Composite Quasi-Maximum Likelihood Estimation of Dynamic Panels with Group-Specific Heterogeneity and Spatially Dependent Errors

Chu, Ba

Carleton University

2017

Composite Quasi-Maximum Likelihood Estimation of Dynamic Panels with Group-Specific Heterogeneity and Spatially Dependent Errors*

Ba Chu[†]
Carleton University

May 29, 2017

*I would like to thank Professor Ron P. Smith, Professor Lanh T. Tran, Professor Robert de Jong, Francesco Bravo, Vadim Marmer, Kevin Song, Hiro Kasahara, Paul Schrimpf, Matt Webb, Raul Razo-Garcia, and participants of various seminars and conferences for insightful discussions and suggestions. I am indebted to Professor Pham Dinh Tao, Professor Le Thi Hoai An, and Ta Minh Thuy for suggesting me useful references for the DC programming and DCA, which was successfully implemented in this paper. Special thanks go to Joshua Brault for his excellent research assistance. I am also grateful to Andrew Schoenrock and Sylvain Pitre at Carleton Research Computing Services for their technical support during the courses of this project. All possible errors remain mine. Computations were partly made on the supercomputer *Mammouth Parallèle 2* from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

[†]Department of Economics, Carleton University, B-857 Loeb Building, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada. Phone: +1 (613) 520 2600 (ext. 1546). Fax: +1 (613) 520 3906. E-mail: ba_chu@carleton.ca. Web Page: http://http-server.carleton.ca/~bchu/.

Abstract

This paper proposes a new method to estimate dynamic panel data models with spatially dependent errors that allows for known/unknown group-specific patterns of slope heterogeneity. Analysis of this model is conducted in the framework of composite quasi-likelihood (CL) maximization. The proposed CL estimator is robust against some misspecification of the unobserved individual/group-specific fixed effects. Since our CL method is based on the idea of doing regressions involving common-group stochastic trends, no endogeneity problem will arise. Therefore, unlike existing methods the proposed estimator does not require the use of intrumental variables nor bias correction/reduction. Clustering and estimation of the parameters of interest involve a large-scale non-convex mixed-integer programming problem, which can then be solved via a new efficient approach developed based on DC (Difference-of-Convex functions) programming and the DCA (DC algorithm). Suppose that the number of time periods and the size of spatial domain grow simultaneously, asymptotic theory is derived for both cases where the covariates are stationary and nonstationary. An extensive Monte Carlo simulation is also provided to examine the finite-sample performance of the proposed estimator. Our method is then applied to study the long-run relationship between saving and investment rates. The empirical findings reconcile various empirical approaches to capital mobility in the literature; and there exists substantial capital mobility in some countries while no conclusion about capital mobility can be drawn in other countries. Applied economists can easily implement the method by using the companion software to this paper.

Keywords: Large dynamic panels, spatial data, group-specific heterogeneity, clustering, asymptotics, large-scale non-convex mixed-integer program, difference of convex (d.c.) functions, DCA, Variable Neighborhood Search (VNS)

1 Introduction

This paper proposes a new method for estimation and inference of dynamic panel data models with unobserved group-specific patterns of slope heterogeneity and spatially dependent errors. Unobserved heterogeneity and spatial dependence across individuals/units have been the main focus of many econometric papers in panel data, and been well motivated from empirical economic problems, for example, in recent studies of empirical growth [see, e.g., Durlauf, Johnson, and Temple (2005), Corrado, Martin, and Weeks (2005), Meliciani and Peracchi (2006), Alexiadis (2013), Durlauf and Quah (1999), Phillips and Sul (2007, 2009)]

A panel model with grouped heterogeneity in the slopes represents a viable approach to summarize grouped data as it is a compromise between a parsimonious model and one with too many parameters. With data clustered in units, one can estimate three different models. In the first model, one can ignore the grouped structure in the units and estimate a regression with the data pooled. The estimates from this 'pooled' model will be biased if the units differ much, but with the pooled data, the model will become the most parsimonious in terms of the number of parameters estimated. At the other extreme, one could estimate one regression model for each unit, then take the average of all the estimated slope parameters if these parameters vary randomly around a constant - this approach is called the mean-group estimator [see Pesaran and Smith (1995); Pesaran, Smith, and Im (1996) and Fotheringham, Charlton, and Brunsdon. (1997)]. Pesaran, Shin, and Smith (1999) also propose the pooled mean group estimator for autoregressive distributed lag models (ARDL) that allow for both common parameters and heterogeneous parameters. However, this option produces a way more parameters, and the estimates of the slope parameters will be highly variable if there are not many observations for each unit. The grouped slope heterogeneity approach represents a middle ground between these two extremes, thus it can be viewed as a compromise between completely ignoring the structure of the data and fully taking this structure into account by estimating many different models.

To be specific, a simple linear spatial-error specification with dynamic grouped patterns of heterogeneity takes the following form as a special case of a general ARDL model defined by (2.2) in Section 2:

$$\Delta y_{i,t} = \mu_i + \phi_{g_i}\left(y_{i,t-1} - \boldsymbol{\theta}_{g_i}^\top \boldsymbol{x}_{i,t}\right) + \lambda^*_{g_i}\Delta y_{i,t-1} + \boldsymbol{\delta}^{*\,\top}_{g_i}\Delta \boldsymbol{x}_{i,t-1} + \epsilon_{i,t},\ i = 1,\dots,N,\ t = 1,\dots,T, \quad (1.1)$$

where g_i represents a group assignment that assigns each individual, i, to some specific group, say $g_i \in \{1,\dots,G\}$; here G is the number of groups to be specified a priori; μ_i, $i = 1,\dots,N$, are individual-specific fixed effects; ϕ_{g_i}, $\boldsymbol{\theta}_{g_i}$, $\lambda^*_{g_i}$, and $\boldsymbol{\delta}^*_{g_i}$, $i = 1,\dots,N$, are common-group slope parameters; the explanatory variables $\boldsymbol{x}_{i,t}$ are contemporaneously independent of $\epsilon_{i,t}$; moreover, $\epsilon_{i,t}$, $i \in \{1,\dots,N\}$ and $t \in \{1,\dots,T\}$, are identically distributed over space and time; for every given $i \in \{1,\dots,N\}$, $\epsilon_{i,t}$ are contemporaneously independent, and for every given $t \in \{1,\dots,T\}$, $\epsilon_{i,t}$ are spatially dependent [across locations], which effectively implies that $\epsilon_{i,t}$ and $\epsilon_{j,s}$, $t \neq s$, are independent if i and j are associated with different locations - because if $\epsilon_{i,t}$ and $\epsilon_{j,s}$ are dependent, then $\epsilon_{j,s}$ and $\epsilon_{j,t}$ are also dependent as $\epsilon_{i,t}$ are spatially dependent, this indeed leads to a contradiction.

As shown in Section 3 the proposed estimation procedure does not require any particular pattern of spatial dependence to be specified for the error terms; it merely assumes that the innovation terms $\epsilon_{i,t}$, $i = 1,\dots,N$ and $t = 1,\dots,T$, behave in such a manner that $\sqrt{N}\epsilon_{*,t} \sim N\left(0, \sigma^2_{\epsilon,N}\right)$ independently over time periods, where $\epsilon_{*,t} \doteq \frac{1}{N}\sum_{i=1}^N \epsilon_{i,t}$ and $\sigma^2_{\epsilon,N} = \frac{1}{N}Var\left(\sum_{i=1}^N \epsilon_{i,t}\right) < \infty$ as N becomes sufficiently large under some mixing assumption about the random field $\epsilon_{i,t}$, whence the composite quasi-likelihood function can then be constructed. It is worth noting at this point that, in this estimation procedure, the normalized variance, $\sigma^2_{\epsilon,N}$, of the spatial sum of errors and the average fixed effect $\mu_* = \frac{1}{N}\sum_{i=1}^{N}\mu_{g_i}$ are treated as nuisance parameters. While μ_* is estimated by the maximum composite likelihood, $\sigma^2_{\epsilon,N}$ can also be estimated directly by using robust ('clustered') standard errors formulas (see, e.g., Arellano (1987); Conley (1999); Driscoll and Kraay (1998); Kelejian and Prucha (2007)).

Heuristics

Since the parameters are common within each group, say g, the $y_{i,t}$'s and $\boldsymbol{x}_{i,t}$'s of units within the group g all have a common regression relationship, so are their common-group stochastic trends. To estimate the common slope parameters within the group g, one could just regress the common-group stochastic trend of all $\Delta y_{i,t}$'s in this group on its lags and the common-group stochastic trend of all $\boldsymbol{x}_{i,t}$'s in the same group and their lags. As in Pesaran (2006), these latent common-group stochastic trends can be proxied by common-group cross-sectional averages. Importantly the regressions involving common-group cross-sectional averages do not induce an endogeneity problem which is often the consequence of doing the within-group or time-differencing transformations in dynamic panel data models. Thus the estimates will be asymptotically unbiased even for T is less than N. This intuition will be elucidated in Section 3, and formalized in Section 4.

For latent underlying group structures, estimates of the group memberships and the associated common-group slope parameters can be obtained in principle by running many regressions involving common-group cross-sectional averages for each partitioning of the set $[1, \ldots, N]$ into G groups, then choose the parameter values associated with the regression that achieves the minimum sum of squared residuals amongst all the partitionings. However the number of regressions to run will be very large if N is large (in fact, it is equal to a Sterling number of the second kind); this renders the so-called 'many-regressions' method infeasible. However, this method of running many regressions can be casted into a non-convex mixed integer programming problem as described in Section 5.

Relation to the Existing Works

Hahn and Moon (2010) and Bester and Hansen (2016) show that the bias of grouped fixed effects (GFE) estimators asymptotically vanishes in nonlinear panel data models with finitely supported fixed effects (i.e., individual-specific fixed effects are common with each group, and differ across groups). The GFEs can be severely biased when individual specific heterogeneity is incorrectly assumed to be constant within each group. Bonhomme, Lamadon, and Manresa's (2016) method to discretize unobserved fixed effects can only reduce the bias of GFE estimators when the number of groups is allowed to grow with the number of individuals. Therefore, for the GFEs to be asymptotically unbiased and normal, Bester and Hansen (2016) rely on the assumption that the maximum discrepancy between two individuals within groups goes to zero as the number of cross-sections becomes large.

In a typical dynamic linear panel, our proposed composite likelihood estimator does not suffer from this type of bias arising due to misspecification of individual specific heterogeneity because -

unlike GFE estimators which require the individual/group-specific fixed effects to be concentrated out prior to estimation of the parameters of interest - the current estimation paradigm involves only the average fixed effect μ_* instead. Moreover, it is worth noting at this point that, since a within-group transformation can cause endogeneity if lagged dependent variables are included, the proposed estimator does not rely on within-group transformation, thus it also does not suffer from an endogeneity bias. Therefore, instrumental variables (IV's) or bias correction are not required to implement our method. In a dynamic panel with long time horizon the IV estimation strategy may not be feasible as the number of lagged variables that can be used as IV's is large, thus another issue related to choice of optimal IV's needs to be dealt with. Bias correction/reduction methods (see, e.g., Hahn and Kuersteiner (2002) and Dhaene and Jochmans (2015)) require preliminary estimators of the fixed effects for estimates of the bias, thus a misspecification in the fixed effects can deteriorate the quality of bias estimates.

Works on panel data models with unknown patterns of group heterogeneity are pretty recent.[1] Su, Shi, and Phillips (2016) propose a new variant of Tibshirani's (1996) LASSO, namely classifier-LASSO, to perform group classification and estimation of regression slope coefficients simultaneously in a single step. However, this estimator often induces non-negligible asymptotic bias when it is applied to dynamic panels or panel regressions where some regressors are endogenous and the time horizon T is smaller, thus bias corrections of Hahn and Kuersteiner's (2002) type are needed. Wang, Phillips, and Su (2016) propose a penalized least-squares criterion function using a new hybrid Panel-CARDS penalty function for simultaneous classification and estimation, effectively extending Ke, Fan, and Wu's (2015) CARDS procedure for cross-sectional data to panel data.

Lin and Ng (2012) propose a conditional K-means procedure, which extends Forgy's (1965) K-means algorithm, to estimate linear panel data models, but asymptotic theory is not derived. Bonhomme and Manresa (2015) propose two minimum sum-of-squares clustering algorithms based on the K-means algorithm and Hansen and Mladenović's (1997) Variable Neighborhood Search (VNS) algorithm to perform group classification and estimation in panels with time-variant grouped patterns of heterogeneity. In their asymptotic theory the GFE estimators are not influenced by the effect of group membership estimation because the probability of misclassifying at least one individual unit decays very fast as long as both N and T go to infinity such that $N/T^{\delta} \downarrow 0$ for some $\delta > 0$. When a lagged dependent variable is included as a covariate in a model with additive time-invariant individual fixed-effects in addition to the time-varying grouped effects the infeasible fixed-effects estimator suffers from the incidental parameters problem (Nickell (1981)); IVs are then needed to produce consistent estimates for the parameters of interest. They also demonstrate

[1] In a related thematic approach, finite mixture models can be employed to model the probability that an individual belongs to a group. Thus, estimation of and inference on these membership probabilities can be performed via the mixture parameters (see, e.g., Kalai, Moitra, and Valiant (2010); Kasahara and Shimotsu (2009); Sun (2005))

that their proposed algorithms can achieve approximately the same optimal solutions to the least-squares clustering problem as other global optimizing algorithms (for example, the *branch and bound* algorithm) in panel datasets with a small number of groups.

Ando and Bai (2016) deal with linear panel data models with grouped factor structure and a large number of explanatory variables. The group membership of each individual can be estimated along with other parameters of the model. A LASSO approach is applied to select significant explanatory variables, and optimal group memberships can be found by using the K-means algorithm.

Nonlinear panel data models with discretized fixed effects are considered in Bonhomme, Lamadon, and Manresa (2016). Druedahl, Jørgensen, and Kristensen (2016) consider a nonparametric GFE estimator for nonlinear panel data models with finitely supported fixed effects. Vogt and Linton (2017) develop methods to classify nonparametric regression functions into clusters based on the premise that there are groups of individuals who share the same regression function.

It is worth mentioning at this point that, in most of earlier works on this topic, units are often cross-sectionally independent - this is somehow a unrealistic assumption. Therefore, group classification is done in a purely data-driven manner by minimizing some unpenalized/penalized sum-of-squared-errors objective function. The proposed procedure is based on the premise that units have common-group stochastic trends, thus it is natural to let the innovation terms of dynamic panel data models have some weak cross-sectional dependence. Due to the presence of these cross-sectionally dependent errors the growth rate of N relative to T that is required for asymptotic normality and the 'oracle' property of group membership estimates will also depend on the degree of weak cross-sectional dependence.

Computational Consideration

In the above-mentioned papers, various clustering techniques (such as K-means or classifier-LASSO) have been employed to partition panel data with latent group patterns while optimizing the associated objective function for estimates of group-specific coefficients. A common feature of these methods is that the problem is nonconvex and often nonsmooth (such as the K-means), thus falling into one of the most difficult areas of the optimization field. The proposed criterion function is also globally non-convex, and minimization of non-convex criterion functions of this type is a NP-hard (Non-deterministic and Polynomial-time hard) problem with possibly many local minima (Garey and Johnson, 1979). Existing methods including the VNS and the K-means algorithm can feasibly search for 'good' local solution while exact solutions are often not known for large datasets with many individuals clustered into many groups; and the K-means often performs poorly when there are outliers in the data.

The proposed method is novel in the sense that the clustering and estimation procedure based on the composite quasi-likelihood function can be formulated in terms of a non-convex mixed-integer programming problem, which can then be efficiently solved by the DC (Difference-of-Convex functions) programming and the DCA (DC Algorithms) as described in Section 5. The DC programming and DCA were developed by Le Thi Hoai An and Pham Dinh Tao (2003); Pham Dinh and Souad (1988); Pham Dinh Tao and Le Thi Hoai An (1998), and have been implemented to successfully solve many large-scale (smooth or non-smooth) non-convex programming problems in various fields, especially in Machine Learning where they often provide global optima and are demonstrated to be more robust and efficient than the standard methods including the K-means algorithm (see, e.g., Le Thi Hoai An, Belghiti, and Pham Dinh Tao (2007); Le Thi Hoai An, Le Hoai Minh, and Pham Dinh Tao (2014); Liu and Shen (2006) and references therein). Interested readers are referred to Le Thi Hoai An (2014); Pham Dinh Tao and Le Thi Hoai An (1997) for some background and rationale behind the DCA.

Outline

The remaining of this paper is organized as follows. Section 2 introduces the model and main assumptions leading to the formulation of the maximum composite likelihood estimation. Section 3 explains the main CL estimation method for dynamic panel data models with group-specific heterogeneity where the group structure can be left unspecified. The asymptotic properties of the proposed estimator are presented in Section 4. It is worth noting at this point that the proof of consistency for the estimators based on the VNS-DCA clustering relies on the properties of the DC program with combinatorial constraints and the Karush-Kuhn-Tucker (KKT) conditions for local optima. When the covariates are stationary the estimates of the true coefficients ϕ_{0,g_i} and $\boldsymbol{\theta}_{g_i}$, $i = 1, \dots, N$, converge in distribution to normal random variables at rate $\sqrt{NT}$; this rate of convergence is the same as the rate that one could obtain when the parameters are homogenous, which is not a surprise as the number of groups remains fixed for any sufficiently large N.

When the covariates are highly persistent the rate of the distributional convergence pertaining to the long-run coefficient $\boldsymbol{\theta}_{g_i}$ is T (instead of $T\sqrt{N}$ as one would expect), which is essentially similar to the rate achieved with fixed N in Pesaran, Shin, and Smith (1999). This slow rate of convergence is possibly the price that one has to pay for having an estimator free from any parametric specification of the cross-sectional variance-covariance matrix. Derivation of the asymptotic theory is based on the premise that the spatial domain V_N and time grow to infinity jointly such that $|V_N|/T$ converges at a rate depending on the polynomial decay rate of the mixing coefficient.

Section 5 provides a detailed description of the main VNS-DCA algorithm and the derivations of the DC program used for clustering and estimation as a one-step procedure. Section 6 contains some information-based criteria for selecting the optimal partitioning and the optimal number of groups. A summary of the simulation study examining the performance of the proposed estimator in finite samples is contained in Section 7. Overall, it was found that, as long as the stability condition [cf. Assumption 4.2 in Section 2 below] holds, the estimator provides relatively small finite-sample biases and mean squared errors for a variety of sample sizes and spatio-temporal error processes. An empirical study of Feldstein and Horioka's (1980) saving-investment puzzle is provided in Section 8, reconciling previous empirical findings about the long-run correlation of the saving and investment rates. Interested practitioners can find the software package to implement the proposed algorithms for clustering and estimation in Section 10. Finally, results of technical flavour but essential for the theoretical justification of the proposed estimation procedure are given in two appendices at the end of the paper.

Some following notations are commonly used: vectors/matrices are written in boldface; $\|\cdot\|$ denotes the Euclidean norm; $\mathbb{I}_d$ represents the identity matrix of dimension d; $\lambda_{\min}(\boldsymbol{A})$ denotes the minimum eigenvalue of $\boldsymbol{A}$; ι_n is a $n \times 1$ column vector of ones and $\mathbb{I}_n$ is a $n \times n$ identity matrix; for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, let $< \boldsymbol{x}, \boldsymbol{y} >$ represent the scalar product of $\boldsymbol{x}$ and $\boldsymbol{y}$; $|V|$ is the cardinality of a set, V; the Euclidean distance between two subsets, A and B, is defined as $d(A, B) = \min\{\|\boldsymbol{a} - \boldsymbol{b}\| : \boldsymbol{a} \in A, \ \boldsymbol{b} \in B\}$; the diameter of a set, A, is denoted by $\mathrm{diam}(A) = \max\{\|\boldsymbol{a} - \boldsymbol{b}\| : \ \boldsymbol{a}, \boldsymbol{b} \in A\}$; A^c denotes the complement of a set, A; $A \backslash B = \{\boldsymbol{s} : \ \boldsymbol{s} \in A \text{ and } \boldsymbol{s} \notin B\}$; C_0 represents a generic constant that may vary from one equation to another; $\lfloor x \rfloor$ stands for the integer part of a (rational) number; $\mathbf{1}(A)$ denotes a characteristic function that takes value 1 if A is true and 0 otherwise; $\xrightarrow{d}$, $\xrightarrow{w}$, and $\xrightarrow{p}$ in order signify the distributional convergence, the weak convergence, and the convergence in probability; $o_p(\cdot)$ and $O_p(\cdot)$ are standard symbols for stochastic orders of magnitude; $\|X\|_\gamma = (E[|X|^\gamma])^{1/\gamma}$ denotes the Hölder norm; $(a, b)^+ = \max(a, b)$ and $(a, b)^- = \min(a, b)$; w.p.1 stands for "with probability approaching 1"; $\mathrm{vec}(A)$ denotes the vectorization of a matrix, A;

2 Model and Assumptions

Consider the following autoregressive distributed lag model for a panel data observed on T time periods, $t = 1, \ldots, T$, and N individuals (units), $i = 1, \ldots, N$:

$$y_{i,t} = \sum_{j=1}^{p} \lambda_{g_i,j} y_{i,t-j} + \sum_{j=0}^{q} \boldsymbol{\delta}_{g_i,j}^{\top} \boldsymbol{x}_{i,t-j} + \mu_i + \epsilon_{i,t}, \tag{2.1}$$

where the d_x covariates ($\boldsymbol{x}_{i,t}$) and the p lags of $y_{i,t}$ (viz. $y_{i,t-1}, \ldots, y_{i,t-p}$) are contemporaneously uncorrelated with the errors $\epsilon_{i,t}$; $\lambda_{g_i,j}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, p$ and $\boldsymbol{\delta}_{g_i,j}$ for $i = 1, \ldots, N$ and $j = 0, \ldots, q$ are group-specific autoregression and regression coefficients respectively. Units are divided into G mutually exclusive, exhaustive groups; and the group membership variables $g_i \in \{1, \ldots, G\}$ are defined via an onto mapping $g : \ \{1, \ldots, N\} \to \{1, \ldots, G\}$. To study the potential long-run relationship between $y_{i,t}$ and $\boldsymbol{x}_{i,t}$ within each group, we rewrite (2.1) in the following error-correction form:

$$\Delta y_{i,t} = \mu_i + \phi_{g_i}(y_{i,t-1} - \boldsymbol{\theta}_{g_i}^\top \boldsymbol{x}_{i,t}) + \sum_{j=1}^{p-1} \lambda_{g_i,j}^* \Delta y_{i,t-j} + \sum_{j=0}^{q-1} \boldsymbol{\delta}_{g_i,j}^{*\top} \Delta \boldsymbol{x}_{i,t-j} + \epsilon_{i,t}, \tag{2.2}$$

where $\phi_{g_i} = -\left(1 - \sum_{j=1}^{p} \lambda_{g_i,j}\right)$, $\boldsymbol{\theta}_{g_i} = -\frac{\sum_{j=0}^{q} \boldsymbol{\delta}_{g_i,j}}{\phi_{g_i}}$, $\lambda_{g_i,j}^* = -\sum_{m=j+1}^{p} \lambda_{g_i,m}$ for $j = 1, \ldots, p-1$, and $\boldsymbol{\delta}_{g_i,j}^* = -\sum_{m=j+1}^{q} \boldsymbol{\delta}_{g_i,m}$ for $j = 1, \ldots, q-1$.

Suppose that each unit ,i, is associated with a location, say s_i, on a d_v-dimensional Euclidean space, V_N, equipped with an Euclidean metric, $\|\cdot\|$, measuring the distance between any two locations in V_N. Here, for clarity of exposition, V_N is assumed to be a sublattice, of the standard d_v-dimensional integer lattice $\mathbb{Z}^{d_v}$, indexed by N; the other cases where V_N is some sublattice of $\mathbb{R}^{d_v}$ follow similarly as long as the distance between any two points in V_N is greater than or equal to one (see, e.g., Jenish and Prucha (2012)).

The variables, $(y_{i,t}, \boldsymbol{x}_{i,t}^\top)^\top$ are spatially dependent at some point in time, t, if their measurements at two different locations depend on each other, and this dependence is assumed to be weaker as the distance between the locations becomes further. For the model to remain parsimonious and tractable, we can allow for spatial dependence in the process $(y_{i,t}, \boldsymbol{x}_{i,t}^\top)^\top$, $t = 1, \ldots, T$, by assuming that, on a specific time period, the errors $\epsilon_{i,t}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, at two different locations are dependent whilst they are independent at different points in time. First, we make the following assumptions:

Assumption 2.1. *The errors, $\epsilon_{i,t}$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$, defined by (2.2), are independent across time and, at some given point in time, they are dependent across locations such that $\epsilon_{s_i,t} \sim N(0, \sigma_{s_i})$.*

It is important to note that the normality of cross-sectional error terms, $\epsilon_{s_i,t}$, $i = 1, \ldots, N$, in Assumption 2.1 could be relaxed when N is sufficiently large since the CLT for strongly mixing random fields (see, e.g., Bulinski and Shashkin (2007)) warrants that $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_{s_i,t}$ converges in distribution to a normal random variable, thus a good approximation to the composite likelihood can be used instead.

Assumption 2.2. *The model (2.1) is stable in that the roots of*

$$\sum_{j=1}^{p} \lambda_{g_i,j} z^j = 1, \quad i = 1, \ldots, N$$

lie outside the unit circle.

Assumption 2.2 is originally employed in Pesaran, Shin, and Smith (1999) to ensure that the order of integration of $y_{i,t}$ is at most equal to the maximum of the orders of integration of the elements of the vector $\boldsymbol{x}_{i,t}$. This condition also warrants the existence of a long-run relationship between $y_{i,t}$ and $\boldsymbol{x}_{i,t}$ within each group. Let $\boldsymbol{w}_{i,t} = (\Delta y_{i,t-1}, \dots, \Delta y_{i,t-p+1}, \Delta\boldsymbol{x}_{i,t}^\top, \dots, \Delta\boldsymbol{x}_{i,t-q+1}^\top)^\top$ denote a vector of $d_w = p + d_x q - 1$ auxiliary covariates, and let $\boldsymbol{\lambda}_{g_i} = (\lambda_{g_i,1}, \dots, \lambda_{g_i,p-1}, \boldsymbol{\delta}_{g_i}^\top, \dots, \boldsymbol{\delta}_{g_i,q-1}^\top)^\top$ be their coefficients. We can rewrite (2.2) as

$$\Delta y_{s_i,t} = \mu_i + \phi_{g_i}\xi_{s_i,t}(\boldsymbol{\theta}_{g_i}) + \boldsymbol{\lambda}_{g_i}^\top \boldsymbol{w}_{s_i,t} + \epsilon_{s_i,t}, \tag{2.3}$$

where $\xi_{s_i,t}(\boldsymbol{\theta}_{g_i}) = y_{s_i,t-1} - \boldsymbol{\theta}_{g_i}^\top \boldsymbol{x}_{s_i,t}$. Our objects for inference are the long-run coefficients $\boldsymbol{\theta}_{g_i}$ and the long-run adjustment speed parameter ϕ_{g_i} with $i = 1, \dots, N$.

It is important to note that, since the joint likelihood of the model is not the same as the product of the likelihoods for each unit (or group), estimation will involve a large unknown variance-spatial covariance matrix of $\epsilon_{s_i,t}$, $i = 1, \dots, N$, thus becomes infeasible. Moreover the expectations of the score functions of the concentrated log-joint likelihood function are not zero due to the absence of the complete orthogonality between $\epsilon_{s_i,t}$ and $\xi_{s_i,t}(\boldsymbol{\theta}_{g_i})$, thus resulting in biases that do not disappear asymptotically. Therefore, we shall instead construct the composite log-likelihood function. To simplify notations, we assume that the nuisance parameters are group-invariant (i.e., $\boldsymbol{\lambda}_{g_1} = \cdots \boldsymbol{\lambda}_{g_N} = \boldsymbol{\lambda}$.) Note that this simplification does not much change our mathematical arguments, thus our asymptotic results will still remain valid even when these nuisance parameters vary over groups. To see this, notice that in the representation of the composite errors (3.2), the projections of $\boldsymbol{x}_{*,t}$'s and $\xi_{*,t}$'s on the span of $\{\boldsymbol{w}_{*,t} : \ t = 1, \dots, T\}$ are of lower asymptotic orders when T goes to infinity and the cluster sizes grow sufficiently large. In fact, in Section 7, simulation results confirm that the algorithm for clustering and estimation based on the objective function imposing group-invariant nuisance parameters performs well even when data are generated from a d.g.p. with group-variant nuisance parameters.

3 Estimation with Known/Unknown Group Memberships

As discussed in the Introduction the panel data model is estimated by using a composite likelihood method. The general principle of composite likelihood methods is to simplify complex dependence relationships by computing marginal or conditional distributions associated with some subsets of data, and multiplying these together to form an inference function. Employing composite likelihood methods can reduce the computational complexity so that it is possible to deal with large datasets and even very complex models where the use of standard likelihood or Bayesian methods is not feasible. Composite likelihood estimators also have good theoretical properties, and behave well in many complex applications (see, e.g., Reid (2013); Varin, Reid, and Firth (2011) for recent reviews of this subject matter.) Following Lindsay (1988), let $\{f(\boldsymbol{y}; \boldsymbol{\theta}), \ \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{\theta} \in \Theta\}$, where $\mathcal{Y} \subset \mathbb{R}^n$ and

$\Theta \subset \mathbb{R}^d$ with $n \geq 1$ and $d \geq 1$, be a parametric model. Consider a set, $\{\mathcal{A}_1, \ldots, \mathcal{A}_k, \ldots, \mathcal{A}_K\}$, of marginal or conditional events associated with likelihoods, $\mathcal{L}_k(\boldsymbol{\theta}; \boldsymbol{y}) \propto f(\boldsymbol{y} \in \mathcal{A}_k; \boldsymbol{\theta})$. A composite likelihood is formally defined as a weighted product $\prod_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\theta}; \boldsymbol{y})^{w_k}$, where w_k, $k = 1, \ldots, K$, represent some non-negative composite weights to be chosen.

We first present the main procedure based on composite likelihood to estimate Model (2.1) when group memberships of individuals/units are given (i.e., each individual belongs to a specified group.) By Assumption 2.1, $\sqrt{N}\epsilon_{*,t} = \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} \epsilon_{s_i,t} \overset{i.i.d.}{\approx} N(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon^2 = \lim_{N\uparrow\infty} \frac{1}{N} Var\left(\sum_{i=1}^{N} \epsilon_{s_i,t}\right) < \infty$ if the spatial dependence among $\epsilon_{s_i,t}$, $i = 1, \ldots, N$, is weak. Therefore, all the likelihoods associated with conditional events, $\mathcal{A}_t(x) \doteq \{(\epsilon_{s_1,t}, \ldots, \epsilon_{s_N,t}) \in \mathbb{R}^T : \sqrt{N}\epsilon_{*,t} = x\}$ with $t = 1, \ldots, T$, are Gaussian.

Let $V_{N,i}$ represent a set of locations for units in group $i \in \{1, \ldots, G\}$ so that $V_N = \bigcup_{i=1}^{G} V_{N,i}$, $L_{N,i} = |V_{N,i}|$, and $g_{*,i} = \frac{L_{N,i}}{N}$ for $i = 1, \ldots, G$. Define $\Delta y_{*,t} = \frac{1}{N}\sum_{i=1}^{N} \Delta y_{s_i,t}$, $\boldsymbol{w}_{*,t} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_{s_i,t}$, $\mu_* = \sum_{i=1}^{G} g_{*,i}\mu_i$, and $\xi_{*,t}(\boldsymbol{\theta}_i) \doteq \xi_{*,t}(\boldsymbol{\theta}_i) = \frac{1}{L_{N,i}}\sum_{\boldsymbol{j}\in V_{N,i}} \xi_{\boldsymbol{j},t}(\boldsymbol{\theta}_i)$, where $\mu_i = \mu_{g(V_{N,i})}$ and $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{g(V_{N,i})}$, $i = 1, \ldots, G$. Collecting all the unknown parameters into a vector, say $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\lambda}^\top, \mu_*)$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_i, \ldots, \phi_G)^\top$ with $\phi_i = \phi_{g(V_{N,i})}$. Setting the composite weights $\{w_t,\ t = 1, \ldots, T\}$ to ones the composite log-likelihood function can then be written as

$$Q_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \mu_*, \sigma_\epsilon^2) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma_\epsilon^2 - \frac{N}{2\sigma_\epsilon^2}\sum_{t=1}^{T} \epsilon_{*,t}^2(\boldsymbol{\Theta}),$$

where $\epsilon_{*,t}(\boldsymbol{\Theta}) = \Delta y_{*,t} - \mu_* - \sum_{i=1}^{G} g_{*,i}\phi_i\xi_{*,t}(\boldsymbol{\theta}_i) - \boldsymbol{\lambda}^\top \boldsymbol{w}_{*,t}$.

Remark 3.1. *Intuitively, while clustering with the least squares criterion function (as in Bonhomme and Manresa (2015)) is based on the premise that - for given values of the parameters - an individual is assigned to a group if its temporal summation of squared errors associated with that group is less than its temporal summations of squared errors associated with all other groups, the CL criterion assigns a subset of individuals, say $\mathfrak{C}$, to a group, say $\mathfrak{g}$, if the temporal summation of squared $\mathfrak{C}$-mean errors (or* centroids *in the language of machine learning) associated with group $\mathfrak{g}$ is less than the temporal summations of squared $\mathfrak{C}$-mean errors associated with all other groups such that the* *mean errors of any pair of groups are as little correlated as possible. To see this point, notice that*

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}^{2}(\boldsymbol{\Theta}) &= \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{c=1}^{G}\frac{1}{N}\sum_{i=1}^{N}u_{i,c}\epsilon_{i,t}(\boldsymbol{\theta}_c)\right)^2 \\
&= \sum_{c=1}^{G}\sum_{t=1}^{T}\underbrace{\left(\frac{1}{N}\sum_{i=1}^{N}u_{i,c}\epsilon_{i,t}(\boldsymbol{\theta}_c)\right)^2}_{\text{squared mean error of group } c} \\
&+ \sum_{c=1}^{G}\sum_{g\neq c}^{G}\underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}u_{i,c}\epsilon_{i,t}(\boldsymbol{\theta}_c)\right)\left(\frac{1}{N}\sum_{i=1}^{N}u_{i,g}\epsilon_{i,t}(\boldsymbol{\theta}_g)\right)}_{\text{correlation between the mean errors of two groups}},
\end{aligned}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$*. When groups are mutually independent, the CL criterion function is the same as the summation of all the sums of squared errors obtained from G regressions of common-group stochastic trends.*

Concentrating out the nuisance parameters $\boldsymbol{\lambda}$ and let $\boldsymbol{\Omega} = (\boldsymbol{\psi}^\top, \sigma_\epsilon^2)^\top$ with $\boldsymbol{\psi} = (\boldsymbol{\theta}^\top, \boldsymbol{\phi}^\top, \mu_*)^\top$, one obtains the concentrated composite log-likelihood function:

$$Q_{N,T}(\boldsymbol{\Omega}) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma_\epsilon^2 - \frac{N}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\epsilon_{*,t}^2(\boldsymbol{\psi}), \tag{3.1}$$

where

$$\begin{aligned}
\epsilon_{*,t}(\boldsymbol{\psi}) = \Delta y_{*,t} &- \left(\sum_{s=1}^{T}\Delta y_{*,s}\boldsymbol{w}_{*,s}^\top\right)\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\boldsymbol{w}_{*,t} \\
&- \sum_{i=1}^{G}g_{*,i}\phi_i\left(\xi_{*,t}(\boldsymbol{\theta}_i) - \left(\sum_{s=1}^{T}\xi_{*,s}^{(i)}(\boldsymbol{\theta}_i)\boldsymbol{w}_{*,s}^\top\right)\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\boldsymbol{w}_{*,t}\right) \\
&- \mu_*\left(1 - \sum_{s=1}^{T}\boldsymbol{w}_{*,s}^\top\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\boldsymbol{w}_{*,t}\right).
\end{aligned} \tag{3.2}$$

To derive the first-order conditions of the log-likelihood maximization, we define the following vectors: $\underbrace{\boldsymbol{A}_t}_{Gd_x\times 1} = \left(\boldsymbol{x}_{*,t}^{(1)\top} - \boldsymbol{w}_{*,t}^\top\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{x}_{*,s}^{(1)\top}, \ldots, \boldsymbol{x}_{*,t}^{(G)\top} - \boldsymbol{w}_{*,t}^\top\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\right.$ $\left.\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{x}_{*,s}^{(G)\top}\right)^\top$, where $\boldsymbol{x}_{*,t}^{(i)} = \frac{1}{L_{N,i}}\sum_{\boldsymbol{j}\in V_{N,i}}\boldsymbol{x}_{\boldsymbol{j},t}$ for $i = 1, \ldots, G$, $\underbrace{\boldsymbol{B}_t(\boldsymbol{\theta})}_{G\times 1} = \left(\left(\sum_{s=1}^{T}\xi_{*,s}^{(1)}(\boldsymbol{\theta}_1)\boldsymbol{w}_{*,s}^\top\right)\right.$

$\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right)^{-1}\boldsymbol{w}_{*,t}-\xi_{*,t}^{(1)}(\boldsymbol{\theta}_1),\ldots,\left(\sum_{s=1}^{T}\xi_{*,s}^{(G)}(\boldsymbol{\theta}_G)\boldsymbol{w}_{*,s}^{\top}\right)\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right)^{-1}\boldsymbol{w}_{*,t}-\xi_{*,t}^{(G)}(\boldsymbol{\theta}_G)\right)^{\top}$,

and $C_t=\sum_{s=1}^{T}\boldsymbol{w}_{*,s}^{\top}\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right)^{-1}\boldsymbol{w}_{*,t}-1$. Some algebraic manipulations yield

$$\frac{\partial Q_{N,T}(\boldsymbol{\Omega})}{\partial\boldsymbol{\theta}} = -diag(g_{*,i}\phi_i\boldsymbol{I}_{d_x},\ i=1,\ldots,G)\frac{N}{\sigma_\epsilon^2}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi})\boldsymbol{A}_t, \tag{3.3}$$

$$\frac{\partial Q_{N,T}(\boldsymbol{\Omega})}{\partial\boldsymbol{\phi}} = -diag(g_{*,i},\ i=1,\ldots,G)\frac{N}{\sigma_\epsilon^2}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi})\boldsymbol{B}_t(\boldsymbol{\theta}), \tag{3.4}$$

$$\frac{\partial Q_{N,T}(\boldsymbol{\Omega})}{\partial\mu_*} = -\frac{N}{\sigma_\epsilon^2}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi})C_t. \tag{3.5}$$

One can now obtain the estimates ($\widetilde{\boldsymbol{\theta}}$, $\widetilde{\boldsymbol{\phi}}$, and $\widetilde{\mu_*}$) of the true parameters ($\boldsymbol{\theta}_0$, $\boldsymbol{\phi}_0$, and μ_{*0}) by finding the roots of (3.3)-(3.5).

Now we shall adapt the composite-likelihood-based procedure described above to the case when group memberships of individuals are not specified a priori. Suppose that the number of groups (*or* clusters) G is given. Let $\boldsymbol{U}=(u_{i,c})\in\mathbb{R}^{G\times N}$, $i=1,\ldots,N$ and $c=1,\ldots,G$, be a $G\times N$ matrix whose elements are defined by $u_{i,c}=1$ if individual $i\in[1,N]$ belongs to group $c\in[1,G]$, and $u_{i,c}=0$ otherwise. Because each individual can only be assigned to one group, we need to impose the constraint $\sum_{c=1}^{G}u_{i,c}=1$ for every $i=1,\ldots,N$. Moreover, let $\Delta_S=\{\boldsymbol{u}\in[0,1]^G:\ \sum_{c=1}^{G}u_c=1\}$ represent the $(G-1)$-simplex in $\mathbb{R}^G$ and Δ_S^N is the Cartesian product of N simplices Δ_S, thus $\boldsymbol{U}\in\Delta_S^N\bigcap\{0,1\}^{G\times N}$. With this matrix of group membership variables, define the composite error as

$$\epsilon_{*,t}(\boldsymbol{\Theta},\boldsymbol{U})=\Delta y_{*,t}-\mu_*-\boldsymbol{\lambda}^{\top}\boldsymbol{w}_{*,t}-\sum_{c=1}^{G}\frac{1}{N}\sum_{i=1}^{N}u_{i,c}\phi_c\xi_{i,t}(\boldsymbol{\theta}_c). \tag{3.6}$$

The composite log-likelihood function is then given by

$$Q_{N,T}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\lambda},\mu_*,\sigma_\epsilon^2,\boldsymbol{U})=-\frac{T}{2}\log 2\pi-\frac{T}{2}\log\sigma_\epsilon^2-\frac{N}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\epsilon_{*,t}^2(\boldsymbol{\Theta},\boldsymbol{U}). \tag{3.7}$$

By concentrating the nuisance parameters $\boldsymbol{\lambda}$ out, one obtains that

$$\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U})=\Delta y_{*,t}^{(w)}-\sum_{c=1}^{G}\frac{1}{N}\sum_{i=1}^{N}u_{i,c}\phi_c\left(y_{i,t-1}^{(w)}-\boldsymbol{\theta}_c^{\top}\boldsymbol{x}_{i,t}^{(w)}\right)-\mu_*1_t^{(w)},$$

where

$$Z_t^{(w)}=Z_t-\left(\sum_{t=1}^{T}Z_t\boldsymbol{w}_{*,t}^{\top}\right)\left(\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^{\top}\right)^{-1}\boldsymbol{w}_{*,t}$$

with Z_t being either $y_{i,t-1}$ or $\boldsymbol{x}_{i,t}$, and

$$1_t^{(w)} = 1 - \left(\sum_{t=1}^{T} \boldsymbol{w}_{*,t}^\top\right)\left(\sum_{t=1}^{T} \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1} \boldsymbol{w}_{*,t}.$$

Moreover, notice that

$$\Delta y_{*,t}^{(w)} = \sum_{c=1}^{G} \frac{1}{N} \sum_{i=1}^{N} u_{0,i,c}\phi_{0,c}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) + \mu_{0,*}1_t^{(w)},$$

where all the subscripts '0' signify the true [unknown] quantities as usual. The objective function $Q_{N,T}$ defined by (3.7) is invariant with respect to all permutations of the group labels; let $\sigma^{(per)}: [1, G] \to P \in \mathcal{P}$ denote a permutation operator, which is a bijective mapping from the set, $[1, G]$, of the original group labels to some set, P, of permuted group labels, and $\mathcal{P}$ is the collection of all the sets of permuted group labels. It then follows that the concentrated composite error $\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U})$ can also be expressed as

$$\begin{aligned}\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) = &\sum_{c=1}^{G} \phi_{\sigma^{(per)}(c)}(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})\frac{1}{N}\sum_{i=1}^{N} u_{i,\sigma^{(per)}(c)}\boldsymbol{x}_{i,t}^{(w)} \\ &+ \sum_{c=1}^{G}(\phi_{0,c} - \phi_{\sigma^{(per)}(c)})\frac{1}{N}\sum_{i=1}^{N} u_{i,\sigma^{(per)}(c)}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) + (\mu_{0,*} - \mu_*)1_t^{(w)} \\ &+ \sum_{c=1}^{G}\phi_{0,c}\frac{1}{N}\sum_{i=1}^{N}(u_{0,i,c} - u_{i,\sigma^{(per)}(c)})\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) + \epsilon_{0,*,t}^{(w)}, \end{aligned} \quad (3.8)$$

where $\epsilon_{0,*,t} = \epsilon_{*,t}(\boldsymbol{\Theta}_0, \boldsymbol{U}_0)$.

The CML estimates $\widehat{\boldsymbol{\psi}}$, $\widehat{\sigma}_\epsilon^2$, and $\widehat{\boldsymbol{U}}$ of $\boldsymbol{\psi}_0$, $\sigma_{\epsilon,0}^2$, and $\boldsymbol{U}_0$ respectively are defined as the solutions to the following large-scale non-convex mixed-integer programming problem:

$$\min\left\{Q_{N,T}(\boldsymbol{\psi}, \sigma_\epsilon^2, \boldsymbol{U}): \ \boldsymbol{\psi} \in \Theta_\psi \subset \mathbb{R}^{G(d_x+1)+1}, \ \sigma_\epsilon^2 \in \Theta_\sigma \subset \mathbb{R}, \text{ and } \boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G\times N}\right\}. \quad (3.9)$$

Intuition behind the proposed CL estimator. Suppose that $y_{i,t}$ and $\boldsymbol{x}_{i,t}$ share a common relationship in Group $c \in [1, \ldots, G]$. The common group stochastic trends that can be reasonably proxied by the observable vector of groupwise cross-sectional averages $(y_{*,t}^{(c)}, \boldsymbol{x}_{*,t}^{(c)})$, $c \in [1, \ldots, G]$, also obey the same relationship, i.e., $\Delta y_{*,t}^{(c)} = \mu_c + \phi_c\xi_{*,t}^{(c)}(\boldsymbol{\theta}_c) + \boldsymbol{\lambda}_c^\top \boldsymbol{w}_{*,t}^{(c)} + \epsilon_{*,t}^{(c)}$. Since $\epsilon_{*,t}^{(c)}$ will be close to zero as the group size is sufficiently large, one needs to blow it up by $\sqrt{N}$ so that $\sqrt{N}\sum_{c=1}^{G}\epsilon_{*,t}^{(c)} \sim N(0, \sigma_\epsilon^2)$. Therefore the CL estimator can be viewed as the minimizer of the temporal average of the squares of the errors from regressions involving the common stochastic trends of $y_{i,t}$ and $\boldsymbol{x}_{i,t}$ in G groups. For given N units, there are many ways to partition these N units into G groups. The estimated group memberships are associated with the group partition that minimizes the sum of squared residuals obtained from G regressions involving common group stochastic trends.

4 Asymptotic Theory

4.1 Known Group Membership

First of all, it is important to note that the parameter spaces (Θ_θ, Θ_ϕ, Θ_μ, and Θ_σ) of $(\boldsymbol{\theta}_0^\top, \boldsymbol{\phi}_0^\top, \mu_{*,0}, \sigma_{\epsilon,0}^2)^\top$ are compact throughout the paper. We study the asymptotic behaviour of $\widetilde{\boldsymbol{\psi}}(=\widetilde{\boldsymbol{\psi}}_{N,T})$ in two different cases. In the first case, it is assumed that, for each $\boldsymbol{i} \in V_N$, $\boldsymbol{x}_{\boldsymbol{i},t}$ is a stationary time series; and in addition the spatio-temporal processes $\{\boldsymbol{x}_{\boldsymbol{j},t} : \boldsymbol{j} \in V_{N,i} \text{ and } t \in [1,T]\}$, $i = 1,\ldots,G$, are mixing and satisfy the following assumption:

Assumption 4.1. *Within each group, i, the random variables $\{\boldsymbol{x}_{\boldsymbol{j},t} : \boldsymbol{j} \in V_{N,i}$ and $t \in [1,T]\}$ are identically distributed across time and space. Moreover,*

(a) *the mixing coefficient - as represented by $\alpha(\cdot)$ in Definition 1 - for $\{(\boldsymbol{x}_t^{(c)}, \epsilon_{\boldsymbol{j},t}) : \boldsymbol{j} \in V_{N,c}$ and $t \in [1,T]\}$, $c = 1,\ldots,G$, where $\boldsymbol{x}_t^{(c)}$ are the common-group stochastic trends of all the $\boldsymbol{x}_{\boldsymbol{j},t}$'s in group c, decays to zero at some rate such that $\alpha(\tau) \le C_\theta \tau^{-\theta_\alpha}$ for some*

$$\theta_\alpha \ge \max\left(\frac{p(d_v+1)\gamma_\eta}{(p-q)(\gamma_\eta-2)} + (d_v+1)\gamma_M, \frac{d_v+1}{1-\frac{2}{\gamma_\eta}}, \frac{2p}{p-4} - \gamma_M\right),$$

where the generic constants $C_\theta > 0$, $\gamma_\eta > 2$, $p > 4$, $q = \frac{2p}{p-2}$, d_v is the dimension of V_N, and $\gamma_M \ge 1$ is given in Definition 1;

(b) $\max\left(E\|\boldsymbol{x}_{\boldsymbol{i},1}\|^p, E\|\boldsymbol{x}_{\boldsymbol{i},1}\|^{\gamma_\eta}, E\|\boldsymbol{x}_{\boldsymbol{i},1}\|^4\right) < \infty$;

(c) $\max\left(|V_N|^{\gamma_M} T^{\gamma_M+1-\theta_\alpha}, |V_N|^{\gamma_M(1-2/\gamma_\eta)} T^{\epsilon-1/2}, T^{(\gamma_M+\theta_\alpha-1)\epsilon-\frac{1}{2}(\theta_\alpha-\gamma_M-1)}|V_N|^{\gamma_M}\right) \downarrow 0$ *for some* $\epsilon \in \left(0, \min\left(\frac{1}{2}, \frac{\theta_\alpha-\gamma_M-1}{2(\gamma_M+\theta_\alpha-1)}\right)\right)$.

Assumption 4.2. *Let $\boldsymbol{X}_{N,T,t}(\boldsymbol{\theta}) = \left(\boldsymbol{x}_{*,t}^{(1)\top}, \ldots, \boldsymbol{x}_{*,t}^{(G)\top}, -\xi_{*,t}^{(1)}(\boldsymbol{\theta}_1), \ldots, -\xi_{*,t}^{(G)}(\boldsymbol{\theta}_G), -1\right)^\top$. The minimum eigenvalue of the non-stochastic limiting matrix, $\boldsymbol{Q}_{zz}$, of $\frac{1}{T}\sum_{t=1}^T \boldsymbol{X}_t(\boldsymbol{\theta}_0)\boldsymbol{X}_t(\boldsymbol{\theta}_0)^\top$ is strictly positive.*

A few remarks are now in order. Condition 4.1(a) imposes a specific degree of weak spatio-temporal dependence on the covariates and the error term. Condition 4.1(b) is rather standard - it requires some moments of the covariate being bounded. Condition 4.1(c) allows both T and N

go to infinity and the divergence speed of N relative to T depends on the structure of the spatial processes and the decay rate of the mixing coefficient. This condition is weaker than the condition [proposed in Hahn and Moon (2010)] that allows N to be some exponential function of T (as such, N needs to be much greater than T) under some common types of weak serial dependence.

We now present the consistency of $\widetilde{\boldsymbol{\psi}}$ for the stationarity case:

Theorem 1 (Consistency)**.** *Suppose that Assumptions 2.1, 2.2, 4.1, and 4.2 hold. Then,* $|\widetilde{\sigma}_\epsilon^2 - \sigma_\epsilon^2| = o_p(1)$ *and* $\|\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\| = o_p\left(N^{-1/2}\right)$.

Theorem 2 (Asymptotic Normality)**.** *Let the conditions for Theorem 1 hold. Then,*

$$\sqrt{NT}(\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \xrightarrow{d} \sigma_\epsilon N\left(\mathbf{0}, [\boldsymbol{D}_{\phi_0}\boldsymbol{D}_g\boldsymbol{Q}_{zz}\boldsymbol{D}_{\phi_0}\boldsymbol{D}_g]^{-1}\right),$$

where $\underbrace{\boldsymbol{D}_\phi}_{(G(d_x+1)+1)\times(G(d_x+1)+1)} = diag(\boldsymbol{\phi}\otimes\mathbb{I}_{d_x}, \mathbb{I}_{G+1})$; $\underbrace{\boldsymbol{D}_g}_{(G(d_x+1)+1)\times(G(d_x+1)+1)} = diag(\boldsymbol{g}\otimes\mathbb{I}_{d_x}, \boldsymbol{g}, 1)$ *with* $\boldsymbol{g} = (g_{*,1}, \ldots, g_{*,G})^\top \in (0,1)^G$.

Remark 4.1. *Since the CL criterion function is nonlinear in the coefficients $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ of the error-correction representation defined by Eq. (2.2), it is not obvious to see the $\sqrt{NT}$-consistency of the CL estimators. To get some intuition about Theorem 2, we consider a linear panel data model with fixed effects and group heterogeneity in the slope coefficient: $y_{i,t} = \alpha_i + \theta_g x_{i,t} + \epsilon_{i,t}$ for all i in group $g \in [1, 2, \ldots, G]$. Define $\alpha_* = \frac{1}{N}\sum_{i=1}^N \alpha_i$, $\overline{x}_g = \frac{1}{T}\sum_{t=1}^T x_{*,t}^{(g)}$, $\overline{y} = \frac{1}{T}\sum_{t=1}^T y_{*,t}$, $\overline{z}_g = \frac{1}{T}\sum_{t=1}^T x_{*,t}^{(g)} y_{*,t}$, and $\overline{x}_{g,c} = \frac{1}{T}\sum_{t=1}^T x_{*,t}^{(g)} x_{*,t}^{(c)}$, where $x_{*,t}^{(c)} = \frac{1}{N}\sum_{i=1}^N u_{i,c}x_{i,t}$ and $y_{*,t} = \frac{1}{N}\sum_{i=1}^N y_{i,t}$. The 'oracle' CL estimate of $\boldsymbol{\psi}_0 = (\alpha_{0,*}, \theta_{0,1}, \ldots, \theta_{0,G})^\top$ is given by*

$$\widetilde{\boldsymbol{\psi}} = \begin{bmatrix} 1 & \overline{x}_1 & \cdots & \overline{x}_G \\ 1 & \overline{x}_{1,1} & \cdots & \overline{x}_{1,g} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \overline{x}_{G,1} & \cdots & \overline{x}_{G,G} \end{bmatrix}^{-1} \begin{bmatrix} \overline{y} \\ \overline{z}_1 \\ \vdots \\ \overline{z}_G \end{bmatrix}.$$

One can then obtain that:

$$\sqrt{NT}\left(\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) = \underbrace{\begin{bmatrix} 1 & \overline{x}_1 & \cdots & \overline{x}_G \\ 1 & \overline{x}_{1,1} & \cdots & \overline{x}_{1,g} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \overline{x}_{G,1} & \cdots & \overline{x}_{G,G} \end{bmatrix}^{-1}}_{\mathcal{A}} \begin{bmatrix} \sqrt{\frac{N}{T}}\sum_{t=1}^T \epsilon_{*,t} \\ \sqrt{\frac{N}{T}}\sum_{t=1}^T x_{*,t}^{(1)}\epsilon_{*,t} \\ \sqrt{\frac{N}{T}}\sum_{t=1}^T x_{*,t}^{(G)}\epsilon_{*,t} \end{bmatrix}, \tag{4.1}$$

where $\epsilon_{*,t} = \frac{1}{N}\sum_{i=1}^N \epsilon_{i,t}$. According to Pesaran (2006), cross-sectional means can well approximate stochastic trends. Therefore, by naively assuming $x_{i,t}$ to have an additive structure: $x_{i,t} = x_{g,t} + x_i$ with $E[x_i] = 0$ for each unit i *in group* g*, one can obtain from law of large numbers that $x_{*,t}^{(g)} \approx x_{g,t}$. Moreover, since $\epsilon_{i,t}$ is a random error, we previously argued that $\sqrt{N}\epsilon_{*,t}$ can be approximated by a normal random variable, say $\mathcal{N}_t$. By applying a central limit theorem, it then follows that*

$\sqrt{\frac{N}{T}}\sum_{t=1}^{T}\epsilon_{*,t}$ and $\sqrt{\frac{N}{T}}\sum_{t=1}^{T}x_{*,t}^{(g)}\epsilon_{*,t}$, $g=1,\ldots,G$, can be approximated by mean-zero normal random variables as long as $x_{*,t}^{(g)}$ and $\epsilon_{*,t}$ are uncorrelated. Since the matrix $\mathcal{A}$ can converge to a finite, positive definite matrix, one then obtains the $\sqrt{NT}$-consistency. The same intuition can carry over to general error-correction models.

In the second case when one assumes that, in each location, $\boldsymbol{i}\in V_N$, $\boldsymbol{x}_{\boldsymbol{i},t}$ is an integrated process of order 1; moreover the spatio-temporal processes $\{\boldsymbol{x}_{\boldsymbol{j},t}:\ \boldsymbol{j}\in V_{N,i} \text{ and } t\in[1,T]\}$, $i=1,\ldots,G$, are heterogeneous across groups. To be precise, we state the following assumption:

Assumption 4.3. *Let $\boldsymbol{x}_{\boldsymbol{i},t}=\sum_{s=1}^{t}\boldsymbol{\eta}_{\boldsymbol{i},s}$, where $\boldsymbol{\eta}_{\boldsymbol{i},s}$ is a mixing centered spatio-temporal process and, within each group, say $i\in[1,G]$ the random variables $\{\boldsymbol{\eta}_{\boldsymbol{j},t},\ \boldsymbol{j}\in V_{N,i}$ and $t\in[1,T]\}$ are identically distributed across time and space. Moreover,*

(a) *the mixing coefficient $(\alpha(\cdot))$ for $\{(\boldsymbol{\eta}_{\boldsymbol{j},t},\epsilon_{\boldsymbol{j},t}):\ \boldsymbol{j}\in V_{N,i}$ and $t\in[1,T]\}$, $i=1,\ldots,G$, decays to zero at some rate such that $\alpha(\tau)\le C_\theta\tau^{-\theta_\alpha}$ with some*

$$\theta_\alpha\ge\max\left(\frac{p(d_v+1)\gamma_\eta}{(p-q)(\gamma_\eta-2)}+(d_v+1)\gamma_M,\frac{d_v+1}{1-\frac{2}{\gamma_\eta}},\frac{2p}{p-4}-\gamma_M\right),$$

where the generic constants $C_\theta>0$, $\gamma_\eta>2$, $p>4$, $q=\frac{2p}{p-2}$, d_v is the dimension of V_N, and $\gamma_M\ge 1$ is given in Definition 1;

(b) $\max\left(E\|\boldsymbol{\eta}_{\boldsymbol{i},1}\|^p,E\|\boldsymbol{\eta}_{\boldsymbol{i},1}\|^{\gamma_\eta},E\|\boldsymbol{\eta}_{\boldsymbol{i},1}\|^4\right)<\infty$;

(c) $\max\left(|V_N|^{\gamma_M}T^{\gamma_M+1-\theta_\alpha},|V_N|^{\gamma_M(1-2/\gamma_\eta)}T^{\epsilon-1/2},T^{(\gamma_M+\theta_\alpha-1)\epsilon-\frac{1}{2}(\theta_\alpha-\gamma_M-1)}|V_N|^{\gamma_M}\right)\downarrow 0$ *for some* $\epsilon\in\left(0,\min\left(\frac{1}{2},\frac{\theta_\alpha-\gamma_M-1}{2(\gamma_M+\theta_\alpha-1)}\right)\right)$.

Assumption 4.4. *Let $\boldsymbol{X}_{N,T,t}(\boldsymbol{\theta}_0)$ be the same as in Assumption 4.2. The minimum eigenvalue of the stochastic limiting matrix:*

$$\boldsymbol{Q}_{zz}=plim_{N,T\uparrow\infty}\left\{diag\left(T^{-1/2}\mathbb{I}_{G.d_x},N^{-1/2}\mathbb{I}_{G+1}\right)\left(\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{X}_{N,T,t}\boldsymbol{X}_{N,T,t}^\top\right)diag\left(T^{-1/2}\mathbb{I}_{G.d_x},N^{-1/2}\mathbb{I}_{G+1}\right)^\top\right\}$$

is positive.

A few remarks are in order. Assumption 4.3(a) requires that the mixing coefficient should vanish at a rate depending on the orders of the moments specified in Assumption 4.3(b). Assumption 4.3(c) refers to the growth rates of V_N and T, which also depend on the dimension and structure of V_N as well as the decay rate of the mixing coefficient. Assumption 4.4 bears some congruence with the standard assumption [employed in the OLS regression] about the positive-definiteness of the square matrix involving regressors

Theorem 3 (Consistency)**.** *Suppose that Assumptions 2.1, 2.2, 4.3, and 4.4 hold. Then,* $|\widetilde{\sigma}_\epsilon^2 - \sigma_{\epsilon,0}^2| = o_p(1)$, $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_p\left(T^{-1/2}\right)$, $\|\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\| = o_p\left(N^{-1/2}\right)$, *and* $|\widetilde{\mu}_* - \mu_{*,0}| = o_p\left(N^{-1/2}\right)$.

To derive the limiting distribution of $\widetilde{\boldsymbol{\psi}}$, we define some further notations.

$$
\begin{aligned}
\underbrace{\mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi})}_{G.d_x \times G} &= \text{diag}(\boldsymbol{g}\mathbb{I}_{d_x})\text{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}) \left\{ \frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^{T} \boldsymbol{A}_t \boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top \right\} \text{diag}(\boldsymbol{g}), \\
\underbrace{\mathcal{H}_{N,T}^{(ac)}(\boldsymbol{\phi})}_{G.d_x \times 1} &= \text{diag}(\boldsymbol{g}\mathbb{I}_{d_x})\text{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}) \left\{ \frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^{T} \boldsymbol{A}_t C_t \right\}, \\
\underbrace{\mathcal{H}_{N,T}^{(bc)}}_{G \times 1} &= \text{diag}(\boldsymbol{g}) \left\{ \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{B}_t(\boldsymbol{\theta}_0) C_t \right\}, \\
\underbrace{\mathcal{H}_{N,T}^{(aa)}(\boldsymbol{\phi})}_{G.d_x \times G.d_x} &= \text{diag}(\boldsymbol{g}\mathbb{I}_{d_x})\text{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}) \left\{ \frac{N}{T^2} \sum_{t=1}^{T} \boldsymbol{A}_t \boldsymbol{A}_t^\top \right\} \text{diag}(\boldsymbol{g}\mathbb{I}_{d_x})\text{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}), \\
\underbrace{\mathcal{H}_{N,T}^{(bb)}}_{G \times G} &= \text{diag}(\boldsymbol{g}) \left\{ \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{B}_t(\boldsymbol{\theta}_0) \boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top \right\} \text{diag}(\boldsymbol{g}).
\end{aligned}
$$

Let $\mathcal{H}_{N,T}(\boldsymbol{\phi}_0) = \begin{pmatrix} \mathcal{H}_{N,T}^{(aa)}(\boldsymbol{\phi}_0) & \mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi}_0) & \mathcal{H}_{N,T}^{(ac)}(\boldsymbol{\phi}_0) \\ \mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi}_0)^\top & \mathcal{H}_{N,T}^{(bb)} & \mathcal{H}_{N,T}^{(bc)} \\ \mathcal{H}_{N,T}^{(ac)}(\boldsymbol{\phi}_0)^\top & \mathcal{H}_{N,T}^{(bc)\top} & 1 \end{pmatrix}$. Lemma 25 effectively implies that

$$\lim_{N,T\uparrow\infty} \mathcal{H}_{N,T}(\boldsymbol{\phi}_0) = \mathcal{H}(\boldsymbol{\phi}_0),$$

where $\mathcal{H}(\boldsymbol{\phi}_0)$ is a positive-definite stochastic matrix.

