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

Kaicheng Chen[a,1]

[a]*School of Economics, Shanghai University of Finance and Economics*

Abstract

Panel data allows for the modeling of unobserved heterogeneity, significantly raising the number of nuisance parameters and making high dimensionality a practical issue. Meanwhile, temporal and cross-sectional dependence in panel data further complicates high-dimensional estimation and inference. This paper proposes a toolkit for high-dimensional panel models with large cross-sectional and time sample sizes. To reduce the dimensionality, I propose a weighted LASSO using two-way cluster-robust penalty weights. Although consistent, the convergence rate of LASSO is slow due to the cluster dependence, rendering inference challenging in general. Nevertheless, asymptotic normality can be established in a semiparametric moment-restriction model by leveraging a clustered-panel cross-fitting approach and, as a special case, in a partial linear model using the full sample. In a panel estimation of the government spending multiplier, I demonstrate how high dimensionality could be hidden and how the proposed toolkit enables flexible modeling and robust inference.

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

1. Introduction

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

- The dimension of observable and potentially relevant variables can be large relative to the sample. For example, in the 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 [2]. 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 made (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 characteristics. Even if the unknown function is assumed to be approximately sparse so that it can be well-approximated by a linear combination of, for example, polynomially transformed covariates, the number of unknown parameters increases at a polynomial rate. [3]

- The modeling of heterogeneity can also raise the number of nuisance parameters. In demand analysis, income effects are specific to different products if the homothetic preference assumption fails. For difference-in-difference analysis, allowing unit-specific trends and heterogeneous trends across the covariates can relax/test the parallel trend assumption. For models with unobserved heterogeneity that appears in a nonlinear way, either treating them as parameters to be estimated (fixed effects) or modeling them in a flexible way (correlated random effects) contributes to high dimensionality. [4].

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

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

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

Without considering the features of panel data and the unobserved heterogeneity, it is a classic partial linear model that has been well-studied in previous semiparametric literature. With high dimensionality, sparse approximation and regularization approaches have been widely employed to reduce the dimensionality. Essentially, regularization, also known as the machine learning approach, trades off bias for smaller

---

[2] 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).

[3] For instance, the 2nd-order polynomial transformation of k-dimensional covariates 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.

[4] 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) and Chiang et al. (2023), for example.

variance to achieve desirable rates of convergence. However, due to the bias introduced by regularization and overfitting, inference can be challenging. Typically, some bias-correction procedures are involved to obtain estimators with better statistical properties and to conduct valid inference.

In the case of panel data, it is soon realized that at least three challenges would appear with the existing high-dimensional approaches. First of all, the statistical properties of many regularized estimators remain unknown with panel data where dependence exists across space/unit and time. Secondly, some bias-correction procedures for inference, such as sample-splitting/cross-fitting, are very particular about the sampling assumption, and the existing approaches are not valid under two-way dependence in the panel. Thirdly, the unobserved individual and time effects may appear in a flexible way, which further complicates estimation and inference.

For the first challenge, I proposed a variant of LASSO that uses regressor-specific penalty weights robust to two-way cluster dependence and weak temporal dependence across clusters. Such a LASSO approach is labeled as the two-way cluster-LASSO, corresponding to the heteroskedasticity-robust LASSO in Belloni et al. (2012) and the cluster-LASSO in Belloni et al. (2016). This approach theoretically derives the common penalty level λ up to a constant and a small-order sequence that does not vary across different data-generating processes. Therefore, data-driven tuning, such as information criterion and cross-validation, is not needed, which makes it more computationally efficient and avoids non-trivial theories that take data-driven tuning into account[5]. A common and important condition for obtaining the desirable statistical properties of LASSO selection/estimation is the so-called "regularization event", which states that the overall penalty level is sufficiently large to dominate the "noise" in the high-dimensional estimation (but not too large at the same time to avoid under-selection and slow rate of convergence). However, existing approaches for ensuring such an event with high probability are not applicable in this case due to the two-way cluster dependence. Instead, by considering the component structure characterization of the two-way dependence and rewriting the error terms using Hajek-type decomposition, I am able to leverage the moderate deviation theorems by Peña et al., 2009 and Gao et al., 2022 and the concentration inequality by Fuk and Nagaev (1971) for bounding the tail probability of the "noise" term. Combining with existing non-asymptotic bounds for the LASSO approach in Belloni et al. (2012), I derive the rate of convergence for the (post) two-way cluster-LASSO.

According to the rate of convergence results, the proposed (post) LASSO estimator is consistent for the slope coefficients of the sparse model. However, it is also revealed that the convergence rate is not as fast as the common rates for LASSO estimation because of the two-way cluster dependence[6]. While the LASSO estimates are still consistent when N, T diverge at the same rate, they converge slower than $(N \wedge T)^{-1/2} = (NT)^{-1/4}$, which is a common rate requirement for inferential theory. This is where the sec-

---

[5]For example, the statistical properties of LASSO methods given in Bickel et al. (2009); Bühlmann and Van De Geer (2011); Chernozhukov et al. (2021); Chetverikov et al. (2021); Babii et al. (2023); Gao et al. (2024) all rely on some proper choices of the penalty level and, among those, only Chernozhukov et al. (2021); Chetverikov et al. (2021) provide theoretical justification for their data-driven tuning.

[6]The intuition for the slow rate of convergence is illustrated in Appendix.

ond challenge arises: if a faster rate of convergence is not achievable due to the two-way cluster dependence, some bias-correction approaches are needed to relax the rate requirement for valid inference[7]. One common approach from the semiparametric literature is to add a correction term to the original identifying moment function(Newey, 1994). It results in an orthogonalized moment condition, which often features multiplicative error terms. Although the orthogonality property allows the nuisance estimations to be noisier, it generally does not ensure valid inference when there is growing dimensionality in the unknown parameters. An extra bias-correction approach, sample splitting or its generalization, cross-fitting, has been used for inference in high-dimensional regression models. The idea of sample splitting is to split the sample in a proper way and use the sub-samples separately to estimate nuisance parameters and main parameters of interest. If the sub-samples are independent of each other, then the first-step estimation of the nuisance parameters will be independent of the sample used for the second-step estimation. With this property, the error term that causes the bias can vanish with a less stringent rate requirement on the nuisance estimation. Intuitively, the dependence between the two steps is eliminated so that a potentially over-fitted nuisance estimate from the first step does not pollute the second step as much as it would otherwise do. However, sample-splitting as well as cross-fitting is very sensitive to the sampling assumption. Building upon recent development of cross-fitting approaches for dependent data (Chiang et al. (2022); Semenova et al. (2023a)), I propose a clustered-panel cross-fitting scheme, and I show that the constructed main and auxiliary samples are "approximately" and independent of each other. Effectively, this inferential procedure extends the double/debiased machine learning (DML, hereafter) approach by Chernozhukov et al. (2018) to panel data models, and it is labeled as the panel DML. Asymptotic normality for the panel DML estimator is established given high-level assumptions on the convergence rates regarding the first-step estimator. It is shown that the crude requirement on the rate of convergence can be relaxed to $(N \wedge T)^{-1/4}$, which admits the first-step estimation through the two-way cluster LASSO.

For the third challenge caused by the unobserved heterogeneity, existing approaches assume either that the unknown function g_0 in 1.1 is linear in (c_i, d_t) (Belloni et al., 2016; Kock and Tang, 2019), or that g_0 is linear in the interactive fixed effects (Vogt et al., 2022). To allow for flexible function forms while remaining tractable, I propose to model (c_i, d_t) as correlated random effects through a generalized Mundlak device while assuming the unknown function to be approximately sparse. In that way, a very rich form of heterogeneity is permitted. While not all of those nonlinear and heterogeneous effects are relevant, and the identities of the truly relevant effects are unknown to the researcher, suitable machine learning approaches, e.g., the two-way cluster-LASSO, can be used to select the relevant effects. However, there is one more subtle issue: common approaches, including the Mundlak device, that deal with the unobserved heterogeneous effects introduce cross-sectional and temporal sample averages, which in turn bring back dependence across cross-fitting sub-

---

[7]There are many bias-correction methods for high-dimensional models. See, for example, the low-dimensional projection adjustment in Zhang and Zhang (2014), the de-sparsification procedure in van de Geer et al. (2014), the decorrelating matrix adjustment in Javanmard and Montanari (2014), the double selection approach in Belloni et al. (2014), the decorrelated score construction in Ning and Liu (2017), the Neyman orthogonal moment construction in Chernozhukov et al. (2018).

samples. Furthermore, even if it remains valid under extra conditions, cross-fitting often causes a loss of efficiency due to the exclusion of observations. On the other hand, without cross-fitting, valid inference remains challenging for high-dimensional panel models in general. Nevertheless, in the case of the partial linear panel model, I further show that inferential theory can be established using the full sample.

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

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

Notation.

Here is a collection of frequently used notations in this paper. Some extra notations are defined along with the context. E and P are as generic expectation and probability operators. $\mathcal{P}_{NT}$ is an expanding collection of all data-generating processes P that satisfy certain conditions. P_{NT} is a sequence of probability laws such that $P_{NT} \in \mathcal{P}_{NT}$ for each (N, T). The dependence on (N, T) and P_{NT} will be suppressed whenever clear in the context. $\|.\|$ is the Euclidean (Frobenius) norm for a matrix. Let $\mathbf{x}$ be a generic $k \times 1$ real vector, then the l^q norm is denoted as $\|\mathbf{x}\|_q := \left(\sum_{j=1}^{k} x_j^q \right)^{1/q}$ for $1 \leq q < \infty$; $\|\mathbf{x}\|_\infty := \max_{1 \leq j \leq k} |x_j|$. The $L^q(P)$

norm is denoted as $\|f\|_{P,q} := \left(\int \|f(\omega)\|^q dP(\omega) \right)^{1/q}$ where f is a random element with probability law P. I denote the empirical average of f_{it} over $i = 1, ..., N$ and $t = 1, ..., T$ as $\mathbb{E}_{NT}[f_{it}] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} f_{it}$ and the empirical L^2 norm as $\|f_{it}\|_{NT,2} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \|f_{it}\|^2 \right)^{1/2}$. Correspondingly, I denote the empirical average of f_{it} over the sub-sample $i \in I_k$ and $t \in S_l$ as $\mathbb{E}_{kl}[f_{it}] = \frac{1}{N_k T_l} \sum_{i \in I_k, t \in S_l} f_{it}$ and the empirical L^2 norm over the subsample as $\|f_{it}\|_{kl,2} = \left(\frac{1}{N_k T_l} \sum_{i \in N_l} \sum_{t \in T_l} \|f_{it}\|^2 \right)^{1/2}$, where I_k, S_l are sub-sample index sets and N_k, T_l are sub-sample sizes that will be introduced next section.

2. Two-Way Cluster LASSO

In the existing literature, not much is known in terms of statistical properties for high-dimensional methods under cluster dependence in both cross-section and time. In this section, a variant of the $l1$-regularization methods, also known as the LASSO, will be proposed and examined. To focus on the LASSO approach under two-way dependence, I consider a simple conditional expectation model of a scalar outcome given a potentially high-dimensional vector of covariates. Let (Y_{it}, X_{it}) be a sample with $i = 1, ..., N$ and $t = 1, ..., T$. The conditional expectation model can be expressed as follows:

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

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

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

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

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

where $\mu = \mathrm{E}[W_{it}]$, 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. $\{\gamma_t\}_{t \geq 1}$ is strictly stationary and is beta-mixing at a geometric rate:

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

for some constants $\kappa > 0$ and $c_\kappa \geq 0$, where $\beta(X, Y) = \frac{1}{2}\|P_{X,Y} - P_X \times P_Y\|_{TV}$ and $\|.\|_{TV}$ denotes the total variation norm.

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[8], then $X_{ij} = f(\xi_i, \zeta_j, \iota_{ij})$ where $(\xi_i)_{i\geq 1}, (\zeta_j)_{j\geq 1}, (\iota_{ij})_{i\geq 1, j\geq 1}$ are mutually independent, uniformly distributed i.i.d. random variables[9]. 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 that the exchangeability is violated. Therefore, by allowing γ_t to be correlated, it introduces temporal dependence across all clusters, making the characterization more sensible in the panel data setting. The beta-mixing condition restricts the temporal dependence of the common time effects to decay at an exponential rate, which is common in literature and can be generated by common autoregressive models as in Baraud et al. (2001). The relaxation of the independence condition on $(\gamma_t)_{t\geq 1}$ can be viewed as a generalization of the component structure representation, as argued by Chiang et al. (2024). It is clear that under Assumption AHK, W_{it} and W_{is} are correlated for $t \neq s$ by sharing the same cross-sectional cluster. Similarly, by sharing the same temporal cluster, W_{it} and W_{jt} are dependent for any $i \neq j$. Furthermore, even if sharing neither the cross-sectional nor the temporal dimensions, observations can still be dependent due to correlated time effects γ_t. It is important to notice that the components in 2.2 simply characterize the dependence in panel data. Differing from factor models or models with unobserved heterogeneity, they do not affect the identification of the regression model.

Due to the potential high dimensionality in X, traditional nonparametric methods are not appropriate for estimating the unknown function f due to the curse of dimensionality. A common approach to reduce the dimensionality is by taking into account the sparsity information in models through regularization. Although the unknown function f is an infinite-dimensional parameter, which is not exactly sparse, it is common to view it from a sparse approximation perspective, following Belloni et al. (2012):

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

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

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

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

Assumption ASM views the high-dimensional linear regression as an approximation. It requires a subset of the parameters ζ_0 to be zero while controlling the size of the approximation error. Compared to the

---

[8]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 $\pi = \pi'$.

[9]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).

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

Under Assumption ASM, we can rewrite the model 2.1 as

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

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

$$\widehat{\zeta} = \arg\min_{\zeta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - f'_{it}\zeta)^2 + \frac{\lambda}{NT} \|\omega\zeta\|_1. \tag{2.5}$$

We note that the level of penalty term is determined by both λ and ω. A large penalty level can reduce avoid noisy estimation due to overfitting but meanwhile too large a penalty level can cause under selection which introduces missing variable bias. This is clearly a tradeoff between overfitting variance and regularization bias. Therefore, to obtain the desirable property of LASSO estimation, λ and ω need to be determined jointly both in theory and in practice. A common choice of ω is $\mathbb{E}_{NT}[\dot{f}'_{it}\dot{f}_{it}]$ where $\dot{f}_{it}$ be the demeaned f_{it} using the sample mean[10]. With this choice of ω, regressors are standardized, so the model selection is not affected by the scale of the regressors. In theory, given ω, λ needs to be chosen in a way that the following event happens with high probability:

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

where $C_\lambda > 1$ is some unknown constant. Condition 2.6 is referred to as the "regularization event" in the literature. If the error term V_{it} is conditionally Gaussian or sub-Gaussian, then the Gaussian tail inequality implies an asymptotic order for λ (see Bickel et al., 2009, Bühlmann and Van De Geer, 2011, and Theorem 29.3 of Hansen, 2022). However, this same choice may not guarantee this event when the error term exhibit non-Gaussian patterns such as asymmetry and heavy tails. To enable applications in broader settings, recent literature has investigated in new tools that ensures validity of LASSO with weaker conditions. Particularly,

---

[10]The demeaning is done because of the inclusion of the intercept term which is not penalized.

for panel data models, Babii et al. (2023) derives a Fuk-Nagaev type concentration inequality for panel data which ensures the validity of LASSO under non-Gaussianity and weak temporal dependence. Under functional dependence measure, Gao et al. (2024) establish new Fuk-Nagaev type concentration inequality and Chernozhukov et al. (2021) extend Gaussian approximation results to ensure LASSO validity with the presence of temporal and cross-sectional dependence. In practice, however, all approaches mentioned above require further estimation of data-driven tuning for the penalty level using cross-validation, information criterion, or bootstrap. These tuning methods are usually computationally costly and hard to justify in theory, and they are further complicated by the temporal and cross-sectional dependence in panel [11]. Furthermore, the restriction imposed for the functional dependence measure in Chernozhukov et al. (2021) and Gao et al. (2024) excludes the two-way cluster dependence considered here as we allow the dependence to not decay over cross-sectional or time.

Belloni et al. (2012) view ω as a self-normalizer and leverage moderate deviation theorems for the self-normalized sums to choose the penalty level that ensures Condition 2.6. While this approach does not require extra data tuning once ω is chosen properly according to the theory, the validity of the approach is restricted: existing moderate deviation theorems only work for independent or weakly dependent random variables. For panel data with cross-sectional independence, one can cluster within each cross-sectional unit or construct temporal blocks that are approximately independent, but there is no existing moderate deviation theorem with self-normalizer that works for two-way dependence. Instead, I utilize the component structure characterization of the dependence and consider a Hoeffding-type decomposition of the high-dimensional error term $f_{it}'V_{it}$: $a_i = \mathrm{E}[f_{it}'V_{it}|\alpha_i]$, $g_t = \mathrm{E}[f_{it}'V_{it}|\gamma_t]$, $e_{it} = f_{it}'V_{it} - a_i - g_t$ [12]. The goal is to design a self-normalizer as the penalty weight such that it accounts for the randomness coming from all three components and, importantly, it should be adaptive whether the components are non-degenerate or degenerate.

To see how this would work, let's first define a generic three-term regressor-specific penalty weight as follows:

$$\omega_j = \sqrt{\omega_{a,j}^2 + \omega_{g,j}^2 + \omega_{e,j}^2}. \tag{2.7}$$

where $\omega_{a,j}$, $\omega_{g,j}$, and $\omega_{e,j}$ are non-negative weights correspond to the three components. We want to derive

---

[11]Chernozhukov et al. (2021) do provide theoretical justification for choosing λ using the multiplier block bootstrap method.

