Standard Errors for Two-Way Clustering with Serially Correlated Time Effects*

Harold D. Chiang† Bruce E. Hansen‡ Yuya Sasaki§

Abstract

We propose improved standard errors and an asymptotic distribution theory for two-way clustered panels. Our proposed estimator and theory allow for arbitrary serial dependence in the common time effects, which is excluded by existing two-way methods, including the popular two-way cluster standard errors of Cameron, Gelbach, and Miller (2011) and the cluster bootstrap of Menzel (2021). Our asymptotic distribution theory is the first which allows for this level of interdependence among the observations. Under weak regularity conditions, we demonstrate that the least squares estimator is asymptotically normal, our proposed variance estimator is consistent, and t-ratios are asymptotically standard normal, permitting conventional inference. The main results extend to two-way fixed-effect models; we argue that two-way clustering is still necessary even if two-way fixed effects are included in estimation. We present simulation evidence that confidence intervals constructed with our proposed standard errors obtain superior coverage performance relative to existing methods. We illustrate the relevance of the proposed method in an empirical application to a standard Fama-French three-factor regression.

Keywords: panel data, serial correlation, standard errors, two-way clustering.

*We benefited from useful comments by A. Colin Cameron and seminar participants at Essex, Kobe, LSE, Michigan State, North Carolina State, Singapore Management University, and UC Davis, and participants in AMES in East and South-East Asia 2022, Cemmap/SNU Workshop on Advances in Econometrics 2022, CIREQ Montréal Econometrics Conference 2022, and NAWM 2023. All the remaining errors are ours. Hansen thanks the National Science Foundation and Phipps Chair for research support. Sasaki thanks Brian and Charlotte Grove Chair for research support. The Stata command is available to install by `ssc install xtregtwo`.

†Harold D. Chiang: hdchiang@wisc.edu. Department of Economics, University of Wisconsin-Madison, William H. Sewell Social Science Building, 1180 Observatory Drive, Madison, WI 53706-1393, USA

‡Bruce E. Hansen: bruce.hansen@wisc.edu. Department of Economics, University of Wisconsin-Madison, William H. Sewell Social Science Building, 1180 Observatory Drive, Madison, WI 53706-1393, USA

§Yuya Sasaki: yuya.sasaki@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA

1 Introduction

A standard panel data set has observations double-indexed over firms i and time t.[1] A panel is said to have a two-way dependence structure if there is dependence across individuals at any given time, and across time for any given individual. A common model for two-way dependence is the components structure $U_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$, where α_i is a firm effect, γ_t is a time effect, and ε_{it} is an idiosyncratic effect. It is typical to view the time effects γ_t as omitted macroeconomic variables, such as the state of the business cycle. Therefore, they are unlikely to be serially independent. Consequently, it is reasonable to treat γ_t as a serially correlated time-series process.

Serial correlation in the common time-effects, however, creates an extra layer of serial dependence beyond two-way dependence. It induces dependence among observations which do not share a common firm or time index. This fundamentally complicates the dependence structure, rendering existing theory and methods inappropriate. Most importantly for practice, existing two-way clustered inference methods do not allow time effects with arbitrary serial correlation. Moreover, no formal asymptotic theory has previously been developed under the two-way clustering setting with serially correlated time effects. This paper sheds new light on literature of two-way clustering by formally investigating asymptotic theory for serially correlated time effects in this context and proposing novel and theoretically supported method of inference under such settings.

The most popular inference method for two-way dependent panels is the two-way clustered standard errors of Cameron, Gelbach, and Miller (2011), which we shall henceforth refer to as CGM. A recent related method is the bootstraps of Menzel (2021); see also Davezies, D'Haultfœuille, and Guyonvarch (2021, Sec. 3.3) for a generalization to empirical processes. Both of these approaches allow for the components structure $U_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$, but only under the additional strong condition that the time effects γ_t are serially independent. The CGM standard errors explicitly calculate the variance allowing for traditional two-way dependence, not allowing for dependence induced by serially

[1] The index i can refer to any entity, such as firms, individuals, or households. For simplicity we will refer to these entities as "firms".

correlated time effects. Consequently, these methods exclude, by construction, the possibility that the common time component γ_t is an unmodelled macroeconomic effect.

An important intuitive extension due to Thompson (2011) allows γ_t to be serially correlated up to a known fixed number of lags and suggests to estimate the asymptotic variance by including unweighted, lagged autocovariance estimates to the CGM estimator. This relaxes the CGM assumptions by allowing serial correlation structures that are of m-dependence. In practice, however, it is difficult to implement since the serial dependence structure is not known a priori. Also, even under this m-dependence setting, no asymptotic distribution theory was provided. This is particularly troubling since serial correlated time effects induces a complicated dependence structure and thus was unclear whether this inference procedure is theoretically justified and under which conditions it is so. Indeed, our simulations unveil that this unweighted, fixed number of lags approach shows unsatisfactory finite sample performance under various DGPs; see Section 5. In addition, based on his own simulations, Thompson (2011) recommends omitting the correction for serial correlation unless the time dimension is large. Thus in practice, the Thompson estimator actually implemented by most (if not all) empirical researchers reduces to the CGM two-way estimator.

Furthermore, an asymptotic distribution theory for regression with two-way clustering with general serial correlated time effects is missing. Cameron, Gelbach, and Miller (2011) assert an asymptotic theory for estimation, but do not examine the impact of two-way dependence, nor examine standard error estimation. As previously mentioned, Thompson (2011) does not provide a distribution theory even under the m-dependence setting. Davezies, D'Haultfœuille, and Guyonvarch (2021), MacKinnon, Nielsen, and Webb (2021), and Menzel (2021) do provide rigorous theory, yet only for settings without serial dependence.

Clustered inference can alternatively be based on unstructured one-way dependence (over either i or t, but not both simultaneously) using the popular clustered variance estimator of Liang and Zeger (1986) and Arellano (1987). These methods, however, cannot account for two-way dependence. An alternative framework is one-way-cluster dependence across i with weak serial dependence across t

(Driscoll and Kraay, 1998). A yet alternative framework has been provided by Vogelsang (2012) and Hidalgo and Schafgans (2021), which study panels with cross-sectional and temporal dependence under a different set of conditions. They allow dependence across firms and time, but the dependence between observations within a time period, as well as the dependence of a cross-sectional unit observed over time, both decay as observations get further apart in time and space. Consequently, these alternative frameworks do not allow arbitrary two-way clustering.

Clustered standard errors have become ubiquitous in applied economic research, as evidenced by a perusal of current applied journals, and by the enormous citations to several of the above-mentioned papers. Petersen (2009) provides an excellent review of these popular methods and their use in empirical research through 2009. Our perusal of current applied journals reveals that nearly all applications use either Liang-Zeger-Arelleno one-way clustering or CGM two-way clustering. While Thompson (2011) is also highly cited, our review indicates that empirical applications do not employ his correction for correlated time effects, but rather use the simpler CGM two-way clustering.

In this article, we modify the CGM and Thompson two-way clustered standard error to accommodates time effects with arbitrary stationary serial dependence. Our approach allows for cluster dependence within individuals i, within time periods t, and allows the common time component γ_t to be serially dependent of arbitrary order. Ours is the first approach which allows this complexity of two-way dependence. This is accomplished by a correction involving kernel smoothing over the autocorrelations, with the number of autocorrelation lags increasing with sample size. To select the lag truncation parameter, we propose a simple rule based on Andrews (1991).

We provide an asymptotic theory of inference under weak regularity conditions, including the assumption that the time effects γ_t are strictly stationary and mixing. We show that the least squares estimator is asymptotically normal, our proposed variance estimator is consistent, and t-ratios are asymptotically standard normal, permitting conventional inference. The proofs of these results are far from trivial. For example, our consistency proof for our proposed cluster-robust variance estimator is nonstandard. We show that the problem can be re-written into a claim of bounding a fourth-order sum

of cross-moments of dependent time series, for which a mixing bound due to Yoshihara (1976) can be applied. Furthermore, the same proof uses novel projection arguments which simplify the derivations.

We explore the performance of our proposed method in a simple simulation experiment which compares the coverage probability of confidence intervals constructed with six different standard error methods. We find that our proposed method has the best performance relative to the competitors in each simulation design considered, and in some settings the difference is substantial.

We also illustrate the relevance of the method with an empirical application to estimation of the slope coefficients in a standard Fama-French three-factor regression using two panels of stock returns. We find that our proposed standard errors are different – and larger – than conventional standard errors, for four of six regression estimates examined.

The rest of this paper is organized as follows. Section 2 discusses two-way dependence with correlated time effects and provides an informal overview of the method with a practical guide. Section 3 presents the formal theoretical results. We discuss extensions to fixed-effect models in Section 4. Section 5 provides simulation evidence on the practical performance of our proposed method. Section 6 presents an empirical application to a standard Fama-French regression. The appendix collects a mathematical proof of the main result, auxiliary lemmas and their proofs, and additional details omitted from the main text. The Stata command is available to install by `ssc install xtregtwo`.

2 Two-Way Dependence with Correlated Time Effects

2.1 Least Squares Estimation

Let (Y_{it}, X_{it}') be a panel of observations over $i = 1, ..., N$ and $t = 1, ..., T$, where Y_{it} is real-valued and X_{it} is a $k \times 1$ vector. The model is the linear regression equation

$$Y_{it} = X_{it}'\beta + U_{it} \tag{2.1}$$

with

$$E[X_{it}U_{it}] = 0. \tag{2.2}$$

The standard estimator for β is least squares

$$\widehat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} Y_{it} \right). \tag{2.3}$$

The least squares residuals are $\widehat{U}_{it} = Y_{it} - X_{it}'\widehat{\beta}$.

We are interested in the variance of $\widehat{\beta}$. It can be calculated explicitly under the auxiliary assumption that the regressors are fixed and the error is strictly exogenous.[2] We only use this assumption to motivate our covariance matrix estimator, however, and will not be needed for our asymptotic distribution theory.

The variance of $\widehat{\beta}$ can be written as follows. Define the firm sums $R_i = \sum_{t=1}^{T} X_{it} U_{it}$, the time sums $S_t = \sum_{i=1}^{N} X_{it} U_{it}$, and the cross-sums $G_m = \sum_{t=1}^{T-m} S_t S_{t+m}'$ and $H_m = \sum_{i=1}^{N} \sum_{t=1}^{T-m} X_{it} U_{it} X_{i,t+m}' U_{i,t+m}$. With a little algebra we obtain the following decomposition.

$$V_{NT} = var(\widehat{\beta}) = \widehat{Q}^{-1} \Omega_{NT} \widehat{Q}^{-1} \tag{2.4}$$

where

$$\widehat{Q} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X_{it}' \tag{2.5}$$

and

$$\Omega_{NT} = \frac{1}{(NT)^2} \sum_{i=1}^{N} E\left[R_i R_i'\right] \tag{2.6}$$

$$+ \frac{1}{(NT)^2} \sum_{t=1}^{T} E\left[S_t S_t'\right] \tag{2.7}$$

$$- \frac{1}{(NT)^2} \sum_{i=1}^{N} \sum_{t=1}^{T} E[X_{it} X_{it}' U_{it}^2] \tag{2.8}$$

$$+ \frac{1}{(NT)^2} \sum_{m=1}^{T-1} E\left[G_m + G_m' - H_m - H_m'\right]. \tag{2.9}$$

The expression (2.6)-(2.9) decomposes the variance of the least squares estimator into four components: (2.6) is the variance of the firm sums; (2.7) is the variance of the time sums; (2.8) is a correction

[2]The strict exogeneity condition is not required for our main theory and is used only for an illustration purpose in the current section for the exact variance calculations.

for double-counting of the common variance in (2.6) and (2.7); and (2.9) is the autocovariances of the time sums, corrected for double-counting.

2.2 Variance Estimation

Estimators of the variance matrix V_{NT} take the general form

$$\widehat{V}_{NT} = \widehat{Q}^{-1}\widehat{\Omega}_{NT}\widehat{Q}^{-1} \tag{2.10}$$

where $\widehat{\Omega}_{NT}$ is some estimator of Ω_{NT}. Different estimators make distinct assumptions on the covariances in (2.6)-(2.9) which lead to distinct estimators for Ω_{NT} in (2.10). The Liang-Zeger-Arellano one-way cluster estimator assumes that observations are independent across i, implying that (2.7)+(2.8)+(2.9) equals zero. The "cluster within t" estimator assumes that observations are independent across t, implying that (2.6)+(2.8)+(2.9) equals zero. The CGM two-way estimator assumes that observations it and js are independent if $i \neq j$ or $t \neq s$, implying that (2.9) equals zero. The respective estimators take the same form as the assumed non-zero expressions in (2.6)-(2.9). For example, the CGM variance estimator of Ω_{NT} is

$$\frac{1}{(NT)^2}\left(\sum_{i=1}^{N}\widehat{R}_i\widehat{R}_i' + \sum_{t=1}^{T}\widehat{S}_t\widehat{S}_t' - \sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'\widehat{U}_{it}^2\right) \tag{2.11}$$

where $\widehat{R}_i = \sum_{t=1}^{T} X_{it}\widehat{U}_{it}$ and $\widehat{S}_t = \sum_{i=1}^{N} X_{it}\widehat{U}_{it}$.

Thompson (2011) assumes that the across-firm autocovariances are non-zero for small lags m, but zero for lags beyond a known constant M. This implies that the sum over m in (2.9) can be truncated above $m = M$. This motivates his estimator of Ω_{NT}, which is (2.11) plus

$$\frac{1}{(NT)^2}\sum_{m=1}^{M}\left(\widehat{G}_m + \widehat{G}_m' - \widehat{H}_m - \widehat{H}_m'\right)$$

where $\widehat{G}_m = \sum_{t=1}^{T-m}\widehat{S}_t\widehat{S}_{t+m}'$ and $\widehat{H}_m = \sum_{t=1}^{T-m}\sum_{i=1}^{N} X_{it}\widehat{U}_{it}X_{i,t+m}'\widehat{U}_{i,t+m}$. Thompson (2011) does not discuss selection of M, other than to indicate that it is known a priori. For his simulations and empirical applications he sets $M = 2$, which we take to be his default choice.

We illustrate the dependence patterns assumed by the different estimators in Figure 1. Each panel shows an array with each entry depicting firm/time pairs (i, t), with the star $\star$ marking the

reference point $(i,t) = (1,1)$, and dependence structures indicated by the grey shading. Panel (A) illustrates the case of independent observations (which corresponds to the unclustered Eicker-Huber-White estimator) where the observation $(i,t) = (1,1)$ is uncorrelated with all other observations. Panel (B) illustrates the case of independence across firms (which corresponds to the Liang-Zeger-Arellano one-way cluster estimator) where the observation $(1,1)$ is correlated with $(1,t)$ for $t > 1$, but is uncorrelated with all other observations. Panel (C) similarly illustrates the case of independence across time (the "cluster within t" estimator). Panel (D) illustrates the case where the observation $(1,1)$ is correlated with $(1,t)$ for $t > 1$ and with $(i,1)$ for $i > 1$ (corresponding to the CGM two-way clustered estimator). Panel (E) illustrates the case where two-way clustering is augmented to allow dependence between $(1,1)$ and (i,t) for all $t \leq 3$. This corresponds to Thompson's estimator. Finally, panel (F) illustrates the case where observation $(1,1)$ is correlated with all other observations. The dark-to-light shading is meant to imply that the correlation between $(1,1)$ and (i,t) is expected to diminish for $t > 1$.

2.3 Correlated Time Effects

To understand the source of cross-firm and cross-time dependence it is illuminating to consider the linear components model $Y_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$ under the assumption of i.i.d. firm effects α_i and idiosyncratic effects ε_{it}. If the time effect γ_t is also i.i.d. then observations (i,t) and (j,s) are independent if $i \neq j$ and $t \neq s$. However, if γ_t is serially dependent, then observations (i,t) and (j,s) can be dependent for arbitrary indices.

In most applications the time effect γ_t is a proxy for omitted macroeconomic factors, and is therefore unlikely to be i.i.d. or have truncated serial dependence. Most macroeconomic variables have untruncated autocorrelation functions.

We empirically illustrate the importance of this feature. Consider two variables involved in a standard market value equation: log Tobin's average Q (market value divided by the stock of non-R&D assets), and log of the R&D stock (relative to the stock of non-R&D assets). Panel regressions

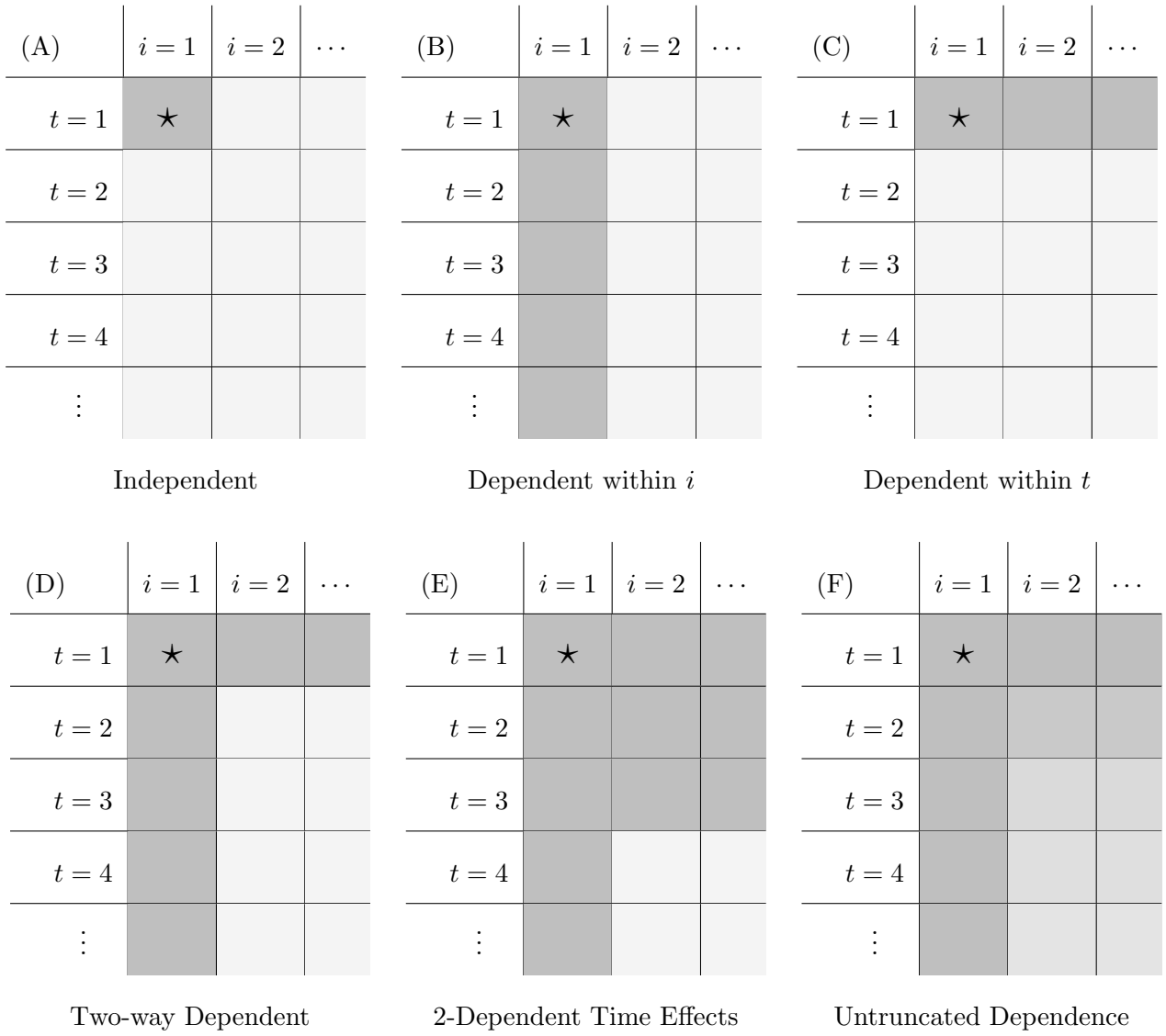


Figure 1: Clustered Dependence Structures.