Theorem 4 (Asymptotic Normality)**.** *Let Assumptions 2.1, 2.2, 4.3, and 4.4 hold. Then,*

$$\begin{pmatrix} T(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \sqrt{NT}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \\ \sqrt{NT}(\widetilde{\mu}_* - \mu_{*,0}) \end{pmatrix} \xrightarrow{w} MN\left(\mathbf{0}, \mathcal{H}(\boldsymbol{\phi}_0)^{-1}\right),$$

where $MN(\cdot,\cdot)$ *stands for a mixed-normal random variable.*

4.2 Unknown Group Membership

We start by defining some common notations that will be used for the rest of this section. Let $\boldsymbol{u}_c = (u_{1,c}, \ldots, u_{N,c})^\top$ be a $N \times 1$ vector of group membership indicators associated with group labelled

c; $\xi^{(w)}_{0,*,t}(\boldsymbol{u}_c) \doteq \xi^{(w)}_{*,t}(\boldsymbol{\theta}_{0,c}, \boldsymbol{u}_c) = \frac{1}{N}\sum_{i=1}^{N} u_{i,c}\xi^{(w)}_{i,t}(\boldsymbol{\theta}_{0,c})$; $\boldsymbol{x}^{(w)}_{*,t}(\boldsymbol{u}_c) = \frac{1}{N}\sum_{i=1}^{N} u_{i,c}\boldsymbol{x}^{(w)}_{i,t}$; $\boldsymbol{\xi}^{(w)}_{*,t}(\boldsymbol{U}, \sigma^{(per)}) = (\xi^{(w)}_{*,t}(\boldsymbol{u}_{\sigma^{(per)}(1)}), \dots, \xi^{(w)}_{*,t}(\boldsymbol{u}_{\sigma^{(per)}(G)})^\top$; $\boldsymbol{x}^{(w)}_{*,t}(\boldsymbol{U}, \sigma^{(per)}) = (\boldsymbol{x}^{(w)\top}_{*,t}(\boldsymbol{u}_{\sigma^{(per)}(1)}), \dots, \boldsymbol{x}^{(w)\top}_{*,t}(\boldsymbol{u}_{\sigma^{(per)}(G)})^\top$;

$$\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0) = \left(\boldsymbol{x}^{(w)\top}_{*,t}(\boldsymbol{U}, \widetilde{\sigma}^{(per)}), \boldsymbol{\xi}^{(w)\top}_{*,t}(\boldsymbol{U}, \widetilde{\sigma}^{(per)}), 1^{(w)}_t, \boldsymbol{\xi}^{(w)\top}_{*,t}(\boldsymbol{U}_0, \sigma^{(per)}) - \boldsymbol{\xi}^{(w)\top}_{*,t}(\boldsymbol{U}, \widetilde{\sigma}^{(per)})\right)^\top;$$

$\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}) = \mathrm{diag}(\phi_{\widetilde{\sigma}^{(per)}(1)}, \dots, \phi_{\widetilde{\sigma}^{(per)}(G)})$; $\boldsymbol{\theta}^{(\sigma^{(per)})} = (\boldsymbol{\theta}^\top_{\sigma^{(per)}(1)}, \dots, \boldsymbol{\theta}^\top_{\sigma^{(per)}(G)})^\top$;
$\boldsymbol{\phi}^{(\sigma^{(per)})} = (\phi_{\sigma^{(per)}(1)}, \dots, \phi_{\sigma^{(per)}(G)})^\top$;

Therefore, in view of (3.8), one obtains that

$$\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon^{(w)}_{0,*,t} = \left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right) \mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right) \boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0).$$

In addition, let

$$\begin{aligned} H(\widehat{\boldsymbol{U}}, \boldsymbol{U}_0) &= \Bigg(\max_{\sigma^{(per)}\in\sigma(\mathcal{P})} \min_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \frac{1}{N}\sum_{c=1}^{G}\sum_{i=1}^{N} |\widehat{u}_{i,\widetilde{\sigma}^{(per)}(c)} - u_{0,i,\sigma^{(per)}(c)}|, \\ &\qquad \max_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \frac{1}{N}\sum_{c=1}^{G}\sum_{i=1}^{N} |\widehat{u}_{i,\widetilde{\sigma}^{(per)}(c)} - u_{0,i,\sigma^{(per)}(c)}|\Bigg)^+ \\ &= \left(\min_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \frac{1}{N}\sum_{c=1}^{G}\sum_{i=1}^{N} |\widehat{u}_{i,\widetilde{\sigma}^{(per)}(c)} - u_{0,i,c}|, \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \frac{1}{N}\sum_{c=1}^{G}\sum_{i=1}^{N} |\widehat{u}_{i,c} - u_{0,i,\sigma^{(per)}(c)}|\right)^+ \end{aligned}$$

where $\sigma(\mathcal{P})$ is the set of all permutation operators operating on $\mathcal{P}$, denote the *optimal matching* distance between $\widehat{\boldsymbol{U}}$ and $\boldsymbol{U}_0$, and

$$\begin{aligned} H(\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}_0) &= \left(\max_{\sigma^{(per)}\in\sigma(\mathcal{P})} \min_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \left(\sum_{c=1}^{G} \left\|\widehat{\boldsymbol{\psi}}_{\widetilde{\sigma}^{(per)}(c)} - \boldsymbol{\psi}_{0,\sigma^{(per)}(c)}\right\|^2\right)^{\frac{1}{2}},\right. \\ &\qquad \left.\max_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left(\sum_{c=1}^{G} \left\|\widehat{\boldsymbol{\psi}}_{\widetilde{\sigma}^{(per)}(c)} - \boldsymbol{\psi}_{0,\sigma^{(per)}(c)}\right\|^2\right)^{\frac{1}{2}}\right)^+ \\ &= \left(\min_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})} \left(\sum_{c=1}^{G} \left\|\widehat{\boldsymbol{\psi}}_{\widetilde{\sigma}^{(per)}(c)} - \boldsymbol{\psi}_{0,c}\right\|^2\right)^{\frac{1}{2}}, \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left(\sum_{c=1}^{G} \left\|\widehat{\boldsymbol{\psi}}_c - \boldsymbol{\psi}_{0,\sigma^{(per)}(c)}\right\|^2\right)^{\frac{1}{2}}\right)^+, \end{aligned}$$

where $\widehat{\boldsymbol{\psi}}_c = (\widehat{\boldsymbol{\theta}}_c^\top, \widehat{\phi}_c, \widehat{\mu}_*)^\top$ and $\boldsymbol{\psi}_{0,c} = (\boldsymbol{\theta}_{0,c}^\top, \phi_{0,c}, \mu_{*,0})^\top$, be the *optimal matching* distance between $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}_0$.

For the *stationary* case, we first need to state the following assumption:

Assumption 4.5. *Suppose that* $\lim_{N\uparrow\infty,T\uparrow\infty}\inf_{H(\boldsymbol{U},\boldsymbol{U}_0)>\eta_u}\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)^\top\right)>0$. *Moreover, let* $\boldsymbol{F}_t^{(1)}(\boldsymbol{U})=\left(\boldsymbol{x}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)}),\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)}),1_t^{(w)}\right)^\top\subset\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)$, *assume that*

$$\lim_{N\uparrow\infty,T\uparrow\infty}\inf_{H(\boldsymbol{U},\boldsymbol{U}_0)<\eta_u}\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t^{(1)}(\boldsymbol{U})\boldsymbol{F}_t^{(1)\top}(\boldsymbol{U})\right)>0.$$

Assumption 4.5 states that groups must be well-separated in the sense that, if the two matrices of group-indicating variables $\boldsymbol{U}$ and $\boldsymbol{U}_0$ are mismatched, then the square matrix containing differences $\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U}_0,\sigma^{(per)})-\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)})$ will be a positive-definite matrix. The second part of Assumption 4.5 is similar to the standard assumption employed in the OLS regression.

Theorem 5. *Under Assumptions 4.1, 4.2 and 4.5, it holds that* $\sqrt{N}H(\widehat{\boldsymbol{\psi}},\boldsymbol{\psi}_0)\xrightarrow{p}0$, $H(\widehat{\boldsymbol{U}},\boldsymbol{U}_0)\xrightarrow{p}0$, *and* $|\widehat{\sigma}_\epsilon^2-\sigma_{\epsilon,0}^2|\xrightarrow{p}0$.

Theorem 6 below gives the expected bias [in terms of the *optimal matching* distance] of the estimates $\widehat{\boldsymbol{U}}(\boldsymbol{\psi})$ uniformly over all $\boldsymbol{\psi}$ in a neighborhood of $\boldsymbol{\psi}_0$. The rate at which the expected bias goes to zero depends on the decay rate of the mixing coefficient.

Theorem 6. *Let* $\{(\boldsymbol{x}_{\boldsymbol{i},t},\epsilon_{\boldsymbol{i},t}):\ \boldsymbol{i}\in V_N,\ t\in[1,T]\}$ *represent a mixing vector-valued spatio-temporal process and* $\widehat{\boldsymbol{U}}(\boldsymbol{\psi})=\mathrm{argmin}_{\boldsymbol{U}\in\Delta_S^N\bigcap\{0,1\}^{G\times N}}\frac{N}{T}\sum_{t=1}^{T}\epsilon_{*,t}^2(\boldsymbol{\psi},\boldsymbol{U})$. *Suppose that (a) within each group,* $c\in[1,G]$, $\{\boldsymbol{x}_{\boldsymbol{i},t},\epsilon_{\boldsymbol{i},t}\}$, $\boldsymbol{i}\in V_N$ *and* $t\in[1,T]$ *are identically distributed over time and space; (b) the mixing coefficient* $\alpha(\tau)<C_0\tau^{-\theta_\alpha}$, $\theta_\alpha>\left(\frac{4\gamma_M}{3},\frac{2d_v+1}{1-2/\delta_\alpha}\right)^+$ *for some* $\delta_\alpha>2$; *(c)* $\|\boldsymbol{x}_{\boldsymbol{i},t}\epsilon_{\boldsymbol{i},t}\|_{\delta_\alpha}<\infty$; *(d)* $E[\exp(\ell|\epsilon_{\boldsymbol{s},t}|)]\le C_\ell$ *and* $E[\exp(\ell\|\boldsymbol{x}_{\boldsymbol{s},t}\|)]\le C_\ell$ *for a constant* $C_\ell>0$ *and* $\ell>0$ *small enough. Then, it holds that*

$$E\left[\sup_{\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)}H\left(\widehat{\boldsymbol{U}}(\boldsymbol{\psi}),\boldsymbol{U}_0\right)\right]\ \le\ C_0\left\{T^{-C_\alpha}+N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}+\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\},$$

where $\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)$ *is an open ball centered at* $\boldsymbol{\psi}_0$ *with an arbitrarily small radius,* η_ψ, *in terms of the optimal matching distance.*

Remark 4.2. *It is important to note that most conditions in Theorem 6 above are standard bounded moment and mixing decay rate conditions except Condition (d). The sub-exponential tails of* $\epsilon_{\boldsymbol{s},t}$ *and* $\boldsymbol{x}_{\boldsymbol{s},t}$ *assumed there are needed to apply the truncation technique that yields the first term* T^{-C_α} *in the decay rate. The same condition is employed in Mammen, Rothe, and Schienle (2012). This condition can be satisfied if the stationary density functions of* $\epsilon_{\boldsymbol{s},t}$ *and* $\boldsymbol{x}_{\boldsymbol{s},t}$ *have compact supports.*

In light of Theorem 6, we readily derive the decay rate of the *optimal matching* distance between the parameter estimates $\widehat{\boldsymbol{\psi}}$ and the oracle estimates (i.e. the estimates constructed by maximizing the CL function using the true unknown groups) $\widetilde{\boldsymbol{\psi}}$. The main result is stated in Theorem 7. Again, this decay rate depends on the decay rate of the mixing coefficient.

Theorem 7. *Let all the conditions in Theorem 6 hold. Under Assumptions 4.1, 4.2 and 4.5, it holds that*

$$H\left(\widehat{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\psi}}\right) = O_p\left(NT^{-C_\alpha} + N^{\gamma_M+1}\log(T)T^{\frac{\gamma_M}{2}-\frac{3}{8}\theta_\alpha} + N\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right),$$

where C_α and C_M are some sufficiently large constants.

For the *nonstationary* case, let

$$\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U}, \boldsymbol{U}_0) \doteq \text{diag}\left(\sqrt{\frac{N}{T}}\mathbb{I}_{Gd_x}, \mathbb{I}_{2G+1}\right)\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)^\top\text{diag}\left(\sqrt{\frac{N}{T}}\mathbb{I}_{Gd_x}, \mathbb{I}_{2G+1}\right).$$

It then follows from Lemma 25 that

$$\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U}, \boldsymbol{U}_0) \xrightarrow{w} \boldsymbol{\Lambda}(\boldsymbol{U}, \boldsymbol{U}_0),$$

where the limit $\boldsymbol{\Lambda}(\boldsymbol{U}, \boldsymbol{U}_0)$ is a stochastic matrix. We first state a variant of Assumption 4.5 about group well-separability in Assumption 4.6.

Assumption 4.6. $\liminf_{H(\boldsymbol{U},\boldsymbol{U}_0)>\eta_u}\boldsymbol{\Lambda}(\boldsymbol{U}, \boldsymbol{U}_0) > 0$ *for every* $\eta_u > 0$.

Theorem 8. *Let $\boldsymbol{x}_{i,t} = \sum_{s=1}^{t}\sum_{i=1}^{N}\boldsymbol{\eta}_{i,s}$ and $\{(\boldsymbol{\eta}_{\boldsymbol{i},t}, \epsilon_{\boldsymbol{i},t}) : \boldsymbol{i} \in V_N,\ t \in [1, T]\}$ is a mixing vector-valued spatio-temporal process Suppose that (a) within each group, $c \in [1, G]$, $\{\boldsymbol{\eta}_{\boldsymbol{i},t}, \epsilon_{\boldsymbol{i},t}\}$, $\boldsymbol{i} \in V_N$ and $t \in [1, T]$ are identically distributed over time and space; (b) the mixing coefficient $\alpha(\tau) < C_0\tau^{-\theta_\alpha}$, $\theta_\alpha > \left(\frac{4\gamma_M}{3}, \frac{2d_v+1}{1-2/\delta_\alpha}\right)^+$ for some $\delta_\alpha > 2$; (c) $\|\boldsymbol{\eta}_{\boldsymbol{i},t}\epsilon_{\boldsymbol{i},t}\|_{\delta_\alpha} < \infty$; (d)* (sub-exponential tails) *$E[\exp(\ell|\epsilon_{\boldsymbol{s},t}|)] \leq C_\ell$ and $E[\exp(\ell\|\boldsymbol{\eta}_{\boldsymbol{s},t}\|)] \leq C_\ell$ for a constant $C_\ell > 0$ and $\ell > 0$ small enough; (e) $N/T \longrightarrow$ const. Then, under Assumptions 2.1, 2.2, 4.3, and 4.6 it holds that $\sqrt{T}H\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0\right) = o_p(1)$, $\sqrt{N}H\left(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}_0\right) = o_p(1)$, $\sqrt{N}|\widehat{\mu}_* - \mu_{*,0}| = o_p(1)$, $|\widehat{\sigma}^2_\epsilon - \sigma^2_{\epsilon,0}| = o_p(1)$.*

Theorems 9 and 10 below present the decay rates for the expected uniform bias of the estimates of the true group-indicating variables, and for the *optimal matching* distance between the parameter estimates $\widehat{\boldsymbol{\psi}}$ and the oracle estimates $\widetilde{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}_0$.

Theorem 9. *Let $\widehat{\boldsymbol{U}}(\boldsymbol{\psi}) = argmin_{\boldsymbol{U}\in\Delta_S^N\bigcap\{0,1\}^{G\times N}}\frac{N}{T}\sum_{t=1}^{T}\epsilon^2_{*,t}(\boldsymbol{\psi}, \boldsymbol{U})$. Suppose that the conditions of Theorem 8 hold. Then,*

$$E\left[\sup_{\boldsymbol{\psi}\in\mathcal{N}_{N,T}(\boldsymbol{\psi}_0,\eta_\psi)}H\left(\widehat{\boldsymbol{U}}(\boldsymbol{\psi}), \boldsymbol{U}_0\right)\right] \leq C_0\left\{N^{-C_\alpha} + T^{-C_\alpha} + N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha} + \exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{\log^2(T)}\right)\right\}$$

for every positive C_α *and* C_{ϵ_η}, *where* $\mathcal{N}_{N,T}(\boldsymbol{\psi}_0, \eta_\psi) = \bigcap_{\substack{\eta_\theta>0,\eta_\phi>0,\eta_\mu>0 \\ \left(\eta_\theta^2+\eta_\phi^2+\eta_\mu^2\right)^{\frac{1}{2}}=\eta_\psi}} \mathcal{B}_T(\boldsymbol{\theta}_0, \eta_\theta) \times \mathcal{B}_N(\boldsymbol{\phi}_0, \eta_\phi) \times \mathcal{B}_N(\mu_{*,0}, \eta_\mu)$ *with* $\mathcal{B}_T(\boldsymbol{\theta}_0, \eta_\theta) = \{\boldsymbol{\theta} \in \Theta_\theta : \sqrt{T}H(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \eta_\theta\}$, $\mathcal{B}_N(\boldsymbol{\phi}_0, \eta_\phi) = \{\boldsymbol{\phi} \in \Theta_\phi : \sqrt{N}H(\boldsymbol{\phi}, \boldsymbol{\phi}_0) < \eta_\phi\}$, *and* $\mathcal{B}_N(\mu_{*,0}, \eta_\mu) = \{\mu_* \in \Theta_\mu : \sqrt{N}|\mu_* - \mu_{*,0}| < \eta_\mu\}$.

Theorem 10. *Let all the conditions stipulated in Theorem 9 hold. Then, it holds under Assumption 4.4 that*

$$H\left(diag\left(\sqrt{T}\mathbb{I}_{G\times d_x}, \sqrt{N}\mathbb{I}_{G+1}\right)\widehat{\boldsymbol{\psi}}, diag\left(\sqrt{T}\mathbb{I}_{G\times d_x}, \sqrt{N}\mathbb{I}_{G+1}\right)\widetilde{\boldsymbol{\psi}}\right)$$
$$= O_p\left(N^{\frac{1-C_\alpha}{2}} + N^{1/2}T^{-\frac{C_\alpha}{2}} + N^{\gamma_M+\frac{1}{2}}\log(T)T^{\frac{\gamma_M}{2}-\frac{3}{8}\theta_\alpha} + N^{\frac{1}{2}}\exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{2\log^2(T)}\right)\right),$$

where C_α *and* C_{ϵ_η} *are some positive constant.*

5 Computation: A New VNS-DCA Algorithm

Some background material on the gist of the DC programming and DCA is provided in Appendix G. First, recall the notations defined earlier: $\Delta y_{*,t}^{(w)} = \Delta y_{*,t} - \left(\sum_{t=1}^T \Delta y_{*,t}\boldsymbol{w}_{*,t}^\top\right)\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\boldsymbol{w}_{*,t}$, $y_{i,t-1}^{(w)} = y_{i,t-1} - \left(\sum_{t=1}^T y_{i,t}\right)\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}\right)^{-1}\boldsymbol{w}_{*,t}$, $\boldsymbol{x}_{i,t}^{(w)} = \boldsymbol{x}_{i,t} - \left(\sum_{t=1}^T \boldsymbol{x}_{i,t}\boldsymbol{w}_{*,t}^\top\right)\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\boldsymbol{w}_{*,t}$, and $1_t^{(w)} = 1 - \left(\sum_{t=1}^T \boldsymbol{w}_{*,t}^\top\right)\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\boldsymbol{w}_{*,t}$. The *concentrated* composite innovations were defined as

$$\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) = \Delta y_{*,t}^{(w)} - \sum_{c=1}^G \frac{1}{N}\sum_{i=1}^N u_{i,c}\phi_c\left(y_{i,t-1}^{(w)} - \boldsymbol{\theta}_c^\top \boldsymbol{x}_{i,t}^{(w)}\right) - \mu_* 1_t^{(w)}. \tag{5.1}$$

For a given $\boldsymbol{U} \in \Delta_S^N$, local minimum values, $\widehat{\boldsymbol{\psi}}(\boldsymbol{U})$, of $\mathcal{E}_{N,T}(\boldsymbol{\psi}, \boldsymbol{U}) = \frac{1}{T}\sum_{t=1}^T \epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U})$ satisfy the Karush-Kuhn-Tucker (KKT) conditions. Since $\frac{\partial}{\partial \mu_*}\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) = -1_t^{(w)}$, it then follows that

$$\widehat{\mu}_* = \left(\sum_{t=1}^T 1_t^{(w)2}\right)^{-1}\left\{\sum_{t=1}^T \Delta y_{*,t}^{(w)} 1_t^{(w)} - \sum_{c=1}^G \frac{1}{N}\sum_{i=1}^N u_{i,c}\widehat{\phi}_c\left(\sum_{t=1}^T y_{i,t-1}^{(w)} 1_t^{(w)} - \widehat{\boldsymbol{\theta}}_c^\top \sum_{t=1}^T \boldsymbol{x}_{i,t}^{(w)} 1_t^{(w)}\right)\right\}. \tag{5.2}$$

Thus, $\widehat{\boldsymbol{\gamma}}(\boldsymbol{U}) = \left(\widehat{\boldsymbol{\theta}}(\boldsymbol{U}), \widehat{\boldsymbol{\phi}}(\boldsymbol{U})\right)$ are the minimum values of

$$\mathcal{E}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{T}\sum_{t=1}^T \epsilon_{*,t}^2(\boldsymbol{\gamma}, \boldsymbol{U}), \tag{5.3}$$

where

$$\epsilon_{*,t}(\boldsymbol{\gamma},\boldsymbol{U}) = A_t - \sum_{c=1}^{G}\frac{1}{N}\sum_{i=1}^{N} u_{i,c}\phi_c\left\{B_{i,t} - \boldsymbol{\theta}_c^\top \boldsymbol{C}_{i,t}\right\}$$

with

$$\begin{aligned}
A_t &\doteq A_{N,T,t} = \Delta y_{*,t}^{(w)} - \left(\sum_{t=1}^{T} 1_t^{(w)2}\right)^{-1}\left\{\sum_{t=1}^{T}\Delta y_{*,t}^{(w)} 1_t^{(w)}\right\} 1_t^{(w)}\\
B_{i,t} &\doteq B_{N,T,i,t} = y_{i,t-1}^{(w)} - \left(\sum_{t=1}^{T} 1_t^{(w)2}\right)^{-1}\left\{\sum_{t=1}^{T} y_{i,t-1}^{(w)} 1_t^{(w)}\right\} 1_t^{(w)}\\
\boldsymbol{C}_{i,t} &\doteq \boldsymbol{C}_{N,T,i,t} = \boldsymbol{x}_{i,t}^{(w)} - \left(\sum_{t=1}^{T} 1_t^{(w)2}\right)^{-1}\left\{\sum_{t=1}^{T} \boldsymbol{x}_{i,t}^{(w)} 1_t^{(w)}\right\} 1_t^{(w)}.
\end{aligned}$$

Let

$$\begin{aligned}
\mathcal{E}_{1,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) = \mathcal{A}_0 + \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i=1}^{N}\Big\{&u_{i,c}^2\phi_c^2\mathcal{A}_{1,i} + u_{i,c}^2\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{B}_{1,i}\boldsymbol{\theta}_c - 2u_{i,c}^2\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{C}_{1,i} - 2Nu_{i,c}\phi_c\mathcal{D}_{1,i}\\
&+2Nu_{i,c}\phi_c\boldsymbol{\theta}_c^\top\mathcal{F}_{1,i}\Big\}, \quad (5.4)
\end{aligned}$$

where $\mathcal{A}_0 = \frac{1}{T}\sum_{t=1}^{T} A_t^2$; $\mathcal{A}_{1,i} \doteq \mathcal{A}_{1,N,T,i} = \frac{1}{T}\sum_{t=1}^{T} B_{i,t}^2$; $\mathcal{B}_{1,i} \doteq \mathcal{B}_{1,N,T,i} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{C}_{i,t}\boldsymbol{C}_{i,t}^\top$; $\mathcal{C}_{1,i} \doteq \mathcal{C}_{1,N,T,i} = \frac{1}{T}\sum_{t=1}^{T} B_{i,t}\boldsymbol{C}_{i,t}$; $\mathcal{D}_{1,i} \doteq \mathcal{D}_{1,N,T,i} = \frac{1}{T}\sum_{t=1}^{T} A_t B_{i,t}$; and $\mathcal{F}_{1,i} \doteq \mathcal{F}_{N,T,i} = \frac{1}{T}\sum_{t=1}^{T} A_t\boldsymbol{C}_{i,t}$.

$$\mathcal{E}_{2,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i\neq j}^{N}\left\{u_{i,c}u_{j,c}\phi_c^2\mathcal{A}_{2,i,j} - u_{i,c}u_{j,c}\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{B}_{2,i,j} + u_{i,c}u_{j,c}\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right\}, \quad (5.5)$$

where $\mathcal{A}_{2,i,j} \doteq \mathcal{A}_{2,N,T,i,j} = \frac{1}{T}\sum_{t=1}^{T} B_{i,t}B_{j,t}$; $\mathcal{B}_{2,i,j} \doteq \mathcal{B}_{2,N,T,i,j} = \frac{1}{T}\sum_{t=1}^{T}\left(B_{i,t}\boldsymbol{C}_{j,t} + B_{j,t}\boldsymbol{C}_{i,t}\right)$; and $\mathcal{C}_{2,i,j} \doteq \mathcal{D}_{2,N,T,i,j} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{C}_{i,t}\boldsymbol{C}_{j,t}^\top$.

$$\begin{aligned}
\mathcal{E}_{3,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{1}{N^2}\sum_{c\leq \ell}^{G}\sum_{i\neq j}^{N}\Big(&u_{i,c}u_{j,\ell}\phi_c\phi_\ell\mathcal{A}_{2,i,j} - u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{B}_{3,i,j} - u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_\ell^\top\mathcal{C}_{3,i,j}\\
&+u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\Big), \quad (5.6)
\end{aligned}$$

where $\mathcal{B}_{3,i,j} = \frac{1}{T}\sum_{t=1}^{T} B_{j,t}\boldsymbol{C}_{i,t}$; and $\mathcal{C}_{3,i,j} \doteq \mathcal{C}_{3,N,T,i,j} = \mathcal{B}_{2,i,j} - \mathcal{B}_{3,i,j}$.

Using the relation $2g_1g_2 = (g_1+g_2)^2 - (g_1^2+g_2^2)$, we can immediately verify that $\mathcal{E}_{1,N,T}(\boldsymbol{\gamma},\boldsymbol{U})$, $\mathcal{E}_{2,N,T}(\boldsymbol{\gamma},\boldsymbol{U})$, and $\mathcal{E}_{3,N,T}(\boldsymbol{\gamma},\boldsymbol{U})$ defined in (5.4)-(5.6) are d.c. functions.

Assumption 5.1. *For ease of exposition, let the parameters $\boldsymbol{\gamma}$ take values in symmetric boxes, $\boldsymbol{\phi} \in \prod_{c=1}^{G}[-\ell_{\phi,c}, \ell_{\phi,c}]$ and*

$\boldsymbol{\theta} \in \prod_{c=1}^{G} \prod_{i=1}^{d_x} [-\ell_{\theta,c,i} \leq \theta_{i,c} \leq \ell_{\theta,c,i}]$.

Let

$$\mathcal{H}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) \equiv \mathcal{H}_{\rho,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) \doteq N^2 \left(\mathcal{H}_{1,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) + \mathcal{H}_{2,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) + \mathcal{H}_{3,N,T}(\boldsymbol{\gamma}, \boldsymbol{U})\right),$$

$$\begin{aligned} F_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = N^2 \left\{\mathcal{E}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) - \mathcal{A}_0\right\} &= N^2 \left\{\mathcal{E}_{1,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) + \mathcal{E}_{2,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) + \mathcal{E}_{3,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) - \mathcal{A}_0\right\} \\ &= \widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) - \mathcal{H}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}), \end{aligned}$$

where

$$\widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) \doteq \frac{5}{2}\rho \sum_{c=1}^{G} \sum_{i=1}^{N} u_{i,c}^2 + 2\rho \sum_{c=1}^{G} \phi_c^2 + 2\rho \sum_{c=1}^{G} \boldsymbol{\theta}_c^\top \boldsymbol{\theta}_c.$$

Note that the following equivalence between mixed-integer sets and polyhedral sets (see, e.g., Hoang (1995)): $\{\boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G\times N}\} \equiv \left\{\boldsymbol{U} \in \Delta_S^N : \; g(\boldsymbol{U}) \leq 0\right\}$, where $g(\boldsymbol{U}) = \sum_{c=1}^{G} \sum_{i=1}^{N} u_{i,c}(1 - u_{i,c})$ is finite concave on $\mathbb{R}^{G\times N}$ and nonnegative on Δ_S^N. In view of Le Thi Hoai An, Huynh Van Ngai, and Pham Dinh Tao (2012, Theorem 1), we immediately obtain that

$$\begin{aligned} &\min_{\boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G\times N}} \min_{\substack{\boldsymbol{\phi} \in \prod_{c=1}^{G}[-\ell_{\phi,c}, \ell_{\phi,c}] \\ \boldsymbol{\theta} \in \prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i}, \ell_{\theta,c,i}]}} F_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) \\ &\qquad = \min_{\boldsymbol{U} \in \Delta_S^N} \min_{\substack{\boldsymbol{\phi} \in \prod_{c=1}^{G}[-\ell_{\phi,c}, \ell_{\phi,c}] \\ \boldsymbol{\theta} \in \prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i}, \ell_{\theta,c,i}]}} \left\{\widetilde{F}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = F_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) + \widetilde{\gamma} g(\boldsymbol{U})\right\} \end{aligned}$$

for some $\widetilde{\gamma} > 0$. The function $\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \mathcal{H}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) - \widetilde{\gamma} g(\boldsymbol{U})$ is convex for some appropriately chosen $\boldsymbol{\rho}$, which is stipulated by Lemmas 8-10. The gradient $\nabla \widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U})$ of $\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U})$ is given by

$$\begin{aligned} \frac{\partial}{\partial u_{i,c}} \widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) &= 5\rho_u u_{i,c} + 2N\phi_c \frac{1}{T} \sum_{t=1}^{T} \epsilon_{*,t}(\boldsymbol{\gamma}, \boldsymbol{U}) \left(B_{i,t} - \boldsymbol{\theta}_c^\top \boldsymbol{C}_{i,t}\right) + \widetilde{\gamma}(2u_{i,c} - 1), \\ \frac{\partial}{\partial \phi_c} \widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) &= 4\rho_\phi \phi_c + 2\frac{N}{T} \sum_{i=1}^{N} u_{i,c} \sum_{t=1}^{T} \epsilon_{*,t}(\boldsymbol{\gamma}, \boldsymbol{U}) \left(B_{i,t} - \boldsymbol{\theta}_c^\top \boldsymbol{C}_{i,t}\right), \\ \frac{\partial}{\partial \boldsymbol{\theta}_c} \widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) &= 4\rho_\theta \boldsymbol{\theta}_c - 2\phi_c \frac{N}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} u_{i,c} \epsilon_{*,t}(\boldsymbol{\gamma}, \boldsymbol{U}) \boldsymbol{C}_{i,t} \end{aligned}$$

for all $i = 1, \ldots, N$ and $c = 1, \ldots, G$.

This then leads to the following d.c. programming problem:

$$\min\Big\{\chi_{\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i},\ell_{\theta,c,i}]\times\prod_{c=1}^{G}[-\ell_{\phi,c},\ell_{\phi,c}]\times\Delta_S^N}(\boldsymbol{\gamma},\boldsymbol{U})+\widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\gamma},\boldsymbol{U})-\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma},\boldsymbol{U}):\ \boldsymbol{U}\in\mathbb{R}^{N\times G},$$
$$\boldsymbol{\gamma}=(\boldsymbol{\theta},\boldsymbol{\phi})\in\mathbb{R}^{G\times d_x\times G}\Big\}. \quad (5.7)$$

Remark 5.1. *The above DC decomposition uses the concentrated composite errors. This DC decomposition has some merits in terms of the execution speed as the objective function has fewer parameters to be optimized than the full composite likelihood function. To minimize the full composite likelihood function the problem (5.7) then needs the convex functions $\widetilde{\mathcal{G}}_{N,T}$ and $\widetilde{\mathcal{H}}_{N,T}$ provided in Appendix F. instead. All the algorithms described below can effectively be employed to minimize the sum of squared composite errors; and the computer program provided is specifically written for this minimization problem using the DC representation derived in Appendix F.*

The DCA applied to the problem (5.7) is described in Algorithm 1 below.

Algorithm 1 DCA

procedure DC–A
  Choose an initial point to start recursion, say $\{\boldsymbol{U}^{(0)},\boldsymbol{\theta}^{(0)},\boldsymbol{\phi}^{(0)}\}$, and an error tolerance level, ϵ
  Set $\ell\leftarrow 0$
  repeat
    $\{\boldsymbol{\lambda}^{(\ell)},\boldsymbol{\gamma}^{(\ell)},\boldsymbol{V}^{(\ell)}\}\in\nabla\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\theta}^{(\ell)},\boldsymbol{\phi}^{(\ell)},\boldsymbol{U}^{(\ell)})$
    $\min\Big\{\widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{U})-<\{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{U}\},\{\boldsymbol{\lambda}^{(\ell)},\boldsymbol{\gamma}^{(\ell)},\boldsymbol{V}^{(\ell)}\}>:\ \{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{U}\}\in\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i},\ell_{\theta,c,i}]$
    $\times\prod_{c=1}^{G}[-\ell_{\phi,c},\ell_{\phi,c}]\times\Delta_S^N\Big\}$
    (i.e., set $\boldsymbol{U}^{(\ell+1)}=\mathrm{Proj}_{\Delta_S^N}\left(\frac{\boldsymbol{V}^{(\ell)}}{5\rho_u}\right)$, $\boldsymbol{\theta}^{(\ell+1)}=\mathrm{Proj}_{\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i},\ell_{\theta,c,i}]}\left(\frac{\boldsymbol{\lambda}^{(\ell)}}{4\rho_\theta}\right)$,
    and $\boldsymbol{\phi}^{(\ell+1)}=\mathrm{Proj}_{\prod_{c=1}^{G}[-\ell_{\phi,c},\ell_{\phi,c}]}\left(\frac{\boldsymbol{\gamma}^{(\ell)}}{4\rho_\phi}\right)$)
    $\{\boldsymbol{\gamma}^{**},\boldsymbol{U}^{**}\}=\{\boldsymbol{\theta}^{(\ell+1)},\boldsymbol{\phi}^{(\ell+1)},\boldsymbol{U}^{(\ell+1)}\}$
    $\{\boldsymbol{\gamma}^{*},\boldsymbol{U}^{*}\}=\{\boldsymbol{\theta}^{(\ell)},\boldsymbol{\phi}^{(\ell)},\boldsymbol{U}^{(\ell)}\}$
    $\ell\leftarrow\ell+1$
  until $\|\{\boldsymbol{\gamma}^{**},\boldsymbol{U}^{**}\}-\{\boldsymbol{\gamma}^{*},\boldsymbol{U}^{*}\}\|\le\epsilon$
return $\{\boldsymbol{\theta}^{(\ell+1)},\boldsymbol{\phi}^{(\ell+1)},\boldsymbol{U}^{(\ell+1)}\}$
end procedure

$\mathrm{Proj}_{\Delta_S^N}(\boldsymbol{v})$ denotes the projection of $\boldsymbol{v}$ onto the Cartesian product of standard unit G-dimensional simplices; there and many efficient algorithms to compute this projection, for example, the spectral projected gradient algorithm (Júdice, Raydan, Rosa, and Santos, 2008). Other projections onto rectangles can be straight-forwardly computed.

In Algorithm 1, there are two important implementation issues that warrant discussion. The first issue is how to choose ρ as small as possible so that the function $\mathcal{H}_{N,T}(\boldsymbol{\gamma},\boldsymbol{U})$ is still convex and the concave part $-\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\gamma},\boldsymbol{U})$ of the d.c. decomposition becomes less important so as to enhance the efficiency of the DCA. Algorithm 2 to update ρ is suggested by Le Thi Hoai An, Le Hoai Minh, and Pham Dinh Tao (2014). The second issue is to choose a 'good' starting point. For the DCA to work, a starting point must not be a local optimal point as the DCA is stationary at that point. The variable neighbourhood search (VNS) algorithm proposed by Hansen and Mladenović (1997) can potentially generate good starting points for the DCA. The VNS is an effective heuristic scheme for combinatorial and global optimization, which can easily implemented using any local search algorithm as a subroutine. The main principle of the VNS is to explore pre-determined distant neighborhoods of the current incumbent solution, and jump from there to a new one if there is an improvement found through a local search routine. A typical VNS routine requires a set of neighborhoods to be specified.

The structure of all non-intersecting neighborhoods [of $\boldsymbol{\gamma}$] in the hyper-rectangle $\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i},\ell_{\theta,c,i}]\times\prod_{c=1}^{G}[-\ell_{\phi,c},\ell_{\phi,c}]$ can be defined by $\mathcal{H}_k(\boldsymbol{\gamma})\doteq H_k(\boldsymbol{\gamma})/H_{k-1}(\boldsymbol{\gamma})$, where $H_k(\boldsymbol{\gamma})=\prod_{c=1}^{G}\prod_{i=1}^{d_x}[\ell_{\theta,c,i}^{(k)},\omega_{\theta,c,i}^{(k)}]\times\prod_{c=1}^{G}[\ell_{\phi,c}^{(k)},\omega_{\phi,c}^{(k)}]$, $k=1,\dots,k_{\max}$, with

$$\begin{aligned}
\ell_{\theta,c,i}^{(k)} &\doteq \ell_{\theta,c,i}^{(k)}(\theta_{c,i})=\theta_{c,i}-\frac{k}{k_{\max}}(\theta_{c,i}+\ell_{\theta,c,i}),\\
\omega_{\theta,c,i}^{(k)} &\doteq \omega_{\theta,c,i}^{(k)}(\theta_{c,i})=\theta_{c,i}+\frac{k}{k_{\max}}(\ell_{\theta,c,i}-\theta_{c,i}),\\
\ell_{\phi,c}^{(k)} &\doteq \ell_{\phi,c}^{(k)}(\phi_c)=\phi_c-\frac{k}{k_{\max}}(\phi_c+\ell_{\phi,c}),\\
\omega_{\phi,c}^{(k)} &\doteq \omega_{\phi,c}^{(k)}(\phi_c)=\phi_c+\frac{k}{k_{\max}}(\ell_{\phi,c}-\phi_c).
\end{aligned}$$

Let $\kappa(\boldsymbol{U},\boldsymbol{U}')$ denote the Hamming distance between $\boldsymbol{U}$ and $\boldsymbol{U}'$ (i.e., the number of pairwise different columns of these $G\times N$ matrices). The system of all neighborhoods [of $\boldsymbol{U}$] induced by this metric in Δ_S^N is then given by $\mathcal{N}_\ell(\boldsymbol{U})\doteq\left\{\boldsymbol{U}'\in\Delta_S^N:\ \kappa(\boldsymbol{U},\boldsymbol{U}')=\ell\right\}$, $\ell=1,\dots,\ell_{\max}$, $\ell_{\max}\doteq N$. Therefore, one can choose $\mathcal{N}_{k,\ell}(\boldsymbol{\gamma},\boldsymbol{U})\doteq\mathcal{H}_k(\boldsymbol{\gamma})\times\mathcal{N}_\ell(\boldsymbol{U})$, $\forall\ k=1,\dots,k_{\max};\ \ell=1,\dots,\ell_{\max}$, as a structure of neighborhoods [of $\boldsymbol{\gamma}\times\boldsymbol{U}$] in $\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i},\ell_{\theta,c,i}]\times\prod_{c=1}^{G}[-\ell_{\phi,c},\ell_{\phi,c}]\times\Delta_S^N$.

The VNS using the defined neighborhood system is reminiscent of the divide-and-conquer strategy used in a branch-and-bound optimization algorithm - breaking the search space into smaller pieces, then optimizing the objective function on these pieces. Unlike branch-and-bound algorithms the VNS also allows the system of neighborhoods to vary at each iteration. The basic VNS procedure is described in Algorithm 3 below. In this algorithm, local searches can be performed by using Simulated Annealing (SA) (see, e.g., Guyon (1995, p. 212)) instead of the K-means algorithm. The K-means - despite of its appealing computational efficiency - has certain shortcomings, such as it is very sensitive to outliers so that the computed clusters are different from actual ones, and it does not often reach global optimum even when being ignited by different initial values (see, e.g.,

Algorithm 2 Update $\boldsymbol{\rho} = (\rho_u, \rho_\phi, \rho_\theta)$

Initialize the routine using $\{\boldsymbol{U}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}\}$, and choose a step size, $\tau_\rho \in (0, 1)$.
Set $\ell \leftarrow 0$ and $\boldsymbol{\rho}^{(0)} \leftarrow \boldsymbol{\rho}_0$, where $\boldsymbol{\rho}_0$ satisfies Lemmas 8-10.
repeat
  $\boldsymbol{\rho}^{(\ell+1)} = \tau_\rho \boldsymbol{\rho}^{(\ell)}$
  $\{\boldsymbol{\lambda}^{(\ell)}, \boldsymbol{\gamma}^{(\ell)}, \boldsymbol{V}^{(\ell)}\} \in \nabla \widetilde{\mathcal{H}}_{\rho^{(\ell+1)},N,T}(\boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}^{(\ell)}, \boldsymbol{U}^{(\ell)})$
  Set $\boldsymbol{U}^{(\ell+1)} = \mathrm{Proj}_{\Delta_S^N}\left(\frac{\boldsymbol{V}^{(\ell)}}{5\rho_u^{(\ell+1)}}\right)$, $\boldsymbol{\theta}^{(\ell+1)} = \mathrm{Proj}_{\prod_{c=1}^{G}\prod_{i=1}^{d_x}[-\ell_{\theta,c,i}, \ell_{\theta,c,i}]}\left(\frac{\boldsymbol{\lambda}^{(\ell)}}{4\rho_\theta^{(\ell+1)}}\right)$,
  and $\boldsymbol{\phi}^{(\ell+1)} = \mathrm{Proj}_{\prod_{c=1}^{G}[-\ell_{\phi,c}, \ell_{\phi,c}]}\left(\frac{\boldsymbol{\gamma}^{(\ell)}}{4\rho_\phi^{(\ell+1)}}\right)$
  $\{\boldsymbol{\gamma}^{**}, \boldsymbol{U}^{**}\} = \{\boldsymbol{\theta}^{(\ell+1)}, \boldsymbol{\phi}^{(\ell+1)}, \boldsymbol{U}^{(\ell+1)}\}$
  $\{\boldsymbol{\gamma}^{*}, \boldsymbol{U}^{*}\} = \{\boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}^{(\ell)}, \boldsymbol{U}^{(\ell)}\}$
  $\ell \leftarrow \ell + 1$
  $\boldsymbol{\rho}^{(\ell)} \leftarrow \boldsymbol{\rho}^{(\ell+1)}$
until $F_{N,T}(\boldsymbol{\gamma}^{**}, \boldsymbol{U}^{**}) > F_{N,T}(\boldsymbol{\gamma}^{*}, \boldsymbol{U}^{*})$
if $\ell > 1$ **then**
  return $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^{(\ell)}$ and $\{\boldsymbol{U}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}\} \leftarrow \{\boldsymbol{U}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}^{(\ell)}\}$
else
  return $\boldsymbol{\rho}^{(0)}$ and $\{\boldsymbol{U}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)}\}$
end if

Tan, Steinbach, and Kumar (2005); Wu (2012)). A properly designed SA-based algorithm can be more efficient than the K-means algorithm in obtaining a globally optimal solution to the clustering problem (Brown and Huntley (1992); Klein and Dubes (1989); Selim and Alsultan (1991)). The annealing process, as implemented via the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953)), always allows for some possibility of moving out of a local optimum by probably accepting a 'worse' local value of the objective function. Therefore the SA can eventually generate near global optimum after a number of runs required to first "melt" the system being optimized at a high effective temperature, then to lower the temperature gradually until the system "freezes" and no further changes to the system can be found. In fact the DCA merely needs a 'good' starting point, which must not be a local optimum, to proceed; and ideally, this 'good' starting point is a near global optimum. The SA procedure is given in Algorithm 4.

Algorithm 3 VNS

procedure VARIABLE NEIGHBORHOOD SEARCH (VNS) PROCEDURE
Choose initial values, $\{\boldsymbol{\gamma}^{(0)}, \boldsymbol{U}^{(0)}\}$, and an error tolerance level, ϵ
$\ell \leftarrow 0$
$k \leftarrow 0$
    do
loop:
        Randomly generate a point $\{\boldsymbol{\gamma}'^{(\ell)}, \boldsymbol{U}'^{(\ell)}\} \in \mathcal{N}_{k,\ell}\left(\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\right) \doteq \mathcal{H}_k(\boldsymbol{\gamma}^{(\ell)}) \times \mathcal{N}_\ell(\boldsymbol{U}^{(\ell)})$
        Do a local search starting at $\{\boldsymbol{\gamma}'^{(\ell)}, \boldsymbol{U}'^{(\ell)}\}$ in $\mathcal{N}_{k,\ell}\left(\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\right)$ and obtain a local optimum,
        $\{\boldsymbol{\gamma}^{(\ell+1)}, \boldsymbol{U}^{(\ell+1)}\}$
        if $F_{N,T}(\boldsymbol{\gamma}^{(\ell+1)}, \boldsymbol{U}^{(\ell+1)}) < F_{N,T}(\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)})$ **then**
            $\{\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\} \leftarrow \{\boldsymbol{\gamma}^{(\ell+1)}, \boldsymbol{U}^{(\ell+1)}\}$
            goto *loop*
        else
            $\ell \leftarrow \ell + 1$
            $k \leftarrow k + 1$
        end if
    while $(\ell \le \ell_{\max}$ **AND** $k \le k_{\max})$ **OR** $\|\{\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\} - \{\boldsymbol{\gamma}^{(\ell-1)}, \boldsymbol{U}^{(\ell-1)}\}\| \le \epsilon$
return $\{\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\}$
end procedure

The algorithm proposed by Selim and Alsultan (1991) is employed for randomly generating a neighboring group assignment, $\boldsymbol{U}'^{(\ell)}$, of $\boldsymbol{U}'$.

6 Empirical Choice of the Number of Groups

In the present maximum likelihood paradigm the optimal selection of the number of groups can be implemented by the following analogues of AIC and BIC:

$$\begin{aligned}
\mathrm{AIC}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{U}}) &= \log\left(\frac{N}{T}\sum_{t=1}^{T}\widehat{\epsilon}_{*,t}^{2}\left(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{U}}\right)\right) + \frac{\left|\widehat{\boldsymbol{\Theta}}\right|}{T}, \\
\mathrm{BIC}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{U}}) &= \log\left(\frac{N}{T}\sum_{t=1}^{T}\widehat{\epsilon}_{*,t}^{2}\left(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{U}}\right)\right) + \frac{\left|\widehat{\boldsymbol{\Theta}}\right|\log T}{T},
\end{aligned}$$

where $\widehat{\boldsymbol{U}}$ consists of estimates for the group membership indicators, and $\widehat{\boldsymbol{\Theta}}$ is the vector containing estimates for the model parameters associated with the group classification provided by $\widehat{\boldsymbol{U}}$; and

Algorithm 4 SA

procedure SIMULATED ANNEALING (SA) PROCEDURE
Initialize the algorithm using $\{\boldsymbol{\gamma}^{(0)}, \boldsymbol{U}^{(0)}\}$
set an initial temperature, Te, a temperature length, TL, and a cooling speed, α
$\ell \leftarrow 0$
repeat
  for $i = 1$ `to` $\lfloor TL/a \rfloor$ **do**
    apply Selim and Alsultan's (1991) algorithm to randomly draw a neighboring point, $\{\boldsymbol{\gamma}^{*}, \boldsymbol{U}^{*}\}$, of $\{\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)}\}$
    compute $\Delta F_{N,T} = F_{N,T}(\boldsymbol{\gamma}^{*}, \boldsymbol{U}^{*}) - F_{N,T}(\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{U}^{(\ell)})$
    if $\Delta F_{N,T} \leq 0$ **then**
      $\boldsymbol{\gamma}^{(\ell)} \leftarrow \boldsymbol{\gamma}^{*}$
      $\boldsymbol{U}^{(\ell)} \leftarrow \boldsymbol{U}^{*}$
    else
      randomly draw $q = \text{Uniform}(0, 1)$
      if $q < \exp(-\Delta F_{N,T}/Te)$ **then**
        $\boldsymbol{\gamma}^{(\ell)} \leftarrow \boldsymbol{\gamma}^{*}$
        $\boldsymbol{U}^{(\ell)} \leftarrow \boldsymbol{U}^{*}$
      end if
    end if
  end for
  $\ell \leftarrow \ell + 1$
  set a new 'cooling' temperature, $Te = Te \times \alpha$
until a stopping criterion is met
return the solution corresponding to the minimum function
end procedure

$\widehat{\epsilon}^{2}_{*,t}\left(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{U}}\right)$, $t = 1, \ldots, T$, denote the estimated residuals associated with $\epsilon_{*,t}(\boldsymbol{\Theta}, \boldsymbol{U}_0)$, $t = 1, \ldots, T$.

7 Monte Carlo Study

7.1 Monte Carlo Design

This simulation study provides some evidence on the small-sample performance of the proposed estimator. The design is based on an ARDL(1,1) model, where the covariate can be $I(0)$ or $I(1)$, with errors being generated using linear/nonlinear SAR processes. Suppose that the covariate is $I(0)$, for a given error-generating process, two different sets of parameters are imposed on the ARDL model in order to examine the impact of the stability condition on the finite-sample performance of the estimator; the same experiment is also replicated for the case where the covariate is $I(1)$. To be specific, we consider the following data generating process (d.g.p.) with four heterogeneous groups:

$$\Delta y_{\boldsymbol{s}_i,t} = \phi_i(y_{\boldsymbol{s}_i,t-1} - \theta_i x_{\boldsymbol{s}_i,t}) + \lambda_i \Delta y_{\boldsymbol{s}_i,t-1} + \gamma_i \Delta x_{\boldsymbol{s}_i,t} + \mu_i + \epsilon_{\boldsymbol{s}_i,t},\ i = 1, \ldots, 4, \tag{7.1}$$

where $\boldsymbol{s}_i = (s_{i,1}, s_{i,2})$ indicates a location on a rectangular lattice, say V_i; the covariate $x_{\boldsymbol{s}_i,t}$ takes either

$$x_{\boldsymbol{s}_i,t} = \begin{cases} 0.6x_{\boldsymbol{s}_i,t-1}+\eta_{\boldsymbol{s}_i,t} & \text{if } |x_{\boldsymbol{s}_i,t-1}|<1, \\ -0.6x_{\boldsymbol{s}_i,t-1}+\eta_{\boldsymbol{s}_i,t} & \text{if } |x_{\boldsymbol{s}_i,t-1}|\geq 1 \end{cases} \tag{7.2}$$

or the unit-root process

$$x_{\boldsymbol{s}_i,t} = x_{\boldsymbol{s}_i,t-1} + \eta_{\boldsymbol{s}_i,t}. \tag{7.3}$$

In the first scenario, it is assumed that the errors are generated by linear SAR processes. To specify the error-generating processes, note that the lattice V_i has a lexicographical order, thus there exists a bijection between the elements of V_i and the counting set $\{1, 2, \ldots, |V_i|\}$. The errors $\epsilon_{\boldsymbol{s}_i,t},\ \boldsymbol{s}_i \in V_i,\ t = 1, \ldots, T$, are generated by a linear SAR process, which can then be represented as

$$\epsilon_{\ell_i,t} = \rho_i \sum_{h_i=1, h_i\neq \ell_i}^{|V_i|} w_{\ell_i,h_i} \epsilon_{h_i,t} + e_{\ell_i,t},\ \ell_i \in \{1, 2, \ldots, |V_i|\},\ i = 1, \ldots, 4, \tag{7.4}$$

where $e_{\ell_i,t} \overset{i.i.d.}{\sim} N(0, \sigma_i^2)$ with $\sigma_i^2 \overset{i.i.d.}{\sim}$ Uniform$(0.5, 1.5)$ and $\boldsymbol{W}_i = \{w_{\ell_i,h_i}\},\ \ell_i \in \{1, 2, \ldots, |V_i|\},\ h_i \in \{1, 2, \ldots, |V_i|\}$, is a spatial weight matrix; similarly, the d.g.p. for $\eta_{\boldsymbol{s}_i,t},\ \boldsymbol{s}_i \in V_i,\ t = 1, \ldots, T$, is given by

$$\eta_{\ell_i,t} = -\rho_i \sum_{h_i=1, h_i\neq \ell_i}^{|V_i|} w_{\ell_i,h_i} \eta_{h_i,t} + \xi_{\ell_i,t},\ \ell_i \in \{1, 2, \ldots, |V_i|\},\ i = 1, \ldots, 4, \tag{7.5}$$

where $\xi_{\ell_i,t} \overset{i.i.d.}{\sim} N(0, \sigma_i^2)$ with $\sigma_i^2 \overset{i.i.d.}{\sim}$ Uniform$(0.5, 1)$.

In the second scenario, it is assumed that the errors are generated by nonlinear spatial autoregressions; the following d.g.p.'s are similar to the one used by Hallin, Lu, and Tran (2004):

$$\epsilon_{s_{i,1},s_{i,2},t} = \sin\left(\epsilon_{s_{i,1}-1,s_{i,2},t} + \epsilon_{s_{i,1},s_{i,2}-1,t} + \epsilon_{s_{i,1}+1,s_{i,2},t} + \epsilon_{s_{i,1},s_{i,2}+1,t}\right) + e_{s_{i,1},s_{i,2},t}, \tag{7.6}$$

$$\eta_{s_{i,1},s_{i,2},t} = \sin\left(\eta_{s_{i,1}-1,s_{i,2},t} + \eta_{s_{i,1},s_{i,2}-1,t} + \eta_{s_{i,1}+1,s_{i,2},t} + \eta_{s_{i,1},s_{i,2}+1,t}\right) + \xi_{s_{i,1},s_{i,2},t}, \tag{7.7}$$

where, for every $\boldsymbol{s}_i = (s_{i,1}, s_{i,2}) \in V_i$, $e_{s_{i,1},s_{i,2},t} \overset{i.i.d.}{\sim} N(0, \sigma_{\boldsymbol{s}_i}^2)$, $\sigma_{\boldsymbol{s}_i} \overset{i.i.d.}{\sim}$ Uniform$(0.5, 1)$ and $\xi_{s_{i,1},s_{i,2},t} \overset{i.i.d.}{\sim} N(0, \sigma_{\boldsymbol{s}_i}^2)$, $\sigma_{\boldsymbol{s}_i} \overset{i.i.d.}{\sim}$ Uniform$(0.5, 1.5)$, are independently generated.

With the d.g.p.'s defined above in mind, we conduct the following Monte Carlo experiments:

Experiment 1: Data are generated according to (7.1), (7.2), (7.4) and (7.5). The spatial weight matrices $\boldsymbol{W}_i,\ i = 1, \ldots, 4$, are of rook-contiguity (or queen-contiguity) type, constructed from actual maps of counties in the four U.S. states: Georga ($i = 1$), Kansas ($i = 2$), Missouri ($i = 3$), and

Texas ($i = 4$). In this experiment the numbers of 'neighbouring' counties in these states are set to $N_1 = |V_1| = 45$, $N_2 = |V_2| = 30$, $N_3 = |V_3| = 30$, and $N_4 = |V_4| = 70$ respectively, and the following sets of parameters will be used:

$$\{\phi_1, \phi_2, \phi_3, \phi_4\} = \{-0.9, -0.5, -0.2, -0.7\},$$

$$\{\theta_1, \theta_2, \theta_3, \theta_4\} = \{-2., -1., 1., 8.\},$$

$$\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{-1., -0.05, 0.05, 1.\},$$

$$\{\gamma_1, \gamma_2, \gamma_3, \gamma_4\} = \{-1., -0.04, 0.04, 1.\},$$

$$\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{-0.05, 0.05, -1., 1.\},$$

$$\{\rho_1, \rho_2, \rho_3, \rho_4\} = \{0.4, 0.05, 0.6, 0.1\}.$$

This experiment illustrates the situation whereby the stability condition nearly breaks down.

Experiment 2: This experiment is similar to Experiment 1 except that the numbers of 'neighbouring' counties in the above-mentioned states are now set equal to $|V_1| = 100$, $|V_2| = 60$, $|V_3| = 65$, and $|V_4| = 150$. The experiment demonstrates how the proposed estimator performs as the cross-sectional dimension grows relative to the number of time periods.

Experiment 3: This experiment is similar to Experiment 1 except for the set of parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{-0.5, -0.05, 0.05, 0.5\}$. This experiment illustrates the situation whereby the stability condition certainly holds true.

Experiment 4: This experiment is the same as Experiment 3 except that the numbers of 'neighbouring' counties specified in Experiment 2 are being used.

Experiment 5: Data are generated according to (7.1), (7.2), (7.6) and (7.7). The same sets of parameters specified in Experiments 1 and 2 are being used. This experiment illustrates the robustness of the proposed estimator when the error-generating d.g.p.'s change. It is important to note at this point that simulating sample paths from a nonlinear SAR, such as (7.6) or (7.7), is not a straight-forward task. Since the principle of contraction mapping warrants that the trigonometric sine function has a fixed point, one could simulate the processes (7.6) and (7.7) using the fixed-point iteration method. We shall briefly describe the algorithm to simulate (7.6) as (7.7) can be simulated in the same way. For each $i \in \{1, 2, 3, 4\}$ and $t \in \{1, \dots, T\}$, to generate $m_i \times n_i$ observations of $\epsilon_{\boldsymbol{s}_i,t}$ on a rectangular region, one can perform the following steps.

Step 1: Set all the initial values of $\epsilon_{\boldsymbol{s}_i,t}$ to zero and generate an array, $\{e_{s_{i,1},s_{i,2},t}\}_{\substack{s_{i,1}=1,\dots,100+m_i \\ s_{i,2}=1,\dots,100+n_i}}$, of mixed-normal random variables.

Step 2: Start from the values generated in Step 1 the process is iterated, say 30 times, for example,

$$\epsilon^{(k+1)}_{s_{i,1},s_{i,2},t} = \sin\left(\epsilon^{(k)}_{s_{i,1}-1,s_{i,2},t} + \epsilon^{(k)}_{s_{i,1},s_{i,2}-1,t} + \epsilon^{(k)}_{s_{i,1}+1,s_{i,2},t} + \epsilon^{(k)}_{s_{i,1},s_{i,2}+1,t}\right) + e_{s_{i,1},s_{i,2},t},\ k = 1, \ldots, 29.$$

Step 3: Take $\{\epsilon^{(30)}_{s_{i,1},s_{i,2},t}\}_{\substack{s_{i,1}=75,\ldots,74+m_i \\ s_{i,2}=75,\ldots,74+n_i}}$ as the simulated sample (and discard $\{\epsilon^{(30)}_{s_{i,1},s_{i,2},t}\}_{\substack{s_{i,1}=1,\ldots,74 \\ s_{i,2}=1,\ldots,74}}$ to allow for a warming-up zone.

Experiment 6: We repeat all the above experiments with (7.2) being replaced by (7.3) to examine the finite-sample performance of the estimator when the covariate is nonstationary.

7.2 Monte Carlo Results

7.2.1 Known group memberships

In this case, group memberships of individuals are known, thus no group classification is needed. *Stationary Covariate:* The vector of true parameters defined in Experiment 1 indicates that the stability condition does not hold in Group $i = 4$. Both the simulated biases and MSEs of the estimates shrink to zero slowly even for a large number of time periods in both small and large spatial groups; and the estimates in Groups $i = 2$ and 3 seem to be much less biased than in Groups $i = 1$ and 4 where the stability condition does not strictly hold. This pattern still persists for a variety of d.g.p.'s generating errors (cf. the first two panels in Tables 1-6).

For the true parameters defined in Experiment 3, both the simulated biases and MSEs are small for relatively large numbers of time periods in both small and large groups. However, comparing the last two panels of Tables 4 and 5 the biases are clearly less severe for the case with nonlinear SAR errors than the case with linear SAR errors, especially when the group sizes are large.
Nonstationary Covariate: The simulated biases and MSEs of the estimates of the long-run slope coefficients in Groups $i = 1, 2$ and 3 seem not much affected by the failure of the stability condition in Group $i = 4$. This is particularly true for the case with nonlinear SAR errors (cf. Tables 7-12). When the d.g.p.'s for all the groups are stable the simulated biases and MSEs become infinitesimal when the errors follow nonlinear SAR processes.

7.2.2 Unknown group memberships

We implement the VNS-DCA procedure to minimize the criterion function. To measure the performance of the VNS-DCA as a clustering algorithm, we report the Rand index in Table 13. The Rank index (named after William M. Rand) measures the number of pairwise agreements. For every unit, say x_i, $i = 1, \ldots, N$, let $G_I(x_i)$ represent its initial group label, and $G_C(x_i)$ represent its group label obtained from a clustering algorithm. According to Rand (1971) the Rand index is defined, in mathematical terms, as

$$\text{RandI} = \frac{a+d}{a+b+c+d},$$

where

$$\begin{aligned} a &= \left|\{i, j \in [1, N]: \ G_I(x_i) = G_I(x_j) \text{ and } G_C(x_i) = G_C(x_j)\}\right|, \\ b &= \left|\{i, j \in [1, N]: \ G_I(x_i) = G_I(x_j) \text{ and } G_C(x_i) \neq G_C(x_j)\}\right|, \\ c &= \left|\{i, j \in [1, N]: \ G_I(x_i) \neq G_I(x_j) \text{ and } G_C(x_i) = G_C(x_j)\}\right|, \\ d &= \left|\{i, j \in [1, N]: \ G_I(x_i) \neq G_I(x_j) \text{ and } G_C(x_i) \neq G_C(x_j)\}\right|. \end{aligned}$$

Note that RandI $\in [0, 1]$, where '0' indicates that the two clusters of data do not agree on any pair of points, and '1' indicates that the two clusters bearing possibly different labels are exactly the same.

Suppose data are generated by the d.g.p. defined by (7.1) and (7.2) with SAR errors using queen-contiguity weights. Table 13 reports improved simulated RandI's as the number of sampled locations increases. Therefore the VNS-DCA performs clustering computations efficiently in Experiments 3 and 4. The number of repetitions in each simulation is 500; and most of the computational time is spent on finding 'good' starting points through implementing the VNS algorithm while the DCA performs quite efficiently (usually converges to an optimum after about 800 to 1500 iterations). The computational time increases polynomially with the number of time periods. According to Tables 13 and 14, the proposed estimators perform well in terms of both biases and mean squared errors.

In addition, Tables 17 and 18 report the finite-sample performance of the proposed procedure when data are generated by (7.1) with the covariate following a unit-root process (7.3) and SAR errors using queen-continuity weights. The Rand index clearly improves as the number of time periods increases in comparison with the case when the covariate follows a stationary process (cf. Tables 15 and 16). The empirical biases and MSE's also have much faster decay rates in this case, especially for big clusters. Therefore the method could perform really well when covariates are nonstationary. The same simulations using SAR errors with rook-continuity weights are repeated for data generated from the d.g.p. (7.1) and (7.3); Results in Tables 19 and 20 show even smaller biases and mean squared errors, confirming that the rates of convergence significantly depend on patterns of weak spatio-temporal dependence as conjectured by the main theorems.