[12]By Hoeffding decomposition, e_{it} can be further decomposed into a term featuring $E[f_{it}V_{it}|\alpha_i,\gamma_t]$ and a residual term. This is mainly relevant when the both a_i and g_t are degenerate and $E[f_{it}V_{it}|\alpha_i,\gamma_t]$ is not, referred to as the non-Gaussianity case discussed in Menzel (2021). Also see recently Davezies et al. (2025) and Hounyo and Lin (2025). This case still remains as a challenge and is a ongoing topic of research in the literature. We will focus on the Gaussian case under both non-degenerate and degenerate case.

a bound for the tail probability of the estimation error: for each $j = 1, ..., p,$

$$P\left(\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1}f_{it,j}V_{it}\right| > z\right) = P\left(\left|\frac{E_N[a_{i,j}] + E_T[g_{t,j}] + E_{NT}[e_{it,j}]}{\sqrt{\omega_{a,j}^2 + \omega_{g,j}^2 + \omega_{e,j}^2}}\right| > z\right)$$

$$\leq P\left(\left|\frac{E_N[a_{i,j}]}{\omega_{a,j}}\right| > \frac{z}{c_\omega}\right) + P\left(\left|\frac{E_T[g_{t,j}]}{\omega_{g,j}}\right| > \frac{z}{c_\omega}\right) + P\left(\left|\frac{E_{NT}[e_{it,j}]}{\omega_{e,j}}\right| > \frac{z}{c_\omega}\right) \tag{2.8}$$

where $a_{i,j}$ are the entries of a_i for $j = 1, ..., p$; $g_{t,j}$ and $e_{it,j}$ are defined similarly; $c_\omega := \frac{\omega_{a,j}+\omega_{g,j}+\omega_{e,j}}{\sqrt{\omega_{a,j}^2+\omega_{g,j}^2+\omega_{e,j}^2}} \in$
$\left[1, \sqrt{3}\right]$ is a scaling constant. To see why 2.8 holds, we note that for any positive numbers A, B, C, a, b, c
and $d = \frac{a+b+c}{\sqrt{a^2+b^2+c^2}}$, we have

$$\left\{\frac{A + B + C}{\sqrt{a^2 + b^2 + c^2}} > z\right\} \subseteq \left\{\frac{A}{a} > \frac{z}{d}\right\} \cup \left\{\frac{B}{b} > \frac{z}{d}\right\} \cup \left\{\frac{C}{c} > \frac{z}{d}\right\}.$$

To prove it, suppose $\neg\left\{\left\{\frac{A}{a} > \frac{z}{d}\right\} \cup \left\{\frac{B}{b} > \frac{z}{d}\right\} \cup \left\{\frac{C}{c} > \frac{z}{d}\right\}\right\}$, which implies $\left\{\frac{A}{a} \leq \frac{z}{d}\right\} \cap \left\{\frac{B}{b} \leq \frac{z}{d}\right\} \cap$
$\left\{\frac{C}{c} \leq \frac{z}{d}\right\}$. Then, we have $\frac{A+B+C}{\sqrt{a+b+c}} = \frac{A}{a}\frac{a}{\sqrt{a+b+c}} + \frac{B}{b}\frac{b}{\sqrt{a+b+c}} + \frac{C}{c}\frac{c}{\sqrt{a+b+c}} \leq \frac{z}{d}d = z$, which is $\neg\left\{\frac{A+B+C}{\sqrt{a+b+c}} > z\right\}$,
so it is proved by contrapositive.

The inequality 2.8 shows that it suffices to consider the tail probability of each component when choosing the penalty level. This greatly simplifies the problem since each of the components possesses much more tractable statistical properties compared to the original error term. It is shown in Appendix that, $a_{i,j}$ is independent over i, $\omega_{g,j}$ is weakly dependent over t, $e_{it,j}$ is independent conditional on $\{\gamma_t\}$, along with other useful properties. Therefore, the moderate deviation theorem for self-normalized sums of independent variables for (Peña et al., 2009) provides a guidance for the choice of $\omega_{a,j}$, and a conditional version helps constructing $\omega_{e,j}$. For the sum of weakly dependent variables, the self-normalizer in Gao et al. (2022) can be taken as a candidate for $\omega_{g,j}$. With the observation above, the infeasible regressor-specific penalty weights are proposed as follows:

$$\omega_{a,j}^2 = \frac{1}{N}\sum_{i=1}^{N}a_{i,j}^2, \quad \omega_{g,j}^2 = \frac{N}{T^2}\sum_{b=1}^{B}\left(\sum_{t\in H_b}g_{t,j}\right)^2, \quad \omega_{e,j}^2 = \frac{1}{NT^2}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}e_{it,j}\right)^2 \tag{2.9}$$

where B is the number of clusters/blocks, h is the block length and H_b is the associated index set. We note that $\omega_{a,j}$ can also be seen as a sample variance estimator (without centering around the sample mean). Furthermore, $\omega_{e,j}$ can be seen as a cluster variance estimator without estimation error, and $\omega_{g,j}$ can be seen as a correlated-cluster variance estimator (e.g., Bester et al., 2008). It may seem natural to use their sample analogs as the feasible penalty weights, by estimating the components and then plugging in. However, it

turns out that a direct sample analog of the infeasible penalty weights is not a good idea in certain settings. To see that, let's first focus on the estimation of the component without considering the estimation error in V_{it}, since the latter one is not the main source of the problem. For what follows, we consider the component estimators given V_{it}: let $v_{it,j} := f_{it,j} V_{it}$ and define

$$\tilde{a}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} v_{it,j}, \quad \tilde{g}_{t,j} = \frac{1}{N} \sum_{i=1}^{N} v_{it,j}, \quad \tilde{e}_{it,j} = v_{it,j} - \tilde{a}_{i,j} - \tilde{g}_{t,j}.$$

We also define the variances $\sigma_{a,j}^2 = E[a_{i,j}^2]$, $\sigma_{g,j}^2 = E[g_{j,t}^2]$, $\sigma_{e,j}^2 = E[e_{it,j}^2]$ and the long-run variances $\Sigma_{g,j} = \sum_{l=-\infty}^{\infty} E[g_{t,j} g_{t+l,j}]$, $\Sigma_{e,j} = \sum_{l=-\infty}^{\infty} E[e_{it,j} e_{it+l,j}]$. It can be shown that, under regularity conditions,

$$\omega_{a,j}^2 = \frac{1}{N} \sum_{i=1}^{N} a_{i,j}^2 = \sigma_{a,j}^2 + O_P(1/N) \tag{2.10}$$

$$\tilde{\omega}_{a,j}^2 := \frac{1}{N} \sum_{i=1}^{N} \tilde{a}_{i,j}^2 = \sigma_{a,j}^2 + \Sigma_{e,j}/T + O_P(1/N). \tag{2.11}$$

When the component $a_{i,j}$ is non-degenerate, $\tilde{\omega}_{a,j}^2$ has a non-degenerate limit $\sigma_{a,j}^2$ as $N, T \to \infty$, which is the same as the limit of $\omega_{a,j}^2$. In that sense, it is a valid choice of penalty weight. Similarly, we can show

$$\omega_{g,j}^2 = \frac{N}{T} \frac{1}{T} \sum_{b=1}^{B} \left(\sum_{t \in H_b} g_{t,j} \right)^2 = \frac{N}{T} \left(\Sigma_{g,j} + O_P(1/T) \right) \tag{2.12}$$

$$\tilde{\omega}_{g,j}^2 = \frac{N}{T} \frac{1}{T} \sum_{b=1}^{B} \left(\sum_{t \in H_b} \tilde{g}_{t,j} \right)^2 = \frac{N}{T} \left(\Sigma_{g,j} + \Sigma_{e,j}/N + O_P(1/T) \right). \tag{2.13}$$

This penalty weight choice is valid when the component $g_{t,j}$ is non-degenerate. Actually, as long as one of the components among $a_{i,j}$ and $g_{t,j}$ is non-degenerate, the penalty weight choice above is valid asymptotically because the non-degenerate component dominates and determines the scaling. However, if both $a_{i,j}$ and $g_{t,j}$ are degenerate, $\Sigma_{a,j} = \Sigma_{g,j} = 0$ and so $\Sigma_{e,j}$ dominates. After rescaling, $T \tilde{\omega}_{a,j}^2 = \Sigma_{e,j} + o_P(1)$ and $N \tilde{\omega}_{g,j}^2 = \Sigma_{e,j} + o_P(1)$ as $N, T \to \infty$ with $N/T \to c$ for $0 < c < \infty$, while $N \omega_{a,j}^2 = o_P(1)$ and $T \omega_{g,j}^2 = o_P(1)$. Meanwhile, it can be shown that

$$T \omega_{e,j}^2 = \frac{1}{NT} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} e_{it,j} \right)^2 = \Sigma_{e,j} + o_P(1)$$

Therefore, we see that the limit of the $\tilde{\omega}_{a,j}^2 + \tilde{\omega}_{g,j}^2$ is already two times of the limit of $\omega_j^2 = \omega_{a,j}^2 + \omega_{g,j}^2 + \omega_{e,j}^2$, making the choice of $\tilde{\omega}_j$ too conservative under degeneracy. Furthermore, we note that the sample analog

11

of $\omega_{e,j}^2$ is mechanically downward biased:

$$\tilde{\omega}_{e,j}^2 := \frac{1}{NT^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \tilde{e}_{it,j} \right)^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \left(v_{it,j} - \frac{1}{T} \sum_{t=1}^{T} v_{it,j} - \frac{1}{N} \sum_{i=1}^{N} v_{it,j} \right) \right)^2$$

$$= \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it,j} \right)^2 = o_P(1/T) \quad \text{under non-degeneracy}$$

$$= o_P(1/NT) \quad \text{under degeneracy}$$

$\tilde{\omega}_{e,j}^2$ is actually mechanically zero if we recenter the components when constructing the penalty weights. For similar reasons, the finite sample performance of $\tilde{\omega}_{g,j}^2$ would not be satisfying when the size of each correlated cluster H_b is chosen too large. Clearly, the above candidates would not generate desirable penalty weights.

Due to the aforementioned issues, I propose the following set of feasible penalty weights:

$$\hat{\omega}_j := \max(\hat{\omega}_{a,j}^2 - \hat{\omega}_{e,j}^2, 0) + \max(\hat{\omega}_{g,j}^2 - \hat{\omega}_{e,j}^2, 0) + \hat{\omega}_{e,j}^2 \tag{2.14}$$

$$\hat{\omega}_{a,j}^2 := \frac{1}{N} \sum_{i=1}^{N} \hat{a}_{i,j}^2 = \frac{1}{NT^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{v}_{it,j} \right)^2 \tag{2.15}$$

$$\hat{\omega}_{g,j}^2 := \frac{N}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M} \right) \hat{g}_{t,j} \hat{g}_{s,j} = \frac{1}{NT^2} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M} \right) \left(\sum_{i=1}^{N} \hat{v}_{it,j} \right) \left(\sum_{i=1}^{N} \hat{v}_{is,j} \right) \tag{2.16}$$

$$\hat{\omega}_{e,j}^2 := \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M} \right) \hat{e}_{it,j} \hat{e}_{is,j}$$

$$= \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} k\left(\frac{|t-s|}{M} \right) (\hat{v}_{it,j} - \hat{a}_i - \hat{g}_t)(\hat{v}_{is,j} - \hat{a}_i - \hat{g}_s) \tag{2.17}$$

where $k\left(\frac{|t-s|}{M} \right)$ is a Bartlett kernel and M is a bandwidth with a requirement $M = o(T^{1/4})$. A similar adjustment has been used in Menzel (2021) and Hounyo and Lin (2025) where the counterparts of $\max(\hat{\omega}_{a,j}^2 - \hat{\omega}_{e,j}^2, 0)$ and $\max(\hat{\omega}_{g,j}^2 - \hat{\omega}_{e,j}^2, 0)$ are used to estimate the variances to determine presence of the components in order for bootstrapping the components. Here these are used as self-normalizers adaptive to both non-degeneracy and degeneracy scenarios. We also note that $\hat{\omega}_{a,j}^2$ has been used as the feasible penalty weights in the cluster-LASSO method of Belloni et al. (2016) but they don't need the extra subtraction adjustment because $\hat{\omega}_{a,j}^2$ itself is in some sense adaptive under one-way cluster dependence, as hinted by 2.11. Additionally, we observe that $\hat{\omega}_{g,j}^2$ is effectively the variance estimator of Driscoll and Kraay (1998), which has been shown to account for cross-sectional cluster dependence and weak serial dependence. $\hat{\omega}_{e,j}^2$ is similar to the "average of HACs" estimator studied in Vogelsang (2012) except that it is regarding the component e instead of the original error term v.

Since the error term v is unobserved, we need to first obtain the residual $\hat{v}$ from some initial estimation and then iterate the estimation. The detailed implementation is given as Algorithm 1 in Appendix. The validity of this proposal relies on whether the feasible penalty weights under the iterative estimation converges to the same limit of the infeasible penalty weights. We establish the validity formally in Appendix. For the main theorem below, we maintain a high-level assumption on the feasible penalty weights following Belloni et al., 2012, 2016: Let $\hat{\omega}$ be the feasible diagonal weights. Suppose there exists $0 < 1/c_1 < l \leq 1$ and $1 \leq u < \infty$ such that $l \to 1$ and

$$l\omega_j^{1/2} \leq \hat{\omega}_j^{1/2} \leq u\omega_j^{1/2}, \text{ uniformly over } j = 1, ..., p, \tag{2.18}$$

where $\{\omega_j\}$ and $\{\hat{\omega}_j\}$ are diagonal entries of ω and $\hat{\omega}$, respectively. Given this choice of penalty weights, it is shown later in theory that the following choice of common penalty level λ delivers the event 2 with probability converging to 1,

$$\lambda = 2C_\lambda \sqrt{NT} \Phi^{-1}\left(1 - \frac{\gamma}{2p}\right), \tag{2.19}$$

where $C_\lambda = c_1 c_\omega > 1$; γ is a small order sequence. Other than C_λ and γ, there are no unknown tuning parameters in this weighted LASSO estimation. In practice, C_λ is chosen as a constant close to 1 and $\gamma = \alpha/\log(N \vee T)$ with α taken as a significance level[13]. The order of γ affects the convergence rate of the LASSO estimator: the theory only requires $\gamma = o(1)$ for LASSO to be consistent, but a faster rate of decay of γ will result in a slower convergence rate of LASSO.

The key identification condition is given as follows. In the low-dimensional case, the identification in a linear regression is given by the non-singularity of $E[f'_{it} f_{it}]$, which implies its empirical counterpart $\mathbb{E}_{NT}[f'_{it} f_{it}]$ is non-singular with high probability. However, as the dimension of f_{it} grows larger than the sample size, $\mathbb{E}_{NT}[f'_{it} f_{it}]$ is singular almost surely. Fortunately, it turns out that under sparsity and $L1$ regularization, we only need certain sub-matrices to be well-behaved for identification. Define

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

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

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

The sparse eigenvalue assumption follows from Belloni et al. (2012). It implies a restricted eigenvalue condition, which represents a modulus of continuity between the prediction norm and the norm of δ within

---

[13]For example, $\alpha = 0.1$ is more liberal than $\alpha = 0.05$. The choices of C_λ and γ around the proposed practical choices do not matter much in finite sample.

a restricted set. More primitive sufficient conditions are discussed in Bickel et al. (2009) and Belloni et al. (2012).

Assumption REG (Regularity Conditions). *(i)* $\log(p/\gamma) = o\left(T^{1/6}/(\log T)^2\right)$. *(ii) For some $\mu > 1, \delta > 0$,* $\max_{j \leq p} \mathrm{E}[|f_{it,j}|^{8(\mu+\delta)}] < \infty$, $\mathrm{E}[|V_{it}|^{8\mu+\delta}] < \infty$. *(iii)* $\mathrm{E}\left[\left(\sum_{t=1}^{T} e_{it,j}\right)^2 |\{\gamma_t\}_{t=1}^T\right] > 0$ *almost surely.* *(iv) For each j, either (1)* $\mathrm{E}\left(a_{i,j}^2\right) + \mathrm{E}\left(g_{t,j}^2\right) > \epsilon$ *for some $\epsilon > 0$ or (2)* $f_{it,j}V_{it}$ *is i.i.d over i, t. (v)* $1 \leq \max_{j \leq p} \omega_j / \min_{j \leq p} \omega_j = O(1)$.

Assumption REG(i) imposes a restriction on the dimension of f_{it}, p, while allowing it to be greater than the sample size. The moment conditions in Assumption REG(ii) are common in the literature. REG(iii) is a non-degeneracy condition, which is the main case of interest.

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

Theorem 2.1. *Suppose Assumptions AHK, ASM, REG hold for model 2.1 as $N, T \to \infty$ jointly with $N/T \to c$. Then, by setting λ as 2.19 and ω_j as 2.7 and 2.9, we have (i) the event 2.6 happens with probability approaching one. Additionally, suppose that Assumption SE holds and $\widehat{\omega}$ satisfies condition 2.18. Let $\widehat{\zeta}$ be the two-way cluster-LASSO estimator or the post-LASSO estimator based on the two-way cluster-LASSO selection. Then, (ii)* $\|\widehat{\zeta}\|_0 = O_P(s)$, *and (iii)* $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(f_{it}'\widehat{\zeta} - f_{it}'\zeta_0\right)^2 = O_P\left(\frac{s\log(p/\gamma)}{l_{NT}}\right)$, $\left\|\widehat{\zeta} - \zeta_0\right\|_1 = O_P\left(s\sqrt{\frac{\log(p/\gamma)}{l_{NT}}}\right)$, *and* $\left\|\widehat{\zeta} - \zeta_0\right\|_2 = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{l_{NT}}}\right)$, *where $l_{NT} = N \wedge T$ for the case (1) and $l_{NT} = NT$ for the case (2) of Assumption REG(iv).*

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

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

3. Clustered-Panel Cross-Fitting and Inference

In this section, I will first propose a cross-fitting scheme suitable for clustered panel data. The goal of the cross-fitting scheme is to split the sample in a proper way so that the two resulting sub-samples are independent or, at least, "approximately" independent. Under the AHK characterization in Assumption AHK, W_{it} are cluster-dependent over both cross-section and time. Importantly, the cluster dependence does not vanish as the distance between observations (if there is any ordering) increases. If γ_t is weakly dependent, which is the focus of this paper, then the dependence between observations that don't share the same cluster in either dimension dies out as the temporal distance grows. In that case, intuitively, one can split the sample so that the sub-samples do not share the same cluster and are far apart in temporal distance. This is exactly how this scheme works:

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

(i) *Select some positive integers (K, L). Randomly partition the cross-sectional index set $\{1, 2, ..., N\}$ into K folds $\{I_1, I_2, ..., I_K\}$ and partition the temporal index set $\{1, 2, ..., T\}$ into L adjacent folds $\{S_1, S_2, ..., S_L\}$ so that $\bigcup_{k=1}^{K} I_k = \{1, ..., N\}$, $\bigcup_{l=1}^{L} S_l = \{1, ..., T\}$[14].*

(ii) *For each $k = 1, ..., K$ and $l = 1, .., L$, construct the main sample $W(k, l) = \{W_{it} : i \in I_k, t \in S_l\}$ and the auxiliary sample $W(-k, -l) = \{W_{it} : i \in \bigcup_{k' \neq k} I_{k'}, t \in \bigcup_{l' \neq l, l \pm 1} S_{l'}\}$.*

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 $L = 8$.