of the former on the latter have been the focus of Griliches (1981) and many subsequent papers (e.g., Hall, Jaffe, and Trajtenberg, 2005; Bloom, Schankerman, and Van Reenen, 2013; Arora, Belenzon, and Sheer, 2021). Using a panel of 727 firms for the years 1981–2001 from Bloom, Schankerman, and Van Reenen (2013) (see Appendix H.2 for details) we estimated time effects for each series. Estimated time effects are plotted in the two panels of Figure 2, with log R&D stock on the left and log Tobin's Q on the right. The graphs reveal considerable serial correlation. Their estimated first-

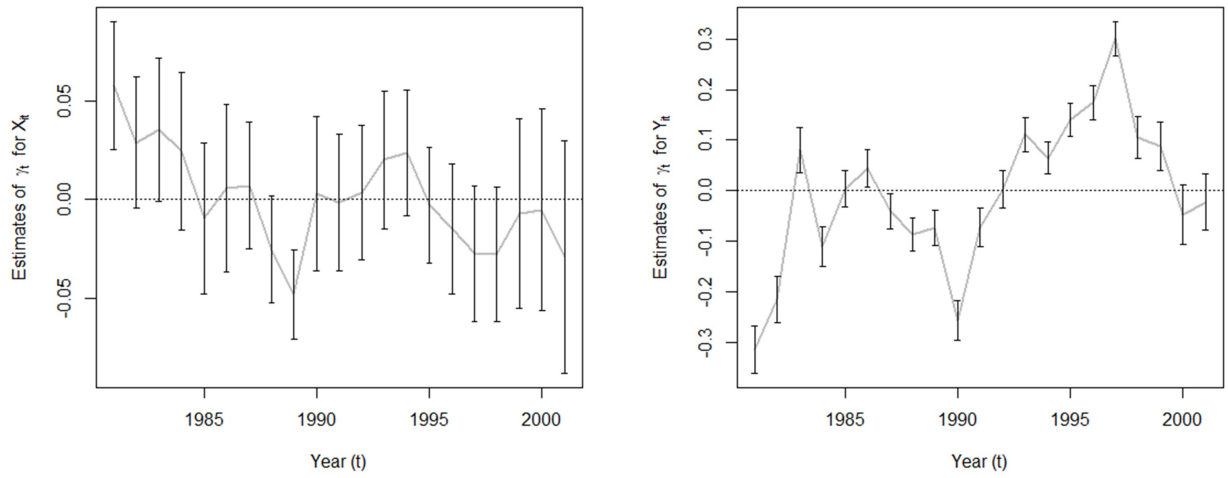


Figure 2: Estimates of the common time effects for log R&D stock (left) and log Tobin's average Q (right). The vertical lines indicate pointwise 95% confidence intervals.

order autocorrelations are 0.425 (with a standard error of 0.003), and 0.467 (with a standard error of 0.007), respectively, which are quite large. Furthermore, the autocorrelations are strong at multiple lags as illustrated in their autocorrelograms, which are displayed in Figure 3. Together, this means that the time effects γ_t for these series have substantial serial dependence, which is not well described by finite M-dependence. This implies that the dependence structures assumed by Cameron, Gelbach, and Miller (2011), Menzel (2021), and Thompson (2011) are incorrect, and rather need to be modified to allow for serial correlation of arbitrary order, as we propose in the next section.

2.4 Variance Estimation with Serially Correlated Time Effects

As described in the previous section, serially correlated time effects γ_t imply that the cross-firm autocorrelations G_m in the variance decomposition (2.9) are non-zero at potentially any lag m. However, we cannot estimate these correlations well at all lags m for fixed T, for the same reasons as arise in time-series variance estimation. Under the assumption that the time effects γ_t are strictly stationary

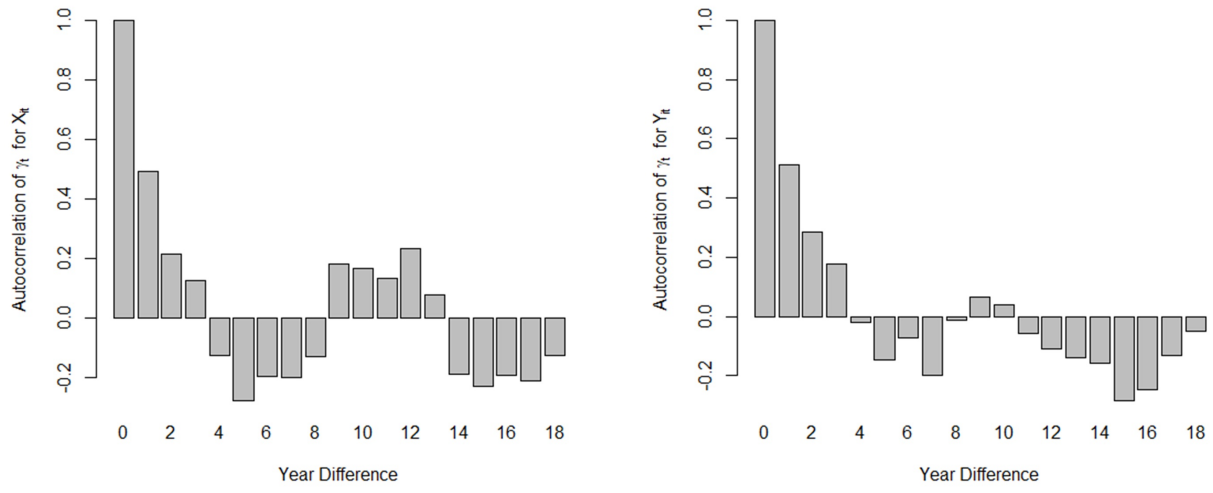


Figure 3: Autocorrelograms of the common time effects for log R&D stock (left) and log Tobin's average Q (right).

and weakly dependent (meaning that the autocorrelation function decays to zero) then it is sufficient to focus on the small lags m, using a weighted average of the terms in (2.9), with the number of terms increasing with sample size. This motivates the following estimator of Ω_{NT}

$$\widehat{\Omega}_{NT} = EVC\left(\frac{1}{(NT)^2}\left(\sum_{i=1}^{N}\widehat{R}_i\widehat{R}_i' + \sum_{t=1}^{T}\widehat{S}_t\widehat{S}_t' - \sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'\widehat{U}_{it}^2\right) + \frac{1}{(NT)^2}\sum_{m=1}^{M}w(m,M)\left(\widehat{G}_m + \widehat{G}_m'\right)\right) \tag{2.12}$$

where $w(m,M)$ is a weight function and $EVC(\cdot)$ is an eigenvalue correction.[3] Inserted into (2.10) we obtain our covariance estimator for $\widehat{\beta}$. This variance estimator is a generalization of the Cameron, Gelbach, and Miller (2011), Thompson (2011), and Newey and West (1987) estimators. The covariance matrix estimator $\widehat{V}_{NT}$ can be multiplied by degree-of-freedom adjustments if desired, but we are unaware of any finite sample justification for a particular choice. The estimator simplifies to that of Cameron, Gelbach, and Miller (2011) when $M = 0$, and to that of Thompson (2011) when $M = 2$

[3]The $EVC(\cdot)$ replaces any negative eigenvalue by zero to ensure positive semidefiniteness.

and $w(m, M) = 1$. Taking the square root of the diagonal elements of $\widehat{V}_{NT}$ yields standard errors for the elements of $\widehat{\beta}$. We omit the $\widehat{H}_m$ terms in (2.12), as they are asymptotically negligible.

The estimator (2.12) depends on the choice of weights $w(m, M)$. Standard choices include the uniform (truncated) weights $w(m, M) = 1$, and the triangular (Newey-West) weights $w(m, M) = 1 - m/(M+1)$, the latter popularized by Newey and West (1987) for time-series data. We recommend the triangular weights. One advantage of this choice for time series applications as emphasized by Newey and West (1987) is that this ensures a non-negative variance estimator. Unfortunately, the clustered estimator (2.12) is not necessarily non-negative, even for $M = 0$, as observed by Cameron, Gelbach, and Miller (2011). However, the estimator (2.12) is considerably less likely to be negative when the weights are triangular than uniform, which is an important practical advantage.

The variance estimator (2.12) critically depends on the number M, which is often called the lag truncation parameter. It is useful to note that M does not need to be integer-valued. The choice of M leads to a bias/precision trade-off, with larger values of M leading to less bias in the estimator $\widehat{\Omega}_{NT}$ of Ω_{NT}, but less precision. In principle it is desirable to use a larger value of M when the errors U_{it} are highly serially correlated, and a smaller value of M otherwise, but the extent of serial correlation is generally unknown, leading to the need for an empirical-based choice of M.

In the context of time-series variance estimation Andrews (1991) proposed a data-driven choice of M based on minimizing the asymptotic mean square error of the variance estimator, which is equivalent to the expression in (2.12). We can therefore apply his method for selection of M, treating the time-sums of the regression scores as time-series observations. Andrews' formula depends on the specific choice of weight function; we assume triangular weights.

For $j = 1, ..., k$, let X_{jit} be the jth element of X_{it}. Define the time-sums $S_{jt} = \sum_{i=1}^{n} X_{jit}\widehat{U}_{it}$ of the regression scores. Fit by least squares the AR(1) equations $S_{jt} = \widehat{\rho}_j S_{j,t-1} + \widehat{e}_{jt}$. The Andrews rule[4]

[4]This is calculated from Andrews' equation (6.4), setting his weights w_a to equal the inverse squared variances of the estimated AR(1) processes, which is appropriate for least squares estimation.

for the lag truncation M is

$$\widehat{M} = 1.8171 \cdot \left(\frac{\sum_{j=1}^{k} \frac{\widehat{\rho}_j^2}{(1-\widehat{\rho}_j)^4}}{\sum_{j=1}^{k} \frac{\left(1-\widehat{\rho}_j^2\right)^2}{(1-\widehat{\rho}_j)^4}} \right)^{1/3} T^{1/3}. \tag{2.13}$$

For the case of a scalar regressor this simplifies to

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

For an even simpler choice, Stock and Watson (2020) suggested the following rule-of-thumb. Setting $\rho = 0.25$ in the above formula (which occurs in a regression when both the regressor and regression error are AR(1) processes with AR(1) coefficients 0.5) the Andrews rule simplifies to $M = 0.75 \cdot T^{1/3}$. For example, for $T = 50$, 100, and 200, respectively, the Stock-Watson rule is $M = 2.7$, 3.5, and 4.4, respectively. The Stock-Watson can be used in place of the Andrews rule (2.13) if desired.

3 Econometric Theory

Consider a panel $\{D_{it} : 1 \leq i \leq N; 1 \leq t \leq T\}$ of random vectors consisting of observed and/or unobserved variables that are relevant to the data generating process of the researcher's interest. For instance, in the linear regression model presented in Section 2, set $D_{it} = (Y_{it}, X_{it}', U_{it})'$. With a Borel-measurable function f (generally unknown to the researcher), we consider the framework of panel dependence in D_{it} generated through the stationary process

$$D_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it}), \tag{3.1}$$

where α_i, γ_t, and ε_{it} are random vectors of arbitrary dimension, with the sequences $\{\alpha_i\}$, $\{\gamma_t\}$ and $\{\varepsilon_{it}\}$ mutually independent, α_i is i.i.d. across i, and ε_{it} is i.i.d. across (i,t).[5] The sequence γ_t is a strictly stationary serially correlated process.

[5]These i.i.d. conditions may be relaxed in some ways, but we leave it for future research. The existing literature on two-way clustering explicitly or implicitly assumes these i.i.d. conditions, and we continue to impose them in this paper. Our focus, therefore, is to relax the i.i.d. assumption on the γ_t component. One way to relax the i.i.d. conditions on the other components is to impose an MDS-type condition, in which case our main results remain to hold. Another is to allow for spatial mixing provided a researcher has spatial information associated with panel data. Also, see Section 7.

The representation (3.1) generalizes the Aldous-Hoover-Kallenberg (AHK) representation (Kallenberg, 2006) which has been widely used for two-way clustering theory. See (e.g., Davezies et al., 2021; MacKinnon et al., 2021; Menzel, 2021). Indeed, it has been argued that the AHK representation is a natural modelling framework for two-way clustered data (MacKinnon, Nielsen, and Webb, 2021).[6] A limitation of the AHK representation is that the time effects γ_t are mutually independent. We relax this assumption, by directly assuming that (3.1) holds, allowing γ_t to be serially dependent. This is a strict generalization of the AHK representation.

In this section we provide an asymptotic distribution theory for the least squares estimator, our proposed covariance matrix estimator, and associated test statistics. We start in Section 3.1 by examining a multivariate mean, followed in Section 3.2 with linear regression.

3.1 Estimation of the Mean

In this subsection, we focus on estimation of a multivariate mean. Let X_{it} be an $m \times 1$ random vector satisfying equation (3.1) for $D_{it} = X_{it}$. The standard estimator of the population mean $\theta = E[X_{it}]$ is the sample mean $\widehat{\theta} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}$.

To obtain an asymptotic representation for $\widehat{\theta}$, we use a Hoeffding-type decomposition. For simplicity and without loss of generality assume that $E[X_{it}] = 0$. Define the random vectors $a_i = E[X_{it} \mid \alpha_i]$, $b_t = E[X_{it} \mid \gamma_t]$, and $e_{it} = X_{it} - a_i - b_t$. This gives rise to the following construct:

$$X_{it} = a_i + b_t + e_{it}. \tag{3.2}$$

This expresses X_{it} as a linear function of a firm effect a_i, time effect b_t, and error e_{it}. However, as (3.2) is a derived relationship, the error e_{it} is not (in general) i.i.d. The decomposition has the following properties. The random vectors a_i and b_t are independent since they are each functions of the independent sequences $\{\alpha_i\}$ and $\{\gamma_t\}$. The sequence $\{a_i\}$ is i.i.d., and the sequence $\{b_t\}$ is strictly

[6] MacKinnon, Nielsen, and Webb (2021) state "[a] natural stochastic framework for the regression model with multiway clustered data is that of separately exchangeable random variables." Since separately exchangeable random variables may be represented by the AHK (cf. Kallenberg, 2006), we make this assertion.

stationary. By iterated expectations we deduce the following: (1) a_i, b_t, and e_{it} are each mean zero; (2) $E[e_{jt} \mid \alpha_i] = 0$ and $E[e_{is} \mid \gamma_t] = 0$ for any i, j, t, and s; (3) $E[a_i e'_{jt}] = 0$ and $E[b_t e'_{is}] = 0$ for any i, j, t, and s; (4) the sequences $\{a_i\}$, $\{b_t\}$, and $\{e_{it}\}$ are mutually uncorrelated; (5) conditional on (γ_t, γ_s), e_{it} and e_{js} are independent for $j \neq i$.

Taking averages we find that

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{T}\sum_{t=1}^{T} b_t + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}. \tag{3.3}$$

The uncorrelatedness of the sequences implies that the three sums are uncorrelated. Hence the variance of $\widehat{\theta}$ equals the sum of the variance matrices of the three components. The variance of the first component equals N^{-1} times the variance matrix of a_i, the variance of the second component approximately equals T^{-1} times the long-run variance matrix of b_t (since the latter is serially correlated), and the variance of the third component approximately equals $(NT)^{-1}$ times the long-run variance matrix of e_t. The technical details are deferred to Appendix A.

We now present sufficient conditions for this decomposition to be valid.

Assumption 1. For some $r > 1$ and $\delta > 0$, (i) $X_{it} = f(\alpha_i, \gamma_t, \varepsilon_{it})$ where $\{\alpha_i\}$, $\{\gamma_t\}$, and $\{\varepsilon_{it}\}$ are mutually independent sequences, α_i is i.i.d across i, ε_{it} is i.i.d across (i,t), and γ_t is strictly stationary. (ii) $E[||X_{it}||^{4(r+\delta)}] < \infty$. (iii) γ_t is an α-mixing sequence with size $2r/(r-1)$, that is, $\alpha(\ell) = O(\ell^{-\lambda})$ for a $\lambda > 2r/(r-1)$.

Assumption 1 (i) assumes that the observed random vectors are generated following the nonlinear, nonseparable, factor structure (3.1). Assumption 1 (ii) imposes moment conditions. Assumption 1 (iii) imposes weak dependence on the time effects. The moment and mixing conditions here are standard in time-series theory, including that of Newey and West (1987) and Hansen (1992).

Define the variance matrices:

$$\Sigma_a = E[a_i a_i'] \tag{3.4}$$

$$\Sigma_b = \sum_{\ell=-\infty}^{\infty} E[b_t b'_{t+\ell}] \tag{3.5}$$

$$\Sigma_e = \sum_{\ell=-\infty}^{\infty} E[e_{it} e'_{i,t+\ell}] \tag{3.6}$$

which are independent of i and t. Given the decomposition (3.2), we can write the variance of the sample mean as a weighted sum of the variance components (3.4)-(3.6).

Theorem 1. *Suppose that Assumption 1 holds. Then* $||\Sigma_a|| < \infty$, $||\Sigma_b|| < \infty$, *and* $||\Sigma_e|| < \infty$, *and as* $(N, T) \to \infty$,

$$var(\widehat{\theta}) = \frac{1}{N}\Sigma_a + \frac{1}{T}\Sigma_b\,(1 + o(1)) + \frac{1}{NT}\Sigma_e\,(1 + o(1))\,.$$

Furthermore, $\widehat{\theta} \xrightarrow{p} \theta$ *as* $N, T \to \infty$.

A proof is provided in Appendix A. Theorem 1 shows that the asymptotic variance of the sample mean depends on three components, inversely proportional to the number of firms N, time dimension T, and their product NT. When either $\Sigma_a > 0$ or $\Sigma_b > 0$ (which occurs when there is a non-degenerate firm or time effect) then the third term in the asymptotic variance is of lower stochastic order. However, in the special case where X_{it} is i.i.d., then $\Sigma_a = 0$ and $\Sigma_b = 0$ so the first two terms equal zero, the long-run variance of e_{it} simplifies to $\Sigma_e = var(X_{it})$, and the variance expression simplifies to $var(\widehat{\theta}) = var(X_{it})/NT$. Consequently, the rate of convergence of the sample mean depends on the cluster structure.

For our distribution theory we require the following additional condition.

Assumption 2. One of the the following two conditions holds.

(i) Either $\Sigma_a > 0$ or $\Sigma_b > 0$, and $N/T \to c \in (0, \infty)$ as $(N, T) \to \infty$.

or

(ii) X_{it} are independent and identically distributed across i and t, and $var(X_{it}) > 0$.

Assumption 2 (i) requires the presence of at least one-way clustering. This assumption is analogous to the positive definiteness condition in Davezies et al. (2021, Propositions 4.1–4.2) and equation (16) of MacKinnon, Nielsen, and Webb (2021). Our results will continue to hold even if N and T diverge at different rates, but we make the homogeneous rate assumption for ease of exposition. On the other

hand, our results will not hold under fixed N or fixed T. Assumption 2 (ii) is the contrary case of no clustering. As we show below, our results hold under either condition. While Assumption 2 is sufficient for our results, it is probably stronger than necessary, but is used for its simplicity and tractability. Assumption 2 does rule out possible scenarios, including cases which lead to non-Gaussian limit distributions (e.g., Menzel, 2021, Example 1.7) – see our discussion in Section 7. The non-Gaussian cases can be characterized by daga generating processes that consist of degenerate additive factors of i, degenerate additive factors of t, and a small number of interactive factors between i and t (Chiang, Hansen, and Sasaki, 2022). With this said, it is legitimate to rule out non-Gaussian cases as our focus is on standard errors (as in the title of this article), which would not make sense under non-Gaussian limit distributions.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then,*

$$var(\widehat{\theta})^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, I_m).$$

A proof is provided in Appendix B. Theorem 2 shows that the self-normalized sample mean is asymptotically normal. Self-normalization is used to allow for differing rates of convergence due to clustering structure.

In this section, we focus on the setting where there is precisely one unit of observation per cluster intersection. In applications, there may be heterogeneous per-cluster numbers of observations, e.g., unbalanced panels. The above theory will straightforwardly extend to such cases – see Appendix G.

