta_0-\tilde{\zeta})' f_{it}' V_{it}^Y \\ &\qquad -\theta_0(\zeta_0-\tilde{\zeta})' f_{it}' V_{it}^D + V_{it}^Z f_{it}(\beta_0-\tilde{\beta}) + V_{it}^Z f_{it}(\pi_0-\tilde{\pi})\Big|^2 \to 0. \end{aligned}$$

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

$$
\begin{aligned}
&\left(\mathrm{E}\left[\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)}\left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|^2\right]\right)^{1/2}\\
=&\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)}\left|\psi(W_{it};\theta_0,\eta)-\psi(W_{it};\theta_0,\eta_0)\right|\right\|_{P,2}=2\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0),\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)}\left|\psi(W_{it};\theta_0,\eta)\right|\right\|_{P,2}\\
\le&2\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0)}\left|Z_{it}-f_{it}\zeta\right|\right\|_{4,P}\left\|\sup_{\beta\in\mathcal{N}_m(\beta_0),\pi\in\mathcal{N}_m(\pi_0)}\left|Y_{it}-f_{it}\beta-\theta_0(D_{it}-f_{it}\pi)\right|\right\|_{4,P}\\
\le&2\left(\left\|Z_{it}\right\|_{4,P}+\left\|\sup_{\zeta\in\mathcal{N}(\zeta_0)}f_{it}\zeta\right\|_{4,P}\right)\left(\left\|Y_{it}\right\|_{4,P}+\left\|\sup_{\beta\in\mathcal{N}_m(\beta_0)}f_{it}\beta\right\|_{4,P}+\theta_0\left\|D_{it}\right\|_{4,P}+\theta_0\left\|\sup_{\pi\in\mathcal{N}_m(\pi_0)}f_{it}\pi\right\|_{4,P}\right)<\infty.
\end{aligned}
$$

So, we can apply the dominated convergence theorem to obtain

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

Then, by Markov inequality, we have

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

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

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

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 $\widehat{\Omega}_d\xrightarrow{p}c\sum_{m=1}^{\infty}\mathrm{E}_P[g_tg_{t+m}]$. As is shown in the proof of Theorem 4.2 (Lemmas A.5-A.7), the only step in showing these claims that involves cross-fitting technique is to show the same term R_{NT} to converges to 0 in probability. Otherwise, the arguments are basically the same. So we do not repeat those here. Combining these results, we obtain $\widehat{\Omega}=\mathrm{E}_P(a_i^2)+c\mathrm{E}_P(g_t2)+c\sum_{m=1}^{\infty}\mathrm{E}_P(g_tg_{t+m})=\Lambda_a\Lambda_a+c\Lambda_g\Lambda_g$.

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 Lemma A.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. □