8 Empirical Application

An open economy can effectively finance its investment by borrowing abroad since domestic saving [as the main source of funds for investment] flow to wherever there are profitable investment projects. Therefore, high correlation between domestic saving and investment - both measured as percentages of gross domestic product (GDP) - empirically established in a regression model for open economies is well known as the Feldstein-Horioka puzzle (henceforth FHP). This puzzle started when Feldstein and Horioka (1980) (FH) showed, by using the cross-section data of 16 Organization for Economic Cooperation and Development (OECD) economies for the period 1960-1974, that temporally averaged national saving and domestic investment were highly correlated. They interpreted this high long-run correlation as an evidence of low international capital mobility. The FHP - which Obstfeld and Rogoff (2001) view as one of the six major puzzles in international macroeconomics - still persists as estimates of the saving-investment (SI) association for small open economies have remained quite high despite ongoing financial market integration and globalization over recent decades (see, e.g., Chang and Smith (2014)). The question as to whether the apparently high capital mobility is a chimera or an elusive reality is still attracting much attention because capital mobility is critical both for the efficient allocation of capital to the most productive locations and for consumption smoothing. It is also relevant for policy issues such as large current account deficits or the role of net overseas balances.

Another plausible interpretation of the close long-run relationship between the investment and saving ratios [first established by Feldstein and Horioka (1980)] is provided by Coakley, Kulasi, and Smith (1996); Jansen (1996). They argued that, since saving and investment behave like unit-root processes, the long-run SI correlation should reflect the intertemporal budget constraint or solvency constraint, which essentially requires that the current account (saving minus investment) must be a stationary process as debt cannot explode. This solvency constraint in turn implies that saving and investment are cointegrated with a unit cointegrating vector. As a result the long-run SI correlation in a cross-section regression should be equal to one. Thus, it may well be that the FH coefficient is not a puzzle, but merely a consequence of the solvency constraint. Jansen (1998) deems that the long-run correlation can provide a test of the relevance of the intertemporal budget constraint, which is one of the cornerstones of modern open-economy macroeconomics. Non-binding of the intertemporal budget constraint implies that the saving and investment rates are not cointegrated (i.e., saving and investment are not correlated in the long-run). This constitutes evidence in favour of international capital mobility by the Feldstein-Horioka criterion.

A variety of econometric specifications has been employed to estimate the SI-correlation . Jansen (1996) applies a vector error-correction model (ECM) - which is consistent with intertemporal general equilibrium models - to the OECD countries, and find that saving and investment are cointegrated across countries. However the degrees of long-run SI-correlation display some variety across countries when including more recent observations into the sample. This heterogeneity [in the SI relationship] between countries can be explained by differences in their economic structures, sizes, cyclical positions, government policies, and macroeconomic openness. To control for the potentially important effects of heterogeneity in saving and investment ratios, panel estimation techniques (such as the dynamic fixed-effects estimator, Pesaran and Smith's (1995) mean-group estimator, and Pesaran, Shin, and Smith's (1999) pooled mean-group estimator) are commonly employed (see, e.g., Coakley, Fuertes, and Spagnolo (2004); Pelgrin and Schich (2008)).

This empirical study revisits the long-run SI-relationship by applying the proposed CL estimation approach to a quarterly dataset consisting of 27 OECD countries (Australia , Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Israel, Italy, Japan, South Korea, Mexico, Netherlands, New Zealand, Norway, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, UK, USA) and four non-OECD countries and organizations (South Africa, the European Union (EU-28), Latvia, Costa Rica) from 1995 Q1 to 2015 Q2.[2] The period covered in this dataset is associated with the era when international capital movements and deregulation of domestic financial markets become more and more popular. Thus, one can expect that, for the countries under our study, the long-run relationship between saving and investment rates has been rather deteriorated.

In view of Jansen (1996) and Pelgrin and Schich (2008), we shall consider the following group-wise ECM with a maximum lag of one:

$$\Delta I_{i,t} = \alpha_i + \sum_{c=1}^{G} \phi_c u_{i,c} \left(I_{i,t-1} - \theta_c S_{i,t}\right) + \sum_{c=1}^{G} \gamma_c u_{i,c} \Delta S_{i,t} + \epsilon_{i,t}, \tag{8.1}$$

where $I_{i,t}$ and $S_{i,t}$ represent the investment rate and the saving rate of country i in period t; and $\epsilon_{i,t} \sim N(0, \sigma_i^2)$; α_i is the country-specific fixed effect; ϕ_c is the error-correction coefficient associated with group c; θ_c is the long-run SI-correlation coefficient associated with group c; γ_c is the short-run SI-correlation coefficient associated with group c; and $u_{i,c}$, $i = 1, \ldots, 31$ and $c = 1, \ldots, G$, are indicators of group memberships. The model (8.1) takes into consideration possible heterogeneity between groups of countries with common characteristics, economic policies, and structures by allowing for group variations in the SI-correlation coefficients, whereas, in many other studies, these coefficients are assumed to be either equal across countries when temporally pooling observations together (Feldstein and Horioka, 1980; Jansen, 1998), or completely different across countries (Coakley, Fuertes, and Spagnolo, 2004; Pelgrin and Schich, 2008). Other types of country-specific heterogeneity can also be accounted for by including fixed effects and error variances that

[2] All the data used for this empirical study are downloaded from the OECD data bank at http://www.oecd-ilibrary.org.

differ across countries.

We start by examining the persistence property of $I_{i,t}$ and $S_{i,t}$. Table 21 presents the augmented Dickey-Fuller (ADF) test results. The p-values reported are greater than the 5 percent level for virtually all series. These findings are consistent with the existing evidence that saving and investment ratios have their dynamics indistinguishable from unit-root processes.

Next, we conduct estimation and inference of the error-correction model (8.1). The VNS-DCA procedure searches for globally optimal points of the CL criterion function over the domains $[-2, 2]$ of θ's, $[-2, -0.001]$ of ϕ's, and $[-1, 1]$ of α's and γ's. The estimation results are reported in Tables 22 and 23. Since the composite error attains its minimum value when the number of groups G is 4, we shall then consider the case where there are 4 optimal groups. The estimate of the EC coefficient $\widehat{\phi}_4 \approx 0$ means that, in the fourth group of countries the saving and investment rates are not cointegrated in the long-run, this implies high international capital mobility by the initial interpretation of Feldstein and Horioka (1980). Since $\widehat{\phi}_4$ is very close to zero the estimate of the true θ_4 then becomes irrelevant; thus, this results in a wide confidence interval, $[-435.43, 444.79]$ (cf. Table 22). The estimates of the other EC coefficients ϕ_1, ϕ_2, and ϕ_3 are significantly different from zero, there exists a long-run relationship between the saving rate and investment rate. In the second group the estimate of the cointegrating vector is not much different from $(1, -1)$ the current account is stationary in the long-run. Therefore, according to Coakley, Kulasi, and Smith (1996) and Jansen (1996) the close long-run relationship between the saving and investment rates should be viewed as a solvency condition that must be satisfied rather than as evidence against capital immobility, thus no conclusion about capital mobility can be drawn for the countries in this group. In the first and third groups the estimates of the cointegrating vectors are significantly different from $(1, -1)$ the current account is non-stationary in the long-run. This result is evidence in favour of international capital mobility. Moreover, we can conclude - by inspecting the estimates for the short-run correlation coefficients γ_1 and γ_3 - that low short-run correlations imply that capital is sufficiently mobile in the countries belonging to the first and third groups. However the degree of capital mobility in these groups is less than in the fourth group.

In addition the geographical sketch of countries on the world map (cf. Figure 1) shows that there is a little neighborhood effect in the long-run SI-relationship. Nowadays, many countries can have common economic structure or fiscal policy due to trade linkages or globalization, not necessarily due to geographical closeness. As noticed in Figure 2, countries in the fourth group have the lowest capital control - this is consistent with our finding that there is no long-run relationship between the investment and saving ratios, thus one could expect high capital mobility in this group. In the first and second groups the capital control indices became rather high after the year of 2004, which provides moderate evidence in favour of capital mobility. Therefore, high average capital control indices for the first and second groups suggest that there are long-run SI-relationships, which is consistent with our finding reported earlier that capital is mobile to a certain degree in the first group whilst it remains inconclusive in the second group.

9 Conclusion

A common problem potentially arising when panel data are spatially dependent is that parameters are not homogeneous over space, but instead vary over different locations where an economic activity takes place. The present paper deals with this type of issue by estimating an error-correction form of an ARDL model that allows for group-specific patterns of unobserved heterogeneity. The inference procedure is based on maximum composite likelihood, thus can bypass full specification of the variance-covariance matrix for the error term, which is often required in the traditional maximum [joint] likelihood paradigm. It is demonstrated [through asymptotic theory and Monte Carlo simulation] that the proposed estimator is asymptotically valid and has good finite-sample performance. The compelling issue of choosing the optimal number of groups and the optimal way of grouping can also be dealt with by using the analogues - based on the composite likelihood function - of the AIC and the BIC. Group-specific time patterns of heterogeneity can of course be allowed, but this complication is not discussed in this current study.

10 Software

A GUI software package to implement the method proposed in this paper can be downloaded from http://http-server.carleton.ca/~bchu/ecmg.htm (source code available upon request).

Table 1: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
Experiment 1 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.86109	-0.106376	0.068128	-0.212314	-0.06528	0.040126	0.013735	0.333403	0.207969
150	-0.72167	-0.088056	0.045632	-0.218835	-0.05005	0.042901	-0.017278	0.309425	0.197628
250	-0.65316	-0.077994	0.001431	-0.191006	-0.02688	0.045023	0.001188	0.294455	0.188259
350	-0.64158	-0.044944	0.065820	-0.216888	-0.00593	0.042386	0.011435	0.299321	0.178516
450	-0.66119	-0.080777	0.038246	-0.214588	-0.03850	0.048995	0.019717	0.332659	0.143224
550	-0.78069	-0.103053	-0.004172	-0.199955	-0.00401	0.061558	-0.010862	0.318524	0.217725
650	-0.60281	-0.077536	-0.008501	-0.176315	-0.04533	0.046447	0.019646	0.292884	0.178645
750	-0.56149	-0.023177	0.036038	-0.196446	-0.03419	0.074197	-0.039323	0.314156	0.138525
Experiment 2 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.943347	0.523168	0.234133	-0.223378	-1.54787	0.046866	-0.029766	-0.70695	0.35397
150	-0.673831	-0.111102	0.057198	-0.213776	-0.06720	0.034929	-0.011663	0.32481	0.16215
250	-0.564914	-0.101189	0.079891	-0.218222	-0.01697	0.052751	0.033071	0.32475	0.16525
350	-0.620736	-0.073731	0.024378	-0.221785	-0.05171	0.023342	-0.014370	0.31309	0.15915
450	-0.688307	-0.095300	0.016818	-0.245358	0.02800	0.021830	-0.000344	0.31039	0.19924
550	-0.670902	-0.093836	0.068862	-0.245341	-0.04130	0.044821	-0.035420	0.32380	0.18687
650	-0.690446	-0.093473	0.012114	-0.195285	-0.05326	0.038125	-0.004501	0.30867	0.20244
750	-0.731355	-0.075748	0.033510	-0.193993	-0.07986	0.0449219	-0.029733	0.32392	0.17823
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.110187	-0.067890	-0.078441	-0.001113	-0.250219	-0.535216	0.082166	0.060488	0.015810
150	-0.045497	-0.023332	-0.024274	-0.001698	0.006884	-0.194721	0.052197	-0.015488	0.020460
250	-0.019351	-0.013274	-0.016588	-0.001520	-0.028486	-0.158752	0.038874	-0.026079	0.021075
350	-0.013855	-0.009533	-0.012156	-0.000485	-0.016486	-0.087744	0.042756	-0.004560	0.021392
450	-0.009807	-0.009852	-0.010282	-0.001130	-0.028529	-0.047788	0.039635	-0.002043	0.021594
550	-0.004898	-0.005949	-0.007844	-0.000706	-0.031241	-0.039467	0.039654	0.004700	0.021781
650	-0.002667	-0.001809	-0.006196	-0.000637	-0.041310	-0.065682	0.037479	0.015756	0.021915
750	-0.001270	-0.003121	-0.006111	-0.000585	-0.043350	-0.022452	0.035340	0.022582	0.021924
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.092588	-0.078116	-0.077676	-0.004888	-0.544623	-0.098329	0.060390	0.106963	0.015661
150	-0.033139	-0.021422	-0.027837	-0.000733	-0.059642	-0.035168	0.040841	-0.005210	0.019703
250	-0.016082	-0.011987	-0.017922	0.000818	-0.049598	0.009395	0.033612	0.020992	0.020504
350	-0.008162	-0.005361	-0.012352	0.000812	-0.066102	-0.003925	0.031797	0.026391	0.020996
450	-0.000513	-0.004085	-0.009919	0.000457	-0.082942	0.042681	0.030141	0.027726	0.021239
550	0.003646	-0.004057	-0.008363	0.000633	-0.084011	0.045922	0.025307	0.044636	0.021374
650	0.006919	-0.003900	-0.007256	0.000152	-0.094483	0.066114	0.026684	0.035941	0.021503
750	0.006309	-0.001032	-0.005382	7.27E-05	-0.091688	0.040157	0.029304	0.043090	0.021661

Table 2: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
Experiment 1 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	2.061030	0.238045	0.226894	0.339639	0.192405	0.161238	0.216442	0.248103	0.228033	7488.58
150	2.099090	0.261776	0.150218	0.246824	0.187279	0.168096	0.230186	0.202591	0.162543	1.71E+41
250	2.256810	0.226814	0.179554	0.226734	0.198282	0.159402	0.205081	0.230600	0.243666	7.44E+78
350	2.156730	0.256681	0.173522	0.261128	0.196061	0.149078	0.220355	0.219168	0.192356	4.04E+116
450	1.839240	0.278468	0.169363	0.291816	0.200510	0.173884	0.216671	0.252268	0.171778	2.73E+154
550	2.198300	0.239358	0.216220	0.223929	0.186552	0.185950	0.252353	0.243869	0.209425	1.24E+192
650	2.132060	0.254277	0.205297	0.206942	0.179464	0.196177	0.222694	0.194514	0.192571	8.36E+229
750	1.860390	0.197142	0.172236	0.209194	0.180146	0.174973	0.214772	0.219093	0.178035	5.53E+267
Experiment 2 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	2.384320	212.738000	13.758200	0.269305	1152.860000	0.160198	0.351016	549.327000	10.862600	6559.82
150	1.982950	0.253339	0.145084	0.214627	0.182533	0.142790	0.205465	0.222483	0.169886	1.56E+41
250	2.426260	0.271670	0.238700	0.238971	0.200262	0.154655	0.256025	0.233882	0.160257	6.49E+78
350	2.000270	0.223061	0.162757	0.241096	0.201454	0.140084	0.237468	0.210311	0.163328	3.63E+116
450	2.582510	0.252512	0.202008	0.275982	0.256633	0.189170	0.317486	0.239224	0.218675	2.19E+154
550	2.426350	0.238815	0.185976	0.265754	0.200350	0.139763	0.222850	0.226840	0.166296	1.23E+192
650	2.100940	0.273460	0.159649	0.206538	0.211218	0.186339	0.200136	0.211990	0.205575	8.83E+229
750	2.182040	0.256051	0.181827	0.217218	0.189705	0.187588	0.249155	0.228272	0.189553	5.08E+267
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.368280	0.463023	0.049834	0.007872	6.653660	49.719200	66.073800	2.199690	0.059562	0.010184
150	0.081101	0.100259	0.008181	0.001855	0.542767	2.508030	6.304430	0.468265	0.048007	0.013849
250	0.044113	0.052732	0.004153	0.001073	0.277699	1.387700	2.670080	0.246744	0.047652	0.014541
350	0.029081	0.040395	0.002893	0.000740	0.175204	0.778334	1.813750	0.156959	0.048048	0.014838
450	0.022085	0.030767	0.002141	0.000578	0.133426	0.450826	1.246650	0.106520	0.048363	0.014997
550	0.017186	0.024377	0.001691	0.000491	0.103069	0.396044	0.961720	0.091526	0.048825	0.01509
650	0.014513	0.019639	0.001408	0.000424	0.083507	0.269695	0.741135	0.071040	0.049166	0.015162
750	0.011942	0.016315	0.001262	0.000364	0.075899	0.203642	0.674555	0.061651	0.049069	0.015225
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.336490	0.499184	0.045415	0.008392	29.993100	38.285400	73.256500	1.995760	0.057128	0.004828
150	0.066222	0.111923	0.008543	0.002018	0.487734	3.733990	5.550900	0.415036	0.045350	0.006514
250	0.037101	0.060085	0.004390	0.001123	0.226384	1.019560	2.447260	0.204707	0.045568	0.006854
350	0.027267	0.039802	0.003075	0.000786	0.150483	0.665346	1.417690	0.129436	0.046570	0.006975
450	0.020279	0.031139	0.002171	0.000599	0.117064	0.431270	1.045770	0.103245	0.046913	0.007063
550	0.016296	0.023378	0.001734	0.000485	0.092761	0.320943	0.708957	0.077289	0.047121	0.007114
650	0.013217	0.019180	0.001400	0.000396	0.080031	0.235476	0.679439	0.061645	0.047398	0.007150
750	0.011669	0.016570	0.001143	0.000354	0.070570	0.234721	0.565636	0.054522	0.047895	0.007168

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 3: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
Experiment 1 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.858399	-0.088697	0.051561	-0.251999	-0.073639	0.033987	-0.002832	0.316786	0.173743
150	-0.701195	-0.076070	0.037866	-0.228812	-0.043758	0.053390	0.016597	0.309517	0.168948
250	-0.663698	-0.041974	0.044686	-0.213368	-0.019480	0.040885	-0.004228	0.316052	0.177749
350	-0.698057	-0.087288	0.028046	-0.225867	-0.065472	0.054591	0.005902	0.300461	0.165818
450	-0.612329	-0.089999	0.027263	-0.213758	-0.035901	0.031392	0.006506	0.314442	0.171976
550	-0.699745	-0.069692	0.029242	-0.201872	-0.074569	0.058946	-0.011789	0.298964	0.176924
650	-0.623609	-0.042775	0.055336	-0.213672	-0.055918	0.053703	0.005600	0.299446	0.156117
750	-0.626769	-0.071302	0.021131	-0.195425	-0.063051	0.041090	0.016664	0.300204	0.162929
Experiment 2 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.763373	-0.110235	0.081199	-0.227623	-0.075237	0.037752	0.025481	0.324663	0.185544
150	-0.703738	-0.084416	0.015175	-0.227624	-0.048532	0.039578	-0.000360	0.300304	0.163391
250	-0.703967	-0.093266	0.062224	-0.196770	-0.033552	0.062668	0.021123	0.329863	0.175616
350	-0.670058	-0.114683	0.024481	-0.205503	-0.021432	0.044461	0.020606	0.317632	0.176906
450	-0.715018	-0.068980	0.058923	-0.223209	-0.040791	0.056332	-0.010557	0.321255	0.172536
550	-0.671723	-0.076533	0.050059	-0.234680	-0.044193	0.044087	0.021208	0.301685	0.164421
650	-0.662065	-0.117619	0.062229	-0.226658	-0.056909	0.030790	0.003325	0.327446	0.181650
750	-0.647779	-0.063958	0.049050	-0.221746	-0.032152	0.012884	-0.015911	0.323810	0.177286
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.087895	-0.073020	-0.074460	-0.001997	-0.241938	-0.643001	0.032662	0.025829	0.016142
150	-0.035869	-0.029701	-0.026189	0.000696	-0.024583	-0.075703	0.038675	0.020277	0.020158
250	-0.013167	-0.015026	-0.016090	0.001048	-0.034377	-0.187792	0.036391	0.029213	0.021046
350	-0.006550	-0.005223	-0.010406	0.000806	-0.047235	-0.131618	0.040012	0.008716	0.021520
450	0.000266	-0.002780	-0.006648	0.000274	-0.065099	-0.081072	0.030614	0.033891	0.021862
550	0.002312	0.001495	-0.005860	0.000100	-0.068602	-0.061908	0.030486	0.031070	0.021907
650	0.003062	0.000076	-0.004954	0.000158	-0.056065	-0.041058	0.034975	0.029493	0.021987
750	0.004951	0.005124	-0.004059	-0.000061	-0.070581	-0.039654	0.035146	0.043064	0.022064
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.089830	-0.101346	-0.078642	-0.006556	-0.297800	-0.486428	0.032444	0.107005	0.015765
150	-0.018141	-0.042031	-0.023972	-0.002195	-0.075558	0.048009	0.031493	0.015916	0.020249
250	-0.014801	-0.024221	-0.011953	-0.001133	-0.046614	0.011725	0.030177	0.008297	0.021163
350	-0.006422	-0.016582	-0.006240	-0.000029	-0.050570	0.038305	0.038714	0.025374	0.021597
450	-0.000105	-0.018953	-0.005170	-0.000627	-0.066211	0.063712	0.030313	0.029299	0.021733
550	-0.002778	-0.014857	-0.003442	0.000424	-0.072256	0.047690	0.030318	0.035364	0.021815
650	0.000942	-0.009095	-0.003085	0.000752	-0.067760	0.054388	0.030497	0.033301	0.021825
750	0.004076	-0.004472	-0.003347	0.000480	-0.092348	0.063148	0.027805	0.039053	0.021818

Table 4: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
Experiment 1 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	1.874520	0.295715	0.543656	0.491988	0.233785	0.283593	0.249145	0.313727	0.247791	7329.74
150	2.117680	0.269733	0.183599	0.228730	0.186245	0.183630	0.235978	0.228160	0.184288	1.62E+41
250	2.001020	0.218764	0.175765	0.242151	0.180947	0.148920	0.198553	0.226997	0.184359	6.98E+78
350	1.957610	0.261443	0.181741	0.233389	0.197940	0.170263	0.251513	0.209617	0.175732	3.67E+116
450	1.946160	0.237436	0.173029	0.232767	0.176523	0.154068	0.224669	0.209865	0.179699	2.17E+154
550	2.210790	0.272458	0.179004	0.249963	0.211208	0.177857	0.214837	0.224487	0.232034	1.35E+192
650	1.853950	0.249563	0.196796	0.253207	0.177171	0.168365	0.235829	0.205364	0.184621	7.57E+229
750	1.784650	0.241543	0.179591	0.231717	0.187445	0.154874	0.214978	0.205378	0.183574	5.62E+267
Experiment 2 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	1.898220	0.545762	0.331516	0.836769	0.231477	0.328107	0.274496	0.303754	0.266136	5646.86
150	2.000690	0.230504	0.169907	0.229354	0.178778	0.156767	0.217029	0.212232	0.168470	1.39E+41
250	2.360900	0.274960	0.179689	0.229213	0.201960	0.167710	0.242299	0.241965	0.191065	6.53E+78
350	2.139580	0.272874	0.174808	0.208555	0.190818	0.172576	0.217213	0.227648	0.190945	3.30E+116
450	2.308180	0.243625	0.202452	0.249083	0.236306	0.206432	0.214332	0.241470	0.173112	1.90E+154
550	2.224590	0.282695	0.192194	0.250986	0.203446	0.176593	0.270079	0.216051	0.182502	1.27E+192
650	2.356580	0.293268	0.185516	0.240022	0.207784	0.154649	0.290928	0.227304	0.176041	8.12E+229
750	2.223230	0.253720	0.161511	0.243489	0.184041	0.161936	0.246289	0.230593	0.192008	4.80E+267
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.317611	0.447992	0.047439	0.007840	7.398800	64.397500	60.899000	2.283360	0.059775	0.010255
150	0.071471	0.100726	0.008547	0.001971	0.500439	4.236710	7.406730	0.440798	0.047415	0.013861
250	0.040180	0.057286	0.004354	0.001186	0.251307	3.611480	4.024490	0.243155	0.047813	0.014503
350	0.026722	0.037043	0.003065	0.000835	0.173618	1.210100	1.870260	0.152877	0.048825	0.014841
450	0.020104	0.028302	0.002069	0.000626	0.127250	0.559407	1.210540	0.110200	0.049576	0.015015
550	0.017489	0.022834	0.001696	0.000486	0.103609	0.349531	0.887500	0.084307	0.049425	0.015122
650	0.014373	0.019724	0.001510	0.000403	0.087617	0.267190	0.742428	0.070074	0.049601	0.015187
750	0.012802	0.016847	0.001254	0.000355	0.077922	0.221034	0.716152	0.057015	0.049738	0.015233
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.312760	0.470171	0.044835	0.007728	6.387040	33.734000	40.105900	2.292450	0.056954	0.004917
150	0.066093	0.101020	0.008108	0.001746	0.495119	2.529910	4.829530	0.428887	0.047303	0.006548
250	0.037576	0.058946	0.003914	0.001070	0.225720	1.263170	2.471900	0.215994	0.048016	0.006895
350	0.024298	0.041783	0.002741	0.000780	0.135799	0.712740	1.624060	0.140107	0.048951	0.007018
450	0.019969	0.031458	0.002101	0.000592	0.105571	0.353870	0.986849	0.109316	0.049040	0.007092
550	0.015469	0.025272	0.001708	0.000501	0.091352	0.343692	0.874810	0.085894	0.049069	0.007128
650	0.013166	0.021070	0.001394	0.000429	0.078976	0.236987	0.667492	0.061006	0.048838	0.007163
750	0.011296	0.017499	0.001201	0.000370	0.071569	0.211944	0.526128	0.055434	0.048630	0.007188

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 5: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Nonlinear SAR Errors

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
True Parameters defined in Experiment 1 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)									
50	-0.970693	-0.069464	0.085820	-0.229508	-0.094959	0.049079	0.018548	0.307814	0.172932
150	-0.782395	-0.145234	0.006716	-0.257922	-0.012191	0.040787	0.036339	0.322878	0.142735
250	-0.598178	-0.029355	0.014677	-0.195009	-0.034324	0.082560	0.070225	0.255610	0.185074
350	-0.552613	-0.049905	0.016743	-0.179277	-0.039556	0.042632	0.072894	0.296988	0.149054
450	-0.655592	-0.033425	0.019777	-0.158854	-0.075098	0.044455	0.007561	0.302537	0.185796
550	-0.698142	-0.031075	0.045309	-0.167126	-0.004099	0.058869	0.019851	0.295063	0.187364
650	-0.708413	-0.047610	0.001637	-0.172818	-0.053839	0.073108	0.022735	0.320742	0.141764
750	-0.556860	0.007222	0.035597	-0.243045	-0.022273	0.014504	0.031651	0.263577	0.154588
True Parameters defined in Experiment 1 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)									
50	-0.732286	-0.096945	0.057881	-0.132431	-0.057375	0.051464	-0.024427	0.360652	0.197779
150	-0.722716	-0.072764	0.061557	-0.235562	-0.056656	0.043586	-0.035096	0.321864	0.176154
250	-0.723979	-0.110542	0.051729	-0.232778	-0.036237	0.039303	-0.026582	0.337413	0.165967
350	-0.732472	-0.078580	0.024620	-0.190927	-0.044191	0.035718	0.028698	0.305701	0.204090
450	-0.766140	-0.089992	0.072140	-0.253662	-0.029252	0.025863	0.008322	0.334413	0.193960
True Parameters defined in Experiment 3 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)									
50	-0.088561	-0.072839	-0.078064	-0.001684	-0.207593	-0.623717	0.697244	0.094563	-0.097617
150	-0.032329	-0.028899	-0.026174	0.000246	-0.080181	-0.024025	0.013967	0.009645	-0.032925
250	-0.024551	-0.010698	-0.015302	0.000484	-0.040640	-0.025790	-0.002152	0.026099	-0.019623
350	-0.012196	-0.007457	-0.009191	-0.000175	-0.038071	-0.004878	-0.000862	0.016825	-0.011662
450	-0.009700	-0.008063	-0.005909	0.000290	-0.027451	0.000672	0.004944	0.004925	-0.007727
550	-0.009402	-0.007019	-0.006708	-0.000145	-0.021847	0.022580	0.017967	0.004376	-0.008769
True Parameters defined in Experiment 3 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)									
50	-0.110275	-0.103319	-0.018277	0.005745	-0.192519	-0.119767	0.190836	0.096212	-0.025917
150	-0.006048	-0.019684	-0.017021	0.003994	-0.068541	-0.110964	0.042790	0.080071	-0.022804
250	-0.001276	-0.004757	-0.011416	-0.000747	-0.051512	0.007802	0.004713	0.028922	-0.014156
350	-0.004430	-0.001101	-0.010341	0.000670	-0.017297	-0.039598	0.000488	0.037260	-0.013522
450	-0.005908	-0.001769	-0.006414	0.001687	-0.002752	-0.013991	-0.026319	0.029580	-0.009105
550	-0.004151	-0.003921	-0.005579	0.002284	-0.008911	-0.019388	-0.011917	0.020173	-0.007925
650	-0.006022	-0.002770	-0.004142	0.002318	-0.010583	-0.016669	-0.000291	0.019550	-0.006171
750	0.003963	-0.005300	-0.004837	0.002129	-0.014054	-0.006079	-0.012081	0.011144	-0.006637

Table 6: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Stationary Covariate and Nonlinear SAR Errors

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
				True Parameters defined in Experiment 1 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)						
50	1.728210	0.288209	0.217136	0.212903	0.196828	0.155598	0.253231	0.251616	0.247764	16178.8
150	1.876970	0.293202	0.213753	0.264166	0.255835	0.176938	0.262756	0.216224	0.166469	2.63E+41
250	1.219210	0.193416	0.149974	0.181570	0.175572	0.131969	0.222332	0.159459	0.173545	1.42E+79
350	1.038330	0.212277	0.171436	0.142657	0.150073	0.160658	0.173758	0.166489	0.153584	5.59E+116
450	1.552950	0.209377	0.222323	0.235925	0.203917	0.153663	0.248256	0.207164	0.176115	3.65E+154
550	1.568460	0.222412	0.153130	0.173305	0.171814	0.124806	0.182638	0.207226	0.172342	1.97E+192
650	1.402050	0.219868	0.155079	0.202297	0.161028	0.147972	0.200994	0.216402	0.166047	1.19E+230
750	1.438890	0.208126	0.162966	0.260664	0.194095	0.118434	0.202310	0.188357	0.155805	8.89E+267
				True Parameters defined in Experiment 1 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)						
50	2.059520	0.257048	0.419943	2.839950	0.181442	0.390532	0.206021	0.893595	0.626746	6026.54
150	2.424370	0.276616	0.237308	0.267437	0.237295	0.186378	0.228351	0.232332	0.177481	1.53E+41
250	2.397670	0.297403	0.186172	0.268930	0.226620	0.168849	0.205104	0.227012	0.213569	6.64E+78
350	2.210270	0.240743	0.190967	0.212326	0.204623	0.151513	0.244701	0.234525	0.195267	4.00E+116
450	2.507860	0.228128	0.185655	0.301912	0.208411	0.174457	0.251623	0.227277	0.189432	1.99E+154
				True Parameters defined in Experiment 3 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)						
50	0.342224	0.136916	0.074713	0.006490	10.158500	35.619900	39.783300	1.884670	0.116173	0.011404
150	0.077128	0.032491	0.013254	0.001734	0.246708	1.200690	1.790750	0.323320	0.021248	0.015315
250	0.042963	0.016458	0.007537	0.001077	0.106464	0.604857	0.859245	0.183681	0.011936	0.016017
350	0.031918	0.011499	0.004946	0.000766	0.079473	0.239812	0.456663	0.122209	0.007954	0.016400
450	0.022800	0.008894	0.003639	0.000549	0.051940	0.179322	0.333514	0.081298	0.005781	0.016608
550	0.018347	0.006511	0.002975	0.000426	0.041856	0.126314	0.270491	0.058633	0.004719	0.016721
				True Parameters defined in Experiment 3 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)						
50	0.317147	0.131605	0.081244	0.006806	2.113940	7.739370	6.942440	1.338710	0.126109	0.001646
150	0.068957	0.031286	0.014025	0.001967	0.310101	2.115000	1.132980	0.278899	0.021727	0.002117
250	0.038976	0.016930	0.006883	0.001231	0.132943	0.186328	0.320449	0.124286	0.010899	0.002215
350	0.026769	0.010255	0.004615	0.000771	0.068841	0.132879	0.122417	0.071972	0.007242	0.002262
450	0.017107	0.007680	0.003376	0.000661	0.044838	0.084980	0.070107	0.041733	0.005256	0.002272
550	0.014116	0.005378	0.002433	0.000584	0.038652	0.064570	0.054273	0.028143	0.003817	0.002292
650	0.010824	0.004359	0.002240	0.000450	0.028168	0.051994	0.034113	0.027418	0.003422	0.002303
750	0.008219	0.004140	0.001906	0.000373	0.024981	0.039532	0.029416	0.017862	0.002869	0.002312

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 7: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
Experiment 1 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.772388	-0.060366	0.006447	-0.208354	-0.061439	0.020733	-0.032845	0.291188	0.199183
150	-0.695992	-0.045514	0.033029	-0.210665	-0.031436	0.032953	-0.004965	0.298536	0.183180
250	-0.692131	-0.067166	0.029266	-0.227474	-0.022225	0.028950	0.007652	0.314914	0.172948
350	-0.636684	-0.043484	0.044083	-0.218829	-0.049687	0.050792	0.030509	0.299835	0.167369
450	-0.667085	-0.030590	0.036145	-0.239834	-0.020722	0.028993	0.007589	0.304018	0.134115
550	-0.587248	-0.060256	0.031286	-0.217165	-0.046386	0.047504	0.015816	0.302195	0.159778
650	-0.595764	-0.088938	0.031584	-0.184563	-0.044686	0.044455	-0.000281	0.309229	0.184283
750	-0.642338	-0.058718	0.019819	-0.217283	-0.020445	0.043001	-0.009625	0.314011	0.174190
Experiment 2 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.771287	-0.078598	0.044942	-0.217584	-0.040968	0.036464	0.005318	0.314548	0.214595
150	-0.734493	-0.123720	0.044498	-0.216069	-0.036558	0.070029	0.005018	0.328518	0.167745
250	-0.676299	-0.086576	0.049178	-0.219873	-0.026853	0.044533	0.001522	0.314691	0.195681
350	-0.735008	-0.079657	0.026759	-0.204788	-0.034029	0.047367	-0.015848	0.325943	0.154354
450	-0.722427	-0.085812	0.069341	-0.227772	-0.012065	0.042268	0.004839	0.336346	0.176720
550	-0.639320	-0.092204	0.029024	-0.201654	-0.031049	0.033382	0.016458	0.325093	0.175283
650	-0.683882	-0.083747	0.039014	-0.221795	-0.055973	0.059398	0.009081	0.300366	0.173214
750	-0.699532	-0.049467	0.066042	-0.237147	-0.031839	0.035950	-0.020321	0.322232	0.167899
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.332934	-0.210182	-0.223219	-0.014492	-0.018860	-0.046950	0.027803	0.038687	0.038045
150	-0.104285	-0.027839	-0.067039	-0.004716	-0.002632	-0.000147	0.014571	0.011013	0.017579
250	-0.050399	0.004364	-0.039018	-0.003676	-0.002315	-0.006227	0.012071	0.005766	0.019321
350	-0.038685	0.024155	-0.026150	-0.003899	-0.000554	-0.006146	0.012713	0.006737	0.020357
450	-0.024770	0.030836	-0.018862	-0.003175	-0.000566	0.004119	0.012062	0.004538	0.020958
550	-0.018543	0.045704	-0.013415	-0.003161	-0.000945	0.004160	0.013147	0.003598	0.021392
650	-0.012275	0.047291	-0.010133	-0.002574	-0.000921	0.004339	0.013060	0.003299	0.021638
750	-0.008899	0.047146	-0.008516	-0.001723	-0.000799	0.003394	0.011159	0.002749	0.021715
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.298269	-0.134550	-0.227004	-0.013736	-0.053053	0.106922	0.069731	0.041179	0.040849
150	-0.092834	0.003013	-0.063940	-0.004222	-0.006221	-0.013396	0.012909	0.022191	0.016723
250	-0.055191	0.029356	-0.038036	-0.002323	-0.007987	-0.016554	0.014924	0.007958	0.018890
350	-0.031210	0.044964	-0.024452	-0.001811	-0.007258	-0.012448	0.014602	0.004280	0.020000
450	-0.024181	0.062836	-0.016313	-0.001518	-0.007952	-0.006682	0.014126	0.003975	0.020657
550	-0.011440	0.072551	-0.010962	-0.001491	-0.006364	-0.006231	0.013984	0.004238	0.021086
650	-0.004967	0.072511	-0.008052	-0.001503	-0.005792	-0.002295	0.014154	0.003852	0.021364
750	-0.002973	0.085185	-0.006110	-0.001287	-0.006178	0.000491	0.012504	0.004934	0.021547

Table 8: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
Experiment 1 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	1.784080	0.314256	0.226271	0.395060	0.262656	0.214268	0.342757	0.247709	0.233542	1563.1
150	1.801640	0.221993	0.167259	0.232390	0.178434	0.138025	0.222561	0.198050	0.190342	3.73E+40
250	1.921630	0.256821	0.177929	0.251787	0.203761	0.170792	0.249554	0.216414	0.186497	1.66E+78
350	1.979670	0.231476	0.177982	0.247388	0.195132	0.168977	0.204435	0.212569	0.185672	8.73E+115
450	1.797290	0.240577	0.177953	0.246300	0.166291	0.162179	0.217141	0.206922	0.155120	4.81E+153
550	1.763160	0.256588	0.165858	0.216826	0.180196	0.158660	0.233045	0.194648	0.165821	3.30E+191
650	1.917760	0.258593	0.164770	0.215954	0.190036	0.161094	0.232589	0.206317	0.182574	1.98E+229
750	1.760470	0.215243	0.168880	0.256738	0.195297	0.168007	0.245497	0.236121	0.188579	1.23E+267
Experiment 2 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	2.060190	0.260103	0.188610	0.342972	0.215741	0.185969	0.247222	0.244404	0.238971	1351.6
150	2.229940	0.333464	0.185489	0.241789	0.187628	0.183464	0.266003	0.230014	0.178057	3.34E+40
250	2.047340	0.276874	0.188172	0.234435	0.177164	0.160468	0.240272	0.216724	0.187430	1.42E+78
350	2.333000	0.257242	0.187210	0.256265	0.201816	0.170270	0.257236	0.245844	0.189163	8.06E+115
450	2.449120	0.311398	0.189733	0.233707	0.212801	0.167307	0.271557	0.231721	0.184160	4.69E+153
550	2.223170	0.288348	0.177577	0.235175	0.185694	0.160029	0.229321	0.239951	0.162904	2.82E+191
650	2.074080	0.253916	0.191666	0.260881	0.199513	0.157434	0.236659	0.218112	0.170691	1.92E+229
750	2.313200	0.253062	0.203446	0.253871	0.197592	0.180396	0.256337	0.235891	0.197359	1.23E+267
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.487350	0.509373	0.127937	0.011103	0.770408	2.077960	4.046830	0.366759	0.129155	0.009921
150	0.086940	0.091094	0.015924	0.002578	0.031488	0.179751	0.498726	0.023236	0.044953	0.013760
250	0.041659	0.044711	0.006429	0.001435	0.010799	0.052716	0.203904	0.007836	0.043367	0.014498
350	0.029621	0.033607	0.003995	0.000937	0.005415	0.061210	0.121219	0.003504	0.045357	0.014830
450	0.021292	0.027919	0.002729	0.000738	0.003162	0.015406	0.086194	0.001971	0.046745	0.015003
550	0.016762	0.025046	0.002198	0.000598	0.002186	0.012389	0.080681	0.001397	0.048044	0.015107
650	0.013687	0.024275	0.001781	0.000505	0.001604	0.010502	0.066370	0.000999	0.048626	0.015178
750	0.011976	0.021099	0.001556	0.000440	0.001261	0.007577	0.055808	0.000799	0.048803	0.015244
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.466390	0.493153	0.125493	0.009900	0.703763	3.563340	5.650820	0.356213	0.097226	0.004687
150	0.076306	0.095829	0.014659	0.002358	0.030725	0.194170	0.450985	0.021062	0.040098	0.006487
250	0.041382	0.055485	0.006637	0.001367	0.010460	0.063261	0.226598	0.007430	0.041107	0.006835
350	0.027891	0.041468	0.004253	0.000944	0.005526	0.030095	0.137587	0.003463	0.043630	0.006970
450	0.021156	0.035793	0.002829	0.000695	0.003293	0.019807	0.097577	0.001978	0.045199	0.007068
550	0.017492	0.032296	0.002248	0.000580	0.002296	0.014443	0.080739	0.001393	0.046578	0.007122
650	0.014244	0.028863	0.001800	0.000505	0.001757	0.022080	0.071459	0.000963	0.047348	0.007159
750	0.013783	0.030813	0.001533	0.000438	0.001446	0.012177	0.059306	0.000820	0.047936	0.007181

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 9: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
Experiment 1 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.750374	-0.059340	0.036238	-0.189004	-0.030481	0.049368	0.015927	0.296970	0.213682
150	-0.596223	-0.037154	0.045307	-0.205836	-0.033283	0.036377	-0.004280	0.314710	0.173781
250	-0.666517	-0.062015	0.025652	-0.214916	-0.041981	0.031282	0.013327	0.303673	0.181018
350	-0.739513	-0.085616	0.019156	-0.220549	-0.031562	0.067190	0.024533	0.295508	0.181225
450	-0.623897	-0.041777	0.007002	-0.187998	-0.021524	0.030250	0.021204	0.294637	0.178217
550	-0.550194	-0.045520	0.013775	-0.188376	-0.026025	0.038938	0.021005	0.286747	0.179588
650	-0.648418	-0.047359	0.023947	-0.214107	-0.031769	0.048818	-0.005876	0.288040	0.193557
750	-0.610700	-0.046447	0.030282	-0.216796	-0.018638	0.029702	0.000064	0.302119	0.177816
Experiment 2 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.767932	-0.102504	0.038142	-0.208172	-0.038660	0.025830	-0.001000	0.321977	0.188069
150	-0.658856	-0.064264	0.042636	-0.228541	-0.034421	0.038988	0.006874	0.297403	0.190425
250	-0.677868	-0.082119	0.052670	-0.228813	-0.043217	0.035842	-0.011712	0.323002	0.178721
350	-0.652315	-0.086769	0.058895	-0.231834	-0.049911	0.056498	0.015868	0.308270	0.169527
450	-0.618034	-0.073036	0.025667	-0.208497	-0.043298	0.063612	0.008473	0.304012	0.165474
550	-0.721412	-0.094638	0.032878	-0.216123	-0.036784	0.039188	0.007039	0.313777	0.187874
650	-0.676821	-0.085765	0.046754	-0.218858	-0.030699	0.070698	0.003455	0.315766	0.175367
750	-0.607330	-0.077979	0.028573	-0.248658	-0.013803	0.048530	-0.011977	0.295002	0.200063
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	-0.323332	-0.204390	-0.226941	-0.016450	-0.060166	-0.058117	0.044352	-0.028006	0.032992
150	-0.109502	-0.035118	-0.066658	-0.006866	-0.006436	0.005972	0.014114	0.000664	0.017515
250	-0.058141	0.002419	-0.039037	-0.003599	-0.007875	-0.005268	0.015267	0.002809	0.019248
350	-0.040712	0.025539	-0.024888	-0.002955	-0.007934	0.004258	0.013338	0.001258	0.020657
450	-0.035537	0.039735	-0.016089	-0.002164	-0.005481	0.000834	0.013660	-0.001043	0.021317
550	-0.022099	0.042947	-0.012499	-0.001763	-0.005539	-0.000117	0.013132	-0.001356	0.021581
650	-0.018917	0.052485	-0.009119	-0.001709	-0.002015	-0.002426	0.012180	0.000197	0.021749
750	-0.011633	0.057715	-0.006167	-0.001530	-0.002465	0.000526	0.011310	0.001421	0.022037
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	-0.319205	-0.188288	-0.207703	-0.013701	-0.093781	-0.016124	0.019829	0.043420	0.048372
150	-0.092007	-0.023151	-0.060868	-0.003592	-0.020264	-0.015444	0.021437	0.001367	0.016895
250	-0.059914	0.020700	-0.032654	-0.003162	-0.008259	-0.008314	0.017653	0.004542	0.019453
350	-0.041768	0.032592	-0.018330	-0.002506	-0.005846	-0.003991	0.016698	0.005212	0.020809
450	-0.027339	0.041132	-0.012858	-0.001658	-0.006438	-0.000128	0.015659	0.003810	0.021062
550	-0.021883	0.053394	-0.006908	-0.001972	-0.005404	-0.003963	0.015852	0.004001	0.021596
650	-0.011323	0.062252	-0.005209	-0.001362	-0.006420	-0.006080	0.014775	0.002116	0.021697
750	-0.010509	0.069957	-0.003996	-0.001468	-0.004755	0.000844	0.012591	0.002868	0.021810

Table 10: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
Experiment 1 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	1.718020	0.230223	0.157970	0.255591	0.161161	0.165465	0.221358	0.194128	0.219575	1554.32
150	1.690030	0.208471	0.179957	0.224608	0.180050	0.172149	0.214264	0.221890	0.172904	3.92E+40
250	1.969860	0.231707	0.179631	0.248554	0.174893	0.156223	0.222335	0.229651	0.189242	1.57E+78
350	2.181670	0.287893	0.180237	0.249106	0.180368	0.163994	0.237818	0.219436	0.194289	9.04E+115
450	1.797500	0.234302	0.178447	0.228286	0.202433	0.170167	0.211633	0.207135	0.193975	5.01E+153
550	1.631810	0.216623	0.164009	0.203291	0.180200	0.139823	0.223747	0.197282	0.198348	3.02E+191
650	1.827440	0.252081	0.189178	0.238634	0.185416	0.163750	0.199683	0.208002	0.194197	1.87E+229
750	1.829510	0.266355	0.152826	0.226327	0.180630	0.141238	0.243661	0.220539	0.175255	1.39E+267
Experiment 2 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	2.131960	0.297478	0.204748	0.390080	0.230395	0.192362	0.270955	0.254558	0.251135	1372.99
150	2.096920	0.230272	0.163723	0.238251	0.197437	0.150039	0.234048	0.195598	0.188017	3.31E+40
250	2.448480	0.254388	0.215078	0.267266	0.176388	0.167501	0.225804	0.221680	0.187835	1.31E+78
350	2.318060	0.303061	0.184964	0.258281	0.209987	0.192934	0.254420	0.231384	0.179608	7.71E+115
450	2.281800	0.240938	0.180969	0.241780	0.192493	0.164916	0.263411	0.211147	0.206284	4.88E+153
550	2.192640	0.268200	0.178558	0.234374	0.186121	0.174338	0.220539	0.232863	0.209650	2.87E+191
650	2.277900	0.290470	0.185857	0.238317	0.191953	0.206126	0.239182	0.236510	0.215025	1.87E+229
750	2.109920	0.252035	0.173455	0.270233	0.213992	0.153196	0.254738	0.211144	0.198459	1.24E+267
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.506027	0.534659	0.131719	0.009732	0.677963	4.108360	7.510570	0.348094	0.124931	0.009888
150	0.083221	0.093116	0.014714	0.002372	0.029124	0.151607	0.465938	0.021103	0.044092	0.013779
250	0.042952	0.047258	0.006762	0.001407	0.011193	0.059394	0.196051	0.007124	0.043693	0.014464
350	0.028293	0.036062	0.004128	0.000883	0.005820	0.026954	0.120033	0.003488	0.046846	0.014821
450	0.020880	0.028838	0.002777	0.000688	0.003125	0.022463	0.090966	0.002083	0.048189	0.014999
550	0.017342	0.026122	0.002104	0.000571	0.002267	0.012420	0.072812	0.001328	0.048817	0.015113
650	0.014288	0.025499	0.001816	0.000464	0.001538	0.010051	0.066007	0.001003	0.049295	0.015195
750	0.012921	0.024114	0.001527	0.000422	0.001229	0.008266	0.053374	0.000779	0.050256	0.015242
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.461307	0.600062	0.118638	0.010954	0.681219	2.745920	3.940910	0.315828	0.096580	0.004778
150	0.080414	0.093677	0.014393	0.002354	0.034382	0.163862	0.476128	0.021043	0.040775	0.006508
250	0.041367	0.052245	0.006153	0.001340	0.010378	0.056904	0.215778	0.006877	0.043299	0.006879
350	0.027448	0.038426	0.003902	0.000906	0.005080	0.029174	0.139556	0.003564	0.046774	0.007011
450	0.022047	0.033594	0.002730	0.000715	0.003139	0.019925	0.104372	0.002089	0.046995	0.007090
550	0.017144	0.031196	0.002123	0.000597	0.002233	0.016444	0.089748	0.001429	0.048887	0.007135
650	0.015520	0.028337	0.001844	0.000517	0.001610	0.012396	0.074753	0.001063	0.048936	0.007172
750	0.013649	0.026350	0.001485	0.000440	0.001458	0.009525	0.060045	0.000891	0.049124	0.007200

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 11: Simulated Biases of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Nonlinear SAR Errors

T	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\mu}_*)$
True Parameters defined in Experiment 1 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)									
50	-0.723526	-0.040998	0.046629	-0.104300	-0.062130	0.077459	0.043379	0.356527	0.132647
150	-0.567370	-0.046924	-0.006174	-0.191241	-0.051944	0.082715	0.001029	0.307892	0.139970
250	-0.441991	-0.040921	0.007777	-0.163293	-0.078092	0.080684	0.091359	0.254318	0.133108
350	-0.707159	-0.031938	0.041027	-0.212994	-0.040094	0.111859	0.042953	0.255755	0.188774
450	-0.648502	-0.053715	-0.015168	-0.242682	-0.069870	0.039701	0.030075	0.273338	0.134138
550	-0.609942	0.030887	0.009425	-0.175573	-0.039331	0.033564	-0.081273	0.250582	0.193532
650	-0.581432	-0.065585	0.019990	-0.218107	-0.039708	0.031023	0.036678	0.322678	0.147192
750	-0.619435	-0.022735	-0.002821	-0.192225	-0.000853	-0.000329	0.010283	0.329151	0.130911
True Parameters defined in Experiment 1 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)									
50	-0.646916	-0.038591	-0.001939	-0.157836	-0.059178	0.037025	0.040185	0.229984	0.231449
150	-0.812449	-0.154695	0.067059	-0.255797	-0.038682	0.030016	0.033121	0.340511	0.157837
250	-0.551709	-0.088185	0.062548	-0.161069	-0.008340	0.020018	-0.010888	0.366303	0.156433
350	-0.548448	-0.089430	0.030634	-0.246120	-0.019109	0.098373	0.042833	0.291950	0.173772
450	-0.604822	-0.035208	0.045075	-0.236033	-0.046778	0.030309	-0.038611	0.306416	0.160123
550	-0.708774	-0.117796	0.027318	-0.199916	-0.076301	-0.000482	-0.004988	0.322035	0.176691
650	-0.655900	-0.091911	0.073396	-0.232343	-0.054881	0.036292	0.027440	0.305545	0.155581
750	-1.004600	-0.127519	0.101725	-0.201340	-0.074953	0.079328	0.033038	0.370229	0.165211
True Parameters defined in Experiment 3 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)									
50	-0.226658	-0.243265	-0.133711	-0.009333	0.003264	0.064979	-0.153564	-0.026662	-0.163471
150	-0.019068	-0.036048	-0.038491	0.003547	-0.008931	0.018213	-0.005528	-0.003632	-0.049383
250	0.003083	-0.004560	-0.005731	0.004262	-0.001570	0.001037	-0.000855	-0.002406	-0.008321
350	0.002159	-0.002231	-0.002147	0.002171	-0.000544	-0.000048	0.002856	-0.002007	-0.002323
450	0.003258	-0.000438	-0.000672	0.002132	0.000877	-0.000191	0.001962	-0.002340	-0.001419
550	0.001636	-0.000787	0.001204	0.001808	-0.000415	0.001482	0.000774	-0.001102	0.000293
650	0.001966	-0.000620	-0.000689	0.000789	-0.001484	0.002071	-0.002052	0.000816	-0.001454
750	0.001076	-0.000369	-0.000588	0.001306	-0.002092	0.002496	-0.001924	0.001331	-0.001331

Table 12: Simulated MSE's of Estimates for the D.G.P. with Known Group Memberships: Nonstationary Covariate and Nonlinear SAR Errors

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\mu}_*)$	Temp. Ave. Error*
True Parameters defined in Experiment 1 ($m_i = n_i = 5$, $i = 1, \ldots, 4$)										
50	1.466860	0.295354	0.269690	0.291832	0.215242	0.975649	0.270796	0.524737	1.061990	2701.36
150	1.563000	0.167259	0.161158	0.201019	0.183243	0.154263	0.193984	0.210872	0.185222	5.61E+40
250	1.442620	0.315143	0.116822	0.246616	0.191720	0.190472	0.190428	0.161411	0.191952	2.87E+78
350	1.713420	0.201995	0.127012	0.210029	0.143973	0.141357	0.205595	0.144391	0.164993	1.27E+116
450	1.548100	0.245364	0.166674	0.243103	0.147247	0.123204	0.203177	0.214479	0.141302	7.39E+153
550	1.520010	0.162698	0.162997	0.207051	0.206016	0.187106	0.214274	0.168558	0.195972	6.10E+191
650	1.725780	0.234293	0.153574	0.228572	0.165952	0.139216	0.176861	0.222916	0.197904	2.98E+229
750	1.693850	0.227462	0.174560	0.241774	0.154371	0.165789	0.200635	0.232889	0.149284	2.16E+267
True Parameters defined in Experiment 1 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)										
50	1.511030	0.362425	0.210047	0.627462	0.214376	0.247412	0.351005	0.196661	0.401426	1052.45
150	2.957730	0.325895	0.191065	0.283897	0.286984	0.159210	0.266135	0.283765	0.183317	2.67E+40
250	1.994570	0.264033	0.188677	0.248677	0.153454	0.136016	0.210630	0.257604	0.190057	1.28E+78
350	2.368800	0.270339	0.163505	0.189996	0.173890	0.191413	0.171048	0.227594	0.155334	7.53E+115
450	2.118350	0.200034	0.164534	0.242241	0.167970	0.169146	0.183241	0.235458	0.200467	3.70E+153
550	2.182980	0.239595	0.189539	0.233886	0.193546	0.143679	0.194847	0.230771	0.168508	2.75E+191
650	2.935830	0.307808	0.195729	0.213703	0.201561	0.234406	0.299095	0.205404	0.187226	1.81E+229
750	2.777710	0.361537	0.230852	0.237976	0.233104	0.216640	0.211317	0.276054	0.196220	1.21E+267
True Parameters defined in Experiment 3 ($m_i = 10$ and $n_i = 20$, $i = 1, \ldots, 4$)										
50	0.349936	0.212711	0.094793	0.009885	0.142102	0.271360	0.745268	0.156800	0.179689	0.001583
150	0.041027	0.015733	0.012396	0.002354	0.007646	0.017754	0.047755	0.013066	0.022906	0.002115
250	0.007670	0.004515	0.002664	0.001338	0.001880	0.004710	0.009541	0.003138	0.005075	0.002226
350	0.003799	0.001789	0.001367	0.000715	0.000735	0.002058	0.003784	0.001082	0.002577	0.002274
450	0.000668	0.000459	0.000445	0.000384	0.000442	0.000901	0.000978	0.000587	0.000996	0.002285
550	0.000221	0.000148	0.000188	0.000205	0.000264	0.000451	0.000347	0.000226	0.000374	0.002304
650	0.000188	0.000083	0.000173	0.000094	0.000161	0.000259	0.000200	0.000132	0.000321	0.002314
750	0.000074	0.000045	0.000055	0.000064	0.000107	0.000162	0.000130	0.000082	0.000101	0.002322

* abbrev. for the temporal average error defined as $N\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\widehat{\boldsymbol{\psi}})$.