3.2 Linear Regression

We now revisit the linear panel regression model (2.1)–(2.2) and the OLS estimator (2.3). In this subsection we set $D_{it} = (Y_{it}, X_{it}', U_{it})'$, so that the framework (3.1) is

$$(Y_{it}, X_{it}', U_{it})' = f(\alpha_i, \gamma_t, \varepsilon_{it}). \tag{3.7}$$

Set $a_i = E[X_{it}U_{it} \mid \alpha_i]$ and $b_t = E[X_{it}U_{it} \mid \gamma_t]$. Let Σ_a and Σ_b be the variance/long-run variance matrices of a_i and b_t, respectively.

Assumption 3. For some $\delta > 0$ and $r > 1$, (i) $\{(Y_{it}, X_{it}', U_{it}) : 1 \le i \le N, 1 \le t \le T\}$ are generated following (3.7), where $\{\alpha_i\}$, $\{\gamma_t\}$, and $\{\varepsilon_{it}\}$ are mutually independent sequences, α_i is i.i.d across i, ε_{it} is i.i.d across (i,t), and γ_t is strictly stationary. (ii) $Q = E[X_{it}X_{it}'] > 0$, $E[\|X_{it}\|^{8(r+\delta)}] < \infty$, and $E[\|U_{it}\|^{8(r+\delta)}] < \infty$. (iii) γ_t is a β-mixing sequence with size $2r/(r-1)$, that is, $\beta(\ell) = O(\ell^{-\lambda})$ for a $\lambda > 2r/(r-1)$. (iv) One of the following two conditions hold: (1) Either $\Sigma_a > 0$ or $\Sigma_b > 0$, and $N/T \to c \in (0, \infty)$ as $(N,T) \to \infty$; or (2) (X_{it}, U_{it}) are independent and identically distributed across i and t, and $var(X_{it}U_{it}) > 0$. (v) For each $M \ge 1$ and $1 \le m \le M$, $w(m,M) = 1 - [m/(M+1)]$. (vi) $M/\min\{N,T\}^{1/2} = o(1)$.

Assumptions 3 (i)–(v) above are the counterparts of Assumptions 1 and 2, extended to the regression model. The moment and mixing conditions are standard in time series regression. Assumptions 3 (v)–(vi) are needed for consistent variance estimation. The α-mixing condition of Assumption 1 has been strengthened to β-mixing in Assumption 3. This is because our proof of consistent variance estimation relies on a deep fourth-order summability result due to Yoshihara (1976) which relies on β-mixing.

Under these assumptions, the asymptotic variance of $\widehat{\beta}$ is

$$V_{NT} = Q^{-1}\Omega_{NT}Q^{-1}$$

where Ω_{NT} is defined in (2.6)-(2.9). Our proposed estimator of V_{NT} is (2.10) with (2.12).

For our theory we focus on a vector-valued parameter $\theta = R'\beta$ for some $k \times m$ matrix R. This includes individual coefficients when $m = 1$. The estimator of θ is $\widehat{\theta} = R'\widehat{\beta}$, its asymptotic variance is $\Sigma_{NT} = R'V_{NT}R$, with estimator $\widehat{\Sigma}_{NT} = R'\widehat{V}_{NT}R$. For the case $m = 1$, a standard error for $\widehat{\theta}$ is $\widehat{\sigma}_{NT} = \sqrt{R'\widehat{V}_{NT}R}$. We now establish consistency of our variance estimator

Theorem 3. *If Assumption 3 holds for model* (2.1)–(2.2)*, then*

$$\Sigma_{NT}^{-1}\widehat{\Sigma}_{NT} \xrightarrow{p} I_k.$$

A proof is provided in Appendix C. It relies on some technical lemmas in Appendix F that are of potential independent interest. This result shows that our proposed variance estimator is consistent.

Notice that we state consistency as a self-normalized matrix ratio. This allow for the differing rates of convergence covered by Assumption 3.

Theorem 3 is new. It is the first demonstration of consistent variance estimation under two-way clustering with serially dependent time effects.

The proof of Theorem 3 includes some technical innovations. Of particular note is the use of the fourth-order summability condition of Yoshihara (1976), combined with projections on the individual-specific and time-specific factors. The summability condition is needed to calculate the variance of the variance estimator, which is a fourth-order sum.

Theorem 4. *If Assumption 3 holds for the model* (2.1)–(2.2)*, then*

$$\Sigma_{NT}^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, I_m) \tag{3.8}$$

and

$$\widehat{\Sigma}_{NT}^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, I_m). \tag{3.9}$$

A proof is provided in Appendix D.

Theorem 4 shows that the least squares estimator $\widehat{\theta}$ is asymptotically normal. Normality holds when the estimator is standardized by its asymptotic covariance matrix Σ_{NT} or by its estimator $\widehat{\Sigma}_{NT}$. This latter result shows (for the case $m = 1$) that t-ratios constructed with our standard errors are asymptotically $N(0, 1)$. It also shows (for the case $m > 1$) that Wald-type tests constructed with our covariance matrix estimator are asymptotically χ^2_m. Hence conventional inference methods can be used with the least squares estimator $\widehat{\beta}$, our variance estimator $\widehat{V}_{NT}$, and our standard errors.

Theorem 4 is new. It is the first result which rigorously demonstrates asymptotic normality of least squares estimators and t-ratios under two-way clustering with serially correlated time effects. The asymptotic normality presented here adapts to the unknown convergence rate. It is, however, worthy to note that this result is pointwise in DGP. In the absence of correlated time effects, Menzel (2021) discusses the issues of uniform inference. In a linear panel data context, Lu and Su (2022) provides uniformly valid inference procedure. Although it remains unclear to us whether their approaches can

be adapted to our framework, it is an interesting future research avenue to investigate the potential uniformity properties of the asymptotics under various sets of sequences of DGPs.

In contrast, test statistics constructed with the popular CGM variance estimator will not have conventional asymptotic distributions when the time effects γ_t are serially correlated. The CGM variance estimator is inconsistent in this situation, so test statistics will have distorted asymptotic distributions.

4 Fixed-Effect Models

For panel data, researchers often include one-way or two-way fixed effects. This section has two contents regarding fixed-effect models. First, Section 4.1 argues that two-way clustering is still necessary in general even if a researcher includes two-way fixed effects. Second, Section 4.2 extends our theory to two-way fixed-effect regressions.

4.1 Two-Way Clustering Is Still Necessary

Some empirical economists believe that it is unnecessary to cluster standard errors if fixed effects are included in estimation. In this section, we argue that fixed effects will not generally solve the problem of two-way cluster dependence.

For instance, consider the two-way fixed-effect model:

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{it} + U_{it}, \text{ where} \\ X_{it} &= \alpha_{i1}\gamma_{t2} + \alpha_{i2}\gamma_{t1} + \varepsilon_{it0} \\ U_{it} &= \alpha_{i0} + \gamma_{t0} + \alpha_{i1}\gamma_{t3} + \alpha_{i3}\gamma_{t1} + \varepsilon_{it1} \end{aligned} \tag{4.1}$$

Note that X_{it} and U_{it} are generated by the latent variables $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}, \alpha_{i3})$, $\gamma_t = (\gamma_{t0}, \gamma_{t1}, \gamma_{t2}, \gamma_{t3})$ and $\varepsilon_{it} = (\varepsilon_{it0}, \varepsilon_{it1})$. Here, α_{i0} and γ_{t0} are additive fixed effects. Suppose that α_{i0}, α_{i1}, α_{i2}, α_{i3}, γ_{t0}, γ_{t1}, γ_{t2}, γ_{t3}, ε_{it0}, and ε_{it1} are mutually independent with mean 0 and variance 1.

To abstract away from finite-sample issues, consider the *population* double differences[7]

$$\widetilde{X}_{it} = X_{it} - \mu_i^X - \mu_t^X + \mu^X \qquad \left(\text{where } \overline{\mu}_i^X = E[X_{it}|\alpha_i], \mu_t^X = E[X_{it}|\gamma_t], \,\&\, \mu^X = E[X_{it}]\right)$$

$$\widetilde{U}_{it} = U_{it} - \mu_i^U - \mu_t^U + \mu^U \qquad \left(\text{where } \overline{\mu}_i^U = E[U_{it}|\alpha_i],\ \mu_t^U = E[U_{it}|\gamma_t],\ \&\ \mu^U = E[U_{it}]\right)$$

Note that they reduce to

$$\widetilde{X}_{it} = \alpha_{i1}\gamma_{t2} + \alpha_{i2}\gamma_{t1} + \varepsilon_{it0}$$

$$\widetilde{U}_{it} = \alpha_{i1}\gamma_{t3} + \alpha_{i3}\gamma_{t1} + \varepsilon_{it1}$$

under the example (4.1). Certainly, the double differencing removes the FEs, α_{i0} and γ_{t0}, but still leaves the strong two-way dependence through $(\alpha_{i1}, \alpha_{i2}, \alpha_{i3})$ and $(\gamma_{t1}, \gamma_{t2}, \gamma_{t3})$. Observe that there is no endogeneity, as $E[\widetilde{U}_{it}|\widetilde{X}_{it}] = 0$. Furthermore, the score

$$\widetilde{X}_{it}\widetilde{U}_{it} = \alpha_{i1}^2\gamma_{t2}\gamma_{t3} + \alpha_{i1}\alpha_{i3}\gamma_{t1}\gamma_{t2} + \alpha_{i1}\alpha_{i2}\gamma_{t1}\gamma_{t3} + \alpha_{i2}\alpha_{i3}\gamma_{t1}^2 +$$
$$(\alpha_{i1}\gamma_{t2} + \alpha_{i2}\gamma_{t1})\varepsilon_{it1} + (\alpha_{i1}\gamma_{t3} + \alpha_{i3}\gamma_{t1})\varepsilon_{it0}$$

entails the non-degenerate projections

$$a_i = \alpha_{i2}\alpha_{i3} \qquad \text{and} \qquad b_t = \gamma_{t2}\gamma_{t3}$$

with respect to α_i and γ_t, respectively.

Hence, in this example, the components of α_i and γ_t are not eliminated by the two-way fixed effects regression of Y on X, so the score $\ddot{X}_{it}\ddot{U}_{it}$ is still two-way dependent, and two-way clustering is necessary for calculation of the covariance matrix. Furthermore, the score is not degenerate so our theory to be presented in Section 4.2 applies.

We would like to emphasize that we have used (4.1) as an illustrative example, and are not suggesting (4.1) as our assumed DGP. We used example (4.1) only to illustrate that, in general, two-way fixed-effect estimation does not eliminate two-way clustered dependence. In contrast, if (4.1) were known to be the true error structure, then a natural estimator would be interactive fixed effects (e.g.,

[7]For a formal account of the discrepancy between the *population* and *sample* double differences, see Section 4.2.

Bai, 2009), but that model is mis-specified within the general setting of two-way clustered dependence, and is hence not generally recommended.

4.2 Theory under Two-Way Fixed-Effect Models

This section shows that our standard errors extend to two-way fixed-effect models. In different settings, the existing literature has considered inference for two-way fixed-effect models. Verdier (2020) studies linear regression with two-way fixed effects for fixed T for sparsely matched data. Juodis (2021) considers bootstrap-based inference for linear models with two-way fixed effects under large N and T with a different set of assumptions.

Consider the two-way fixed-effect model

$$Y_{it} = X_{it}'\beta + \xi_i + \eta_t + U_{it}, \tag{4.2}$$

$$(X_{it}', \xi_i, \eta_t, U_{it})' = f(\alpha_i, \gamma_t, \varepsilon_{it}), \tag{4.3}$$

where ξ_i and η_t are fixed effects and $E[U_{it}] = 0$.[8] Define the within-transformed outcome and within-transformed regressors by

$$\begin{aligned}\ddot{Y}_{it} =& Y_{it} - \frac{1}{N}\sum_{i'=1}^{N} Y_{i't} - \frac{1}{T}\sum_{t'=1}^{T} Y_{it'} + \frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T} Y_{i't'}, \qquad \text{and} \\ \ddot{X}_{it} =& X_{it} - \frac{1}{N}\sum_{i'=1}^{N} X_{i't} - \frac{1}{T}\sum_{t'=1}^{T} X_{it'} + \frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T} X_{i't'},\end{aligned}$$

respectively. The within transformations induce complex dependence structure between transformed variables. Using idempotency of the within transformation matrices (see Ch. 17.8 of Hansen 2022), the two-way within estimator $\widehat{\beta}$ for β is defined as the OLS estimator of $\ddot{Y}_{it}$ on $\ddot{X}_{it}$ and thus satisfies

$$\widehat{\beta} - \beta = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{X}_{it}\ddot{X}_{it}'\right)^{-1} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{X}_{it}U_{it}.$$

(Note that $\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{X}_{it}\ddot{U}_{it} = \sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{X}_{it}U_{it}$ where $\ddot{U}_{it} = U_{it} - \frac{1}{N}\sum_{i'=1}^{N} U_{i't} - \frac{1}{T}\sum_{t'=1}^{T} U_{it'} + \frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T} U_{i't'}$.) Also, define the variance estimator $\widehat{\Sigma}_{NT}$ as in (2.10) with $(\ddot{Y}_{it}, \ddot{X}_{it}')$ in place

[8] Note that (4.3) implicitly requires that $\xi_i = f_2(\alpha_i)$ and $\eta_t = f_3(\gamma_t)$.

of (Y_{it}, X_{it}') and with the two-way within estimator $\widehat{\beta}$ defined in this section. Denote the population counterpart of the within-transformed regressor by

$$\widetilde{X}_{it} = X_{it} - E[X_{it}|\gamma_t] - E[X_{it}|\alpha_i] + E[X_{it}].$$

Note that $(\widetilde{X}_{it}', U_{it})$ only depends on α_i, γ_t, and ε_{it} and satisfies $E[\widetilde{X}_{it}] = 0$.

Theorem 5 (Regression models with two-way fixed effects)**.** *Suppose Assumption 3 (ii), (iii), (iv)(1), (v), (vi) holds for (X_{it}', U_{it}), and the outcome variable Y_{it} is generated following* (4.2)–(4.3) *where α_i, γ_t, and ε_{it} are defined in the same way as in Assumption 3. In addition, assume that $E[\widetilde{X}_{it}U_{it}] = 0$ and $\|X_{it}\|_\infty \leq K$ for a constant K that is independent of N and T. Then, the conclusions in Theorems 3 and 4 continue to hold, and thus*

$$\widehat{\Sigma}_{NT}^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, I_m).$$

A proof is provided in Appendix E. It is worthy noting that the proof is not a mere extension of the previous theorems, but requires nontrivial technicalities involving maximal inequalities in the time-series framework.

5 Simulations: Comparisons of Alternative Standard Errors

In this section, we use simulated data to examine the performance of our proposed robust variance estimator in comparison with six existing alternatives.

We generate data based on the linear model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + U_{it},$$

where the right-hand side variables $(X_{it}, U_{it})'$ are generated through the panel dependence structure

$$X_{it} = w_\alpha \alpha_i^x + w_\gamma \gamma_t^x + w_\varepsilon \varepsilon_{it}^x \qquad \text{and}$$
$$U_{it} = w_\alpha \alpha_i^u + w_\gamma \gamma_t^u + w_\varepsilon \varepsilon_{it}^u.$$

We set $(\beta_0, \beta_1) = (0.1, 0.1)$ throughout. For the weight parameters, we use $(w_\alpha, w_\gamma, w_\varepsilon) = (0.0, 0.0, 0.5)$ to generate i.i.d. data and also use $(w_\alpha, w_\gamma, w_\varepsilon) = (0.15, 0.20, 0.15)$ to generate dependent data. The latent components $(\alpha_i^x, \alpha_i^u, \varepsilon_{it}^x, \varepsilon_{it}^u)$ are all mutually independent $N(0,1)$.

The latent common time effects (γ_t^x, γ_t^u) are dynamically generated according to the AR(1) design:

$$\gamma_t^x = \rho\gamma_{t-1}^x + \widetilde{\gamma}_t^x \text{ where } \widetilde{\gamma}_t^x \text{ are independent draws from } N(0, 1-\rho^2); \text{ and}$$
$$\gamma_t^u = \rho\gamma_{t-1}^u + \widetilde{\gamma}_t^u \text{ where } \widetilde{\gamma}_t^u \text{ are independent draws from } N(0, 1-\rho^2).$$

The initial values are drawn from $N(0,1)$. We vary the AR parameter $\rho \in \{0.25, 0.50, 0.75\}$ across sets of simulations.

For each realization of observed data $\{(Y_{it}, X_{it}) : 1 \le i \le N, 1 \le t \le T\}$ constructed according to the data generating process described above, we estimate (β_0, β_1) by OLS. Our objective is to evaluate the performance of our proposed robust variance estimator $\widehat{V}_{NT} = \widehat{Q}^{-1}\widehat{\Omega}_{NT}\widehat{Q}^{-1}$, where $\widehat{Q}$ and $\widehat{\Omega}_{NT}$ are given in (2.5) and (2.12), and the tuning parameter $\widehat{M}$ is chosen according to the rule (2.13). Through simulation studies, we examine the performance of this robust variance estimator (hereafter referred to as CHS; Chiang-Hansen-Sasaki) in comparison with six existing alternative variance estimators which are in popular use for panel data analysis. They include the heteroskedasticity robust estimator (EHW; Eicker-Huber-White; also known as HC0), the cluster robust estimator within i (CRi; which corresponds to the Liang-Zeger-Arellano estimator), the cluster robust estimator within t (CRt), the two-way cluster robust estimator (CGM; Cameron, Gelbach, and Miller, 2011), the wild bootstrap estimator (MNW; MacKinnon, Nielsen, and Webb, 2021) for CGM, the bootstrap estimator (M; Menzel, 2021),[9] and the two-way cluster robust estimator with 2-dependence (T; Thompson, 2011).

Table 1 reports simulation results. Reported values are the coverage frequencies for the slope parameter β_1 for the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and

[9]Implementation of M requires some tuning parameters. We set the number of bootstrap iterations and the model selection tuning parameters, κ_a and κ_g following the simulation code for regressions by Menzel (2021). In addition to the default method of M, we also ran M without its model selection feature to find its results the same as those of the default method. Hence, we only report results by the default method of M.

I.I.D. Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	MNW	M	T	CHS
(I)	50	100	—	0.947	0.939	0.942	0.933	0.947	0.999	0.913	0.927
(II)	75	75	—	0.951	0.945	0.947	0.940	0.949	0.999	0.915	0.933
(III)	100	50	—	0.953	0.950	0.945	0.940	0.953	0.999	0.897	0.931

Dependence Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	MNW	M	T	CHS
(IV)	50	100	0.25	0.337	0.740	0.847	0.925	0.944	0.935	0.922	0.929
(V)	75	75	0.25	0.310	0.653	0.882	0.926	0.940	0.936	0.919	0.929
(VI)	100	50	0.25	0.287	0.536	0.885	0.912	0.931	0.914	0.890	0.912
(VII)	50	100	0.50	0.302	0.685	0.786	0.888	0.910	0.898	0.919	0.917
(VIII)	75	75	0.50	0.273	0.590	0.823	0.878	0.897	0.892	0.911	0.910
(IX)	100	50	0.50	0.246	0.471	0.821	0.857	0.885	0.862	0.880	0.887
(X)	50	100	0.75	0.225	0.579	0.639	0.770	0.801	0.779	0.879	0.877
(XI)	75	75	0.75	0.207	0.493	0.664	0.749	0.776	0.770	0.872	0.863
(XII)	100	50	0.75	0.187	0.398	0.660	0.715	0.748	0.718	0.836	0.823

Table 1: Coverage probabilities for the slope parameter β_1 for the OLS with the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show results under the i.i.d. and dependence designs, respectively. The sample size is indicated by (N, T). The parameter ρ indicates the AR coefficient in the dependence design. EHW stands for Eicker–Huber–White, CRi stands for cluster robust within i, CRt stands for cluster robust within t, CGM stands for Cameron-Gelbach-Miller, MNW stands for MacKinnon-Nielsen-Webb, M stands for Menzel, T stands for Thompson, and CHS stands for Chiang-Hansen-Sasaki.

bottom panels show coverage probability results under the i.i.d. design and the dependence design, respectively. In each group of three consecutive rows, the panel sample sizes (N, T) vary by rows. Cells are shaded based on the proximity of the simulated coverage probability to the nominal probability of 0.95; the darker shades indicate more correct coverage.