---

[14]For simplicity, I assume N and T are divisible by K and L, respectively. In practice, if N is not divisible by K, the size for each cross-sectional block can be chosen differently with some length equal to floor(N/K) and others equal to ceil(N/K). and the same applies to the temporal dimension.
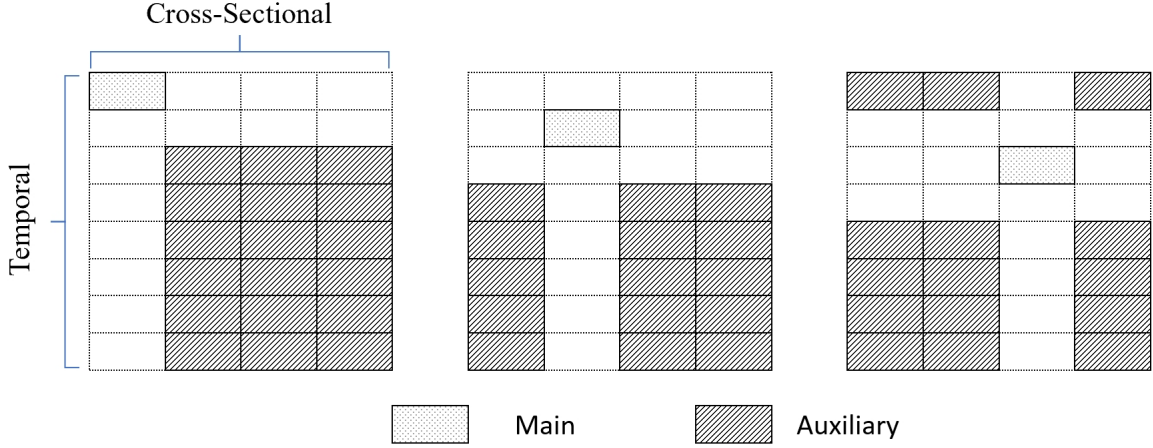
Figure 1: Clustered-Panel cross-fitting with $K = 4$ and $L = 8$. Three graphs from left to right correspond to the main and auxiliary sample constructions with $(k, l) = (1, 1)$, $(k, l) = (2, 2)$, $(k, l) = (3, 3)$. For a simple illustration, observations in the main sample are all adjacent in the cross-sectional dimension but it is not necessary in practice; the same applies to the auxiliary sample.

Since the sub-samples $W(k, l)$ and $W(-k, -l)$ do not share any cluster, they are free from cluster dependence, and what's left is the weak dependence over time. Unless imposing $m-$dependence, the sub-samples above will not be independent. However, under certain regularity conditions regarding the weak dependence, it can be shown through the coupling technique that as long as the temporal distance between the sub-samples diverges at a certain rate, there exist coupling sub-samples that are independent of each other while having the same marginal distributions as the constructed sub-samples with probability converging to 1. Such a result is provided in Lemma 3.1.

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 Assumption AHK holds 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)*

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

The proof of Lemma 3.1 is given in Appendix. Lemma 3.1 shows that the main and auxiliary samples from the proposed clustered-panel cross-fitting scheme are approximately independent as N, T diverge. Note that the hypothetical samples $\tilde{W}(k, l)$ and $\tilde{W}(-k, -l)$ only serve as technical tools and do not matter in practice. 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).

It turns out such properties of the cross-fitting scheme are very useful for inference in a high-dimensional panel model. For what follows, I will define an inferential procedure for high-dimensional panel with a two-step estimator and cluster-robust variance estimator using the proposed cross-fitting scheme in the context of a semi-parametric moment restriction model, as an extension of the prototypical DML approach in Chernozhukov et al. (2018). The clustered-panel cross-fitting scheme is used to remove the dependence between

the first and the second steps of estimation so as to relax the rate requirement for the first-step estimation of the high-dimensional nuisance parameters.

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 $\varphi(W_{it}; \theta, \eta)$ with the following properties:

$$\mathrm{E}[\psi(W_{it}; \theta_0, \eta_0)] = 0, \tag{3.1}$$

$$\partial_r \mathrm{E}\left[\psi\left(W_{it}; \theta_0, \eta_0 + r(\eta - \eta_0)\right)\right]|_{r=0} = 0. \tag{3.2}$$

Essentially, $\psi(W_{it}; \theta, \eta)$ is adjusted for the fact that η_0 needs to be estimated and, as a result, 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. In model 1.1, $\varphi(W_{it}; \theta, \eta) = D_{it} U_{it}$ and $\psi(W_{it}; \theta, \eta) = \left(D_{it} - \mathrm{E}[D_{it}|X_{it}, c_i, d_t]\right)\left(Y_{it} - D_{it}\theta - g(X_{it}, c_i, d_t)\right)$. In the treatment effect model with unconfoundedness conditional on covariates and unobserved heterogeneous effects, $\varphi(W_{it}; \theta, \eta) = \mathrm{E}[Y_{it}|D_{it} = 1, X_{it}, c_i, d_t] - \mathrm{E}[Y_{it}|D_{it} = 0, X_{it}, c_i, d_t] - \theta^{\mathrm{ATE}}$ and $\psi(W_{it}; \theta, \eta)$ is the moment function corresponding to the well-known augmented inverse probability weighting estimator, which is doubly robust.

The next definition summarizes the panel DML estimation and inference procedures for a semiparametric moment restriction model:

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

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

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

(iii) *For each k and l, use the sample $W(-k, -l)$ for the first-step estimation and obtain $\widehat{\eta}_{kl}$, then construct $\overline{\psi}_{kl}(\theta) = \mathbb{E}_{kl}[\psi(W_{it}; \theta, \widehat{\eta}_{kl})]$ using $W(k, l)$ for each (k, l), where $\mathbb{E}_{kl}[.]$ is defined in Notation.*

(iv) *Finally, obtain the DML estimator $\widehat{\theta}$ as the solution to*

$$\frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \overline{\psi}_{kl}(\theta) = 0. \tag{3.3}$$

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

A formal study its the statistical property and required assumptions are provided in the Online Supplementary Material. As is shown in the next section, the cross-fitting approach may not be compatible when we consider unobserved heterogeneities in the panel model, unless a strong functional assumption is imposed.

4. Partial Linear Model with Unobserved Heterogeneity

In this section, a partial linear model with non-additive unobserved heterogeneous effects is considered. The proposed toolkit is flexible enough to allow for models with instrumental variables used for identification, so I consider the following model: for $i = 1, ..., N$ and $t = 1, ..., T$,

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

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

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

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

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

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

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

To justify its use, we shall recall the idea of the conventional Mundlak device. Due to the correlation between (c_i, d_t) and the covariates, the endogeneity issue arises if we don't control for the unobserved heterogeneity. To explicitly model the correlation between the random effects and the covariates, Mundlak (1978) proposes an auxiliary regression between the random effects and the cross-sectional sample average and shows that if the random effects enter the model linearly then the resulting estimator GLS estimator is equivalent to the common within-estimator. Wooldridge (2021) further shows that the equivalence relations exist among the POLS estimators resulting from the Mundlak device, within-transformation, and the fixed-effects dummies. Therefore, if the within-transformation and including fixed-effects dummies are sensible

ways of dealing with unobserved heterogeneity, then allowing the Mundlak device to have a more flexible function form should also be reasonable and more robust. A similar assumption is considered in Wooldridge and Zhu (2020).

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

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

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

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

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

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

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

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

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

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

By defining a new error term $V_{it}^g := \left(L_{2,it} - E[L_{2,it}]\right)\left(\eta_{Y,2} - \eta_{D,2}\theta_0\right) + U_{it}$, a new approximation error

19

$r_{it} = r_{it}^Y + r_{it}^D \theta_0$, the vector of observables $f_{it} := (L_{1,it}, 1)$ with dimension denoted by p, and the nuisance vectors $\beta_0 := (\eta_{Y,1}, E[L_{2,it}]\eta_{Y,2})$, $\pi_0 := (\eta_{D,1}, E[L_{2,it}]\eta_{D,2})$, we can rewrite the model above as

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

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

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

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

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

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

where $V_{it}^Y = \left(L_{2,it} - E[L_{2,it}]\right)\eta_{Y,2} + U_{it}^Y$ and $V_{it}^D = \left(L_{2,it} - E[L_{2,it}]\right)\eta_{D,2} + U_{it}^D$. For $l = Z, Y, D$, let ω_l be the infeasible penalty weights for the two-way cluster LASSO estimation of $(\zeta_0, \beta_0, \pi_0)$, as defined in 2.7 with V_{it} replaced by V_{it}^l. Correspondingly, let $\widehat{V}^l$ be the residuals, and $\widehat{\omega}_l$ be the feasible penalty weights. The two-step debiased estimator $\widehat{\theta}$ for θ_0 using the full-sample is defined as the solution of $\mathbb{E}_{NT}[\psi_{it}(\theta, \widehat{\eta})] = 0$ where $\widehat{\eta}$ are the (post) two-way cluster LASSO estimators for η_0 obtained in the first step using the full-sample.

For statistical analysis, the following notations are used: $a_i = E[V_{it}^Z V_{it}^g|\alpha_i]$, $g_t = E[V_{it}^Z V_{it}^g|\gamma_t]$, $\Sigma_a = E[a_i a_i']$, $\Sigma_g = \sum_{l=-\infty}^{\infty} E[g_t g_{t+l}']$; $a_{i,j,l} = E[f_{it,j} V_{it}^l|\alpha_i]$, $g_{t,j,l} = E[f_{it,j} V_{it}^l|\gamma_t]$ for $l = Z, Y, D$; $A_0 = E[V_{it}^Z V_{it}^D]$, $\Omega_0 = \Sigma_a + c\Sigma_g$.

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

(i) A_0 is non-singular.

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

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

(iv) Either $\lambda_{min}[\Sigma_a] > 0$ or $\lambda_{min}[\Sigma_g] > 0$; $\min_{j \leq p} E[a_{i,j}^l]^2 > 0$, $\min_{j \leq p} E[g_{t,j}^l]^2 > 0$,

and $\min_{j \leq p} E\left[\left(\sum_{t=1}^T e_{it,j}^l\right)^2 |\{\gamma_t\}_{t=1}^T\right] > 0$ almost surely for $l = D, Y, Z$.

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

(vi) The feasible penalty weights $\widehat{\omega}_l$ satisfy condition 2.18 for $l = D, Y, Z$.

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

Theorem 4.1. *Suppose, for $P = P_{NT}$ for each (N, T), the following conditions hold for model 4.1 and $W_{it} = \left(Y_{it}, D_{it}, X_{it}, Z_{it}, U_{it}, c_i, d_t, \epsilon_i, \epsilon_t\right)$: (i) Assumptions AHK, SE, GMD, REG-P; (ii) sparse approximation in 4.4 and 4.5 with $s = o\left(\frac{\sqrt{N \wedge T}}{\log(p/\gamma)}\right)$, $\|r_{it}^l\|_{NT,2} = o_P\left(\sqrt{\frac{1}{N \wedge T}}\right)$ for $l = Y, D$. Then, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$, $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$, $V := A_0^{-1} \Omega_0 A_0^{-1}$.*

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

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

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

where $\widehat{A}_{NT} := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (Z_{it} - f_{it}' \widehat{\zeta})(D_{it} - f_{it}' \widehat{\pi})$ and

$$\widehat{\Omega}_{\text{A}} := \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{r=1}^{T} \psi_{it}(\widehat{\theta}, \widehat{\eta}) \psi_{ir}(\widehat{\theta}, \widehat{\eta})',$$

$$\widehat{\Omega}_{\text{DK}} := \frac{1}{NT^2} \sum_{t=1}^{T} \sum_{r=1}^{T} k\left(\frac{|t-r|}{M}\right) \sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{it}(\widehat{\theta}, \widehat{\eta}) \psi_{jr}(\widehat{\theta}, \widehat{\eta})',$$

$$\widehat{\Omega}_{\text{NW}} := \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{r=1}^{T} k\left(\frac{|t-r|}{M}\right) \psi_{it}(\widehat{\theta}, \widehat{\eta}) \psi_{ir}(\widehat{\theta}, \widehat{\eta})'.$$

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

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

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

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

Theorems 4.1 and 4.2 together show that valid inference is available without cross-fitting in this partial linear model.

5. Monte Carlo Simulation

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

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

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

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

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

$$\textbf{DGP(ii)} - \textbf{Partial Linear Model}: \quad Y_{it} = D_{it}\theta_0 + \left(X_{it}\beta_0 + c_i + d_t\right)^2 + U_{it},$$

$$D_{it} = \frac{\exp(X_{it}\pi_0)}{1 + \exp(X_{it}\pi_0)} + V_{it}, \quad c_i = \bar{D}_i + \bar{X}_i\xi_0 + \epsilon_i^c, \quad d_t = \bar{D}_t + \bar{X}_t\zeta_0 + \epsilon_t^d,$$

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

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

$$\textbf{DGP(i)} - \textbf{Additive Components}: \quad X_{it,j} = w_1\alpha_{i,j} + w_2\gamma_{t,j} + w_3\varepsilon_{it,j},$$

$$U_{it} = w_1\alpha_i^u + w_2\gamma_t^u + w_3\varepsilon_{it}^u, \quad V_{it} = w_1\alpha_i^v + w_2\gamma_t^v + w_3\varepsilon_{it}^v,$$

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

For the partial linear model, the Mundlak device will be used for estimation. It is well-known that the Mundlak device is mechanically equivalent to within-transformation in a linear panel model, in which the within-transformation would also remove the additive components in DGP(i) and eliminate the two-way dependence in the within-transformed random variables. When the true model is partially linear in

22

the covariates, the Mundlak device also projects out many underlying components and removes most of the dependence driven by the additive components. To illustrate it is not necessarily the case in general, a multiplicative component structure is considered as follows:

$$\textbf{DGP(ii)} - \textbf{Multiplicative Components}: \quad X_{it,j} = w_1\alpha_{i,j} + w_2\gamma_{t,j} + w_3\varepsilon_{it,j},$$

$$U_{it} = \frac{w_4}{c_p}\sum_{j=1}^{p}\left[\alpha_i^u\gamma_{t,j} + \alpha_{i,j}\gamma_t^u\right] + w_5\varepsilon_{it}^u, \quad V_{it} = \frac{w_4}{c_p}\sum_{j=1}^{p}\left[\alpha_i^v\gamma_{t,j} + \alpha_{i,j}\gamma_t^v\right] + w_5\varepsilon_{it}^v,$$

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

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

The simulation results are based on 1000 Monte Carlo replications. It is a relatively small number of replications but it is necessitated by the high computational cost of multiple high-dimensional estimation and inference procedures, particularly with cross-fitting. For variance estimation, bandwidth parameters M of the Bartlett kernel are required. I use the min-MSE rule from Andrews (1991) for both purposes. For a generic scalar score v_{it}, the formula is given as $\hat{M} = 1.8171\left(\frac{\hat{\rho}^2}{(1-\hat{\rho}^2)^2}\right)^{1/3}T^{1/3} + 1$ where $\hat{\rho}$ is the OLS estimator from the regression $\bar{v}_t = \rho\bar{v}_{t-1} + \eta_t$ where $\bar{v}_t = \frac{1}{N}\sum_{i=1}^{N}\hat{v}_{it}$ and $\hat{v}_{it} = \hat{U}_{it}\hat{V}_{it}$.

Table 5.1 presents a set of baseline results that are obtained for a decent number of regressors ($p = 200$)

Table 5.1: DGP(i) with $N = T = 25$, $s = 4$, $p = 200$, $\iota = 0.5$, $\rho = 0.5$, $c_\beta = c_\pi = 0.5$

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

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

among which 4 are associated with non-zero slope coefficients. The number of covariates is much larger than either cross-sectional or temporal dimensions. On the other hand, the number of non-zero coefficients can be regarded as a small order of the sample sizes, approximately satisfying the sparsity condition. In the first step, model selections are done using different LASSO approaches reported in the second column. The number of selected regressors for both Y and D are reported in the third and fourth columns. First, comparing the results obtained without using cross-fitting, it is shown that when the number of regressors is not extremely large relative to the sample size, the POLS estimator dominates the sparse methods through different LASSOs in terms of Monte Carlo bias, standard deviation, and coverage probability obtained using DKA standard error, even though the true model is sparse. Among the sparse methods, the proposed two-way cluster-LASSO exhibits the smallest bias and best coverage, while its standard deviation is slightly larger than the heteroskedastic-robust LASSO and root-LASSO. In terms of selection, the proposed method selects the number of regressors closest to the true number of relevant regressors while other sparse methods over-select to different extents.

When cross-fitting is employed, all methods have witnessed a significant improvement in terms of sample coverage. This is particularly true for LASSO-based methods that are not designed for dependent data. This is not too surprising because those non-robust sparse methods tend to over-select, and the cross-fitting is designed to remove the overfitting bias and to restore asymptotic normality. As a cost of cross-fitting, the Monte Carlo standard deviation increased, indicating the efficiency loss due to the exclusion of sub-samples in the first-step estimation. It is also worth emphasizing that the CHS- and DKA-type variance estimators designed for cross-fitting approaches play an important role in the desirable sample coverage. In some unreported simulations, it is shown that inference based on the cross-fitting variance estimators proposed in

Table 5.2: DGP(i) with $N = T = 25$, $s = 4$, $p = 600$, $\iota = 0.5$, $\rho = 0.5$, $c_\beta = c_\pi = 0.5$

Cross Fitting	First-Step Estimator	First-Step Ave.		Second-Step			Coverage (%)	
		Sel. Y	Sel. D	Bias	SD	RMSE	CHS	DKA
No	POLS	600	600	0.008	0.221	0.221	26.6	38.6
	H LASSO	39.5	39.8	0.073	0.049	0.087	51.2	78.9
	R LASSO	25.1	25.3	0.079	0.055	0.097	52.4	79.1
	C LASSO	14.0	15.2	0.058	0.096	0.112	68.8	78.4
	TW LASSO	6.9	7.5	0.033	0.098	0.103	81.6	88.1
Yes	H LASSO	24.7	24.8	0.059	0.128	0.141	95.7	97.6
	R LASSO	12.0	12.1	0.058	0.128	0.141	95.8	97.4
	C LASSO	10.7	11.4	0.047	0.127	0.136	95.7	97.5
	TW LASSO	7.0	7.5	0.067	0.125	0.142	90.7	95.1

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

Chernozhukov et al. (2018) and Chiang et al. (2022) suffers from severe under-coverage. This is not surprising but the implication is more subtle: while two-way dependence potentially affects both estimation and inference, its negative impact on the inference is more salient.