Table 13: Simulated Biases of Estimates for the D.G.P. with Unknown Group Memberships: Stationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{RandI}$	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\boldsymbol{U}})^*$
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.836711	-0.200000	0.095032	0.020655	0.148046	0.009603	0.011330	-0.009128	-0.014196	1.16E-18
150	0.836693	-0.199999	0.089935	0.019214	0.110930	0.009278	0.005221	-0.008646	-0.014399	1.10E-18
250	0.836689	-0.299999	0.083727	0.017611	0.104466	0.008208	0.009424	-0.007622	-0.013958	1.09E-18
350	0.836687	-0.199999	0.067225	0.016771	0.136913	0.007224	0.013576	-0.006649	-0.013525	1.18E-18
450	0.836686	-0.080000	0.011006	0.015864	0.089830	0.006397	0.007531	-0.005702	-0.013107	1.18E-18
550	0.836684	-0.060000	0.012274	0.015444	0.035880	0.005543	0.003146	-0.004791	-0.012700	1.19E-18
650	0.836684	-0.070000	0.013526	0.015076	0.041917	0.004674	0.002536	-0.003875	-0.012295	1.24E-18
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.925562	-0.099428	0.339384	-0.003816	0.008937	0.004996	0.004546	-0.004841	-0.086793	4.09E-19
150	0.924752	-0.199423	0.092875	-0.001761	0.005797	0.004227	0.004346	-0.005541	-0.037867	1.05E-19
250	0.924410	-0.099419	0.087759	-0.002013	0.004822	0.002850	0.001880	-0.002841	-0.021134	1.08E-19
350	0.929492	-0.002970	-0.003031	-0.005825	0.002198	0.025807	0.001741	-0.001121	-0.018386	4.77E-21

* abbrev. for the *optimal matching* biases of estimates of the group indicators $\boldsymbol{U}_0$, measured by $\min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \frac{1}{N} \sum_{c=1}^{G} \sum_{i=1}^{N} \{\widehat{u}_{i,\sigma^{(per)}(c)} - u_{0,i,c}\}$

Table 14: Simulated MSE of Estimates for the D.G.P. with Unknown Group Memberships: Stationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\boldsymbol{U}})^*$
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	0.150000	0.026653	0.011333	0.084379	0.005106	0.000850	0.002475	0.504371	0.000214
150	0.120000	0.027223	0.011497	0.013927	0.005343	0.000902	0.002644	0.427144	5.93E-06
250	0.051000	0.017473	0.006156	0.019005	0.005351	0.000924	0.002145	0.427167	1.20E-06
350	0.049000	0.017699	0.006138	0.023703	0.005357	0.000946	0.002047	0.527188	7.86E-07
450	0.037000	0.008900	0.003650	0.008162	0.004606	0.000964	0.001948	0.327207	5.57E-07
550	0.024000	0.002809	0.001677	0.003390	0.004364	0.000982	0.001649	0.127225	4.31E-07
650	0.020000	0.002627	0.001698	0.003149	0.003266	0.000999	0.001650	0.127243	3.55E-07
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	0.089949	0.040047	0.004028	0.008898	0.009925	0.002190	0.011651	0.009422	0.000143
150	0.074986	0.033616	0.003405	0.007984	0.008924	0.001507	0.008281	0.005682	1.19E-05
250	0.056772	0.016365	0.001779	0.004932	0.002923	0.000677	0.002310	0.001174	4.99E-06
350	0.001103	0.001148	0.004241	0.000604	0.001251	0.017232	0.005461	0.000723	5.01E-06

* abbrev. for the *optimal matching* MSE of estimates of the group indicators $\boldsymbol{U}_0$, measured by $\min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \frac{1}{N} \sum_{c=1}^{G} \sum_{i=1}^{N} \left(\widehat{u}_{i,\sigma^{(per)}(c)} - u_{0,i,c}\right)^2$

Table 15: Simulated Biases of Estimates for the D.G.P. with Unknown Group Memberships: Stationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{RandI}$	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\boldsymbol{U}})$
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.836716	-0.099100	0.054333	0.014412	0.510008	0.011888	0.015607	-0.010132	-0.014834	1.07E-18
150	0.836694	-0.090900	0.069532	0.013345	0.114049	0.011624	0.014626	-0.009505	-0.015013	1.11E-18
250	0.836689	-0.071100	0.053505	0.011841	0.106889	0.010447	0.008930	-0.008493	-0.014565	1.17E-18
350	0.837688	-0.070900	0.047049	0.011199	0.129387	0.009519	0.007094	-0.007537	-0.014129	1.17E-18
450	0.837696	-0.069100	0.040171	0.010565	0.090897	0.008602	0.007043	-0.006593	-0.013708	1.21E-18
550	0.839185	-0.058700	0.032759	0.009754	0.061515	0.007703	0.005036	-0.005651	-0.010292	1.12E-18
650	0.839885	-0.053000	0.023551	0.009078	0.042456	0.006716	0.004967	-0.004720	-0.009883	1.18E-18
750	0.840684	-0.049200	0.014789	0.008342	0.041850	0.005821	0.003774	-0.003812	-0.004781	1.18E-18
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.838567	-0.094636	0.029774	0.008948	0.087649	0.029731	-0.003346	-0.005173	-0.021587	8.27E-19
150	0.834924	-0.094514	0.039710	0.009719	0.075975	0.023436	-0.001940	-0.004988	-0.020252	8.37E-19
250	0.839924	-0.059450	0.048313	0.010082	0.064472	0.024444	-0.000148	-0.004727	-0.019214	8.41E-19
350	0.843923	-0.039450	0.045543	0.009858	0.054150	0.024203	0.002094	-0.004170	-0.009798	8.67E-19
450	0.893273	-0.039449	0.012743	0.009267	0.044112	0.013975	0.001310	-0.003606	-0.003746	8.82E-19
550	0.910910	-0.018448	0.007001	0.008746	0.023409	0.013712	0.001569	-0.003054	-0.002511	8.89E-19
650	0.910780	-0.009447	0.007723	0.008356	0.021432	0.013477	0.001806	-0.002501	-0.002792	8.53E-19

Table 16: Simulated MSE of Estimates for the D.G.P. with Unknown Group Memberships: Stationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\boldsymbol{U}})$
Experiment 3 ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	0.099010	0.025183	0.008270	0.007523	0.006102	0.000871	0.002963	0.513780	0.000197647
150	0.048120	0.025772	0.008439	0.020195	0.006379	0.000830	0.003854	0.535010	5.99E-06
250	0.028970	0.016027	0.008515	0.018185	0.006388	0.000854	0.003955	0.235030	1.21E-06
350	0.030110	0.015255	0.008564	0.012916	0.006394	0.000875	0.004057	0.215060	7.84E-07
450	0.029790	0.006465	0.008597	0.017394	0.006398	0.000895	0.004058	0.135080	5.65E-07
550	0.025080	0.005650	0.008624	0.009631	0.006401	0.000813	0.003059	0.105090	4.32E-07
650	0.020310	0.004836	0.008644	0.005818	0.006404	0.000730	0.002060	0.085110	3.61E-07
750	0.018930	0.001006	0.008662	0.002746	0.006406	0.000747	0.001061	0.055130	2.94E-07
Experiment 4 ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	0.019387	0.008612	0.001449	0.004357	0.007006	0.000478	0.001206	0.700020	3.32E-05
150	0.019366	0.009711	0.001643	0.006349	0.008043	0.000559	0.001390	0.685560	4.27E-06
250	0.009364	0.010319	0.001762	0.008105	0.008393	0.000603	0.001416	0.405013	2.16E-06
350	0.003936	0.009037	0.001775	0.009838	0.008394	0.000608	0.001416	0.205020	1.23E-07
450	0.004936	0.009043	0.001785	0.011560	0.008395	0.000613	0.001416	0.201026	9.53E-08
550	0.002936	0.010484	0.001792	0.003265	0.008395	0.000619	0.001417	0.055032	7.66E-08
650	0.000269	0.009054	0.001799	0.004965	0.008396	0.000624	0.001417	0.009038	6.45E-08

Table 17: Simulated Biases of Estimates for the D.G.P. with Unknown Group Memberships: Nonstationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{RandI}$	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\boldsymbol{U}})$
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)										
50	0.895825	-0.399503	0.129798	-0.040096	-0.163200	0.091291	0.243564	-0.152902	0.150471	6.60E-19
150	0.931093	-0.299501	0.097043	-0.052502	-0.111101	0.070058	0.143022	-0.136157	0.085895	5.07E-19
250	0.940236	-0.199501	0.066033	-0.040046	-0.050991	0.077410	0.123526	-0.128520	0.038520	3.76E-19
350	0.951340	-0.187501	0.052012	-0.038449	-0.043039	0.069474	0.120338	-0.107952	0.030947	3.63E-19
450	0.946767	-0.099500	0.020433	-0.030807	-0.021219	0.031004	0.112045	-0.095472	0.020290	3.37E-19
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)										
50	0.864367	-0.117727	0.007791	-0.536374	-0.167275	1.007590	0.558425	-0.530089	-3.907340	2.73E-18
150	0.918841	0.005550	0.000560	-0.001022	0.005762	0.000117	1.15E-08	2.12E-09	-0.000561	6.69E-21
250	0.919706	-2.30E-07	5.52E-09	2.37E-08	0.000647	1.60E-08	-5.07E-09	-5.93E-06	-4.11E-08	1.64E-21
350	0.917975	-3.82E-07	3.21E-09	1.12E-08	-1.35E-07	3.63E-08	1.82E-08	-7.53E-09	4.16E-08	6.80E-21
450	0.938044	-8.16E-07	-6.26E-09	5.39E-08	-4.12E-08	-1.27E-07	-3.33E-08	-7.72E-09	5.64E-08	4.77E-20

Table 18: Simulated MSE of Estimates for the D.G.P. with Unknown Group Memberships: Nonstationary Covariate and Linear SAR Errors with Queen-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\boldsymbol{U}})$
Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)									
50	0.037714	0.132322	0.021985	0.085987	0.099318	0.079034	0.009677	0.069538	9.29E-05
150	0.048762	0.095269	0.074415	0.084054	0.019932	0.048921	0.007974	0.010090	7.96E-05
250	0.046392	0.072235	0.018656	0.080438	0.029932	0.039696	0.009325	0.012185	6.50E-05
350	0.036971	0.062368	0.022342	0.077061	0.019731	0.032408	0.009740	0.010455	1.45E-05
450	0.027732	0.055562	0.027623	0.047778	0.011731	0.026706	0.001011	0.008597	2.53E-05
Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)									
50	0.226348	0.043843	0.662631	0.221120	0.207491	2.686050	0.854394	0.822771	0.369873
150	0.002578	2.64E-05	3.32E-13	0.002771	0.000405	3.50E-05	1.14E-06	4.34E-09	2.63E-05
250	1.55E-11	5.07E-15	8.73E-05	2.83E-13	7.88E-10	2.49E-09	1.37E-13	9.92E-15	2.81E-13
350	3.36E-11	7.78E-15	1.02E-13	1.03E-12	4.86E-09	3.01E-13	1.31E-13	4.48E-14	4.79E-13
450	8.82E-12	9.83E-15	9.34E-14	1.99E-12	4.27E-14	3.47E-13	4.29E-15	8.27E-15	9.58E-13

Table 19: Simulated Biases of Estimates for the D.G.P. with Unknown Group Memberships: Nonstationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{RandI}$	$\widetilde{Bias}(\widehat{\phi}_1)$	$\widetilde{Bias}(\widehat{\phi}_2)$	$\widetilde{Bias}(\widehat{\phi}_3)$	$\widetilde{Bias}(\widehat{\phi}_4)$	$\widetilde{Bias}(\widehat{\theta}_1)$	$\widetilde{Bias}(\widehat{\theta}_2)$	$\widetilde{Bias}(\widehat{\theta}_3)$	$\widetilde{Bias}(\widehat{\theta}_4)$	$\widetilde{Bias}(\widehat{\boldsymbol{U}})$
			Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)							
50	0.803562	-0.092262	-0.001256	-0.078915	-0.036258	-0.000337	-0.000559	0.000022	0.028714	1.06E-18
150	0.836695	-0.091908	-0.002776	-0.011379	-0.062262	-0.000204	-0.000654	-0.000048	0.029758	1.07E-18
250	0.836792	-0.021837	-0.004199	-0.013775	-0.053089	0.000071	-0.000548	-0.000013	0.019577	1.14E-18
350	0.846890	-0.030215	-0.001547	-0.011620	-0.010785	0.000224	-0.000101	-0.000019	0.014932	1.14E-18
450	0.856910	-0.011811	-0.000888	-0.010174	-0.011118	0.000394	-0.000129	-0.000027	0.009076	1.11E-18
			Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)							
50	0.765065	-0.183985	0.027641	-0.920037	-0.325931	2.055080	0.965660	-0.868013	-6.436870	3.32E-18
150	0.823150	0.010192	0.002008	-0.022168	0.011088	0.046714	0.029226	-0.023231	-0.142422	7.51E-20
250	0.909706	-0.000787	0.000979	-0.001818	-0.001416	0.001221	0.000216	-7.71E-09	-0.001042	6.12E-20
350	0.957973	-8.14E-07	5.52E-09	2.35E-08	-4.07E-08	1.49E-08	-5.02E-09	-6.94E-06	-4.10E-08	7.24E-21
450	0.957678	-2.30E-07	3.16E-09	1.07E-08	-1.35E-07	3.40E-08	1.81E-08	-7.48E-09	4.16E-08	8.02E-22

Table 20: Simulated MSE of Estimates for the D.G.P. with Unknown Group Memberships: Nonstationary Covariate and Linear SAR Errors with Rook-Contiguity Weights

T	$\widetilde{MSE}(\widehat{\phi}_1)$	$\widetilde{MSE}(\widehat{\phi}_2)$	$\widetilde{MSE}(\widehat{\phi}_3)$	$\widetilde{MSE}(\widehat{\phi}_4)$	$\widetilde{MSE}(\widehat{\theta}_1)$	$\widetilde{MSE}(\widehat{\theta}_2)$	$\widetilde{MSE}(\widehat{\theta}_3)$	$\widetilde{MSE}(\widehat{\theta}_4)$	$\widetilde{MSE}(\widehat{\boldsymbol{U}})$
			Experiment 3 using (7.3) instead of (7.2) ($N_1 = 45$, $N_2 = 30$, $N_3 = 30$, and $N_4 = 70$)						
50	0.009112	0.000547	0.001821	0.003870	0.000386	0.000066	0.000008	0.001562	2.43E-05
150	0.009065	0.000570	0.001611	0.004957	0.000212	0.000075	0.000004	0.001420	2.99E-06
250	0.011053	0.000581	0.001671	0.003652	0.000228	0.000052	0.000005	0.001129	1.75E-06
350	0.003905	0.000589	0.001704	0.005266	0.000242	0.000046	0.000005	0.000933	1.39E-06
450	0.004905	0.000595	0.001530	0.004886	0.000255	0.000030	0.000005	0.001035	1.15E-06
			Experiment 4 using (7.3) instead of (7.2) ($N_1 = 100$, $N_2 = 60$, $N_3 = 65$, and $N_4 = 150$)						
50	0.301930	0.058145	1.130780	0.378353	5.786110	1.511700	1.304870	58.980300	0.343519
150	0.005589	0.001878	0.027947	0.005613	0.123075	0.044864	0.032433	1.039320	0.008055
250	0.004744	4.41E-05	0.000151	0.007093	6.80E-05	2.13E-06	4.47E-09	4.95E-05	0.000582
350	8.78E-12	5.09E-15	9.21E-14	2.76E-13	2.73E-13	1.30E-13	9.90E-15	9.64E-13	7.84E-10
450	3.35E-11	7.75E-15	1.00E-13	1.03E-12	3.72E-14	4.19E-15	4.47E-14	2.81E-13	4.85E-09

Table 21: Unit Root Tests

	Unit root ADF test*	
Country	$I_{i,t}$	$S_{i,t}$
Australia	0.1529	0.3528
Belgium	0.0482	0.8366
Canada	0.4372	0.2867
Costa Rica	0.4589	0.0477
Czech Republic	0.3651	0.2809
Denmark	0.2294	0.4473
Estonia	0.2478	0.5324
European Union (EU-28)	0.5762	0.1357
Finland	0.1356	0.9741
France	0.2363	0.3993
Germany	0.3279	0.0857
Greece	0.9345	0.6244
Hungary	0.4449	0.0895
Israel	0.0810	0.0455
Italy	0.6776	0.5049
Japan	0.3629	0.5700
Korea	0.0197	0.1168
Latvia	0.1844	0.5216
Mexico	0.0527	0.2478
Netherlands	0.4776	0.2230
New Zealand	0.0128	0.0845
Norway	0.3402	0.1501
Portugal	0.9577	0.5390
Slovak Republic	0.1069	0.0060
Slovenia	0.7136	0.0508
South Africa	0.4728	0.3390
Spain	0.8018	0.4308
Sweden	0.0452	0.1721
Switzerland	0.0000	0.7048
UK	0.2104	0.3151
USA	0.5978	0.6530

* ADF is the Augmented Dickey-Fuller test with optimal lag orders ($\leq$ 11) selected by the Schwartz information criterion. Numbers reported are MacKinnon (1996) one-sided p-values.

Table 22: Investment-Saving Error-Correction Model (ECM) Estimates

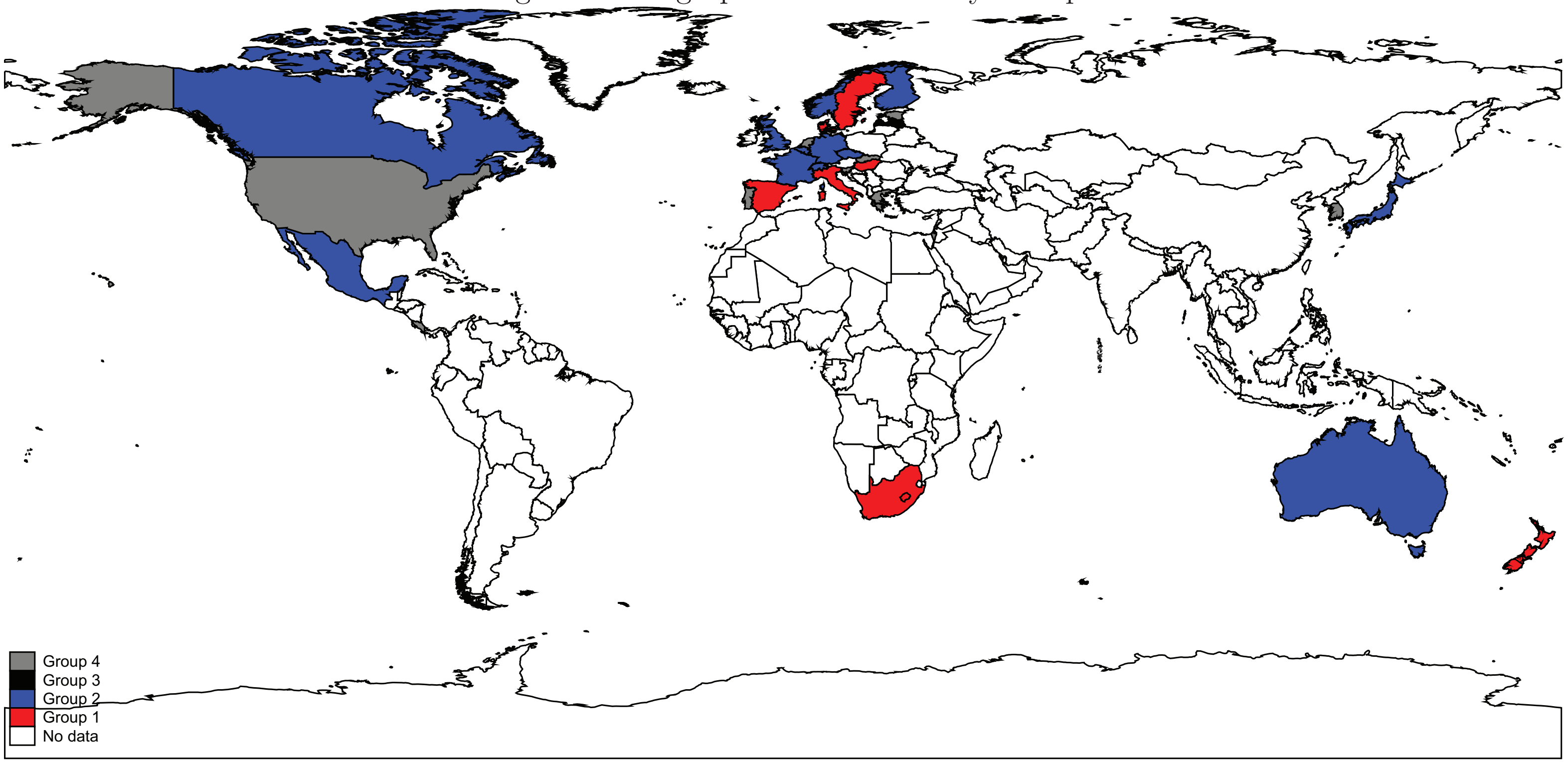
G	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	$\widehat{\phi}_4$	$\widehat{\phi}_5$	$\widehat{\phi}_6$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\gamma}_5$	$\widehat{\gamma}_6$	Composite error
2	-0.001 [-0.4148,0.4168]*	-1.0089 [-1.6152,-0.4025]	–	–	–	–	-0.7310 [-551.74,550.28]	0.4660 [0.0735,0.8586]	–	–	–	–	0.1812	-0.0112	–	–	–	–	0.0026
3	-0.3898 [-1.7557,0.9760]	-0.001 [-0.4039,0.4059]	-0.4608 [-0.8393,-0.0824]	–	–	–	-0.4127 [-5.6886, 4.8631]	-0.4679 [-565.57,564.63]	-0.0444 [-1.1959,1.1070]	–	–	–	-0.1833	0.3406	0.3617	–	–	–	0.0019
4	-0.8441 [-1.4443,-0.2439]	-0.9785 [-1.3922,-0.5647]	-0.1365 [-0.8329,0.5597]	-0.001 [-0.2497, 0.2517]	–	–	0.1701 [-0.4510,0.7913]	1.0573 [0.6879,1.4267]	0.4336 [-8.8299,9.6971]	-0.8177 [-435.43,444.79]	–	–	0.0784	0.4491	-0.0750	-0.0931	–	–	0.0005
5	-0.4407 [-1.0298,0.1484]	-0.0465 [-1.3475,1.2544]	-0.1560 [-0.8147,0.5025]	-0.001 [-0.6292,0.6312]	-1.2015 [-1.6539,-0.7490]	–	-0.2933 [-1.7819,1.1951]	-0.0851 [-32.36,32.19]	-0.3379 [-6.2712,5.5953]	-0.0647 [-697.95,697.83]	0.8283 [0.4359,1.2207]	–	0.2212	0.0855	0.2959	-0.2268	0.3686	–	0.0008
6	0.5167	-0.3255	0.0277	-0.3094	-1.1471	-0.0873	-0.7834	-0.3776	-0.4112	0.0010	0.8522	0.0665	0.1107	-0.0149	0.5724	-0.0758	0.1516	-0.4504	0.0006

Note: (*) Script-size numbers in square brackets are the bounds of 95% confidence intervals.

Table 23: Investment-Saving Error-Correction Model (ECM) Group Classifications

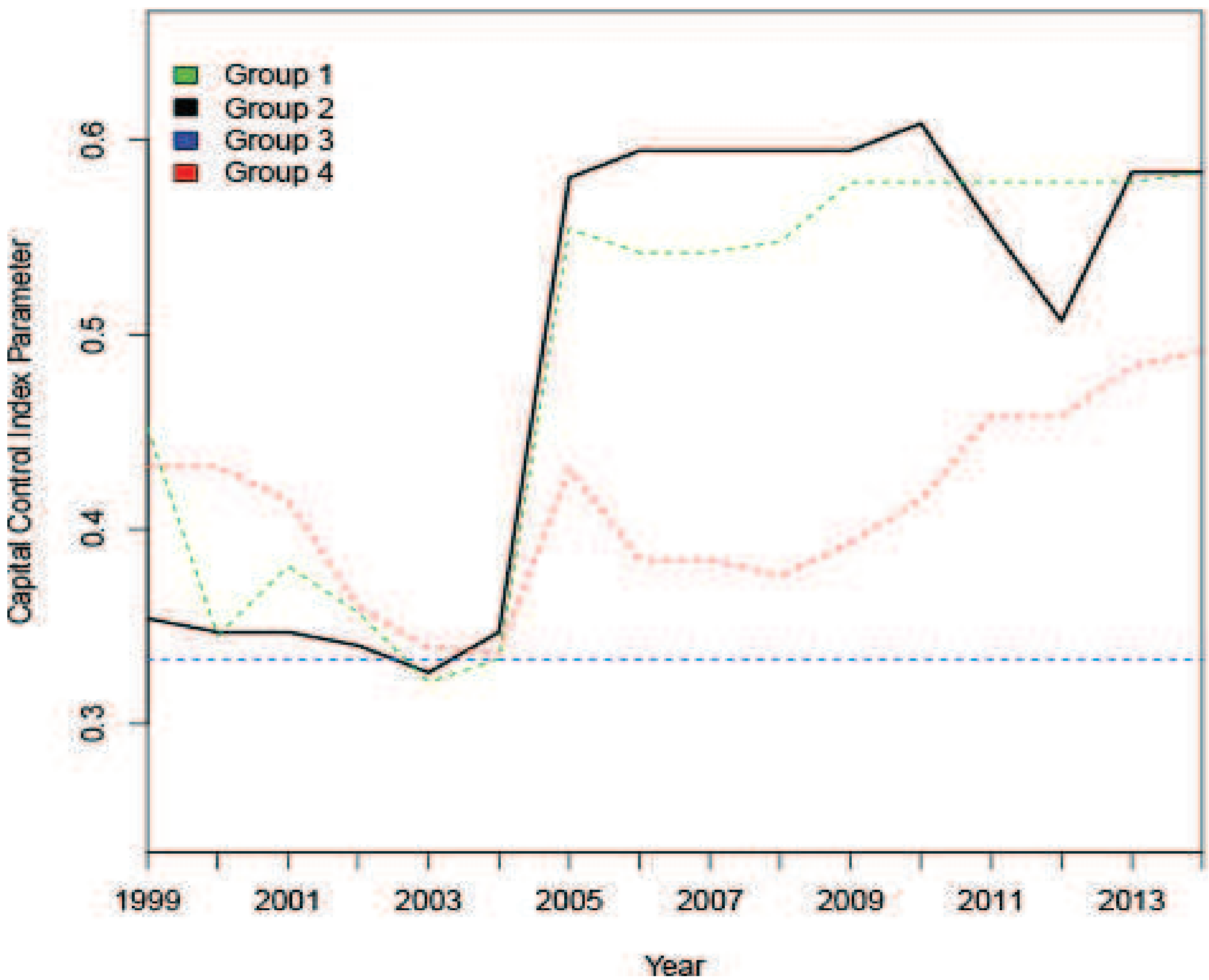
G	Group estimates
2	*Group 1:* Australia, France, Germany, Italy, Korea, Mexico, Netherlands, New Zealand, Portugal, Slovak Republic, Spain, UK, USA, Estonia, Israel, South Africa, Euro28, Costa Rica *Group 2:* Belgium, Canada, Czech Republic, Denmark, Finland, Greece, Hungary, Japan, Norway, Sweden, Switzerland, Slovenia, Latvia
3	*Group 1:* Spain, Israel,South Africa *Group 2:* Australia, Belgium, Canada, Czech Republic, Greece, Italy, Korea, Netherlands, New Zealand, Portugal, Sweden, UK, USA, Estonia, Costa Rica *Group 3:* Denmark, Finland, France, Germany, Hungary, Japan, Mexico, Norway, Slovak Republic, Switzerland, Slovenia, Euro28, Latvia
4	*Group 1:* Denmark, Hungary, Italy, New Zealand, Spain, Sweden, South Africa *Group 2:* Australia, Belgium, Canada, Czech Republic, Finland, France, Germany, Japan, Mexico, Norway, Switzerland, UK, Euro28 *Group 3:* Latvia *Group 4:* Greece, Korea, Netherlands, Portugal, Slovak Republic, USA, Estonia, Israel, Slovenia, Costa Rica
5	*Group 1:* Denmark, Hungary, Korea, New Zealand, Portugal, South Africa, Latvia *Group 2:* Netherlands, Slovenia *Group 3:* Greece, Italy, Slovak Republic *Group 4:* Germany, Spain, Estonia, Israel, Costa Rica *Group 5:* Australia, Belgium, Canada, Czech Republic, Finland, France, Japan, Mexico, Norway, Switzerland, UK, USA, Euro28

Figure 1: Geographical Locations by Group



Note: This group map is sketched using Bonhomme and Manresa's (2015) Stata code.

Figure 2: Average Capital Control Indices by Group



Note: The indices reported here are constructed based on Miniane's (2004) methodology.

Appendix A. Some Important Definitions

Definition 1. *A random field,* $\{X_{\boldsymbol{i}},\ \boldsymbol{i} \in V_n\}$*, on a sublattice indexed by* n*, say* V_n*, in the standard integer lattice* $\mathbb{Z}^{d_v}$ *is mixing with the mixing coefficient* $\alpha_s(\cdot)$ *if there exists a function* $\alpha(\tau) \downarrow 0$ *as* $\tau \uparrow \infty$ *such that, for any pair of subsets,* $S, S' \subset V_n$*,*

$$\alpha_s(\mathcal{B}(S), \mathcal{B}(S')) = \sup\left\{\left|P(A\bigcap B) - P(A)P(B)\right|,\ A \subset \mathcal{B}(S) \text{ and } B \subset \mathcal{B}(S')\right\} \leq M_\alpha(|S|, |S'|)\alpha(d(S, S')),$$

where $\mathcal{B}(S)$ *is the Borel* σ*-field generated by the random elements* $\{X_{\boldsymbol{i}},\ \boldsymbol{i} \in S\}$ *and* $M_\alpha(\cdot, \cdot)$ *is a symmetric positive function non-decreasing in its arguments. Throughout this paper, we assume that* $M_\alpha(\cdot, \cdot)$ *satisfies one of the following conditions:*

$$M_\alpha(n, m) \leq C_0 \min(n, m) \tag{A-1}$$

$$M_\alpha(n, m) \leq C_0(n + m_1)^{\gamma_M} \text{ for some } \gamma_M \geq 1. \tag{A-2}$$

Conditions (A-1) and (A-2) correspond to the ones used by Neaderhouser (1980) and Takahata (1983) respectively. They are satisfied by many spatial models (see, e.g., Rosenblatt (1985) or Guyon (1995)). It is important to note that, if $M_\alpha(n, m) = 1$ *for every* $n, m \geq 1$*, then we call* $\{X_{\boldsymbol{i}},\ \boldsymbol{i} \in V_n\}$ *a strongly mixing random field. There are many random fields which do not satisfy the strong-mixing condition, but they do satisfy the mixing condition (see, e.g., Neaderhouser (1980)).*

Appendix B. Auxiliary Results

Lemma 1. *Let* $\{\eta_{\boldsymbol{s}},\ \boldsymbol{s} \in V_n\}$ *represent a mixing centered random field. Suppose that* $\eta_{\boldsymbol{s}},\ \boldsymbol{s} \in V_n$ *are identically distributed across locations such that* $E[|\eta_{\boldsymbol{s}}|^{\gamma_\eta}] < \infty$ *for some* $\gamma_\eta > 2$*; and* $\sum_{r=1}^{diam(V_n)} r^{d_v - 1}\alpha(r)^{1 - \frac{2}{\gamma_\eta}} < \infty$*. Then,*

$$E\left|\sum_{\boldsymbol{s} \in V_n} \eta_{\boldsymbol{s}}\right|^2 \leq C_0|V_n|.$$

Proof. For brevity, define $S(V_n) = \sum_{\boldsymbol{s} \in V_n} \eta_{\boldsymbol{s}}$. One has

$$\begin{aligned}
\frac{E|S(V_n)|^2}{|V_n|} &= E[\eta_{\boldsymbol{s}}^2] + \frac{1}{|V_n|}\sum_{\boldsymbol{s}, \boldsymbol{w} \in V_n, \boldsymbol{s} \neq \boldsymbol{w}} E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}] \\
&= E[\eta_{\boldsymbol{s}}^2] + \frac{1}{|V_n|}\sum_{\boldsymbol{s} \in V_n}\sum_{r=1}^{\mathrm{diam}(V_n)}\sum_{\boldsymbol{w} \in V_n, \|\boldsymbol{w} - \boldsymbol{s}\| = r} E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}] = E[\eta_{\boldsymbol{s}}^2] + \mathcal{A}_n.
\end{aligned}$$

By Lemma 12, one gets $\mathcal{A}_n \le C_0 \frac{1}{|V_n|} \sum_{\boldsymbol{s}\in V_n} \sum_{r=1}^{\mathrm{diam}(V_n)} |\{\boldsymbol{w} \in V_n : \ \|\boldsymbol{w}-\boldsymbol{s}\| = r\}| \, \alpha(r)^{1-2/\gamma_\eta}$. Invoking Lemma 11, it then follows that $\mathcal{A}_n \le C_0 \sum_{r=1}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta} < \infty$. The lemma is proved. ☐

Lemma 2. *Let $\{\eta_{\boldsymbol{s}}, \ \boldsymbol{s} \in V_n\}$ be defined as in Lemma 1 and $S(V_n) = \sum_{\boldsymbol{s}\in V_n} \eta_{\boldsymbol{s}}$. Suppose that $E[|\eta_{\boldsymbol{s}}|^{2\gamma_\eta}] < \infty$ for some $\gamma_\eta > 2$. If $\sum_{r=1}^{diam(V_n)} r^{d_v-1}\alpha(r)^{1-\frac{2}{\gamma_\eta}} < \infty$ and $|V_n|^{1/2} \sum_{r=|V_n|^{\frac{1}{2d_v}}}^{diam(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta} < \infty$, then*

$$E[S(V_n)^3] \le C_0 |V_n|^{3/2}. \tag{B-1}$$

Proof. One can immediately obtain

$$E[S(V)^2] = |V_n| E[\eta_{\boldsymbol{s}}^3] + \sum_{\boldsymbol{s},\boldsymbol{w}\in V_n, \boldsymbol{s}\neq \boldsymbol{w}} E[\eta_{\boldsymbol{s}}^2 \eta_{\boldsymbol{w}}] + \sum_{\substack{\boldsymbol{s},\boldsymbol{w},\boldsymbol{z}\in V_n \\ \boldsymbol{s}\neq\boldsymbol{w} \\ \boldsymbol{s}\neq\boldsymbol{z} \\ \boldsymbol{w}\neq\boldsymbol{z}}} E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}] = |V_n| E[\eta_{\boldsymbol{s}}^3] + \mathcal{A}_n + \mathcal{B}_n. \tag{B-2}$$

(Note that the symbols $\mathcal{A}_n$ and $\mathcal{B}_n$ are meant specifically in this proof and different from those defined elsewhere.) By Lemma 12, one has $|E[\eta_{\boldsymbol{s}}^2\eta_{\boldsymbol{w}}]| \le C_0 \|\eta_{\boldsymbol{s}}^2\|_{\gamma_\eta} \|\eta_{\boldsymbol{w}}\|_{\gamma_\eta} M_\alpha(1,1)^{1-2/\gamma_\eta} \alpha(\|\boldsymbol{s} - \boldsymbol{w}\|)^{1-2/\gamma_\eta}$. Thus, in view of Lemmas 11 and 12,

$$\mathcal{A}_n \le C_0 \sum_{\boldsymbol{s}\in V_n} \sum_{r=1}^{\mathrm{diam}(V_n)} \sum_{\|\boldsymbol{s}-\boldsymbol{w}\|=r, \boldsymbol{w}\in V_n} \alpha(r)^{1-2/\gamma_\eta} \le 2d_v C_0 |V_n| \sum_{r=1}^{\mathrm{diam}(V_n)} (2r+1)^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}$$
$$\le C_0|V_n|. \tag{B-3}$$

Next, to bound $\mathcal{B}_n$, a decomposition of the summation indices yields

$$\begin{aligned}
\mathcal{B}_n &= \sum_{\boldsymbol{s}\in V_n} \left(\sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|\le c_n}} + \sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|> c_n}} \right) \left(\sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|\le c_n}} + \sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|> c_n}} \right) E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}] \\
&= \sum_{\boldsymbol{s}\in V_n} \left(\sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|\le c_n}} \sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|\le c_n}} + \sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|\le c_n}} \sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|> c_n}} + \sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|> c_n}} \sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|\le c_n}} \right. \\
&\quad \left. + \sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|> c_n}} \sum_{\substack{\boldsymbol{z}\in V_n \\ \|\boldsymbol{z}-\boldsymbol{s}\|> c_n}} \right) E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}] = \mathcal{B}_{n,1} + \mathcal{B}_{n,2} + \mathcal{B}_{n,3} + \mathcal{B}_{n,4}.
\end{aligned} \tag{B-4}$$

Notice that, by Lemma 11, for a given $\boldsymbol{s} \in V_n$, $\sum_{\substack{\boldsymbol{s}\in V_n \\ 1\le\|\boldsymbol{w}-\boldsymbol{s}\|\le c_n}} = \sum_{r=1}^{c_n} \sum_{\substack{\boldsymbol{w}\in V_n \\ \|\boldsymbol{w}-\boldsymbol{s}\|=r}} \le 2d_v \sum_{r=1}^{c_n} (2r +$

$1)^{d_v-1} \leq 2d_v 2^{d_v-2}(2^{d_v-1}+1)\sum_{r=1}^{c_n} r^{d_v-1} < C_0 c_n^{d_v}$, where the last inequality holds by the formula: $\sum_{k=1}^{n} k^p \approx \frac{n^{p+1}}{p+1}$; and, by Lemma 12, $|E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}]| \leq C_0\alpha\left(\min(\|\boldsymbol{w}-\boldsymbol{s}\|, \|\boldsymbol{z}-\boldsymbol{s}\|)\right)^{1-2/\gamma_\eta}$. It immediately follows that

$$\mathcal{B}_{n,1} \leq C_0|V_n|c_n^{d_v}\sum_{r=1}^{c_n} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}. \tag{B-5}$$

Since $\mathcal{B}_{n,2} = \sum_{\boldsymbol{s}\in V_n}\sum_{\substack{\boldsymbol{w}\in V_n\\ \|\boldsymbol{w}-\boldsymbol{s}\|\leq c_n}}\left(\sum_{\substack{\boldsymbol{z}\in V_n\\ \|\boldsymbol{z}-\boldsymbol{s}\|>c_n\\ \|\boldsymbol{z}-\boldsymbol{w}\|\leq c_n}} + \sum_{\substack{\boldsymbol{z}\in V_n\\ \|\boldsymbol{z}-\boldsymbol{s}\|>c_n\\ \|\boldsymbol{z}-\boldsymbol{w}\|>c_n}}\right) E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}] = \mathcal{B}_{n,2,a}+\mathcal{B}_{n,2,b}$, where $|E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}\eta_{\boldsymbol{z}}]| \leq C_0\alpha\left(\min(\|\boldsymbol{z}-\boldsymbol{s}\|, \|\boldsymbol{z}-\boldsymbol{w}\|)\right)^{1-2/\gamma_\eta}$, one has

$$\begin{aligned}
\mathcal{B}_{n,2,a} &\leq C_0\sum_{\boldsymbol{s}\in V_n}\sum_{\substack{\boldsymbol{w}\in V_n\\ \|\boldsymbol{w}-\boldsymbol{s}\|\leq c_n}}\sum_{\substack{\boldsymbol{z}\in V_n\\ \|\boldsymbol{z}-\boldsymbol{s}\|\leq c_n}} \alpha\left(\|\boldsymbol{z}-\boldsymbol{w}\|\right)^{1-2/\gamma_\eta} \leq C_0|V_n|c_n^{d_v}\sum_{r=1}^{c_n} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}\\
\mathcal{B}_{n,2,b} &\leq C_0\sum_{\boldsymbol{s}\in V_n}\sum_{\substack{\boldsymbol{w}\in V_n\\ \|\boldsymbol{w}-\boldsymbol{s}\|\leq c_n}}\sum_{r=c_n}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta} \leq C_0|V_n|c_n^{d_v}\sum_{r=c_n}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}.
\end{aligned}$$

Thus,

$$\mathcal{B}_{n,2} \leq C_0|V_n|c_n^{d_v}\sum_{r=1}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}; \tag{B-6}$$

and similarly, one also has

$$\mathcal{B}_{n,3} \leq C_0|V_n|c_n^{d_v}\sum_{r=1}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}. \tag{B-7}$$

By the same argument, we can also show that

$$\begin{aligned}
\mathcal{B}_{n,4} &\leq C_0\sum_{\boldsymbol{s}\in V_n}\sum_{\substack{\boldsymbol{w}\in V_n\\ \|\boldsymbol{w}-\boldsymbol{s}\|>c_n}}\sum_{r=c_n}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}\\
&\leq C_0|V_n|^2\sum_{r=c_n}^{\mathrm{diam}(V_n)} r^{d_v-1}\alpha(r)^{1-2/\gamma_\eta}.
\end{aligned}\tag{B-8}$$

The lemma readily follows from (B-2)-(B-8) by choosing $c_n = |V_n|^{\frac{1}{2d_v}}$. ☐

Lemma 3. *Let $S(V_n) = \sum_{\boldsymbol{s}\in V_n}\eta_{\boldsymbol{s}}$ be defined as in Lemma 1 above. Suppose that $\alpha(\tau) \leq C_\theta\tau^{-\theta_\alpha}$ for some $\theta_\alpha \geq \max\left(\frac{pd_v\gamma_\eta}{(p-q)(\gamma_\eta-2)} + d_v\gamma_M, \frac{d_v}{1-\frac{2}{\gamma_\eta}}, \frac{1}{1-\frac{\delta}{\delta+2}-\frac{2}{p}} - \gamma_M\right)$, where $\gamma_\eta > 2$, $p > \delta+2$, $q = \frac{p(2+\delta)}{2p-2-\delta}$ for* *some* $\delta > 0$, *and* γ_M *is defined in Definition 1. Moreover, assume that* $\max\left(E|\eta_{\boldsymbol{i}}|^p, E|\eta_{\boldsymbol{i}}|^{\gamma_\eta}, E|\eta_{\boldsymbol{i}}|^{\delta+2}\right) < \infty$. *Then,*

$$E[|S(V_n)|^{2+\delta}] < C_*|V_n|^{1+\frac{\delta}{2}},$$

where C_* *is some sufficiently large generic constant such that* $C_* > \frac{4c_\delta A_\delta C_u}{1-\tau_0-4c_\delta \xi A_\delta}$, $c_\delta = \begin{cases} 1 & \text{if } \delta<1, \\ 2^{\delta-1} & \text{if } \delta\geq 1 \end{cases}$, C_u *is the generic constant chosen in Lemma 1,* $A_\delta > 0$ *is to ensure that (B-9) holds,* $\xi \in \left(0, \frac{1-\tau_0}{4c_\delta A_\delta}\right)$, *and* τ_0 *is some generic constant chosen less than 1.*

Proof. As the argument based on the decomposition of summation indices (used in the proof of Lemma 2) is rather cumbersome to apply in this current context, especially when δ is greater than 3, we shall here base the proof on an inductive argument, reminiscent of the one used in Bulinski and Shashkin (2007). First, note that, for a given δ, one can always choose an $A_\delta > 0$ to ensure that

$$(x+y)^2(1+x+y)^\delta \leq x^{2+\delta} + y^{2+\delta} + A_\delta\left((1+x)^\delta y^2 + x^2(1+y)^\delta\right) \text{ for any } x, y \geq 0. \tag{B-9}$$

Let $h(n) = \min\{k \in \mathbb{Z}_+ :\ 2^k \geq n\}$, $n \in \mathbb{N}$. For any sublattice, $V_n \subset \mathbb{Z}^{d_v}$, having edges of lengths, at most equal to $\ell_1, \ldots, \ell_{d_v}$, we define $h(V_n) = \sum_{i=1}^{d_v} h(\ell_i)$. We need to show that, for some $C*$ large enough and all sublattices, $V_n \subset \mathbb{Z}^{d_v}$,

$$E\left[S^2(V_n)\left(1+S(V_n)\right)^\delta\right] \leq C_*|V_n|^{1+\frac{\delta}{2}}. \tag{B-10}$$

When $h(V_n) = 0$ (i.e., $|V_n| = 1$), (B-10) is obviously true. Suppose that (B-10) holds for every U_n such that $h(U_n) \leq h_0$. One needs to verify that it also holds for any V_n such that $h(V_n) = h_0 + 1$, say. Let $\ell_+(V_n)$ represent the maximum length of the longest edge of V_n. Draw a hyperplane orthogonal to this longest edge, cutting this edge into two intervals of lengths, $\lfloor\frac{\ell_+(V_n)}{2}\rfloor$ and $\ell_+(V_n) - \lfloor\frac{\ell_+(V_n)}{2}\rfloor$. The hyperplane then divides V_n into two non-overlapping sublattices, say $V_{1,n}$ and $V_{2,n}$, with $h(V_{1,n}), h(V_{2,n}) \leq h_0$.

Let $Q_{1,n} = S(V_{1,n})$ and $Q_{2,n} = S(V_{2,n})$. By (B-9) and Lemma 13, one obtains that, for some $\tau_0 < 1$,

$$\begin{aligned} E\left[S^2(V_n)\left(1+S(V_n)\right)^\delta\right] &= E\left[(Q_{1,n}+Q_{2,n})^2(1+Q_{1,n}+Q_{2,n})^\delta\right] \\ &\leq C_*(|V_{1,n}|^{1+\frac{\delta}{2}} + |V_{2,n}|^{1+\frac{\delta}{2}}) + A_\delta\left\{E\left[(1+|Q_{1,n}|)^\delta Q_{2,n}^2\right] + E\left[(1+|Q_{2,n}|)^\delta Q_{1,n}^2\right]\right\} \\ &\leq C_*\tau_0|V_n|^{1+\frac{\delta}{2}} + A_\delta E\left[|1+Q_{1,n}|^\delta Q_{2,n}^2\right] + A_\delta E\left[|1+Q_{2,n}|^\delta Q_{1,n}^2\right]. \end{aligned} \tag{B-11}$$

We still need to bound $E\left[|1+Q_{1,n}|^\delta Q_{2,n}^2\right]$ and $E\left[|1+Q_{2,n}|^\delta Q_{1,n}^2\right]$. We shall now proceed with the former as the latter is quite similar. Introduce the subset $\mathcal{U}_n = \left\{\boldsymbol{s} \in V_{2,n} :\ d(\boldsymbol{s}, V_{1,n}) \leq \xi|V_n|^{1/d_v}\right\}$

for some $\xi \in \left(0, \frac{1-\tau_0}{4c_\delta A_\delta}\right)$, where c_δ is defined in (B-13). An application of the elementary inequality ($(a+b)^r \le c_r(a^r + b^r)$, $c_r = 1$ if $r < 1$ and $c_r = 2^{r-1}$ if $r \ge 1$) yields

$$\begin{aligned} E\left[|1+Q_{1,n}|^\delta Q_{2,n}^2\right] &= E\left[|1+Q_{1,n}|^\delta \left(S(\mathcal{U}_n) + S(V_{2,n}\backslash \mathcal{U}_n)\right)^2\right] \\ &\le 2E\left[|1+Q_{1,n}|^\delta S^2(\mathcal{U}_n)\right] + 2E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right] \\ &\le 2c_\delta E[S^2(\mathcal{U}_n)] + 2c_\delta E\left[|Q_{1,n}|^\delta S^2(\mathcal{U}_n)\right] + 2E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right] \\ &\le 2c_\delta E[S^2(\mathcal{U}_n)] + 2c_\delta \left(E\left[|Q_{1,n}|^{2+\delta}\right]\right)^{\frac{\delta}{\delta+2}} \left(E[S^{2+\delta}(\mathcal{U}_n)]\right)^{\frac{2}{\delta+2}} + 2E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right] \\ &\le 2c_\delta E[S^2(\mathcal{U}_n)] + 2c_\delta \xi C_* |V_n|^{1+\frac{\delta}{2}} + 2E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right]. \end{aligned} \quad \text{(B-12)}$$

Since $E[S^2(\mathcal{U}_n)] \le C_u|V_n|$, where C_u is some given constant, by Lemma 1, one then has

$$E\left[|1+Q_{1,n}|^\delta Q_{2,n}^2\right] \le 2c_\delta C_u|V_n| + 2c_\delta \xi C_* |V_n|^{1+\frac{\delta}{2}} + 2E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right]. \quad \text{(B-13)}$$

Moreover, notice that

$$\begin{aligned} E\left[|1+Q_{1,n}|^\delta S^2(V_{2,n}\backslash \mathcal{U}_n)\right] &\le \left|\sum_{\boldsymbol{i},\boldsymbol{j} \in V_{2,n}\backslash \mathcal{U}_n, \boldsymbol{i}\ne \boldsymbol{j}} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| + \left|\sum_{\boldsymbol{i} \in V_{2,n}\backslash \mathcal{U}_n} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}^2\right]\right| \\ &= \mathcal{A}_n + \mathcal{B}_n. \end{aligned} \quad \text{(B-14)}$$

To bound $\mathcal{A}_n$, note that, for each $\boldsymbol{i} \in V_{2,n}\backslash \mathcal{U}_n$,

$$\left|\sum_{\boldsymbol{j} \in V_{2,n}\backslash \mathcal{U}_n, \boldsymbol{i}\ne \boldsymbol{j}} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| \le \left|\sum_{\boldsymbol{j} \in V_{2,n}^{(1)}} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| + \left|\sum_{\boldsymbol{j} \in V_{2,n}^{(2)}, \boldsymbol{j}\ne \boldsymbol{i}} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| = \mathcal{A}_{1,n} + \mathcal{A}_{2,n},$$

where $V_{2,n}^{(1)} = \{\boldsymbol{j} \in V_{2,n}\backslash \mathcal{U}_n : \ \|\boldsymbol{j} - \boldsymbol{i}\| \ge \xi |V_n|^{1/d_v}\}$ and $V_{2,n}^{(2)} = \{\boldsymbol{j} \in V_{2,n}\backslash \mathcal{U}_n : \ \|\boldsymbol{j} - \boldsymbol{i}\| < \xi |V_n|^{1/d_v}\}$. For each pair, $\boldsymbol{i} \ne \boldsymbol{j} \in V_{2,n}\backslash \mathcal{U}_n$, define truncated random variables, $\eta_{1,\boldsymbol{i}} = \eta_{\boldsymbol{i}} \mathbf{1}(|\eta_{\boldsymbol{i}}| \le M(\boldsymbol{i},\boldsymbol{j}))$ and $\eta_{2,\boldsymbol{i}} = \eta_{\boldsymbol{i}} - \eta_{1,\boldsymbol{i}}$, where $M(\boldsymbol{i},\boldsymbol{j}) = (d(\{\boldsymbol{j}\}, \{\boldsymbol{i}\} \bigcup V_{1,n}))^{\theta_\eta}$ with $\frac{qd_v}{p-q} \le \theta_\eta \le (\theta_\alpha - d_v\gamma_M)(1-2/\gamma_\eta) - d_v$. Also let $|1+Q_{1,n}|^\delta = |1+Q_{1,n}|^\delta \mathbf{1}(|Q_{1,n}| \le L_n) + |1+Q_{1,n}|^\delta \mathbf{1}(|Q_{1,n}| > L_n)$, where $L_n = |V_n|^{1/2}$, one obtains that

$$\begin{aligned} E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right] &= Cov\left(|1+Q_{1,n}|^\delta \mathbf{1}(|Q_{1,n}| \le L_n)\eta_{1,\boldsymbol{i}}, \eta_{\boldsymbol{j}}\right) + Cov\left(|1+Q_{1,n}|^\delta \mathbf{1}(|Q_{1,n}| > L_n)\eta_{1,\boldsymbol{i}}, \eta_{\boldsymbol{j}}\right) \\ &\quad + Cov\left(|1+Q_{1,n}|^\delta \eta_{2,\boldsymbol{i}}, \eta_{\boldsymbol{j}}\right) = I + II + III, \end{aligned} \quad \text{(B-15)}$$

where, by Lemma 12,

$$I \leq C_\alpha \|\eta_{\boldsymbol{i}}\|_{\gamma_\eta} \left\| \eta_{1,\boldsymbol{i}} |1+Q_{1,n}|^\delta \mathbf{1}(|Q_{1,n}| \leq L_n) \right\|_{\gamma_\eta} \left\{ M_\alpha(1, |V_{1,n}|+1)\alpha\left(d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})\right)\right\}^{1-2/\gamma_\eta}$$
$$\leq C_\alpha C_{\gamma_\eta} L_n^\delta M(\boldsymbol{i},\boldsymbol{j}) \left\{ M_\alpha(1, |V_{1,n}|+1)\alpha\left(d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})\right)\right\}^{1-2/\gamma_\eta},$$

where $C_{\gamma_\eta} = \|\eta_{\boldsymbol{i}}\|_{\gamma_\eta}$ and C_α is the generic constant defined by Lemma 12. By the same argument, one can prove that

$$II \leq C_\alpha C_{\gamma_\eta} M(\boldsymbol{i},\boldsymbol{j}) \left(\frac{E\left[|1+Q_{1,n}|^{\delta\gamma_\eta} Q_{1,n}^2\right]}{L_n^2} \right)^{1/\gamma_\eta} \left\{ M_\alpha(1, |V_{1,n}|+1)\alpha\left(d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})\right)\right\}^{1-2/\gamma_\eta}$$
$$\leq C_\alpha C_{\gamma_\eta} C_*^{1/\gamma_\eta} M(\boldsymbol{i},\boldsymbol{j}) L_n^{-2/\gamma_\eta} |V_{1,n}|^{\frac{\delta}{2}+\frac{1}{\gamma_\eta}} \left\{ M_\alpha(1, |V_{1,n}|+1)\alpha\left(d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})\right)\right\}^{1-2/\gamma_\eta}.$$

An application of Hölder's inequality, one has

$$III \leq \left(E[|1+Q_{1,n}|^{2+\delta}]\right)^{\frac{\delta}{2+\delta}} \left(E|\eta_{2,\boldsymbol{i}}|^q\right)^{1/q} \|\eta_{\boldsymbol{i}}\|_p \leq C_p^{1+p/q} C_*^{\frac{\delta}{2+\delta}} \frac{|V_{1,n}|^{\frac{\delta}{2}}}{M(\boldsymbol{i},\boldsymbol{j})^{p/q-1}},$$

where $C_p = \|\eta_{\boldsymbol{i}}\|_p$. Therefore, in view of (B-15),

$$E\left[|1+Q_{1,n}|^\delta \eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right] \leq C_\alpha C_{\gamma_\eta} M(\boldsymbol{i},\boldsymbol{j}) \left\{ L_n^\delta + C_*^{1/\gamma_\eta} L_n^{-2/\gamma_\eta} |V_{1,n}|^{\frac{\delta}{2}+\frac{1}{\gamma_\eta}} \right\}$$
$$\left\{ M_\alpha(1, |V_{1,n}|+1)\alpha\left(d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})\right)\right\}^{1-2/\gamma_\eta} + C_p^{1+p/q} C_*^{\frac{\delta}{2+\delta}} \frac{|V_{1,n}|^{\frac{\delta}{2}}}{M(\boldsymbol{i},\boldsymbol{j})^{p/q-1}}.$$

It then follows that

$$\mathcal{A}_{1,n} \leq C_\alpha C_\theta^{1-2/\gamma_\eta} C_{\gamma_\eta} \left\{ L_n^\delta + C_*^{1/\gamma_\eta} L_n^{-2/\gamma_\eta} |V_{1,n}|^{\frac{\delta}{2}+\frac{1}{\gamma_\eta}} \right\} \left\{ M_\alpha(1, |V_n|+1) \right\}^{1-2/\gamma_\theta}$$
$$\sum_{\boldsymbol{j}\in V_{2,n}^{(1)}, \boldsymbol{j}\neq\boldsymbol{i}} d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})^{\theta_\eta - \theta_\alpha(1-2/\gamma_\eta)} + C_p^{1+p/q} C_*^{\frac{\delta}{2+\delta}} |V_{1,n}|^{\frac{\delta}{2}} \sum_{\boldsymbol{j}\in V_{2,n}^{(1)}, \boldsymbol{j}\neq\boldsymbol{i}} d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})^{-\theta_\eta \frac{p-q}{q}}.$$

Notice that, by Lemma 11, one can effectively show that

$$\sum_{\boldsymbol{j}\in V_{2,n}^{(1)}, \boldsymbol{j}\neq\boldsymbol{i}} d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})^{\theta_\eta - \theta_\alpha(1-2/\gamma_\eta)} \leq \sum_{m\geq \xi|V_n|^{1/d_v}} |\{\boldsymbol{j}\in V_{2,n}\backslash \mathcal{U}_n : \ \|\boldsymbol{j}-\boldsymbol{i}\| = m\}|\, m^{\theta_\eta - \theta_\alpha(1-2/\gamma_\eta)}$$
$$\leq 4d_v 2^{2d_v-3} \sum_{m\geq \xi|V_n|^{1/d_v}} m^{\theta_\eta + d_v - 1 - \theta_\alpha(1-2/\gamma_\eta)} \approx \frac{4d_v 2^{2d_v-3}}{\theta_\alpha(1-2/\gamma_\eta) - \theta_\eta - d_v} \left(\xi |V_n|^{\frac{1}{d_v}}\right)^{\theta_\eta + d_v - \theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}$$

and

$$\sum_{\boldsymbol{j}\in V_{2,n}^{(1)},\boldsymbol{j}\neq\boldsymbol{i}} d(\{\boldsymbol{j}\},\{\boldsymbol{i}\}\bigcup V_{1,n})^{-\theta_\eta\frac{p-q}{q}} \leq 4d_v 2^{2d_v-3}\sum_{m\geq\xi|V_n|^{1/d_v}} m^{d_v-1-\theta_\eta\frac{p-q}{q}} \approx \frac{4d_v 2^{2d_v-3}}{\theta_\eta\frac{p-q}{q}-d_v}\left(\xi|V_n|^{1/d_v}\right)^{d_v-\theta_\eta\frac{p-q}{q}}.$$

It then follows that

$$\mathcal{A}_{1,n} \;\leq\; 4d_v 2^{2d_v-3}\left\{\frac{\xi^{\theta_\eta+d_v-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}C_\alpha C_\theta^{1-2/\gamma_\eta}C_{\gamma_\eta}(C_*^{1/\gamma_\eta}+1)}{\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)-\theta_\eta-d_v}+\frac{\xi^{d_v-\theta_\eta\frac{p-q}{q}}C_p^{1+p/q}C_*^{\frac{\delta}{2+\delta}}}{\theta_\eta\frac{p-q}{q}-d_v}\right\}|V_n|^{\delta/2}. \tag{B-16}$$

Next, to derive the upper bound for $\mathcal{A}_{2,n}$, note that, for every pair, $\boldsymbol{i},\boldsymbol{j}\in V_{2,n}^{(2)}$, an application of the triangle inequality yields $\left|E\left[|1+Q_{1,n}|^\delta\eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| \leq \left|E\left[|1+Q_{1,n}|^\delta\eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]-E\left[|1+Q_{1,n}|^\delta\right]E\left[\eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| + \left|E\left[|1+Q_{1,n}|^\delta\right]E\left[\eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\right]\right| = \mathcal{A}_{2,a,n}+\mathcal{A}_{2,b,n}$. First, by Lemma 12 and Hölder's inequality, one can show that

$$\begin{aligned}\mathcal{A}_{2,a,n} &\leq C_\alpha\left(E[|1+Q_{1,n}|^{2+\delta}]\right)^{\frac{\delta}{2+\delta}}\|\eta_{\boldsymbol{i}}\eta_{\boldsymbol{j}}\|_{\frac{p}{2}}\left\{M_\alpha(2,|V_{1,n}|)\alpha(d(\{\boldsymbol{i},\boldsymbol{j}\},V_{1,n}))\right\}^{1-\frac{\delta}{2+\delta}-\frac{1}{p}}\\ &\leq C_\alpha C_p^2 C_*^{\frac{\delta}{\delta+2}}\xi^{-\theta_\alpha\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)}|V_n|^{\frac{\delta}{2}+(\gamma_M-\theta_\alpha)\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)}\end{aligned}$$

and

$$\mathcal{A}_{2,b,n}\leq C_\alpha C_{\gamma_\eta}^2 C_*^{\frac{\delta}{2+\delta}}|V_n|^{\frac{\delta}{2}}\alpha(\|\boldsymbol{i}-\boldsymbol{j}\|)^{1-\frac{2}{\gamma_\eta}}.$$

Therefore, one has

$$\begin{aligned}\mathcal{A}_{2,n} &\leq C_\alpha C_p^2 C_*^{\frac{\delta}{\delta+2}}\xi^{-\theta_\alpha\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)}|V_n|^{\frac{\delta}{2}+(\gamma_M-\theta_\alpha)\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)}\left|\{\boldsymbol{j}\in V_{2,n}^{(2)},\ \boldsymbol{j}\neq\boldsymbol{i}\}\right|\\ &\quad + C_\alpha C_{\gamma_\eta}^2 C_*^{\frac{\delta}{2+\delta}}|V_n|^{\frac{\delta}{2}}\sum_{\boldsymbol{j}\in V_{2,n}^{(2)},\boldsymbol{j}\neq\boldsymbol{i}}\|\boldsymbol{i}-\boldsymbol{j}\|^{-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}.\end{aligned}$$

Moreover, an application of Lemma 12 and an elementary inequality (i.e., $\sum_{k=1}^n k^p\approx\frac{n^{p+1}}{p+1},\ p>-1$ as $n\uparrow\infty$) yields

$$\begin{aligned}\left|\{\boldsymbol{j}\in V_{2,n}^{(2)},\ \boldsymbol{j}\neq\boldsymbol{i}\}\right| = \left|\{\boldsymbol{j}\in V_{2,n}\backslash\mathcal{U}_n:\ \|\boldsymbol{j}-\boldsymbol{i}\|<\xi|V_n|^{1/d_v}\}\right| &\leq \left|\{\boldsymbol{j}\in\mathbb{Z}^{d_v}:\ \|\boldsymbol{j}-\boldsymbol{i}\|<\xi|V_n|^{1/d_v}\}\right|\\ &\leq 2^{2(d_v-1)}\xi^{d_v}|V_n|\end{aligned}$$

and

$$\sum_{\boldsymbol{j}\in V_{2,n}^{(2)},\boldsymbol{j}\neq\boldsymbol{i}} \|\boldsymbol{i}-\boldsymbol{j}\|^{-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)} = \sum_{r=1}^{\xi|V_n|^{\frac{1}{d_v}}} \sum_{\boldsymbol{j}\in V_{2,n}^{(2)},\ \|\boldsymbol{j}-\boldsymbol{i}\|=r} r^{-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)} = \sum_{r=1}^{\xi|V_n|^{1/d_v}} \left|\{\boldsymbol{j}\in V_{2,n}^{(2)}:\ \|\boldsymbol{j}-\boldsymbol{i}\|=r\}\right| r^{-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}$$

$$\leq 2d_v 2^{2d_v-3}\sum_{r=1}^{\xi|V_n|^{\frac{1}{d_v}}} r^{d_v-1-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)} \leq 2d_v 2^{2d_v-3}\max\left(1, (\xi|V_n|)^{d_v-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}\right).$$

It then follows that

$$\mathcal{A}_{2,n} \leq C_\alpha C_*^{\frac{\delta}{\delta+2}}\left\{2^{2(d_v-1)}C_p^2\xi^{d_v-\theta_\alpha\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)} + 2d_v 2^{2d_v-3}C_{\gamma_\eta}^2\right\}|V_n|^{\frac{\delta}{2}}. \tag{B-17}$$

Finally, an application of Hölder's inequality yields $E[|1+Q_{1,n}|^\delta\eta_{\boldsymbol{i}}^2] \leq \left(E|1+Q_{1,n}|^{2+\delta}\right)^{\frac{\delta}{\delta+2}}\|\eta_{\boldsymbol{i}}\|_{2+\delta}^2 \leq C_*^{\frac{\delta}{\delta+2}}C_{\delta+2}^2|V_n|^{\frac{\delta}{2}}$, where $C_{\delta+2} = \|\eta_{\boldsymbol{i}}\|_{\delta+2}$. Therefore, one has

$$\mathcal{B}_n \leq C_*^{\frac{\delta}{\delta+2}}C_{\delta+2}^2|V_n|^{\frac{1+\delta}{2}} \tag{B-18}$$

Collecting all the results derived in (B-13)-(B-18), we have

$$E\left[|1+Q_{1,n}|^\delta Q_{2,n}^2\right] \leq 2c_\delta C_u|V_n| + 2\left\{c_\delta\xi C_* + 4d_v 2^{2d_v-3}\left\{\frac{\xi^{\theta_\eta+d_v-\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)}C_\alpha C_\theta^{1-2/\gamma_\eta}C_{\gamma_\eta}(C_*^{1/\gamma_\eta}+1)}{\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)-\theta_\eta-d_v}\right.\right.$$

$$\left.\left.+\frac{\xi^{d_v-\theta_\eta\frac{p-q}{q}}C_p^{1+p/q}C_*^{\frac{\delta}{2+\delta}}}{\theta_\eta\frac{p-q}{q}-d_v}\right\} + C_\alpha C_*^{\frac{\delta}{\delta+2}}\left\{2^{2(d_v-1)}C_p^2\xi^{d_v-\theta_\alpha\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)} + 2d_v 2^{2d_v-3}C_{\gamma_\eta}^2\right\} + C_*^{\frac{\delta}{\delta+2}}C_{\delta+2}^2\right\}|V_n|^{\frac{1+\delta}{2}}.$$

In view of (B-11), some algebraic manipulations yield

$$E\left[S^2(V_n)\left(1+S(V_n)\right)^\delta\right] \leq \left((\tau_0+4c_\delta\xi A_\delta)C_* + B_1 C_*^{\frac{\delta}{\delta+2}} + C_1 C_*^{\frac{1}{\gamma_\eta}} + D_1\right)|V_n|^{\frac{1+\delta}{2}} + 4c_\delta C_u A_\delta|V_n|, \tag{B-19}$$

where

$$B_1 = 4A_\delta\left\{4d_v 2^{2d_v-3}\frac{\xi^{d_v}-\theta_\eta\frac{p-q}{q}C_p^{1+\frac{p}{q}}}{\theta_\eta\frac{p-q}{q}-d_v} + C_\alpha\left\{2^{2(d_v-1)}C_p^2\xi^{d_v-\theta_\alpha\left(1-\frac{\delta}{\delta+2}-\frac{2}{p}\right)} + 2d_v 2^{2d_v-3}C_{\gamma_\eta}^2\right\} + C_{\delta+2}^2\right\}$$

and $C_1 = D_1 = 16A_\delta d_v 2^{2d_v-3} \frac{\xi^{\theta_\eta + d_v - \theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)} C_\alpha C_\theta^{1-\frac{2}{\gamma_\eta}} C_{\gamma_\eta}}{\theta_\alpha\left(1-\frac{2}{\gamma_\eta}\right)-\theta_\eta-d_v}$. The right-hand side of (B-19) is a root-polynomial function of C_*, thus will become less than C_* if C_* is large. The inductive argument has been proved. □

Lemma 4 (FCLT for Strongly Mixing Spatio-Temporal Data)**.** *Let* $S(V_n, \lfloor T\tau \rfloor) = \sum_{t=1}^{\lfloor T\tau \rfloor} \sum_{\boldsymbol{i}\in V_n} \eta_{\boldsymbol{i},t}$ *represent a partial-sum process of mixing centered spatio-temporal random fields,* $\{\eta_{\boldsymbol{i},t},\ \boldsymbol{i} \in V_n, t \in [1, T]\}$*. Suppose that* $\{\eta_{\boldsymbol{i},t},\ \boldsymbol{i} \in V_n, t \in [1, T]\}$ *are identically distributed across both space and time. Moreover, let the following conditions holds: (a)* $\alpha(\tau) \le C_\theta \tau^{-\theta_\alpha}$ *for some* $\theta_\alpha \ge \max\left(\frac{p(d_v+1)\gamma_\eta}{(p-q)(\gamma_\eta-2)} + (d_v+1)\gamma_M, \frac{d_v+1}{1-\frac{2}{\gamma_\eta}}, \frac{1}{1-\frac{\delta}{\delta+2}-\frac{2}{p}} - \gamma_M\right)$*, where* $\gamma_\eta > 2$*,* $p > \delta + 2$*,* $q = \frac{p(2+\delta)}{2p-2-\delta}$ *for some* $\delta > 0$*,* d_v *is the dimension of* V_n*, and* γ_M *is given in Definition 1; (b)* $\max\left(E|\eta_{\boldsymbol{i}}|^p, E|\eta_{\boldsymbol{i}}|^{\gamma_\eta}, E|\eta_{\boldsymbol{i}}|^{\delta+2}\right) < \infty$*; (c)* $|V_n|^{\gamma_M} T^{\gamma_M+1-\theta_\alpha} \downarrow 0$*. Then,*

$$\frac{1}{\sigma\sqrt{T}|V_n|^{1/2}} S(V_n, \lfloor T\tau \rfloor) \xrightarrow{w} W(\tau),$$

where $\sigma^2 = \lim_{n,T\uparrow\infty} \frac{1}{T|V_n|} E[S^2(V_n, T)] < \infty$ *and* $W(\tau)$ *is the Brownian motion.*

Proof. Let $D[0, 1]$ denote the Skorohod space of càdlàg functions on $[0, 1]$. (All the properties that we need can be found in Billingsley (1968).) The partial-sum process $S(V_n, \lfloor T\tau \rfloor)$ can be considered as a random function in $D[0, 1]$. Therefore, the FCLT for $S(V_n, \lfloor T\tau \rfloor)$ is reminiscent of Billingsley (1968, Theorem 20.1). As in Deo (1975) the weak convergence of $S(V_n, \lfloor T\tau \rfloor)$ to the Brownian motion requires the following conditions: Define $\overline{S}(V_n, \lfloor T\tau \rfloor) = \frac{1}{\sigma\sqrt{T}|V_n|^{1/2}} S(V_n, \lfloor T\tau \rfloor)$,

(i) $\lim_{n,T\uparrow\infty} E[\overline{S}^2(V_n, \lfloor T\tau \rfloor)] = \tau$ for each $\tau \in (0, 1]$,

(ii) $\overline{S}^2(V_n, \lfloor T\tau \rfloor)$ is uniformly integrable for each $\tau \in (0, 1]$,

(iii) $\overline{S}(V_n, \lfloor T\tau \rfloor)$ has asymptotically independent increments,

(iv) $\overline{S}(V_n, \lfloor T\tau \rfloor)$ is tight in $D[0, 1]$ (see Billingsley (1968, Theorem 19.2)).