We observe the following four points in these results. First, under the i.i.d. design (rows (I)–(III)), EHW, CRi, CRt, CGM, MNW and CHS produce accurate coverage probabilities. On the other hand, M yields over-coverage consistently across different sample sizes, and T yields under-coverage especially under small T. Second, EHW and CRi tend to behave poorly in general when there are two ways of cluster dependence as in rows (IV)–(XII). Third, CRt also tends to behave poorly under larger extents of serial dependence as in rows (X)–(XII). Likewise, CGM, MNW, and M perform less preferably as the serial dependence becomes stronger. These results are consistent with the fact that these methods do not account for serially correlated common time effects. Fourth, in contrast, T and CHS behave more robustly under stronger serial dependence especially when T is large. Whenever T and ρ are smaller, as in rows (III) and (VI), T incurs under-coverage and hence CHS outperforms T in general. The last observation that T performs poorly for small sample sizes is consistent with the similar observations made by Thompson (2011, page 7) in his Monte Carlo simulation studies. In summary, we demonstrate that confidence intervals constructed with our proposed standard errors lead to robustly superior coverage performance relative to the existing methods.

We ran additional simulations beyond those presented in this section. Their results are found in Appendix I. Specifically, Appendix I.1 illustrates power analyses, Appendix I.2 presents simulations for the two-way fixed-effect estimator, and Appendix I.3 allow for multiple observations per (i, t) intersection.

6 An Empirical Application

In this section, we highlight differences across the alternative standard error estimates for estimates of a simple asset pricing model. Consider the Fama-French three-factor model

$$R_{it} - R_{ft} = \beta_1(R_{Mt} - R_{ft}) + \beta_2 SMB_t + \beta_3 HML_t + e_{it},$$

where R_{it} is the total return of portfolio/stock i in month t, R_{ft} is the risk-free rate of return in month t, R_{Mt} is the total market portfolio return in month t, SMB_t is the size premium (small$-$big), HML_t is the value premium (high$-$low), and $\beta = (\beta_1, \beta_2, \beta_3)'$ are the factor coefficients.

We use two data sets of portfolio/stock returns. They are (A) 44 industry portfolios excluding four financial sectors (banking, insurance, real estate, and trading) and (B) individual stocks. For each of these data sets, (A) and (B), we use the monthly panel of length 120 from January 2000 to December 2009. For the individual stock data set (B), we use the balanced portion of the panel data, consisting of $N = 779$ stocks. The risk-free rate is based on the monthly 30-day T-bill beginning-of-month yield. See Appendix H.3 for the source of data.

Let $\ddot{Y}_{it}$ and $\ddot{X}_{it}$ denote the within-transformations[10] of $R_{it} - R_{ft}$ and $(R_{Mt} - R_{ft}, SMB_t, HML_t)'$, respectively. ($\ddot{X}_{it}$ is homogeneous in the cross section.) We estimate β by the within-estimator

$$\widehat{\beta} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it} \ddot{X}_{it}' \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it} \ddot{Y}_{it}$$

to remove any additive portfolio/stock fixed effects. Thus, our proposed standard errors are computed based on $\widehat{V}_{NT} = \widehat{Q}^{-1} \widehat{\Omega}_{NT} \widehat{Q}^{-1}$ with $\widehat{Q}$ and $\widehat{\Omega}_{NT}$ from (2.5) and (2.12) with Y_{it}, X_{it} and $\widehat{U}_{it}$ replaced by $\ddot{Y}_{it}$, $\ddot{X}_{it}$ and $\ddot{Y}_{it} - \ddot{X}_{it}\widehat{\beta}$, respectively. Table 2 summarizes estimates $\widehat{\beta}$ of β along with alternative standard error estimates of them for each of the two data sets described above.

For each row of Table 2, the standard error due to the Eicker-Huber-White (EHW) estimator is smaller than any other standard error. On the other hand, the two-way cluster robust estimators

[10] Technically, our theory does not allow for these demeaned variables as they are (N, T)-dependent. But we expect that there should be no change to the theory if we allow for array data. Formalizing this would be a useful future research direction.

(A) 44 Industry Portfolios. $(N, T) = (44, 119)$.

		Standard Errors						
	$\widehat{\beta}$	EHW	CRi	CRt	CGM	M	T	CHS
MKT	0.959	0.022	0.055	0.030	0.059	0.021	0.058	0.059
SMB	0.076	0.029	0.035	0.041	0.045	0.029	0.055	0.053
HML	0.358	0.030	0.066	0.049	0.076	0.028	0.082	0.080

(B) Individual Stocks. $(N, T) = (779, 119)$.

		Standard Errors						
	$\widehat{\beta}$	EHW	CRi	CRt	CGM	M	T	CHS
MKT	1.157	0.012	0.021	0.033	0.037	0.035	0.037	0.033
SMB	0.474	0.020	0.025	0.051	0.053	0.055	0.067	0.071
HML	0.173	0.016	0.029	0.053	0.058	0.054	0.053	0.052

Table 2: Estimates of the factor coefficients with six alternative standard error estimates of them. EHW stands for Eicker–Huber–White, CRi stands for cluster robust within i, CRt stands for cluster robust within t, CGM stands for Cameron-Gelbach-Miller, M stands for Menzel, T stands for Thompson, and CHS stands for Chiang-Hansen-Sasaki.

of Cameron-Gelbach-Miller (CGM) and Thompson (T) and our proposed estimator (CHS) tend to yield the largest standard errors in each row. The remaining three standard errors stay in the middle between these two groups. The estimator by Menzel (M) behaves similarly to EHW in panel (A) while it behaves similarly to CGM in panel (B). This puzzling outcome arises from the model selection feature of M. In fact, M would also behave similarly to CGM in panel (A) as well if the tuning parameter of M were chosen to take a much smaller value.[11] Because of these idiosyncratic behaviors of M that

[11] As in the simulation section, we set the number of bootstrap iterations and the model selection tuning parameters for M following the simulation code for regressions by Menzel (2021).

depend on discrete outcomes of model selection which in turn depend on tuning parameters, we will hereafter focus on the other estimators in comparing the results.

Observe in panel (A) that the statistical significance of the coefficient β_2 of SMB meaningfully diminishes as the standard error estimator becomes more robust (again, except for M). Specifically, it is significant at the 5% level with EHW, CRi and M, but it becomes insignificant at this level with CRt, CGM, T, or CHS. Furthermore, while it is significant at the 10% level with EHW, CRi, CRt, and CGM, it becomes insignificant at this level with T and CHS. This part of the estimation results shows a case where accounting for serial correlation in common time effects may even overturn conclusions from statistical inference based on the other standard errors. Accounting for arbitrary untruncated time correlation, however, CHS yields a slightly smaller standard error estimate than T for this case.

Finally, we remark that the standard errors of MNW are not displayed here, because their method uses existing standard errors such as CGM and their bootstrap is used to compute the critical values for the t-statistic based on them.

7 Summary and Discussions

In this paper, we propose new robust standard error estimators for panel data. The new estimators account for the cluster dependence within i, the cluster dependence within t, and serial dependence in the common time effects. In particular, all the existing robust standard error estimators fail to accommodate untruncated serial dependence in the common time effects, while this feature is relevant to empirical data used in economics and finance. Simulation studies show that the new standard errors produce robustly superior coverage performance than existing alternatives, including the heteroskedasticity robust estimator (Eicker-Huber-White; also known as HC0), the cluster robust estimator within i, the cluster robust estimator within t, the two-way cluster robust estimator (Cameron, Gelbach, and Miller, 2011), the bootstrap estimator (Menzel, 2021), and the two-way cluster robust estimator with 2-dependence (Thompson, 2011).

In the rest of this section, we discuss limitations of our method and potentials of future research

in relation to the existing literature. Since the seminal work by Cameron, Gelbach, and Miller (2011) and Thompson (2011), a few important papers have proposed methods of robust inference in two way cluster dependence.

Davezies, D'Haultfœuille, and Guyonvarch (2021) derive Donsker results under multiway cluster dependence. In the current paper, we only derive limit distributions for finite-dimensional parameters that are relevant to many empirical applications in economics and finance. In other words, Davezies, D'Haultfœuille, and Guyonvarch (2021) provide a generalization of existing results by allowing for empirical processes, we on the other hand provide a generalization in a different direction by allowing for serial dependence in common time effects. Combining these two directions of generalization is left for future research.

Menzel (2021) proposes a method of (conservative) inference with uniform validity over a large class of distributions including the case of non-Gaussian degeneracy under two-way cluster dependence. In the current paper, for the purpose of providing a simple method of inference via analytic standard error formulas, we focus on the case of Gaussian degeneracy as well as non-degenerate cases. In other words, Menzel (2021) provides a generalization of existing results by allowing for uniformity over a large class, we on the other hand provide a generalization in a different direction by allowing for serial dependence in common time effects. Combining these two directions of generalization is also left for future research.

Chiang, Kato, and Sasaki (2023) derive a high-dimensional central limit theorem under multiway cluster dependence. In the current paper, we only consider finite-dimensional parameters that are relevant to many empirical applications in economics and finance. In other words, Chiang, Kato, and Sasaki (2023) provide a generalization of existing results by allowing for high dimensionality, we on the other hand provide a generalization in a different direction by allowing for serial dependence in common time effects. Again, combining these two directions of generalization is left for future research.

The independence conditions in Assumption 1 (i) may be relaxed in a couple of directions. One

way is relax the i.i.d. assumption on the ε_{it} factor in (3.1). With this said, the existing literature explicitly or implicitly makes this assumption, and we continue to focus on i.i.d. ε_{it}. Another way is to relax the i.i.d. assumption on the α_i factor in (3.1). For instance, if a researcher obtains spatial information associated with panel data, then it may be a possibility to allow for spatial α-mixing (e.g., Jenish and Prucha, 2009). We leave these extensions for future research.

Our standard errors allowing for serial correlation in time effects effectively use the Newey-West-type long-run variance estimation, but it is well known that in some cases a relatively long time series may be required for such estimators to perform well (e.g., Lazarus, Lewis, Stock, and Watson, 2018). Inference based on moving-block bootstrap (e.g., Gonçalves, 2011) may improve the finite-sample performance, and we suggest it as another direction for future research. Another promising recent proposal by Chen and Vogelsang (2023) is to use fixed-b asymptotic theory to derive bias corrections and improve the distributional approximation.

Appendix

A Proof of Theorem 1

Proof. Following the argument in the text, without loss of generality set $\theta = 0$, and make the decompositions (3.2) and (3.3). Since the sums in (3.3) are uncorrelated, we find

$$var(\widehat{\theta}) = var\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right) + var\left(\frac{1}{T}\sum_{t=1}^{T} b_t\right) + var\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\right). \tag{A.1}$$

We take each term separately.

First, recall that a_i is i.i.d., mean zero, and has variance matrix Σ_a. Since $a_i = E[X_{it} \mid \alpha_i]$, an application of the conditional Jensen inequality shows that

$$||\Sigma_a|| \leq E[||a_i||^2] = E[||E[X_{it} \mid \alpha_i]||^2] \leq E[E[||X_{it}||^2 \mid \alpha_i]] = E[||X_{it}||^2],$$

under Assumption 1. Hence $||\Sigma_a|| < \infty$ as claimed. We calculate that

$$var\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right) = \frac{1}{N}\Sigma_a. \tag{A.2}$$

Second, by a standard variance decomposition for stationary time series,

$$var\left(\frac{1}{T}\sum_{t=1}^{T} b_t\right) = \frac{1}{T}\sum_{\ell=-(T-1)}^{T-1}\left(1-\frac{|\ell|}{T}\right)E[b_t b_{t+\ell}'] \\ = \frac{1}{T}\Sigma_b\left(1+o(1)\right). \tag{A.3}$$

The second equality holds if the sum (3.5) converges, which we now demonstrate. Since $b_t = E[X_{it} \mid \gamma_t]$, an application of the conditional Jensen inequality shows that $E[||b_t||^s] = E[||E[X_{it} \mid \gamma_t]||^s] \leq E[E[||X_{it}||^s \mid \gamma_t]] = E[||X_{it}||^s]$ for $s \geq 1$, and in particular $E[||b_t||^{4(r+\delta)}] < \infty$. Also, as b_t is a function only of γ_t, it has the same mixing coefficients. By an application of Theorem 14.13 (ii) in Hansen (2022), we find

$$||\Sigma_b|| = \left|\left|\sum_{\ell=-\infty}^{\infty} E[b_t b_{t+\ell}']\right|\right| \leq 8\left(E||b_t||^{4(r+\delta)}\right)^{1/2(r+\delta)}\sum_{\ell=-\infty}^{\infty}\alpha(\ell)^{1-1/2(r+\delta)} < \infty$$

under Assumption 1. Hence $||\Sigma_b|| < \infty$ as claimed, and the bound (A.3) follows.

Third, the law of iterated expectations implies that for $j \neq i$

$$E[e_{it}e_{js}'] = E[E[e_{it}e_{js}' \mid \gamma_t, \gamma_s]] = E[E[e_{it} \mid \gamma_t]E[e_{js}' \mid \gamma_s]] = 0$$

the second equality since conditional on (γ_t, γ_s), e_{it} is independent of e_{js} for $j \neq i$. Combined with the stationarity of e_{it}, this implies that

$$var\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\right) = \frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E[e_{it}e_{is}'] \\ = \frac{1}{NT}\sum_{\ell=-(T-1)}^{T-1}\left(1-\frac{|\ell|}{T}\right)E[e_{it}e_{i,t+\ell}'] \\ = \frac{1}{NT}\Sigma_e\left(1+o(1)\right). \tag{A.4}$$

The final equality holds if the sum (3.6) converges, which we now demonstrate. Conditional on α_i, e_{it} is an α-mixing process with the same mixing coefficients as γ_t. Furthermore, the moments of e_{it} are bounded by those of X_{it}. Using iterated expectations, Jensen's inequality, the fact that $E[e_{it} \mid \alpha_i] = 0$, Theorem 14.13 (ii) in Hansen (2022), and again Jensen's inequality,

$$||E[e_{it}e_{i,t+\ell}']|| \leq E||E[e_{it}e_{i,t+\ell}' \mid \alpha_i]|| \leq 8\left(E||e_{it}||^{4(r+\delta)}\right)^{1/2(r+\delta)}\alpha(\ell)^{1-1/2(r+\delta)}.$$

This (under Assumption 1) shows that

$$||\Sigma_e|| = \left|\left| \sum_{\ell=-\infty}^{\infty} E[e_{it} e'_{i,t+\ell}] \right|\right| \leq 8 \left(E||e_{it}||^{4(r+\delta)} \right)^{1/2(r+\delta)} \sum_{\ell=-\infty}^{\infty} \alpha(\ell)^{1-1/2(r+\delta)} < \infty.$$

Thus $||\Sigma_e|| < \infty$ and the bound (A.4) follows.

Together, (A.1)-(A.4) establish that

$$var(\widehat{\theta}) = \frac{1}{N}\Sigma_a + \frac{1}{T}\Sigma_b \left(1 + o(1)\right) + \frac{1}{NT}\Sigma_e \left(1 + o(1)\right)$$

as claimed.

Since $||\Sigma_a|| < \infty$, $||\Sigma_b|| < \infty$, and $||\Sigma_e|| < \infty$, it follows that $var(\widehat{\theta}) \to 0$ as $N, T \to \infty$. By Chebyshev's inequality, we deduce that $\widehat{\theta} \xrightarrow{p} \theta$, completing the proof. □

B Proof of Theorem 2

Proof. Without loss of generality, set $\theta = 0$. The proof is different under Assumption 2 parts (i) and (ii). First, take case (ii). If X_{it} is i.i.d. then $\sqrt{NT}\widehat{\theta} \xrightarrow{d} N(0, var(X_{it}))$ by Lyapunov's central limit theorem. Also, $var(\widehat{\theta}) = var(X_{it})/NT$. Combining, we find the stated result. Hence, for the remainder of the proof we focus on case (i).

Using (3.2),

$$\sqrt{N}\widehat{\theta} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i + \sqrt{\frac{N}{T}} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} b_t + \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} e_{it}. \tag{B.1}$$

The first term in (B.1) consists of a sum of the i.i.d. zero-mean random vectors a_i with finite variance Σ_a. By Lyapunov's central limit theorem, we deduce

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \xrightarrow{d} N(0, \Sigma_a). \tag{B.2}$$

The second term in (B.1) consists of the α-mixing sequence b_t (see Theorem 14.12 in Hansen (2022)). Applying the central limit theorem for α-mixing sequences (cf. Hansen, 2022, Theorem 14.15) under Assumption 1(iii), which implies $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-1/2(r+\delta)} < \infty$, we deduce

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} b_t \xrightarrow{d} N(0, \Sigma_b). \tag{B.3}$$

The asymptotic variance Σ_b was shown finite in Theorem 1. The asymptotic distributions in (B.2) and (B.3) are independent since the sequences a_i and b_t are independent.

Equation (A.4) shows that the variance of the third term in (B.1) is $O(T^{-1})$, and hence this term is $o_p(1)$. Together, (B.1)-(B.3) plus $N/T \to c$ show that

$$\sqrt{N}\widehat{\theta} \xrightarrow{d} N(0, \Sigma_a) + \sqrt{c}N(0, \Sigma_b) = N(0, \Sigma), \tag{B.4}$$

where $\Sigma \equiv \Sigma_a + c\Sigma_b$. Assumption 2(i) implies that $\Sigma > 0$.

Theorem 1 and $N/T \to c$ establish that

$$N \cdot var(\widehat{\theta}) = \Sigma_a + \frac{N}{T}\Sigma_b(1 + o(1)) + \frac{1}{T}\Sigma_e(1 + o(1)) \to \Sigma.$$

Together,

$$\Big(var(\widehat{\theta})\Big)^{-1/2}\widehat{\theta} = \Big(N \cdot var(\widehat{\theta})\Big)^{-1/2}\sqrt{N}\widehat{\theta} \xrightarrow{d} \Sigma^{-1/2}N(0, \Sigma) = N(0, I_m).$$

This is the stated result. □

C Proof of Theorem 3

Proof. The proof branches into the two cases, (1) and (2), of Assumption 3 (iv). For readibility, we defer some of the lengthy technical calculations to Lemmas 1-3 in Appendix F.

First, consider the case where Assumption 3 (iv) (1) holds. From Theorem 1, we have

$$N\Omega_{NT} = \Sigma_a + \frac{N}{T}\Sigma_b + o(1) \to \Sigma > 0,$$

where $\Sigma = \Sigma_a + c\Sigma_b$. Thus

$$N\Sigma_{NT} = NR'Q^{-1}\Omega_{NT}Q^{-1}R \to R'Q^{-1}\Sigma Q^{-1}R > 0. \tag{C.1}$$

Next, by combining Lemmas 1 and 2 from Appendix F, we have

$$N\widehat{\Omega}_{NT} = \underbrace{\frac{1}{NT^2}\sum_{i=1}^{N}\widehat{R}_i\widehat{R}_i'}_{=E[a_ia_i']+o_p(1)} + \underbrace{\frac{1}{NT^2}\sum_{t=1}^{T}\widehat{S}_t\widehat{S}_t'}_{=cE[b_tb_t']+o_p(1)} - \underbrace{\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'\widehat{U}_{it}^2}_{=O_p(T^{-1})}$$

$$+ \underbrace{\frac{1}{NT^2}\sum_{m=1}^{M} w(m,M)\left(\widehat{G}_m + \widehat{G}_m'\right)}_{=c\sum_{m=-\infty}^{\infty} E[b_t b_{t+m}'] + cE[b_t b_t'] + o_p(1)}$$

$$\xrightarrow{p} E[a_i a_i'] + cE[b_t b_t'] + c\sum_{m=-\infty}^{\infty} E[b_t b_{t+m}'] + cE[b_t b_t']$$

$$= \Sigma.$$

Now, consider $\widehat{Q}$. Setting $\theta = \text{vec}(X_{it}X_{it}')$ and $\widehat{\theta} = \text{vec}(\widehat{Q})$, which satisfy the conditions of Theorem 2 under Assumption 3. By the proof of Theorem 2 and the continuous mapping theorem, we obtain $||\widehat{Q}^{-1} - Q^{-1}|| = o_p(1)$ and $\widehat{Q}^{-1} = O_p(1)$ by Assumption 3 (ii). Together we have established that

$$N\widehat{\Sigma}_{NT} = NR'\widehat{Q}^{-1}\widehat{\Omega}_{NT}\widehat{Q}^{-1}R \xrightarrow{p} R'Q^{-1}\Sigma Q^{-1}R. \tag{C.2}$$

Equations (C.1) and (C.2) together imply that

$$\Sigma_{NT}^{-1}\widehat{\Sigma}_{NT} \xrightarrow{p} \left(R'Q^{-1}\Sigma Q^{-1}R\right)^{-1}\left(R'Q^{-1}\Sigma Q^{-1}R\right) = I_k.$$

This is the stated result.