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

We have seen the case with exact sparsity in Tables 5.1 and 5.2. As claimed in the theory, the proposed estimation and inference procedures are also valid under approximate sparsity. Table 5.3 reports the simulation results under DGP(ii) where the true model is nonlinear in the the control variables and correlated random effects. The functional form of the nonlinearity is not given and is approximated by the second-order polynomial series. While only 10 observable covariates are considered, the Mundlak device and the polynomial transformation generate 560 regressors that will be included in the approximately sparse linear model. Due to the large number of regressors relative to the overall sample size, the approach based on the POLS estimation has the largest Monte Carlo standard deviation and root mean square error, and it suffers from severe under-coverage. Compared to the POLS and other sparse methods, the proposed two-

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

Cross Fitting	First-Step Estimator	First-Step Ave.		Second-Step			Coverage (%)	
		Sel. Y	Sel. D	Bias	SD	RMSE	CHS	DKA
No	POLS	560	560	0.012	0.173	0.173	54.4	67.4
	H LASSO	12.2	3.4	0.032	0.126	0.130	87.2	90.8
	R LASSO	11.0	3.3	0.030	0.127	0.130	86.2	91.0
	C LASSO	12.3	24.7	0.030	0.127	0.130	87.8	91.8
	TW LASSO	9.3	3.1	0.023	0.127	0.129	87.8	93.6
Yes	H LASSO	9.4	2.7	0.022	0.154	0.156	96.0	98.4
	R LASSO	6.9	2.0	0.016	0.155	0.156	96.2	98.2
	C LASSO	9.6	3.3	0.005	0.154	0.154	96.3	98.6
	TW LASSO	6.8	1.2	0.014	0.153	0.154	97.4	98.6

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

way cluster-LASSO method selects the most sparse model while having the smallest bias and root mean square error, and it also achieves the best coverage. When the clustered-panel cross-fitting is employed in the inference procedure, we find that the Monte Carlo coverage probability for confidence intervals based on CHS-type standard errors improves significantly, and the confidence intervals based on DKA-type standard errors switch from slight under-coverage to over-coverage. As the correlated random effects are used for estimation, they project out most of the components that drive the two-way cluster dependence under DGP(ii). In that case, the adjustment in the standard error formulas due to cross-fitting can be conservative.

6. Empirical Application

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

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

the high dimensionality from heterogeneity and flexible modeling could be hidden.

Due to the endogeneity in the variation of the regional military procurement, Nakamura and Steinsson (2014) achieves identification through an instrumental variable (IV) approach. As argued by the authors, the national military spending is largely determined by geopolitical events so it is likely exogenous to the unobserved factors of regional military spending and it affects the regional military spending disproportionally. In other words, the identifying assumption is that the buildups and drawdowns in national military spending are not due to unbalanced military development across regions. Based on this observation, a share-shift type IV is considered and the share is estimated by regressing the regional military spending on the national military spending allowing for region-specific constant slope coefficients.[15] To focus on the main idea, the shares are taken as given and the resulting instrument variable is treated as observable instead of generated regressors.

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

Baseline model :
$$Y_{it} = \theta_0 D_{it} + \pi_i W_t + c_i + d_t + U_{it},$$

Partial linear model :
$$Y_{it} = \theta_0 D_{it} + g(X_{it}, W_t, c_i, d_t) + U_{it}$$

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

In the baseline specification of Nakamura and Steinsson (2014), W_t are not included in the baseline model. In their alternative specifications, W_t is chosen as the real interest rate or the change in national oil price. These two variables are never included together in the original study. Note that allowing the unit-specific slope coefficients for controls generates many nuisance parameters: with 51 state groups[16], one control would increase 51 parameters and two controls would generate 102 parameters, without considering

---

[15]All quantities, unless specifically defined, are in terms of two-year growth rate of the real per capita values. Per capita is in terms of total population. Nakamura and Steinsson (2014) also presents results when per capita is calculated using the working age population as a robustness check.

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

interactions or higher order terms. With a sample size of less than 2000, the high dimensionality in nuisance parameters could result in a noisy estimate of θ_0. In this paper, to obtain a more precise estimate and make the excludability assumption of the IV more plausible, besides the controls from the original study, I also consider additional controls. As is shown in Table 3 of Nakamura and Steinsson (2014), the change in state population is likely not affected by the treatment (the regional military spending), so it is immune to the "bad control" problem; But it could affect the treatment and the outcome, so it is included in X_{it}. By considering more flexible function forms and additional exogenous control variables, the excludability condition of the instruments is more plausible. On the other hand, the high-dimensionality arose from the flexible function form and the unobserved heterogeneity necessitates the use of high-dimensional methods. Moreover, state-level yearly variables of those macroeconomic characteristics are often considered to be cluster-dependent in both cross-sectional and time groups due to correlated time shocks and state-unobserved factors. These concerns justify the use of robust estimation and inference methods proposed in this paper.

Table 6.1: Multiplier estimates from the original model

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

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

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

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

The main comparisons are done in Tables 6.2 and 6.3. In Table 6.2, no cross-fitting is performed in the

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

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

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

first stage. The number of parameters associated with regressors generated by the polynomials transformations are reported in column (4) and the number of selected parameters associated with Z are reported in column (6)[17]. Overall, with more controls and the polynomial transformation of the observables, the standard errors are generally larger than those in 6.1. With no transformations of the original regressors, the estimates obtained by four different methods are similar and are consistent with the baseline results. It is noticeable that the proposed approach TW LASSO achieves an estimate that is consistent with the baseline results and has the least variability. As the flexibility and number of nuisance parameters increase with the higher-order polynomial transformations, the number of selected regressors increases across all methods. While the standard errors of most approaches become larger and the estimates deviate from the baseline results, the proposed approach remains less noisy. This indicates that many higher-order polynomials included in the extended model for robustness in the function form may not matter much. While the existing approaches tend to over-select those terms under potential two-way dependence, the proposed method is robust against over-selection.

As in the Monte Carlo simulation, the results obtained with cross-fitting are also examined. Although

---

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

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

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

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

the theoretical results for inference procedure based on cross-fitting methods with the presence of the Mundlak device are not formally given in this paper, the conjecture is that it is still valid under the same set of conditions given in Section 4. Table 6.3 demonstrates the comparison between various sparse methods with the clustered-panel cross-fitting [18]. It reveals a similar pattern as in Table 6.2: The variability of different methods increases as the model approximated by higher-order polynomial series, except for the proposed approach which witnesses more accuracy as the approximation is made more flexible.

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

7. Conclusion and Discussion

The inferential theory for high-dimensional models is particularly relevant in panel data settings where the modeling of unobserved heterogeneity commonly leads to high-dimensional nuisance parameters. This

---

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

paper enriches the toolbox of researchers in dealing with high-dimensional panel models. Particularly, I propose a package of tools that deal with the estimation and inference in high-dimensional panel models that feature two-way cluster dependence and unobserved heterogeneity. I first develop a weighted LASSO approach that is robust to two-way cluster dependence in the panel data. As is shown in the statistical analysis of the two-way cluster LASSO, the convergence rates are slow due to the cluster dependence, making it challenging for inference purposes. However, by utilizing a cross-fitting method designed for a two-way clustered panel, the rate requirement for the first step can be substantially relaxed, making the proposed two-way cluster-LASSO a feasible first-step estimator for the panel-DML inference procedure in a high-dimensional semiparametric model. Individually, both the two-way cluster-LASSO and the clustered-panel cross-fitting can be of independent interest; Together, they extend the DML approach to panel data settings. In the partial linear panel model with high dimensionality as a special case, I further consider the unobserved heterogeneity and inferential theory using the full sample. The validity of the full-sample estimation and inference is established under a slightly stronger sparsity condition compared to the general case with cross-fitting.

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

While both theoretical and simulation results support the proposed approaches, some limitations remain in certain scenarios. The feasible penalty weight estimation is highly non-trivial due to two-way cluster dependence and high dimensionality. The statistical analysis of the two-way cluster LASSO relies on high-level assumptions about the feasible penalty weights. A devoted exploration of such issues requires a more comprehensive treatment and is an important direction for 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.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817.

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

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

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

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

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

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

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

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

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and dantzig selector. Ann. Statist. 37, 1705–1732.

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. CEPR Discussion Paper .

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

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

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

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., 2021. Lasso-driven inference in time and space. Ann. Statist. 49, 1702–1735.

Chetverikov, D., Liao, Z., Chernozhukov, V., 2021. On cross-validated lasso in high dimensions. Ann. Statist. 49, 1300–1317.

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

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

Chiang, H.D., Rodrigue, J., Sasaki, Y., 2023. 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., 2025. Analytic inference with two-way clustering. URL: https://arxiv.org/abs/2506.20749, arXiv:2506.20749.

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.

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

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

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

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

Gao, L., Shao, Q.M., Shi, J., 2022. Refined cramér-type moderate deviation theorems for general self-normalized sums with applications to dependent random variables and winsorized mean. Ann. Statist. 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. Ann. Statist. 42, 1166 – 1202.

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

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

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

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

Hounyo, U., Lin, J., 2025. Projection-based wild bootstrap under general two-way cluster dependence with serial dependence. doi:10.2139/ssrn.5361213.

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

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

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., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation. Econometrica 55, 703–708.

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

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

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 hetero-
geneous treatment effects in high-dimensional dynamic panels under weak dependence". Quantitative
Economics 14, 471–510.

Strassen, V., 1965. The existence of probability measures with given marginals. Ann. Math. Statist. 36,
423–439.

Vogelsang, T.J., 2012. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear
panel models with fixed-effects. Journal of Econometrics 166, 303–319.

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.

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.

8. Appendix

This appendix provides extra results and proofs for the two-way cluster-LASSO and cross-fitting. Proofs
for the panel DML and the partial linear model are provided in the Online Supplementary Material.

The Slow Rate of Convergence in an Oracle Case

To illustrate the slow rate of convergence under the component structure representation, we consider the
simplest multivariate mean model $Y_{it} = \theta_0 + f(\alpha_i, \gamma_t, \epsilon_{it})$ where Y_{it} is a high-dimensional vector with dimen-
sion $s = o(NT)$ and $\theta_0 = E[Y_{it}]$. To estimate the high-dimensional vector θ_0, we consider the sample mean
estimator $\widehat{\theta} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}$. Consider the following decomposition:

$$\widehat{\theta} - \theta_0 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (a_i + g_t + e_{it}) = \frac{1}{N} \sum_{i=1}^{N} a_i + \frac{1}{T} \sum_{t=1}^{T} g_t + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}, \qquad (8.1)$$

where $a_i := E[Y_{it} - \theta_0 | \alpha_i]$, $g_t := E[Y_{it} - \gamma_t]$, and $e_{it} := Y_{it} - \theta_0 - a_i - g_t$. For illustration purposes, suppose
those components are i.i.d sequences and independent of each other. Then it can be shown that, under some
regularity conditions, $\|\widehat{\theta} - \theta_0\|_2 = O_P\left(\sqrt{\frac{s}{N \wedge T}}\right)$, which is slower than the common rate requirement for
inferential theory.

Algorithm 1: Implementation of the Two-Way Cluster-LASSO

    i Let $\hat{\omega}_{j,0}^2$ be the produce of the sample variance of $f_{it,j}$ and the sample variance of Y, for each j. Set $\lambda_0 = 2C_\lambda\sqrt{NT}\hat{\sigma}_V\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right)$. Obtain the initial residuals $\tilde{V}_0$ using post-LASSO with the penalty level and weights set as above.

    ii Given residual $\tilde{V}_l$ for $l = 0, 1, 2, ...$, calculate $\hat{\omega}_{j,l}$ according to 2.14 for each j. Run post-LASSO with $\hat{\omega}_{j,l}$ and $\lambda = 2C_\lambda\sqrt{NT}\hat{\sigma}_V\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right)$ and update the residual as $\tilde{V}_{l+1}$.

   iv Repeat Step ii until it converges with the number of iteration greater or equal to 2. Obtains the (post) LASSO estimates from the last iteration.

Some Lemmas for Two-Way Clustering

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

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

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

  (i) $\{a_i\}_{i\geq1}$ is a sequence of i.i.d random vectors, $\{g_t\}_{t\geq1}$ are strictly stationary and β-mixing with the the mixing coefficient $\beta_g(m) \leq \beta_\gamma(m)$ for all $m \geq 1$; for each i, $\{e_{it}\}_{t\geq1}$ 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 $\mathrm{E}[e_{it}|\gamma_r] = 0$ and $\mathrm{E}[e_{it}|\alpha_j]$ for any i, t, j, r. It is less obvious to see $\mathrm{E}[e_{it}|\gamma_r] = 0$ for some $r \neq t$:

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

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

Using the properties above, one can derive the LLN and CLT for two-way clustered panel data. The following two lemmas are regarding the LLN and CLT for two-way clustered data with correlated time effects.

Lemma 8.1. *Suppose that W_{it} satisfy Assumption AHK and $\mathrm{E}\left[\|F(W_{it})\|^{4(r+\delta)}\right] < \infty$. Then,*

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

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

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

Lemma 8.2. *With the same setting as in Lemma 8.1, further assume that either (1) $\lambda_{min}\left[\Sigma_a\right] > 0$ or $\lambda_{min}\left[\Sigma_g\right] > 0$, or (2) W_{it} is i.i.d. over i and t. Then, as $N, T \to \infty$ and $N/T \to c$, $\sqrt{N}\left(\mathrm{E}_{NT}[F(W_{it})] - \mathrm{E}[F(W_{it})]\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_a + c\Sigma_g)$*

Lemmas 8.1 and 8.2 follow from Theorems 1 and 2 in Chiang et al. (2024).

Proof of Theorem 2.1

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

We first consider the non-degeneracy case: $\min_{j \le p} \mathrm{E}(a_{i,j}^2) > 0$ or $\min_{j \le p} \mathrm{E}(g_{t,j}^2) > 0$. Due to the inequality 2.8, we have, for each $j = 1, ..., p$:

$$\mathrm{P}\left(\frac{1}{NT}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_j^{-1}f_{it,j}V_{it}\right| > \frac{\lambda}{2c_1 NT}\right) \le \mathrm{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{a,j}^{-1}a_{i,j}\right| > \frac{\lambda\sqrt{N}}{2C_\lambda NT}\right)$$

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

$$:= p_{1,j}(\lambda) + p_{2,j}(\lambda) + p_{3,j}(\lambda)$$

where $C_\lambda = c_1 c_\omega$, $\omega_{a,j}^2 = \frac{1}{N}\sum_{i=1}^{N}a_{i,j}^2$, $\omega_{g,j}^2 = \frac{N}{T^2}\sum_{b=1}^{B}\left(\sum_{t \in H_b}g_{t,j}\right)^2$, $\omega_{e,j}^2 = \frac{1}{NT^2}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}e_{it,j}\right)^2$.
Then, the union-bound inequality implies

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

37

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

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

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

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

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

$$\sum_{j=1}^{p} p_{1,j}(\lambda) = P\left(\left|\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega_{a,j}^{-1} a_{i,j}\right| > \frac{\lambda \sqrt{N}}{2C_\lambda NT}\right)$$

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

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

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

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

Furthermore, by the strict stationarity and the non-degeneracy condition of case (3), we can verify that for some $v > 0$, $E\left[\sum_{t=r}^{r+m} g_{t,j}\right]^2 \geq v^2 m$ for all $t \geq 1, r \geq 0, m \geq 1$. By Assumption REG(ii) and Hölder's inequality, we have $E|f_{it,j} V_{it}|^{4(\mu+\delta)} < \infty$ for some $\mu > 1$, $\delta > 0$. Then, by Theorem 3.2 of Gao et al. (2022) with $\tau = 1$ and $\alpha = \frac{1}{1+2\tau}$, we have

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

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

$$\sum_{j=1}^{p} p_{2,j}(\lambda) = P\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\sqrt{T/N}\right)\omega_{g,j}\right)^{-1} g_{t,j}\right| > \frac{\lambda\sqrt{N}}{2C_\lambda NT}\right) \le \gamma\left[1 + O(1)\left(\frac{1}{l_{g,T}}\right)^2\right]$$

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

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

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

By Lemma 8.1, $|\Sigma_{e,j}| < \infty$. Furthermore, as is shown below, ω_j is bounded from below by some constant $a > 0$.

Define $\Xi_{e,j}^{\gamma} := \frac{(E[|\bar{e}_{i,j}|^2|\{\gamma_t\}_{t=1}^{T}])^{1/2}}{(E[|\bar{e}_{i,j}|^3|\{\gamma_t\}_{t=1}^{T}])^{1/3}}$. Since $\bar{e}_{i,j}$ is conditionally independent, we can apply the conditional version of self-normalization theorem for independent variables from Peña et al. (2009) as follows:

$$P\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{e,j}^{-1}\bar{e}_{i,j}\right| > x \mid \{\gamma_t\}_{t=1}^{T}\right) \le 2(1 - \Phi(x))\left[1 + O(1)\left(\frac{1+x}{N^{1/6}\Xi_{e,j}^{\gamma}}\right)^3\right]$$

for $x \in [0, N^{1/6}\Xi_{e,j}^{\gamma}]$. Under Assumption REG(iii), $1/\Xi_{e,j}^{\gamma} = O_P((E[|\bar{e}_{i,j}|^3 \mid \{\gamma_t\}_{t=1}^{T}])^{1/3})$. Under Assumption REG(ii), $E[(E[|\bar{e}_{i,j}|^3 \mid \{\gamma_t\}_{t=1}^{T}])^{1/3}] \le (E[|\bar{e}_{i,j}|^3])^{1/3} < \infty$, uniformly over $j \le p$ by Jensen's inequality, iterated expectation, and Hölder's inequality. Therefore, by integrating out $\{\gamma_t\}_{t=1}^{T}$, we obtain that

$$P\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\omega_{e,j}^{-1}\bar{e}_{i,j}\right| > x\right) \le 2(1 - \Phi(x))\left[1 + O(1)\left(\frac{1+x}{N^{1/6}}\right)^3\right]$$

for $x \in [0, AN^{1/6}]$ for some constant $A < \infty$. Let $l_{e,N}$ be some increasing sequence. If $AN^{1/6}/l_{e,N} - 1 > 0$

and $x \in [0, AN^{1/6}/l_{e,N} - 1]$, then

$$= P\left(\left|\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega_{e,j}^{-1} \bar{e}_{i,j}\right| > x\right) \le 2(1 - \Phi(x))\left[1 + O(1)\left(\frac{1}{l_{e,N}}\right)^3\right]$$

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

$$\sum_{j=1}^{p} p_{3,j}(\lambda) = P\left(\left|\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_{e,j}^{-1} e_{it,j}\right| > \frac{\lambda \sqrt{N}}{2C_\lambda NT}\right)$$

$$\le 2p(1 - \Phi(\Phi^{-1}(1 - \gamma/2p))) \le \gamma[1 + O(1)(1/l_{e,N})^3]$$

given that $\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right) \in [0, AN^{1/6}/l_{e,N} - 1]$ and $AN^{1/6}/l_{e,N} - 1 > 0$. Note that $\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right) \lesssim \sqrt{\log(p/\gamma)} = o\left(N^{1/12}/\log N\right)$ under Assumption REG(i) and $N/T \to c$ as $N, T \to \infty$. Therefore, it suffices to take $l_{e,N} = O(\log N)$, and it follows that $\sum_{j=1}^{p} p_{3,kl}(\lambda) \to 0$ as $\gamma \to 0$ and $(N, T) \to \infty$.