Verification of **(i)***:* $\frac{1}{T|V_n|} Var\left(S(V_n, \lfloor T\tau \rfloor)\right) = \frac{\lfloor T\tau \rfloor}{T} E[\eta_{\boldsymbol{i},t}^2] + \frac{1}{T|V_n|} \sum_{s=1}^{\lfloor T\tau \rfloor} \sum_{t=1,t\neq s}^{\lfloor T\tau \rfloor} \sum_{\boldsymbol{i},\boldsymbol{j}\in V_n} E[\eta_{\boldsymbol{i},s}\eta_{\boldsymbol{j},t}] + \frac{1}{T|V_n|} \sum_{s=1}^{\lfloor T\tau \rfloor} \sum_{\boldsymbol{i},\boldsymbol{j}\in V_n,\boldsymbol{i}\neq\boldsymbol{j}} E[\eta_{\boldsymbol{i},s}\eta_{\boldsymbol{j},s}] = \frac{\lfloor T\tau \rfloor}{T} E[\eta_{\boldsymbol{i},t}^2] + \mathcal{A}_{n,T} + \mathcal{B}_{n,T}$.

$$\mathcal{A}_{n,T} = \frac{\lfloor T\tau \rfloor}{T} \frac{1}{|V_n|} \sum_{t=1}^{\lfloor T\tau \rfloor} \sum_{\boldsymbol{i},\boldsymbol{j}\in V_n} E[\eta_{\boldsymbol{i},0}\eta_{\boldsymbol{j},t}] = \frac{\lfloor T\tau \rfloor}{T} \left(\sum_{t=1}^{\lfloor T\tau \rfloor} \frac{1}{|V_n|} \sum_{\boldsymbol{i}\in V_n} E[\eta_{\boldsymbol{i},0}\eta_{\boldsymbol{i},t}] + \sum_{t=1}^{\lfloor T\tau \rfloor} \frac{1}{|V_n|} \sum_{\boldsymbol{i},\boldsymbol{j}\in V_n,\boldsymbol{i}\neq\boldsymbol{j}} E[\eta_{\boldsymbol{i},0}\eta_{\boldsymbol{j},t}] \right)$$

$$= \frac{\lfloor T\tau \rfloor}{T} (\mathcal{A}_{1,n,T} + \mathcal{A}_{2,n,T}), \quad \text{(B-20)}$$

where $\mathcal{A}_{1,n,T} \leq C_0 \sum_{\tau=1}^{\infty} \|\eta_{\boldsymbol{i},0}\|_{\gamma_\eta} \|\eta_{\boldsymbol{i},t}\|_{\gamma_\eta} \alpha(\tau)^{1-2/\gamma_\eta} < \infty$ by Lemma 12 and Conditions (a) and (b); and, by Lemma 11,

$$\mathcal{A}_{2,n,T} = \sum_{m=1}^{\mathrm{diam}(V_n \times [1,\lfloor T\tau \rfloor])} \frac{1}{|V_n|} \sum_{\boldsymbol{i} \in V_n} \sum_{\substack{t \in [1,\lfloor T\tau \rfloor] \\ \boldsymbol{j} \in V_n \\ d(\{\boldsymbol{i},0\},\{\boldsymbol{j},t\})=m}} E[\eta_{\boldsymbol{i},0}\eta_{\boldsymbol{j},t}]$$

$$\leq C_0 \sum_{m=1}^{\mathrm{diam}(V_n \times [1,\lfloor T\tau \rfloor])} \frac{1}{|V_n|} \sum_{\boldsymbol{i} \in V_n} |\{\{\boldsymbol{j},t\} \in V_n \times [1,\lfloor T\tau \rfloor]:\ d(\{\boldsymbol{i},0\},\{\boldsymbol{j},t\}) = m\}| \, \|\eta_{\boldsymbol{i},0}\|_{\gamma_\eta} \|\eta_{\boldsymbol{j},t}\|_{\gamma_\eta} \alpha(m)^{1-2/\gamma_\eta}$$

$$\leq C_0 \sum_{m=1}^{\infty} m^{d_v} \alpha(m)^{1-2/\gamma_\eta} < \infty.$$

By using the same argument, one can also verify that

$$\mathcal{B}_{n,T} = \frac{\lfloor T\tau \rfloor}{T} \frac{1}{|V_n|} \sum_{\substack{\boldsymbol{i},\boldsymbol{j} \in V_n \\ \boldsymbol{i} \neq \boldsymbol{j}}} E[\eta_{\boldsymbol{i},s}\eta_{\boldsymbol{j},s}] = \frac{\lfloor T\tau \rfloor}{T} \mathcal{B}_{1,n,T}, \tag{B-21}$$

where $\mathcal{B}_{1,n,T} \leq C_0 \sum_{m=1}^{\infty} m^{d_v - 1} \alpha(m)^{1-2/\gamma_\eta} < \infty$.

Notice that $\frac{\lfloor T\tau \rfloor}{T} \to \tau$ and $\sigma^2 = E[\eta_{\boldsymbol{i},t}^2] + \lim_{n,T\uparrow\infty}(\mathcal{A}_{1,n,T} + \mathcal{A}_{2,n,T} + \mathcal{B}_{1,n,T}) < \infty$, Condition (i) has been verified.

Verification of **(ii)***:* Is is sufficient to show that $\overline{S}^2(V_n, T)$ is uniformly integrable. An application of the Tchebyshev inequality and Lemma 3 yields that, for some $\delta > 0$,

$$\begin{aligned} E\left[\overline{S}^2(V_n,T)\mathbf{1}(|\overline{S}(V_n,T)| \geq C)\right] &= \frac{1}{\sigma^2 T|V_n|} E\left[S^2(V_n,T)\mathbf{1}\left(|S(V_n,T)| \geq \sigma\sqrt{T|V_n|}C\right)\right] \\ &\leq \frac{1}{\sigma^{2+\delta}C^\delta (T|V_n|)^{1+\delta/2}} E\left[|S(V_n,T)|^{2+\delta}\right] \\ &\leq \frac{C_*}{\sigma^{2+\delta}C^\delta} \to 0 \text{ as } C \to \infty. \end{aligned}$$

Verification of **(iii)***:* Let $0 = s_1 \leq t_1 < s_2 \leq t_2 < \cdots < s_m \leq t_m = 1$ denote a partition of the unit interval $[0,1]$. For all Borel sets, $H_1, \ldots, H_m$, of $\mathbb{R}$, one needs to show that

$$\lim_{n,T\uparrow\infty} \Big| P\left(\overline{S}(V_n, \lfloor Tt_i \rfloor) - \overline{S}(V_n, \lfloor Ts_i \rfloor) \in H_i,\ i = 1, \ldots m\right) - \prod_{i=1}^{m} P\left(\overline{S}(V_n, \lfloor Tt_i \rfloor) - \overline{S}(V_n, \lfloor Ts_i \rfloor) \in H_i\right) \Big| = 0. \tag{B-22}$$

Note that, as the event $\{\overline{S}(V_n, \lfloor Tt_i \rfloor) - \overline{S}(V_n, \lfloor Ts_i \rfloor) \in H_i\}$ belongs to the σ-algebra $\mathcal{B}_i$ generated by the sequence $\{\eta_{\boldsymbol{s}}: \ \boldsymbol{s} \in V_n \times [\lfloor Ts_i \rfloor + 1, \lfloor Tt_i \rfloor]\}$, the random element $\xi = \mathbf{1}\left(\overline{S}(V_n, \lfloor Tt_i \rfloor) - \overline{S}(V_n, \lfloor Ts_i \rfloor) \in H_i\right)$ is $\mathcal{B}_i$-measurable. By Lemma 14, one obtains that

$$\left| E\left[\prod_{s=1}^{m} \xi_s\right] - \prod_{s=1}^{m} E[\xi_s] \right| \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} \left| Cov\left(\xi_i - 1, (\xi_j - 1) \prod_{s=j+1}^{m} \xi_s\right) \right|. \tag{B-23}$$

Since $(\xi_j - 1) \prod_{s=j+1}^{m} \xi_s$ is $\bigcup_{i=j}^{m} \mathcal{B}_i$-measurable and $d\left(V_n \times [\lfloor Ts_i \rfloor + 1, \lfloor Tt_i \rfloor], V_n \times \bigcup_{\ell=j}^{m} [\lfloor Ts_\ell \rfloor + 1, \lfloor Tt_\ell \rfloor]\right) = \lfloor T(s_j - t_i) \rfloor > \lfloor Tb \rfloor > 0$ for every $j > i$, where b is some positive number, by Lemma 12, one obtains that

$$\begin{aligned} Cov\left(\xi_i - 1, (\xi_j - 1) \prod_{s=j+1}^{m} \xi_s\right) &\leq C_0 M_\alpha \left(|V_n \times [\lfloor Ts_i \rfloor + 1, \lfloor Tt_i \rfloor]|, \left| V_n \times \bigcup_{\ell=j}^{m} [\lfloor Ts_\ell \rfloor + 1, \lfloor Tt_\ell \rfloor] \right| \right) \alpha(\lfloor Tb \rfloor) \\ &\leq C_0 \left(T|V_n|\right)^{\gamma_M} \alpha(\lfloor Tb \rfloor) \approx \left(T|V_n|\right)^{\gamma_M} \left(\lfloor Tb \rfloor\right)^{-\theta_\alpha} \to 0 \text{ by Condition } (c). \end{aligned}$$

Therefore, in view of (B-23), (B-22) has been proved.

Verification of **(iv)***:* In view of Billingsley (1968, Theorem 8.4) (adapted to $D[0,1]$), the tightness condition will follow if one can prove that, for each positive ϵ, there exist a positive λ and integers, n_0 and T_0, such that $n \geq n_0$ and $T \geq T_0$ together imply

$$P\left(\max_{1 \leq t \leq T} |S(V_n, t)| \geq \sigma\lambda\sqrt{T|V_n|}\right) \leq \frac{\epsilon}{\lambda^2\sigma^2}. \tag{B-24}$$

First, introduce the events $E_1 = \left\{|S(V_n, 1)| \geq \sigma\lambda\sqrt{T|V_n|}\right\}$ and $E_j = \left\{\max_{1 \leq i < j} |S(V_n, i)| < \sigma\lambda\sqrt{T|V_n|} < |S(V_n, j)|\right\}$ for every $j > 1$. It then follows that

$$\begin{aligned} P\left(\max_{1 \leq t \leq T} |S(V_n, t)| \geq \sigma\lambda\sqrt{T|V_n|}\right) &\leq P\left(|S(V_n, T)| \geq (\lambda - \lambda_1)\sigma\sqrt{T|V_n|}\right) \\ &\quad + \sum_{j=1}^{T-1} P\left(E_j \bigcap \left\{|S(V_n, T) - S(V_n, j)| \geq \sigma\lambda\sqrt{T|V_n|}\right\}\right) \end{aligned} \tag{B-25}$$

for any $\lambda_1 < \lambda$. Note that

$$\begin{aligned} & P\left(E_j \bigcap \left\{|S(V_n, T) - S(V_n, j)| \geq \sigma\lambda\sqrt{T|V_n|}\right\}\right) \\ &\quad \leq P\left(E_j \bigcap \left\{|S(V_n, T) - S(V_n, j + k)| \geq \sigma(\lambda_1 - \lambda_2)\sqrt{T|V_n|}\right\}\right) \\ &\quad\quad + P\left(|S(V_n, j + k) - S(V_n, j)| \geq \sigma\lambda_2\sqrt{T|V_n|}\right) = \mathcal{A}_{n,T} + \mathcal{B}_{n,T}, \end{aligned}$$

where $\lambda_2 \in (0, \lambda_1)$ and k takes some value less than T. To bound the right-hand side of (B-25), one first needs to bound $\mathcal{A}_{n,T}$ and $\mathcal{B}_{n,T}$. Let $\mathcal{B}_i^j$ be the σ-algebra generated by $\{\eta_{\boldsymbol{s}} : \boldsymbol{s} \in V_n \times [i,j]\}$. Thus the Bernoulli random variables $\mathbf{1}(E_j)$ is $\mathcal{B}_1^j$-measurable and $\mathbf{1}\left(|S(V_n,T) - S(V_n, j+k)| \geq \sigma(\lambda_1 - \lambda_2)\sqrt{T|V_n|}\right)$ is $\mathcal{B}_{j+k+1}^T$-measurable. Invoking Lemma 12, one obtains

$$
\begin{aligned}
&\Big| P\left(E_j \bigcap |S(V_n,T) - S(V_n,j+k)| \geq \sigma(\lambda_1-\lambda_2)\sqrt{T|V_n|}\right) \\
&\qquad\qquad - P(E_j)P\left(|S(V_n,T) - S(V_n,j+k)| \geq \sigma(\lambda_1-\lambda_2)\sqrt{T|V_n|}\right)\Big| \\
&\qquad\qquad\qquad \leq C_0 M_\alpha\left(|V_n|j, |V_n|(T-j-k)\right)\alpha\left(d\left(V_n\times[1,j], V_n\times[j+k+1,T]\right)\right) \\
&\qquad\qquad\qquad\qquad \leq C_0(T|V_n|)^{\gamma_M}\alpha(k).
\end{aligned}
$$

In addition,

$$
\begin{aligned}
P\left(|S(V_n,T) - S(V_n,j+k)| \geq \sigma(\lambda_1-\lambda_2)\sqrt{T|V_n|}\right) &\overset{(a)}{\leq} \frac{E[|S(V_n,T)-S(V_n,j+k)|^2]}{\sigma^2(\lambda_1-\lambda_2)^2 T|V_n|} \\
&\leq \frac{1}{\sigma^2(\lambda_1-\lambda_2)^2 T|V_n|}\left(\sum_{\boldsymbol{s}\in V_n\times[1,j+k]} E[|\eta_{\boldsymbol{s}}|^2] + \sum_{\boldsymbol{s},\boldsymbol{w}\in V_n\times[1,j+k],\ \boldsymbol{s}\neq\boldsymbol{w}} E[\eta_{\boldsymbol{s}}\eta_{\boldsymbol{w}}]\right) \\
&\overset{(b)}{\leq} |V_n|T\left(\|\eta_{\boldsymbol{s}}\|^2 + 2C_0 d_v 3^{\gamma_M(1-2/\gamma_\eta)}\|\eta_{\boldsymbol{s}}\|^2_{\gamma_\eta}\sum_{r=1}^{\infty}(2r+1)^{d_v}\alpha(r)^{1-2/\gamma_\eta}\right) = |V_n|T\Theta_\eta,
\end{aligned}
$$

where Conditions (a) and (b) ensure that $\Theta_\eta < \infty$; (a) follows from the Tchebyshev inequality; and (b) follows from Lemma 12. It then follows that

$$
\mathcal{A}_{n,T} \leq P(E_j)\frac{\Theta_\eta}{\sigma^2(\lambda_1-\lambda_2)^2} + C_0(T|V_n|)^{\gamma_M}\alpha(k). \tag{B-26}
$$

By the Tchebyshev inequality and Lemma 3, one also has

$$
\begin{aligned}
\mathcal{B}_{n,T} &\leq \sum_{s=j+1}^{j+k} P\left(\left|\sum_{\boldsymbol{i}\in V_n}\eta_{\boldsymbol{i},s}\right| \geq \frac{\sigma\lambda_2\sqrt{|V_n|T}}{k}\right) \\
&\qquad \leq \frac{k^{3+\delta}}{\sigma^{2+\delta}\lambda_2^{2+\delta}(|V_n|T)^{1+\delta/2}} E\left[\left|\sum_{\boldsymbol{i}\in V_n}\eta_{\boldsymbol{i},s}\right|^{2+\delta}\right] \leq C_*\frac{k^{3+\delta}}{\sigma^{2+\delta}\lambda_2^{2+\delta}T^{1+\frac{\delta}{2}}}
\end{aligned} \tag{B-27}
$$

for some $\delta > 0$. In view of (B-25), (B-26), and (B-27), we have

$$P\left(\max_{1\le t\le T}|S(V_n,t)| \ge \sigma\lambda\sqrt{T|V_n|}\right) \le P\left(|S(V_n,T)| \ge (\lambda-\lambda_1)\sigma\sqrt{T|V_n|}\right) + \frac{\Theta_\eta}{\sigma^2(\lambda_1-\lambda_2)^2}\sum_{j=1}^{T-1}P(E_j) + C_0T^{\gamma_M+1}|V_n|^{\gamma_M}\alpha(k) + C_*\frac{k^{3+\delta}}{\sigma^{2+\delta}\lambda_2^{2+\delta}T^{\frac{\delta}{2}}}.$$

Because the events E_j, $j = 1,\dots,T-1$ are disjoint and $\bigcup_{j=1}^{T-1}E_j \subset \{\max_{1\le j\le T}|S(V_n,t)| \ge \sigma\lambda\sqrt{T|V_n|}\}$, one can immediately show that

$$P\left(\max_{1\le t\le T}|S(V_n,t)| \ge \sigma\lambda\sqrt{T|V_n|}\right) \le \left(1-\frac{\Theta_\eta}{\sigma^2(\lambda_1-\lambda_2)^2}\right)^{-1}\left(P\left(|S(V_n,T)| \ge (\lambda-\lambda_1)\sigma\sqrt{T|V_n|}\right) + C_0T^{\gamma_M+1}|V_n|^{\gamma_M}\alpha(k) + C_*\frac{k^{3+\delta}}{\sigma^{2+\delta}\lambda_2^{2+\delta}T^{\frac{\delta}{2}}}\right). \quad \text{(B-28)}$$

Now, let $\lambda_1 = \lambda/2$. For a given $\epsilon > 0$, one can choose λ sufficiently large so that

$$P\left(|S(V_n,T)| \ge \frac{1}{2}\lambda\sigma\sqrt{T|V_n|}\right) \le \frac{\epsilon}{9\lambda^2\sigma^2},$$

which is possible because of the uniform integrability condition (ii). One can also choose $\lambda_2 < \lambda_1$ so that $\frac{\Theta_\eta}{\sigma^2(\lambda_1-\lambda_2)^2} < \frac{2}{3}$. Therefore, (B-28) results in

$$P\left(\max_{1\le t\le T}|S(V_n,t)| \ge \sigma\lambda\sqrt{T|V_n|}\right) \le 3\left(\frac{\epsilon}{9\sigma^2\lambda^2} + C_0|V_n|^{\gamma_M}T^{\gamma_M+1-\theta_\alpha} + C_*\frac{k^{3+\delta}}{\sigma^{2+\delta}\lambda_2^{2+\delta}T^{\frac{\delta}{2}}}\right).$$

If one chooses $k < T$ such that $k^{3+\delta}/T^{\delta/2}$ is arbitrarily small, Condition (c) then implies that

$$P\left(\max_{1\le t\le T}|S(V_n,t)| \ge \sigma\lambda\sqrt{T|V_n|}\right) \le \frac{\epsilon}{\sigma^2\lambda^2}.$$

The tightness condition was verified. □

Lemma 5. *Let $S(U_n,V_n,T) = \sum_{t=1}^{T}\sum_{\boldsymbol{i}\in U_n}\sum_{\boldsymbol{j}\in V_n} w_{\boldsymbol{i},t}\epsilon_{\boldsymbol{j},t}$, where $\{w_{\boldsymbol{i},t},\ \boldsymbol{i}\in U_n\}$ and $\{\epsilon_{\boldsymbol{j},t},\ \boldsymbol{j}\in V_n\}$ are contemporaneously independent centered spatio-temporal processes; and for given $\boldsymbol{i}\in U_n$ and $\boldsymbol{j}\in V_n$, $w_{\boldsymbol{i},t}$ is a causal process and $\{\epsilon_{\boldsymbol{j},t},\ t = 1,\dots,T\}$ are independent over time. In addition, suppose that (a) the processes are identically distributed across both space and time, (b) $w_{\boldsymbol{i},t}$ and $\epsilon_{\boldsymbol{j},t}$ are mixing with the mixing coefficient satisfying $\alpha(\tau) \le C_\theta\tau^{-\theta_\alpha}$ for some $\theta_\alpha \ge \max\left(\frac{pd_v\gamma_\eta}{(p-q)(\gamma_\eta-2)} + d_v\gamma_M, \frac{d_v}{1-\frac{2}{\gamma_\eta}}, \frac{2p}{p-4} - \gamma_M\right)$, where $\gamma_\eta > 2$, $p > 4$, $q = \frac{4p}{2p-4}$, d_v is the dimension of*

V_n, *and* γ_M *is given in Definition 1, (c)* $\max\left(\left(|U_n|+|V_n|\right)^{\gamma_M(1-2/\gamma_\eta)}T^{\epsilon-1/2}, T^{(\gamma_M+\theta_\alpha-1)\epsilon-\frac{1}{2}(\theta_\alpha-\gamma_M-1)}\right.$ $\left.\max(|U_n|,|V_n|)^{\gamma_M}\right)\downarrow 0$ *for some* $\epsilon\in\left(0,\min\left(\frac{1}{2},\frac{\theta_\alpha-\gamma_M-1}{2(\gamma_M+\theta_\alpha-1)}\right)\right)$*. Then,*

$$\frac{1}{\sqrt{T|U_n||V_n|}}S(U_n,V_n,T)\xrightarrow{d}N(0,\sigma^2),$$

where $\sigma^2=\lim_{n\uparrow\infty}\frac{1}{|U_n||V_n|}E\left|\sum_{\boldsymbol{i}\in U_n}w_{\boldsymbol{i},t}\right|^2E\left|\sum_{\boldsymbol{i}\in V_n}\epsilon_{\boldsymbol{i},t}\right|^2$.

Proof. Let $S^*(U_n,V_n,T)=\sum_{t=1}^T w_{*,t}\epsilon_{*,t}$, where $w_{*,t}=\frac{1}{|U_n|}\sum_{\boldsymbol{i}\in U_n}w_{\boldsymbol{i},t}$ and $\epsilon_{*,t}=\frac{1}{|V_n|}\sum_{\boldsymbol{i}\in V_n}\epsilon_{\boldsymbol{i},t}$.

Step I: Divide the time-period index set $[1,T]$ into k_T big blocks, $\{\eta^{(b)}_{n,T,i},\ i=0,\dots,k_T-1\}$, of size $p_T=\lfloor T^{1/2+\epsilon}\rfloor$ and k_T+1 small blocks, $\{\eta^{(s)}_{n,T,i},\ i=0,\dots,k_T\}$, of size $q_T=\lfloor T^{1/2-\epsilon}\rfloor$ for some small $0<\epsilon<\frac{\theta_\alpha-\gamma_M-1}{2(\gamma_M+\theta_\alpha-1)}$, and put

$$\eta^{(b)}_{n,T,i}=\sum_{j=1}^{p_T}w_{*,i(p_T+q_T)+j}\epsilon_{*,i(p_T+q_T)+j},\ i=0,\dots,k_T-1,$$

$$\eta^{(s)}_{n,T,i}=\sum_{j=1}^{p_T}w_{*,i(p_T+q_T)+p_T+j}\epsilon_{*,i(p_T+q_T)+p_T+j},\ i=0,\dots,k_T-1,$$

and the small remaining block $\eta^{(s)}_{n,T,k_T}=\sum_{j=k_T(p_T+q_T)+1}^T w_{*,j}\epsilon_{*,j}$. It then follows that

$$S^*(U_n,V_n,T)=\sum_{i=0}^{k_T-1}\eta^{(b)}_{n,T,i}+\sum_{i=0}^{k_T}\eta^{(s)}_{n,T,i}=\mathcal{B}_{n,T}+\mathcal{S}_{n,T}. \tag{B-29}$$

Step II: Derive the asymptotic variance for $S^*(U_n,V_n,T)$. Notice that $E\left|S^*(U_n,V_n,T)\right|^2=\sum_{t=1}^T E\left|w_{*,t}\epsilon_{*,t}\right|^2+\sum_{t\neq s}E[w_{*,t}\epsilon_{*,t}w_{*,s}\epsilon_{*,s}]=\mathcal{T}_{1,n,T}+\mathcal{T}_{2,n,T}$, where $\frac{|U_n||V_n|}{T}\mathcal{T}_{1,n,T}=\frac{1}{|U_n|}E\left|\sum_{\boldsymbol{i}\in U_n}w_{\boldsymbol{i},t}\right|^2$ $\frac{1}{|V_n|}E\left|\sum_{\boldsymbol{i}\in V_n}\epsilon_{\boldsymbol{i},t}\right|^2<\infty$ as $n\to\infty$ by Lemma 1 and $\mathcal{T}_{2,n,T}=2\sum_{s<t}E[\epsilon_{*,t}]E[w_{*,t}w_{*,s}\epsilon_{*,s}]=0$. One then obtains that

$$\sigma^2=\lim_{n\uparrow\infty}\left(\frac{1}{|U_n|}E\left|\sum_{\boldsymbol{i}\in U_n}w_{\boldsymbol{i},t}\right|^2\right)\left(\frac{1}{|V_n|}E\left|\sum_{\boldsymbol{i}\in V_n}\epsilon_{\boldsymbol{i},t}\right|^2\right).$$

Step III: Let $\mathcal{S}^*_{n,T}=\sqrt{\frac{|U_n||V_n|}{T}}\frac{1}{\sigma}\mathcal{S}_{n,T}$. Then,

$$E\left|\mathcal{S}^*_{n,T}\right|^2=\frac{|U_n||V_n|}{T\sigma^2}E\left|\sum_{i=0}^{k_T-1}\eta^{(s)}_{n,T,i}+\eta_{n,T,k_T}\right|^2$$
$$=\frac{|U_n||V_n|}{T\sigma^2}\left(E|\eta^{(s)}_{n,T,k_T}|^2+k_TE\left|\eta^{(s)}_{n,T,0}\right|^2+2E\left[\sum_{i=0}^{k_T-1}\eta^{(s)}_{n,T,i}\eta^{(s)}_{n,T,k_T}\right]\right),$$

where $E|\eta^{(s)}_{n,T,k_T}|^2 = O\left(\frac{\lfloor T-k_T(p_T+q_T)\rfloor}{|U_n||V_n|}\right)$ and $E\left|\eta^{(s)}_{n,T,0}\right|^2 = O\left(\frac{q_T}{|U_n||V_n|}\right)$ by Lemma 1. In addition, $E\left[\sum_{i=0}^{k_T-1}\eta^{(s)}_{n,T,i}\eta^{(s)}_{n,T,k_T}\right] = \sum_{j=1}^{q_T}\sum_{\ell=k_T(p_T+q_T)+1}^{T} E[w_{*,i(p_T+q_T)+p_T+j}w_{*,\ell}\epsilon_{*,i(p_T+q_T)+p_T+j}\epsilon_{*,\ell}] = 0$. It then follows that $E\left|\mathcal{S}^*_{n,T}\right|^2 = O\left(\frac{\lfloor T-k_T(p_T+q_T)\rfloor}{T} + \frac{q_T k_T}{T}\right) = o(1)$.

Step IV: Let $\mathcal{B}^*_{n,T} = \sqrt{\frac{|U_n||V_n|}{T}}\frac{1}{\sigma}\mathcal{B}_{n,T}$. Show that

$$Q_1 = \left|E\exp\left(i\theta\mathcal{B}^*_{n,T}\right) - \prod_{i=0}^{k_T-1} E\exp\left(i\theta\sqrt{\frac{|U_n||V_n|}{T}}\frac{1}{\sigma}\eta^{(b)}_{n,T,i}\right)\right| = o(1), \text{ where } i = \sqrt{-1}. \tag{B-30}$$

To do so, invoking Lemma 12 yields $Q_1 \le C_0\sum_{j=0}^{k_T-2} M_\alpha\left(|V_n|p_T, (k_T-j-1)p_T|V_n|\right)\alpha(q_T)$ $\le C_0 k_T\left(|V_n|p_T\right)^{\gamma_M}\alpha(q_T) \overset{(a)}{\le} C_0 T^{1/2-\epsilon}|V_n|^{\gamma_M}T^{\gamma_M(1/2+\epsilon)}\alpha(q_T) \overset{(b)}{\le} C_0|V_n|^{\gamma_M}T^{(\gamma_M+\theta_\alpha-1)\epsilon-\frac{1}{2}(\theta_\alpha-\gamma_M-1)}$, where (a) follows because $k_T\lfloor\frac{T}{p_T+q_T}\rfloor \le q_T$, and (b) follows because of Condition (b). Now, invoking Condition (c), we obtain $Q_1 = o(1)$.

Step V: Show that $\sum_{i=0}^{k_T-1} E\left|\sqrt{\frac{|U_n||V_n|}{T}}\frac{1}{\sigma}\eta^{(b)}_{n,T,i}\right|^2 \to 1$. An application of Lemma 1 yields

$$\sum_{i=0}^{k_T-1} E\left|\sqrt{\frac{|U_n||V_n|}{T}}\frac{1}{\sigma}\eta^{(b)}_{n,T,i}\right|^2 = \frac{|U_n||V_n|}{T}\frac{1}{\sigma^2}\sum_{i=0}^{k_T} p_T E[w^2_{*,t}\epsilon^2_{*,t}] = \frac{p_T k_T}{T} \to 1.$$

Step VI: Finally, we need to verify the following uniform integrability condition

$$\frac{|U_n||V_n|}{T}\sum_{i=0}^{k_T-1} E\left[|\eta^{(b)}_{n,T,i}|^2\mathbf{1}\left(|\eta^{(b)}_{n,T,i}| > \lambda\sqrt{\frac{T}{|U_n||V_n|}}\sigma\right)\right] \to 0 \tag{B-31}$$

for every $\lambda > 0$. Invoking the Tchebyshev inequality, the left-hand side of (B-31) is dominated by $\frac{1}{\lambda^2\sigma^2}\left(\frac{|U_n||V_n|}{T}\right)^2\sum_{i=0}^{k_T} E\left|\eta^{(b)}_{n,T,i}\right|^4$. To study this upper bound, one needs to bound $E\left|\eta^{(b)}_{n,T,i}\right|^4$. For ease of notation, we shall write $\widetilde{w}_{*,j} = w_{*,i(p_T+q_T)+j}$ and $\widetilde{\epsilon}_{*,j} = \epsilon_{*,i(p_T+q_T)+j}$. Therefore,

$$\begin{aligned}
E\left|\eta^{(b)}_{n,T,i}\right|^4 &= \sum_{j=1}^{p_T} E\left|\widetilde{w}_{*,j}\widetilde{\epsilon}_{*,j}\right|^4 + \sum_{j_1\neq j_2}^{p_T} E[\widetilde{w}^2_{*,j_1}\widetilde{\epsilon}^2_{*,j_1}\widetilde{w}^2_{*,j_2}\widetilde{\epsilon}^2_{*,j_2}] \\
&\quad + \sum_{j_1\neq j_2}^{p_T} E[\widetilde{w}^3_{*,j_1}\widetilde{\epsilon}^3_{*,j_1}\widetilde{w}_{*,j_2}\widetilde{\epsilon}_{*,j_2}] + \sum_{j_1\neq j_2\neq j_3 \text{ (pairwise)}}^{p_T} E[\widetilde{w}^2_{*,j_1}\widetilde{\epsilon}^2_{*,j_1}\widetilde{w}_{*,j_2}\widetilde{\epsilon}_{*,j_2}\widetilde{w}_{*,j_3}\widetilde{\epsilon}_{*,j_3}] \\
&\quad + \sum_{j_1\neq j_2\neq j_3\neq j_4 \text{ (pairwise)}}^{p_T} E[\widetilde{w}_{*,j_1}\widetilde{\epsilon}_{*,j_1}\widetilde{w}_{*,j_2}\widetilde{\epsilon}_{*,j_2}\widetilde{w}_{*,j_3}\widetilde{\epsilon}_{*,j_3}\widetilde{w}_{*,j_4}\widetilde{\epsilon}_{*,j_4}] \\
&= \mathcal{A}_{n,T} + \mathcal{B}_{n,T} + \mathcal{C}_{n,T} + \mathcal{D}_{n,T} + \mathcal{E}_{n,T}.
\end{aligned} \tag{B-32}$$

An application of Hölder's inequality and Lemma 3 under Condition (b) yields

$$\mathcal{A}_{n,T} < C_* \frac{p_T}{|U_n|^2 |V_n|^2} \tag{B-33}$$

and

$$\mathcal{B}_{n,T} < 2 \sum_{j_1<j_2}^{p_T} E[\widetilde{\epsilon}^2_{*,j_2}] E[\widetilde{w}^2_{*,j_1} \widetilde{\epsilon}^2_{*,j_1} \widetilde{w}^2_{*,j_2}] \leq 2 p_T^2 E[\widetilde{\epsilon}^2_{*,j_1}] E^{1/3}[\widetilde{w}^6_{*,1}] E^{1/3}[\widetilde{\epsilon}^6_{*,j_1}] E^{1/3}[\widetilde{w}^6_{*,1}]$$

$$\leq C_0 \frac{p_T^2}{|U_n|^2 |V_n|^2}. \tag{B-34}$$

Moreover, by the covariance inequality (Lemma 12),

$$\mathcal{C}_{n,T} = \sum_{j_1>j_2} E[\widetilde{\epsilon}^3_{*,j_1}] E[\widetilde{w}^3_{*,j_1} \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2}]$$

$$< C_0 \left| E[\widetilde{\epsilon}^3_{*,j_1}] \right| \|\widetilde{w}^3_{*,j_1}\|_{\gamma_\eta} \|\widetilde{w}_{*,j_2}\|_{\gamma_\eta} \|\widetilde{\epsilon}_{*,j_2}\|_{\gamma_\eta} M_\alpha(|U_n|, |V_n|)^{1-2/\gamma_\eta} \sum_{j_1>j_2} \alpha(|j_1 - j_2|)^{1-2/\gamma_\eta}.$$

It is immediate to verify that all the conditions set out in Lemma 2 hold, therefore, one has $E[\widetilde{\epsilon}^3_{*,j_1}] < C_* \frac{1}{|V_n|^{3/2}}$. Also, by invoking Lemma 3, one obtains $\|\widetilde{w}^3_{*,j_1}\|_{\gamma_\eta} < C_* \frac{1}{|U_n|^{3/2}}$, $\|\widetilde{w}_{*,j_2}\|_{\gamma_\eta} < C_* \frac{1}{|U_n|^{1/2}}$, and $\|\widetilde{\epsilon}_{*,j_2}\|_{\gamma_\eta} < C_* \frac{1}{|V_n|^{1/2}}$. It then follows that

$$\mathcal{C}_{n,T} < C_* \frac{p_T (|U_n| + |V_n|)^{\gamma_M (1-2/\gamma_\eta)}}{|U_n|^2 |V_n|^2} \sum_{\tau=1}^{\infty} \alpha(\tau)^{1-2/\gamma_\eta}. \tag{B-35}$$

Notice that

$$\mathcal{D}_{n,T} = \underbrace{\sum_{\substack{j_1<\min(j_2,j_3) \\ j_2 \neq j_3}}^{p_T} E[\widetilde{w}^2_{*,j_1} \widetilde{\epsilon}^2_{*,j_1} \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2} \widetilde{w}_{*,j_3} \widetilde{\epsilon}_{*,j_3}]}_{=0} + \sum_{\substack{j_1>\min(j_2,j_3) \\ j_2 \neq j_3}} E[\widetilde{\epsilon}^2_{*,j_1}] E[\widetilde{w}^2_{*,j_1} \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2} \widetilde{w}_{*,j_3} \widetilde{\epsilon}_{*,j_3}]$$

$$\leq C_0 E[\widetilde{\epsilon}^2_{*,j_1}] \sum_{j_1>j_2>j_3}^{p_T} \left| E[\widetilde{w}^2_{*,j_1} \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2} \widetilde{w}_{*,j_3} \widetilde{\epsilon}_{*,j_3}] \right|,$$

where - by the same argument as above - one has $E[\widetilde{\epsilon}^2_{*,j_1}] < C \frac{1}{|V_n|}$ and

$$\left| E[\widetilde{w}^2_{*,j_1} \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2} \widetilde{w}_{*,j_3} \widetilde{\epsilon}_{*,j_3}] \right| \leq C_0 \|\widetilde{w}^2_{*,j_1}\|_{\gamma_\eta} \left\| \widetilde{w}_{*,j_2} \widetilde{\epsilon}_{*,j_2} \widetilde{w}_{*,j_3} \widetilde{\epsilon}_{*,j_3} \right\|_{\gamma_\eta} M_\alpha(|U_n|, |V_n|)^{1-2/\gamma_\eta} \alpha(|j_1 - j_2|)^{1-2/\gamma_\eta}$$

with $\|\widetilde{w}_{*,j_1}^2\|_{\gamma_\eta} < C_0 \frac{1}{|U_n|}$ and $\|\widetilde{w}_{*,j_2}\widetilde{\epsilon}_{*,j_2}\widetilde{w}_{*,j_3}\widetilde{\epsilon}_{*,j_3}\|_{\gamma_\eta} < C_0 \frac{1}{|U_n||V_n|}$. It then follows that

$$\mathcal{D}_{n,T} < C_0 \frac{p_T^2(|U_n| + |V_n|)^{\gamma_M(1-2/\gamma_\eta)}}{|U_n|^2|V_n|^2} \sum_{\tau=1}^{\infty} \alpha(\tau)^{1-2/\gamma_\eta}. \tag{B-36}$$

Finally, it is not difficult to see that

$$\mathcal{E}_{n,T} < C_0 \sum_{j_1<j_2<j_3<j_4}^{p_T} E[\widetilde{w}_{*,j_1}\widetilde{\epsilon}_{*,j_1}\widetilde{w}_{*,j_2}\widetilde{\epsilon}_{*,j_2}\widetilde{w}_{*,j_3}\widetilde{\epsilon}_{*,j_3}\widetilde{w}_{*,j_4}\widetilde{\epsilon}_{*,j_4}] = 0. \tag{B-37}$$

Therefore, in view of (B-32)-(B-37), we have

$$E\left|\eta_{n,T,i}^{(b)}\right|^4 < C_0 \left\{ \frac{p_T^2}{|U_n|^2|V_n|^2} + \frac{p_T^2(|U_n| + |V_n|)^{\gamma_M(1-2/\gamma_\eta)}}{|U_n|^2|V_n|^2} \sum_{\tau=1}^{\infty} \alpha(\tau)^{1-2/\gamma_\eta} \right\}.$$

Invoking Condition (c), (B-31) has been verified. The main theorem then follows in view of Steps *I-VI* above. □

Lemma 6. *Let* $\{X_{\boldsymbol{s},t} : \ \boldsymbol{s} \in V_N, \ t \in [1,T]\}$ *be a mixing spatio-temporal process. Suppose that (a)* $X_{\boldsymbol{s},t}$, $\boldsymbol{s} \in V_N$ *and* $t \in [1,T]$ *are identically distributed over time and space; (b)* $E[\exp(\ell\|X_{\boldsymbol{s},t}\|)] \leq C_\ell$ *for a constant* $C_\ell > 0$ *and* $\ell > 0$ *small enough; (c)* $\|X_{\boldsymbol{s},t}\|_{\delta_\alpha} < \infty$ *for some* $\delta_\alpha > 2$*; (d)* $\alpha(\tau) \leq C_0\tau^{-\theta_\alpha}$ *for some* $\theta_\alpha > \left(\frac{4\gamma_M}{3}, \frac{d_v+1}{1-2/\delta_\alpha}\right)^+$. *Then,*

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_{*,t}\right| \geq M\right) \leq \frac{2}{M}T^{-C_\alpha} + C_0 N^{\gamma_M}\log(T)T^{\gamma_M - \frac{3}{4}\theta_\alpha} + 4\max\left\{\exp\left(-\frac{1}{256}\left(\frac{M}{C_\ell}\right)^2\frac{NT}{2C_\sigma}\right),\right.$$
$$\left.\exp\left(-\frac{1}{32}\frac{N}{C_\ell}\frac{T^{1/4}}{\log(T)}\right)\right\}^+,$$

where C_σ *and* C_α *are sufficiently large constants.*

Proof. First, we employ the following truncation: $X_{\boldsymbol{s},t} = X_{\boldsymbol{s},t}^{(<)} + X_{\boldsymbol{s},t}^{(>)}$ with $X_{\boldsymbol{s},t}^{(<)} = X_{\boldsymbol{s},t}\mathbf{1}(|X_{\boldsymbol{s},t}| \leq C_x\log(T))$ and $X_{\boldsymbol{s},t}^{(>)} = X_{\boldsymbol{s},t} - X_{\boldsymbol{s},t}^{(<)}$. It then follows that

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_{*,t}\right| \geq M\right) \leq P\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_{*,t}^{(<)} - E[X_{*,t}^{(<)}]\right| \geq M/2\right)$$
$$+ P\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_{*,t}^{(>)} - E[X_{*,t}^{(>)}]\right| \geq M/2\right) = \mathcal{T}_{<,N,T} + \mathcal{T}_{>,N,T}. \tag{B-38}$$

By the Tchebyshev inequality and Conditions (a) and (b), one could show that

$$\mathcal{T}_{>,N,T} \leq \frac{2}{M} E[|X_{\boldsymbol{s},t}| \mathbf{1}(|X_{\boldsymbol{s},t}| > C_x \log(T))] \leq \frac{2}{M} T^{-C_\alpha}, \tag{B-39}$$

where C_x can always be chosen to make C_α large enough. To bound $\mathcal{T}_{<,N,T}$, let μ_T and b_T denote two divergent sequences so that $T - b_T < 2\mu_T b_T \leq T$, divide $\{X_{*,1}, \ldots, X_{*,T}\}$ into $2\mu_T$ blocks of size b_T. We can always choose b_T and μ_T in such a way that the remainder $\{X_{*,T-2\mu_T b_T}, \ldots, X_{*,T}\}$ can be ignored. Let $(\xi_{\boldsymbol{s},1}, \ldots, \xi_{\boldsymbol{s},b_T}), (\xi_{\boldsymbol{s},b_T+1}, \ldots, \xi_{\boldsymbol{s},2b_T}), \ldots, (\xi_{\boldsymbol{s},(2\mu_T-1)b_T+1}, \ldots, \xi_{\boldsymbol{s},2\mu_T b_T})$ be independent blocks of random elements such that $(\xi_{\boldsymbol{s},jb_T+1}, \ldots, \xi_{\boldsymbol{s},(j+1)b_T})$ and $(X^{(<)}_{\boldsymbol{s},jb_T+1}, \ldots, X^{(<)}_{\boldsymbol{s},(j+1)b_T})$, $j = 1, \ldots, 2\mu_T$, have the same distribution. Moreover, define

$$Z_{*,j} = \sum_{t=(2j-1)b_T+1}^{2jb_T} \xi_{*,t}, \;\; j = 1, 2 \ldots, \mu_T.$$

It then follows that

$$\begin{aligned}
\mathcal{T}_{<,N,T} &\leq 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}\sum_{t=(2j-1)b_T+1}^{2jb_T} X^{(<)}_{*,t} - E[X^{(<)}_{*,t}]\right| \geq \frac{M}{4}\right) \\
&\leq 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}\left\{\sum_{t=(2j-1)b_T+1}^{2jb_T} X^{(<)}_{*,t} - Z_{*,j}\right\}\right| \geq \frac{M}{8}\right) + 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T} Z_{*,j} - E[Z_{*,j}]\right| \geq \frac{M}{8}\right) \\
&= \mathcal{T}^{(a)}_{<,N,T} + \mathcal{T}^{(b)}_{<,N,T}.
\end{aligned}$$

Let $S_{N,j} = [(2j-1)b_T + 1, 2jb_T] \times V_N$, $j = 1, \ldots, \mu_T$, then $d(S_{N,i}, S_{N,j}) \geq b_T$ for every $i \neq j$. Since $Z_{*,j}$ is $\mathcal{B}(S_{N,j})$-measurable, $|Z_{*,j}| \leq C_x b_T \log(T)$, and $S_{N,j}$ contains Nb_T sites, an application of Rio's coupling inequality (Lemma 15) yields

$$\mathcal{T}^{(a)}_{<,N,T} \leq 2C_x \mu_T b_T \log(T) M_\alpha(Nb_T\mu_T, Nb_T)\alpha(b_T) \leq C_0 N^{\gamma_M} \log(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha}. \tag{B-40}$$

In addition, as $|Z_{*,j} - E[Z_{*,j}]| \leq 2C_x b_T \log(T)$, thus $|\widetilde{Z}_{*,j}| = \frac{|Z_{*,j} - E[Z_{*,j}]|}{2C_x b_T \log(T)} \leq 1$ and

$$Var(\widetilde{Z}_{*,j}) = \frac{Var(Z_{*,j})}{4C_x^2 b_T^2 \log^2(T)} \leq \frac{1}{2C_x^2 b_T^2 \log^2(T)} Var\left(\sum_{t=1}^{b_T} X_{*,t}\right). \tag{B-41}$$

Some combinatorics arguments yield

$$
\begin{aligned}
Var\left(\sum_{t=1}^{b_T} X_{*,t}\right) &= \frac{1}{N^2}\sum_{t=1}^{b_T}\sum_{\boldsymbol{s}\in V_N} Var(X_{\boldsymbol{s},t}) \\
&+ \frac{1}{N^2}\sum_{r=1}^{\operatorname{diam}(V_N\times[1,b_T])}\sum_{(\boldsymbol{s},t)\in V_N\times[1,b_T]}\sum_{\substack{(\boldsymbol{w},\tau)\in V_N\times[1,b_T]\\ \|(\boldsymbol{s},t)-(\boldsymbol{w},\tau)\|=r}} Cov(X_{\boldsymbol{s},t}, X_{\boldsymbol{w},\tau}) \\
&\le \frac{1}{N^2}\sum_{t=1}^{b_T}\sum_{\boldsymbol{s}\in V_N} Var(X_{\boldsymbol{s},t}) \\
+ C_0\frac{1}{N^2}\sum_{(\boldsymbol{w},\tau)\in V_N\times[1,b_T]}&\sum_{r=1}^{\operatorname{diam}(V_N\times[1,b_T])} |\{(\boldsymbol{s},t)\in V_N\times[1,b_T]:\ \|(\boldsymbol{s},t)-(\boldsymbol{w},\tau)\|=r\}| \\
&\|X_{\boldsymbol{s},t}\|_{\delta_\alpha}^2\left\{M_\alpha(1,1)\alpha(r)\right\}^{1-\frac{2}{\delta_\alpha}} < C_sigma\frac{b_T}{N}.
\end{aligned}
$$

It then follows from (B-41) that

$$
Var(\widetilde{Z}_{*,j}) \le C_\sigma \frac{1}{N b_T \log^2(T)}.
$$

Invoking Lemma 19, one obtains that

$$
\mathcal{T}^{(b)}_{<,N,T} = P\left(\left|\sum_{j=1}^{\mu_T}\widetilde{Z}_{*,j}\right| \ge \frac{M}{16C_x}\frac{T}{b_T\log(T)}\right) \le 2\max\left\{\exp\left(-\frac{1}{4}\left(\frac{M}{8C_x}\right)^2\frac{Nb_T}{C_\sigma}\right), \exp\left(-\frac{M}{32C_x}\frac{T^{1/4}}{\log(T)}\right)\right\}^+. \tag{B-42}
$$

The main lemma then follows from (B-38)-(B-42). ☐

Lemma 7. *Let $\{(X_{\boldsymbol{i},t}, \epsilon_{\boldsymbol{i},t}) : \ \boldsymbol{i}\in V_N,\ t\in[1,T]\}$ represent a mixing bivariate spatio-temporal process. Suppose that (a) $\{X_{\boldsymbol{i},t}, \epsilon_{\boldsymbol{i},t}\}$, $\boldsymbol{i}\in V_N$ and $t\in[1,T]$ are identically distributed over time and space; (b) $\alpha(\tau) < C_0\tau^{-\theta_\alpha}$, $\theta_\alpha > \left(\frac{4\gamma_M}{3}, \frac{2d_v+1}{1-2/\delta_\alpha}\right)^+$ for some $\delta_\alpha > 2$; (c) $\|X_{\boldsymbol{i},t}\epsilon_{\boldsymbol{i},t}\|_{\delta_\alpha} < \infty$; (d) $E[\exp(\ell\|X_{\boldsymbol{s},t}\|)] \le C_\ell$ for a constant $C_\ell > 0$ and $\ell > 0$ small enough. Then,*

$$
\begin{aligned}
P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\{X_{*,t}\epsilon_{*,t} - E[X_{*,t}\epsilon_{*,t}]\}\right| \ge M\right) &\le C_0\Big\{T^{-C_\alpha} + N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha} \\
&+ \max\left\{\exp\left(-C_\sigma N^2\log^2(T)T^{7/4}\right), \exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}\Big\},
\end{aligned}
$$

where C_σ and C_M are some sufficiently large constants.

Proof. Define the following truncated random variables: $X_{\boldsymbol{i},t} = X^{(<)}_{\boldsymbol{i},t} + X^{(>)}_{\boldsymbol{i},t}$ with $X^{(<)}_{\boldsymbol{i},t} = X_{\boldsymbol{i},t}\mathbf{1}(|X_{\boldsymbol{i},t}| \le$

$C_x \log(T))$ and $X^{(>)}_{\boldsymbol{i},t} = X_{\boldsymbol{i},t}\mathbf{1}(|X_{\boldsymbol{i},t}| > C_x \log(T))$; $\epsilon_{\boldsymbol{i},t} = \epsilon^{(<)}_{\boldsymbol{i},t} + \epsilon^{(>)}_{\boldsymbol{i},t}$ with $\epsilon^{(<)}_{\boldsymbol{i},t} = \epsilon_{\boldsymbol{i},t}\mathbf{1}(|\epsilon_{\boldsymbol{i},t}| \le C_\epsilon \log(T))$ and $\epsilon^{(>)}_{\boldsymbol{i},t} = \epsilon_{\boldsymbol{i},t}\mathbf{1}(|\epsilon_{\boldsymbol{i},t}| > C_\epsilon \log(T))$. Thus, $X\epsilon = X^{(<)}\epsilon^{(<)} + X^{(<)}\epsilon^{(>)} + X^{(>)}\epsilon^{(<)} + X^{(>)}\epsilon^{(>)}$. It then follows that

$$
\begin{aligned}
&P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\{X_{*,t}\epsilon_{*,t} - E[X_{*,t}\epsilon_{*,t}]\}\right| \ge M\right) \le P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\left\{X^{(<)}_{*,t}\epsilon^{(<)}_{*,t} - E\left[X^{(<)}_{*,t}\epsilon^{(<)}_{*,t}\right]\right\}\right| \ge \frac{M}{4}\right)\\
&+P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\left\{X^{(<)}_{*,t}\epsilon^{(>)}_{*,t} - E\left[X^{(<)}_{*,t}\epsilon^{(>)}_{*,t}\right]\right\}\right| \ge \frac{M}{4}\right) + P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\left\{X^{(>)}_{*,t}\epsilon^{(<)}_{*,t} - E\left[X^{(>)}_{*,t}\epsilon^{(<)}_{*,t}\right]\right\}\right| \ge \frac{M}{4}\right)\\
&\quad + P\left(\frac{1}{T}\left|\sum_{t=1}^{T}\left\{X^{(>)}_{*,t}\epsilon^{(>)}_{*,t} - E\left[X^{(>)}_{*,t}\epsilon^{(>)}_{*,t}\right]\right\}\right| \ge \frac{M}{4}\right) = \mathcal{T}_{1,N,T} + \mathcal{T}_{2,N,T} + \mathcal{T}_{3,N,T} + \mathcal{T}_{4,N,T}. \quad \text{(B-43)}
\end{aligned}
$$

To bound $\mathcal{T}_{1,N,T}$, let $w_{*,t} = X^{(<)}_{*,t}\epsilon^{(<)}_{*,t}$. Divide $\{w_{*,1}, \ldots, w_{*,T}\}$ into $2\mu_T$ blocks, $\{w_{*,(j-1)b_T+1}, \ldots, w_{*,jb_T}\}$, $j = 1, \ldots, 2\mu_T$, of size b_T and a smaller remaining block. One can always choose μ_T and b_T such that the remaining block is negligible so that it can be ignored. Define $2\mu_T$ contemporaneously independent blocks, $\{(\xi_{\boldsymbol{i},1}, \zeta_{\boldsymbol{i},1}), \ldots (\xi_{\boldsymbol{i},b_T}, \zeta_{\boldsymbol{i},b_T})\}, \{(\xi_{\boldsymbol{i},b_T+1}, \zeta_{\boldsymbol{i},b_T+1}), \ldots, (\xi_{\boldsymbol{i},2b_T}, \zeta_{\boldsymbol{i},2b_T})\}, \ldots, \{(\xi_{\boldsymbol{i},(2\mu_T-1)b_T}, \zeta_{\boldsymbol{i},(2\mu_T-1)b_T}),$ $\ldots, (\xi_{\boldsymbol{i},2\mu_T b_T}, \zeta_{\boldsymbol{i},2\mu_T b_T})\}$, such that $(\xi_{\boldsymbol{i},(j-1)b_T+1}, \ldots, \xi_{\boldsymbol{i},jb_T})$ and $\{X_{\boldsymbol{i},(j-1)b_T+1}, \ldots, X_{\boldsymbol{i},jb_T}\}$ are identically distributed; and $(\zeta_{\boldsymbol{i},(j-1)b_T+1}, \ldots, \zeta_{\boldsymbol{i},jb_T})$ and $\{\epsilon_{\boldsymbol{i},(j-1)b_T+1}, \ldots, \epsilon_{\boldsymbol{i},jb_T}\}$ are identically distributed. Let $Z_{*,j} = \sum_{t=(2j-1)b_T+1}^{2jb_T} \xi_{*,t}\zeta_{*,t}$, $j = 1, \ldots, \mu_T$. We have

$$
\begin{aligned}
\mathcal{T}_{1,N,T} &\le 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}\sum_{t=(2j-1)b_T+1}^{2jb_T} w_{*,t} - E[w_{*,t}]\right| \ge \frac{M}{8}\right)\\
&\le 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}\left(\sum_{t=(2j-1)b_T+1}^{2jb_T} w_{*,t} - Z_{*,j}\right)\right| \ge \frac{M}{16}\right) + 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}(Z_{*,j} - E[Z_{*,j}])\right| \ge \frac{M}{16}\right)\\
&= \mathcal{T}^{(a)}_{1,N,T} + \mathcal{T}^{(b)}_{1,N,T}. \quad \text{(B-44)}
\end{aligned}
$$

Define $S_{N,j} = [(2j-1)b_T+1, 2jb_T] \times V_N \times V_N$. Then, $d(S_{N,j}, S_{N,k}) \ge b_T$ for $j \ne k$ and $\sum_{t=(2j-1)b_T+1}^{2jb_T} w_{*,t}$ is $\mathcal{B}(S_{N,j})$-measurable. Since $|Z_{*,j}| \le C_x C_\epsilon b_T \log^2(T)$, an application of Lemma 15 yields

$$
\mathcal{T}^{(a)}_{1,N,T} \le C_0 N^{2\gamma_M} \log^2(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha}.
$$

To bound $\mathcal{T}^{(b)}_{1,N,T}$, notice that

$$
\begin{aligned}
Var\left(Z_{*,1}\right) &= \frac{1}{N^4} Var\left(\sum_{(\boldsymbol{i},\boldsymbol{j},t)\in V_{N,T}} X^{(<)}_{\boldsymbol{i},t}\epsilon^{(<)}_{\boldsymbol{j},t}\right) \\
&= \frac{1}{N^4}\sum_{(\boldsymbol{i},\boldsymbol{j},t)\in V_{N,T}} Var(X^{(<)}_{\boldsymbol{i},t}\epsilon^{(<)}_{\boldsymbol{j},t}) + \frac{1}{N^4}\sum_{\substack{(\boldsymbol{i}_1,\boldsymbol{j}_1,s)\in V_{N,T}\\(\boldsymbol{i}_2,\boldsymbol{j}_2,t)\in V_{N,T}\\ \|(\boldsymbol{i}_1,\boldsymbol{j}_1,s)-(\boldsymbol{i}_2,\boldsymbol{j}_2,t)\|\neq 0}} Cov(X^{(<)}_{\boldsymbol{i}_1,s}\epsilon^{(<)}_{\boldsymbol{j}_1,s}, X^{(<)}_{\boldsymbol{i}_2,t}\epsilon^{(<)}_{\boldsymbol{j}_2,t}),
\end{aligned}
$$

where the second term is bounded by
$C_0\frac{1}{N^4}\sum_{(\boldsymbol{i}_1,\boldsymbol{j}_1,s)\in V_{N,T}}\sum_{r=1}^{\mathrm{diam}(V_{N,T})}|\{(\boldsymbol{i}_2,\boldsymbol{j}_2,t)\in V_{N,T}:\ \|(\boldsymbol{i}_2,\boldsymbol{j}_2,t)-(\boldsymbol{i}_1,\boldsymbol{j}_1,s)\|=4\}|\,\|X_{\boldsymbol{i}_1,s}\epsilon_{\boldsymbol{j}_1,s}\|^2_{\delta_\alpha}$
$\{M_\alpha(1,1)\alpha(r)\}^{1-2/\delta_\alpha} \leq C_0\frac{b_T}{N^2}\sum_{r=1}^{\infty} r^{2d_v}\alpha(r)^{1-\frac{2}{\delta_\alpha}}$ in view of Lemma 12. Conditions (b) and (c) imply that

$$
Var\left(Z_{*,1}\right) \leq C_\sigma \frac{b_T}{N^2}.
$$

Let $\widetilde{Z}_{*,j} = \frac{1}{C_x C_\epsilon b_T\log^2(T)}|Z_{*,j}| \leq 1$. Invoking Lemma 19, one can show that

$$
\begin{aligned}
\mathcal{T}^{(b)}_{1,N,T} &\leq 2P\left(\left|\frac{1}{T}\sum_{j=1}^{\mu_T}(Z_{*,j}-E[Z_{*,j}])\right| \geq C_\sigma\mu_T\frac{1}{b_T\log^4(T)N^2}\frac{M}{32C_\epsilon C_x C_\sigma}\frac{N^2\log^2(T)T}{\mu_T}\right) \\
&\leq 4\max\left\{\exp\left(-C_\sigma\frac{N^2\log^2(T)T^2}{\mu_T}\right), \exp\left(-C_M\frac{\mu_T}{\log^2(T)}\right)\right\}.
\end{aligned}
$$

By choosing $\mu_T = O(T^{1/4})$ and $b_T = O\left(\lfloor\frac{T}{2\mu_T}\rfloor\right)$, we obtain that

$$
\mathcal{T}^{(b)}_{1,N,T} \leq 4\max\left\{\exp\left(-C_\sigma N^2\log^2(T)T^{7/4}\right), \exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}.
$$

It then follows from (B-44) that

$$
\mathcal{T}_{1,N,T} \leq C_0 N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha} + 4\max\left\{\exp\left(-C_\sigma N^2\log^2(T)T^{7/4}\right), \exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}. \tag{B-45}
$$

To bound the remaining terms in (B-43), Condition (d) implies that one can choose C_ϵ such that, for C_α large enough, $E\left[|\epsilon_{\boldsymbol{i},t}|^2\mathbf{1}(|\epsilon_{\boldsymbol{i},t}|>C_\epsilon\log(T))\right] \leq T^{-C_\alpha}$ and $E\left[|X_{\boldsymbol{i},t}|^2\mathbf{1}(|X_{\boldsymbol{i},t}|>C_\epsilon\log(T))\right] \leq T^{-C_\alpha}$. Therefore, we have

$$
\mathcal{T}_{2,N,T} \leq \frac{8}{M}\frac{1}{N^2}\sum_{\boldsymbol{i},\boldsymbol{j}\in V_N} E[|X^{(<)}_{\boldsymbol{i},t}\epsilon^{(>)}_{\boldsymbol{j},t}|] \leq \frac{8}{M}\frac{1}{N^2}\sum_{\boldsymbol{i},\boldsymbol{j}\in V_N}\|X_{\boldsymbol{i},t}\|_2\left\|\epsilon^{(>)}_{\boldsymbol{j},t}\right\|_2 \leq C_0 T^{-C_\alpha}. \tag{B-46}
$$

Similarly, we also obtain $\mathcal{T}_{3,N,T} \le C_0 T^{-C_\alpha}$ and $\mathcal{T}_{4,N,T} \le C_0 T^{-2C_\alpha}$. The main lemma then follows from (B-43)-(B-46). $\square$

Lemma 8. *The function*

$$\mathcal{H}_{1,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i=1}^{N}\frac{1}{2}\left(\rho_u u_{i,c}^2 + \rho_\phi \frac{\phi_c^2}{N} + \rho_\theta \frac{\boldsymbol{\theta}_c^\top \boldsymbol{\theta}_c}{N}\right) - \mathcal{E}_{1,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) + \mathcal{A}_0$$

is convex for every $\boldsymbol{\rho} \doteq (\rho_u, \rho_\phi, \rho_\theta)$ *satisfying* (B-47)-(B-48), (B-53)-(B-55).