Second, consider the case where Assumption 3 (iv) (2) holds. Observe that $\Sigma = var(X_{it}U_{it}) > 0$ under Assumption 3 (iv) (2). Then $NT\Omega_{NT} = \Sigma > 0$ and

$$NT\Sigma_{NT} = NTR'Q^{-1}\Omega_{NT}Q^{-1}R = R'Q^{-1}\Sigma Q^{-1}R > 0. \tag{C.3}$$

Lemma 3 in Appendix F implies $NT\widehat{\Omega}_{NT} \xrightarrow{p} NT\Omega_{NT} = \Sigma$. The law of large number for i.i.d. random variables, Assumption 3(ii), and the continuous mapping theorem imply that $||\widehat{Q}^{-1} - Q^{-1}|| = o_p(1)$. Together, this implies that

$$NT\widehat{\Sigma}_{NT} = NTR'\widehat{Q}^{-1}\widehat{\Omega}_{NT}\widehat{Q}^{-1}R \xrightarrow{p} R'Q^{-1}\Sigma Q^{-1}R. \tag{C.4}$$

Equations (C.3) and (C.4) together imply that

$$\Sigma_{NT}^{-1}\widehat{\Sigma}_{NT} \xrightarrow{p} \left(R'Q^{-1}\Sigma Q^{-1}R\right)^{-1}\left(R'Q^{-1}\Sigma Q^{-1}R\right) = I_k.$$

This is the stated result and completes the proof. □

D Proof of Theorem 4

Proof. Assumption 3 (iv) imposes either non-singular clustered dependence (1) or i.i.d. dependence (2). Under the latter the result is classical; hence we focus on condition (1). Some algebra reveals that

$$\begin{aligned}\sqrt{N}(\widehat{\theta}-\theta) &= \sqrt{N}R'\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}U_{it}\right)\\ &= R'Q^{-1}\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}U_{it} - R'\widehat{Q}^{-1}\left(\widehat{Q}-Q\right)Q^{-1}\left(\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}U_{it}\right). \qquad \text{(D.1)}\end{aligned}$$

The first term in (D.1) is a self-normalized sample mean in the random vectors $X_{it}U_{it}$, which satisfy the conditions of Theorem 2. Consequently, for $\Sigma = \Sigma_a + c\Sigma_b$, the first term in (D.1) satisfies

$$R'Q^{-1}\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}U_{it} \xrightarrow{d} R'Q^{-1}N(0,\Sigma) = N(0, R'Q^{-1}\Sigma Q^{-1}R),$$

as shown in (B.4).

Recall that $\widehat{Q}$ is the sample average of the variables $X_{it}X_{it}'$, which satisfy the conditions of Theorem 2. It follows that $||\widehat{Q}-Q|| = o_p(1)$ and $\widehat{Q}^{-1} = O_p(1)$. Consequently, the second term in (D.1) is $o_p(1)$. Together, we deduce that

$$\sqrt{N}(\widehat{\theta}-\theta) \xrightarrow{d} N(0, R'Q^{-1}\Sigma Q^{-1}R). \qquad \text{(D.2)}$$

From Theorem 1,

$$N\Omega_{NT} = \Sigma_a + \frac{N}{T}\Sigma_b + o(1) \to \Sigma.$$

Thus we find that

$$N\Sigma_{NT} \to R'Q^{-1}N\Omega_{NT}Q^{-1}R. \qquad \text{(D.3)}$$

Together, (D.2) and (D.3) imply that

$$\begin{aligned}\Sigma_{NT}^{-1/2}(\widehat{\theta}-\theta) &= (N\Sigma_{NT})^{-1/2}\sqrt{N}(\widehat{\theta}-\theta)\\ &\xrightarrow{d} (R'Q^{-1}N\Omega_{NT}Q^{-1}R)^{-1/2}N(0, R'Q^{-1}\Sigma Q^{-1}R)\end{aligned}$$

$$= N(0, I_m).$$

This is (3.8).

Equation (3.9) follows by combining (3.8) with Theorem 3. □

E Proof of Theorem 5

Proof. We first consider the case of non-degeneracy. Since $X_{it} - E[X_{it}|\gamma_t]$ is independent across i conditionally on $(\gamma_t)_{t=1}^T$ and $\|X_{it}\|_\infty \leq K$, Theorem 2.14.1 in van der Vaart and Wellner (1996) yields

$$E\left[\max_{t=1,...,T}\left|\frac{1}{N}\sum_{i'=1}^{N} X_{it} - E[X_{it}|\gamma_t]\right| |(\gamma_t)_{t=1}^T\right] \lesssim \sqrt{\frac{K^2 \log T}{N}}.$$

Integrating out both sides using Fubini's theorem, we obtain

$$E\left[\max_{t=1,...,T}\left|\frac{1}{N}\sum_{i'=1}^{N} X_{it} - E[X_{it}|\gamma_t]\right|\right] \lesssim \sqrt{\frac{K^2 \log T}{N}} = o(1). \tag{E.1}$$

We are now going to show

$$E\left[\max_{i=1,...,N}\left|\frac{1}{T}\sum_{t'=1}^{T} X_{it'} - E[X_{it}|\alpha_i]\right|\right] = o(1). \tag{E.2}$$

We use Bernstein's big-block-small-block argument (for example, see Step 1 in the Proof of Theorem E.1 in Chernozhukov, Chetverikov, and Kato 2019) to show (E.2). Specifically, let $q = q_T \sim T^{3/4}$, $s = s_T \sim T^{1/3}$, and $m = T/(q+s)$ be positive sequences of integers satisfying $q + s \leq T/2$. It immediately follows that $q, s \to \infty$, $q = o(T)$, $s^2/T = o(1)$, $q^{-1}s^2 \log N = o(1)$, and $m \sim T^{1/4}$. Define $I_1 = \{1, ..., q\}$, $J_1 = \{q+1, ..., q+s\}$, $...I_m = \{(m-1)(q+s)+1, ..., m(q+s)\}$, $J_m = \{(m-1)(q+s)+q+1, ..., m(q+s)\}$, $J_{m+1} = \{m(q+s)+1, ..., T\}$. As $\lambda > 2r/(r-1) > 2$, we have

$$m\beta(s) = mO(s^{-\lambda}) = o(1),$$

where $\beta(\cdot)$ is the β-mixing coefficient. The integers, q and s, will serve as the lengths of big and small blocks, respectively, and m is the number of big blocks. Now, for each $i = 1, ..., N$, one has the decomposition

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{X_{it} - E[X_{it}|\alpha_i]\} = \frac{1}{\sqrt{T}}\sum_{l=1}^{m} L_{il} + \frac{1}{\sqrt{T}}\sum_{l=1}^{m} S_{il} + \frac{1}{\sqrt{T}} S_{i,(m+1)}, \tag{E.3}$$

$$\text{where } L_{il} = \sum_{t \in I_l} \{X_{it} - E[X_{it}|\alpha_i]\}, \quad S_{il} = \sum_{t \in J_l} \{X_{it} - E[X_{it}|\alpha_i]\}.$$

L_{il} (respectively, S_{il}) equals the sum over a big block (respectively, small block). Define $(\widehat{L}_{il})_{l=1}^m$ and $(\widehat{S}_{il})_{l=1}^{m+1}$ to be the decoupled copies of $(L_{il})_{l=1}^m$ and $(S_{il})_{l=1}^{m+1}$, respectively. That is, they are two independent sequences of random vectors such that

$$\widehat{L}_{il} \overset{d}{=} L_{il} \text{ for } l \in \{1, \ldots, m\} \quad \text{and} \quad \widehat{S}_{il} \overset{d}{=} S_{il} \text{ for } l \in \{1, \ldots, m+1\}$$

conditionally on $(\alpha_i)_i$. We now claim that, for any $y \in \mathbb{R}$, it holds that

$$\begin{aligned}
&P\left(\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{l=1}^m \widehat{L}_{il} \le y - o(1) \mid (\alpha_i)_i\right) - o(1) \\
\le &P\left(\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{t=1}^T \{X_{it} - E[X_{it}|\alpha_i]\} \le y \mid (\alpha_i)_i\right) \\
\le &P\left(\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{l=1}^m \widehat{L}_{il} \le y + o(1) \mid (\alpha_i)_i\right) + o(1). \qquad \text{(E.4)}
\end{aligned}$$

We will prove only the second inequality as the first one follows from a mirrored argument. By (E.3), we have

$$\begin{aligned}
&\max_i \frac{1}{\sqrt{T}} \sum_{t=1}^T \{X_{it} - E[X_{it}|\alpha_i]\} \qquad \text{(E.5)} \\
\le &\left|\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{t=1}^T \{X_{it} - E[X_{it}|\alpha_i]\} - \max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{l=1}^m L_{il}\right| + \max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{l=1}^m L_{il} \\
\le &\left|\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_l S_{il}\right| + \left|\max_{1\le i\le N} \frac{1}{\sqrt{T}} S_{i,(m+1)}\right| + \max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{l=1}^m L_{il}. \qquad \text{(E.6)}
\end{aligned}$$

Applying Corollary 2.7 in Yu (1994), we have

$$\sup_{y\in\mathbb{R}} \left| P\left(\max_{1\le i\le N} \sum_{l=1}^m L_{il} \le y \mid (\alpha_i)_i\right) - P\left(\max_{1\le i\le N} \sum_{l=1}^m \widehat{L}_{il} \le y \mid (\alpha_i)_i\right)\right| \le (m-1)\beta(s), \qquad \text{(E.7)}$$

$$\sup_{y>0} \left| P\left(\max_{1\le i\le N} \left|\sum_{l=1}^m S_{il}\right| > y \mid (\alpha_i)_i\right) - P\left(\max_{1\le i\le N} \left|\sum_{l=1}^m \widehat{S}_{il}\right| > y \mid (\alpha_i)_i\right)\right| \le (m-1)\beta(q). \qquad \text{(E.8)}$$

Therefore, for every $\delta_1, \delta_2 > 0$, (E.3) yields

$$P\left(\max_{1\le i\le N} \frac{1}{\sqrt{T}} \sum_{t=1}^T \{X_{it} - E[X_{it}|\alpha_i]\} \le y \mid (\alpha_i)_i\right)$$

$$
\begin{aligned}
\leq & P\left(\max_{1\leq i\leq N}\frac{1}{\sqrt{T}}\sum_{l=1}^{m}\widehat{L}_{il}\leq y+\delta_1+\delta_2\mid(\alpha_i)_i\right)+P\left(\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}\sum_{l=1}^{m}\widehat{S}_{il}\right|>\delta_1\mid(\alpha_i)_i\right)\\
& +P\left(\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}S_{i,(m+1)}\right|>\delta_2\mid(\alpha_i)_i\right)+2(m-1)\beta(s)\\
=: & (i)+(ii)+(iii)+(iv). \qquad \text{(E.9)}
\end{aligned}
$$

Recall that $m\beta(s)=o(1)$, and thus $(iv)=o(1)$. Second, we bound the term (iii) in (E.9) by noting that $S_{i,(m+1)}$ consists of a sum over s terms, each of which is bounded in modulus by $2K$. Thus

$$
\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}S_{i,(m+1)}\right|\lesssim\frac{2sK}{\sqrt{T}}=o(1),
$$

and hence $(iii)=o(1)$ for any $\delta_2>0$.

Next, we bound the term (ii) in (E.9). Since $\|X_{it}\|_\infty\leq K$ and $\widehat{S}_{il}$ is the sum of s terms, it follows that $\|\widehat{S}_{il}\|_\infty\leq 2sK$. Then, applying Theorem 2.14.1. in van der Vaart and Wellner (1996) then yields

$$
E\left[\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}\sum_{l=1}^{m}\widehat{S}_{il}\right|\mid(\alpha_i)_i\right]\lesssim\sqrt{\frac{s^2\log N}{q}}=o(1).
$$

Markov's inequality then implies that the term (ii) is $o(1)$.

As δ_1,δ_2 are arbitrary, this verifies Equation (E.4).

Now, by the independence of $(\widehat{L}_{il})_{l=1}^m$ and $\|\widehat{L}_{il}\|_\infty\leq 2qK$, Theorem 2.14.1 in van der Vaart and Wellner (1996) can be applied to yield that

$$
E\left[\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}\sum_{l=1}^{m}\widehat{L}_{il}\right|\right]\lesssim\sqrt{\frac{m}{T}}\cdot\sqrt{q^2\log N}\lesssim\sqrt{q\log N}.
$$

In the light of

$$
\max_{1\leq i\leq N}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{X_{it}-E[X_{it}|\alpha_i]\}\overset{d}{=}\max_{1\leq i\leq N}\frac{1}{\sqrt{T}}\sum_{l=1}^{m}\widehat{L}_{il},
$$

implied by Equation (E.4), we now conclude

$$
E\left[\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t'=1}^{T}X_{it'}-E[X_{it}|\alpha_i]\right|\right]\lesssim\sqrt{\frac{q\log N}{T}}=o(1).
$$

Note that

$$
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}=E[X_{it}]+O\left(\frac{1}{\sqrt{N\wedge T}}\right) \qquad \text{(E.10)}
$$

by our Theorem 2.

By combining the above uniform rates, under non-degeneracy, we have

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}_{it}U_{it} =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\frac{1}{N}\sum_{i'=1}^{N}X_{i't}-\frac{1}{T}\sum_{t'=1}^{T}X_{it'}+\frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T}X_{i't'}\right)U_{it} \\
=& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-E[X_{it}|\gamma_t]-E[X_{it}|\alpha_i]+E[X_{it}]\right)U_{it}+o_p(1)\cdot\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}U_{it} \\
=& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}U_{it}+o_p(1)\cdot\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}U_{it} \\
=& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}U_{it}+o_p\left(\frac{1}{\sqrt{N\wedge T}}\right),
\end{aligned}$$

where the first equality follows by the definition of $\ddot{X}_{it}$, the second equality follows by (E.1), (E.2) and (E.10), the third equality follows by the definition of $\widetilde{X}_{it}$, and the fourth equality uses $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}U_{it} = O_p((N\wedge T)^{-1/2})$ which follows from our Theorem 2.

Following a similar decomposition and a crude calculation, we have

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}_{it}\ddot{X}_{it}' = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}\widetilde{X}_{it}' + o_p(1).$$

We have shown that

$$\sqrt{N\wedge T}(\widehat{\beta}-\beta) = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}\widetilde{X}_{it}'\right)\frac{1}{\sqrt{N\wedge T}}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}U_{it}+o_p(1).$$

Note that the sums have summands of the forms, $\widetilde{X}_{it}\widetilde{X}_{it}'$ and $\widetilde{X}_{it}U_{it}$, which only depend on $(\alpha_i,\gamma_t,\varepsilon_{it})$. Thus, our Theorems 2 and 3 can be applied. By replicating the Proof of Theorem 4, we have the desired result for the non-degenerate case.

Similarly, under i.i.d. sampling, by applying Theorem 2.14.1 in van der Vaart and Wellner (1996), we have

$$\begin{aligned}
E\left[\max_{t=1,\ldots,T}\left|\frac{1}{N}\sum_{i'=1}^{N}X_{it}-E[X_{it}|\gamma_t]\right|\right]\bigvee E\left[\max_{i=1,\ldots,N}\left|\frac{1}{T}\sum_{t'=1}^{T}X_{it'}-E[X_{it}|\alpha_i]\right|\right] \\
= O\left(\sqrt{\frac{\log NT}{N\wedge T}}\right)=o(1),
\end{aligned}$$

and thus

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}_{it}U_{it} =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\frac{1}{N}\sum_{i'=1}^{N}X_{i't}-\frac{1}{T}\sum_{t'=1}^{T}X_{it'}+\frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T}X_{i't'}\right)U_{it}\\ =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}U_{it}+o_p(1)\cdot\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}U_{it}\\ =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}U_{it}+o_p\left(\frac{1}{\sqrt{NT}}\right),\end{aligned}$$

where the first sum has mean zero and has its summands depending only on ε_{it} and thus is i.i.d. over i and t. The rest follows from the same arguments in the non-degenerate case. □

F Technical Lemmas

This section contains key technical lemmas for consistency of variance estimation under the current asymptotic setting. Throughout this section, for any $a, b \in \mathbb{R}^+ \cup \{0\}$, we use the short-hand notation $a \lesssim b$ to indicate $a \leq Cb$ for some $C < \infty$ independent of (N, T).

F.1 Generalized Newey-West Estimator under Non-Degeneracy

Lemma 1 (Generalized Newey-West Estimator under Non-Degeneracy)**.** *If Assumption 3 holds with (iv)(1), then*

$$\begin{aligned}\frac{1}{N^2T}\sum_{m=1}^{M}w(m,M)(\widehat{G}'_m-\widehat{H}'_m) &= \frac{1}{N^2T}\sum_{m=1}^{M}w(m,M)\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X'_{i',t-m}\\ &\xrightarrow{p}\sum_{m=1}^{\infty}E[X_{it}U_{it}U_{i',t-m}X'_{i',t-m}]\\ &=\sum_{m=1}^{\infty}E[b_tb'_{t-m}]\end{aligned}$$

for $i \neq i'$.

Proof. A sequence of symmetric matrices A_n converges to a symmetric matrix A_0 if and only if $b'A_nb \to b'A_0b$ for all comfortable b. Therefore, it suffices to assume without the loss of generality that $k = 1$.

First notice that by the law of total covariance as well as Assumption 3(i), for $m = 1, ..., M$ and $T = m + 1, ..., T$, it holds that

$$
\begin{aligned}
&E[X_{it}U_{it}U_{i't-m}X_{i',t-m}] \\
=&cov(X_{it}U_{it}, X_{i',t-m}U_{i',t-m}) \\
=&cov(E[X_{it}U_{it} \mid \gamma_t, \gamma_{t-m}], E[X_{i',t-m}U_{i',t-m} \mid \gamma_t, \gamma_{t-m}]) + E[cov(X_{i,t}U_{i,t}, X_{i',t-m}U_{i',t-m} \mid \gamma_t, \gamma_{t-m})] \\
=&cov(E[X_{it}U_{it} \mid \gamma_t], E[X_{i,t-m}U_{i,t-m} \mid \gamma_{t-m}]) + 0 = E[b_t b_{t-m}],
\end{aligned}
$$

where the second to the last equality follows from the independence of α_i. This verifies the last equality on the right hand side of the statement. To show the convergence in probability, consider the decomposition

$$
\begin{aligned}
&\left| \sum_{m=1}^{M} \frac{w(m,M)}{N^2T} \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X_{i',t-m} - \sum_{m=1}^{\infty} E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] \right| \\
\leq &\left| \sum_{m=1}^{M} \frac{w(m,M)}{N^2T} \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} \left\{ X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X_{i',t-m} - X_{it}U_{it}U_{i',t-m}X_{i',t-m} \right\} \right| \\
&+ \left| \sum_{m=1}^{M} \frac{w(m,M)}{N^2T} \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} \left\{ X_{it}U_{it}U_{i',t-m}X_{i',t-m} - E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] \right\} \right| \\
&+ \left| \sum_{m=1}^{M} \frac{1}{N^2T} |w(m,M) - 1| \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] \right| \\
&+ \left| \sum_{m=M+1}^{\infty} \frac{1}{N^2T} \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] \right| + o_p(1) \\
=:&(1) + (2) + (3) + (4) + o_p(1).
\end{aligned} \tag{F.1}
$$

Note that we have used the fact that

$$
\sum_{m=1}^{\infty} E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] = \frac{1}{N^2T} \sum_{m=1}^{\infty} \sum_{t=m+1}^{T} \sum_{i=1}^{N} \sum_{i' \neq i}^{N} E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}] + o(1)
$$

under Assumption 3 (i). It suffices to show that each of the four terms, (1)–(4), is asymptotically negligible.