Put together, we have shown

$$P\left(\max_{j=1,\dots,p} \left|\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_j^{-1} f_{it,j} V_{it}\right| \le \frac{\lambda}{2c_1 NT}\right) \to 1. \tag{8.2}$$

Now we consider the degenerate case: W_{it} is i.i.d over i and t. In this case, we have $a_{i,j} = E[f_{it,j} V_{it} | \alpha_i] = E[f_{it,j} V_{it}] = 0$, $g_{t,j} = E[f_{it,j} V_{it} | \gamma_t] = E[f_{it,j} V_{it}] = 0$, and $e_{it,j} = f_{it,j} V_{it}$. Then, we observe that $p_{1,j}(\lambda) = p_{2,j}(\lambda)$. It follows that

$$P\left(\frac{1}{NT} \left|\sum_{i=1}^{N} \sum_{t=1}^{T} \omega_j^{-1} f_{it,j} V_{it}\right| > \frac{\lambda}{2c_1 NT}\right) = P\left(\frac{1}{\sqrt{NT}} \left|\sum_{i=1}^{N} \sum_{t=1}^{T} (\sqrt{T} \omega_{e,j})^{-1} e_{it,j}\right| > \frac{\lambda}{2c_1 \sqrt{NT}}\right)$$

Then, following the same steps as for $p_{3,j}(\lambda)$, we have that, with $\lambda = 2C_\lambda \sqrt{NT} \Phi^{-1}\left(1 - \frac{\gamma}{2p}\right)$, the probability above converges to zero and so Condition 8.2 is obtained for the degenerate case, too. Note that here we rescale the penalty weights and the common penalty level so that the penalty weights is not degenerate under the iid case. See Lemma 8.4 for the rate of convergence for the penalty weights under the non-degenerate and degenerate cases.

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

40

eigenvalues and Gram matrix as follows:

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

Define the weighted restricted eigenvalues as follows:

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

Assumption ASM, Conditions 2.18, and 8.2, Lemma 6 of Belloni et al. (2012) implies that, for the non-degenerate case,

$$\left\| f'_{it} \left(\widehat{\xi} - \zeta_0 \right) \right\|_{NT,2} \leq \left(u + \frac{1}{c_1} \right) \frac{\sqrt{s} \lambda}{NT K_{c_0}^\omega(M_f)} + 2 \|r\|_{NT,2} = O_P \left(\frac{1}{K_{c_0}^\omega(M_f)} \sqrt{\frac{s \log(p/\gamma)}{N \wedge T}} + \sqrt{\frac{s}{N \wedge T}} \right),$$

$$\left\| \omega \left(\widehat{\xi} - \zeta_0 \right) \right\|_1 \leq \frac{3 c_0 \sqrt{s}}{K_{2c_0}^\omega(M_f)} \left[\left(u + \frac{1}{c_1} \right) \frac{\sqrt{s} \lambda}{NT K_{c_0}^\omega(M_f)} + 2 \|r\|_{NT,2} \right] + 3 c_0 \frac{NT}{\lambda} \|r\|_{NT,2}^2,$$

$$= O_P \left(\frac{s}{K_{2c_0}^\omega(M_f) K_{c_0}^\omega(M_f)} \sqrt{\frac{\log(p/\gamma)}{N \wedge T}} + \sqrt{\frac{s}{N \wedge T}} + \frac{s / \sqrt{N \wedge T}}{\log(p/\gamma)} \right)$$

where $c_0 := \frac{u c_1 + 1}{l c_1 - 1} > 1$. For the degenerate case, the common penalty level λ is rescaled so that we replace $N \wedge T$ with NT in the above rates. This rate difference remains the same for the rest of the proof, and we will focus on the rate of the non-degenerate case.

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

$$K_C^\omega(M_f) \geq \frac{1}{b} K_{bC/a}(M_f). \tag{8.3}$$

Under Assumption REG(v), $b < \infty$ and $1 < b/a < \infty$. By 8.3, we have $1/K_{c_0}^\omega(M_f) \leq b/K_{\bar{C}}(M_f)$ where $\bar{C} := b c_0 / a$. By arguments given in Bickel et al. (2009), Assumption SE implies that $1/K_C(M_f) = O_P(1)$ for any $C > 0$. Therefore,

$$\left\| f'_{it} \left(\widehat{\xi} - \zeta_0 \right) \right\|_{NT,2} = O_P \left(\sqrt{\frac{s \log(p/\gamma)}{N \wedge T}} \right), \quad \left| \omega \left(\widehat{\xi} - \zeta_0 \right) \right|_1 = O_P \left(s \sqrt{\frac{\log(p/\gamma)}{N \wedge T}} \right).$$

By Hölder's inequality and that $\min_j \omega_j \geq a > 0$,

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

where the first inequality follows from Hölder's inequality.

41

The $l2$ rate of convergence will be derived after the sparsity bounds. We now switch the focus to the Post-LASSO. By the finite sample bounds of Lemma 8.3, we have

$$\|f(X_{it}) - f'_{it}\widehat{\zeta}_{PL}\|_{NT,2} = \left(\sqrt{\frac{s}{\phi_{min}(s)(M_f)}} + \sqrt{\frac{\widehat{m}}{\phi_{min}(\widehat{m})(M_f)}} \right) O_P \left(\frac{\lambda}{NT} \right)$$
$$+ O_P \left(\|f(X_{it}) - \left(\mathcal{P}_{\widehat{\Gamma}} f \right)_{it} \|_{NT,2} \right), \tag{8.4}$$

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

$$\|f'_{it}(\widehat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \leq \|f_{it}(X_{it}) - f'_{it}\widehat{\zeta}_{PL}\|_{NT,2} + \|r_{it}\|_{NT,2}, \tag{8.5}$$

$$\|\omega(\widehat{\zeta}_{PL} - \zeta_0)\|_1 \leq \frac{b\sqrt{\widehat{m}+s}}{\sqrt{\phi_{min}(\widehat{m}+s)(M_f)}} \times \|f'_{it}(\widehat{\zeta}_{PL} - \zeta_0)\|_{NT,2} \tag{8.6}$$

$$\|f(X_{it}) - \mathcal{P}_{\widehat{\Gamma}}f(X_{it})\|_{NT,2} \leq \left(u + \frac{1}{c_1} \right) \frac{\lambda\sqrt{s}}{NT K^\omega_{c_0}(M_f)} + 3\|r_{it}\|_{NT,2}. \tag{8.7}$$

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

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

where $a > 0$ has been shown previously.

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

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

Note that $\frac{6c_0 NT\|r_{it}\|_{NT,2}}{\lambda\sqrt{s}} = O_P(1/\log(p \wedge NT)) \xrightarrow{p} 0$. Recall that $1/K^\omega_{c_0}(M_f) \leq b/K_{\bar{C}}(M_f) < \infty$. Let $\mu := \min_m \left\{ \sqrt{\phi_{max}(m)(M_f)/\phi_{min}(m)(M_f)} : m > 18\bar{C}^2 s\phi_{max}(m)(M_f)/K^2_{\bar{C}}(M_f) \right\}$, and let $\bar{m}$ be the integer associated with μ. By the definition of $\mathcal{M}$, it implies that $\bar{m} \in \mathcal{M}$ with probability approaching one, which implies $\bar{m} > \widehat{m}$ due to 8.8. By Lemma 9 (the sub-linearity of sparse eigenvalues) from Belloni et al. (2012) and 8.8, we have

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

Combining the results above with 8.4 and 8.7 to gives

$$\|f(X_{it}) - f'_{it}\widehat{\zeta}_{PL})\|_{NT,2} = O_P\left(\sqrt{\frac{s\mu^2\log(p/\gamma)}{(N\wedge T)K_{\bar{C}}^2}} + \|r_{it}\|_{NT,2} + \frac{\lambda\sqrt{s}}{NTK_{c_0}^\omega(M_f)}\right).$$

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

$$\|f(X_{it}) - f'_{it}\widehat{\zeta}_{PL})\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right).$$

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

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

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

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

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

$$\|\widehat{\zeta}_{PL} - \zeta_0\|_2 \leq \frac{\left\|f'_{it}\left(\widehat{\zeta}_{PL} - \zeta_0\right)\right\|_{NT,2}}{\kappa_1} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right).$$

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

$\square$

Lemma 8.3. *Under Assumption ASM, if* $S_{\max} := \max_{1\leq j\leq p} |\mathbb{E}_{NT}[\omega^{-1}f_{it,j}V_{it}]| \leq \frac{\lambda}{2c_1 NT}, 0 < a = \min_j \omega \leq$

$\max_j \omega = b < \infty$, and $u \geq 1 \geq l \geq 1/c_1$, then

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

Proof of Lemma 8.3. We can decompose $f(X_{it}) - f'_{it}\widehat{\zeta}_{PL}$ as follows:

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

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

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

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

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

$\square$

Feasible Penalty Weights Validity

Let $\breve{\omega}_j$ be the feasible penalty weights without estimation error, defined the same way as 2.14 with $\hat{v}$ replaced by v. The following lemma shows that $\breve{\omega}_j$ has the same probability limit as the infeasible penalty weights.

Lemma 8.4. *Under Assumptions AHK and REG, and additionally, (i) $M = o(T^{1/4})$, (ii) $E[U_{it}|X_{i't'}] = 0$ for any i, t, i', t', (iii) $\sup_{i,t} |v_{it,j}|$ bounded by some absolute constant, then for each $j = 1, ..., p$,*

(i) If $E\left(a_{i,j}^2\right) + E\left(g_{t,j}^2\right) > \epsilon$ for some $\epsilon > 0$, then

$$\omega_{a,j}^2 - \left(\breve{\omega}_{a,j}^2 - \breve{\omega}_{e,j}^2\right) = o_P(1)$$

$$\omega_{g,j}^2 - \left(\breve{\omega}_{g,j}^2 - \breve{\omega}_{e,j}^2\right) = o_P(1)$$

$$\omega_{e,j}^2 - \breve{\omega}_{e,j}^2 = o_P(1)$$

(ii) If $f_{it,j}V_{it}$ is i.i.d, then

$$T\left[\omega_{a,j}^2 - \left(\breve{\omega}_{a,j}^2 - \breve{\omega}_{e,j}^2\right)\right] = o_P(1)$$

$$T\left[\omega_{g,j}^2 - \left(\breve{\omega}_{g,j}^2 - \breve{\omega}_{e,j}^2\right)\right] = o_P(1)$$

$$T\left[\omega_{e,j}^2 - \breve{\omega}_{e,j}^2\right] = o_P(1)$$

Proof of Lemma 8.4. We first consider the case (ii). In this case, since $a_{i,j} = g_{t,j} = 0$ and $e_{it,j} = v_{it,j}$, it is trivially true that $\Sigma_{e,j} = \sigma_{e,j}^2 = E[v_{it,j}^2]$ and

$$T\omega_{a,j}^2 = 0,$$

$$T\omega_{g,j}^2 = 0,$$

$$T\omega_{e,j}^2 = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T k\left(\frac{|t-s|}{M}\right)v_{it,j}v_{is,j}.$$

Due to the independence over i, j, we have

$$E[T\omega_{e,j}^2] = E[v_{it,j}^2].$$

And

$$Var(T\omega_{e,j}^2) \lesssim E[v_{it,j}^4]/N + E[v_{it,j}^4]/(NT)$$

Then, by Chebyshev inequality, we have $T\omega_{e,j}^2 = E[v_{it,j}^2] + o_P(1)$. Then, we consider the feasible penalty

weights without estimation errors:

$$E\left[T\breve{\omega}_{a,j}^2\right] = E\left[\frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}v_{it,j}\right)^2\right] = E[v_{it,j}^2]$$

$$E\left[T\breve{\omega}_{g,j}^2\right] = E\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}k\left(\frac{|t-s|}{M}\right)\left(\sum_{i=1}^{N}v_{it,j}\right)\left(\sum_{i=1}^{N}v_{is,j}\right)\right] = E[v_{it,j}^2]$$

$$Var(T\breve{\omega}_{a,j}^2) \lesssim E[v_{it,j}^4]/N + E[v_{it,j}^4]/(NT)$$

$$Var(T\breve{\omega}_{a,j}^2) \lesssim E[v_{it,j}^4]/N + E[v_{it,j}^4]/T + E[v_{it,j}^4]/(NT)$$

Then, by Chebyshev inequality, we have $T\breve{\omega}_{a,j}^2 = E[v_{it,j}^2]+o_P(1)$ and $T\breve{\omega}_{a,j}^2 = E[v_{it,j}^2]+o_P(1)$ Furthermore,

$$E\left[T\breve{\omega}_{e,j}^2\right] = E\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}k\left(\frac{|t-s|}{M}\right)\tilde{e}_{it,j}\tilde{e}_{is,j}\right] = E[T\omega_{e,j}^2] + O(1/N) + O(1/T) + o(1)$$

$$Var(T\breve{\omega}_{e,j}^2) \lesssim E[v_{it,j}^4]/N$$

By Chebyshev inequality, we have $T\breve{\omega}_{e,j}^2 = E[v_{it,j}^2]$. Then, the second statement of the lemma is proved.

Now we consider the first case. By Markov inequality and the fact that $a_{i,j}$ is i.i.d over i, we have

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}a_{i,j}^2 - \sigma_a^2\right| > \epsilon\right) \leq Var\left(\frac{1}{N}\sum_{i=1}^{N}a_{i,j}^2\right)/\epsilon^2 = E\left[a_{i,j}^4\right]/(N\epsilon) \leq E\left[v_{it,j}^4\right]/(N\epsilon)$$

So, we have $\omega_{a,j}^2 = \sigma_{a,j}^2 + O_P(1/\sqrt{N})$.

Next, consider $\breve{\omega}_{a,j}^2 = \frac{1}{N}\sum_{i=1}^{N}\tilde{a}_{i,j}^2$. We note that

$$\left(a_i + \frac{1}{T}\sum_{t=1}^{T}e_{it}\right) - \tilde{a}_i = \frac{1}{T}\sum_{t=1}^{T}g_t =: R_{g,NT},$$

which does not depend on i and $R_{g,NT} = O_P\left(1/\sqrt{T}\right)$. Then

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{a}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\left(a_i + \frac{1}{T}\sum_{t=1}^{T}e_{it}\right) - R_{g,NT}\right]^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T}e_{it}\right)^2 - \frac{2R_{g,NT}}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T}e_{it}\right) + R_{g,NT}^2.$$

Note that $a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}$ is independent over i conditional on $\{\gamma_t\}$, and that

$$E\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 = \sigma_a^2 + \Sigma_e(1 + o(1/T))/T = \sigma_a^2 + o(1),$$

$$\text{Var}\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 \le E\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^4 \le Ea_i^4 + Ee_{it}^4 < \infty.$$

Moreover,

$$\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 \,\Big|\, \{\gamma_t\}\right) = \frac{1}{N}\text{Var}\left(\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 \,\Big|\, \{\gamma_t\}\right)$$

$$\le \frac{1}{N}E\left(\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^4 \,\Big|\, \{\gamma_t\}\right)$$

$$\le \frac{1}{N}\left(\left(Ea_i^4\right)^{1/4} + \left(E\left[e_{it}^4 \mid \gamma_t\right]\right)^{1/4}\right)^4.$$

By Jensen's inequality,

$$E\left[\left(E\left[e_{it}^4 \mid \gamma_t\right]\right)^{1/4}\right] \le \left(E\left[E\left[e_{it}^4 \mid \gamma_t\right]\right]\right)^{1/4} = E\left[e_{it}^4\right]^{1/4} \le E[v_{it,j}^4]^{1/4}.$$

Then, by the conditional Markov inequality and Fubini's theorem, we have

$$\frac{1}{N}\sum_{i=1}^{N}\left(a_i + \frac{1}{T}\sum_{t=1}^{T} e_{it}\right)^2 = \left(\sigma_a^2 + \frac{1}{T}\Sigma_e\right) + O_P\left(\frac{1}{\sqrt{N}}\right).$$

And so

$$\breve{\omega}_{a,j}^2 = \left(\sigma_a^2 + \frac{1}{T}\Sigma_e\right) + O_P\left(\frac{1}{\sqrt{N}}\right) + O_P\left(\frac{1}{\sqrt{NT}}\right) + O_P\left(\frac{1}{T}\right) = \sigma_a^2 + o_P(1).$$

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

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

We can use the ratio test to examine the convergence of sum:

$$\lim_{q\to\infty} \frac{(q+1)^2\exp(-\frac{\kappa\zeta}{4+\zeta}(q+1))}{q^2\exp(-\frac{\kappa\zeta}{4+\zeta}q)} = \lim_{q\to\infty}\left(\frac{q+1}{q}\right)\exp(-\frac{\kappa\zeta}{4+\zeta}) = \exp(-\frac{\kappa\zeta}{4+\zeta})$$

Since $\kappa > 0$ and $\zeta > 0$, we have $\frac{\kappa\zeta}{4+\zeta} > 0$ and so $\exp(-\frac{\kappa\zeta}{4+\zeta}) < 1$. Thus we conclude the infinite sum does not diverge to infinity. The third condition is ensured by our assumptions directly. Thus, by Proposition 2 of Bester et al. (2008), we have

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

Therefore, we conclude that $\omega_j = \tilde{\omega}_j^{\text{CHS}} + o_P(1)$.