Proof. Write $\mathcal{H}_{1,N,T}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i=1}^{N}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U})$, where $\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{1}{2}\left(\rho_u u_{i,c}^2 + \rho_\phi \frac{\phi_c^2}{N} + \rho_\theta \frac{\boldsymbol{\theta}_c^\top \boldsymbol{\theta}_c}{N}\right) - N^2\mathcal{A}_0 - u_{i,c}^2\phi_c^2\mathcal{A}_{1,i} - u_{i,c}^2\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 2u_{i,c}^2\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{C}_{1,i} + 2Nu_{i,c}\phi_c\mathcal{D}_{1,i} - 2Nu_{i,c}\phi_c\boldsymbol{\theta}_c^\top\mathcal{F}_{1,i}$. One needs to verify that $\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U})$ is convex for each $i \in [1,N]$ and $c \in [1,G]$. It is equivalent to showing that the minimum eigenvalue of the Hessian matrix is strictly positive. The positivity of the minimum eigenvalue of a matrix can be verified by the positive definiteness of all the sub-matrices. Some simple calculations yield the second-order derivatives of $\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U})$:

$$\begin{aligned}
a_{2,2}^{(1)} &\doteq \partial^2_{\phi_c}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\phi}{N} - 2u_{i,c}^2\mathcal{A}_{1,i} - 2u_{i,c}^2\boldsymbol{\theta}_c^\top\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 4u_{i,c}^2\boldsymbol{\theta}_c^\top\mathcal{C}_{1,i},\\
\boldsymbol{a}_{2,3}^{(1)} &\doteq \partial^2_{\phi_c,\boldsymbol{\theta}_c}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = -4u_{i,c}^2\phi_c\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 4u_{i,c}^2\phi_c\mathcal{C}_{1,i} - 2Nu_{i,c}\mathcal{F}_{1,i}(=\boldsymbol{a}_{2,3}^{(1)}),\\
\boldsymbol{a}_{3,3}^{(1)} &\doteq \partial^2_{\boldsymbol{\theta}_c}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\theta}{N}\mathbb{I}_{d_x} - 2u_{i,c}^2\phi_c^2\mathcal{B}_{1,i},\\
a_{1,1}^{(1)} &\doteq \partial^2_{u_{i,c}}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = \rho_u - 2\phi_c^2\mathcal{A}_{1,i} - 2\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 4\phi_c^2\boldsymbol{\theta}_c^\top\mathcal{C}_{1,i},\\
a_{1,2}^{(1)} &\doteq \partial^2_{u_{i,c}\phi_c}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = -4u_{i,c}\phi_c\mathcal{A}_{1,i} - 4u_{i,c}\phi_c\boldsymbol{\theta}_c^\top\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 8u_{i,c}\phi_c\boldsymbol{\theta}_c^\top\mathcal{C}_{1,i} + 2N\mathcal{D}_{1,i} - 2N\boldsymbol{\theta}_c^\top\mathcal{F}_{1,i}(=a_{2,1}^{(1)}),\\
\boldsymbol{a}_{1,3}^{(1)} &= \partial^2_{u_{i,c}\boldsymbol{\theta}_c}\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U}) = -4u_{i,c}\phi_c^2\mathcal{B}_{1,i}\boldsymbol{\theta}_c + 4u_{i,c}\phi_c^2\mathcal{C}_{1,i} - 2N\phi_c\mathcal{F}_{1,i}(=\boldsymbol{a}_{3,1}^{(1)}).
\end{aligned}$$

Let $\mathcal{H}_1 = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \boldsymbol{a}_{1,3}^{(1)\top} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & \boldsymbol{a}_{2,3}^{(1)\top} \\ \boldsymbol{a}_{3,1}^{(1)} & \boldsymbol{a}_{3,2}^{(1)} & \boldsymbol{a}_{3,3}^{(1)} \end{pmatrix}$ denote the Hessian matrix of $\mathfrak{f}_{1,N,T}^{(c,i)}(\boldsymbol{\gamma},\boldsymbol{U})$. The positive definiteness of the first Hessian sub-matrix is warranted by

$$\rho_u \ge \max_{i,c}\left\{2\ell_{\phi,c}^2|\mathcal{A}_{1,i}| + 2\ell_{\phi,c}^2\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}\left(\mathcal{B}_{1,i}\right) + 4\ell_{\phi,c}^2|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{C}_{1,i}|\right\}, \tag{B-47}$$

where $\boldsymbol{\ell}_{\theta,c} = (\ell_{\theta,c,1},\ldots,\ell_{\theta,c,d_x})^\top$. The positive definiteness of the second Hessian sub-matrix is warranted by

$$\begin{aligned}\rho_u \ge \max_{i,c}\Big\{&(2\ell_{\phi,c}^2 + 4\ell_{\phi,c})|\mathcal{A}_{1,i}| + 2\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}(\mathcal{B}_{1,i})\left(\ell_{\phi,c}^2 + 2\ell_{\phi,c}\right) + 4(\ell_{\phi,c}^2 + 2\ell_{\phi,c})|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{C}_{1,i}|\\ &+2N|\mathcal{D}_{1,i}| + 2N|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{F}_{1,i}|\Big\}\end{aligned} \tag{B-48}$$

and

$$\rho_\phi \geq N \max_{i,c} \left\{2|\mathcal{A}_{1,i}|(1+2\ell_{\phi,c}) + 2(1+2\ell_{\phi,c})\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}(\mathcal{B}_{1,i}) + 4(1+2\ell_{\phi,c})|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{C}_{1,i}| \right.$$
$$\left. +2N|\mathcal{D}_{1,i}| + 2N|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{F}_{1,i}|\right\}. \quad \text{(B-49)}$$

In view of Lemma 20 the positive definiteness of the third Hessian sub-matrices is warranted by

$$\frac{1}{d+2}\left(a_{1,1}^{(1)} + a_{1,2}^{(1)} + \boldsymbol{a}_{1,3}^{(1)\top}\imath_{d,d_x}\right) > \max(0, a_{1,2}^{(1)}, \boldsymbol{a}_{1,3}^{(1)\top}), \quad \text{(B-50)}$$
$$\frac{1}{d+2}\left(a_{1,2}^{(1)} + a_{2,2}^{(1)} + \boldsymbol{a}_{2,3}^{(1)\top}\imath_{d,d_x}\right) > \max(0, a_{1,2}^{(1)}, \boldsymbol{a}_{2,3}^{(1)\top}), \quad \text{(B-51)}$$
$$\frac{1}{d+2}\left(\boldsymbol{a}_{1,3}^{(1)} + \boldsymbol{a}_{2,3}^{(1)} + \boldsymbol{a}_{3,3}^{(1)}\imath_{d,d_x}\right) > \max\left(0, \boldsymbol{a}_{1,3}^{(1)}, \boldsymbol{a}_{2,3}^{(1)}, -2\phi_c^2 u_{i,c}^2\mathcal{B}_{1,i}\right) \quad \text{(B-52)}$$

for all $d = 1, \ldots, d_x$, where $\imath_{d,d_x} = (\underbrace{\underbrace{1, \ldots, 1}_{d}, 0, \ldots, 0}_{d_x})^\top$. We immediately verify that $|a_{1,2}^{(1)}| \leq \ell_{a_{12}}^{(1)} \doteq \max_{c,i}\left\{4\ell_{\phi,c}|\mathcal{A}_{1,i}| + 4\ell_{\phi,c}\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}(\mathcal{B}_{1,i}) + 8\ell_{\phi,c}|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{C}_{1,i}| + 2N|\mathcal{D}_{1,i}| - 2N|\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{F}_{1,i}|\right\}$, $|\boldsymbol{a}_{2,3}^{(1)}| \leq \boldsymbol{\ell}_{a_{23}}^{(1)} \doteq \max_{c,i}\left\{4\ell_{\phi,c}(|\mathcal{B}_{1,i}|\boldsymbol{\ell}_{\theta,c} + \mathcal{C}_{1,i}) + 2N\mathcal{F}_{1,i}\right\}$, and $|\boldsymbol{a}_{1,3}^{(1)}| \leq \boldsymbol{\ell}_{a_{13}}^{(1)} \doteq \max_{c,i}\left\{4\ell_{\phi,c}^2|\mathcal{B}_{1,i}||\boldsymbol{\ell}_{\theta,c}| + 4\ell_{\phi,c}^2|\mathcal{C}_{1,i}| + 2N\ell_{\phi,c}|\mathcal{F}_{1,i}|\right\}$. Therefore, Eq. (B-50) is implied by

$$\rho_u \geq \max_{d,c,i}\left\{2\ell_{\phi,c}^2|\mathcal{A}_{1,i}| + 2\ell_{\phi,c}^2\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}(\mathcal{B}_{1,i}) + 4\ell_{\phi,c}^2\left|\boldsymbol{\ell}_{\theta,c}^\top\right||\mathcal{C}_{1,i}| + \ell_{a_{12}}^{(1)} + \boldsymbol{\ell}_{a_{13}}^\top\imath_{d,d_x}\right.$$
$$\left. +(d+2)\max\left(\ell_{a_{12}}^{(1)}, \boldsymbol{\ell}_{a_{13}}^{(1)\top}\right)\right\}. \quad \text{(B-53)}$$

Eq. (B-51) is implied by

$$\rho_\phi \geq N\max_{d,c,i}\left\{2|\mathcal{A}_{1,i}| + 2\|\boldsymbol{\ell}_{\theta,c}\|^2\lambda_{\max}(\mathcal{B}_{1,i}) + 4\boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{1,i}| + \ell_{a_{12}}^{(1)} + \boldsymbol{\ell}_{a_{23}}^{(1)\top}\imath_{d,d_x} + (d+2)\max(\ell_{a_{12}}^{(1)}, \boldsymbol{\ell}_{a_{23}}^{(1)\top})\right\}. \quad \text{(B-54)}$$

Eq. (B-52) is implied by

$$\rho_\theta\iota_{d_x} \geq N\max_{c,i}\left\{2u_{i,c}^2\ell_{\phi,c}^2|\mathcal{B}_{1,i}|\iota_{d_x} + \boldsymbol{\ell}_{a_{13}}^{(1)} + \boldsymbol{\ell}_{a_{23}}^{(1)} + (d_x+2)\max\left(\boldsymbol{\ell}_{a_{13}}^{(1)}, \boldsymbol{\ell}_{a_{23}}^{(1)}, 2\ell_{\phi,c}^2|\mathcal{B}_{1,i}\right)\right\}. \quad \text{(B-55)}$$

□

Lemma 9. *The function*

$$\mathcal{H}_{2,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i\neq j}^{N}\frac{1}{2}\left(\rho_u\frac{(u_{i,c}^2 + u_{j,c}^2)}{N-1} + \rho_\phi\frac{\phi_c^2}{N(N-1)} + \rho_\theta\frac{\boldsymbol{\theta}_c^\top\boldsymbol{\theta}_c}{N(N-1)}\right) - \mathcal{E}_{2,N,T}(\boldsymbol{\gamma}, \boldsymbol{U})$$

is convex for every $\boldsymbol{\rho} \doteq (\rho_u, \rho_\phi, \rho_\theta)$ *satisfying* (B-71)-(B-75).

Proof. Write $\mathcal{H}_{2,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{N^2}\sum_{c=1}^{G}\sum_{i\neq j}^{N} \mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$, where $\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{2}\Big(\rho_u \frac{(u_{i,c}^2+u_{j,c}^2)}{N-1} + \rho_\phi \frac{\phi_c^2}{N(N-1)} + \rho_\theta \frac{\boldsymbol{\theta}_c^\top \boldsymbol{\theta}_c}{N(N-1)}\Big) - u_{i,c}u_{j,c}\phi_c^2 \mathcal{A}_{2,i,j} + u_{i,c}u_{j,c}\phi_c^2\boldsymbol{\theta}_c^\top \mathcal{B}_{2,i,j} - u_{i,c}u_{j,c}\phi_c^2\boldsymbol{\theta}_c^\top \mathcal{C}_{2,i,j}\boldsymbol{\theta}_c$. One then needs to prove that $\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$ is a convex function for some $\boldsymbol{\rho}$ sufficiently large. The second-order partial derivatives of $\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$ can be derived in a straight-forward manner.

$$a_{1,1}^{(2)} \doteq D^2_{u_{i,c}}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_u}{N-1}, \tag{B-56}$$

$$a_{1,2}^{(2)} \doteq D^2_{u_{i,c}u_{j,c}}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \phi_c^2\left(\boldsymbol{\theta}_c^\top\mathcal{B}_{2,i,j} - \mathcal{A}_{2,i,j} - \boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(= a_{2,1}^{(2)}), \tag{B-57}$$

$$a_{1,3}^{(2)} \doteq D^2_{u_{i,c}\phi_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = 2u_{j,c}\phi_c\left(\boldsymbol{\theta}_c^\top\mathcal{B}_{2,i,j} - \mathcal{A}_{2,i,j} - \boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c(= a_{3,1}^{(2)})\right), \tag{B-58}$$

$$\boldsymbol{a}_{1,4}^{(2)} \doteq D^2_{u_{i,c}\boldsymbol{\theta}_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{j,c}\phi_c^2(\mathcal{B}_{2,i,j} - 2\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c)(= \boldsymbol{a}_{4,1}^{(2)}), \tag{B-59}$$

$$a_{2,2}^{(2)} \doteq D^2_{u_{j,c}}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_u}{N-1}, \tag{B-60}$$

$$a_{2,3}^{(2)} \doteq D^2_{u_{j,c}\phi_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = 2u_{i,c}\phi_c\left(\boldsymbol{\theta}_c^\top\mathcal{B}_{2,i,j} - \mathcal{A}_{2,i,j} - \boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(= a_{3,2}^{(2)}), \tag{B-61}$$

$$\boldsymbol{a}_{2,4}^{(2)} \doteq D^2_{u_{j,c}\boldsymbol{\theta}_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{i,c}\phi_c^2\left(\mathcal{B}_{2,i,j} - 2\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(= \boldsymbol{a}_{4,2}^{(2)}), \tag{B-62}$$

$$a_{3,3}^{(2)} \doteq D^2_{\phi_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_\phi}{N(N-1)}, \tag{B-63}$$

$$\boldsymbol{a}_{3,4}^{(2)} \doteq D^2_{\phi_c\boldsymbol{\theta}_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = 2u_{i,c}u_{j,c}\phi_c\left(\mathcal{B}_{2,i,j} - 2\mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(= \boldsymbol{a}_{4,3}^{(2)}), \tag{B-64}$$

$$\boldsymbol{a}_{4,4}^{(2)} \doteq D^2_{\boldsymbol{\theta}_c}\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_\theta}{N(N-1)}\mathbb{I}_{d_x} - 2u_{i,c}u_{j,c}\phi_c^2\mathcal{C}_{2,i,j}. \tag{B-65}$$

Let $\mathcal{H}_2 = \begin{pmatrix} a_{1,1}^{(2)} & a_{1,2}^{(2)} & a_{1,3}^{(2)} & \boldsymbol{a}_{1,4}^{(2)\top} \\ a_{2,1}^{(2)} & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \boldsymbol{a}_{2,3}^{(2)\top} \\ a_{3,1}^{(2)} & a_{3,2}^{(2)} & a_{3,3}^{(2)} & \boldsymbol{a}_{3,4}^{(2)\top} \\ \boldsymbol{a}_{4,1}^{(2)} & \boldsymbol{a}_{4,2}^{(2)} & \boldsymbol{a}_{4,3}^{(2)} & \boldsymbol{a}_{4,4}^{(2)} \end{pmatrix}$ be the Hessian matrix of $\mathfrak{f}_{2,N,T}^{(c,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$. Similar to the proof of Lemma 8, all the sub-matrices of the Hessian matrix $\mathcal{H}_2$ are positively definite if the following conditions hold:

$$a_{1,1}^{(2)}a_{2,2}^{(2)} - a_{1,2}^{(2)}a_{2,1}^{(2)} > 0; \tag{B-66}$$

$$\frac{1}{d+3}(a_{1,1}^{(2)} + a_{1,2}^{(2)} + a_{1,3}^{(3)} + \boldsymbol{a}_{1,4}^{(2)\top}\imath_{d,d_x}) > \max\left(0, a_{1,2}^{(2)}, a_{1,3}^{(2)}, \boldsymbol{a}_{1,4}^{(2)\top}\right) \tag{B-67}$$

for every $d = 0, \ldots, d_x$;

$$\frac{1}{d+3}(a_{2,1}^{(2)} + a_{2,2}^{(2)} + a_{2,3}^{(2)} + \boldsymbol{a}_{2,4}^{(2)\top}\imath_{d,d_x}) > \max\left(0, a_{2,1}^{(2)}, a_{2,3}^{(2)}, \boldsymbol{a}_{2,4}^{(2)\top}\right) \tag{B-68}$$

for every $d = 0, \ldots, d_x$;

$$\frac{1}{d+3}(a_{3,1}^{(2)} + a_{3,2}^{(2)} + a_{3,3}^{(2)} + \boldsymbol{a}_{3,4}^{(2)\top}\imath_{d,d_x}) > \max\left(0, a_{3,1}^{(2)}, a_{3,2}^{(2)}, \boldsymbol{a}_{3,4}^{(2)\top}\right) \tag{B-69}$$

for every $d = 0, \ldots, d_x$; and

$$\frac{1}{d+3}\left(\mathbb{S}_{d,d_x}\boldsymbol{a}_{4,1}^{(2)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{4,2}^{(2)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{4,3}^{(2)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{4,4}^{(2)}\imath_{d,d_x}\right) > \max\left(0, \boldsymbol{a}_{4,1}^{(2)}, \boldsymbol{a}_{4,2}^{(2)}, \boldsymbol{a}_{4,3}^{(2)}, \boldsymbol{a}_{4,4}^{*(2)}\right) \tag{B-70}$$

for every $d = 0, \ldots, d_x$, where $\boldsymbol{a}_{4,4}^{*(2)}$ is the matrix with zero diagonal elements and the off-diagonal elements are the same as those of $\boldsymbol{a}_{4,4}^{(2)}$; and $\underbrace{\mathbb{S}_{d,d_x}}_{d\times d_x} = \begin{pmatrix} 1\,0\,0\,0 & \underbrace{\cdots}_{d_x-5} & 0 \\ 0\,1\,0\,0 & \cdots & 0 \\ 0\,0\,1\,0 & \cdots & 0 \\ \vdots\,\vdots\,\vdots\,\vdots & \ddots & \vdots \\ 0\,0\,0\,0 & \cdots & 1 \end{pmatrix}$ is the selection matrix.

The relation (B-66) is valid when

$$\rho_u \geq (N-1)\max_{c,i,j}\left\{\ell_{\phi,c}^2|\mathcal{A}_{2,i,j}| + \ell_{\phi,c}^2|\boldsymbol{\ell}_{\theta,c}^\top|\mathcal{B}_{2,i,j}| + \ell_{\phi,c}^2\boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c}\right\}. \tag{B-71}$$

Moreover, notice that

$$\begin{aligned}
|a_{1,2}^{(2)}| &\leq \ell_{a_{12}}^{(2)} \doteq \max_{c,i,j}\left\{\ell_{\phi,c}^2\left(\mathcal{A}_{2,i,j} + |\boldsymbol{\ell}_{\theta,c}^\top||\mathcal{B}_{2,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}|\boldsymbol{\ell}_{\theta,c}\right)\right\};\\
|a_{1,3}^{(2)}| &\leq \ell_{a_{13}}^{(2)} \doteq 2\max_{c,i,j}\left\{\ell_{\phi,c}\left(\mathcal{A}_{2,i,j} + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{B}_{2,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c}\right)\right\};\\
|a_{2,3}^{(2)}| &\leq \ell_{a_{23}}^{(2)} \doteq 2\max_{c,i,j}\left\{\ell_{\phi,c}\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{B}_{2,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c}\right)\right\};\\
|\boldsymbol{a}_{1,4}^{(2)}| &\leq \boldsymbol{\ell}_{a_{14}}^{(2)} \doteq \max_{c,i,j}\{\ell_{\phi,c}^2(|\mathcal{A}_{2,i,j}| + 2|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c})\};\\
|\boldsymbol{a}_{2,4}^{(2)}| &\leq \boldsymbol{\ell}_{a_{24}}^{(2)} = \boldsymbol{\ell}_{a_{14}}^{(2)};\\
|\boldsymbol{a}_{3,4}^{(2)}| &\leq \boldsymbol{\ell}_{a_{34}}^{(2)} \doteq 2\max_{c,i,j}\left\{\ell_{\phi,c}\left(|\mathcal{B}_{2,i,j}| + 2|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c}\right)\right\};\\
|\boldsymbol{a}_{4,4}^{*(2)}| &\leq 2\max_{c,i,j}\{\ell_{\phi,c}^2|\mathcal{C}_{2,i,j}^*|\} \text{ with } \mathcal{C}_{2,i,j}^* \text{ is } \mathcal{C}_{2,i,j} \text{ except for zero diagonal elements.}
\end{aligned}$$

Therefore, it follows that (B-67) holds if

$$\rho_u \geq (N-1)\left\{\ell_{a_{12}}^{(2)} + \ell_{a_{13}}^{(2)} + \boldsymbol{\ell}_{a_{14}}^{(2)}\imath_{d,d_x} + (d+3)\max(\ell_{a_{12}}^{(2)}, \ell_{a_{13}}^{(2)}, \boldsymbol{\ell}_{a_{14}}^{(2)\top})\right\}; \tag{B-72}$$

(B-68) holds if

$$\rho_u \geq (N-1)\left\{\ell_{a_{12}}^{(2)} + \ell_{a_{23}}^{(2)} + \boldsymbol{\ell}_{a_{14}}^{(2)\top}\imath_{d,d_x} + (d+3)\max\left(\ell_{a_{21}}^{(2)}, \ell_{a_{23}}^{(2)}, \boldsymbol{\ell}_{a_{14}}^{(2)\top}\right)\right\}; \tag{B-73}$$

(B-69) holds if

$$\rho_\phi \geq N(N-1)\left\{\ell_{a_{13}}^{(2)} + \ell_{a_{23}}^{(2)} + \boldsymbol{\ell}_{a_{34}}^{(2)\top}\imath_{d,d_x} + (d+3)\max\left(\ell_{a_{13}}^{(2)}, \ell_{a_{23}}^{(2)}, \boldsymbol{\ell}_{a_{34}}^{(2)\top}\right)\right\}; \tag{B-74}$$

and, finally, (B-70) holds if

$$\rho_\theta \geq N(N-1)\max_{c,i,j} \|2\ell_{\phi,c}^2 \mathbb{S}_{d,d_x}|\mathcal{C}_{2,i,j}|\imath_{d,d_x} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}_{a14}^{(2)} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}_{a24}^{(2)} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}_{a34}^{(2)} + (d+3)\max\left(0, \boldsymbol{\ell}_{a14}^{(2)}, \boldsymbol{\ell}_{a24}^{(2)}, \boldsymbol{\ell}_{a34}^{(2)}, 2\ell_{\phi,c}^2|\mathcal{C}_{2,i,j}^*|\right)\|_\infty, \quad \text{(B-75)}$$

where $\|\cdot\|_\infty$ is the matrix sup-norm. □

Lemma 10. *The function*

$$\mathcal{H}_{3,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{N^2}\sum_{c\neq\ell}^{G}\sum_{i\neq j}^{N}\frac{1}{2}\left(\rho_u\left(\frac{u_{i,c}^2}{(G-1)(N-1)} + \frac{u_{j,\ell}^2}{(G-1)(N-1)}\right) + \rho_\phi\left(\frac{\phi_c^2}{(G-1)N(N-1)} + \frac{\phi_\ell^2}{(G-1)N(N-1)}\right) + \rho_\theta\left(\frac{\boldsymbol{\theta}_c^\top\boldsymbol{\theta}_c}{(G-1)N(N-1)} + \frac{\boldsymbol{\theta}_\ell^\top\boldsymbol{\theta}_\ell}{(G-1)N(N-1)}\right)\right) - \mathcal{E}_{3,N,T}(\boldsymbol{\gamma}, \boldsymbol{U})$$

is convex for every $\boldsymbol{\rho} \doteq (\rho_u, \rho_\phi, \rho_\theta)$ satisfying (B-76), (B-90)-(B-102).

Proof. Write $\mathcal{H}_{3,N,T}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{N^2}\sum_{c\neq\ell}^{G}\sum_{i\neq j}^{N}\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$, where $\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{1}{2}\left\{\rho_u\left(\frac{u_{i,c}^2}{(G-1)(N-1)} + \frac{u_{j,\ell}^2}{(G-1)(N-1)}\right) + \rho_\phi\left(\frac{\phi_c^2}{(G-1)N(N-1)} + \frac{\phi_\ell^2}{(G-1)N(N-1)}\right) + \rho_\theta\left(\frac{\boldsymbol{\theta}_c^\top\boldsymbol{\theta}_c}{(G-1)N(N-1)} + \frac{\boldsymbol{\theta}_\ell^\top\boldsymbol{\theta}_\ell}{(G-1)N(N-1)}\right)\right\} - u_{i,c}u_{j,\ell}\phi_c\phi_\ell\mathcal{A}_{2,i,j} + u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} + u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j} - u_{i,c}u_{j,\ell}\phi_c\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell$. We then need to find conditions to ensure the convexity of $\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U})$. To start with, let's compute the second-order partial derivatives:

$$\begin{aligned}
a_{1,1}^{(3)} &= D_{u_{i,c}}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_u}{(G-1)(N-1)},\\
a_{1,3}^{(3)} &= D_{u_{i,c}\phi_c}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = -u_{j,\ell}\phi_\ell\mathcal{A}_{2,i,j} + u_{j,\ell}\phi_\ell\left(\boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} + \boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j}\right) - u_{j,\ell}\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell (= a_{3,2}^{(3)}),\\
a_{1,4}^{(3)} &= D_{u_{i,c}\phi_\ell}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = -u_{j,\ell}\phi_c\mathcal{A}_{2,i,j} + u_{j,\ell}\phi_c\left(\boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} + \boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j}\right) - u_{j,\ell}\phi_c\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell (= a_{4,1}^{(3)}),\\
\boldsymbol{a}_{1,5}^{(3)} &= D_{u_{i,c}\boldsymbol{\theta}_c}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{j,\ell}\phi_c\phi_\ell\mathcal{C}_{3,i,j} - u_{j,\ell}\phi_c\phi_\ell\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell (= \boldsymbol{a}_{5,1}^{(3)}),\\
\boldsymbol{a}_{1,6}^{(3)} &= D_{u_{i,c}\boldsymbol{\theta}_\ell}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{j,\ell}\phi_c\phi_\ell\left(\mathcal{B}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right) (= \boldsymbol{a}_{6,1}^{(3)}),\\
a_{2,2}^{(3)} &= D_{u_{j,\ell}}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = \frac{\rho_u}{(G-1)(N-1)},\\
a_{1,2}^{(3)} &= D_{u_{j,\ell},u_{i,c}}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = -\phi_c\phi_\ell\left(\mathcal{A}_{2,i,j} - \boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} - \boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j} + \boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\right) (= a_{2,1}^{(3)}),\\
a_{2,3}^{(3)} &= D_{u_{j,\ell},\phi_c}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = -u_{i,c}\phi_\ell\mathcal{A}_{2,i,j} + u_{i,c}\phi_\ell\left(\boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} + \boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j}\right) - u_{i,c}\phi_\ell\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell (= a_{3,2}^{(3)}),\\
a_{2,4}^{(3)} &= D_{u_{j,\ell},\phi_\ell}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = -u_{i,c}\phi_c\mathcal{A}_{2,i,j} + u_{i,c}\phi_c\left(\boldsymbol{\theta}_\ell^\top\mathcal{B}_{3,i,j} + \boldsymbol{\theta}_c^\top\mathcal{C}_{3,i,j}\right) - u_{i,c}\phi_c\boldsymbol{\theta}_c^\top\mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell (= a_{4,2}^{(3)}),\\
\boldsymbol{a}_{2,5}^{(3)} &= D_{u_{j,\ell},\boldsymbol{\theta}_c}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{i,c}\phi_c\phi_\ell\left(\mathcal{C}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\right) (= \boldsymbol{a}_{5,2}^{(3)},\\
\boldsymbol{a}_{2,6}^{(3)} &= D_{u_{j,\ell},\boldsymbol{\theta}_\ell}^2\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma}, \boldsymbol{U}) = u_{i,c}\phi_c\phi_\ell\left(\mathcal{B}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right) (= \boldsymbol{a}_{6,2}^{(3)}),
\end{aligned}$$

$$
\begin{aligned}
a_{3,4}^{(3)} &= D_{\phi_c,\phi_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = -u_{i,c}u_{j,\ell}\left(\mathcal{A}_{2,i,j} - \boldsymbol{\theta}_\ell^\top \mathcal{B}_{3,i,j} - \boldsymbol{\theta}_c^\top \mathcal{C}_{3,i,j} + \boldsymbol{\theta}_c^\top \mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\right), \\
a_{3,3}^{(3)} &= D_{\phi_c}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\phi}{(G-1)N(N-1)}, \\
\boldsymbol{a}_{3,5}^{(3)} &= D_{\phi_c,\boldsymbol{\theta}_c}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = u_{i,c}u_{j,\ell}\phi_\ell\left(\mathcal{C}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\right)(=\boldsymbol{a}_{5,3}^{(3)}), \\
\boldsymbol{a}_{3,6}^{(3)} &= D_{\phi_c,\boldsymbol{\theta}_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = u_{i,c}u_{j,\ell}\phi_\ell\left(\mathcal{B}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(=\boldsymbol{a}_{6,3}^{(3)}), \\
\boldsymbol{a}_{4,5}^{(3)} &= D_{\phi_\ell,\boldsymbol{\theta}_c}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = u_{i,c}u_{j,\ell}\phi_c\left(\mathcal{C}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_\ell\right)(=\boldsymbol{a}_{5,4}^{(3)}), \\
\boldsymbol{a}_{4,4}^{(3)} &= D_{\phi_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\phi}{(G-1)N(N-1)}, \\
\boldsymbol{a}_{4,6}^{(3)} &= D_{\phi_\ell,\boldsymbol{\theta}_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = u_{i,c}u_{j,\ell}\phi_c\left(\mathcal{B}_{3,i,j} - \mathcal{C}_{2,i,j}\boldsymbol{\theta}_c\right)(=\boldsymbol{a}_{6,4}^{(3)}), \\
\boldsymbol{a}_{5,5}^{(3)} &= D_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_c}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\theta}{(G-1)N(N-1)}\mathbb{I}_{d_x}, \\
\boldsymbol{a}_{5,6}^{(3)} &= D_{\boldsymbol{\theta}_c,\boldsymbol{\theta}_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = u_{i,c}u_{j,\ell}\phi_c\phi_\ell\mathcal{C}_{2,i,j}(=\boldsymbol{a}_{6,5}^{(3)}), \\
\boldsymbol{a}_{6,6}^{(3)} &= D_{\boldsymbol{\theta}_\ell,\boldsymbol{\theta}_\ell}^2 \mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U}) = \frac{\rho_\theta}{(G-1)N(N-1)}\mathbb{I}_{d_x}.
\end{aligned}
$$

Let's denote by $\mathcal{H}_3 = \begin{pmatrix} a_{1,1}^{(3)} & a_{1,2}^{(3)} & a_{1,3}^{(3)} & a_{1,4}^{(3)} & \boldsymbol{a}_{1,5}^{(3)\top} & \boldsymbol{a}_{1,6}^{(3)\top} \\ a_{2,1}^{(3)} & a_{2,2}^{(3)} & a_{2,3}^{(3)} & a_{2,4}^{(3)} & \boldsymbol{a}_{2,5}^{(3)\top} & \boldsymbol{a}_{2,6}^{(3)\top} \\ a_{3,1}^{(3)} & a_{3,2}^{(3)} & a_{3,3}^{(3)} & a_{3,4}^{(3)} & \boldsymbol{a}_{3,5}^{(3)\top} & \boldsymbol{a}_{3,6}^{(3)\top} \\ a_{4,1}^{(3)} & a_{4,2}^{(3)} & a_{4,3}^{(3)} & a_{4,4}^{(3)} & \boldsymbol{a}_{4,5}^{(3)\top} & \boldsymbol{a}_{4,6}^{(3)\top} \\ \boldsymbol{a}_{1,5}^{(3)} & \boldsymbol{a}_{2,5}^{(3)} & \boldsymbol{a}_{3,5}^{(3)} & \boldsymbol{a}_{4,5}^{(3)} & \boldsymbol{a}_{5,5}^{(3)} & \boldsymbol{a}_{6,5}^{(3)\top} \\ \boldsymbol{a}_{1,6}^{(3)} & \boldsymbol{a}_{2,6}^{(3)} & \boldsymbol{a}_{3,6}^{(3)} & \boldsymbol{a}_{4,6}^{(3)} & \boldsymbol{a}_{5,6}^{(3)} & \boldsymbol{a}_{6,6}^{(3)} \end{pmatrix}$ the Hessian matrix of $\mathfrak{f}_{3,N,T}^{(c,\ell,i,j)}(\boldsymbol{\gamma},\boldsymbol{U})$. Since $\boldsymbol{\gamma}$ satisfies Assumption 5.1, one obtains that

$$|a_{1,2}^{(3)}| \le \ell_{a_{12}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,c}\ell_{\phi,\ell}\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\};$$

$$|a_{1,3}^{(3)}| \le \ell_{a_{13}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,\ell}\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\};$$

$$|a_{2,3}^{(3)}| \le \ell_{a_{23}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,\ell}\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\};$$

$$|a_{1,4}^{(3)}| \le \ell_{a_{14}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,c}\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\};$$

$$|a_{2,4}^{(3)}| \le \ell_{a_{24}}^{(3)} = \ell_{a_{14}}^{(3)};$$

$$|a_{3,4}^{(3)}| \le \ell_{a_{34}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right\};$$

$$|\boldsymbol{a}_{1,5}^{(3)}| \le \boldsymbol{\ell}_{a_{15}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,c}\ell_{\phi,\ell}\left(|\mathcal{C}_{3,i,j}| + |\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\};$$

$$|\boldsymbol{a}_{1,6}^{(3)}| \le \boldsymbol{\ell}_{a_{16}}^{(3)} \doteq \max_{i,j,c,\ell}\left\{\ell_{\phi,c}\ell_{\phi,\ell}\left(|\mathcal{B}_{3,i,j}| + |\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,c}\right)\right\};$$

$$|\boldsymbol{a}_{2,5}^{(3)}| \le \boldsymbol{\ell}_{a_{25}}^{(3)} = \boldsymbol{\ell}_{a_{15}}^{(3)};\ |\boldsymbol{a}_{2,6}^{(3)}| \le \boldsymbol{\ell}_{a_{26}}^{(3)} = \boldsymbol{\ell}_{a_{16}}^{(3)};\ |\boldsymbol{a}_{3,5}^{(3)}| \le \boldsymbol{\ell}_{a_{35}}^{(3)} = \frac{1}{\ell_{\phi,c}}\boldsymbol{\ell}_{a_{15}}^{(3)}; |\boldsymbol{a}_{3,6}^{(3)}| \le \boldsymbol{\ell}_{a_{36}}^{(3)} = \frac{1}{\ell_{\phi,c}}\boldsymbol{\ell}_{a_{16}}^{(3)};$$

$$|\boldsymbol{a}_{4,5}^{(3)}| \leq \boldsymbol{\ell}_{a_{45}}^{(3)} = \frac{1}{\ell_{\phi,\ell}}\boldsymbol{\ell}_{a_{15}}^{(3)};\ |\boldsymbol{a}_{4,6}^{(3)}| \leq \boldsymbol{\ell}_{a_{46}}^{(3)} = \frac{1}{\ell_{\phi,\ell}}\boldsymbol{\ell}_{a_{16}}^{(3)};\ |\boldsymbol{a}_{5,6}^{(3)}| \leq \boldsymbol{\ell}_{a_{56}}^{(3)} \doteq \max_{i,j,c,\ell}\{\ell_{\phi,c}\ell_{\phi,\ell}|\mathcal{C}_{2,i,j}|\}.$$

Invoking Lemma 20 the minimum eigenvalue of $\mathcal{H}_3$ is positive is implied by the following inequality constraints:

$$\rho_u \geq (G-1)(N-1)\max_{i,j,c,\ell}\left\{\phi_c\phi_\ell\left(|\mathcal{A}_{2,i,j}| + \boldsymbol{\ell}_{\theta,\ell}^\top|\mathcal{B}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{3,i,j}| + \boldsymbol{\ell}_{\theta,c}^\top|\mathcal{C}_{2,i,j}|\boldsymbol{\ell}_{\theta,\ell}\right)\right\}, \tag{B-76}$$

$$\frac{1}{3}(a_{1,1}^{(3)} + a_{1,2}^{(3)} + a_{1,3}^{(3)}) > \max(0, a_{1,2}^{(3)}, a_{1,3}^{(3)}), \tag{B-77}$$

$$\frac{1}{3}(a_{2,1}^{(3)} + a_{2,2}^{(3)} + a_{2,3}^{(3)}) > \max(0, a_{2,1}^{(3)}, a_{2,3}^{(3)}), \tag{B-78}$$

$$\frac{1}{3}(a_{3,1}^{(3)} + a_{3,2}^{(3)} + a_{3,3}^{(3)}) > \max(0, a_{3,1}^{(3)}, a_{3,2}^{(3)}), \tag{B-79}$$

$$\frac{1}{4}(a_{1,1}^{(3)} + a_{1,2}^{(3)} + a_{1,3}^{(3)} + a_{1,4}^{(3)}) > \max(0, a_{1,2}^{(3)}, a_{1,3}^{(3)}, a_{1,4}^{(3)}), \tag{B-80}$$

$$\frac{1}{4}(a_{2,1}^{(3)} + a_{2,2}^{(3)} + a_{2,3}^{(3)} + a_{2,4}^{(3)}) > \max(0, a_{2,1}^{(3)}, a_{2,3}^{(3)}, a_{2,4}^{(3)}), \tag{B-81}$$

$$\frac{1}{4}(a_{1,3}^{(3)} + a_{3,2}^{(3)} + a_{3,3}^{(3)} + a_{3,4}^{(4)}) > \max(0, a_{3,1}^{(3)}, a_{3,2}^{(3)}, a_{3,4}^{(3)}), \tag{B-82}$$

$$\frac{1}{4}(a_{4,1}^{(3)} + a_{4,2}^{(3)} + a_{4,3}^{(3)} + a_{4,4}^{(4)}) > \max(0, a_{4,1}^{(3)}, a_{4,2}^{(3)}, a_{4,3}^{(3)}), \tag{B-83}$$

$$\frac{1}{4+d+e}\left(a_{1,1}^{(3)} + a_{1,2}^{(3)} + a_{1,3}^{(3)} + a_{1,4}^{(3)} + \boldsymbol{a}_{1,5}^{(3)\top}\imath_{d,d_x} + \boldsymbol{a}_{1,6}^{(3)\top}\imath_{e,d_x}\right) > \max(0, a_{1,2}^{(3)}, a_{1,3}^{(3)}, a_{1,4}^{(3)}, \boldsymbol{a}_{1,5}^{(3)\top}, \boldsymbol{a}_{1,6}^{(3)\top}) \tag{B-84}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$,

$$\frac{1}{4+d+e}\left(a_{2,1}^{(3)} + a_{2,2}^{(3)} + a_{2,3}^{(3)} + a_{2,4}^{(3)} + \boldsymbol{a}_{2,5}^{(3)\top}\imath_{d,d_x} + \boldsymbol{a}_{2,6}^{(3)\top}\imath_{e,d_x}\right) > \max\left(0, a_{2,1}^{(3)}, a_{2,3}^{(3)}, a_{2,4}^{(3)}, \boldsymbol{a}_{2,5}^{(3)\top}, \boldsymbol{a}_{2,6}^{(3)\top}\right) \tag{B-85}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$,

$$\frac{1}{4+d+e}\left(a_{3,1}^{(3)} + a_{3,2}^{(3)} + a_{3,3}^{(3)} + a_{3,4}^{(3)} + \boldsymbol{a}_{3,5}^{(3)\top}\imath_{d,d_x} + \boldsymbol{a}_{3,6}^{(3)\top}\imath_{e,d_x}\right) > \max\left(0, a_{3,1}^{(3)}, a_{3,2}^{(3)}, a_{3,4}^{(3)}, \boldsymbol{a}_{3,5}^{(3)\top}, \boldsymbol{a}_{3,6}^{(3)\top}\right) \tag{B-86}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$,

$$\frac{1}{4+d+e}\left(a_{4,1}^{(3)} + a_{4,2}^{(3)} + a_{4,3}^{(3)} + a_{4,4}^{(3)} + \boldsymbol{a}_{4,5}^{(3)\top}\imath_{d,d_x} + \boldsymbol{a}_{4,6}^{(3)\top}\imath_{e,d_x}\right) > \max\left(0, a_{4,1}^{(3)}, a_{4,2}^{(3)}, a_{4,3}^{(3)}, \boldsymbol{a}_{4,5}^{(3)\top}, \boldsymbol{a}_{4,6}^{(3)\top}\right) \tag{B-87}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$,

$$\frac{1}{4+d+3}\left(\mathbb{S}_{d,d_x}\boldsymbol{a}_{1,5}^{(3)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{2,5}^{(3)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{3,5}^{(3)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{4,5}^{(3)} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{5,5}^{(3)}\imath_{d,d_x} + \mathbb{S}_{d,d_x}\boldsymbol{a}_{6,5}^{(3)}\imath_{e,d_x}\right)$$
$$> \max\left(0, \mathbb{S}_{d,d_x}\boldsymbol{a}_{1,5}^{3}, \mathbb{S}_{d,d_x}\boldsymbol{a}_{2,5}^{(3)}, \mathbb{S}_{d,d_x}\boldsymbol{a}_{3,5}^{(3)}, \mathbb{S}_{d,d_x}\boldsymbol{a}_{4,5}^{(3)}, \mathbb{S}_{d,d_x}\boldsymbol{a}_{6,5}^{(3)}\right) \quad \text{(B-88)}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$, and

$$\frac{1}{4+d+e}\left(\mathbb{S}_{e,d_x}\boldsymbol{a}_{1,6}^{(3)} + \mathbb{S}_{e,d_x}\boldsymbol{a}_{2,6}^{(3)} + \mathbb{S}_{e,d_x}\boldsymbol{a}_{3,6}^{(3)} + \mathbb{S}_{e,d_x}\boldsymbol{a}_{4,6}^{(3)\top} + \mathbb{S}_{e,d_x}\boldsymbol{a}_{6,5}^{(3)}\imath_{d,d_x} + \mathbb{S}_{e,d_x}\boldsymbol{a}_{6,6}^{(3)}\imath_{e,d_x}\right)$$
$$> \max\left(0, \mathbb{S}_{e,d_x}\boldsymbol{a}_{1,6}^{(3)}, \mathbb{S}_{e,d_x}\boldsymbol{a}_{2,6}^{(3)}, \mathbb{S}_{e,d_x}\boldsymbol{a}_{3,6}^{(3)}, \mathbb{S}_{e,d_x}\boldsymbol{a}_{4,6}^{(3)}, \mathbb{S}_{e,d_x}\boldsymbol{a}_{6,5}^{(3)}\right) \quad \text{(B-89)}$$

for all $d = 1, \ldots, d_x$ and $e = 1, \ldots, d_x$.

By some simple calculations, Eqs. (B-77)-(B-89) hold if the following conditions hold:

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{13}}^{(3)} + 3\max(0, \ell_{a_{12}}^{(3)}, \ell_{a_{13}}^{(3)})\right); \quad \text{(B-90)}$$

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{23}}^{(3)} + 3\max\left(\ell_{a_{12}}^{(3)}, \ell_{a_{23}}^{(3)}\right)\right); \quad \text{(B-91)}$$

$$\rho_\phi \geq (G-1)N(N-1)\left(\ell_{a_{13}}^{(3)} + \ell_{a_{23}}^{(3)} + 3\max\left(\ell_{a_{13}}^{(3)}, \ell_{a_{23}}^{(3)}\right)\right); \quad \text{(B-92)}$$

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{13}}^{(3)} + 4\max(\ell_{a_{12}}^{(3)}, \ell_{a_{13}}^{(3)}, \ell_{a_{14}}^{(3)})\right); \quad \text{(B-93)}$$

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{23}}^{(3)} + \ell_{a_{24}}^{(3)} + 4\max(\ell_{a_{12}}^{(3)}, \ell_{a_{23}}^{(3)}, \ell_{a_{24}}^{(3)})\right); \quad \text{(B-94)}$$

$$\rho_\phi \geq (G-1)N(N-1)\left(\ell_{a_{13}}^{(3)} + \ell_{a_{23}}^{(3)} + \ell_{a_{34}}^{(3)} + \max\left(\ell_{a_{13}}^{(3)}, \ell_{a_{23}}^{(3)}, \ell_{a_{34}}^{(3)}\right)\right); \quad \text{(B-95)}$$

$$\rho_\phi \geq (G-1)N(N-1)\left(\ell_{a_{14}}^{(3)} + \ell_{a_{24}}^{(3)} + \ell_{a_{34}}^{(3)} + \max\left(\ell_{a_{14}}^{(3)}, \ell_{a_{24}}^{(3)}, \ell_{a_{34}}^{(3)}\right)\right); \quad \text{(B-96)}$$

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{13}}^{(3)} + \ell_{a_{14}}^{(3)} + \boldsymbol{\ell}_{a_{15}}^{(3)\top}\imath_{d,d_x} + \boldsymbol{\ell}_{a_{16}}^{(3)\top}\imath_{e,d_x}\right.$$
$$\left. + (4+d+e)\max\left(\ell_{a_{12}}^{(3)}, \ell_{a_{13}}^{(3)}, \ell_{a_{14}}^{(3)}, \boldsymbol{\ell}_{a_{15}}^{(3)\top}, \boldsymbol{\ell}_{a_{16}}^{(3)\top}\right)\right); \quad \text{(B-97)}$$

$$\rho_u \geq (G-1)(N-1)\left(\ell_{a_{12}}^{(3)} + \ell_{a_{13}}^{(3)} + \ell_{a_{14}}^{(3)} + \boldsymbol{\ell}_{a_{15}}^{(3)\top}\imath_{d,d_x} + \boldsymbol{\ell}_{a_{16}}^{(3)\top}\imath_{e,d_x}\right.$$
$$\left. + (4+d+e)\max\left(\ell_{a_{12}}^{(3)}, \ell_{a_{13}}^{(3)}, \ell_{a_{14}}^{(3)}, \boldsymbol{\ell}_{a_{15}}^{(3)\top}, \boldsymbol{\ell}_{a_{16}}^{(3)\top}\right)\right); \quad \text{(B-98)}$$

$$\rho_\phi \geq (G-1)N(N-1)\left(\ell_{a_{13}}^{(3)} + \ell_{a_{23}}^{(3)} + \ell_{a_{34}}^{(3)} + \boldsymbol{\ell}_{a_{35}}^{(3)\top}\imath_{d,d_x} + \boldsymbol{\ell}_{a_{36}}^{(3)\top}\imath_{e,d_x}\right.$$
$$\left. + (4+d+e)\max\left(\ell_{a_{13}}^{(3)}, \ell_{a_{23}}^{(3)}, \ell_{a_{34}}^{(3)}, \boldsymbol{\ell}_{a_{35}}^{(3)\top}, \boldsymbol{\ell}_{a_{36}}^{(3)\top}\right)\right); \quad \text{(B-99)}$$

$$\rho_\phi \geq (G-1)N(N-1)\left(\ell^{(3)}_{a_{14}} + \ell^{(3)}_{a_{24}} + \ell^{(3)}_{a_{34}} + \boldsymbol{\ell}^{(3)\top}_{a_{45}}\imath_{d,d_x} + \boldsymbol{\ell}^{(3)\top}_{a_{46}}\imath_{e,d_x}\right.$$
$$\left. +(4+d+e)\max\left(\ell^{(3)}_{a_{14}}, \ell^{(3)}_{a_{24}}, \ell^{(3)}_{a_{34}}, \boldsymbol{\ell}^{(3)\top}_{a_{45}}, \boldsymbol{\ell}^{(3)\top}_{a_{46}}\right)\right); \quad \text{(B-100)}$$

$$\rho_\theta \iota_d \geq (G-1)N(N-1)\left(\mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{15}} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{25}} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{35}} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{45}} + \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)\top}_{a_{65}}\imath_{e,d_x}\right.$$
$$\left. +(4+d+e)\max\left(\mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{15}}, \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{25}}, \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{35}}, \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)}_{a_{45}}, \mathbb{S}_{d,d_x}\boldsymbol{\ell}^{(3)\top}_{a_{65}}\right)\right); \quad \text{(B-101)}$$

and

$$\rho_\theta \iota_e \geq (G-1)N(N-1)\left(\mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{16}} + \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{26}} + \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{36}} + \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{46}} + \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)\top}_{a_{65}}\imath_{d,d_x}\right.$$
$$\left. +(4+d+e)\max\left(\mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{16}}, \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{26}}, \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{36}}, \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{46}}, \mathbb{S}_{e,d_x}\boldsymbol{\ell}^{(3)}_{a_{65}}\right)\right). \quad \text{(B-102)}$$

□

Appendix C. Known Results

Lemma 11. *For any fixed $\boldsymbol{a} \in \mathbb{Z}^d$ with $d \geq 1$,*

$$\left|\{\boldsymbol{b} \in \mathbb{Z}^d : \ \|\boldsymbol{a} - \boldsymbol{b}\| = r\}\right| \leq 2d(2r+1)^{d-1}.$$

Proof. See, e.g., Sunklodas (2008). □

Lemma 12. *Suppose that the random field $\{\eta_{\boldsymbol{s}} : \ \boldsymbol{s} \in V_n\}$ is mixing. Let $\mathcal{L}_r(\mathcal{F})$ denote the class of $\mathcal{F}$-measurable random functions, say $f(X)$, satisfying $\|f(X)\|_r = \{E|f(X)|^r\}^{1/r} < \infty$. Let $U = u(\eta_{\boldsymbol{s}}) \in \mathcal{L}(\mathcal{B}(S))$ and $V = v(\eta_{\boldsymbol{s}}) \in \mathcal{L}(\mathcal{B}(S'))$ be measurable functions of $\eta_{\boldsymbol{s}}$. If $\max(\|U\|_r, \|V\|_s) < \infty$ for some $r, s > 2$, one then has, for some $r > 1$ and $1/s + 1/r < 1$,*

$$|Cov(U,V)| \leq C_0\|U\|_r\|V\|_s \left\{M_\alpha(|S|,|S'|)\alpha(d(S,S'))\right\}^{1-1/r-1/s}.$$

In case where $U < C_1$ and $V < C_2$ almost surely, one has

$$Cov(U,v) \leq C_0C_1C_2M_\alpha(|S|,|S'|)\alpha(d(S,S')).$$

Proof. This lemma is a variant of Davydov's inequality (see, e.g., Truong and Stone (1992)). □

Lemma 13. *There exists a value $\tau_0 = \tau_0(\delta) < 1$ such that, for any subset $U \subset \mathbb{Z}^{d_v}$ with $|U| > 1$, one has that*

$$|U_1|^{1+\frac{\delta}{2}} + |U_2|^{1+\frac{\delta}{2}} \leq \tau_0|U|^{1+\frac{\delta}{2}}, \ \textit{where } U = U_1\bigcup U_2 \textit{ and } U_1\bigcap U_2 \neq \emptyset.$$

Proof. See Bulinski and Shashkin (2006). ☐

Lemma 14. *Let $(\xi_1, \dots, \xi_n)$ be a random vector such that $\max_{i=1,\dots,n-1} |E[\prod_{s=i}^n \xi_s]| < \infty$ and $|C_0\xi_i| \leq 1,\ i = 1, \dots, n$. Then, $|E[\prod_{s=1}^n \xi_s] - \prod_{s=1}^n E[\xi_s]| \leq \sum_{i=1}^{n-1} \sum_{j=i+1}^n \Big| E\left[(\xi_i - 1)(\xi_j - 1)\prod_{s=j+1}^n \xi_s\right]$ $-E[\xi_i - 1]E\left[(\xi_j - 1)\prod_{s=j+1}^n \xi_s\right]\Big|$.*

Proof. See Nakhapetyan (1988). ☐

Lemma 15. *Suppose $S_1, S_2, \dots, S_r$ be sets, each containing m sites with $dist(S_i, S_j) = \inf_{\boldsymbol{u}\in S_i, \boldsymbol{v}\in S_j} \|\boldsymbol{u} - \boldsymbol{v}\| \geq \delta$ for all $i \neq j$, where $1 \leq i \leq j$ and $1 \leq j \leq r$. Suppose that $Y_1, Y_2, \dots, Y_r$ be a sequence of real-valued random variables measurable with respect to Borel fields, $\mathcal{B}(S_1), \mathcal{B}(S_2), \dots, \mathcal{B}(S_r)$, respectively and Y_i takes values in $[a, b]$. Then, there exists a sequence of independent random variables, $Y_1^*, Y_2^*, \dots, Y_r^*$, independent from $Y_1, Y_2, \dots, Y_r$ such that Y_i^* has the same distribution as Y_i and satisfies*

$$\sum_{i=1}^r E|Y_i - Y_i^*| \leq 2r(b - a)M_\alpha((r - 1)m, m)\alpha(\delta).$$

Proof. The proof based on Rio (1995) can be found in Carbon, Tran, and Wu (1997). ☐

Lemma 16 (CLT for Double Arrays of Martingale Difference Sequences (M.D.S.))**.** *Let $u_{n,t}$ be a double arrays of m.d.s. with respect to some sequence, $\mathcal{F}_{n,t}$, $t = 1, \dots, T$, of σ-fields such that $E[u_{n,t}|\mathcal{F}_t] = 0$, and let $\boldsymbol{z}_{n,t}$ be a sequence of G-dimensional random vectors measurable with respect to $\mathcal{F}_{n,t}$. Suppose that (i) $\lim_{n,T\uparrow\infty} \sum_{t=1}^T \boldsymbol{z}_{n,t}\boldsymbol{z}_{n,t}^\top \xrightarrow{p} \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is possibly a stochastic matrix, and (ii) $\lim_{n,T\uparrow\infty} \sum_{t=1}^T E\left[\|\boldsymbol{z}_{n,t}\|^{2+\delta}\right] < \infty$ for some $\delta > 0$. Then,*

$$\sum_{t=1}^T \boldsymbol{z}_{n,t}u_{n,t} \xrightarrow{w} \sigma_u\boldsymbol{\eta}^{1/2}N(\mathbf{0}, \mathbb{I}_G),$$

where $\sigma_u^2 = \lim_{n\uparrow\infty} E[u_{n,t}^2|\mathcal{F}_{n,t}]$, and $\boldsymbol{\eta}$ and $N(\mathbf{0}, \mathbb{I}_G)$ are independent.

Proof. See Rao (1987, p. 50). ☐

Lemma 17. *If a sequence of random variables $\{X_n,\ n \in \mathbb{N}\}$ satisfies $\sum_{n=1}^\infty E|X_n| < \infty$, then $\sum_{n=1}^\infty$ almost surely converges to a random variable $X = O_p(1)$.*

Proof. See Taniguchi, Hirukawa, and Tamaki (2008, Theorem A.2). ☐

Lemma 18. *Let C be a nonempty bounded polyhedral convex set, f be a d.c. function on C, and g be a nonnegative concave function on C. Then, there exists $\gamma_0 \geq 0$ such that, for all $\gamma > \gamma_0$, the following problems have the same optimal value and the same solution set:*

$$(P)\ \inf\{f(x) :\ x \in C,\ g(x) \leq 0\}$$

$$\textit{(P')}\ \inf\{f(x)+\gamma g(x):\ x\in C\}.$$

Proof. See Le Thi Hoai An, Huynh Van Ngai, and Pham Dinh Tao (2012). □

Lemma 19 (Chernoff-type inequality)**.** *Let $X_i,\ i=1,\dots,n$, represent jointly independent centered random variables. Let $S_n=\sum_{i=1}^n X_i$, where $\max_{1\le i\le n}|X_i|\le 1$ and $\max_{1\le i\le n} Var(X_i)\le\sigma^2$. Then,*

$$P\left(|S_n|\ge n\lambda\sigma\right)\le 2\max\left(\exp\left(-n\frac{\lambda^2}{4}\right),\exp\left(-\frac{n\lambda\sigma}{2}\right)\right).$$

Proof. Notice that, by the independence of $X_i,\ i=1,\dots,n$, one immediately has $E[\exp(\theta S_n)]=\prod_{i=1}^n E[\exp(\theta X_i)]$. Using an elementary inequality, $\exp(\theta X_i)\le 1+\theta X_i+\theta^2X_i^2$ for $|\theta|\le 1$, we obtain that $E[\exp(\theta X_i)]\le 1+\theta^2Var(X_i)\le\exp\left(\theta^2Var(X_i)\right)$, thus, $E[\exp(\theta S_n)]\le\exp(n\theta^2\sigma^2)$. Invoking Chernoff's inequality, one can immediately show that

$$P\left(|S_n|\ge n\lambda\sigma\right)\ \le\ 2P\left(S_n\ge n\lambda\sigma\right)\ \le\ 2\exp\left(\min_{0<\theta\le 1}\{-\theta\lambda\sigma+n\theta^2\sigma^2\}\right)\ =\ \begin{cases}2\exp\left(-\frac{n\lambda^2}{4}\right) & \text{if } \lambda\le 2\sigma\\ 2\exp\left(-\frac{n\lambda\sigma}{2}\right) & \text{if } \lambda>2\sigma\end{cases}.$$

□

Lemma 20. *Let $\boldsymbol{A}=(a_{i,j})_{1\le i,j\le n}$ be a matrix satisfying, for $i=1,\dots,n$, $\sum_{k=1}^n a_{i,k}>0$, and $a_{i,j}<\frac{1}{n}\sum_{k=1}^n a_{i,k}$ for every $j\ne i$. Then, $det(\boldsymbol{A})>0$.*

Proof. See Carnicer, Goodman, and Pena (1999, Corollary 4.5). □

Appendix D. Proof of Results in Section 4.1

To start with, we define some notations: $\overline{Q}_{N,T}(\boldsymbol{\Omega})=\frac{1}{T}Q_{N,T}(\boldsymbol{\Omega})$; $\underbrace{\boldsymbol{F}_t(\boldsymbol{\theta})}_{(G(d_x+1)+1)\times 1}=(\boldsymbol{A}_t^\top,\boldsymbol{B}_t(\boldsymbol{\theta})^\top,C_t)^\top$; $\epsilon_{0,*,t}=\frac{1}{N}\sum_{i=1}^N\epsilon_{s_i,t}(\boldsymbol{\Theta}_0)$; $\epsilon_{*,t}(\boldsymbol{\psi}_0)=\epsilon_{0,*,t}-\sum_{s=1}^T\epsilon_{0,*,t}\boldsymbol{w}_{*,s}^\top\left(\sum_{s=1}^T\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\boldsymbol{w}_{*,t}$; Some algebraic manipulations yields

$$\begin{aligned}\overline{Q}_{N,T}(\boldsymbol{\Omega}_0)-\overline{Q}_{N,T}(\boldsymbol{\Omega})=&\frac{1}{2}\left(\frac{\sigma_{\epsilon,0}^2}{\sigma_\epsilon^2}-\log\frac{\sigma_{\epsilon,0}^2}{\sigma_\epsilon^2}-1\right)+\frac{1}{2}\left(\frac{1}{\sigma_\epsilon^2}-\frac{1}{\sigma_{\epsilon,0}^2}\right)\left(\frac{N}{T}\sum_{t=1}^T\epsilon_{0,*,t}^2-\sigma_{\epsilon,0}^2\right)\\&+\frac{1}{\sigma_\epsilon^2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top\boldsymbol{D}_\phi\boldsymbol{D}_g\frac{N}{T}\sum_{t=1}^T\boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)\\&+\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^\top\boldsymbol{D}_\phi\boldsymbol{D}_g\left(\frac{N}{T}\sum_{t=1}^T\boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top\right)\boldsymbol{D}_\phi\boldsymbol{D}_g(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\\&=\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3+\mathcal{T}_4.\qquad\text{(D-1)}\end{aligned}$$

Appendix D..1 Some lemmas

The proof of the main theorems needs the following lemmas.

Lemma 21. *Suppose that, for each $\boldsymbol{i} \in V_N$, $\boldsymbol{x}_{\boldsymbol{i},t}$ is a stationary process. Let Assumptions 2.1, 2.2, and 4.1 hold. Then,*

$$\sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{\theta}_0) \epsilon_{*,t}(\boldsymbol{\psi}_0) = O_p(1).$$

Proof. First of all, note that $\sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0) = \sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{0,*,t}$. Define a $(G(d_x+1)+1)\times 1$ vector, $\boldsymbol{Z}_{N,T,t} = \frac{1}{\sqrt{T}} \left({\boldsymbol{x}_{*,t}^{(1)}}^{\top}, \dots, {\boldsymbol{x}_{*,t}^{(G)}}^{\top}, -\xi_{*,t}(\boldsymbol{\theta}_{0,1}), \dots, -\xi_{*,t}(\boldsymbol{\theta}_{0,G}), -1 \right)^{\top}$, and $u_{0,*,t} = \sqrt{N}\epsilon_{0,*,t}$. Next, one needs to prove that

$$\left\| \sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{0,*,t} - \sum_{t=1}^{T} \boldsymbol{Z}_{N,T,t} u_{0,*,t} \right\| = o_p\left(N^{-1/2}\right). \tag{D-2}$$

Notice that

$$\left| \sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{0,*,t} - \sum_{t=1}^{T} \boldsymbol{Z}_{N,T,t} u_{0,*,t} \right| = \left| \sum_{s=1}^{T} \boldsymbol{Z}_{N,T,s} \boldsymbol{w}_{*,s}^{\top} \left(\sum_{s=1}^{T} \boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top} \right)^{-1} \sum_{t=1}^{T} \boldsymbol{w}_{*,t} u_{0,*,t} \right|.$$

Since $\frac{1}{T}\sum_{s=1}^{T} E\left|\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right| = \frac{1}{T|V_N|^2}\sum_{s=1}^{T}\sum_{\boldsymbol{i},\boldsymbol{j}\in V_N}\left|\boldsymbol{w}_{\boldsymbol{i},s}\boldsymbol{w}_{\boldsymbol{j},s}^{\top}\right| < E[\|\boldsymbol{w}_{\boldsymbol{i},s}\|^2] < \infty$ in view of Assumption 4.1(c), an application of Lemma 17 yields $\left|\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right| = O_{a.s.}(T)$. In addition, by the same argument, one also obtains $\left|\sum_{s=1}^{T}\boldsymbol{Z}_{N,T,s}\boldsymbol{w}_{*,s}^{\top}\right| \leq \left|\sum_{s=1}^{T}\boldsymbol{Z}_{N,T,s}\right| \max_{1\leq t\leq T}\left|\boldsymbol{w}_{*,t}^{\top}\right| = O_p(\sqrt{T})o_p(1) = o_p(\sqrt{T})$ because $E[\boldsymbol{w}_{*,t}] = 0$. Invoking Lemma 5 together with Assumption 4.1 yields $\left|\sum_{t=1}^{T}\boldsymbol{w}_{*,t}u_{0,*,t}\right| = O_p\left(\sqrt{\frac{T}{N}}\right)$. Therefore, we can obtain (D-2).

Assumption 2.2 ensures that, for each $i \in [1, G]$ and $\boldsymbol{j} \in V_{N,i}$, the time series $\xi_{\boldsymbol{j},t}(\boldsymbol{\theta}_{0,i})$ is stationary. Hence, by applying Lemma 17, it is not hard to show that

$$\sum_{t=1}^{T} \boldsymbol{Z}_{N,T,t}\boldsymbol{Z}_{N,T,t}^{\top} \xrightarrow{a.s.} \boldsymbol{Q}_{zz}, \tag{D-3}$$

where the limiting matrix $\boldsymbol{Q}_{zz}$ is non-stochastic. Furthermore, note that every element of the vector $\boldsymbol{Z}_{N,T,t}$ has the $(2+\delta)$-th moment being bounded by $T^{-1-\delta/2}$; for example the k-the element, $x_{k,*,t}^{(1)}$, of $\boldsymbol{x}_{*,t}^{(1)}$ has the $(2+\delta)$-th moment satisfying

$$\frac{1}{|V_{N,1}|^{2+\delta}}E\left|\sum_{\boldsymbol{j}\in V_{N,1}} x_{k,\boldsymbol{j},t}\right|^{2+\delta} \leq \frac{1}{|V_{N,1}|^{2+\delta}}\left\{2^{\delta+1}E\left|\sum_{\boldsymbol{j}\in V_{N,1}}(x_{k,\boldsymbol{j},t}-E[x_{k,\boldsymbol{j},t}])\right|^{2+\delta}\right.$$
$$\left.+2^{\delta+1}|V_{N,1}|^{2+\delta}(E[x_{k,\boldsymbol{j},t}])^{2+\delta}\right\} < \infty,$$

where the last inequality follows because $E\left|\sum_{\boldsymbol{j}\in V_{N,1}}(x_{k,\boldsymbol{j},t}-E[x_{k,\boldsymbol{j},t}])\right|^{2+\delta} \leq C_*|V_{N,1}|^{1+\delta/2}$ by Lemma 3, implying that $E\left|\frac{1}{\sqrt{T}}x_{k,*,t}^{(1)}\right|^{2+\delta} < C_0\frac{1}{T^{1+\delta/2}}$. Therefore, one has

$$E[\|\boldsymbol{Z}_{N,T,t}\|^{2+\delta}] \leq E[\|\boldsymbol{Z}_{N,T,t}\|_1^{2+\delta}] < C_0\frac{1}{T^{1+\delta/2}}. \tag{D-4}$$

The conditions in Lemma 16 hold because of (D-3) and (D-4); it then follows that

$$\sum_{t=1}^{T}\boldsymbol{Z}_{N,T,t}u_{0,*,t} \xrightarrow{d} \sigma_\epsilon N(\boldsymbol{0},\boldsymbol{Q}_{zz}). \tag{D-5}$$

The lemma then follows from (D-4) and (D-5). □

Lemma 22. *Let the assumptions of Lemma 21 hold. Then,*

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top = O_p(1).$$

Proof. We need to show that the block matrices on the diagonal are stochastically bounded as the same argument can also be applied to the other block matrices off the diagonal. Define $\boldsymbol{X}_{*,t} = (\boldsymbol{x}_{*,t}^{(1)\top},\ldots,\boldsymbol{x}_{*,t}^{(G)\top})^\top$. By the same argument as in the proof of Lemma 21, one immediately shows that

$$\left|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\boldsymbol{A}_t^\top - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_{*,t}\boldsymbol{X}_{*,t}^\top\right| = \left|\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_{*,t}\boldsymbol{w}_{*,t}^\top\right)\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\boldsymbol{X}_{*,t}^\top\right)\right|$$
$$= o_p(1).$$

An application of Lemma 17 yields

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_{*,t}\boldsymbol{X}_{*,t}^\top = O_p(1).$$

It then follows that

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\boldsymbol{A}_t^\top = O_p(1).$$

□

Lemma 23. *Under Assumptions 2.1 and 4.3, we have*

$$\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\epsilon_{0,*,t} \xrightarrow{w} \sigma_\epsilon \boldsymbol{K}_g\boldsymbol{\Sigma}_\eta^{1/2}\int_0^1 \boldsymbol{W}_\eta(\tau)dW_\epsilon(\tau),$$

where

$$\boldsymbol{K}_t = diag\left(\frac{1}{\sqrt{g_{*,i}}}\mathbb{I}_{d_x},\ i=1,\ldots,G\right)$$

and

$$\boldsymbol{\Sigma}_\eta = \left\{\boldsymbol{\Sigma}_\eta^{(i,j)}\right\}_{i=1,\ldots,G;j=1,\ldots,G}$$

with

$$\boldsymbol{\Sigma}_\eta^{(i,j)} = plim_{N,T\uparrow\infty}\frac{1}{T\sqrt{|V_{N,i}||V_{N,j}|}}E\left[\boldsymbol{S}_\eta(V_{N,i},T)\boldsymbol{S}_\eta(V_{N,j},T)^\top\right],\ \ where\ \boldsymbol{S}_\eta(V_{N,i},t) = \sum_{s=1}^{t}\sum_{\boldsymbol{j}\in V_{N,i}}\boldsymbol{\eta}_{\boldsymbol{j},s};$$

and $\boldsymbol{W}_\eta(\tau)$ *is a* $G.d_x\times 1$ *vector of Brownian motions with the covariance kernel* $E[\boldsymbol{W}_\eta(\tau)\boldsymbol{W}_\eta(\kappa)^\top] = \min(\tau,\kappa)\mathbb{I}_{G.d_x}$*, which are also independent of* $W_\epsilon(\tau)$.