First, consider term (2). Define

$$Z_{tm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\left\{X_{it}U_{it}U_{i',t-m}X_{i',t-m} - E[X_{it}U_{it}U_{i',t-m}X_{i',t-m} \mid (\alpha_i)_{i=1}^N]\right\} \qquad \text{and}$$

$$\widetilde{Z}_{tm} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\left\{E[X_{it}U_{it}U_{i',t-m}X_{i',t-m} \mid (\alpha_i)_{i=1}^N] - E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}]\right\}$$

for each N, t and m. With this notation, we can bound term (2) as

$$(2) \leq \left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}Z_{tm}\right| + \left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}\widetilde{Z}_{tm}\right|. \tag{F.2}$$

Consider the first term on the right-hand side of (F.2). Observe that $E[Z_{tm} \mid (\alpha_i)_{i=1}^N] = 0$. By Theorem 14.2 in Davidson (1994) with $r = 2(r+\delta)$ (where r on the left-hand side is in terms of the notation by Davidson (1994), and r and δ on the right-hand side satisfy our Assumption 3) and $p = 2$,

$$\left\{E\left[\left|E\left[Z_{tm} \mid (\alpha_i)_{i=1}^N, \mathcal{F}_{-\infty}^{t-\ell}\right]\right|^2 \mid (\alpha_i)_{i=1}^N\right]\right\}^{1/2} \leq 6\alpha(\ell)^{1/2-1/2(r+\delta)}\left\{E[|Z_{tm}|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N]\right\}^{1/2(r+\delta)}$$

almost surely. Then, by Lemma A in Hansen (1992) with $\beta = 2$,

$$\begin{aligned}\left\{E\left[\left|\frac{1}{T}\sum_{t=m+1}^{T}Z_{tm}\right|^2 \mid (\alpha_i)_{i=1}^N\right]\right\}^{1/2} &\lesssim \frac{1}{T}\sum_{\ell=1}^{\infty}\alpha(\ell)^{1/2-1/2(r+\delta)}\left\{\sum_{t=m+1}^{T}\left(E[|Z_{tm}|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N]\right)^{1/(r+\delta)}\right\}^{1/2} \\ &\lesssim T^{-1/2}\left\{E[|Z_{tm}|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N]\right\}^{1/2(r+\delta)}\end{aligned}$$

for each $m \geq 0$. Here, we have used the boundedness $\sum_{\ell=1}^{\infty}\alpha(\ell)^{1/2-1/2(r+\delta)} < \infty$ implied by Assumption 3 (iii). By Minkowski's inequality and the inequality obtained above, we have

$$\begin{aligned}&\frac{T^{1/2}}{M}\left\{E\left[\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}Z_{tm}\right|^2 \mid (\alpha_i)_{i=1}^N\right]\right\}^{1/2} \\ &\leq \frac{T^{1/2}}{M}\sum_{m=1}^{M}|w(m,M)|\left\{E\left[\left|\frac{1}{T}\sum_{t=m+1}^{T}Z_{tm}\right|^2 \mid (\alpha_i)_{i=1}^N\right]\right\}^{1/2} \\ &\lesssim \left\{E[|Z_{tm}|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N]\right\}^{1/2(r+\delta)}\end{aligned}$$

uniformly in T. By Markov's inequality, for any $\varepsilon > 0$,

$$P\left(\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}Z_{tm}\right| > \varepsilon \mid (\alpha_i)_{i=1}^N\right) = O\left(E\left[\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}Z_{tm}\right|^2 \mid (\alpha_i)_{i=1}^N\right]\right)$$

almost surely. Thus, by Fubini theorem, Jensen's inequality, and the bounded moment $\left\{E[|Z_{tm}|^{2(r+\delta)}]\right\}^{1/2(r+\delta)} < \infty$ following Assumption 3 (ii), we have

$$P\left(\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}Z_{tm}\right|>\varepsilon\right)=O\left(\frac{M^2}{T}\right)=o(1)$$

as $M^2/T = o(1)$ under Assumption 3 (vi).

Next, consider the second term on the right-hand side of (F.2). By Minkowski's inequality,

$$\frac{N^{1/2}}{M}\left\{E\left[\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}\widetilde{Z}_{tm}\right|^2\right]\right\}^{1/2}\leq\frac{N^{1/2}}{M}\sum_{m=1}^{M}|w(m,M)|\left\{E\left[\left|\frac{1}{T}\sum_{t=m+1}^{T}\widetilde{Z}_{tm}\right|^2\right]\right\}^{1/2}<\infty$$

uniformly in T. To see the last inequality, set $\theta = 0$ without loss of generality. By the identical distribution of γ_t,

$$E\left[\left|\frac{1}{T}\sum_{t=m+1}^{T}\widetilde{Z}_{tm}\right|^2\right]=O\left(E[|\widetilde{Z}_{tm}|^2]\right).$$

Now, fix any m and denote $W_{ii'} = E[X_{i,t+m}U_{i,t+m}U_{i't}X_{i't}]$. Then

$$\begin{aligned}
E[|\widetilde{Z}_{tm}|^2] &= \frac{1}{N^4}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\sum_{\iota=1}^{N}\sum_{\iota'\neq\iota}^{N}E\left[\left(E[W_{ii'}\mid(\alpha_i)_{i=1}^N]-E[W_{ii'}]\right)\left(E[W_{\iota\iota'}\mid(\alpha_i)_{i=1}^N]-E[W_{\iota\iota'}]\right)\right]\\
&= \frac{1}{N^4}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\sum_{\iota=1}^{N}\sum_{\iota'\neq\iota}^{N}E\left[\left(E[W_{ii'}\mid\alpha_i,\alpha_{i'}]-E[W_{ii'}]\right)\left(E[W_{\iota\iota'}\mid\alpha_\iota,\alpha_{\iota'}]-E[W_{\iota\iota'}]\right)\right]\\
&\leq \frac{2}{N^4}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\sum_{\iota'\neq i}^{N}E\left[\left(E[W_{ii'}\mid\alpha_i,\alpha_{i'}]-E[W_{ii'}]\right)\left(E[W_{i\iota'}\mid\alpha_i,\alpha_{\iota'}]-E[W_{i\iota'}]\right)\right]\\
&+ \frac{2}{N^4}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\sum_{\iota=1}^{N}E\left[\left(E[W_{ii'}\mid\alpha_i,\alpha_{i'}]-E[W_{ii'}]\right)\left(E[W_{\iota i}\mid\alpha_\iota,\alpha_i]-E[W_{\iota i}]\right)\right]\\
&+ \frac{1}{N^4}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}E\left[\left(E[W_{ii'}\mid\alpha_i,\alpha_{i'}]-E[W_{ii'}]\right)^2\right]\\
&\leq C\left(\frac{1}{N}E[|X_{i,t+m}U_{i,t+m}U_{\iota t}X_{\iota t}|^2]\right)\\
&\leq C\left(\frac{1}{N}E[|X_{it}U_{it}|^4]\right)=O\left(\frac{1}{N}\right)
\end{aligned}$$

for some constant $C > 0$, following Assumption 3 (ii) and Jensen's inequality. Note that the second inequality holds since $E[W_{ii'} \mid \alpha_i, \alpha_{i'}]$ and $E[W_{\iota\iota'} \mid \alpha_\iota, \alpha_{\iota'}]$ are independent when $(i, i') \neq (\iota, \iota')$. This

bound holds uniformly over all $m = 1, ..., M$. An application of Markov's inequality combined with the above calculations yields

$$P\left(\left|\sum_{m=1}^{M}\frac{w(m,M)}{T}\sum_{t=m+1}^{T}\widetilde{Z}_{tm}\right| > \varepsilon\right) = O\left(\frac{M^2}{N}\right) = o(1).$$

Now, take term (1), which we bound as follows.

$$(1) \lesssim \left|\sum_{m=1}^{M}\frac{w(m,M)}{N^2T}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}\Big\{(\beta-\widehat{\beta})X_{it}U_{i',t-m}+U_{it}X_{i',t-m}(\beta-\widehat{\beta}) \right.$$
$$\left. +X_{it}X_{i',t-m}(\beta-\widehat{\beta})^2\Big\}X_{i',t-m}\right|. \tag{F.3}$$

The first term on the right-hand side of (F.3) is bounded by

$$|\beta-\widehat{\beta}|\left|\sum_{m=1}^{M}\frac{w(m,M)}{N^2T}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}^2U_{i',t-m}X_{i',t-m}'\right| = O_p\left(\frac{M}{\sqrt{\min\{N,T\}}}\right) = o_p(1)$$

under Assumption 3 (ii) and (vi). Here, we have used $|\widehat{\beta}-\beta| = O_p((\min\{N,T\})^{-1/2})$ implied by the central limit theorem under Assumption 3. A similar argument applies to the second term on the right-hand side of (F.3).

Next, consider term (3), which equals

$$(3) = \left|\sum_{m=1}^{M}|w(m,M)-1|\frac{1}{T}\sum_{t=m+1}^{T}E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}]\right|.$$

Applying Theorem 14.13.2 in Hansen (2022) conditional on $(\alpha_i)_{i=1}^N$, we have

$$\begin{aligned}&\left|E\left[X_{it}U_{it}U_{i',t-m}X_{i',t-m} \mid (\alpha_i)_{i=1}^N\right]\right| \\ &\lesssim \alpha(m)^{(1+2\delta)/(4+4\delta)}\left\{E[|X_{it}U_{it}|^{4(r+\delta)} \mid \alpha_i]\right\}^{1/4(r+\delta)}\left\{E[|U_{i',t-m}X_{i',t-m}|^2 \mid \alpha_{i'}]\right\}^{1/2}\end{aligned}$$

with $\sum_{m=1}^{\infty}\alpha(m)^{(1+2\delta)/(4+4\delta)} < \infty$ following Assumption 3 (iii). Recall that α_i are i.i.d. By integrating out α_i and $\alpha_{i'}$ and applying Jensen's inequality, we have

$$\begin{aligned}\left|E[X_{it}U_{it}U_{i',t-m}X_{i',t-m}]\right| &\le E\left[\left|E\left[X_{it}U_{it}U_{i',t-m}X_{i',t-m} \mid (\alpha_i)_{i=1}^N\right]\right|\right] \\ &\lesssim \alpha(m)^{(1+2\delta)/(4+4\delta)}\left\{E[|X_{it}U_{it}|^{4(r+\delta)}]\right\}^{1/4(r+\delta)}\left\{E[|X_{it}U_{it}|^2]\right\}^{1/2} \\ &\lesssim \alpha(m)^{(1+2\delta)/(4+4\delta)}\end{aligned}$$

for all m. Since $w(m, M) \to 1$ as $T \to \infty$ for each m, the dominated convergence theorem, the above bound, and Assumption 3 (iii) together imply $(3) = o(1)$.

Finally, consider term (4). By the law of total covariance, one can rewrite (4) as

$$(4) = O\left(\left|\sum_{m=M+1}^{\infty} cov(X_{it}U_{it}, X_{i',t-m}U_{i',t-m})\right|\right) = O\left(\left|\sum_{m=M+1}^{\infty} cov(E[X_{it}U_{it} \mid \gamma_t], E[X_{i',t-m}U_{i',t-m} \mid \gamma_{t-m}])\right|\right).$$

Thus, $(4) = o(1)$ follows from an application of Lemma 6.17 in White (1984) as $m \to \infty$. □

F.2 Eicker-White CRVE under Non-Degeneracy

Lemma 2 (Eicker-White CRVE under non-degeneracy)**.** *If Assumption 3 holds with (iv)(1), then*

$$\begin{aligned}\frac{1}{N^2T}\sum_{t=1}^{T}\widehat{S}_t\widehat{S}_t' =& \frac{1}{N^2T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{i'=1}^{N} X_{it}\widehat{U}_{it}\widehat{U}_{i't}X_{i't}' \\ &\xrightarrow{p} E\left[(E[X_{it}U_{it} \mid \gamma_t])(E[X_{it}U_{it} \mid \gamma_t])'\right] \\ &= E[b_t b_t'], \end{aligned} \tag{F.4}$$

and

$$\begin{aligned}\frac{1}{NT^2}\sum_{i=1}^{N}\widehat{R}_i\widehat{R}_i' =& \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} X_{it}\widehat{U}_{it}\widehat{U}_{it'}X_{it'}' \\ &\xrightarrow{p} E\left[(E[X_{it}U_{it} \mid \alpha_i])(E[X_{it}U_{it} \mid \alpha_i])'\right] \\ &= E[a_i a_i']. \end{aligned} \tag{F.5}$$

Proof. Throughout the proof, assume $k = 1$ without loss of generality.

Proof of (F.4)**:** Notice that we have

$$\begin{aligned}E[X_{it}U_{it}U_{i't}X_{i't}] &= cov(X_{it}U_{it}, U_{i't}X_{i't}) \\ &= E[cov(X_{it}U_{it}, U_{i't}X_{i't}|\gamma_t)] + cov(E[X_{it}U_{it}|\gamma_t], E[X_{i't}U_{i't}|\gamma_t]) \\ &= 0 + E\left[(E[X_{it}U_{it} \mid \gamma_t])(E[X_{it}U_{it} \mid \gamma_t])\right]\end{aligned}$$

by the law of total covariance. Thus, it suffices to bound the right-hand side of

$$\left|\frac{1}{N^2T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{i'=1}^{N} X_{it}\widehat{U}_{it}\widehat{U}_{i't}X_{i't} - E[X_{it}U_{it}U_{i't}X_{i't}]\right|$$

$$\leq \left| \frac{1}{N^2 T} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{i'=1}^{N} \{ X_{it} \widehat{U}_{it} \widehat{U}_{i't} X_{i't} - X_{it} U_{it} U_{i't} X_{i't} \} \right| + \left| \frac{1}{N^2 T} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{i'=1}^{N} X_{it} U_{it} U_{i't} X_{i't} - E[X_{it} U_{it} U_{i't} X_{i't}] \right| =: (1) + (2). \tag{F.6}$$

First, consider term (2). Set

$$Z_t = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left\{ X_{it} U_{it} U_{i't} X_{i't} - E[X_{it} U_{it} U_{i't} X_{i't} \mid (\alpha_i)_{i=1}^N] \right\} \qquad \text{and}$$

$$\widetilde{Z}_t = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left\{ E[X_{it} U_{it} U_{i't} X_{i't} \mid (\alpha_i)_{i=1}^N] - E[X_{it} U_{it} U_{i't} X_{i't}] \right\}$$

for each t, N. We decompose term (2) as

$$(2) \leq \left| \frac{1}{T} \sum_{t=1}^{T} Z_t \right| + \left| \frac{1}{T} \sum_{t=1}^{T} \widetilde{Z}_t \right| =: (3) + (4). \tag{F.7}$$

Second, consider term (3). Observe that $E[Z_t \mid (\alpha_i)_{i=1}^N] = 0$. By Theorem 14.2 in Davidson (1994) with $r = 2(r + \delta)$ (where r on the left-hand side is in terms of the notation by Davidson (1994), and r and δ on the right-hand side satisfy our Assumption 3) and $p = 2$, we have

$$\left\{ E\left[\left| E\left[Z_t \mid (\alpha_i)_{i=1}^N, \mathcal{F}_{-\infty}^{t-\ell} \right] \right|^2 \mid (\alpha_i)_{i=1}^N \right] \right\}^{1/2} \leq 6\alpha(\ell)^{1/2 - 1/2(r+\delta)} \left\{ E[|Z_t|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N] \right\}^{1/2(r+\delta)}$$

almost surely. Then, by Lemma A in Hansen (1992) with $\beta = 2$, it holds that

$$\begin{aligned} \left\{ E\left[\left| \frac{1}{T} \sum_{t=1}^{T} Z_t \right|^2 \mid (\alpha_i)_{i=1}^N \right] \right\}^{1/2} &\lesssim \frac{1}{T} \sum_{\ell=1}^{\infty} \alpha(\ell)^{1/2 - 1/2(r+\delta)} \left\{ \sum_{t=1}^{T} \left(E[|Z_t|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N] \right)^{1/(r+\delta)} \right\}^{1/2} \\ &\lesssim T^{-1/2} \left\{ E[|Z_t|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N] \right\}^{1/2(r+\delta)}. \end{aligned}$$

Thus, we have

$$T^{1/2} \left\{ E\left[\left| \frac{1}{T} \sum_{t=1}^{T} Z_t \right|^2 \mid (\alpha_i)_{i=1}^N \right] \right\}^{1/2} \lesssim \left\{ E[|Z_{tm}|^{2(r+\delta)} \mid (\alpha_i)_{i=1}^N] \right\}^{1/2(r+\delta)}$$

uniformly in T. By Markov's inequality, we obtain for any $\varepsilon > 0$

$$P\left(\left| \frac{1}{T} \sum_{t=1}^{T} Z_t \right| > \varepsilon \mid (\alpha_i)_{i=1}^N \right) = O\left(E\left[\left| \frac{1}{T} \sum_{t=1}^{T} Z_t \right|^2 \mid (\alpha_i)_{i=1}^N \right] \right)$$

almost surely. Therefore, by Fubini theorem, Jensen's inequality, and the bounded moment $E[|Z_t|^{2(r+\delta)}] < \infty$ that holds under Assumption 3 (ii), we have

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| > \varepsilon\right) = O\left(\frac{1}{T}\right) = o(1),$$

showing (3) $= o_p(1)$ in (F.7).

Third, consider term (4) in (F.7). By the identical distribution of the γ_t, we have $E[|T^{-1}\sum_{t=1}^{T}\widetilde{Z}_t|^2] = O(E[|\widetilde{Z}_t|^2]) = O(N^{-1})$, the final equality by a direct calculation. Markov's inequality implies that

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} \widetilde{Z}_t\right| > \varepsilon\right) = O\left(\frac{1}{N}\right) = o(1),$$

showing (4) $= o_p(1)$ in (F.7). Thus, we obtain (2) $= o_p(1)$ in (F.6).

Finally, consider term (1) in (F.6). Note that

$$(1) \lesssim \left|\frac{1}{N^2T}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'=1}^{N} X_{it}\left\{(\beta-\widehat{\beta})X_{it}U_{i't} + U_{it}X_{i't}(\beta-\widehat{\beta}) + X_{it}X_{i't}(\beta-\widehat{\beta})^2\right\}X_{i't}\right|.$$

Similar to (F.3), the first two terms are $O_p(|\widehat{\beta}-\beta|) = O_p((\min\{N,T\})^{-1/2})$ while the third term is $O_p(|\widehat{\beta}-\beta|^2) = O_p((\min\{N,T\})^{-1})$. It therefore follows that (1) $= O_p((\min\{N,T\})^{-1/2})$ in (F.6).

Proof of (F.5)**:** Observe that

$$\begin{aligned}
&\left|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} X_{it}\widehat{U}_{it}\widehat{U}_{it'}X_{it'} - E\left[(E[X_{it}U_{it} \mid \alpha_i])(E[X_{it}U_{it} \mid \alpha_i])\right]\right| \\
&\leq \left|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left\{X_{it}\widehat{U}_{it}\widehat{U}_{it'}X_{it'} - X_{it}U_{it}U_{it'}X_{it'}\right\}\right| \\
&+ \left|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T}\{X_{it}U_{it}U_{it'}X_{it'} - E[X_{it}U_{it}U_{it'}X_{it'}]\}\right| \\
&+ \left|\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} E[X_{it}U_{it}U_{it'}X_{it'}] - E\left[(E[X_{it}U_{it} \mid \alpha_i])(E[X_{it}U_{it} \mid \alpha_i])\right]\right| = (5)+(6)+(7). \quad \text{(F.8)}
\end{aligned}$$

Term (5) can be shown to be $O_p(|\widehat{\beta}-\beta|) = o_p(1)$ similarly to term (1) in (F.6). Consider term (7). By the law of total covariances,

$$E[X_{it}U_{it}U_{it'}X_{it'}] = cov(E[X_{it}U_{it}|\alpha_i], E[X_{it'}U_{it'}|\alpha_i]) + E[cov(X_{it}U_{it}, X_{it'}U_{it'}|\alpha_i)].$$

From this equality follows

$$\begin{aligned}
&\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} E\left[X_{it}U_{it}U_{it'}X_{it'}\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\{cov(E[X_{it}U_{it} \mid \alpha_i], E[X_{it'}U_{it'} \mid \alpha_i]) + E[cov(X_{it}U_{it}, X_{it'}U_{it'} \mid \alpha_i)]\} \\
&= E\left[(E[X_{it}U_{it} \mid \alpha_i])(E[X_{it}U_{it} \mid \alpha_i])\right] + o(1). \qquad \text{(F.9)}
\end{aligned}$$

To see the second equality, note that, by Assumption 3 (i)–(iii) and an application of Theorem 14.13 (ii) in Hansen (2022) along with Jensen's inequality, for any $t, t' \in \{1, ..., T\}$, we have

$$E[cov(X_{it}U_{it}, X_{it'}U_{it'} \mid \alpha_i)] \leq 8\left(E[|X_{it}U_{it}|^{4(r+\delta)}]\right)^{1/2(r+\delta)} \alpha(|t-t'|)^{1-1/2(r+\delta)}.$$

Moreover, Assumption 1 (iii) implies $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-1/2(r+\delta)} < \infty$. Therefore,

$$T^{-2}\sum_{t=1}^{T}\sum_{t'=1}^{T} E[cov(X_{it}U_{it}, X_{it'}U_{it'} \mid \alpha_1)] = o(1)$$

follows, which in turn implies the second equality in (F.9). This shows that (7) $= o(1)$ in (F.8).