Next, we consider $\breve{\omega}_{t,j}^2$. We note that $g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j} - \tilde{g}_t = \frac{1}{N}\sum_{i=1}^{N} a_i =: R_{a,NT} = O_P(1/\sqrt{N})$, which does not depend on i or t. Then, we have

$$T/N\breve{\omega}_{g,j}^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right)\left(g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j} - R_{a,NT}\right)\left(g_s + \frac{1}{N}\sum_{i=1}^{N} e_{is,j} - R_{a,NT}\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right)\left(g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j}\right)\left(g_s + \frac{1}{N}\sum_{i=1}^{N} e_{is,j}\right) + o_P(1)$$

where the last $o_P(1)$ term follows from $M = o(T^{1/4}) = o(\sqrt{T})$ and $R_{a,NT} = O_P(1/\sqrt{N})$. Due to the mutual uncorrelatedness and that $e_{it,j}$ is uncorrelated over i, the long-run variance of $g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j}$ can be written as

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E(g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j})(g_s + \frac{1}{N}\sum_{i=1}^{N} e_{is,j}) = \Sigma_{g,j}(1+o(1)) + \Sigma_{e,j}(1+o(1))/N = \Sigma_{g,t} + o(1)$$

Note that conditional on $\{\alpha_i\}$, $g_t + \frac{1}{N}\sum_{i=1}^{N} e_{it,j}$ is beta-mixing at a geometric rate under Assumption AHK. We also note that $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right) g_t \frac{1}{N}\sum_{i=1}^{N} e_{is,j}$ is mean zero. Let $\bar{e}_{s,j} = \frac{1}{N}\sum_{i=1}^{N} e_{it,j}$, then we can write

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} k\left(\frac{|t-s|}{M}\right) g_t\bar{e}_{s,j} = \frac{1}{T}\sum_{l=-M}^{M} k(|l|/M)\sum_{t=1}^{T-|l|} g_{t,j}\bar{e}_{t+h,j}$$

Consider $z_{t,j}(h) := g_{t,j}\bar{e}_{t+h,j}$, which is also beta-mixing sequence with the geometric rate. For each h, we

can apply Theorem 14.13 of Hansen (2022)

$$Var\left(\frac{1}{T}\sum_{t=1}^{T-|l|}z_{t,j}(h)|\{\alpha_i\}\right) = (E[z_{t,j}^4|\{\alpha_i\}])^{1/2}O(1/T)$$

By Jensen's inequality and iterated expectation, we have $E[(E[z_{t,j}^4|\{\alpha_i\}])^{1/2}] \leq E[z_{t,j}^4]^{1/2} \leq \infty$. It follows that, $Var\left(\frac{1}{T}\sum_{t=1}^{T-|l|}z_{t,j}(h)\right) = O(1/T)$. Let $\bar{z}_{j,l}(h) = \frac{1}{T}\sum_{t=1}^{T-|l|}z_{t,j}(h)$ By Holder's inequality, we have

$$Var\left(\frac{1}{T}\sum_{l=-M}^{M}k(|l|/M)\bar{z}_{j,l}(h)\right) = E\left[\sum_{l=-M}^{M}k\left(\frac{|l|}{M}\right)\bar{z}_{j,l}(h)\right]^2$$

$$= \sum_{l=-M}^{M}k\left(\frac{|l|}{M}\right)\sum_{l'=-M}^{M}k\left(\frac{|l'|}{M}\right)E\left[\bar{z}_{j,l}(h)\bar{z}_{j,l'}(h)\right] \leq \sum_{l=-M}^{M}k\left(\frac{|l|}{M}\right)\sum_{l'=-M}^{M}k\left(\frac{|l'|}{M}\right)\sqrt{Var(\bar{z}_{j,l'})Var(\bar{z}_{j,l})}$$

$$= O(M^2/T) = o(1).$$

Therefore, we have $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}k\left(\frac{|t-s|}{M}\right)g_t\bar{e}_{s,j} = o_P(1)$. Similarly, we can show $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}k\left(\frac{|t-s|}{M}\right)\bar{e}_{t,j}\bar{e}_{s,j} = \Sigma_e/N$. With the extra requirement on $M = o(T^{1/4})$, we can verify the conditions of Theorem 2 from Newey and West (1987), we have $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}k\left(\frac{|t-s|}{M}\right)g_{t,j}g_{s,j} = \Sigma_g + o_P(1)$. It follows that $\breve{\omega}_{g,j}^2 - \omega_{g,j}^2 = o_P(1)$.

Lastly, consider $\omega_{e,j}^2 - \breve{\omega}_{e,j}^2$. Under the extra strict heterogeneity and uniform bound on the scores, we can apply the results of Theorem 5 from Chiang et al. (2024): $\max_i |a_{i,j} - \frac{1}{T}\sum_{t=1}^{T}v_{it,j}| = o_P(1)$ and $\max_i |g_{t,j} - \frac{1}{N}\sum_{i=1}^{N}v_{it,j}| = o_P(1)$. Therefore,

$$\breve{\omega}_{e,j}^2 = \frac{1}{NT^2}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}e_{it} + a_{i,j} - \frac{1}{T}\sum_{t=1}^{T}v_{it,j} + g_{t,j} - \frac{1}{N}\sum_{i=1}^{N}v_{it,j}\right)^2$$

$$= \frac{1}{NT^2}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}e_{it}\right)^2 + o_P(1) = \omega_{e,j}^2 + o_P(1)$$

This finishes the proof of the first case and therefore completes the proof.

$\square$

Cross-Fitting Validity

Let (X, Y) be random elements taking values in Euclidean space $S = (S_1 \times S_2)$ with probability laws P_X and P_Y, respectively. Let $\|v\|_{TV}$ denote the total variation norm of a signed measure v on a measurable space (S, Σ) where Σ is a σ-algebra on S:

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

The following lemma, quoted from Semenova et al. (2023a)(Lemma A.3), is a result that 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 8.5. *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 8.6. *Let $X_1, ..., X_q$ and Y be random elements taking values in Polish space $S = (S_1 \times ... \times S_m \times S_y)$.*

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

Proof of Lemma 8.6. By Lemma 8.5, we have

$$\beta((X_1, ..., X_m), Y) = \frac{1}{2}\left\|P_{(X_1,...,X_q),Y} - P_{(X_1,...,X_m)} \times P_Y\right\|_{TV}$$

$$= \mathrm{P}((X_1, ..., X_m, Y) \neq (\tilde{X}_1, ..., \tilde{X}_m, \tilde{Y})) \leq \sum_{i=1}^{m} \mathrm{P}((X_i, Y) \neq (\tilde{X}_i, \tilde{Y})) = \sum_{i=1}^{m} \beta(X_i, Y),$$

where the inequality follows from the union bound. $\square$

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

Proof of Lemma 3.1. By Lemma 8.5, for each (k, l) we have

$$\mathrm{P}\{(W(k, l), W(-k, -l)) \neq (\tilde{W}(k, l), \tilde{W}(-k, -l))\}$$

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

$$\leq \sum_{i \in I_k} \beta\left(\{W_{it}\}_{t \in S_l}, \bigcup_{k' \neq k, l' \neq l, l \pm 1} \{W_{it}\}_{i \in I_{k'}, t \in S_{l'}}\right) \leq \sum_{k' \neq k, l' \neq l, l \pm 1} \sum_{j \in I_{k'}} \sum_{i \in I_k} \beta\left(\{W_{it}\}_{t \in S_l}, \{W_{jt}\}_{t \in S_{l'}}\right)$$

where the last two inequalities follow from Lemma 8.6. Note that for $s, m \geq 1$, we have

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

where the last inequality follows from Assumption 2. Therefore,

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

which in turn gives

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

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

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

$\square$

Online Supplementary Material

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

Kaicheng Chen

This Online Supplementary Material is structured as follows. Section 8.1 provides a formal discussion and statistical analysis of the panel DML method. Section 8.2 collects proofs of the theorems of in Section 4.

8.1. Results for the Panel DML Method

In this section, a formal study of the panel DML method as an extension of the prototypical DML approach from Chernozhukov et al. (2018) is provided. Differing from the existing literature, the approach in this paper is made robust to two-way cluster dependence characterized by Assumption AHK. To restrict the focus,[19] I will assume a non-degeneracy condition. First, I define the decomposition components and their corresponding (long-run) variance-covariance matrices as follows:

$$a_i := \mathrm{E}\left[\psi\left(W_{it}; \theta_0, \eta_0\right) | \alpha_i\right], \quad \Sigma_a := \mathrm{E}[a_i a_i'],$$

$$g_t := \mathrm{E}\left[\psi\left(W_{it}; \theta_0, \eta_0\right) | \gamma_t\right], \quad \Sigma_g := \sum_{l=-\infty}^{\infty} \mathrm{E}[g_t g_{t+l}'],$$

$$e_{it} := \psi\left(W_{it}; \theta_0, \eta_0\right) - a_i - g_t, \quad \Sigma_e := \sum_{l=-\infty}^{\infty} \mathrm{E}[e_{it} e_{i,t+l}'].$$

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

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

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

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

---

[19]Although the panel DML procedure also works well for the i.i.d data, a formal result is not given in this paper, and the main theorem excludes such cases. The DML inference procedure for the i.i.d case has been studied in Chernozhukov et al. (2018) with the only difference in the cross-fitting scheme and the variance estimation. The proposed cross-fitting scheme is trivially valid for the i.i.d case, and the two-way cluster robust variance estimators considered in this paper, which are adapted from Chiang et al. (2024); Chen and Vogelsang (2024), can also be shown valid under i.i.d assumption following similar arguments in Chiang et al. (2024); Chen and Vogelsang (2024).

(i) $\psi(W; \theta, \eta)$ *is linear in* θ: $\psi(W; \theta, \eta) = \psi^a(W, \eta)\theta + \psi^b(W, \eta)$.

(ii) $\psi(W; \theta, \eta)$ *satisfy the Neyman orthogonality conditions 3.1 and 3.2 with respect to the probability measure P, or, more generally, 3.2 can be replaced by a* λ_{NT} *near-orthogonality condition*

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

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

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

Assumption DML1(i) restricts the focus of this paper to models with linear orthogonal moment conditions, which covers the model in Section 4. For nonlinear orthogonal moment conditions, Chernozhukov et al. (2018) has shown that the DML estimator has the same desirable properties under more complicated regularity conditions. Focusing on the linear cases allows us to pay more attention to issues specifically attributed to panel data. Assumption DML1(ii) slightly relaxes the orthogonality condition 3.2 by a near-orthogonality condition, which is useful for the approximate sparse model with approximation errors. Assumption DML1(iii) imposes a mild smoothness assumption on the orthogonal moment condition and Assumption DML1(iv) is a common condition for identification.

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

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

$$m_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E} \left\| \psi(W_{it}; \theta_0, \eta) \right\|^q \right)^{1/q} \le c_m,$$

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

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

$$r_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left\| \mathrm{E}[\psi^a(W_{it}; \eta) - \psi^a(W_{it}; \eta_0)] \right\| \le \delta_{NT},$$

$$r'_{NT} := \sup_{\eta \in \mathcal{T}_{NT}} \left(\mathrm{E} \left\| \psi(W_{it}; \theta_0, \eta) - \psi(W_{it}; \theta_0, \eta_0) \right\|^2 \right)^{1/2} \le \delta_{NT},$$

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

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

some models, e.g., the partial linear model, λ'_{NT} can be exactly 0, then it is possible to obtain the weakest possible rate requirement for the first-step estimator, i.e. $\varepsilon_{NT} = o(1)$.

Theorem 8.1 (Asymptotic Normality and Variance). *Suppose Assumptions AHK, ND, DML1, DML2 hold for any $P \in \mathcal{P}_{NT}$, then for some $\delta_{NT} \geq N^{-1/2}$, as $(N, T) \to \infty$ jointly and $N/T \to c$,*

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

where $V := A_0^{-1} \Omega A_0^{-1'}$ and $\Omega := \Sigma_a + c\Sigma_g$.

We observe that the convergence rate of the two-step estimator $\hat{\theta}$ resulting from the panel DML procedure is non-standard. It is $\sqrt{N}$-consistent instead of $\sqrt{NT}$-consistent. This is because the cluster dependence introduced by the unit and time components does not decay over time or space. Intuitively, with more persistence, the information carried by data is accumulated more slowly. It is a common feature in the literature of robust inference with cluster dependence[20] and it is also related to inferential theory under strong cross-sectional dependence (e.g., Gonçalves, 2011).

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

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

where $\hat{\hat{A}} := \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{1}{N_k T_l} \sum_{i \in I_k, s \in S_l} \psi^a(W_{it}; \hat{\eta}_{kl})$ and, with $k\left(\frac{m}{M}\right) := 1 - \frac{m}{M}$ for $m = 0, 1, ..., M-1$

---

[20]For example, see MacKinnon et al., 2021, Chiang et al., 2022, Chiang et al., 2024, Chen and Vogelsang, 2024 among many others.

and 0 otherwise (i.e., Bartlett kernel) and the bandwidth parameter M chosen from 1 to T_l,

$$\widehat{\Omega}_A := \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{1}{N_k T_l^2} \sum_{i\in I_k, t\in S_l, r\in S_l} \psi(W_{it}; \widehat{\theta}, \widehat{\eta}_{kl}) \psi(W_{ir}; \widehat{\theta}, \widehat{\eta}_{kl})',$$

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

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

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

Theorem 8.2 can be seen as a generalization of the consistency result for the CHS variance estimator in Chiang et al. (2024) by allowing for the estimated nuisance parameters in the moment functions. A remaining practical issue is that $\widehat{V}$ is not ensured to be positive semi-definite. It has been shown in Chen and Vogelsang (2024) that negative variance estimates happen with a non-trivial number of times under certain data-generating processes. Accordingly, an alternative two-term variance estimator was proposed in Chen and Vogelsang (2024). Following the same idea, I propose an alternative variance estimator by dropping the double-counting term $\widehat{\Omega}_{NW}$:

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

The estimator is referred to as the DKA variance estimator because it is a sum of Driscoll-Krray and Arellano variance estimators.[21] Similar approaches can be found in MacKinnon et al. (2021). It relies on the fact that the double-counting term is of small order asymptotically when the panel is two-way clustering. Similar to other two-term cluster-robust variance estimators, it has the computational advantage of guaranteeing positive semi-definiteness but at the cost of inconsistency in the case of no clustering or clustering at the intersection. For theoretical results and more detailed discussions on the trade-off between the ensured

---

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

positive-definiteness and the risk of being too conservative/losing power, readers are referred to MacKinnon et al. (2021) and Chen and Vogelsang (2024). The next theorem 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.

Theorem 8.3 (Alternative Consistent Variance Estimator). *Under the same conditions as Theorem 8.2, we have, as $N, T \to \infty$ and $N/T \to c$ where $0 < c < \infty$, $\widehat{V}_{\text{DKA}} = \widehat{V}_{\text{CHS}} + o_P(1)$.*

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

8.1.1. *Proof of Theorem 8.1*

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

Let $\widehat{\theta}$ be a solution from equation 3.3. To simplify the notation, we denote

$$\widehat{A}_{kl} = \mathbb{E}_{kl}[\psi^a(W_{it}, \widehat{\eta}_{kl})], \ \widehat{\overline{A}} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \widehat{A}_{kl}, \ A_0 = \text{E}[\psi^a(W_{it}; \eta_0)],$$

$$\widehat{B}_{kl} = \mathbb{E}_{kl}[\psi^b(W_{it}, \widehat{\eta}_{kl})], \ \widehat{\overline{B}} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \widehat{B}_{kl}, \ B_0 = \text{E}[\psi^b(W_{it}; \eta_0)],$$

$$\widehat{\overline{\psi}}(\theta) = \widehat{\overline{A}}\theta + \widehat{\overline{B}}, \ \bar{\psi}(\theta, \eta) = \mathbb{E}_{NT}\psi(W_{it}; \theta, \eta).$$

Claim 8.1. *On event* $\{\mathcal{E}_{\eta} \cap \mathcal{E}_{cp}\}$, $\|\widehat{\overline{A}} - A_0\| = O_P(N^{-1/2} + r_{NT})$.

By Claim 8.1 and Assumption DML1(iv) that all singular values of A_0 are bounded below by zero, it follows that all singular values of $\widehat{\overline{A}}$ are also bounded below from zero, on event $\mathcal{E}_{\eta}$. Then, by the linearity in Assumption DML1(i), we can write $\widehat{\theta} = -\widehat{\overline{A}}^{-1}\widehat{\overline{B}}$, $\theta_0 = -A_0^{-1}B_0$. Then, by basic algebra, we have

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \sqrt{N}(-\widehat{\overline{A}}^{-1}\widehat{\overline{B}} - \theta_0) = -\sqrt{N}\widehat{\overline{A}}^{-1}(\widehat{\overline{B}} + \widehat{\overline{A}}\theta_0) = -\sqrt{N}\widehat{\overline{A}}^{-1}\widehat{\overline{\psi}}(\theta_0)$$

$$= \sqrt{N}A_0^{-1}\bar{\psi}(\theta_0, \eta_0) + \sqrt{N}A_0^{-1}\left(\widehat{\overline{\psi}}(\theta_0) - \bar{\psi}(\theta_0, \eta_0)\right)$$

$$+ \sqrt{N}\left[\left(A_0 + \widehat{\overline{A}} - A_0\right)^{-1} - A_0^{-1}\right]\left(\bar{\psi}(\theta_0, \eta_0) + \widehat{\overline{\psi}}(\theta_0) - \bar{\psi}(\theta_0, \eta_0)\right)$$

Claim 8.2. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\|\widehat{\bar{\psi}}(\theta_0) - \bar{\psi}(\theta_0, \eta_0)\| = O_P(r'_{NT}/\sqrt{N} + \lambda_{NT} + \lambda'_{NT})$.