Proof. We merely need to study the limiting distribution of each term in the random vector $\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\epsilon_{0,*,t}$ as the limiting joint distribution of the random vector per se can be derived by applying the Cramér-Wold device. First, noticing that the i-th term of $\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\epsilon_{0,*,t}$ can be written as

$$\begin{aligned}\mathfrak{A}_{i,N,T} &= \frac{N}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}_{*,t}^{(i)} - \sum_{s=1}^{T}\boldsymbol{x}_{*,s}^{(i)}\boldsymbol{w}_{*,s}^\top\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\boldsymbol{w}_{*,t}\right)\epsilon_{0,*,t} \\ &= \frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(i)}\epsilon_{0,*,t} - \frac{N}{T}\sum_{s=1}^{T}\boldsymbol{x}_{*,s}^{(i)}\boldsymbol{w}_{*,s}^\top\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^\top\right)^{-1}\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\epsilon_{0,*,t} = \mathcal{A}_{i,N,T} + \mathcal{B}_{i,N,T}. \quad \text{(D-6)}\end{aligned}$$

Define $S_\epsilon(V_N,t) = \sum_{s=1}^{t}\sum_{\boldsymbol{j}\in V_N}\epsilon_{\boldsymbol{j},s}$. Invoking Lemma 4 together with the Cramér-Wold device, one obtains that

$$\begin{aligned}\mathcal{A}_{i,N,T} &= \frac{1}{L_{N,i}T}\sum_{t=1}^{T}\boldsymbol{S}_\eta(V_{N,i},t)[S_\epsilon(V_N,t)-S_\epsilon(V_N,t-1)]\\ &\approx \frac{1}{\sqrt{g_{*,i}L_{N,i}NT}}\sum_{t=1}^{T}\int_{t/T}^{\frac{t+1}{T}}\boldsymbol{S}_\eta(V_{N,i},\lfloor T\tau\rfloor)\Delta S_\epsilon(V_N,\lfloor T\tau\rfloor)\\ &\xrightarrow{w} \frac{1}{\sqrt{g_{*,i}}}\sigma_\epsilon \boldsymbol{\Sigma}_\eta^{(i)^{1/2}}\int_0^1 \boldsymbol{W}_\eta^{(i)}(\tau)dW_\epsilon(\tau),\end{aligned}$$

where $\boldsymbol{\Sigma}_\eta^{(i)} = \text{plim}_{N,T\uparrow\infty}\frac{1}{T|V_{N,i}|}E\left[\boldsymbol{S}_\eta(V_{N,i},T)\boldsymbol{S}_\eta(V_{N,i},T)^\top\right] < \infty$, and $\boldsymbol{W}_\eta^{(i)}(\tau)$ is a $d_x\times 1$ vector of Brownian motions with the covariance kernel $E\left[\boldsymbol{W}_\eta^{(i)}(\tau)\boldsymbol{W}_\eta^{(i)}(\kappa)^\top\right] = \min(\tau,\kappa)\mathbb{I}_{d_x}$, which are also independent of $W_\epsilon(\tau)$. To bound $\mathcal{B}_{i,N,T}$, note that

$$\begin{aligned}\sum_{s=1}^{T}\boldsymbol{x}_{*,s}^{(i)}\boldsymbol{w}_{*,s}^\top = \left(\sum_{s=1}^{T}|\boldsymbol{x}_{*,s}^{(i)}|\right)\max_{1\le t\le T}|\boldsymbol{w}_{*,t}|^\top\\ \approx \frac{T^{\frac{3}{2}}}{|V_{N,i}|^{\frac{1}{2}}}\left\{\sum_{s=1}^{T}\int_{\frac{s}{T}}^{\frac{s+1}{T}}\left(\frac{1}{\sqrt{T|V_{N,i}|}}\boldsymbol{S}_\eta(V_{N,i},\lfloor T\tau\rfloor)\right)\frac{1}{T}\right\}\max_{1\le t\le T}|\boldsymbol{w}_{*,t}|^\top,\end{aligned}$$

where $\sum_{s=1}^{T}\int_{\frac{s}{T}}^{\frac{s+1}{T}}\left(\frac{1}{\sqrt{T|V_{N,i}|}}\boldsymbol{S}_\eta(V_{N,i},\lfloor T\tau\rfloor)\right)\frac{1}{T} \xrightarrow{w} \boldsymbol{\Sigma}_\eta^{(i)^{\frac{1}{2}}}\int_0^1 \boldsymbol{W}_\eta^{(i)}(\tau)d\tau$ by Lemma 4 and the continuous mapping theorem; and, for every $t\in[1,T]$, $\boldsymbol{w}_{*,t} = o_p(1)$, which can immediately be shown by employing Lemma 3. One then has

$$\sum_{s=1}^{T}\boldsymbol{x}_{*,s}^{(i)}\boldsymbol{w}_{*,s}^\top = o_p\left(T^{3/2}N^{-1/2}\right).$$

Furthermore, by the same argument as in the proof of Lemma 21, one obtains that

$$\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\epsilon_{0,*,t} = O_p\left(T^{1/2}N^{-1}\right)$$

and

$$\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top = O_p(T).$$

It then follows that $\mathcal{B}_{i,N,T} = o_p(N^{-1/2})$. Therefore, it has been shown that

$$\mathfrak{A}_{i,N,T} \xrightarrow{w} \frac{1}{\sqrt{g_{*,i}}}\sigma_\epsilon\boldsymbol{\Sigma}_\eta^{(i)^{1/2}}\int_0^1 \boldsymbol{W}_\eta^{(i)}(\tau)dW_\epsilon(\tau).$$

□

Lemma 24. *Let* $\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0) = (-\xi_{*,t}(\boldsymbol{\theta}_{0,1}), \ldots, -\xi_{*,t}(\boldsymbol{\theta}_{0,G}), -1)^\top$. *Under Assumptions 2.1 and 4.3, we have*

$$\sqrt{\frac{N}{T}} \sum_{t=1}^{T} (\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top, C_t)^\top \epsilon_{0,*,t} \xrightarrow{w} \sigma_\epsilon N\left(\mathbf{0}, \boldsymbol{Q}_{\xi\xi}\right),$$

where $\boldsymbol{Q}_{\xi\xi} = plim_{N,T\uparrow\infty} \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)^\top$.

Proof. By the same argument used to verify (D-6), one can show that

$$\sqrt{\frac{N}{T}} \left\| \sum_{t=1}^{T} (\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top, C_t)^\top \epsilon_{0,*,t} - \sum_{t=1}^{T} \boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)\epsilon_{0,*,t} \right\| = o_p(N^{-1/2}). \tag{D-7}$$

In the same spirit as Lemma 22, an application of the central limit theorem for martingale difference sequences yields

$$\frac{N}{T} \sum_{t=1}^{T} \boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)\epsilon_{0,*,t} \xrightarrow{d} \sigma_\epsilon N\left(\mathbf{0}, \boldsymbol{Q}_{\xi\xi}\right), \tag{D-8}$$

where $\boldsymbol{Q}_{\xi\xi}$ is a non-stochastic asymptotic variance-covariance matrix with finite elements. The lemma then follows from (D-7) and (D-8). □

Lemma 25. *Under Assumptions 2.1 and 4.3, we have*

$$\frac{N}{T^2} \sum_{t=1}^{T} \boldsymbol{A}_t \boldsymbol{A}_t^\top \xrightarrow{w} (\boldsymbol{g} \otimes \boldsymbol{\iota}_{d_x})(\boldsymbol{g} \otimes \boldsymbol{\iota}_{d_x})^\top \boldsymbol{\Sigma}_\eta \int_0^1 \boldsymbol{W}_\eta(\tau)\boldsymbol{W}_\eta(\tau)^\top d\tau, \tag{D-9}$$

where $\boldsymbol{\iota}_{d_x}$ *is the* $d_x \times 1$ *vector of ones,* $\boldsymbol{\Sigma}_\eta$ *and* $\boldsymbol{W}_\eta(\tau)$ *are as defined in Lemma 23;*

$$\frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^{T} \boldsymbol{A}_t \boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top = O_p(1); \tag{D-10}$$

$$\frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^{T} \boldsymbol{A}_t C_t = O_p(1); \tag{D-11}$$

and

$$\frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top, C_t)^\top (\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top, C_t) \xrightarrow{p} E\left[\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)^\top\right], \tag{D-12}$$

where $\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)$ *is defined in Lemma 24.*

Proof. We shall show (D-9), (D-10) and (D-11) as the proof for (D-12) is pretty similar to Lemma

24. As in Lemma 22, let $\boldsymbol{X}_{*,t} = (\boldsymbol{x}_{*,t}^{(1)\top}, \ldots, \boldsymbol{x}_{*,t}^{(G)\top})^\top$. Write

$$\frac{N}{T^2}\sum_{t=1}^T \boldsymbol{A}_t\boldsymbol{A}_t^\top = \frac{N}{T^2}\sum_{t=1}^T \boldsymbol{X}_{*,t}\boldsymbol{X}_{*,t}^\top - \frac{N}{T^2}\sum_{t=1}^T \boldsymbol{X}_{*,t}\boldsymbol{w}_{*,t}^\top\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{X}_{*,t}^\top = \mathcal{T}_{N,T,1} + \mathcal{T}_{N,T,2}.$$

First, notice that

$$\begin{aligned}
\frac{N}{T^2}\sum_{t=1}^T \boldsymbol{x}_{*,t}^{(i)}\boldsymbol{x}_{*,t}^{(j)\top} &= \frac{N}{T\sqrt{L_{N,i}L_{N,j}}}\sum_{t=1}^T \left\{\frac{1}{\sqrt{TL_{N,i}}}\boldsymbol{S}_\eta(V_{N,i},t)\right\}\left\{\frac{1}{\sqrt{TL_{N,j}}}\boldsymbol{S}_\eta(V_{N,j},t)^\top\right\} \\
&\approx \frac{1}{\sqrt{g_{*,i}g_{*,j}}}\sum_{t=1}^T \int_{\frac{t}{T}}^{\frac{t+1}{T}} \left\{\frac{1}{\sqrt{TL_{N,i}}}\boldsymbol{S}_\eta(V_{N,i},\lfloor T\tau\rfloor)\right\}\left\{\frac{1}{\sqrt{TL_{N,j}}}\boldsymbol{S}_\eta(V_{N,j},\lfloor T\tau\rfloor)^\top\right\}\frac{1}{T} \\
&\xrightarrow{w} \frac{1}{\sqrt{g_{*,i}g_{*,j}}}\boldsymbol{\Sigma}_\eta^{(i)1/2}\boldsymbol{\Sigma}_\eta^{(j)1/2}\int_0^1 \boldsymbol{W}_\eta^{(i)}(\tau)\boldsymbol{W}_\eta^{(j)}(\tau)d\tau,
\end{aligned}$$

where $\boldsymbol{W}_\eta^{(i)}(\tau)$ is a $d_x \times 1$ vector of Brownian motions (as defined during the proof of Lemma 23) and both $\boldsymbol{W}_\eta^{(i)}(\tau)$ and $\boldsymbol{W}_\eta^{(j)}(\tau)$ are possibly correlated. Invoking the Cramér-Wold device, one immediately obtains that

$$\mathcal{T}_{N,T,1} \xrightarrow{w} (\boldsymbol{g}\otimes\boldsymbol{\iota}_{d_x})(\boldsymbol{g}\otimes\boldsymbol{\iota}_{d_x})^\top\boldsymbol{\Sigma}_\eta\int_0^1 \boldsymbol{W}_\eta(\tau)\boldsymbol{W}_\eta(\tau)^\top d\tau. \tag{D-13}$$

Moreover, by the same argument as in the proof of Lemma 23, it follows that $\sum_{t=1}^T \boldsymbol{X}_{*,t}\boldsymbol{w}_{*,t}^\top = o_p\left(T^{3/2}N^{-1/2}\right)$ and $\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top = O_p(T)$. Therefore, one has $\mathcal{T}_{N,T,2} = o_p(1)$ and (D-9) has been verified.

To show (D-10), write

$$\begin{aligned}
\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^T \boldsymbol{A}_t\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top &= \frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^T \boldsymbol{X}_{*,t}\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)^\top \\
&\quad - \frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^T \boldsymbol{X}_{*,t}\boldsymbol{w}_{*,t}^\top\left(\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{w}_{*,t}^\top\right)^{-1}\sum_{t=1}^T \boldsymbol{w}_{*,t}\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)^\top = \mathfrak{T}_{N,T,1} + \mathfrak{T}_{N,T,2}.
\end{aligned}$$

To bound each element of $\mathfrak{T}_{N,T,1}$, an application of Lemmas 3 and 4 immediately yields

$$
\begin{aligned}
\left|\sum_{s=1}^{T} \boldsymbol{x}_{*,s}^{(i)} \boldsymbol{\xi}_{N,T,s}(\boldsymbol{\theta}_0)^\top\right| &\leq \left(\sum_{s=1}^{T} |\boldsymbol{x}_{*,s}^{(i)}|\right) \max_{1\leq t\leq T} |\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)|^\top \\
&\approx \frac{T^{\frac{3}{2}}}{|V_{N,i}|^{\frac{1}{2}}} \left\{\sum_{s=1}^{T} \int_{\frac{s}{T}}^{\frac{s+1}{T}} \left(\frac{1}{\sqrt{T|V_{N,i}|}} \boldsymbol{S}_\eta(V_{N,i}, \lfloor T\tau \rfloor)\right) \frac{1}{T}\right\} \max_{1\leq t\leq T} |\boldsymbol{\xi}_{N,T,t}(\boldsymbol{\theta}_0)|^\top \\
&= O_p\left(\frac{T^{3/2}}{N^{1/2}}\right) O_p(1) \text{ by invoking the continuous mapping theorem.}
\end{aligned}
$$

Therefore, one has that $\mathfrak{T}_{N,T,1} = O_p(1)$, and $\mathfrak{T}_{N,T,2} = o_p(1)$ by the same argument. Similarly, one could prove (D-11). $\square$

Appendix D..2 Proof of Theorem 1

Introduce the following open balls centered at the true parameters: $B(\sigma_{\epsilon,0}^2, \delta_\sigma) = \{\sigma_\epsilon^2 \in \Theta_\sigma : |\sigma_\epsilon^2 - \sigma_{\epsilon,0}^2| < \delta_\sigma\}$, $B_N(\boldsymbol{\theta}_0, \delta_\theta) = \{\boldsymbol{\theta} \in \Theta_\theta : \sqrt{N}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_\theta\}$, $B_N(\boldsymbol{\phi}_0, \delta_\phi) = \{\boldsymbol{\phi} \in \Theta_\phi : \sqrt{N}\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| < \delta_\phi\}$, and $B_N(\mu_{*,0}, \delta_\mu) = \{\mu_* \in \Theta_\mu : \sqrt{N}|\mu_* - \mu_{*,0}| < \delta_\mu\}$, where δ_σ, δ_θ, δ_ϕ, and δ_μ are the radiuses. Let B^c denote the complement of any ball, B, in a parameter space, define

$$
A(\boldsymbol{\psi}_0, \delta_\psi) = \bigcup_{\substack{0<\delta_\theta,\delta_\phi,\delta_\mu<\infty \\ (\delta_\theta^2+\delta_\phi^2+\delta_\mu^2)^{1/2}=\delta_\psi}} B_N^c(\boldsymbol{\theta}_0, \delta_\theta) \times B_N^c(\boldsymbol{\phi}_0, \delta_\phi) \times B_N^c(\mu_{*,0}, \delta_\mu).
$$

It then follows that

$$
\begin{aligned}
&P\left(\widetilde{\sigma}_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma),\ \widetilde{\boldsymbol{\theta}} \in B_N^c(\boldsymbol{\theta}_0, \delta_\theta),\ \widetilde{\boldsymbol{\phi}} \in B_N^c(\boldsymbol{\phi}_0, \delta_\phi),\ \widetilde{\mu}_* \in B_N^c(\mu_{*,0}, \delta_\mu) \text{ for every } 0 < \delta_\sigma, \delta_\theta, \delta_\phi, \delta_\mu < \infty\right) \\
&\qquad \leq P\left(\sup_{\substack{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma) \\ \boldsymbol{\psi} \in A(\boldsymbol{\psi}_0, \delta_\psi)}} \overline{Q}_{N,T}(\boldsymbol{\Omega}) \geq \overline{Q}_{N,T}(\boldsymbol{\Omega}_0)\right).
\end{aligned}
$$

Therefore, one needs to verify that either

$$
\lim_{N,T\uparrow\infty} P\left(\sup_{\substack{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma) \\ \boldsymbol{\psi} \in A(\boldsymbol{\psi}_0, \delta_\psi)}} \overline{Q}_{N,T}(\boldsymbol{\Omega}) \geq \overline{Q}_{N,T}(\boldsymbol{\Omega}_0)\right) = 0
$$

or

$$\lim_{N,T\uparrow\infty} P\left(\inf_{\substack{\sigma_\epsilon^2\in B^c(\sigma_{\epsilon,0}^2,\delta_\sigma)\\ \boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}} \left[\overline{Q}_{N,T}(\boldsymbol{\Omega}_0)-\overline{Q}_{N,T}(\boldsymbol{\Omega})\right]\geq 0\right)=1 \tag{D-14}$$

holds.

We examine the terms defined in (D-1). First, note that $\inf_{\sigma_\epsilon^2\in B^c(\sigma_{\epsilon,0}^2,\delta_\sigma)}\mathcal{T}_1>0$ and $\inf_{\sigma_\epsilon^2\in B^c(\sigma_{\epsilon,0}^2,\delta_\sigma)}\mathcal{T}_2=o_p(1)$ for every $\delta_\sigma>0$ by Assumption 2.1 and the weak law of large numbers. Moreover, since $B_N^c(\boldsymbol{\theta}_0,\delta_\theta)$, $B_N^c(\boldsymbol{\phi}_0,\delta_\phi)$, and $B_N^c(\mu_{*,0},\delta_\mu)$ are compact sets, then, for each triplet, $0<\delta_\theta,\delta_\phi,\delta\mu<\infty$, there exist a vector, $\boldsymbol{\psi}^*=\left(\boldsymbol{\theta}^{*\top},\boldsymbol{\phi}^{*\top},\mu_*^*\right)^\top\in B_N^c(\boldsymbol{\theta}_0,\delta_\theta)\times B_N^c(\boldsymbol{\phi}_0,\delta_\phi)\times B_N^c(\mu_{*,0},\delta_\mu)$, such that $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0+N^{-1/2}\boldsymbol{d}_\theta$, $\boldsymbol{\phi}^*=\boldsymbol{\phi}_0+N^{-1/2}\boldsymbol{d}_\phi$, and $\mu_*^*=\mu_{*,0}+N^{-1/2}d_\mu$ respectively; thus, in view of Lemma 21, one has

$$\begin{aligned}\inf_{\substack{\sigma_\epsilon^2\in B^c(\sigma_{\epsilon,0}^2,\delta_\sigma)\\ \boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}}\mathcal{T}_3 &\geq \frac{1}{\sup_{\sigma_\epsilon^2\in B(\sigma_{\epsilon,0}^2,\delta_\sigma)}\sigma_\epsilon^2}(\boldsymbol{d}_\theta^\top,\boldsymbol{d}_\phi^\top,d_\mu)\inf_{\boldsymbol{\phi}\in B_N(\boldsymbol{\phi}_0,\delta_\phi)}\boldsymbol{D}_\phi\boldsymbol{D}_g\frac{\sqrt{N}}{T}\sum_{t=1}^T\boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)\\ &=O_p(T^{-1/2})=o_p(1),\end{aligned}$$

where $\boldsymbol{d}_\theta$, $\boldsymbol{d}_\phi^\top$, and d_μ do not vary with N because if they do, then, for some arbitrarily small constant, ν, there exist sufficiently large integers, $T_0=T_0(\nu)$ and $N_0=N_0(\nu)$, such that

$$\left|\frac{\sqrt{N_0}}{T_0}\sum_{t=1}^{T_0}\boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)\right|<\nu$$

with probability 1. Since $\delta_\psi=(\delta_\theta^2+\delta_\phi^2+\delta_\mu^2)^{1/2}$ is an arbitrarily positive constant (neither depending on N nor T), it may happen that $\left\|(\boldsymbol{d}_{\theta,N_0}^\top,\boldsymbol{d}_{\phi,N_0}^\top,d_{\mu,N_0})\right\|<\delta_\psi$, where the subscript N_0 emphasizes the dependence on N_0, so that $\boldsymbol{\psi}^*\notin B_N^c(\boldsymbol{\theta}_0,\delta_\theta)\times B_N^c(\boldsymbol{\phi}_0,\delta_\phi)\times B_N^c(\mu_{*,0},\delta_\mu)$; we then have a contradiction.

Finally, an application of the minimum eigenvalue inequality yields

$$\begin{aligned}\inf_{\substack{\sigma_\epsilon^2\in B^c(\sigma_{\epsilon,0}^2,\delta_\sigma)\\ \boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}}\mathcal{T}_4 &\geq \frac{1}{2\sup_{\sigma_\epsilon^2\in B(\sigma_{\epsilon,0}^2,\delta_\sigma)}\sigma_\epsilon^2}\inf_{\boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}\|\sqrt{N}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\boldsymbol{D}_\phi\boldsymbol{D}_g\|^2\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^T\boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top\right)\\ &>\frac{1}{2\sup_{\sigma_\epsilon^2\in B(\sigma_{\epsilon,0}^2,\delta_\sigma)}\sigma_\epsilon^2}\delta_\psi^2\inf_{\boldsymbol{\phi}\in B_N(\boldsymbol{\phi}_0,\delta_\phi)}\lambda_{\min}\left(\boldsymbol{D}_\phi\boldsymbol{D}_g\boldsymbol{D}_g^\top\boldsymbol{D}_\phi^\top\right)\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^T\boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top\right)>0,\end{aligned}$$

where the last inequality follows because of Lemma 22 and Assumption 4.2. Collecting all the above arguments, we have proved (D-14).

Appendix D..3 Proof of Theorem 2

Let $\mathcal{Q}_{N,T}(\boldsymbol{\psi}) = \sigma_\epsilon^2 \frac{\partial \overline{Q}_{N,T}(\boldsymbol{\Omega})}{\partial \boldsymbol{\psi}}$, where $\frac{\partial \overline{Q}_{N,T}(\boldsymbol{\Omega})}{\partial \boldsymbol{\psi}}$ is defined by (3.3)-(3.5). In view of the consistency of $\widehat{\boldsymbol{\Omega}} = (\widetilde{\boldsymbol{\psi}}^\top, \widetilde{\sigma}_\epsilon^2)^\top$ (as established in Theorem 1), by applying the mean-value expansion of $\mathcal{Q}_{N,T}(\boldsymbol{\psi})$ around $\boldsymbol{\psi}_0$, we have

$$\mathbf{0} = \frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} = \frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} + \frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top}(\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0),$$

where $\boldsymbol{\psi}^* = (\boldsymbol{\theta}^{*\top}_{N,T}, \boldsymbol{\phi}^{*\top}_{N,T}, \mu_*^*)^\top$ is a vector of the mean values such that

$$P\left(\boldsymbol{\theta}^* \in B_N(\boldsymbol{\theta}_0, \delta_\theta),\ \boldsymbol{\phi}^* \in B_N(\boldsymbol{\phi}_0, \delta_\phi),\ \text{and } \mu_*^* \in B_N(\mu_{*,0}, \delta_\mu)\right) \approx 1$$

for sufficiently large integers, N and T, where $B_N(\boldsymbol{\theta}_0, \delta_\theta)$, $B_N(\boldsymbol{\phi}_0, \delta_\phi)$, and $B_N(\mu_{*,0}, \delta_\mu)$ are the open balls defined in the proof of Theorem 1. It then follows that

$$\sqrt{NT}(\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) = \left(-\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top}\right)^{-1} \left(\sqrt{\frac{T}{N}}\frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right). \tag{D-15}$$

By the same argument as in in the proof of Lemma 21, one can immediately show that

$$\sqrt{\frac{T}{N}}\frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} = -\boldsymbol{D}_{\phi_0}\boldsymbol{D}_g\sqrt{\frac{N}{T}}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0) \xrightarrow{d} \sigma_\epsilon N(\mathbf{0}, \boldsymbol{D}_{\phi_0}\boldsymbol{D}_g\boldsymbol{Q}_{zz}\boldsymbol{D}_{\phi_0}\boldsymbol{D}_g). \tag{D-16}$$

In addition,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}^*)\boldsymbol{F}_t(\boldsymbol{\theta}^*)^\top &= \frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top + \frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)(\boldsymbol{F}_t(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0))^\top \\ &+ \frac{1}{T}\sum_{t=1}^T (\boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0))\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top + \frac{1}{T}\sum_{t=1}^T (\boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0))(\boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0))^\top,\end{aligned}$$

where $\boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0) = \left(\underbrace{\mathbf{0}}_{Gd_x \times 1}{}^\top, \boldsymbol{A}_t^\top \underbrace{\text{diag}((\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_{0,i}),\ i = 1, \ldots, G)}_{Gd_x \times G}, 0\right)^\top$. Therefore, by Theorem 1 and the same argument used in the proof of Lemma 21, one can show that $\max_{1 \le t \le T} |\boldsymbol{F}(\boldsymbol{\theta}^*) - \boldsymbol{F}_t(\boldsymbol{\theta}_0)| = o_p(N^{-1/2})$ and $\max_{1 \le t \le T} |\boldsymbol{F}_t(\boldsymbol{\theta}_0)| = O_p(1)$. This then implies that

$$\frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}^*)\boldsymbol{F}_t(\boldsymbol{\theta}^*)^\top = \frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top + o_p(N^{-1/2}).$$

Since $\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} = \boldsymbol{D}_g \boldsymbol{D}_{\phi^*} \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}^*)\boldsymbol{F}_t(\boldsymbol{\theta}^*)^\top\right) \boldsymbol{D}_{\phi^*}\boldsymbol{D}_g$, where $\|\boldsymbol{D}_{\phi^*} - \boldsymbol{D}_{\phi_0}\| = o_p(N^{-1/2})$, an application of Lemma 22 yields

$$\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} \xrightarrow{p} \boldsymbol{D}_{\phi_0}\boldsymbol{D}_g \boldsymbol{Q}_{zz}\boldsymbol{D}_{\phi_0}\boldsymbol{D}_g.$$

The theorem then follows by the continuous mapping theorem.

Appendix D..4 Proof of Theorem 3

Define open balls centered at the true parameters, $B(\sigma^2_{\epsilon,0}, \delta_\sigma) = \{\sigma^2_\epsilon \in \Theta_\sigma : |\sigma^2_\epsilon - \sigma_{\epsilon,0}| < \delta_\sigma\}$, $B_T(\boldsymbol{\theta}_0, \delta_\theta) = \{\boldsymbol{\theta} \in \Theta_\theta : \sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_\theta\}$, $B_N(\boldsymbol{\phi}_0, \delta_\phi) = \{\boldsymbol{\phi} \in \Theta_\phi : \sqrt{N}\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\| < \delta_\phi\}$, and $B_N(\mu_{*,0}, \delta_\mu) = \{\mu_* \in \Theta_\mu : \sqrt{N}|\mu_* - \mu_{*,0}| < \delta_\mu\}$, where δ_σ, δ_θ, δ_ϕ, and δ_μ are the radiuses of the respective balls. Let B^c denote the complement of any ball, B, in a parameter space, define

$$A(\boldsymbol{\psi}_0, \delta_\psi) = \bigcup_{\substack{0<\delta_\theta,\delta_\phi,\delta_\mu<\infty \\ (\delta_\theta^2+\delta_\phi^2+\delta_\mu^2)^{1/2}=\delta_\psi}} B^c_T(\boldsymbol{\theta}_0, \delta_\theta) \times B^c_N(\boldsymbol{\phi}_0, \delta_\phi) \times B^c_N(\mu_{*,0}, \delta_\mu).$$

We need to prove along the lines of the proof of Theorem 1 that

$$\lim_{N,T\uparrow\infty} P\left(\inf_{\substack{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma) \\ \boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}} \left[\overline{Q}_{N,T}(\boldsymbol{\Omega}_0) - \overline{Q}_{N,T}(\boldsymbol{\Omega})\right] \geq 0\right) = 1. \tag{D-17}$$

To examine the terms defined in (D-1), $\inf_{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma)} \mathcal{T}_1 > 0$ and $\inf_{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma)} \mathcal{T}_2 = o_p(1)$ as in the proof of Theorem 1. For the third term, notice that

$$\begin{aligned}
\inf_{\substack{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma) \\ \boldsymbol{\psi}\in A(\boldsymbol{\psi}_0,\delta_\psi)}} \mathcal{T}_3 \geq \frac{1}{2\inf_{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma)} \sigma^2_\epsilon} & \left\{ \inf_{\substack{\boldsymbol{\theta}\in B^c_T(\boldsymbol{\theta}_0,\delta_\theta) \\ \boldsymbol{\phi}\in B^c_N(\boldsymbol{\phi}_0,\delta_\phi)}} (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \mathrm{diag}(\boldsymbol{g}\otimes \mathbb{I}_{d_x})\mathrm{diag}(\boldsymbol{\phi}\otimes \mathbb{I}_{d_x})\frac{N}{T}\sum_{t=1}^T \boldsymbol{A}_t \epsilon_{*,t}(\boldsymbol{\phi}_0) \right. \\
& \left. + \inf_{\substack{\boldsymbol{\theta}\in B^c_T(\boldsymbol{\theta}_0,\delta_\theta) \\ \boldsymbol{\phi}\in B^c_N(\boldsymbol{\phi}_0,\delta_\phi)}} (\boldsymbol{\phi}-\boldsymbol{\phi}_0)^\top \mathrm{diag}(\boldsymbol{g})\frac{N}{T}\sum_{t=1}^T \boldsymbol{B}(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\phi}_0) + \inf_{\mu_*\in B^c_N(\mu_{*,0},\delta_\mu)} (\mu_* - \mu_{*,0})\frac{N}{T}\sum_{t=1}^T C_t \epsilon_{*,t}(\boldsymbol{\phi}_0) \right\} \\
& = \frac{1}{2\inf_{\sigma^2_\epsilon \in B^c(\sigma^2_{\epsilon,0},\delta_\sigma)} \sigma^2_\epsilon}(\mathcal{T}_{3,a} + \mathcal{T}_{3,b} + \mathcal{T}_{3,c}).
\end{aligned}$$

Because $B^c_N(\boldsymbol{\theta}_0, \delta_\theta)$, $B^c_N(\boldsymbol{\phi}_0, \delta_\phi)$, and $B^c_N(\mu_{*,0}, \delta_\mu)$ are compact sets, then, for each triplet, $0 < \delta_\theta, \delta_\phi, \delta\mu < \infty$, there exist a vector, $\boldsymbol{\psi}^* = \left(\boldsymbol{\theta}^{*\top}, \boldsymbol{\phi}^{*\top}, \mu^*_*\right)^\top \in B^c_N(\boldsymbol{\theta}_0, \delta_\theta) \times B^c_N(\boldsymbol{\phi}_0, \delta_\phi) \times B^c_N(\mu_{*,0}, \delta_\mu)$, such that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 + T^{-1/2}\boldsymbol{d}_\theta$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0 + N^{-1/2}\boldsymbol{d}_\phi$, and $\mu_*^* = \mu_{*,0} + N^{-1/2}d_\mu$ respectively to satisfy

$$\begin{aligned}
\mathcal{T}_{3,a} &= \boldsymbol{d}_\theta^\top \text{diag}(\boldsymbol{g} \otimes \mathbb{I}_{d_x})\boldsymbol{\phi} \in B_N^c(\boldsymbol{\phi}_0, \delta_\phi)\text{diag}(\boldsymbol{\phi} \otimes \mathbb{I}_{d_x})\frac{N}{T^{3/2}}\sum_{t=1}^T \boldsymbol{A}_t \epsilon_{*,t}(\boldsymbol{\phi}_0) \\
\mathcal{T}_{3,b} &= \boldsymbol{d}_\phi^\top \text{diag}(\boldsymbol{g})\frac{\sqrt{N}}{T}\sum_{t=1}^T \boldsymbol{B}(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\phi}_0) \\
\mathcal{T}_{3,c} &= d_\mu \frac{\sqrt{N}}{T}\sum_{t=1}^T C_t \epsilon_{*,t}(\boldsymbol{\phi}_0).
\end{aligned}$$

As in the proof of Theorem 1, one can argue that $\boldsymbol{d}_\theta$, $\boldsymbol{d}_\phi$, and $d_m u$ do not vary with T and N. Invoking Lemmas 23 and 24, one readily has $\mathcal{T}_{3,a} = O_p(T^{-1/2})$, $\mathcal{T}_{3,b} = O_p(T^{-1/2})$, and $\mathcal{T}_{3,c} = O_p(T^{-1/2})$. Therefore, $\inf_{\substack{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma) \\ \boldsymbol{\psi} \in A(\boldsymbol{\psi}_0, \delta_\psi)}} \mathcal{T}_3 \geq 0$ in probability. Finally, to bound $\mathcal{T}_4$, define the normalization matrix $\boldsymbol{K}_{N,T} = \text{diag}\left(T^{1/2}\mathbb{I}_{G.d_x}, N^{1/2}\mathbb{I}_{G+1}\right)$. By the minimum eigenvalue inequality together with Assumption 4.4, one can obtain that, as N and T become large,

$$\begin{aligned}
\inf_{\substack{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma) \\ \boldsymbol{\psi} \in A(\boldsymbol{\psi}_0, \delta_\psi)}} \mathcal{T}_4 \geq \frac{1}{2\inf_{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma)} \sigma_\epsilon^2} &\left(\boldsymbol{K}_{N,T}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)\right)^\top \boldsymbol{D}_\phi \boldsymbol{D}_g \boldsymbol{K}_{N,T}^{-1} \left(\frac{N}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top\right) \\
\boldsymbol{K}_{N,T}^{-1}\boldsymbol{D}_\phi \boldsymbol{D}_g \boldsymbol{K}_{N,T}(\boldsymbol{\psi} - \boldsymbol{\psi}_0) \geq \frac{1}{2\inf_{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma)} \sigma_\epsilon^2} &\inf_{\boldsymbol{\psi} \in A(\boldsymbol{\psi}_0, \delta_\psi)} \left\| \left(\boldsymbol{K}_{N,T}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)\right)^\top \boldsymbol{D}_\phi \boldsymbol{D}_g \right\|^2 \\
&\lambda_{\min}\left(\boldsymbol{K}_{N,T}^{-1}\left(\frac{N}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{\theta}_0)\boldsymbol{F}_t(\boldsymbol{\theta}_0)^\top\right)\boldsymbol{K}_{N,T}^{-1}\right) \\
&\geq \frac{1}{2\inf_{\sigma_\epsilon^2 \in B^c(\sigma_{\epsilon,0}^2, \delta_\sigma)} \sigma_\epsilon^2}\delta_\psi^2 \inf_{\boldsymbol{\phi} \in B_N^c(\boldsymbol{\phi}_0, \delta_\phi)} \lambda_{\min}\left(\boldsymbol{D}_\phi \boldsymbol{D}_g \boldsymbol{D}_g \boldsymbol{D}_\phi\right)\lambda_{\min}(\boldsymbol{Q}_{zz}) > 0,
\end{aligned}$$

where the stochastic limiting matrix $\boldsymbol{Q}_{zz}$ exists because of Lemma 25. We have verified (D-17).

Appendix D..5 Proof of Theorem 4

By using the same notations as in the proof of Theorem 2, in view of the consistency of $\widehat{\boldsymbol{\Omega}} = (\widetilde{\boldsymbol{\psi}}^\top, \widetilde{\sigma}_\epsilon^2)^\top$ established in Theorem 3, an application of the first-order Taylor expansion of $\mathcal{Q}_{N,T}(\boldsymbol{\psi})$ around $\boldsymbol{\theta}_0$ yields

$$\mathbf{0} = \frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\theta}_T^*, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_T^*$ is some point lying in the ball $B_T(\boldsymbol{\theta}_0, \delta_\theta)$. Thus,

$$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \left(-\frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\theta}_T^*, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right)^{-1} \frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta}}. \tag{D-18}$$

Note that

$$\epsilon_{*,t}(\boldsymbol{\psi}) = \epsilon_{*,t}(\boldsymbol{\psi}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}) \boldsymbol{A}_t + (\boldsymbol{\phi} - \boldsymbol{\phi}_0)^\top \mathrm{diag}(\boldsymbol{g}) \boldsymbol{B}_t(\boldsymbol{\theta}_0) + (\mu_* - \mu_{*,0}) C_t. \tag{D-19}$$

One then obtains, in view of (3.3), that

$$\begin{aligned}
\frac{\partial \mathcal{Q}_{N,T}(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta}} &= -\mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x}) \frac{N}{T} \sum_{t=1}^T \boldsymbol{A}_t \epsilon_{*,t}(\boldsymbol{\psi}_0) \\
&\quad - \mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x}) \left\{ \frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^T \boldsymbol{A}_t \boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top \right\} \mathrm{diag}(\boldsymbol{g}) \sqrt{NT} (\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \\
&\quad - \mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x}) \left\{ \frac{N^{1/2}}{T^{3/2}} \sum_{t=1}^T \boldsymbol{A}_t C_t \right\} \sqrt{NT} (\widetilde{\mu}_* - \mu_{*,0}).
\end{aligned}$$

In addition, from Lemma 25 and Theorem 3, we also have

$$-\frac{1}{T} \frac{\partial^2 \mathcal{Q}_{N,T}(\boldsymbol{\theta}_T^*, \widetilde{\boldsymbol{\phi}}, \widetilde{\mu}_*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x}) \left\{ \frac{N}{T^2} \sum_{t=1}^T \boldsymbol{A}_t \boldsymbol{A}_t^\top \right\} \mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x}) = \mathcal{H}_{N,T}^{(aa)}(\boldsymbol{\phi}_0) + o_p(1).$$

It then follows from (D-18) and (D-19) that, as $\widetilde{\boldsymbol{\phi}}$ is consistent by Theorem 3,

$$\begin{aligned}
&(\mathcal{H}_{N,T}^{(aa)}(\boldsymbol{\phi}_0) + o_p(1)) T (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + (\mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi}_0) + o_p(1)) \sqrt{NT} (\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) + (\mathcal{H}_{N,T}^{(ac)}(\boldsymbol{\phi}_0) + o_p(1)) \sqrt{NT} (\widetilde{\mu}_* - \mu_{*,0}) \\
&\qquad = -\mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x}) \mathrm{diag}(\boldsymbol{\phi}\mathbb{I}_{d_x}) \frac{N}{T} \sum_{t=1}^T \boldsymbol{A}_t \epsilon_{*,t}(\boldsymbol{\psi}_0) + o_p(1), \qquad \text{(D-20)}
\end{aligned}$$

where $\mathcal{H}_{N,T}^{(aa)}(\boldsymbol{\phi}_0) = O_p(1)$, $\mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi}_0) = O_p(1)$, and $\mathcal{H}_{N,T}^{(ac)}(\boldsymbol{\phi}_0) = O_p(1)$ in view of Lemma 25. By the same argument leading to (D-18), one can derive that

$$\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 = \left(-\frac{\partial^2 \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_N^*, \widetilde{\mu}_*)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^\top} \right)^{-1} \frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_0, \widetilde{\mu}_*)}{\partial \boldsymbol{\phi}}, \tag{D-21}$$

where $\boldsymbol{\phi}_N^*$ is lying in an open ball, $B_N(\boldsymbol{\phi}_0, \delta_\phi)$, centered at $\boldsymbol{\phi}_0$ and

$$
\begin{aligned}
&-\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_N^*, \widetilde{\mu}_*)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^\top} = \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})^\top\right\}\mathrm{diag}(\boldsymbol{g}) \\
&= \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{B}_t(\boldsymbol{\theta}_0)\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top\right\}\mathrm{diag}(\boldsymbol{g}) + \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{B}_t(\boldsymbol{\theta}_0)\left(\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{B}_t(\boldsymbol{\theta}_0)\right)^\top\right\}\mathrm{diag}(\boldsymbol{g}) \\
&\qquad + \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{B}_t(\boldsymbol{\theta}_0)\right)\boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top\right\}\mathrm{diag}(\boldsymbol{g}) \\
&\qquad + \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{B}_t(\boldsymbol{\theta}_0)\right)\left(\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{B}_t(\boldsymbol{\theta}_0)\right)^\top\right\}\mathrm{diag}(\boldsymbol{g}) = \mathcal{H}_{N,T}^{(bb)} + \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3.
\end{aligned}
$$

Since $\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{B}_t(\boldsymbol{\theta}_0) = \mathrm{diag}\left((\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0})^\top,\ i = 1, \ldots, G\right)\boldsymbol{A}_t$, using the same argument as in Lemma 23 together with Theorem 3 yields that

$$
\begin{aligned}
\mathcal{J}_1 &= \mathrm{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{B}_t(\boldsymbol{\theta}_0)\boldsymbol{A}_t^\top\right\}\mathrm{diag}\left((\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0}),\ i = 1, \ldots, G\right)\mathrm{diag}(\boldsymbol{g}) = O_p\left(T^{1/2}N^{-1/2}\right)o_p\left(T^{-1/2}\right) \\
&= o_p(N^{-1/2}).
\end{aligned}
$$

Analogously, one also obtains that $\mathcal{J}_2 = o_p(N^{-1/2})$ and

$$
\begin{aligned}
\mathcal{J}_3 &= \mathrm{diag}(\boldsymbol{g})\mathrm{diag}\left((\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0})^\top,\ i = 1, \ldots, G\right)\left\{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}_t\boldsymbol{A}_t^\top\right\}\mathrm{diag}\left((\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0}),\ i = 1, \ldots, G\right)\mathrm{diag}(\boldsymbol{g}) \\
&= o_p(T^{-1})O_p(TN^{-1}) = o_p(N^{-1}).
\end{aligned}
$$

It then follows that

$$
-\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_N^*, \widetilde{\mu}_*)}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^\top} = \mathcal{H}_{N,T}^{(bb)} + o_p(1). \tag{D-22}
$$

In view of (3.4), we have

$$
\frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_0, \widetilde{\mu}_*)}{\partial \boldsymbol{\phi}} = -\mathrm{diag}(\boldsymbol{g})\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})\epsilon_{*,t}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_0, \widetilde{\mu}_*),
$$

where

$$\begin{aligned}
\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})\epsilon_{*,t}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\phi}_0,\widetilde{\mu}_*) &= \boldsymbol{B}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)+\left(\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})-\boldsymbol{B}_t(\boldsymbol{\theta}_0)\right)\epsilon_{*,t}(\boldsymbol{\psi}_0)+\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})\left(\epsilon_{*,t}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\phi}_0,\widetilde{\mu}_*)-\epsilon_{*,t}(\boldsymbol{\psi}_0)\right)\\
&= \boldsymbol{B}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)+\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\boldsymbol{A}_t\epsilon_{*,t}(\boldsymbol{\psi}_0)\\
&\quad+\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})\boldsymbol{A}_t^\top\operatorname{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x})\operatorname{diag}(\boldsymbol{g}\mathbb{I}_{d_x})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+(\widetilde{\mu}_*-\mu_{*,0})\boldsymbol{B}_t(\widetilde{\boldsymbol{\theta}})C_t\\
&= \boldsymbol{B}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)+\boldsymbol{B}_t(\boldsymbol{\theta}_0)\boldsymbol{A}_t^\top\operatorname{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x})\operatorname{diag}(\boldsymbol{g}\mathbb{I}_{d_x})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+\boldsymbol{B}_t(\boldsymbol{\theta}_0)C_t(\widetilde{\mu}_*-\mu_{*,0})\\
&\quad+\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\boldsymbol{A}_t\epsilon_{*,t}(\boldsymbol{\psi}_0)+\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\boldsymbol{A}_tC_t(\widetilde{\mu}_*-\mu_{*,0})\\
&\quad+\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\boldsymbol{A}_t\boldsymbol{A}_t^\top\operatorname{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x})\operatorname{diag}(\boldsymbol{g}\mathbb{I}_{d_x})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).
\end{aligned}$$

By Lemma 23 and Theorem 3, one obtains in view of (D-20) that

$$\begin{aligned}
\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\frac{N^{1/2}}{T^{1/2}}\sum_{t=1}^T\boldsymbol{A}_t\epsilon_{*,t}(\boldsymbol{\psi}_0) &= o_p\left(N^{-1/2}\right),\\
\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\left\{\frac{N^{1/2}}{T^{1/2}}\sum_{t=1}^T\boldsymbol{A}_tC_t\right\}(\widetilde{\mu}_*-\mu_{*,0}) &= O_p\left(N^{-1/2}\right),\\
\operatorname{diag}\left((\widetilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i,0})^\top,\ i=1,\ldots,G\right)\left\{\frac{N^{1/2}}{T^{1/2}}\sum_{t=1}^T\boldsymbol{A}_t\boldsymbol{A}_t^\top\right\}\operatorname{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x})\operatorname{diag}(\boldsymbol{g}\mathbb{I}_{d_x})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) &= O_p\left(N^{-1/2}T^{-1/2}\right).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\sqrt{\frac{T}{N}}\frac{\partial\mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}},\boldsymbol{\phi}_0,\widetilde{\mu}_*)}{\partial\boldsymbol{\phi}} &= -\operatorname{diag}(\boldsymbol{g})\sqrt{\frac{N}{T}}\sum_{t=1}^T\boldsymbol{B}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)\\
&\quad-\operatorname{diag}(\boldsymbol{g})\left(\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^T\boldsymbol{B}_t(\boldsymbol{\theta}_0)\boldsymbol{A}_t^\top\right)\operatorname{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x})\operatorname{diag}(\boldsymbol{g}\mathbb{I}_{d_x})T(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\
&\quad-\operatorname{diag}(\boldsymbol{g})\left\{\frac{1}{T}\sum_{t=1}^T\boldsymbol{B}_t(\boldsymbol{\theta}_0)C_t\right\}\sqrt{NT}(\widetilde{\mu}_*-\mu_{*,0})+o_p(1). \quad\text{(D-23)}
\end{aligned}$$

Therefore, in view of (D-21), we have

$$\begin{aligned}
&(\mathcal{H}_{N,T}^{(bb)}+o_p(1))\sqrt{NT}(\widetilde{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)+\mathcal{H}_{N,T}^{(ab)}(\boldsymbol{\phi}_0)^\top T(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\
&\quad+\mathcal{H}_{N,T}^{(bc)}\sqrt{NT}(\widetilde{\mu}_*-\mu_{*,0}) = -\operatorname{diag}(\boldsymbol{g})\sqrt{\frac{N}{T}}\sum_{t=1}^T\boldsymbol{B}_t(\boldsymbol{\theta}_0)\epsilon_{*,t}(\boldsymbol{\psi}_0)+o_p(1). \quad\text{(D-24)}
\end{aligned}$$

By the same argument leading to (D-21), it can be shown that

$$\widetilde{\mu}_* - \mu_{*,0} = \left(-\frac{\partial^2 \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\phi}}, \mu^*_{*,N})}{\partial \mu_*^2}\right)^{-1} \frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\phi}}, \mu_{*,0})}{\partial \mu_*}, \tag{D-25}$$

where $\mu^*_{*,N}$ is some point in an open ball, $B_N(\mu_{*,0}, \delta_\mu)$, centered at $\mu_{*,0}$, and

$$-\frac{1}{N}\frac{\partial^2 \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\phi}}, \mu^*_{*,N})}{\partial \mu_*^2} = 1 + o_p(1).$$

By the same argument leading to (D-24), one readily has

$$\begin{aligned}\sqrt{\frac{T}{N}}\frac{\partial \mathcal{Q}_{N,T}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\phi}}, \mu_{*,0})}{\partial \mu_*} = -\sqrt{\frac{N}{T}}\sum_{t=1}^{T} C_t \epsilon_{*,t}(\boldsymbol{\psi}_0) - \left(\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^{T} C_t \boldsymbol{A}_t^\top\right) \mathrm{diag}(\widetilde{\boldsymbol{\phi}}\mathbb{I}_{d_x})\mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x})T(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \\ - \left(\frac{1}{T}\sum_{t=1}^{T} C_t \boldsymbol{B}_t(\boldsymbol{\theta}_0)^\top\right) \mathrm{diag}(\boldsymbol{g})\sqrt{NT}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0).\end{aligned}$$

Therefore, from (D-25), one obtains that

$$(\mathcal{H}^{(ac)}_{N,T}(\boldsymbol{\phi}_0)^\top + o_p(1))T(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \mathcal{H}^{(bc)\top}_{N,T}\sqrt{NT}(\widetilde{\boldsymbol{\phi}}-\boldsymbol{\phi}_0) + \sqrt{NT}(\widetilde{\mu}_* - \mu_{*,0}) = -\sqrt{\frac{N}{T}}\sum_{t=1}^{T} C_t \epsilon_{*,t}(\boldsymbol{\psi}_0) + o_p(1). \tag{D-26}$$

Collecting up the terms defined by (D-21), (D-24), and (D-26), we have

$$\begin{aligned}\mathcal{H}_{N,T}(\boldsymbol{\phi}_0)\begin{pmatrix} T(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \sqrt{NT}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \\ \sqrt{NT}(\widetilde{\mu}_* - \mu_{*,0}) \end{pmatrix} \\ = -\mathrm{diag}\left(\mathrm{diag}(\boldsymbol{g}\mathbb{I}_{d_x})\mathrm{diag}(\boldsymbol{\phi}_0\mathbb{I}_{d_x}), \mathrm{diag}(\boldsymbol{g}), 1\right)\sum_{t=1}^{T}\begin{pmatrix} \frac{\sqrt{N}}{T}\boldsymbol{A}_t \\ \frac{1}{\sqrt{T}}\boldsymbol{B}_t \\ \frac{1}{\sqrt{T}}C_t \end{pmatrix}\sqrt{N}\epsilon_{*,t}(\boldsymbol{\psi}_0) = \mathcal{M}_{N,T}(\boldsymbol{\phi}_0),\end{aligned}$$

whence it follows that

$$\begin{pmatrix} T(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \sqrt{NT}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \\ \sqrt{NT}(\widetilde{\mu}_* - \mu_{*,0}) \end{pmatrix} = -\mathcal{H}_{N,T}(\boldsymbol{\phi}_0)^{-1}\mathcal{M}_{N,T}(\boldsymbol{\phi}_0).$$

Invoking Lemmas 23, 24, and 25, one can prove that $\mathcal{M}_{N,T}(\boldsymbol{\phi}_0) \xrightarrow{w} MN\left(\mathbf{0}, \mathcal{H}(\boldsymbol{\phi}_0)\right)$. The main theorem then follows by applying the continuous mapping theorem.

Appendix E. Proof of Results in Section 4.2

We start by defining some common notations that will be used for the rest of this section. Let $\boldsymbol{u}_c = (u_{1,c}, \ldots, u_{N,c})^\top$ be a $N \times 1$ vector of group membership indicators associated with group labelled c; $\xi^{(w)}_{0,*,t}(\boldsymbol{u}_c) \doteq \xi^{(w)}_{*,t}(\boldsymbol{\theta}_{0,c}, \boldsymbol{u}_c) = \frac{1}{N}\sum_{i=1}^N u_{i,c}\xi^{(w)}_{i,t}(\boldsymbol{\theta}_{0,c})$; $\boldsymbol{x}^{(w)}_{*,t}(\boldsymbol{u}_c) = \frac{1}{N}\sum_{i=1}^N u_{i,c}\boldsymbol{x}^{(w)}_{i,t}$.

Appendix E..1 Proof of Theorem 5

An application of Lemma 21 yields

$$\sqrt{\frac{N}{T}}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)\epsilon^{(w)}_{0,*,t} = O_p(1).$$

Therefore, it follows that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\{\epsilon^2_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon^{(w)2}_{0,*,t}\} = &\left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)\\ &\operatorname{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)\frac{1}{T}\sum_{t=1}^T \boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)^\top \operatorname{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)\\ &\left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)^\top + O_p\left((NT)^{-1/2}\right).\end{aligned} \quad \text{(E-1)}$$

Let $\mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi) = \{\boldsymbol{\psi} \in \Theta_\psi : \ H(\boldsymbol{\psi}, \boldsymbol{\psi}_0) < \eta_\psi\}$ represent an open ball centered at $\boldsymbol{\psi}_0$ with radius η_ψ, and $\mathcal{B}(\boldsymbol{U}_0, \eta_u) = \{\boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G\times N} : \ H(\boldsymbol{U}, \boldsymbol{U}_0) < \eta_u\}$ be an open ball centered at $\boldsymbol{U}_0$ with radius η_u. We denote by $\mathcal{B}^c(\boldsymbol{\psi}_0, \eta_\psi)$ and $\mathcal{B}^c(\boldsymbol{U}_0, \eta_u)$ the complements of $\mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)$ and $\mathcal{B}(\boldsymbol{U}_0, \eta_u)$ respectively. Since $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{U}})$ are the minimum values of $\frac{1}{T}\sum_{t=1}^T \epsilon^2_{*,t}(\boldsymbol{\psi}, \boldsymbol{U})$, it then follows that

$$P\left(\widehat{\boldsymbol{\psi}} \in \mathcal{B}^c(\boldsymbol{\psi}_0, \eta_\psi), \widehat{\boldsymbol{U}} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)\right) \le P\left(\inf_{\substack{\widehat{\boldsymbol{\psi}} \in \mathcal{B}^c(\boldsymbol{\psi}_0, \eta_\psi)\\ \widehat{\boldsymbol{U}} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)}} \frac{1}{T}\sum_{t=1}^T\{\epsilon^2_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon^{(w)2}_{0,*,t}\} \le 0\right). \quad \text{(E-2)}$$

In view of (E-1), an application of the eigenvalue inequality yields

$$\frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^{2}(\boldsymbol{\psi},\boldsymbol{U})-\epsilon_{0,*,t}^{(w)2}\}\geq C_0\inf_{H(\boldsymbol{U},\boldsymbol{U}_0)>\eta_u}\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)^{\top}\right)H(\widehat{\boldsymbol{\psi}},\boldsymbol{\psi}_0)^2 + O_p\left((NT)^{-1/2}\right)>0$$

by Assumption 4.5. In view of (E-2), one obtains that $H(\widehat{\boldsymbol{\psi}},\boldsymbol{\psi}_0)<\eta_\psi$ and $H(\widehat{\boldsymbol{U}},\boldsymbol{U}_0)<\eta_u$ w.p.1 for some arbitrarily small constants, η_ψ and η_u.

One can now refine the rates that $H(\widehat{\boldsymbol{\psi}},\boldsymbol{\psi}_0)\xrightarrow{p}0$. First let's define new open ball nested in $\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)$, i.e., $\mathfrak{B}(\boldsymbol{\psi}_0,\eta'_\psi/\sqrt{N})\subset\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)$. Some algebra yields

$$\overline{Q}_{N,T}(\boldsymbol{\psi}_0,\sigma_{\epsilon,0}^2,\boldsymbol{U}_0)-\overline{Q}_{N,T}(\boldsymbol{\psi},\sigma_{\epsilon}^2,\boldsymbol{U})=\frac{1}{2}\left(\frac{\sigma_{\epsilon,0}^2}{\sigma_\epsilon^2}-\log\frac{\sigma_{\epsilon,0}^2}{\sigma_\epsilon^2}-1\right)+\frac{1}{2}\left(\frac{1}{\sigma_\epsilon^2}-\frac{1}{\sigma_{\epsilon,0}^2}\right)\left(\frac{N}{T}\sum_{t=1}^{T}\epsilon_{0,*,t}^2-\sigma_{\epsilon,0}^2\right) + \frac{1}{2\sigma_\epsilon^2}\frac{N}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi},\boldsymbol{U})-\epsilon_{0,*,t}^{(w)2}\}=\mathcal{T}_1+\mathcal{T}_2(N,T)+\mathcal{T}_3(N,T),\quad\text{(E-3)}$$

where $\mathcal{T}_1>0$ for every $|\sigma_\epsilon^2-\sigma_{\epsilon,0}^2|>\eta_\sigma$ with some arbitrarily small $\eta_\sigma>0$, and $\mathcal{T}_2(N,T)=o_p(1)$ by the same argument in Theorem 1. Moreover, by Lemma 21, we have

$$\sqrt{\frac{N}{T}}\sum_{t=1}^{T}\underbrace{\left(\boldsymbol{x}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)}),\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)}),1_t^{(w)}\right)}_{\boldsymbol{F}_t^{(1)}(\boldsymbol{U})^{\top}}\epsilon_{0,*,t}^{(w)}=O_p(1)$$

for every $\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})$; and

$$\Bigg\{\max_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})}\inf_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\frac{N}{T}\sum_{t=1}^{T}\left(\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U}_0,\sigma^{(per)})-\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)})\right)\epsilon_{0,*,t}^{(w)}\right|, \max_{\sigma^{(per)}\in\sigma(\mathcal{P})}\inf_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})}\left|\frac{N}{T}\sum_{t=1}^{T}\left(\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U}_0,\sigma^{(per)})-\boldsymbol{\xi}_{*,t}^{(w)\top}(\boldsymbol{U},\widetilde{\sigma}^{(per)})\right)\epsilon_{0,*,t}^{(w)}\right|\Bigg\}^{+}=o_p(1)$$

for every $\boldsymbol{U} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)$. Therefore, we have that, for every $\boldsymbol{U} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)$ and $\boldsymbol{\psi} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)$,

$$\Bigg\{ \max_{\widetilde{\sigma}^{(per)} \in \sigma(\mathcal{P})} \inf_{\sigma^{(per)} \in \sigma(\mathcal{P})} \Bigg| \Big((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top} \Big)$$

$$\operatorname{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1} \right) \sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0) \epsilon_{0,*,t}^{(w)} \Bigg|,$$

$$\max_{\sigma^{(per)} \in \sigma(\mathcal{P})} \inf_{\widetilde{\sigma}^{(per)} \in \sigma(\mathcal{P})} \Bigg| \Big((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top} \Big)$$

$$\operatorname{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1} \right) \sqrt{\frac{N}{T}} \sum_{t=1}^{T} \boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0) \epsilon_{0,*,t}^{(w)} \Bigg| \Bigg\}^{+} = o_p(1).$$

It then follows from (E-3) that

$$\overline{Q}_{N,T}(\boldsymbol{\psi}_0, \sigma_{\epsilon,0}^2, \boldsymbol{U}_0) - \overline{Q}_{N,T}(\boldsymbol{\psi}, \sigma_\epsilon^2, \boldsymbol{U}) \geq C_0 \inf_{H(\boldsymbol{U}, \boldsymbol{U}_0) < \eta_u} \lambda_{\min} \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{F}_t^{(1)}(\boldsymbol{U}) \boldsymbol{F}_t^{(1)}(\boldsymbol{U})^\top \right) N H(\boldsymbol{\psi}, \boldsymbol{\psi}_0)^2 + o_p(1)$$

for every $\boldsymbol{\psi} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)$ and $\boldsymbol{U} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)$. It then follows that

$$P\left(|\widehat{\sigma}_\epsilon^2 - \sigma_{\epsilon,0}^2| > \eta_\sigma, \widehat{\boldsymbol{\psi}} \in \mathfrak{B}\left(\boldsymbol{\psi}_0, \eta'_\psi / \sqrt{N} \right) \right)$$
$$\leq P\left(\inf_{\boldsymbol{\psi} \in \mathfrak{B}^c\left(\boldsymbol{\psi}_0, \eta'_\psi/\sqrt{N}\right), \boldsymbol{U} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)} \left\{ \overline{Q}_{N,T}(\boldsymbol{\psi}_0, \sigma_{\epsilon,0}^2, \boldsymbol{U}_0) - \overline{Q}_{N,T}(\boldsymbol{\psi}, \sigma_\epsilon^2, \boldsymbol{U}) \right\} \leq 0 \right)$$

because $\inf_{\boldsymbol{\psi} \in \mathfrak{B}^c\left(\boldsymbol{\psi}_0, \eta'_\psi/\sqrt{N}\right), \boldsymbol{U} \in \mathcal{B}(\boldsymbol{U}_0, \eta_u)} \left\{ \overline{Q}_{N,T}(\boldsymbol{\psi}_0, \sigma_{\epsilon,0}^2, \boldsymbol{U}_0) - \overline{Q}_{N,T}(\boldsymbol{\psi}, \sigma_\epsilon^2, \boldsymbol{U}) \right\} \geq \eta_\psi'^2 > 0$. This completes the proof.

Appendix E..2 Proof of Theorem 6

First, note that discrete constraints of the form $\boldsymbol{U} \in \{0,1\}^{G \times N}$ in the combinatorial optimization problem (3.9) are equivalent to a system of d.c. constraints: $\boldsymbol{U} \in [0,1]^{G \times N}$, $g(\boldsymbol{U}) = \sum_{c=1}^{G} \sum_{i=1}^{N} u_{i,c}(1-u_{i,c}) \leq 0$. Clearly, $g(\boldsymbol{U})$ is finitely concave on $\mathbb{R}^{G \times N}$, non-negative on Δ_S^N. It immediately follows that $\Delta_S^N \bigcap \{0,1\}^{G \times N} = \{\boldsymbol{U} \in \Delta_S^N : \ g(U) = 0\} = \{\boldsymbol{U} \in \Delta_S^N : \ g(U) \leq 0\}$. By Lemma 18 the following problems are equivalent:

$$(P_\Delta) \ \inf \left\{ \frac{N}{T} \sum_{t=1}^{T} \epsilon_{*,t}^2 (\boldsymbol{\psi}, \boldsymbol{U}) : \ \boldsymbol{\psi} \in \Theta_\psi \subset \mathbb{R}^{G(d_x+1)+1}, \ \boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G \times N} \right\}$$

$$(P_\gamma)\ \inf\left\{\frac{N}{T}\sum_{t=1}^{T}\epsilon_{*,t}^2\left(\boldsymbol{\psi},\boldsymbol{U}\right)+\gamma g(\boldsymbol{U}):\ \boldsymbol{\psi}\in\Theta_\psi\subset\mathbb{R}^{G(d_x+1)+1},\ \boldsymbol{U}\in\Delta_S^N\right\}$$

for all $\gamma>\gamma_0$, where γ_0 is some positive constant.