Finally, take term (6) in (F.8). Define

$$Z_i = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}[X_{it}U_{it}U_{it'}X_{it'} - E\left[X_{it}U_{it}U_{it'}X_{it'} \mid (\gamma_t)_{t=1}^{T}\right]]$$

and

$$\begin{aligned}
\widetilde{Z}_i &= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} E\left[X_{it}U_{it}U_{it'}X_{it'} \mid (\gamma_t)_{t=1}^{T}\right] - E\left[X_{it}U_{it}U_{it'}X_{it'}\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\{h_i(\gamma_t, \gamma_{t'}) - E[h_i(\gamma_t, \gamma_{t'})]\}
\end{aligned}$$

where $h_i(\gamma_t, \gamma_{t'}) = E\left[X_{it}U_{it}U_{it'}X_{it'} \mid \gamma_t, \gamma_{t'}\right]$. We have the bound:

$$(6) \leq \left|\frac{1}{N}\sum_{i=1}^{N} Z_i\right| + \left|\frac{1}{N}\sum_{i=1}^{N} \widetilde{Z}_i\right| =: (8) + (9). \qquad \text{(F.10)}$$

Take (8). Note that conditional on $(\gamma_t)_{t=1}^{T}$, Z_i are mutually independent and mean zero. Jensen's and Markov's inequalities then imply that (8) $= o_p(1)$. Now, consider term (9). Note that $E[\widetilde{Z}_i] = 0$ and

$$E\left[\left|\frac{1}{N}\sum_{i=1}^{N} \widetilde{Z}_i\right|^2\right] \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} E\left|\widetilde{Z}_i\widetilde{Z}_{i'}\right| \leq E[\widetilde{Z}_i^2],$$

the final inequality by Cauchy-Schwarz and the fact that $\widetilde{Z}_i$ are identically distributed since α_i are.

By a direct calculation,

$$E[\widetilde{Z}_i^2] = \left[\frac{1}{T^4}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{t''=1}^{T}\sum_{t'''=1}^{T} h_i(\gamma_t,\gamma_{t'})h_i(\gamma_{t''}\gamma_{t'''})\right] - \left(\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} E[h_i(\gamma_t,\gamma_{t'})]E[h_i(\gamma_{t''}\gamma_{t'''})]\right)^2$$
$$\leq \frac{4}{T^4}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{t''=1}^{T}\sum_{t'''=1}^{T} h_i(\gamma_t,\gamma_{t'})h_i(\gamma_{t''}\gamma_{t'''}).$$

To control the right hand side, we apply Lemma 2 in Yoshihara (1976) with $\delta = 2(r+\delta)-2$, $\delta' = 2r-2$, $r = 2(r+\delta)$ (where r and δ on the left-hand side are in terms of the notation by Yoshihara (1976), and r and δ on the right-hand side satisfy our Assumption 3). With this setting, note that

$$\frac{2+\delta'}{\delta'} = \frac{2r}{2(r-1)} < \frac{2r}{r-1} < \lambda,$$

where r here is from Assumption 3, and thus Assumption 3 (iii) implies that

$$\beta(\ell) = O(\ell^{-\lambda}) = O(\ell^{-(2+\delta')/\delta'}),$$

which is sufficient for the mixing requirement in the Lemma 2 of Yoshihara (1976). We now verify Condition (2.4) in Yoshihara (1976), which requires

$$E\left[\left|E[X_{it}U_{it}U_{it'}X_{it'}|\gamma_t,\gamma_{t'}]\right|^{2(r+\delta)}\right] < \infty.$$

This follows from our Assumption 3(ii), Jensen's inequality, and Cauchy-Schwarz's inequality, as

$$E\left[\left|E[X_{it}U_{it}U_{it'}X_{it'}|\gamma_t,\gamma_{t'}]\right|^{2(r+\delta)}\right] \leq E\left[\left|X_{it}U_{it}U_{it'}X_{it'}\right|^{2(r+\delta)}\right]$$
$$\leq \left\{E\left[\left|X_{it}U_{it}\right|^{4(r+\delta)}\right]\cdot E\left[\left|U_{it'}X_{it'}\right|^{4(r+\delta)}\right]\right\}^{1/2} < \infty.$$

We next verify Condition (2.3) in Yoshihara (1976), which requires

$$\int\int \left|E[X_{it}U_{it}U_{it'}X'_{it'}|\gamma_t = u,\gamma_{t'} = v]\right|^{2(r+\delta)} dF(u)dF(v) < \infty,$$

where $F(\cdot)$ is the common CDF of γ_i. This holds since, by Cauchy-Schwarz and Assumption 3(i)

$$\left|E[X_{it}U_{it}U_{it'}X_{it'}|\gamma_t = u,\gamma_{t'} = v]\right|^2 \leq E[(X_{it}U_{it})^2|\gamma_t = u]E[(U_{it'}X_{it'})^2|\gamma_{t'} = v],$$

and thus by Jensen's inequality,

$$\begin{aligned}
&\int\int |E[X_{it}U_{it}U_{it'}X_{it'}|\gamma_t = u, \gamma_{t'} = v]|^{2(r+\delta)}\, dF(u)dF(v) \\
&\leq \int\int \left|E[(X_{it}U_{it})^2|\gamma_t = u] \cdot E[(U_{it'}X_{it'})^2|\gamma_{t'} = v]\right|^{(r+\delta)} dF(u)dF(v) \\
&\leq \int E[|X_{it}U_{it}|^{2(r+\delta)}|\gamma_t = u]dF(u) \cdot \int E[|U_{it'}X_{it'}|^{2(r+\delta))}|\gamma_{t'} = v]dF(v) \\
&= \left(E[E[|X_{it}U_{it}|^{2(r+\delta)}|\gamma_t]]\right)^2 = \left(E[|X_{it}U_{it}|^{2(r+\delta)}]\right)^2 < \infty
\end{aligned}$$

under Assumption 3(ii). Applying Lemma 2 of Yoshihara (1976) following Equation (10) in Dehling and Wendler (2010) now yields

$$E\left[\frac{4}{T^4}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{t''=1}^{T}\sum_{t'''=1}^{T} h_i(\gamma_t, \gamma_{t'})h_i(\gamma_{t''}\gamma_{t'''})\right] = O(T^{-1-\delta/(r-1)(r+\delta)}) = o(1),$$

where the last equality follows because $\delta/(r-1)(r+\delta) > 0$. This result and Markov's inequality together imply (9) $= o_p(1)$ in (F.10), and hence (6) $= o_p(1)$ in (F.10). This completes the proof of (F.5). □

F.3 Generalized Newey-West and Eicker-White CRVE under Degeneracy

Lemma 3 (Generalized Newey-West and Eicker-White CRVE under Degeneracy)**.** *If Assumption 3 holds with (iv)(2), then*

$$\frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\widehat{U}_{it}\widehat{U}_{it}X'_{it} - \frac{var(X_{11}U_{11})}{NT} = o_p\left(\frac{1}{NT}\right), \tag{F.11}$$

$$\begin{aligned}
&\frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{t=1}^{T} X_{it}\widehat{U}_{it}\widehat{U}_{i't}X'_{i't} \\
&\quad + \sum_{m=1}^{M}\frac{w(m,M)}{(NT)^2}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\left(X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X'_{i',t-m} + X_{i',t-m}\widehat{U}_{i',t-m}\widehat{U}_{it}X'_{it}\right) \\
&\quad - \frac{var(X_{11}U_{11})}{NT} = o_p\left(\frac{1}{T}\right),
\end{aligned} \tag{F.12}$$

and

$$\frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t'=1}^{T} X_{it}\widehat{U}_{it}\widehat{U}_{it'}X'_{it'} - \frac{var(X_{11}U_{11})}{NT} = o_p\left(\frac{1}{N}\right). \tag{F.13}$$

Proof. The first statement (F.11) follows from the consistency of $\widehat{\beta}$ implied by its asymptotic normality from the first part of Theorem 4 (which does not rely on the current lemma) and the law of large numbers for i.i.d. random variables.

The second statement (F.12) follows from

$$\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{t=1}^{T}X_{it}\widehat{U}_{it}\widehat{U}_{i't}X'_{i't} = \underbrace{\left(\frac{N-1}{N}\right)\frac{1}{N(N-1)T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}\sum_{t=1}^{T}X_{it}\widehat{U}_{it}\widehat{U}_{i't}X'_{i't}}_{=O_p((N^2T)^{-1/2})} + \underbrace{\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}\widehat{U}_{it}\widehat{U}_{it}X'_{it}}_{=var(R_{11})/N+O_p(N^{-1}(NT)^{-1/2})},$$

where the first term on the right-hand side is $o_p(1)$ by the law of large numbers for i.i.d. data, and the second term on the right-hand side is $o_p(1/N)$ by the first statement (F.11) . The consistency of $\widehat{\beta}$ and the law of large numbers for i.i.d. data imply that

$$\begin{aligned}
&\frac{1}{N^2T}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X'_{i',t-m} \\
&= \frac{1}{N^2T}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}U_{it}U_{i',t-m}X'_{i',t-m} + O_p\left(\|\widehat{\beta}-\beta\|\right) \\
&= O_p\left(\frac{1}{\sqrt{N^2T}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}$$

Thus, by Assumption 3 (v)(vi), we have

$$\sum_{m=1}^{M}\frac{w(m,M)}{(NT)^2}\sum_{t=m+1}^{T}\sum_{i=1}^{N}\sum_{i'\neq i}^{N}X_{it}\widehat{U}_{it}\widehat{U}_{i',t-m}X'_{i',t-m} = O_p\left(\frac{M}{T\sqrt{NT}}\right) = o_p\left(\frac{1}{T}\right).$$

Combining these probability limits together yields the second statement (F.12) .

The third statement (F.13) can be shown similarly to the first part of the second statement. □

G Heterogeneous Per-Cluster Numbers of Observations

The main results presented in Section 3.1 in the main text straightforwardly extends to a more general case with possibly zero or multiple observations per cluster intersection. Suppose that the (i,t)-th cluster contains J_{it} units $\{X_{it1},\ldots,X_{itJ_{it}}\}$ of observations, where J_{it} is considered a random variable.

As in the main text, let $\theta = E[X_{itj}] = 0$ without loss of generality. In this extended setting, the sample mean is defined by

$$\widetilde{\vartheta} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} J_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J_{it}} X_{itj}.$$

The standard approach to clustered data is to treat the cluster sum $\overline{X}_{it} = \sum_{j=1}^{J_{it}} X_{itj}$ as the effective observation for the (i,t)-th unit. Accordingly, define their projections $\overline{a}_i = E[\overline{X}_{it}|\alpha_i]$ and $\overline{b}_t = E[\overline{X}_{it}|\gamma_t]$. Let their (long-run) variances be denoted by $\overline{\Sigma}_a = E[\overline{a}_i\overline{a}_i']$ and $\overline{\Sigma}_b = \sum_{\ell=-\infty}^{\infty} E[\overline{b}_t\overline{b}_{t+\ell}']$. With $\widehat{\mu}_J = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} J_{it}$, we now extend Assumptions 1 and 2 in the baseline model as follows.

Assumption 4. (i) Assumption 1 holds with X_{it} replaced by $\overline{X}_{it}$. (ii) Assumption 2 holds with Σ_a and Σ_b replaced by $\overline{\Sigma}_a$ and $\overline{\Sigma}_b$, respectively. (iii) $\widehat{\mu}_J \xrightarrow{p} \mu_J \in (0,\infty)$.

Note that more low level conditions on sampling formulated in terms of J_{it} and X_{itj} can be done following the approach taken in Assumption 1 in Davezies, D'Haultfœuille, and Guyonvarch (2018). The following theorem states that the conclusions of Theorems 1 and 2 continue to hold under this generalized setting.

Theorem 6. *Suppose that Assumption 4 holds. Then,*

$$\widehat{\vartheta} \xrightarrow{p} \theta \qquad \textit{and} \qquad var(\widehat{\vartheta})^{-1/2}(\widehat{\vartheta} - \theta) \xrightarrow{d} N(0, I_m).$$

Proof. Under Assumption 4 (i), it immediately follows that the conclusion of Theorem 1 holds with θ, $\widehat{\theta}$, a_i, b_t, and e_{it} replaced by $\vartheta = E[\overline{X}_{it}]$, $\widehat{\vartheta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\overline{X}_{it}$, $\overline{a}_i$, $\overline{b}_t$, and $\overline{e}_{it} = \overline{X}_{it} - \overline{a}_i - \overline{b}_t$, respectively. Since $\widetilde{\vartheta} = \widehat{\mu}_J^{-1}\widehat{\vartheta}$, it follows that $\widehat{\vartheta} \xrightarrow{p} \theta$ as $N, T \to \infty$ under Assumption 4 (iii).

Under Assumption 4 (i)–(ii), the conclusion of Theorem 2 holds with $\widehat{\theta}$ replaced by $\widehat{\vartheta}$. Since $\widetilde{\vartheta} = \widehat{\mu}_J^{-1}\widehat{\vartheta}$, it follows that $var(\widetilde{\vartheta})^{-1/2}(\widetilde{\vartheta} - \theta) \xrightarrow{d} N(0, I_m)$ holds under Assumption 4 (iii). □

H Additional Information about Data Analyses

H.1 Data used for Section 2.3

For the analysis in Section 2.3, we use the data from Bloom, Schankerman, and Van Reenen (2013). This data set is publicly available as a supplementary material of Bloom, Schankerman, and Van Reenen (2013) from the Econometric Society. We use two variables contained in spillovers.dta. The log Tobin's average Q is available as the variable named lq. The log R&D stock divided by capital stock is available as the variable named grd_k.

H.2 Estimation of γ for the Preliminary Analysis in Section 2.3

In the preliminary analysis of the market value data to motivate our novel standard error formula, we estimate $\widehat{\gamma}_t$ for each t and its partial autocorrelation coefficient in Section 2.3. The current appendix section presents a concrete estimation procedure used to obtain the estimates in Section 2.3.

To eliminate the firm effect α_i, we first apply the within transformation of Y_{it} and obtain $\ddot{Y}_{it} = \gamma_t - \overline{\gamma} + \varepsilon_i - \overline{\varepsilon}_i$ for each (i,t), where $\overline{\gamma} = T^{-1}\sum_{t=1}^{T}\gamma_t$ and $\overline{\varepsilon}_i = T^{-1}\sum_{t=1}^{T}\varepsilon_{it}$. We can then estimate γ_t up to location $\overline{\gamma}$ by $\widehat{\gamma}_t = N^{-1}\sum_{i=1}^{N}\ddot{Y}_{it}$ for each t. The partial autocorrelation of $\{\gamma_t\}_t$ in can be also estimated by running the first-order autoregression of $\ddot{Y}_{it}$. We report the HC0 standard error in Section 2.3 because our robust standard error formula is yet to be introduced as of Section 2.3. The autocorrelogram is produced based on the series $\{\widehat{\gamma}_t\}_{t=1}^{T}$.

H.3 Data used for Section 6

For the analysis in Section 6, we use the data from Gagliardini, Ossola, and Scaillet (2016). This data set is publicly available as a supplementary material of Gagliardini, Ossola, and Scaillet (2016) from the Econometric Society. We combined data from multiple files located in the folder named GOS_dataCodes_paper. The returns of the 44 industry portfolios are available in Wspace_44Indu.mat, the returns of the 9936 individual stocks are available in Wspace_CRSPCMST_ret.mat, the Fama-French factors are available in Wspace_Fact.mat, and the risk-free rates (monthly 30-day T-bill yields)

are available in RiskFree.mat.

I Additional Simulations

I.1 Power

The baseline simulation studies presented in Section 5 focuses on the coverage probabilities. In this section, we present additional simulation studies focusing on the power.

We continue to use the same simulation designs as in Section 5. Instead of computing the coverage probabilities, however, we now compute the rejection probabilities for the hypothesis $H_0 : \beta_1 = b$ for various values of $b \in [0.5, 1.5]$ with the nominal size of 5%. Recall that the true value is $\beta_1 = 1$. We use the sample size of $N = T = 50$ throughout, and run 10,000 Monte Carlo iterations for each set of simulations.

Figure 4 illustrates the power curves for EHW, CRi, CRt, CGM, MNW, M, T, and CHS. Panel (A) illustrates power curves under the i.i.d. design. Panels (B), (C), and (D) illustrate power curves under the dependence designs with $\rho = 0.25$, 0.50, and 0.75, respectively.

The sizes are complementary to the coverage probabilities reported in Section 5. As the hypothesized value of β_1 deviates away from the true value $\beta_1 = 1$, the power increases for each method. Some methods show higher power than the others, but only at the expense of size distortions.

I.2 Simulations for the Two-Way Fixed-Effect Estimator

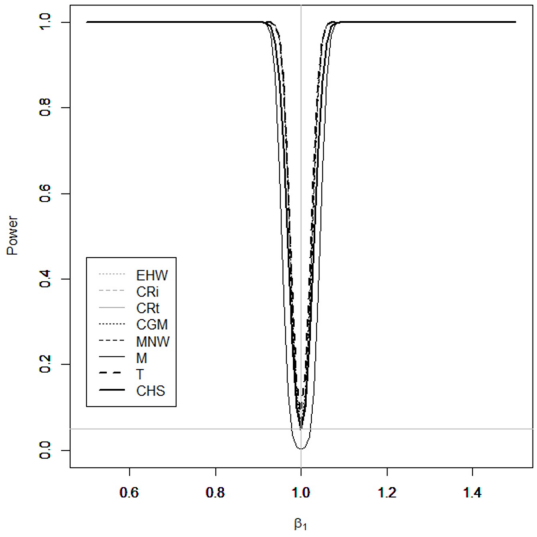
The baseline simulation studies presented in Section 5 focuses on the OLS. In this section, we present additional simulation studies focusing on the two-way fixed-effect estimator for fixed-effect models.

Motivated by the example model (4.1) in Section 4.1, consider the following data generating process.