By Assumption DML2(i) and Jensen's inequality, we have $\|A_0\| \le m'_{NT} \le c_m$. Then, Claim 8.2 implies that

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

Since $E[\bar{\psi}(\theta_0, \eta_0)] = 0$, by Lemma 8.2, we have $\sqrt{N} \bar{\psi}(\theta_0, \eta_0) \overset{d}{\to} \mathcal{N}(0, \Omega)$ where $\Omega = \Sigma_a + c\Sigma_g$ and $\|\Omega\| < \infty$. By Claims B.1, B.2, and the asymptotic normality of $\sqrt{N} \bar{\psi}(\theta_0, \eta_0)$, we have

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

and $\sqrt{N} \left(\widehat{\theta} - \theta_0 \right) = A_0^{-1} \mathcal{N}(0, \Omega) + O_P \left(N^{-1/2} + r_{NT} + r'_{NT} + \sqrt{N} \lambda_{NT} + \sqrt{N} \lambda'_{NT} \right) \overset{d}{\to} A_0^{-1} \mathcal{N}(0, \Omega)$. The proofs for Claim B.1 and Claim 8.2 are given in Appendix C of the Online Supplementary Material. $\square$

8.1.2. *Proof of Theorem 8.2*

Proof. By the same arguments for Theorem 8.1, we have $P(\mathcal{E}_\eta \cap \mathcal{E}_{cp}) = 1 - P(\mathcal{E}_\eta^c \cup \mathcal{E}_{cp}^c) \ge 1 - o(1)$. By Claim B.1, we have $\|\widehat{\bar{A}} - A_0\| = O_P(N^{-1/2} + r_{NT})$ on event $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$. Therefore, due to $\|A_0^{-1}\| \le a_0^{-1}$ ensured by Assumption DML1(iv) and $\Omega < \infty$ as shown in Claim 8.2, it suffices to show $\|\widehat{\Omega}_{CHS} - \Omega\| = o_P(1)$. Furthermore, since K, L are fixed constants, it suffices to show for each (k, l) that $\|\widehat{\Omega}_{CHS,kl} - \Omega\| = o_P(1)$ where

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

Since a sequence of symmetric matrices Ω_n converges to a symmetric matrix Ω_0 if and only if $e' \Omega_n e \to$

6

$e'\Omega_0 e$ for all comfortable e, it suffices to assume without loss of generality that the dimension of ψ to be 1. To simplify the expression, we denote

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

Claim 8.3. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{a,kl} - \Sigma_a\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$.

Claim 8.4. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{b,kl} - c\mathrm{E}[g_t g_t]\right| = O_P\left(N^{-1/2} + r'_{NT}\right)$.

Claim 8.5. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{c,kl}\right| = O_P\left(T^{-1}\right)$.

Claim 8.6. *On event* $\{\mathcal{E}_\eta \cap \mathcal{E}_{cp}\}$, $\left|\widehat{\Omega}_{d,kl} - c\sum_{m=1}^\infty \mathrm{E}[g_t g_{t+m}]\right| = o_P(1)$.

Combining the Claims 8.3 - 8.6 completes the proof of Theorem 8.2. $\qquad\square$

8.1.3. *Proof of Theorem 8.3*

Proof. 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 $\widehat{\Omega}_{\mathrm{NW},kl} = \widehat{\Omega}_{c,kl} + \widehat{\Omega}_{e,kl} - \widehat{\Omega}_{d,kl}$ where $\widehat{\Omega}_{c,kl}$ and $\widehat{\Omega}_{d,kl}$ are defined in the proof of Theorem 8.2, 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 8.6 also follows. It implies that $\widehat{\Omega}_{e,kl} = \widehat{\Omega}_{d,kl} + o_P(1)$. By Claim 8.5, we have $\widehat{\Omega}_{c,kl} = O_P(T^{-1})$. Therefore, we conclude that $\widehat{\Omega}_{\mathrm{NW},kl} = o_P(1)$. $\qquad\square$

Proof of Claim 8.1.

Proof. Fix any (k, l), we have

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

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

7

notation, we denote $\ddot{\psi}_{it}^{a,kl} := \psi^a(W_{it}, \hat{\eta}_{kl}) - \mathrm{E}[\psi^a(W_{it}, \hat{\eta}_{kl})|W(-k,-l)]$. Consider $\|\Delta_{A,1}\|$:

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

$$\leq \left(\frac{1}{N_k T_l}\right)^2 \sum_{i\in I_k, t\in S_l, r\in S_l} \left|\mathrm{E}\left[\langle \ddot{\psi}_{it}^{a,kl}, \ddot{\psi}_{is}^{a,kl}\rangle |W(-k,-l)\right]\right| + \sum_{t\in S_l, i\in I_k, j\in I_k} \left|\mathrm{E}\left[\langle \ddot{\psi}_{it}^{a,kl}, \ddot{\psi}_{jt}^{a,kl}\rangle |W(-k,-l)\right]\right|$$

$$+ \sum_{t\in S_l, i\in I_k} \left|\mathrm{E}\left[\langle \ddot{\psi}_{it}^{a,kl}, \ddot{\psi}_{it}^{a,kl}\rangle |W(-k,-l)\right]\right| + 2\sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j\in I_k} \left|\mathrm{E}\left[\langle \ddot{\psi}_{it}^{a,kl}, \ddot{\psi}_{j,t+m}^a\rangle |W(-k,-l)\right]\right|$$

$$+ 2\sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i\in I_k} \left|\mathrm{E}\left[\langle \ddot{\psi}_{it}^{a,kl}, \ddot{\psi}_{i,t+m}^a\rangle |W(-k,-l)\right]\right| =: \left(\frac{1}{N_k T_l}\right)^2 (a(1) + a(2) + a(3) + 2a(4) + 2a(5)).$$

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

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

$$= \mathrm{E}\left[\|\ddot{\psi}_{it}^{a,kl}\|^2 |W(-k,-l)\right].$$

Therefore, we have

$$a(1) \leq N_k T_l^2 \mathrm{E}\left[\|\ddot{\psi}_{it}^{a,kl}\|^2 |W(-k,-l)\right], \quad a(2) \leq N_k^2 T_l \mathrm{E}\left[\|\ddot{\psi}_{it}^{a,kl}\|^2 |W(-k,-l)\right],$$

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

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

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

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

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

For each $i \in I_k, t \in S_l$, we can decompose $\ddot{\psi}_{d,i,t}^{a,kl} = a_i^{kl} + g_t^{kl} + e_{it}^{kl}$ where $a_i = \mathrm{E}[\ddot{\psi}_{d,i,t}^{a,kl}|\alpha_i]$, $g_i = \mathrm{E}[\ddot{\psi}_{d,i,t}^{a,kl}|\gamma_t]$, and $e_{it} = \ddot{\psi}_{d,i,t}^{a,kl} - a_i - g_t$. Conditional on $W(-k,-l)$, $(a_i^{kl}, g_t^{kl}, e_{it}^{kl})$ are mutually uncorrelated, $a_i \perp a_j$ for

8

$i \neq j$, and g_t^{kl} is also beta-mixing with $\beta_g(m) \leq \beta_\gamma(m)$. Therefore, we have

$$\mathrm{E}\left[\ddot{\psi}_{d,i,t}^{a,kl}\ddot{\psi}_{d,j,t+m}^{a,kl}|W(-k,-l)\right] = \mathrm{E}\left[g_t^{kl}g_{t+m}^{kl} + e_{it}^{kl}e_{j,t+m}^{kl}|W(-k,-l)\right]$$

$$= \mathrm{E}\left[g_t^{kl}g_{t+m}^{kl}|W(-k,-l)\right] + \mathrm{E}\left[\mathrm{E}\left[e_{it}^{kl}e_{j,t+m}^{kl}|\alpha_i,\alpha_j,W(-k,-l)\right]|W(-k,-l)\right]$$

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

$$\left|\mathrm{E}\left[\mathrm{E}\left[e_{it}^{kl}e_{j,t+m}^{kl}|\alpha_i,\alpha_j,W(-k,-l)\right]|W(-k,-l)\right]\right| \leq \mathrm{E}\left[\left|\mathrm{E}\left[e_{it}^{kl}e_{j,t+m}^{kl}|\alpha_i,\alpha_j,W(-k,-l)\right]\right||W(-k,-l)\right]$$

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

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

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

Then, we have

$$\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k, i\neq j}\sum_{d=1}^D\mathrm{E}\left[\ddot{\psi}_{d,i,t}^{a,kl}\ddot{\psi}_{d,j,t+m}^{a,kl}|W(-k,-l)\right]$$

$$\lesssim c_m^2\frac{1}{N_k^2 T_l}\sum_{m=1}^{T_l-1}\sum_{t=\min(S_l)}^{\max(S_l)-m}\sum_{i,j\in I_k, i\neq j}\sum_{d=1}^D\alpha_\gamma(m)^{1-2/q} \leq c_m^2 D\sum_{m=1}^\infty c_\kappa exp(-\kappa m)^{1-2/q} \leq \frac{c_m^2 Dc_\kappa}{exp(\kappa(1-2/q))-1} < \infty,$$

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

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

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

Proof of Claim 8.2

Proof. Since K and L are finite, it suffices to show for any k, l,

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

To simplify the notation, we denote

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

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

$$b(1) = \left\| \frac{\sqrt{N_k}}{N_k T_l} \sum_{i \in I_k, t \in S_l} \left[\ddot{\psi}_{it}^{kl} - \mathrm{E}[\ddot{\psi}_{it}^{kl} | W(-k, -l)] \right] \right\|$$

$$b(2) = \left\| \frac{1}{N_k T_l} \mathrm{E} \left[\psi(W_{it}; \theta_0, \hat{\eta}_{kl}) | W(-k, -l) \right] - \mathrm{E} \left[\psi(W_{it}; \theta_0, \eta_0) \right] \right\|.$$

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

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

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

$$\mathrm{E}[b(1)^2 | W(-k, -l)] \le \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \left| \mathrm{E} \left[\langle \tilde{\psi}_{it}^{kl}, \tilde{\psi}_{is}^{kl} \rangle | W(-k, -l) \right] \right|$$

$$+ \sum_{t \in S_l, i \in I_k, j \in I_k} \left| \mathrm{E} \left[\langle \tilde{\psi}_{it}^{kl}, \tilde{\psi}_{jt}^{kl} \rangle | W(-k, -l) \right] \right| + \sum_{t \in S_l, i \in I_k} \left| \mathrm{E} \left[\langle \tilde{\psi}_{it}^{kl}, \tilde{\psi}_{it}^{kl} \rangle | W(-k, -l) \right] \right|$$

$$+ 2 \sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i,j \in I_k} \left| \mathrm{E} \left[\langle \tilde{\psi}_{it}^{kl}, \tilde{\psi}_{j,t+m}^{kl} \rangle | W(-k, -l) \right] \right| + 2 \sum_{m=1}^{T_l-1} \sum_{t=\min(S_l)}^{\max(S_l)-m} \sum_{i \in I_k} \left| \mathrm{E} \left[\langle \tilde{\psi}_{it}^{kl}, \tilde{\psi}_{i,t+m}^{kl} \rangle | W(-k, -l) \right] \right|$$

$$=: c(1) + c(2) + c(3) + 2c(4) + 2c(5).$$

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

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

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

$$\left(\mathrm{E} \left[\|\tilde{\psi}_{it}^{kl}\|^2 | W(-k, -l) \right] \right)^{1/2}$$

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

$$\le 2 \left(\mathrm{E} \left[\|\ddot{\psi}_{it}^{kl}\|^2 | W(-k, -l) \right] \right)^{1/2} \le 2r'_{NT}.$$

10

Therefore, we have

$$c(1) \leq \mathrm{E}\left[\|\tilde{\psi}_{it}^{kl}\|^2 | W(-k, -l)\right] = O(r_{NT}'^2), \qquad c(2) \leq c\mathrm{E}\left[\|\tilde{\psi}_{it}^{kl}\|^2 | W(-k, -l)\right] = O(r_{NT}'^2),$$

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

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

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

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

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

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

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

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

where $\tilde{r} \in (0, 1)$. We note that $f_{kl}(0) = 0$ on the event $\mathcal{E}_{cp}$. On the event $\mathcal{E}_\eta \cap \mathcal{E}cp$ and under Assumption DML1(ii)(near-orthogonality), we have $\left\|f'_{kl}(0)\right\| \leq \lambda_{NT}$ and $\left\|f''_{kl(0)}\right\| \leq \lambda'_{NT}$. Therefore, we have shown that $b(2) = O_P(\lambda_{NT}) + O_P(\lambda'_{NT})$. Combining the bounds for $b(1)$ and $b(2)$ completes the proof of Claim 8.2.

$\square$

Proof of Claim 8.3

Proof. By triangle inequality, we have

$$\left|\hat{\Omega}_{a,kl} - \Sigma_a\right| \leq \left|I_{a,1}^{(kl)}\right| + \left|I_{a,2}^{(kl)}\right| + \left|I_{a,2}^{(kl)}\right|,$$

$$I_{a,1}^{(kl)} := \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \left\{\hat{\psi}_{it}^{(kl)}\hat{\psi}_{ir}^{(kl)} - \psi_{it}^{(0)}\psi_{ir}^{(0)}\right\},$$

$$I_{a,2}^{(kl)} := \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \left\{\psi_{it}^{(0)}\psi_{ir}^{(0)} - \mathrm{E}[\psi_{it}^{(0)}\psi_{ir}^{(0)}]\right\},$$

$$I_{a,2}^{(kl)} := \frac{1}{N_k T_l^2} \sum_{i \in I_k, t \in S_l, r \in S_l} \mathrm{E}[\psi_{it}^{(0)}\psi_{ir}^{(0)}] - \mathrm{E}[a_i a_i].$$

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

$$\mathrm{E}\left|h_i(\gamma_t, \gamma_r)\right|^{2+\delta} < \infty, \tag{8.9}$$

$$\int \int |h_i(u,v)|^{2+\delta} \, dF(u)dF(v) < \infty, \tag{8.10}$$

for some $\delta > 0$ and $F(.)$ is the common CDF of γ_t. Consider condition 8.9. By Minkowski's inequality,

Jensen's inequality, and the law of iterated expectation, we have

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

where the second inequality follows from Hölder's inequality and the identical distribution of γ_t. Let $\delta = \frac{p-4}{2}$, then $\left(\mathrm{E} \left| \psi_{it}^{(0)} \right|^{4+2\delta} \right)^{\frac{1}{2+\delta}} < c_m$ and $\mathrm{E} \left| \psi_{it}^{(0)} \right|^2 \le c_m^2$ follows from Assumption DML2(i). Therefore, condition 8.9 is satisfied.

Consider condition 8.10. By Minkowski's inequality and Jensen's inequality, we have

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

$$\le \left(\int \int \left| \mathrm{E}[\psi_{it}^{(0)} \psi_{ir}^{(0)} | \gamma_t = u, \gamma_r = v] \right|^{2+\delta} dF(u) dF(v) \right)^{\frac{1}{2+\delta}} + \mathrm{E} \left| \psi_{it}^{(0)} \psi_{ir}^{(0)} \right|$$

$$\le \left(\int \int \left(\mathrm{E} \left[\left| \psi_{it}^{(0)} \right|^2 | \gamma_t = u \right] \right)^{\frac{2+\delta}{2}} \left(\mathrm{E} \left[\left| \psi_{ir}^{(0)} \right|^2 | \gamma_r = v \right] \right)^{\frac{2+\delta}{2}} dF(u) dF(v) \right)^{\frac{1}{2+\delta}} + \mathrm{E} \left| \psi_{it}^{(0)} \right|^2$$

$$\le \left(\int \int \mathrm{E} \left[\left| \psi_{it}^{(0)} \right|^{2+\delta} | \gamma_t = u \right] \mathrm{E} \left[\left| \psi_{ir}^{(0)} \right|^{2+\delta} | \gamma_r = v \right] dF(u) dF(v) \right)^{\frac{1}{2+\delta}} + \mathrm{E} \left| \psi_{it}^{(0)} \right|^2$$

$$= \left(\mathrm{E} \left| \psi_{it}^{(0)} \right|^{4+2\delta} \right)^{\frac{1}{2+\delta}} + \mathrm{E} \left| \psi_{it}^{(0)} \right|^2$$

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 8.10 is also satisfied with $\delta = \frac{p-4}{2}$. By Lemma 3.4 in Dehling and Wendler (2010), we conclude

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

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

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

13

By conditional Markov inequality, we have

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

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

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

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

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

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

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

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

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

$$\lesssim R_{kl}\left\{\left\|\psi_{it}^{(0)}\right\|_{kl,2} + R_{kl}\right\},$$

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

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

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

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

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

14

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

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

where the last inequality follows from Assumption DML2(ii). Then, by Markov inequality and Lemma 6.1 from Chernozhukov et al. (2018), 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

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

$\square$

Proof of Claim 8.4

Proof. By triangle inequality, we have

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

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

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

$$I_{b,3}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{t \in S_l, i \in I_k, j \in I_k} \mathrm{E}[\psi_{it}^{(0)} \psi_{jt}^{(0)}] - c\mathrm{E}[g_t g_t'],$$

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

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

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

Due to identical distribution of γ_t, $\mathrm{E}[g_t g_t']$ 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

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

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

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

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

$$\mathrm{E}\left|\tilde{I}_{b,2,t}^{(kl)}\right|^2 = \frac{1}{N_k^4}\sum_{i,j\in I_k}\sum_{i',j'\in I_k}\mathrm{E}\left[\left(\mathrm{E}[\zeta_{ij,t}|\alpha_i,\alpha_j] - \mathrm{E}[\zeta_{ij,t}]\right)\left(\mathrm{E}[\zeta_{i'j'}|\alpha_{i'},\alpha_{j'}] - \mathrm{E}[\zeta_{i'j'}]\right)\right]$$

$$\lesssim \frac{1}{N_k}\mathrm{E}[\zeta_{ij,t}]^2 < \frac{1}{N_k}\mathrm{E}\left[\psi_{it}^{(0)}\right]^4 = O(1/N_k).$$

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

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

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

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

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

$$\lesssim \frac{1}{\sqrt{T_l}}\left(\mathrm{E}\left[|I_{b,2,t}^{(kl)}|^{\frac{q}{2}}|\{\alpha_i\}_{i\in I_k}\right]\right)^{\frac{2}{q}}$$

16

By conditional Markov inequality, we have

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

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

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

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

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

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

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

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

where $R_{kl} = \left\| \widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)} \right\|_{kl,2}$. We have shown in the proof of Claim 8.3 that $\left\| \psi_{it}^{(0)} \right\|_{kl,2} = O_P(1)$ and $R_{kl}^2 = O_P\left(N^{-1} + (r'_{NT})^2 \right)$. So $\left| I_{b,1}^{kl} \right| = O_P\left(N^{-1/2} + r'_{NT} \right)$. To summarize

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

which completes the proof of Claim 8.4.

$\square$

Proof of Claim 8.5

Proof. 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

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

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

$$I_{c,3}^{(kl)} := \frac{K/L}{N_k T_l^2} \sum_{i \in I_k, t \in S_l} \mathrm{E}[\psi_{it}^{(0)} \psi_{it}^{(0)}],$$

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

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

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

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

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

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

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

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

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

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

where $R_{kl} = \left\|\widehat{\psi}_{it}^{(kl)} - \psi_{it}^{(0)}\right\|_{kl,2}$. We have shown in the proof of Claim 8.3 that $\left\|\psi_{it}^{(0)}\right\|_{kl,2} = O_P(1)$ and $R_{kl}^2 = O_P\left(N^{-1} + (r'_{NT})^2\right)$. So, $\left|I_{c,1}^{(kl)}\right| = O_P\left(N^{-1/2}/T + r'_{NT}/T\right)$. To summarize

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

which completes the proof of Claim 8.5.

□

Proof of Claim 8.6

Proof. By triangle inequality, we have

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

where

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

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

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

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

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

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

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

$$\mathrm{E}\left[\psi_{it}^{(0)} \psi_{j,t+m}^{(0)} \right] = cov(\psi_{it}^{(0)}, \psi_{j,t+m}^{(0)})$$

$$= \mathrm{E}[cov(\psi_{it}^{(0)}, \psi_{j,t+m}^{(0)})|\gamma_t, \gamma_{t+m}] + cov(\mathrm{E}[\psi_{it}^{(0)}|\gamma_t], \mathrm{E}[\psi_{j,t+m}^{(0)}|\gamma_{t+m}])$$

$$= 0 + \mathrm{E}[g_t g_{t+m}'],$$

where the last equality follows from the properties of Hajek-type decomposition components. 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 $I_{d,5}^{(kl)} = o(1)$.

19

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

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

By iterated expectation and Jensen's inequality, we have

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

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

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

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

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

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

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

20

where

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

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

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

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

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

$$\mathrm{E} \left| \tilde{I}_{d,2,tm}^{(kl)} \right|^2 = \frac{1}{N_k^4} \sum_{i,j\in I_k} \sum_{i',j'\in I_k} \mathrm{E} \left[\left(\mathrm{E}[\zeta_{ijm}|\alpha_i,\alpha_j] - \mathrm{E}[\zeta_{ij,t}] \right) \left(\mathrm{E}[\zeta_{i'j'}|\alpha_{i'},\alpha_{j'}] - \mathrm{E}[\zeta_{i'j'}] \right) \right]$$

$$\lesssim \frac{1}{N_k} \mathrm{E}[\zeta_{ijm}]^2 < \frac{1}{N_k} \mathrm{E} \left[\psi_{it}^{(0)} \right]^4 = O(1/N_k).$$

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

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

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

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

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

21

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

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

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

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

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

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

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

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

To summarize

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

$$= o_P(1).$$

which completes the proof of Claim 8.6.

$\square$

8.2. Results for the Partial Linear Model

8.2.1. **Proof of Theorem 4.1**

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

$$A_{NT} = \frac{1}{NT}(V^Z)'V^D, \ \widehat{A}_{NT} = \frac{1}{NT}(Z - f\widehat{\zeta}_0)'(D - f\widehat{\pi}_0),$$

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

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

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

For the asymptotic normality of $\sqrt{N \wedge T}\left(\widehat{\theta} - \theta_0\right)$, we need to show the following statements: (i) $A_{NT} \overset{p}{\to} A_0 = \mathrm{E}[V_{it}^Z V_{it}^D]$; (ii) $\sqrt{N \wedge T}\psi_{NT} \overset{d}{\to} \mathcal{N}(0, \Omega_0)$; (iii) $\sqrt{N \wedge T}\left[\widehat{\psi}_{NT} - \psi_{NT}\right] = o(1)$; (iv) $\widehat{A}_{NT} - A_{NT} = o_P(1)$. With statements (i) - (iv) and the identification condition in Assumption REG-P(i) such that A_0 is non-singular, $\sqrt{N \wedge T}\left(\widehat{\theta} - \theta_0\right) \overset{d}{\to} \mathcal{N}\left(0, A_0^{-1}\Omega_0 A_0^{-1\prime}\right)$. Then, the conclusion of the theorem follows.

First, we note that Assumptions REG-P(ii) and AHK imply that $(\bar{F}_i, \bar{F}_t)$ are functions of only $(\alpha_i, \gamma_t, \varepsilon_{it})$, and so are f_{it} and V_{it}^l for $l = g, D, Y, Z$. Therefore, the results based on Hajek-type decomposition are still applicable. Furthermore, under Assumptions REG-P(ii), $\bar{F}_i$ is a function of only (c_i, ϵ_i^c) and $\bar{F}_t$ is a function of only (d_t, ϵ_t^d), so $f_{it} = (L_{1,it}, 1)$ is a function of $(X_{it}, c_i, d_t, \epsilon_i^c, \epsilon_t^d)$. By definition, $\mathrm{E}[U^D|X_{it}, c_i, d_t] = 0$. Given that $(\epsilon_i^c, \epsilon_t^d)$ are independent shocks, $\mathrm{E}[U^D|X_{it}, c_i, d_t, \epsilon_i^c, \epsilon_t^d] = 0$. Therefore, $\mathrm{E}[f_{it}U_{it}^D] = 0$. By definition of $L_{2,it}$ and that $(\epsilon_i^c, \epsilon_t^d)$ are independent shocks, $\mathrm{E}[f_{it}L_{2,it}] = \mathrm{E}[f_{it}]\mathrm{E}[L_{2,it}]$. Therefore, $\mathrm{E}[f_{it}'V_{it}^D] = \mathrm{E}\left[f_{it}[(L_{2,it} - \mathrm{E}[L_{2,it}])\eta_{D,2} + U_{it}^D]\right] = 0$. Similarly, we have $\mathrm{E}[f_{it}'V_{it}^Y] = 0$.

Statement (i) follows from Lemma 8.1 under Assumptions AHK and REG-P(iii). For Statement (ii), we first observe that $V_{it}^Z = Z_{it}(1 - \zeta_0)$ where $\zeta_0 = \left(\mathrm{E}[f_{it}'f_{it}]\right)^{-1}\mathrm{E}[f_{it}'Z_{it}]$. Due to the exogeneity condition $\mathrm{E}[Z_{it}V^g] = 0$, we have $\mathrm{E}[V_{it}^Z V_{it}^g] = 0$. With the additional Assumption REG-P(iv), Statement (ii) follows from Lemma 8.2.

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

$$
\begin{aligned}
NT|\widehat{\psi}_{NT} - \psi_{NT}| \leq & |(f(\zeta_0 - \widehat{\zeta}))'(f(\beta_0 - \widehat{\beta}) + V^Y + r^Y - \theta_0(f(\pi_0 - \widehat{\pi}) + V^D + r^D))| \\
& + |(Z - f\zeta_0)'(\theta_0(f(\widehat{\pi} - \pi_0)) - f(\beta_0 - \widehat{\beta}) + r^g)| \\
\leq & |(f(\zeta_0 - \widehat{\zeta}))'f(\beta_0 - \widehat{\beta})| + |(f(\zeta_0 - \widehat{\zeta}))'V^Y| + |(f(\zeta_0 - \widehat{\zeta}))'r^Y| \\
& + \theta_0|(f(\zeta_0 - \widehat{\zeta}))'f(\pi_0 - \widehat{\pi})| + \theta_0|(f(\zeta_0 - \widehat{\zeta}))'V^D| + \theta_0|(f(\zeta_0 - \widehat{\zeta}))'r^D| \\
& + \theta_0|(V^Z)'f(\widehat{\pi} - \pi_0)| + \theta_0|(V^Z)'f(\beta_0 - \widehat{\beta})| + \theta_0|(V^Z)'r^g| \quad (8.11)
\end{aligned}
$$

Under Assumption AHK, the sparse approximation conditions as well as Assumption REG-P(ii) - (vii), we can apply Theorem 2.1 to obtain that $\|f_{it}'(\eta_0 - \widehat{\eta})\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N\wedge T}}\right)$, $\|\eta_0 - \widehat{\eta}\|_1 = O_P\left(s\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}\right)$ for $\eta = \zeta, \pi, \beta$, and $P\left(\max_{j=1,\ldots,p}\left|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\omega_{j,l}^{-1}f_{it,j}V_{it}^l\right| \geq \frac{\lambda}{2c_1 NT}\right) \to 0$ for $l = Z, D, Y$ where $\lambda = \frac{2C_\lambda NT}{\sqrt{N\wedge T}}\Phi^{-1}(1 - \gamma/2p)$. By Lemma 8.2, $\omega_{j,l} \overset{p}{\to} \frac{A\wedge T}{N}\Sigma_{a,j,l} + \frac{N\wedge T}{T}\Sigma_{g,j,l}$ where $\min_{j\leq p}\Sigma_{a,j}^l > 0$ by Assumption REG-P(iv) and Lemma 8.1. Therefore, $\min_j\omega_{j,l}^{-1} > 0$, which implies $\frac{1}{NT}\|f'V^l\|_\infty = O_P(\Phi^{-1}(1 - \gamma/2p)/\sqrt{N\wedge T}) = O_P\left(\sqrt{\frac{\log(p/\gamma)}{N\wedge T}}\right)$ for $l = D, Y, Z$.

Consider the first term in 8.11. By Cauchy-Swartz inequality, we have $\frac{\sqrt{N \wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))'f(\beta_0 - \widehat{\beta})| \leq$
$\sqrt{N \wedge T}\|f'_{it}(\zeta_0 - \widehat{\zeta})\|_{NT,2}\|f'_{it}(\beta_0 - \widehat{\beta})\|_{NT,2} = O_P\left(\frac{s\log(p/\gamma)}{\sqrt{N \wedge T}}\right)$. Consider the second term in 8.11. By

Hölder's inequality, we have $\frac{\sqrt{N \wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))'V^Y| \leq \frac{\sqrt{N \wedge T}}{NT}\|\zeta_0 - \widehat{\zeta}\|_1\|f'V^Y\|_\infty = O_P\left(s\frac{\log(p/\gamma)}{\sqrt{N \wedge T}}\right)$. Con-

sider the third term in 8.11. By Cauchy-Swartz inequality, we have $\frac{\sqrt{N \wedge T}}{NT}|(f(\zeta_0 - \widehat{\zeta}))'r^Y| \leq \sqrt{N \wedge T}\|f'_{it}(\zeta_0 -$
$\widehat{\zeta})\|_{NT,2}\|r^Y_{it}\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}}\right)$. For the last term of 8.11, Cauchy-Swartz inequality implies that

$\frac{\sqrt{N \wedge T}}{NT}|(V^Z)'r| \leq \sqrt{N \wedge T}\|V^Y_{it}\|_{NT,2}\|r^Y_{it}\|_{NT,2}$. By Assumption REG-P(ii), we have $|\mathrm{E}[(V^Y_{it})^2]^{4(\mu+\delta)}| <$
∞. Then we can apply Lemma 8.1 and obtain that $\|V^Y_{it}\|_{NT,2} \to (\mathrm{E}[(V^Y_{it})^2])^{1/2}$. Therefore, we have
$\frac{\sqrt{N \wedge T}}{NT}|(V^Z)'r| = o_P(1)$. The arguments for the rest of the terms in 8.11 are similar. Under the sparsity

condition $s = \frac{\sqrt{N \wedge T}}{\log(p/\gamma)}$, we conclude that $\sqrt{N \wedge T}|\widehat{\psi}_{NT} - \psi_{NT}| = o_P(1)$.

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

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

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

8.2.2. *Proof of Theorem 4.2*

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

$$\widehat{\Omega}_{\mathrm{CHS}} := \widehat{\Omega}_a + \widehat{\Omega}_b - \widehat{\Omega}_c + \widehat{\Omega}_d + \widehat{\Omega}'_d,$$

$$\widehat{\Omega}_a := \frac{1}{NT^2}\sum_{i=1}^N\sum_{t=1}^T\sum_{r=1}^T \psi_{it}(\widehat{\theta}, \widehat{\eta})\psi_{ir}(\widehat{\theta}, \widehat{\eta})', \quad \widehat{\Omega}_b := \frac{1}{NT^2}\sum_{t=1}^T\sum_{i=1}^N\sum_{j=1}^N \psi_{it}(\widehat{\theta}, \widehat{\eta})\psi_{jt}(\widehat{\theta}, \widehat{\eta})',$$

$$\widehat{\Omega}_c := \frac{1}{NT^2}\sum_{i=1}^N\sum_{t=1}^T \psi_{it}(\widehat{\theta}, \widehat{\eta})\psi_{it}(\widehat{\theta}, \widehat{\eta})', \quad \widehat{\Omega}_d := \frac{1}{NT^2}\sum_{m=1}^{M-1} k\left(\frac{m}{M}\right)\sum_{t=1}^{T-m}\sum_{i=1}^N\sum_{j=1,j\neq i}^N \psi_{it}(\widehat{\theta}, \widehat{\eta})\psi_{j,t+m}(\widehat{\theta}, \widehat{\eta})'.$$

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

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

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

$$I_{a,1} := \frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \left\{ \psi_{it}(\widehat{\theta},\widehat{\eta})\psi_{ir}(\widehat{\theta},\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0) \right\},$$

$$I_{a,2} := \frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \left\{ \psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0) - \mathrm{E}[\psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0)] \right\},$$

$$I_{a,2} := \frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \left\{ \mathrm{E}[\psi_{it}(\theta_0,\eta_0)\psi_{ir}(\theta_0,\eta_0)] - \mathrm{E}[a_i^2] \right\}.$$

Note that in proving Claim 8.3, the cross-fitting device is only used to show that $I_{a,1}$ is of small order. Since the arguments for showing $I_{a,2}$ and $I_{a,3}$ to be of small order are basically the same as those in the proof of Claim 8.3, they are not repeated here.

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

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

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

By Minkowski's inequality, we have

$$R_{NT} = \left\| \psi_{it}^a(\eta_0)(\widehat{\theta} - \theta_0) + (\psi_{it}^a(\eta_0) - \psi_{it}^a(\widehat{\eta}))(\widehat{\theta} - \theta_0) + \psi_{it}(\theta_0,\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0) \right\|_{NT,2}$$

$$\leq \left\| \psi_{it}^a(\eta_0)(\widehat{\theta} - \theta_0) \right\|_{NT,2} + \left\| (\psi_{it}^a(\eta_0) - \psi_{it}^a(\widehat{\eta}))(\widehat{\theta} - \theta_0) \right\|_{NT,2} + \left\| \psi_{it}(\theta_0,\widehat{\eta}) - \psi_{it}(\theta_0,\eta_0) \right\|_{NT,2}$$

$$=: R_{a,1} + R_{a,2} + R_{a,3},$$

where $\psi_{it}^a(\eta) := \left(Z_{it} - f_{it}\zeta \right)\left(D_{it} - f_{it}\pi \right)$. Under Assumption REG-P(ii), we have

$$\mathrm{E}[\psi_{it}^a(\eta_0)]^2 = \mathrm{E}[V_{it}^Z(V_{it}^D + r_{it}^D)]^2 = O_P(1),$$

and Markov inequality implies that $\left\| \psi_{it}^a(\eta_0) \right\|_{NT,2} = O_P(1)$. By Theorem 4.1, we have $\widehat{\theta} - \theta_0 = O_P\left(\frac{1}{\sqrt{N \wedge T}} \right)$.

Therefore, $R_{a,1} \leq \left\| \psi_{it}^a(\eta_0) \right\|_{NT,2} |\widehat{\theta} - \theta_0| = O_P\left(\frac{1}{\sqrt{N \wedge T}} \right)$. To bound $R_{a,2}$, we note

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

Under Assumption REG-P(iii), we have $\mathrm{E}|V_{it}^D|^{8(\mu+\delta)} < \infty$, which implies $\mathrm{E}\left[\max_{i \leq N, t \leq T} |V_{it}^D|^2 \right] \lesssim (NT)^{\frac{1}{4(\mu+\delta)}}$. By Markov inequality, we have $\max_{i \leq N, t \leq T} |V_{it}^D|^2 = O_P((NT)^{\frac{1}{4(\mu+\delta)}})$. As in the proof of Theorem 4.1, The-

25

orem 2.1 can be applied to obtain $\left\| f'_{it}(\widehat{\zeta} - \zeta_0) \right\|_{NT,2} = O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}} \right)$. Then, we have

$$R_{a,2} = \left\| f'_{it}(\widehat{\zeta} - \zeta_0) V_{it}^D \right\|_{NT,2} \leq \left(\max_{i \leq N, t \leq T} |V_{it}^D|^2 \right)^{1/2} \left\| f'_{it}(\widehat{\zeta} - \zeta_0) \right\|_{NT,2}$$

$$= O_P((NT)^{\frac{1}{8(\mu+\delta)}}) O_P\left(\sqrt{\frac{s\log(p/\gamma)}{N \wedge T}} \right) = O_P((NT)^{\frac{1}{8(\mu+\delta)}}) o_P\left(\frac{1}{(N \wedge T)^{1/4}} \right) = o_P(1).$$

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

It is left to show that $\widehat{\Omega}_b \xrightarrow{p} cE[g_t^2]$, $\widehat{\Omega}_c = o_P(1)$, and $\widehat{\Omega}_d \xrightarrow{p} c\sum_{m=1}^{\infty} E[g_t g_{t+m}]$. As is shown in the proofs of Claim 8.4 in Appendix C of the Online Supplementary Material, the only step in showing these claims that involves the cross-fitting technique is to show the same term R_{NT} to converge to 0 in probability. Otherwise, the arguments are basically the same and not repeated here. Combining these results, we obtain $\widehat{\Omega} \xrightarrow{p} E(a_i^2) + cE(g_t^2) + c\sum_{m=1}^{\infty} E(g_t g_{t+m}) = \Sigma_a + c\Sigma_g$.

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

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

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

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

Following the same arguments as in the proof of Claim 8.6, we have $\widehat{\Omega}_e = \widehat{\Omega}_d + o_P(1)$. We have shown $\widehat{\Omega}_c = o_P(1)$. Therefore, we conclude that $\widehat{\Omega}_{NW} = o_P(1)$. So it is proved. $\qquad \square$