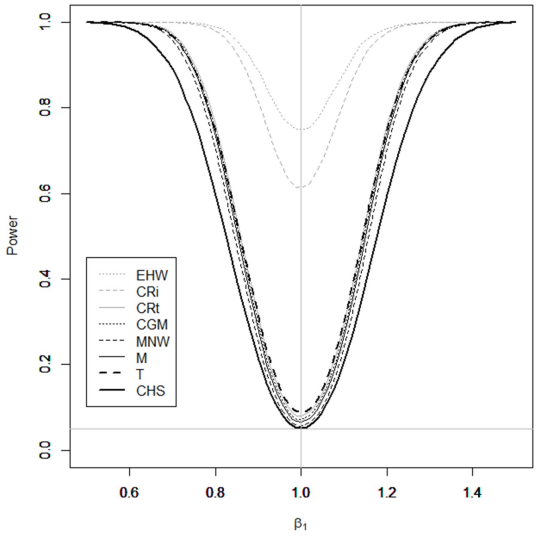
$$Y_{it} = \beta_0 + \beta_1 X_{it} + U_{it},$$

where $(\beta_0, \beta_1)' = (0.1, 0.1)'$ and the right-hand side variables $(X_{it}, U_{it})'$ are generated through the

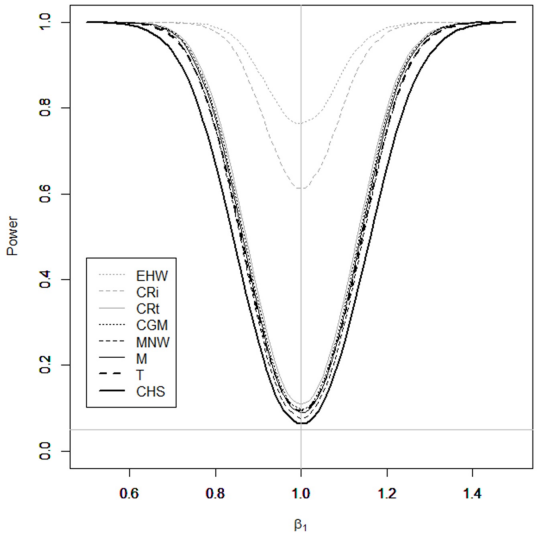
(A) I.I.D. Design



(B) Dependence Design with $\rho = 0.25$



(C) Dependence Design with $\rho = 0.50$



(D) Dependence Design with $\rho = 0.75$

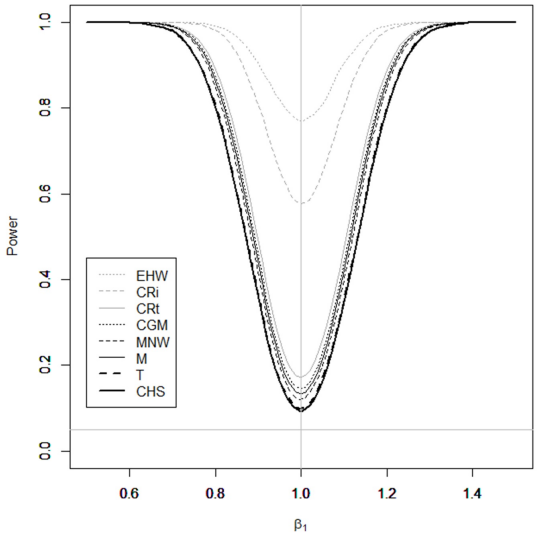


Figure 4: Power curves for EHW, CRi, CRt, CGM, MNW, M, T, and CHS based on 10,000 Monte Carlo iterations. The true parameter value is $\beta_1 = 1$ and the nominal size is 5%. Panel (A) illustrates power curves under the i.i.d. design. Panels (B), (C), and (D) illustrate power curves under the dependence designs with $\rho = 0.25$, 0.50, and 0.75, respectively. The sample size is set to $N = T = 75$ throughout.

panel dependence structure

$$X_{it} = w_1\alpha_{i1}\gamma_{t2} + w_2\alpha_{i2}\gamma_{t1} + w_3\varepsilon_{it0} \qquad \text{and}$$

$$U_{it} = w_4\alpha_{i0} + w_5\gamma_{t0} + w_6\alpha_{i1}\gamma_{t3} + w_7\alpha_{i3}\gamma_{t1} + w_8\varepsilon_{it1}.$$

For the weight parameters, we use $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8) = (0.0, 0.0, 0.5, 0.0, 0.0, 0.0, 0.0, 0.5)$ to generate i.i.d. data and $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8) = (0.1, 0.1, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1)$ to generate dependent data. The latent components $(\alpha_{i0}, \alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \varepsilon_{it0}, \varepsilon_{it1})$ are all mutually independent $N(0, 1)$.

Similarly to the baseline simulation design presented in Section 5, the latent common time effects $(\gamma_{t0}, \gamma_{t1}, \gamma_{t2}, \gamma_{t3})$ are dynamically generated according to the AR(1) design:

$$\gamma_{t0} = \rho\gamma_{(t-1)0} + \widetilde{\gamma}_{t0} \text{ where } \widetilde{\gamma}_{t0} \text{ are independent draws from } N(0, 1-\rho^2);$$

$$\gamma_{t1} = \rho\gamma_{(t-1)1} + \widetilde{\gamma}_{t1} \text{ where } \widetilde{\gamma}_{t1} \text{ are independent draws from } N(0, 1-\rho^2);$$

$$\gamma_{t2} = \rho\gamma_{(t-1)2} + \widetilde{\gamma}_{t2} \text{ where } \widetilde{\gamma}_{t2} \text{ are independent draws from } N(0, 1-\rho^2); \text{ and}$$

$$\gamma_{t3} = \rho\gamma_{(t-1)3} + \widetilde{\gamma}_{t3} \text{ where } \widetilde{\gamma}_{t3} \text{ are independent draws from } N(0, 1-\rho^2).$$

The initial values are drawn from $N(0, 1)$. We vary the AR coefficient $\rho \in \{0.25, 0.50, 0.75\}$ across sets of simulations.

For each realization of observed data $\{(Y_{it}, X_{it}) : 1 \leq i \leq N, 1 \leq t \leq T\}$ constructed according to the data generating process described above, we estimate β_1 by the two-way fixed-effect estimator and compute its standard error as in Section 4.2. As in the baseline simulation studies, we compare our estimator CHS with EHW, CRi, CRt, CGM, MNW, M and T; see Section 5 for details.

Table 3 reports simulation results. Reported values are the coverage frequencies for the slope parameter β_1 for the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show coverage probability results under the i.i.d. design and the dependence design, respectively. In each group of three consecutive rows, the panel sample sizes (N, T) vary by rows. Cells are shaded based on the proximity of the simulated coverage probability to the nominal probability of 0.95; the darker shades indicate more correct coverage.

I.I.D. Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	MNW	M	T	CHS
(I)	50	100	—	0.945	0.936	0.942	0.934	0.947	0.999	0.924	0.930
(II)	75	75	—	0.949	0.945	0.946	0.939	0.952	0.999	0.931	0.932
(III)	100	50	—	0.946	0.943	0.939	0.937	0.948	0.999	0.924	0.926

Dependence Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	MNW	M	T	CHS
(IV)	50	100	0.25	0.417	0.861	0.736	0.932	0.955	0.873	0.930	0.934
(V)	75	75	0.25	0.408	0.809	0.808	0.931	0.951	0.762	0.929	0.932
(VI)	100	50	0.25	0.408	0.726	0.855	0.929	0.953	0.864	0.922	0.927
(VII)	50	100	0.50	0.386	0.830	0.701	0.911	0.938	0.840	0.922	0.927
(VIII)	75	75	0.50	0.370	0.766	0.763	0.902	0.927	0.707	0.911	0.923
(IX)	100	50	0.50	0.357	0.680	0.799	0.891	0.923	0.806	0.900	0.912
(X)	50	100	0.75	0.322	0.751	0.596	0.841	0.880	0.746	0.895	0.908
(XI)	75	75	0.75	0.309	0.667	0.638	0.810	0.846	0.561	0.859	0.890
(XII)	100	50	0.75	0.291	0.567	0.652	0.773	0.815	0.620	0.825	0.860

Table 3: Coverage probabilities for the slope parameter β_1 for the two-way fixed-effect estimator with the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show results under the i.i.d. and dependence designs, respectively. The sample size is indicated by (N, T). The parameter ρ indicates the AR coefficient in the dependence design. EHW stands for Eicker–Huber–White, CRi stands for cluster robust within i, CRt stands for cluster robust within t, CGM stands for Cameron-Gelbach-Miller, MNW stands for MacKinnon-Nielsen-Webb, M stands for Menzel, T stands for Thompson, and CHS stands for Chiang-Hansen-Sasaki.

Observe that we have similar qualitative patterns in these results to those presented for the baseline simulation studies presented in Section 5.

I.3 Simulations with Multiple Observations in A Cluster Intersection

The baseline simulation studies presented in Section 5 focuses the case where each (i,t) intersection in the panel contains one observation. In this section, we present additional simulation studies allowing for multiple observations in each (i,t) intersection.

For each i and t, we independently draw the number J_{it} of observations from Binomial$(5, 0.5)$. We generate data based on the linear model

$$Y_{itj} = \beta_0 + \beta_1 X_{itj} + U_{itj},$$

where the right-hand side variables $(X_{itj}, U_{itj})'$ are generated through the panel dependence structure

$$X_{itj} = w_\alpha \alpha_i^x + w_\gamma \gamma_t^x + w_\varepsilon \varepsilon_{it}^x + w_\eta \eta_{itj}^x \qquad \text{and}$$
$$U_{itj} = w_\alpha \alpha_i^u + w_\gamma \gamma_t^u + w_\varepsilon \varepsilon_{it}^u + w_\eta \eta_{itj}^u.$$

The error terms, ε_{it}^x, ε_{it}^u, η_{itj}^x and η_{itj}^u, are independently drawn from the standard normal distribution. All the other settings remain the same as in Section 5 of the main text. We compare our estimator CHS with EHW, CRi, CRt, CGM, and T. In this subsection, we omit the bootstrap methods (MNW and M) due to their long computational time.

Table 4 reports simulation results. Reported values are the coverage frequencies for the slope parameter β_1 for the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show coverage probability results under the clustered-at-the-intersection-level design (i.e., $(w_\alpha, w_\gamma, w_\varepsilon, w_\eta) = (0.00, 0.00, 0.25, 0.25)$) and the dependence design (i.e., $(w_\alpha, w_\gamma, w_\varepsilon, w_\eta) = (0.1, 0.2, 0.1, 0.1)$), respectively. In each group of three consecutive rows, the panel sample sizes (N, T) vary by rows. Cells are shaded based on the proximity of the simulated coverage probability to the nominal probability of 0.95; the darker shades indicate more correct coverage.

Observe that we have similar qualitative patterns in these results to those presented for the baseline simulation studies presented in Section 5. We make one remark regarding the over-sized results for EHW in the clustered-at-the-intersection-level design. Observe that our design entails dependence within each cluster cell (i,t), although the clustered-at-the-intersection-level design imposes independence across the cluster cells. As such, there are non-trivial dependence that causes the over-sized results for EHW.

I.4 Simulations for the Two-Way Fixed-Effect Estimator with the Presence of Multiple Observations in A Cluster Intersection

In this section, we present yet additional simulation studies based on the two-way fixed-effect estimator for fixed-effect models, as in Appendix I.2, allowing for multiple observations in each (i,t) intersection.

For each i and t, we independently draw the number J_{it} of observations from Binomial$(5, 0.5)$. Consider the following data generating process.

$$Y_{itj} = \beta_0 + \beta_1 X_{itj} + U_{itj},$$

where the right-hand side variables $(X_{itj}, U_{itj})'$ are generated through the panel dependence structure

$$X_{itj} = w_1\alpha_{i1}\gamma_{t2} + w_2\alpha_{i2}\gamma_{t1} + w_3\varepsilon_{it0} + w_4\eta_{itj0} \qquad \text{and}$$

$$U_{itj} = w_5\alpha_{i0} + w_6\gamma_{t0} + w_7\alpha_{i1}\gamma_{t3} + w_8\alpha_{i3}\gamma_{t1} + w_9\varepsilon_{it1} + w_{10}\eta_{itj1}.$$

The error terms, ε_{it0}, ε_{it1}, η_{itj0}, and η_{itj1}, are independently drawn from the standard normal distribution. All the other settings remain the same as in Appendix I.2 of the main text. We compare our estimator CHS with EHW, CRi, CRt, CGM, and T. In this subsection, we omit the bootstrap methods (MNW and M) due to very long computational time of them.

Table 5 reports simulation results. Reported values are the coverage frequencies for the slope parameter β_1 for the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show coverage probability results under the clustered-at-the-intersection-level design (i.e., $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}) = (0, 0, 0.25, 0.25, 0, 0, 0, 0, 0.25, 0.25)$) and the dependence

design (i.e., $(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}) = (0.1, 0.1, 0.05, 0.05, 0.1, 0.2, 0.1, 0.1, 0.05, 0.05)$), respectively. In each group of three consecutive rows, the panel sample sizes (N, T) vary by rows. Cells are shaded based on the proximity of the simulated coverage probability to the nominal probability of 0.95; the darker shades indicate more correct coverage. Observe that we have similar qualitative patterns in these results to those presented in Appendix I.3.

References

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 817–858.

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

Arora, A., S. Belenzon, and L. Sheer (2021): "Knowledge spillovers and corporate investment in scientific research," *American Economic Review*, 111, 871–98.

Bai, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77, 1229–1279.

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying technology spillovers and product market rivalry," *Econometrica*, 81, 1347–1393.

Cameron, C. A., J. B. Gelbach, and D. L. Miller (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Chen, K. and T. J. Vogelsang (2023): "Fixed-b Asymptotics for Panel Models with Two-Way Clustering," *arXiv preprint arXiv:2309.08707*.

Chernozhukov, V., D. Chetverikov, and K. Kato (2019): "Inference on causal and structural parameters using many moment inequalities," *Review of Economic Studies*, 86, 1867–1900.

Chiang, H. D., B. E. Hansen, and Y. Sasaki (2022): "Standard errors for two-way clustering with serially correlated time effects," *arXiv preprint arXiv:2201.11304*.

Chiang, H. D., K. Kato, and Y. Sasaki (2023): "Inference for high-dimensional exchangeable arrays," *Journal of the American Statistical Association*, 118, 1595–1605.

Davezies, L., X. D'Haultfœuille, and Y. Guyonvarch (2018): "Asymptotic Results under Multiway Clustering," ArXiv:1807.07925.

——— (2021): "Empirical process results for exchangeable arrays," *Annals of Statistics*, 49, 845–862.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, OUP Oxford.

Dehling, H. and M. Wendler (2010): "Central limit theorem and the bootstrap for U-statistics of strongly mixing data," *Journal of Multivariate Analysis*, 101, 126–137.

Driscoll, J. C. and A. C. Kraay (1998): "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics*, 80, 549–560.

Gagliardini, P., E. Ossola, and O. Scaillet (2016): "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 84, 985–1046.

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

Griliches, Z. (1981): "Market value, R&D, and patents," *Economics Letters*, 7, 183–187.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005): "Market value and patent citations," *RAND Journal of Economics*, 16–38.

Hansen, B. E. (1992): "Consistent covariance matrix estimation for dependent heterogeneous processes," *Econometrica*, 967–972.

——— (2022): *Econometrics*, Princeton University Press.

Hidalgo, J. and M. Schafgans (2021): "Inference without smoothing for large panels with cross-sectional and temporal dependence," *Journal of Econometrics*, 223, 125–160.

Jenish, N. and I. R. Prucha (2009): "Central limit theorems and uniform laws of large numbers for arrays of random fields," *Journal of Econometrics*, 150, 86–98.

Juodis, A. (2021): "This shock is different: Estimation and inference in misspecified two-way fixed effects panel regressions," Tech. rep., Working Paper.

Kallenberg, O. (2006): *Probabilistic Symmetries and Invariance Principles*, Springer Science & Business Media.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR inference: Recommendations for practice," *Journal of Business & Economic Statistics*, 36, 541–559.

Liang, K.-Y. and S. L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lu, X. and L. Su (2022): "Uniform inference in linear panel data models with two-dimensional heterogeneity," *Journal of Econometrics*.

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

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

Newey, W. K. and K. D. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix," *Econometrica*, 55, 703–708.

Petersen, M. A. (2009): "Estimating standard errors in finance panel data sets: Comparing approaches," *Review of Financial Studies*, 22, 435–480.

Stock, J. H. and M. W. Watson (2020): *Introduction to Econometrics*, New York: Pearson, 4 ed.

Thompson, S. B. (2011): "Simple formulas for standard errors that cluster by both firm and time," *Journal of Financial Economics*, 99, 1–10.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Verdier, V. (2020): "Estimation and inference for linear models with two-way fixed effects and sparsely matched data," *Review of Economics and Statistics*, 102, 1–16.

Vogelsang, T. J. (2012): "Heteroskedasticity, Autocorrelation, and Spatial Correlation Robust Inference in Linear Panel Models with Fixed-Fffects," *Journal of Econometrics*, 166, 303–319.

White, H. (1984): *Asymptotic Theory for Econometricians*, Academic Press.

Yoshihara, K.-i. (1976): "Limiting behavior of U-statistics for stationary, absolutely regular processes," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35, 237–252.

Yu, B. (1994): "Rates of convergence for empirical processes of stationary mixing sequences," *Annals of Probability*, 94–116.

Clustered-at-The-Intersection-Level Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	T	CHS
(I)	50	100	—	0.894	0.944	0.947	0.939	0.920	0.935
(II)	75	75	—	0.886	0.942	0.941	0.935	0.912	0.929
(III)	100	50	—	0.895	0.946	0.943	0.938	0.898	0.928

Dependence Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	T	CHS
(IV)	50	100	0.25	0.222	0.493	0.905	0.920	0.917	0.924
(V)	75	75	0.25	0.180	0.388	0.910	0.918	0.911	0.924
(VI)	100	50	0.25	0.161	0.310	0.904	0.908	0.881	0.909
(VII)	50	100	0.50	0.188	0.435	0.840	0.862	0.911	0.904
(VIII)	75	75	0.50	0.148	0.335	0.847	0.859	0.900	0.898
(IX)	100	50	0.50	0.139	0.269	0.842	0.848	0.875	0.882
(X)	50	100	0.75	0.134	0.328	0.675	0.703	0.869	0.861
(XI)	75	75	0.75	0.116	0.262	0.678	0.695	0.849	0.835
(XII)	100	50	0.75	0.102	0.215	0.681	0.691	0.827	0.807

Table 4: Coverage probabilities for the slope parameter β_1 for the OLS with the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show results under the clustered-at-the-intersection-level design and dependence design, respectively. The sample size is indicated by (N, T). In addition, there are multiple observations in each (i, t) cell, where the number is independently drawn from Binomial$(5, 0.5)$. The parameter ρ indicates the AR coefficient in the dependence design. EHW stands for Eicker–Huber–White, CRi stands for cluster robust within i, CRt stands for cluster robust within t, CGM stands for Cameron-Gelbach-Miller, T stands for Thompson, and CHS stands for Chiang-Hansen-Sasaki.

Clustered-at-The-Intersection-Level Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	T	CHS
(I)	50	100	—	0.888	0.936	0.944	0.935	0.922	0.930
(II)	75	75	—	0.887	0.943	0.943	0.938	0.914	0.935
(III)	100	50	—	0.888	0.946	0.938	0.936	0.896	0.927

Dependence Design: Nominal Probability = 95%

	N	T	ρ	EHW	CRi	CRt	CGM	T	CHS
(IV)	50	100	0.25	0.237	0.853	0.725	0.925	0.920	0.928
(V)	75	75	0.25	0.234	0.806	0.799	0.930	0.923	0.931
(VI)	100	50	0.25	0.233	0.717	0.840	0.915	0.901	0.917
(VII)	50	100	0.50	0.219	0.822	0.687	0.904	0.911	0.921
(VIII)	75	75	0.50	0.208	0.764	0.753	0.897	0.916	0.916
(IX)	100	50	0.50	0.209	0.659	0.784	0.877	0.893	0.898
(X)	50	100	0.75	0.180	0.750	0.584	0.835	0.873	0.899
(XI)	75	75	0.75	0.170	0.664	0.628	0.800	0.883	0.882
(XII)	100	50	0.75	0.156	0.553	0.640	0.757	0.851	0.844

Table 5: Coverage probabilities for the slope parameter β_1 for the two-way fixed-effect estimator with the nominal probability of 95% based on 10,000 Monte Carlo iterations. The top and bottom panels show results under the clustered-at-the-intersection-level design and dependence design, respectively. The sample size is indicated by (N, T). In addition, there are multiple observations in each (i, t) cell, where the number is independently drawn from Binomial$(5, 0.5)$. The parameter ρ indicates the AR coefficient in the dependence design. EHW stands for Eicker–Huber–White, CRi stands for cluster robust within i, CRt stands for cluster robust within t, CGM stands for Cameron-Gelbach-Miller, T stands for Thompson, and CHS stands for Chiang-Hansen-Sasaki.