For given $\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)$ with $\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)$ being an open ball centered at $\boldsymbol{\psi}_0$ with an arbitrarily small radius, η_ψ, one obtains that

$$\widehat{\boldsymbol{U}}(\boldsymbol{\psi})=\text{argmin}_{\boldsymbol{U}\in\Delta_S^N\bigcap\{0,1\}^{G\times N}}\frac{N}{T}\sum_{t=1}^{T}\epsilon_{*,t}^2\left(\boldsymbol{\psi},\boldsymbol{U}\right)=\text{argmin}_{\boldsymbol{U}\in\Delta_S^N}\left\{\frac{N}{T}\sum_{t=1}^{T}\epsilon_{*,t}^2\left(\boldsymbol{\psi},\boldsymbol{U}\right)+\gamma g(\boldsymbol{U})\right\}.$$

Then, $\widehat{\boldsymbol{U}}$ satisfies the KarushKuhnTucker (KKT) conditions (see, e.g., Bonnans and Shapiro (2000, p. 146)):

$$-\nabla\mathcal{Q}_{N,T}(\boldsymbol{\psi},\widehat{\boldsymbol{U}})\in N_{\Delta_S^N}(\widehat{\boldsymbol{U}}),\tag{E-4}$$

where $\mathcal{Q}_{N,T}(\boldsymbol{\psi},\widehat{\boldsymbol{U}})=\frac{N}{T}\sum_{t=1}^{T}\epsilon_{*,t}^2\left(\boldsymbol{\psi},\boldsymbol{U}\right)$ and $N_{\Delta_S^N}(\widehat{\boldsymbol{U}})$ is the normal cone of Δ_S^N at vec$(\widehat{\boldsymbol{U}})$;

$$\boldsymbol{u}^\top\nabla^2\mathcal{Q}_{N,T}(\boldsymbol{\psi},\widehat{\boldsymbol{U}})\boldsymbol{u}\geq 0\text{ for every }\boldsymbol{u}\in T_{\Delta_S^N}(\widehat{\boldsymbol{U}}),\tag{E-5}$$

where $T_{\Delta_S^N}(\widehat{\boldsymbol{U}})$ is the tangent cone of Δ_S^N at vec$(\widehat{\boldsymbol{U}})$.

Because $\boldsymbol{U}_0$ and $\widehat{\boldsymbol{U}}(\boldsymbol{\psi})$ are binary variables, it follows from (E-4) and (E-5) that

$$\begin{aligned}
&E\left[\sup_{\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)}H\left(\widehat{\boldsymbol{U}}(\boldsymbol{\psi}),\boldsymbol{U}_0\right)\right]=\int_0^2 P\left(\sup_{\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)}H\left(\widehat{\boldsymbol{U}}(\boldsymbol{\psi}),\boldsymbol{U}_0\right)>\tau\right)d\tau\\
\leq\ & C_0P\Bigg(\Big(\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\frac{1}{N}\sum_{i=1}^{N}|\widehat{u}_{i,\sigma^{(per)}(c)}(\boldsymbol{\psi})-u_{0,i,c}|,\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\frac{1}{N}\sum_{i=1}^{N}|\widehat{u}_{i,c}(\boldsymbol{\psi})-u_{0,i,\sigma^{(per)}(c)}|\Big)^+\neq 0\\
&\qquad\text{for every }\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi)\text{ and at least one }c\in[1,G]\Bigg)\\
\leq\ & P\Bigg(\Bigg|\sum_{c=1}^{G}\Bigg\{\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U}_0)\phi_c\frac{1}{N}\sum_{i=1}^{N}(u_{0,i,\sigma^{(per)}(c)}-u_{i,c})\\
&\qquad\underbrace{\left\{(\boldsymbol{\theta}_c-\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})^\top\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})\right\}+\gamma\left(\frac{1}{N}\sum_{i=1}^{N}u_{0,i,\sigma^{(per)}(c)}u_{i,c}-1\right)\Bigg\}\Bigg|>0}_{\text{by (E-4)}}\\
&\qquad\text{for every }\underbrace{\gamma<\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\sum_{c=1}^{G}\phi_c^2\frac{1}{NT}\sum_{t=1}^{T}\left((\boldsymbol{\theta}_c-\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})^\top\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})\right)^2}_{\text{by (E-5)}},\\
&\qquad\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_\psi),\text{ and }\boldsymbol{u}\in\Delta_S^N\Bigg)
\end{aligned}$$

$$+ P\left(\left|\sum_{c=1}^{G}\left\{\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U}_0)\phi_{\sigma^{(per)}(c)}\frac{1}{N}\sum_{i=1}^{N}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\right.\right.\right.$$

$$\underbrace{\left.\left.\left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^{\top}\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\}+\gamma\left(\frac{1}{N}\sum_{i=1}^{N}u_{0,i\sigma^{(per)}(c)}u_{i,c}-1\right)\right\}\right|>0}_{\text{by (E-4)}}$$

$$\text{for every }\gamma<\underbrace{\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\sum_{c=1}^{G}\phi^2_{\sigma^{(per)}(c)}\frac{1}{NT}\left((\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^{\top}\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right)^2}_{\text{by (E-5)}},$$

$$\left.\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_{\psi})\text{ and }\boldsymbol{u}\in\Delta_S^N\right)=\mathcal{T}_{1,N,T}+\mathcal{T}_{2,N,T}. \quad \text{(E-6)}$$

To bound $E\left[\sup_{\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_{\psi})}H\left(\widehat{\boldsymbol{U}}(\boldsymbol{\psi}),\boldsymbol{U}_0\right)\right]$, we shall bound $\mathcal{T}_{1,N,T}$ since $\mathcal{T}_{2,N,T}$ can be bounded in the same manner. As the cardinality of $\sigma(\mathcal{P})$ is finite, we only need to work out the rate of convergence for

$$\mathcal{T}'_{1,N,T}=P\left(\sum_{c=1}^{G}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left\{\frac{1}{T}\sum_{t=1}^{T}\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U}_0)\phi_c\frac{1}{N}\sum_{i=1}^{N}(u_{0,i,c}-u_{i,c})\left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^{\top}\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\}\right.\right.$$

$$\left.+\gamma\left(\frac{1}{N}\sum_{i=1}^{N}u_{0,i,c}u_{i,c}-1\right)\right\}>\epsilon_{\eta}\text{ for every }\gamma<\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\sum_{c=1}^{G}\phi_c^2\frac{1}{NT}\sum_{t=1}^{T}\left((\boldsymbol{\theta}_c-\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})^{\top}\boldsymbol{x}_{i,t}^{(w)}\right.$$

$$\left.\left.-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,\sigma^{(per)}(c)})\right)^2,\boldsymbol{\psi}\in\mathcal{B}(\boldsymbol{\psi}_0,\eta_{\psi})\text{ and }\boldsymbol{u}\in\Delta_S^N\right),$$

where ϵ_{η} is some arbitrarily small positive constant. Notice that $\boldsymbol{U}$ is bounded and $\gamma=O_{a.s.}\left(N^{-1}\right)$ by the strong law of large numbers and the compactness of the parameter spaces. An application of Boole's inequality yields

$$\mathcal{T}'_{1,N,T} \leq P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_c(u_{0,i,c}-u_{i,c}) \left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} - \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\}\right.\right.$$
$$\left.\left.(\mu_{0,*}-\mu_*)1_t^{(w)}\right| > \frac{\epsilon_\eta}{4}\right) + P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} (\phi_{0,c}-\phi_{\sigma^{(per)}(c)})(u_{0,i,c}-u_{i,c})\right.\right.$$
$$\left.\left.\left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} - \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\} \xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c},\boldsymbol{u}_{0,c})\right| > \epsilon_\eta/4\right)$$
$$+ P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)}) \left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top \boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_{0,c})\right.\right.$$
$$\left.\left.\left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} - \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\}\right| > \epsilon_\eta/4\right) + P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)}\right.\right.$$
$$\left.\left.(u_{0,i,c}-u_{i,\sigma^{(per)}(c)}) \ \left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} - \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\} \epsilon_{0,*,t}^{(w)}\right| > \epsilon_\eta/4\right) = \mathcal{T}_{1,N,T}^{(a)}+\mathcal{T}_{1,N,T}^{(b)}+\mathcal{T}_{1,N,T}^{(c)}+\mathcal{T}_{1,N,T}^{(d)}.$$

As T becomes large, using the same argument in Lemma 21, it can be verified that $\boldsymbol{x}_{i,t}^{(w)} = \boldsymbol{x}_{i,t}+o_p(1)$, $\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) = \xi_{i,t}(\boldsymbol{\theta}_{0,c}) + o_p(1)$, and $1_t^{(w)} = 1 + o_p(1)$. Therefore, by Boole's inequality, one has

$$\mathcal{T}_{1,N,T}^{(a)}$$
$$\leq P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,c})(\mu_{*,0}-\mu_*)(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})\boldsymbol{x}_{i,t}\right| > \epsilon_\eta/12\right)$$
$$+P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,c})(\mu_{*,0}-\mu_*)(\xi_{i,t}(\boldsymbol{\theta}_{0,c})-E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})])\right| > \epsilon_\eta/12\right)$$
$$+ P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left|\sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)}\left(u_{0,i,c}-u_{i,c}\right)(\mu_{0,*}-\mu_*)E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\right| > \epsilon_\eta/12\right)$$
$$= \mathcal{T}_{1,N,T}^{(a*)}+\mathcal{T}_{1,N,T}^{(a**)}+\mathcal{T}_{1,N,T}^{(a***)}. \quad \text{(E-7)}$$

Since $\left|\phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,c})(\mu_{*,0}-\mu_*)(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})\right| < \infty$, an application of Lemma 6 yields

$$\max\{\mathcal{T}_{1,N,T}^{(a*)},\mathcal{T}_{1,N,T}^{(a**)}\} < C_0\left\{T^{-C_\alpha} + N^{\gamma_M}\log(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}\right.$$
$$\left. + \max\left(\exp\left(-C_{\epsilon_\eta}NT\right), \exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{\log(T)}\right)\right)\right\},$$

where C_α is a sufficiently large constant, and C_{ϵ_η} is some generic constant depending on ϵ_η. Moreover, notice that

$$\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_{\sigma^{(per)}(c)}\left(u_{0,i,c}-u_{i,c}\right)\left(\mu_{0,*}-\mu_*\right)E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\right|$$
$$\leq \eta_\psi \max_{c,i,t}|E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]| \sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_{\sigma^{(per)}(c)}\left(u_{0,i,c}-u_{i,c}\right)\right|.$$

Choosing $\eta_\psi \leq K_{\eta,1}$ with $K_{\eta,1}=\frac{\epsilon_\eta}{12\lim_{\substack{N\uparrow\infty\\T\uparrow\infty}}|E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]|\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,c})\right|}$, one has $\mathcal{T}_{1,N,T}^{(a***)}=0$ for N and T sufficiently large. From (E-7), we can show that

$$\mathcal{T}_{1,N,T}^{(a)}<C_0\left\{T^{-C_\alpha}+N^{\gamma_M}\log(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}+\max\left(\exp\left(-C_{\epsilon_\eta}NT\right),\exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{\log(T)}\right)\right)\right\}. \quad \text{(E-8)}$$

Since $\xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c},\boldsymbol{u}_{0,c})=g_cE[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]+o_p(1)$ with $g_c=\frac{1}{N}\sum_{i=1}^{N}u_{0,i,c}$, we have

$$\mathcal{T}_{1,N,T}^{(b)}<P\Bigg(\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}g_cE[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\Bigg|\sum_{i=1}^{N}\sum_{t=1}^{T}(\phi_{0,c}-\phi_{\sigma^{(per)}(c)})(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})$$
$$(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top\boldsymbol{x}_{i,t}^{(w)}\Bigg|>\frac{\epsilon_\eta}{12}\Bigg)+P\Bigg(\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}g_cE[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]$$
$$\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}(\phi_{0,c}-\phi_{\sigma^{(per)}(c)})(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\left(\xi_{i,t}(\boldsymbol{\theta}_{0,c})-E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\right)\right|>\frac{\epsilon_\eta}{12}\Bigg)$$
$$+P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}g_cE^2[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}(\phi_{0,c}-\phi_{\sigma^{(per)}(c)})(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\right|>\frac{\epsilon_\eta}{12}\right),$$

where the first two terms can be bounded in the same manner as (E-8) and the last term can be made arbitrarily close to zero by choosing $\eta_\psi<(K_{\eta,1},K_{\eta,2})^-$ with

$$K_{\eta,2}=\frac{\epsilon_\eta}{12\sup_{\boldsymbol{U}\in\Delta_S^N}\frac{1}{NT}\sum_{c=1}^{G}g_cE^2[\xi_{i,t}(\boldsymbol{\theta}_{0,c})]\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\right|}.$$

It then follows that

$$\mathcal{T}_{1,N,T}^{(b)}<C_0\left\{T^{-C_\alpha}+N^{\gamma_M}\log(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}+\max\left(\exp\left(-C_{\epsilon_\eta}NT\right),\exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{\log(T)}\right)\right)\right\}. \quad \text{(E-9)}$$

Using exactly the same argument, one also obtains that

$$\mathcal{T}_{1,N,T}^{(c)} \quad < \quad C_0 \left\{ T^{-C_\alpha} + N^{\gamma_M} \log(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha} + \max\left(\exp\left(-C_{\epsilon_\eta} NT\right), \exp\left(-C_{\epsilon_\eta} \frac{T^{1/4}}{\log(T)}\right)\right)\right\}. \tag{E-10}$$

Now, to bound the last term of $\mathcal{T}'_{1,N,T}$. Notice that

$$\begin{aligned}
\mathcal{T}_{1,N,T}^{(d)} \leq P &\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT} \sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left| \phi_{\sigma^{(per)}(c)} \left(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c}\right)^\top \right| \right. \\
&\qquad \left. \left| \sum_{i=1}^{N}\sum_{t=1}^{T} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \;\; \boldsymbol{x}_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \epsilon_{0,*,t}^{(w)} \right| > \frac{\epsilon_\eta}{8} \right) \\
&+ P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT} \sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left| \sum_{i=1}^{N}\sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \epsilon_{0,*,t}^{(w)} \right| > \frac{\epsilon_\eta}{8} \right) \\
&= \mathcal{T}_{1,N,T}^{(d)*} + \mathcal{T}_{1,N,T}^{(d)**}.
\end{aligned} \tag{E-11}$$

Since $\boldsymbol{x}_{i,t}^{(w)} = \boldsymbol{x}_{i,t} + o_p(1)$ and $\epsilon_{0,*,t}^{(w)} = \epsilon_{0,*,t} + o_p(1)$, it then follows that

$$\begin{aligned}
\mathcal{T}_{1,N,T}^{(d)*} \leq P &\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT} \sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left| \phi_{\sigma^{(per)}(c)} \left(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c}\right)^\top \right| \right. \\
&\left. \left| \sum_{i=1}^{N}\sum_{t=1}^{T} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \;\; \{\boldsymbol{x}_{i,t}(\boldsymbol{\theta}_{0,c})\epsilon_{0,*,t} - E[\boldsymbol{x}_{i,t}(\boldsymbol{\theta}_{0,c})\epsilon_{0,*,t}]\} \right| > \frac{\epsilon_\eta}{16} \right) \\
&+ P\left(\sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT} \sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left| \phi_{\sigma^{(per)}(c)} \left(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c}\right)^\top \right| \right. \\
&\qquad \left. \left| \sum_{i=1}^{N}\sum_{t=1}^{T} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \;\; E\left[\boldsymbol{x}_{i,t}(\boldsymbol{\theta}_{0,c})\epsilon_{0,*,t}\right] \right| > \frac{\epsilon_\eta}{16} \right),
\end{aligned}$$

where the last term can be made arbitrarily close to zero by choosing $\eta_\psi < (K_{\eta,1}, K_{\eta,2}, K_{\eta,3})^-$ with $\eta_\psi < \frac{\epsilon_\eta}{16 \sup_{\boldsymbol{U}\in\Delta_S^N} \frac{1}{NT}\sum_{c=1}^{G} \min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \left\| \phi_{\sigma^{(per)}(c)} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(u_{0,i,c} - u_{i,\sigma^{(per)}(c)}\right) E[\boldsymbol{x}_{i,t}(\boldsymbol{\theta}_{0,c})\epsilon_{0,*,t}] \right\|}$, and the first term can be bounded by invoking Lemma 7. It then follows that

$$\begin{aligned}
\mathcal{T}_{1,N,T}^{(d)*} \leq C_0 &\left\{ T^{-C_\alpha} + N^{2\gamma_M} \log^2(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha} \right. \\
&\left. + \max\left\{ \exp\left(-C_\sigma N^2 \log^2(T) T^{7/4}\right), \exp\left(-C_M \frac{T^{1/4}}{\log^2(T)}\right) \right\} \right\}.
\end{aligned}$$

By exactly the same argument, one can also show that

$$\mathcal{T}_{1,N,T}^{(d)**} \leq C_0 \left\{ T^{-C_\alpha} + N^{2\gamma_M} \log^2(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha} + \max\left\{ \exp\left(-C_\sigma N^2 \log^2(T) T^{7/4}\right), \exp\left(-C_M \frac{T^{1/4}}{\log^2(T)}\right) \right\} \right\}.$$

It then follows from (E-11) that

$$\mathcal{T}_{1,N,T}^{(d)} \leq C_0 \left\{ T^{-C_\alpha} + N^{2\gamma_M} \log^2(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha} + \max\left\{ \exp\left(-C_\sigma N^2 \log^2(T) T^{7/4}\right), \exp\left(-C_M \frac{T^{1/4}}{\log^2(T)}\right) \right\} \right\}. \quad \text{(E-12)}$$

In view of (E-7)-(E-12) the main theorem then follows.

Appendix E..3 Proof of Theorem 7

The proof proceeds in the following three main steps:

STEP 1: It can immediately be verified that $\epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) = (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}})^\top \boldsymbol{D}_\phi \boldsymbol{D}_g \boldsymbol{X}_{N,T,t}(\boldsymbol{\theta}) + \epsilon_{*,t}(\widetilde{\boldsymbol{\psi}}, \boldsymbol{U}_0)$. Since $\widetilde{\boldsymbol{\psi}}$ is the minimum value of $\widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) = \frac{N}{T}\sum_{t=1}^T \epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U}_0)$, it must satisfy the equations:

$$\begin{aligned} \frac{N}{T}\sum_{t=1}^T \xi_{*,t}^{(w)}(\widetilde{\boldsymbol{\theta}}_c, \boldsymbol{u}_c)\epsilon_{*,t}(\widetilde{\boldsymbol{\psi}}, \boldsymbol{U}_0) &= 0, \\ \frac{N}{T}\sum_{t=1}^T \boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_{0,c})\epsilon_{*,t}(\widetilde{\boldsymbol{\psi}}, \boldsymbol{U}_0) &= 0, \\ \frac{N}{T}\sum_{t=1}^T 1_t^{(w)}\epsilon_{*,t}(\widetilde{\boldsymbol{\psi}}, \boldsymbol{U}_0) &= 0. \end{aligned}$$

Therefore, an application of the eigenvalue inequality and Theorem 5 yields

$$\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) \geq \lambda_{\min}\left(\frac{N}{T}\sum_{t=1}^T \boldsymbol{X}_{N,T,t}(\boldsymbol{\theta}_0)\boldsymbol{X}_{N,T,t}(\boldsymbol{\theta}_0)^\top\right) H(\widehat{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\psi}}) > C_0 H(\widehat{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\psi}}), \quad \text{(E-13)}$$

where the last inequality follows from Assumption 4.2.

STEP 2: Let $\widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) = \frac{N}{T}\sum_{t=1}^T \epsilon_{*,t}^2(\boldsymbol{\psi}, \widehat{\boldsymbol{U}})$, where $\widehat{\boldsymbol{U}} \doteq \widehat{\boldsymbol{U}}(\boldsymbol{\psi}) = \operatorname{argmin}_{\boldsymbol{U} \in \Delta_S^N \bigcap \{0,1\}^{G\times N}} \frac{N}{T}\sum_{t=1}^T \epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U})$.

One can show that

$$\widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) - \widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) = \sum_{c=1}^{G} \phi_c \frac{1}{N} \sum_{i=1}^{N} (u_{0,i,c} - \widehat{u}_{i,\sigma^{(per)}(c)}) \frac{N}{T} \sum_{t=1}^{T} \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right)$$
$$+ \sum_{c=1}^{G} \phi_c (\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \frac{1}{N} \sum_{i=1}^{N} (\widehat{u}_{i,\sigma^{(per)}(c)} - u_{0,i,c}) \frac{N}{T} \sum_{t=1}^{T} \boldsymbol{x}_{i,t}^{(w)} \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right).$$

By Hölder's inequality, one obtains that

$$\begin{aligned} \left| \widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) - \widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) \right| \leq & \sum_{c=1}^{G} |\phi_c| \left\{ \frac{1}{N} \sum_{i=1}^{N} (u_{0,i,c} - \widehat{u}_{i,\sigma^{(per)}(c)})^2 \right\}^{\frac{1}{2}} \\ & \left\{ \frac{N}{T^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right) \right)^2 \right\}^{\frac{1}{2}} \\ & + \sum_{c=1}^{G} \left| \phi_c (\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \right| \left\{ \frac{1}{N} \sum_{i=1}^{N} (\widehat{u}_{i,\sigma^{(per)}(c)} - u_{0,i,c})^2 \right\}^{\frac{1}{2}} \\ & \left\{ \frac{N}{T^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \boldsymbol{x}_{i,t}^{(w)} \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right) \right)^2 \right\}^{\frac{1}{2}}. \end{aligned} \quad \text{(E-14)}$$

Because all the clusters are sufficiently large and $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{i,t}\xi_{i,t}(\boldsymbol{\theta}_{0,c}) = O_p(1)$ for every $i \in [1, N]$ and $c \in [1, G]$, one can verify that

$$\frac{1}{NT^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right) \right)^2 = O_p(1)$$

and

$$\frac{1}{NT^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \boldsymbol{x}_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \left(\epsilon_{*,t}(\boldsymbol{\psi}, \widehat{\boldsymbol{U}}) + \epsilon_{*,t}(\boldsymbol{\psi}, \boldsymbol{U}_0) \right) \right)^2 = O_p(1).$$

Since the objective functions are invariant with respect to relabelling of groups, by Theorem 6, we obtain that

$$\begin{aligned} & \sup_{\boldsymbol{\psi} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)} \left| \widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) - \widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) \right| \\ & = O_p \left(NT^{-C_\alpha} + N^{\gamma_M + 1} \log(T) T^{\frac{\gamma_M}{2} - \frac{3}{8}\theta_\alpha} + N \exp\left(-C_M \frac{T^{1/4}}{\log^2(T)} \right) \right). \end{aligned} \quad \text{(E-15)}$$

STEP 3: Notice that

$$\begin{aligned}\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) &= \widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) + \underbrace{\widehat{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}})}_{\leq 0} + \widehat{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) \\ &\leq \left\{\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}})\right\} + \left\{\widehat{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}})\right\}. \quad \text{(E-16)}\end{aligned}$$

Some probability event computations yield that

$$\left\{\left|\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}})\right| > \epsilon\right\} \subset \left\{\widehat{\boldsymbol{\psi}} \notin \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)\right\} \bigcup \left\{\widehat{\boldsymbol{\psi}} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi), \left|\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}})\right| > \epsilon\right\}$$

and

$$\left\{\left|\widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}})\right| > \epsilon\right\} \subset \left\{\widetilde{\boldsymbol{\psi}} \notin \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)\right\} \bigcup \left\{\widetilde{\boldsymbol{\psi}} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi), \left|\widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) - \widehat{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}})\right| > \epsilon\right\}.$$

It then follows from (E-16) that

$$\begin{aligned}P\left(\widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) > \epsilon\right) \leq P\left(\widehat{\boldsymbol{\psi}} \notin \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)\right) + P\left(\widetilde{\boldsymbol{\psi}} \notin \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)\right) \\ + 2P\left(\sup_{\boldsymbol{\psi} \in \mathcal{B}(\boldsymbol{\psi}_0, \eta_\psi)} \left|\widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi}) - \widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi})\right| > \frac{\epsilon}{2}\right)\end{aligned}$$

Invoking Theorem 6 together with (E-15), one obtains by letting

$$\epsilon = C_0\left\{NT^{-C_\alpha} + N^{\gamma_M+1}\log(T)T^{\frac{\gamma_M}{2}-\frac{3}{8}\theta_\alpha} + N\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}$$

that

$$0 \leq \widetilde{\mathcal{Q}}_{N,T}(\widehat{\boldsymbol{\psi}}) - \widetilde{\mathcal{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}) = O_p\left(NT^{-C_\alpha} + N^{\gamma_M+1}\log(T)T^{\frac{\gamma_M}{2}-\frac{3}{8}\theta_\alpha} + N\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right). \quad \text{(E-17)}$$

Combining (E-13) and (E-17), we obtain the main theorem.

Appendix E..4 Proof of Theorem 8

We proceed in the following four main steps:

STEP 1: First, by arguing along the lines of the proof of Lemma 23, one obtains that

$$\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\epsilon_{0,*,t}^{(w)} = \frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\epsilon_{0,*,t} = \frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}(\boldsymbol{u}_c)\epsilon_{0,*,t}$$
$$-\frac{N}{T}\left(\sum_{t=1}^{T}\boldsymbol{x}_{*,t}(\boldsymbol{u}_c)\boldsymbol{w}_{*,t}^{\top}\right)\left(\sum_{s=1}^{T}\boldsymbol{w}_{*,s}\boldsymbol{w}_{*,s}^{\top}\right)^{-1}\left(\sum_{t=1}^{T}\boldsymbol{w}_{*,t}\epsilon_{0,*,t}\right) = \mathcal{T}_{1,N,T} + \mathcal{T}_{2,N,T}.$$

Define $N(\boldsymbol{u}_c) \doteq \sum_{i=1}^{N} u_{i,c}$ and $g_c \equiv g(\boldsymbol{u}_c) = \frac{N(\boldsymbol{u}_c)}{N}$. One then obtains that

$$\mathcal{T}_{1,N,T} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} u_{i,c}\boldsymbol{x}_{i,t}\sum_{i=1}^{N}\epsilon_{0,i,t} = \sqrt{g_c}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{N(\boldsymbol{u}_c)}}\sum_{i=1}^{N}u_{i,c}\boldsymbol{x}_{i,t}\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\epsilon_{0,i,t}\right)$$
$$\xrightarrow{w} \sigma_\epsilon\sqrt{g_c}\left(\boldsymbol{\Sigma}_\eta^{(c,c)}\right)^{\frac{1}{2}}\int_0^1 \boldsymbol{W}_\eta^{(c)}(\tau)dW_\epsilon(\tau),$$

where $\boldsymbol{\Sigma}_\eta^{(c,c)}$ is defined as in Lemma 23 and $\boldsymbol{W}_\eta^{(c)}(\tau)$ is a $d_x \times 1$ vector of Brownian motions with the covariance kernel $E[\boldsymbol{W}_\eta^{(c)}(\tau)\boldsymbol{W}_\eta^{(c)}(\kappa)^\top] = \min(\kappa,\tau)\mathbb{I}_{d_x}$. Moreover, note that $\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}(\boldsymbol{u}_c)\boldsymbol{w}_{*,t}^\top \leq \left(\frac{N}{T}\sum_{t=1}^{T}|\boldsymbol{x}_{*,t}(\boldsymbol{u}_c)|\right)\max_{t\in[1,T]}|\boldsymbol{w}_{*,t}^\top| = o_p(1)$ by Lemma 4. It then follows from the weak law of large numbers that $\mathcal{T}_{2,N,T} = o_p(1)$. Therefore, we obtain

$$\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\epsilon_{0,*,t}^{(w)} \xrightarrow{w} \sigma_\epsilon\sqrt{g_c}\left(\boldsymbol{\Sigma}_\eta^{(c,c)}\right)^{\frac{1}{2}}\int_0^1 \boldsymbol{W}_\eta^{(c)}(\tau)dW_\epsilon(\tau). \tag{E-18}$$

Using the same argument as in the proof of Lemma 25, we can also show that

$$\sqrt{\frac{N}{T}}\sum_{t=1}^{T}\xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c}\boldsymbol{u}_c)\epsilon_{0,*,t}^{(w)} = O_p(1), \tag{E-19}$$

$$\frac{N}{T^2}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_k)^\top \xrightarrow{w} g_c g_k\left(\boldsymbol{\Sigma}_\eta^{(c,c)}\right)^{\frac{1}{2}}\left(\boldsymbol{\Sigma}_\eta^{(k,k)}\right)^{\frac{1}{2}}\int_0^1 \boldsymbol{W}_\eta^{(c)}(\tau)\boldsymbol{W}_\eta^{(c)}(\tau)^\top d\tau, \tag{E-20}$$

$$\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c},\boldsymbol{u}_d) = O_p(1), \tag{E-21}$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c},\boldsymbol{u}_c)\xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c},\boldsymbol{u}_d) = O_p(1), \tag{E-22}$$

where all the terms in the limits are stochastic.

STEP 2: By the definitions of $\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)$, one can write:

$$\frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon_{0,*,t}^{(w)2}\} = 2\left(\sqrt{\frac{T}{N}}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)$$
$$\mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right) \mathrm{diag}\left(\sqrt{\frac{N}{T}}\mathbb{I}_{Gd_x}, \mathbb{I}_{2G+1}\right) \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)\epsilon_{0,*,t}^{(w)}$$
$$+\left(\sqrt{\frac{T}{N}}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right) \mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)$$
$$\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U}, \boldsymbol{U}_0)\mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)\left(\sqrt{\frac{T}{N}}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\right.$$
$$\left.\mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)^\top. \quad \text{(E-23)}$$

Eqs. (E-18) and (E-19) imply that

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U}, \boldsymbol{U}_0)\epsilon_{0,*,t}^{(w)} = \left(O_p\left(N^{-1}\right)\iota_{G\times d_x}, O_p\left(\frac{1}{\sqrt{NT}}\right)\iota_{2G+1}\right)$$

uniformly in $\boldsymbol{U}$. Therefore the first term in (E-23) is negligible in probability. Thus, one has that

$$\frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon_{0,*,t}^{(w)2}\} = \left(\sqrt{\frac{T}{N}}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top, (\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top, \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)$$
$$\mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U}, \boldsymbol{U}_0)\mathrm{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}), \mathbb{I}_{2G+1}\right)\left(\sqrt{\frac{T}{N}}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})} - \boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,\right.$$
$$\left.(\boldsymbol{\phi}_0^{(\sigma^{(per)})} - \boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\ \mu_{*,0} - \mu_*, \boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)^\top + o_p(1). \quad \text{(E-24)}$$

STEP 3: Define the following open balls: $\mathcal{B}(\boldsymbol{\theta}_0, \eta_\theta) = \{\boldsymbol{\theta} \in \Theta_\theta : \sqrt{\frac{T}{N}}H(\boldsymbol{\theta}, \boldsymbol{\theta}_0) < \eta_\theta\}$, $\mathcal{B}(\boldsymbol{\phi}_0, \eta_\phi) = \{\boldsymbol{\phi} \in \Theta_\phi : H(\boldsymbol{\phi}, \boldsymbol{\phi}_0) < \eta_\phi\}$, $\mathcal{B}(\mu_{*,0}, \eta_\mu) = \{|\mu - \mu_{*,0}| < \eta_\mu\}$, and $\mathcal{B}_N(\boldsymbol{U}_0, \eta_u) = \{\boldsymbol{U} \in \Delta_S^N : H(\boldsymbol{U}, \boldsymbol{U}_0) < \eta_u\}$. Let's denote by $\mathcal{A}(\boldsymbol{\psi}_0, \eta_\psi) = \bigcup_{\substack{\eta_\theta, \eta_\phi, \eta_\mu \\ (\eta_\theta^2 + \eta_\phi^2 + \eta_\mu^2)^{1/2} = \eta_\psi}} \mathcal{B}^c(\boldsymbol{\theta}_0, \eta_\theta) \times \mathcal{B}^c(\boldsymbol{\phi}_0, \eta_\phi) \times \mathcal{B}^c(\mu_{*,0}, \eta_\mu)$ a union of the complements of the above-defined open balls. It then follows that, for some $\eta_\phi \in (0, 1)$,

$$P\left(\widehat{\boldsymbol{\psi}} \in \mathcal{A}(\boldsymbol{\psi}_0, \eta_\phi), \widehat{\boldsymbol{U}} \in \mathcal{B}_N^c(\boldsymbol{U}_0, \eta_u)\right) \le P\left(\inf_{\substack{\boldsymbol{\psi} \in \mathcal{A}(\boldsymbol{\psi}_0, \eta_\phi) \\ \boldsymbol{U} \in \mathcal{B}_N^c(\boldsymbol{U}_0, \eta_u)}} \frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi}, \boldsymbol{U}) - \epsilon_{0,*,t}^{(w)2}\} \le 0\right) \quad \text{(E-25)}$$

By the eigenvalue inequality, it follows from (E-24) that

$$\inf_{\substack{\boldsymbol{\psi}\in\mathcal{A}(\boldsymbol{\psi}_0,\eta_\phi)\\ \boldsymbol{U}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u)}} \frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi},\boldsymbol{U})-\epsilon_{0,*,t}^{(w)2} \geq C_0 \inf_{\substack{\boldsymbol{\psi}\in\mathcal{A}(\boldsymbol{\psi}_0,\eta_\phi)\\ \boldsymbol{U}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u)}} \{\lambda_{\min}\left(\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U},\boldsymbol{U}_0)\right)$$

$$\left|H\left(\left(\sqrt{\frac{T}{N}}\boldsymbol{\theta},\boldsymbol{\phi},\mu_*\right),\left(\sqrt{\frac{T}{N}}\boldsymbol{\theta}_0,\boldsymbol{\phi}_0,\mu_{*,0}\right)\right)\right|^2\Bigg\}.$$

By sending N and T to infinity, one obtains from Assumption 4.6 that

$$\lim_{N\uparrow\infty,T\uparrow\infty}\inf_{\substack{\boldsymbol{\psi}\in\mathcal{A}(\boldsymbol{\psi}_0,\eta_\phi)\\ \boldsymbol{U}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u)}} \frac{1}{T}\sum_{t=1}^{T}\{\epsilon_{*,t}^2(\boldsymbol{\psi},\boldsymbol{U})-\epsilon_{0,*,t}^{(w)2}\} > C_0\eta_\psi^2 \text{ w.p.1.} \tag{E-26}$$

Eqs. (E-25) and (E-26) imply that

$$P\left(\widehat{\boldsymbol{\psi}}\in\mathcal{A}(\boldsymbol{\psi}_0,\eta_\phi),\widehat{\boldsymbol{U}}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u)\right)\downarrow 0.$$

Thus, one has shown that $\sqrt{\frac{T}{N}}H(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}_0)=o_p(1)$, $H\left((\widehat{\boldsymbol{\phi}},\widehat{\mu}_*),(\boldsymbol{\phi}_0,\mu_{*,0})\right)=o_p(1)$, and $H(\widehat{\boldsymbol{U}},\boldsymbol{U}_0)=o_p(1)$.
STEP 4: To refine the convergence rates of $\widehat{\boldsymbol{\psi}}$, define open balls,

$$\mathcal{N}_T(\boldsymbol{\theta}_0,\eta'_\theta)=\{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_0,\eta_\theta):\ \sqrt{T}H(\boldsymbol{\theta},\boldsymbol{\theta}_0)<\eta'_\theta\},$$

$$\mathcal{N}_N(\boldsymbol{\phi}_0,\eta'_\phi)=\{\boldsymbol{\phi}\in\mathcal{B}(\boldsymbol{\phi}_0,\eta_\phi):\ \sqrt{N}H(\boldsymbol{\phi},\boldsymbol{\phi}_0)<\eta'_\phi\},$$

$$\mathcal{N}_N(\mu_*,\eta'_\mu)=\{\mu_*\in\mathcal{B}(\mu_{*,0},\eta_\mu):\ \sqrt{N}|\mu_*-\mu_{*,0}|<\eta'_\mu\},$$

and

$$\mathcal{B}(\sigma_{\epsilon,0}^2,\eta_\sigma)=\{\sigma_\epsilon^2\in\Theta_\sigma:\ |\sigma_\epsilon^2-\sigma_{\epsilon,0}^2|<\eta_\sigma\}.$$

Let denote by $\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi)=\bigcup_{\substack{\eta'_\theta,\eta'_\phi,\eta'_\mu\\ (\eta'^2_\theta+\eta'^2_\phi+\eta'^2_\mu)^{\frac{1}{2}}=\eta'_\psi}}\mathcal{N}_T^c(\boldsymbol{\theta}_0,\eta'_\theta)\times\mathcal{N}_N^c(\boldsymbol{\phi}_0,\eta'_\phi)\times\mathcal{N}_N^c(\mu_*,\eta'_\mu)$ a union of the complements of these open balls. One obtains that

$$P\left(\widehat{\boldsymbol{\psi}}\in\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi),\widehat{\boldsymbol{U}}\in\mathcal{B}^c_N(\boldsymbol{U}_0,\eta_u),\widehat{\sigma}^2_\epsilon\in\mathcal{B}^c(\sigma^2_{\epsilon,0},\eta_\sigma)\right)$$

$$\leq P\left(\inf_{\substack{\boldsymbol{\psi}\in\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi)\\ \boldsymbol{U}\in\mathcal{B}^c_N(\boldsymbol{U}_0,\eta_u)\\ \sigma^2_\epsilon\in\mathcal{B}^c(\sigma^2_{\epsilon,0},\eta_\sigma)}}\left\{\overline{Q}_{N,T}(\boldsymbol{\psi}_0,\sigma^2_{\epsilon,0},\boldsymbol{U}_0)-\overline{Q}_{N,T}(\boldsymbol{\psi},\sigma^2_\epsilon,\boldsymbol{U})\right\}\leq 0\right) \quad \text{(E-27)}$$

Similar to the argument in STEP 2, notice that

$$\begin{aligned}\frac{N}{T}\sum_{t=1}^{T}\{\epsilon^2_{*,t}(\boldsymbol{\psi},\boldsymbol{U})-\epsilon^{(w)2}_{0,*,t}\} &= 2\left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})}-\boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,(\boldsymbol{\phi}_0^{(\sigma^{(per)})}-\boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\mu_{*,0}-\mu_*,\boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)\\
&\quad \text{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}),\mathbb{I}_{2G+1}\right)\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\epsilon^{(w)}_{0,*,t}\\
&+\left(\sqrt{T}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})}-\boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,\sqrt{N}(\boldsymbol{\phi}_0^{(\sigma^{(per)})}-\boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\sqrt{N}(\mu_{*,0}-\mu_*),\sqrt{N}\boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)\\
&\quad \text{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}),\mathbb{I}_{2G+1}\right)\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U},\boldsymbol{U}_0)\text{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}),\mathbb{I}_{2G+1}\right)\\
&\quad \left(\sqrt{T}(\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})}-\boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,\sqrt{N}(\boldsymbol{\phi}_0^{(\sigma^{(per)})}-\boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\ \sqrt{N}(\mu_{*,0}-\mu_*),\sqrt{N}\boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)^\top. \quad \text{(E-28)}\end{aligned}$$

In view of (E-23), we have that

$$\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\epsilon^{(w)}_{0,*,t}=\left(O_p(1)\,\iota_{G\times d_x},O_p\left(\sqrt{\frac{N}{T}}\right)\iota_{2G+1}\right).$$

uniformly in $\boldsymbol{U}$. Therefore, it is immediate to see that the first term on the right-hand side of (E-28) is probabilistically negligible for every $\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_0,\eta_\theta)$, $\boldsymbol{\phi}\in\mathcal{B}(\boldsymbol{\phi}_0,\eta_\phi)$,, $\mu_*\in\mathcal{B}(\mu_{*,0},\eta_\mu)$, and $\boldsymbol{U}\in\mathcal{B}(\boldsymbol{U}_0,\eta_u)$, i.e.,

$$\begin{aligned}&\left(\max_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left\{\left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})}-\boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,(\boldsymbol{\phi}_0^{(\sigma^{(per)})}-\boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\mu_{*,0}-\mu_*,\boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)\right.\right.\\
&\qquad\left.\text{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}),\mathbb{I}_{2G+1}\right)\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\epsilon^{(w)}_{0,*,t}\right\},\\
&\quad\max_{\sigma^{(per)}\in\sigma(\mathcal{P})}\min_{\widetilde{\sigma}^{(per)}\in\sigma(\mathcal{P})}\left\{\left((\boldsymbol{\theta}^{(\widetilde{\sigma}^{(per)})}-\boldsymbol{\theta}_0^{(\sigma^{(per)})})^\top,(\boldsymbol{\phi}_0^{(\sigma^{(per)})}-\boldsymbol{\phi}^{(\widetilde{\sigma}^{(per)})})^\top,\mu_{*,0}-\mu_*,\boldsymbol{\phi}_0^{(\sigma^{(per)})\top}\right)\right.\\
&\qquad\left.\left.\text{diag}\left(\boldsymbol{D}_\phi(\widetilde{\sigma}^{(per)}),\mathbb{I}_{2G+1}\right)\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{F}_t(\boldsymbol{U},\boldsymbol{U}_0)\epsilon^{(w)}_{0,*,t}\right\}\right)^+=o(1)O_p(1)+O_p\left(\sqrt{\frac{N}{T}}\right).\end{aligned}$$

By the eigenvalue inequality, one can show from (E-28) that

$$\inf_{\substack{\boldsymbol{\psi}\in\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi)\\ \boldsymbol{U}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u)}} \frac{N}{T}\sum_{t=1}^{T}\{\epsilon^2_{*,t}(\boldsymbol{\psi},\boldsymbol{U})-\epsilon^{(w)2}_{0,*,t}\} > o_p(1)+C_0\inf_{\boldsymbol{U}\in\mathcal{B}(\boldsymbol{U}_0,\eta_u)}\lambda_{\min}\left(\boldsymbol{\Lambda}_{N,T}(\boldsymbol{U},\boldsymbol{U}_0)\right)$$

$$\left|\inf_{\boldsymbol{\psi}\in\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi)} H\left((\sqrt{T}\boldsymbol{\theta},\sqrt{N}\boldsymbol{\phi},\sqrt{N}\mu_*),(\sqrt{T}\boldsymbol{\theta}_0,\sqrt{N}\boldsymbol{\phi}_0,\sqrt{N}\mu_{*,0})\right)\right|^2 > \eta'^2_\psi$$

in view of Assumption 4.6. The rest of the proof is immediate by following the same line as the proof of Theorem 7. Hence, in view of (E-27) we have

$$P\left(\widehat{\boldsymbol{\psi}}\in\mathfrak{A}(\boldsymbol{\psi}_0,\eta'_\psi),\widehat{\boldsymbol{U}}\in\mathcal{B}_N^c(\boldsymbol{U}_0,\eta_u),\widehat{\sigma}^2_\epsilon\in\mathcal{B}^c(\sigma^2_{\epsilon,0},\eta_\sigma)\right)\longrightarrow 0.$$

The main theorem then follows.

Appendix E..5 Proof of Theorem 9

The proof of this theorem follows along the same lines as the proof of Theorem 6; some of the arguments need to be modified due to the nonstationarity of $\boldsymbol{x}_{i,t}$. First, by Lemma 25, we have $\gamma = o_p\left(N^{-3/2}\right)+O_p\left(N^{-1}\right)$, where γ is defined in the proof of Theorem 6. It is then sufficient to derive the convergence rates for $\mathcal{T}^{(a)}_{1,N,T}$, $\mathcal{T}^{(b)}_{1,N,T}$, $\mathcal{T}^{(c)}_{1,N,T}$, and $\mathcal{T}^{(d)}_{1,N,T}$. Recall some notations that we have defined during the proof of Theorem 8: $N(\boldsymbol{u}_c)\doteq\sum_{i=1}^{N}u_{i,c}$ and $g_c\equiv g(\boldsymbol{u}_c)=\frac{N(\boldsymbol{u}_c)}{N}$.

To bound $\mathcal{T}^{(a)}_{1,N,T}$, an application of Lemma 4 yields

$$\frac{1}{\sqrt{N}T^{\frac{3}{2}}}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{i,c}\boldsymbol{x}_{i,t}=\sum_{t=1}^{T}\int_{\frac{t}{T}}^{\frac{t+1}{T}}\frac{1}{\sqrt{NT}}\sum_{s=1}^{\lfloor T\tau\rfloor}\sum_{i=1}^{N}u_{i,c}\boldsymbol{\eta}_{i,s}\frac{1}{T}\xrightarrow{w}\sqrt{g_c}\boldsymbol{\Sigma}_\eta^{(c,c)\frac{1}{2}}\int_0^1\boldsymbol{W}_\eta^{(c)}(\tau)d\tau, \quad \text{(E-29)}$$

where

$$\boldsymbol{\Sigma}_\eta^{(c,c)}=\text{plim}_{N,T\uparrow\infty}\frac{1}{TN(\boldsymbol{u}_c)}E\left[\boldsymbol{S}_\eta(N(\boldsymbol{u}_c),T)\boldsymbol{S}_\eta(N(\boldsymbol{u}_c),T)^\top\right],\ \text{where } \boldsymbol{S}_\eta(N(\boldsymbol{u}_c),t)=\sum_{s=1}^{t}\sum_{i=1}^{N}u_{i,c}\boldsymbol{\eta}_{i,s};$$

and $\boldsymbol{W}_\eta^{(c)}(\tau)$ is a $d_x\times1$ vector of Brownian motions with the covariance kernel $E[\boldsymbol{W}_\eta^{(c)}(\tau)\boldsymbol{W}_\eta^{(c)}(\kappa)^\top]=$

$\min(\tau, \kappa)\mathbb{I}_{d_x}$. It then follows that, for a given $\boldsymbol{\psi} \in \mathcal{N}_{N,T}(\boldsymbol{\psi}_0, \eta_\psi)$,

$$\min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \sum_{c=1}^{G} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \phi_{\sigma^{(per)}(c)} (u_{0,i,c} - u_{i,c})(\mu_{0,*} - \mu_*)(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} \right|$$
$$< C_0 \frac{1}{N} \left(\sqrt{N} |\mu_{0,*} - \mu_*| \right) \min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \sum_{c=1}^{G} \left(\sqrt{T} (\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \right) \left| \frac{1}{\sqrt{N} T^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{i=1}^{N} (u_{0,i,c} - u_{i,c}) \boldsymbol{x}_{i,t} \right|.$$

Moreover the weak convergence to a Gaussian process in (E-29) implies that

$$\lim_{N\uparrow\infty, T\uparrow\infty} E\left[\left| \frac{1}{\sqrt{N} T^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{i=1}^{N} (u_{0,i,c} - u_{i,c}) \boldsymbol{x}_{i,t} \right|^{C_\alpha} \right] < \infty$$

for every $C_\alpha \geq 1$. Therefore, by the consistency of $\widehat{\boldsymbol{\psi}}$ as demonstrated in Theorem 8, one obtains that the first term in (E-7) $\mathcal{T}_{1,N,T}^{(a*)} = O\left(N^{-C_\alpha}\right)$; and the convergence rates of the remaining terms $\mathcal{T}_{1,N,T}^{(a**)}$ and $\mathcal{T}_{1,N,T}^{(a***)}$ remain the same. Therefore, it follows that

$$\mathcal{T}_{1,N,T}^{(a)} < C_0 \left\{ T^{-C_\alpha} + N^{-C_\alpha} + N^{\gamma_M} \log(T) T^{\gamma_M - \frac{3}{4}\theta_\alpha} + \max\left(\exp\left(-C_{\epsilon_\eta} NT\right), \exp\left(-C_{\epsilon_\eta} \frac{T^{1/4}}{\log(T)} \right) \right) \right\}. \tag{E-30}$$

To bound $\mathcal{T}_{1,N,T}^{(b)}$, notice that

$$\sum_{c=1}^{G} \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (\phi_{0,c} - \phi_{\sigma^{(per)}(c)})(u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \left\{ (\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \boldsymbol{x}_{i,t}^{(w)} - \xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) \right\} \xi_{*,t}^{(w)}(\boldsymbol{\theta}_{0,c}, \boldsymbol{u}_{0,c})$$
$$= \sum_{c=1}^{G} \left(g_{*,c} E[\xi_{*,t}(\boldsymbol{\theta}_{0,c})] + o_p(1) \right) \left\{ (\phi_{0,c} - \phi_{\sigma^{(per)}(c)})(\boldsymbol{\theta}_{\sigma^{(per)}(c)} - \boldsymbol{\theta}_{0,c})^\top \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \boldsymbol{x}_{i,t}^{(w)} \right\}$$
$$+ \sum_{c=1}^{G} \left(g_{*,c} E[\xi_{*,t}(\boldsymbol{\theta}_{0,c})] + o_p(1) \right) (\phi_{0,c} - \phi_{\sigma^{(per)}(c)}) \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \left(\xi_{i,t}(\boldsymbol{\theta}_{0,c}) - E[\xi_{i,t}(\boldsymbol{\theta}_{0,c})] \right)$$
$$+ \sum_{c=1}^{G} \left(g_{*,c} E[\xi_{*,t}(\boldsymbol{\theta}_{0,c})] + o_p(1) \right) (\phi_{0,c} - \phi_{\sigma^{(per)}(c)}) \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (u_{0,i,c} - u_{i,\sigma^{(per)}(c)}) \left(E[\xi_{i,t}(\boldsymbol{\theta}_{0,,\sigma^{(per)}(c)})] + o_p(1) \right)$$
$$= \widetilde{A}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) + \widetilde{B}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) + \widetilde{C}_{N,T}(\boldsymbol{U}, \sigma^{(per)}).$$

Since
$\min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \left| \widetilde{A}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) + \widetilde{B}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) + \widetilde{C}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) \right| \approx \min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \left| \widetilde{A}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) \right| +$ $\min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \left| \widetilde{B}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) \right| + \min_{\sigma^{(per)} \in \sigma(\mathcal{P})} \left| \widetilde{C}_{N,T}(\boldsymbol{U}, \sigma^{(per)}) \right|$ for every $\boldsymbol{\psi} \in \mathcal{N}_{N,T}(\boldsymbol{\psi}_0, \eta_\psi)$, an ap-

plication of Boole's inequality yields

$$\mathcal{T}_{1,N,T}^{(b)} \leq P\left(\min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \widetilde{A}_{N,T}(\sigma^{(per)}) > \frac{\epsilon_\eta}{12}\right) + P\left(\min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \widetilde{B}_{N,T}(\sigma^{(per)}) > \frac{\epsilon_\eta}{12}\right)$$
$$+ P\left(\min_{\sigma^{(per)}\in\sigma(\mathcal{P})} \widetilde{C}_{N,T}(\sigma^{(per)}) > \frac{\epsilon_\eta}{12}\right) = \mathcal{T}_{1,N,T}^{(b*)} + \mathcal{T}_{1,N,T}^{(b**)} + \mathcal{T}_{1,N,T}^{(b***)}.$$

In view of (E-29), we have $\mathcal{T}_{1,N,T}^{(b*)} = O\left(N^{-C_\alpha}\right)$ for every $C_\alpha > 1$. The other terms $\mathcal{T}_{1,N,T}^{(b**)}$ and $\mathcal{T}_{1,N,T}^{(b***)}$ have the same convergence rates as $\mathcal{T}_{1,N,T}^{(a**)}$ and $\mathcal{T}_{1,N,T}^{(a***)}$. It then follows that

$$\mathcal{T}_{1,N,T}^{(b)} < C_0\left\{T^{-C_\alpha} + N^{-C_\alpha} + N^{\gamma_M}\log(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha} + \max\left(\exp\left(-C_{\epsilon_\eta}NT\right), \exp\left(-C_{\epsilon_\eta}\frac{T^{1/4}}{\log(T)}\right)\right)\right\}. \tag{E-31}$$

To bound $\mathcal{T}_{1,N,T}^{(c)}$, some simple calculations yield

$$\sum_{c=1}^{G}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top \boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_{0,c})\Big\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top\boldsymbol{x}_{i,t}^{(w)}$$
$$-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\Big\} = \frac{1}{N}\sum_{c=1}^{G}\phi_{\sigma^{(per)}(c)}\sqrt{T}\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top\left\{\frac{1}{T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\boldsymbol{x}_{i,t}^{(w)}\boldsymbol{x}_{*,t}^{(w)\top}\right\}$$
$$\sqrt{T}\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right) + \frac{1}{NT}\sum_{c=1}^{G}\phi_{\sigma^{(per)}(c)}\sqrt{T}\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top\frac{1}{\sqrt{T}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})$$
$$\left(\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c}) - E[\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})]\right)\boldsymbol{x}_{*,t}^{(w)} + \frac{1}{\sqrt{N}}\sum_{c=1}^{G}\phi_{\sigma^{(per)}(c)}\sqrt{T}\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top E[\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})]\frac{N^{\frac{1}{2}}}{T^{\frac{3}{2}}}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}$$
$$\frac{1}{N}\sum_{i=1}^{N}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)}) = \mathfrak{T}_{1,N,T}(\sigma^{(per)},\boldsymbol{U}) + \mathfrak{T}_{2,N,T}(\sigma^{(per)},\boldsymbol{U}) + \mathfrak{T}_{3,N,T}(\sigma^{(per)},\boldsymbol{U}).$$

By (D-9) in Lemma 25, one can show that

$$\frac{N}{T^2}\sum_{t=1}^{T}u_{i,c}\boldsymbol{x}_{*,t}^{(w)}\boldsymbol{x}_{*,t}^{(w)\top} \xrightarrow{w} g_c\boldsymbol{\Sigma}_\eta^{(c,c)}\int_0^1\boldsymbol{W}_\eta^{(c)}(\tau)\boldsymbol{W}_\eta^{(c)}(\tau)^\top d\tau.$$

Therefore, an application of the dominated convergence theorem and the Tchebyshev inequality yields

$$P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\mathfrak{T}_{1,N,T}(\sigma^{(per)},\boldsymbol{U})\right| > \frac{\epsilon_\eta}{8}\right) = O\left(N^{-C_\alpha}\right) \text{ for every } C_\alpha > 1.$$

Moreover, by Lemma 4 and the continuous mapping theorem, one can show that

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{i,c}\left(\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})-E[\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})]\right)\boldsymbol{x}_{*,t}^{(w)}\overset{w}{\longrightarrow}g_c\sigma_{\xi}^{(c)}\boldsymbol{\Sigma}_{\eta}^{(c,c)}\int_0^1\boldsymbol{W}_{\eta}^{(c)}(\tau)dW_{\xi}(d\tau),$$

where $\sigma_{\xi}^{(c)2}=\lim_{N\uparrow\infty,T\uparrow\infty}\frac{1}{N(\boldsymbol{u}_c)T}Var\left(\sum_{t=1}^{T}\sum_{i=1}^{N}u_{i,c}\xi_{i,t}(\boldsymbol{\theta}_{0,c})\right)$. Thus, we have

$$P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\mathfrak{T}_{2,N,T}(\sigma^{(per)},\boldsymbol{U})\right|>\frac{\epsilon_\eta}{8}\right)=O\left((NT)^{-C_\alpha}\right)\text{ for every }C_\alpha>1.$$

By the same argument, we also obtain

$$P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\mathfrak{T}_{3,N,T}(\sigma^{(per)},\boldsymbol{U})\right|>\frac{\epsilon_\eta}{8}\right)=O\left(N^{-C_\alpha/2}\right)\text{ for every }C_\alpha>1.$$

It then follows that

$$\mathcal{T}_{1,N,T}^{(c)}<C_0\left(N^{-C_\alpha/2}+(NT)^{-C_\alpha}\right). \tag{E-32}$$

Finally, to bound $\mathcal{T}_{4,N,T}^{(d)}$, notice that

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_{\sigma^{(per)}(c)}(u_{0,i,c}-u_{i,\sigma^{(per)}(c)})\left\{(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top\boldsymbol{x}_{i,t}^{(w)}-\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\right\}\epsilon_{0,*,t}^{(w)}\\&\quad=\frac{1}{N\sqrt{T}}\phi_{\sigma^{(per)}(c)}\sqrt{T}(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c})^\top\frac{N}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_{0,c})\epsilon_{0,*,t}-\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_{\sigma^{(per)}(c)})\epsilon_{0,*,t}\right)\\&\quad+\phi_{\sigma^{(per)}(c)}\frac{1}{T}\sum_{t=1}^{T}\left(\xi_{*,t}^{(w)}(\boldsymbol{u}_{0,c})\epsilon_{0,*,t}-\xi_{*,t}^{(w)}(\boldsymbol{u}_{\sigma^{(per)}(c)})\epsilon_{0,*,t}\right)=\mathfrak{L}_{1,N,T}(\sigma^{(per)},\boldsymbol{U})+\mathfrak{L}_{2,N,T}(\sigma^{(per)},\boldsymbol{U}).\end{aligned}$$

An application of Lemma 23 yields that $\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{*,t}^{(w)}(\boldsymbol{u}_c)\epsilon_{0,*,t}\overset{w}{\longrightarrow}\sqrt{g_c}\sigma_\epsilon\boldsymbol{\Sigma}_{\eta}^{(c,c)}\int_0^1\boldsymbol{W}_{\eta}^{(c,c)}(\tau)dW_\epsilon(\tau)$. Thus, by the dominated convergence theorem and the Tchebyshev inequality, one readily obtains

$$P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\mathfrak{L}_{1,N,T}(\sigma^{(per)},\boldsymbol{U})\right|>\frac{\epsilon_\eta}{8}\right)=O\left((N\sqrt{T})^{-C_\alpha}\right)\text{ for every }C_\alpha>1.$$

Also, by Lemma 7, we can show that

$$P\left(\sup_{\boldsymbol{U}\in\Delta_S^N}\min_{\sigma^{(per)}\in\sigma(\mathcal{P})}\left|\mathfrak{L}_{2,N,T}(\sigma^{(per)},\boldsymbol{U})\right|>\frac{\epsilon_\eta}{8}\right)=O\left(T^{-C_\alpha}+N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}\right.$$
$$\left.+\max\left\{\exp\left(-C_\sigma N^2\log^2(T)T^{7/4}\right),\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}\right).$$

It then follows that

$$\mathcal{T}_{4,N,T}^{(d)}<C_0\left\{T^{-C_\alpha}+N^{2\gamma_M}\log^2(T)T^{\gamma_M-\frac{3}{4}\theta_\alpha}\right.$$
$$\left.+\max\left\{\exp\left(-C_\sigma N^2\log^2(T)T^{7/4}\right),\exp\left(-C_M\frac{T^{1/4}}{\log^2(T)}\right)\right\}\right\}. \quad \text{(E-33)}$$

Collecting all the terms derived in (E-30)-(E-33), we obtain the main theorem.

Appendix E..6 Proof of Theorem 10

Recall some commonly-used notations: $\epsilon_{*,t}=(\boldsymbol{\psi}-\widetilde{\boldsymbol{\psi}})^\top\boldsymbol{D}_\phi\boldsymbol{D}_g\boldsymbol{X}_{N,T,t}(\boldsymbol{\theta})+\epsilon_{*,t}(\widetilde{\boldsymbol{\psi}},\boldsymbol{U}_0)$, where

$$\boldsymbol{X}_{N,T,t}(\boldsymbol{\theta})\doteq\left(\boldsymbol{x}_{*,t}^{(1)\top},\ldots,\boldsymbol{x}_{*,t}^{(G)\top},-\xi_{*,t}^{(1)}(\boldsymbol{\theta}_1),\ldots,-\xi_{*,t}^{(G)}(\boldsymbol{\theta}_G),-1\right)^\top$$

with $\boldsymbol{D}_\phi=\operatorname{diag}(\boldsymbol{\phi}\otimes\mathbb{I}_{d_x},\mathbb{I}_{G+1})$ and $\boldsymbol{D}_g=\operatorname{diag}(\boldsymbol{g}\otimes\mathbb{I}_{d_x},\boldsymbol{g},1)$; and $\widetilde{\mathfrak{Q}}_{N,T}(\boldsymbol{\psi})=\frac{N}{T}\sum_{t=1}^T\epsilon_{*,t}^2(\boldsymbol{\psi},\boldsymbol{U}_0)$. This proof proceeds along the lines of the proof of Theorem 7. As in the Step 1, it can be shown that

$$\widetilde{\mathfrak{Q}}_{N,T}(\widehat{\boldsymbol{\psi}})-\widetilde{\mathfrak{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}})$$
$$\geq\lambda_{\lim}\left(\operatorname{diag}\left(T^{-1/2}\mathbb{I}_{G\times d_x},N^{-1/2}\mathbb{I}_{G+1}\right)\left(\frac{N}{T}\sum_{t=1}^T\boldsymbol{X}_{N,T,t}(\widehat{\boldsymbol{\theta}})\boldsymbol{X}_{N,T,t}(\widehat{\boldsymbol{\theta}})^\top\right)\operatorname{diag}\left(T^{-1/2}\mathbb{I}_{G\times d_x},N^{-1/2}\mathbb{I}_{G+1}\right)\right)$$
$$H\left(\operatorname{diag}\left(\sqrt{T}\mathbb{I}_{G\times d_x},\sqrt{N}\mathbb{I}_{G+1}\right)\widehat{\boldsymbol{\psi}},\operatorname{diag}\left(\sqrt{T}\mathbb{I}_{G\times d_x},\sqrt{N}\mathbb{I}_{G+1}\right)\widetilde{\boldsymbol{\psi}}\right).$$

Since $\widehat{\boldsymbol{\theta}}$ is consistent by Theorem 8, it then follows from Assumption 4.4 that

$$H\left(\operatorname{diag}\left(\sqrt{T}\mathbb{I}_{G\times d_x},\sqrt{N}\mathbb{I}_{G+1}\right)\widehat{\boldsymbol{\psi}},\operatorname{diag}\left(\sqrt{T}\mathbb{I}_{G\times d_x},\sqrt{N}\mathbb{I}_{G+1}\right)\widetilde{\boldsymbol{\psi}}\right)\leq\widetilde{\mathfrak{Q}}_{N,T}(\widehat{\boldsymbol{\psi}})-\widetilde{\mathfrak{Q}}_{N,T}(\widetilde{\boldsymbol{\psi}}). \quad \text{(E-34)}$$

Next, by applying Lemmas 24 and 25, it can be shown that

$$\begin{aligned}
\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^{T}\xi_{i,t}^{(w)}\boldsymbol{x}_{i,t}^{(w)} &= O_p(1), && \text{(E-35)}\\
\frac{1}{T}\sum_{t=1}^{T}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,d}) &= O_p(1),\ \forall\ c,d\in[1,G], && \text{(E-36)}\\
\sqrt{\frac{N}{T}}\sum_{t=1}^{T}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\epsilon_{0,*,t} &= O_p(1), && \text{(E-37)}\\
\frac{N}{T^2}\sum_{t=1}^{T}\boldsymbol{x}_{i,t}^{(w)}\boldsymbol{x}_{i,t}^{(w)\top} &= O_p(1), && \text{(E-38)}\\
\frac{N}{T}\sum_{t=1}^{T}\boldsymbol{x}_{i,t}^{(w)}\epsilon_{0,*,t} &= O_p(1), && \text{(E-39)}\\
\frac{N^{1/2}}{T^{3/2}}\sum_{t=1}^{T}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\boldsymbol{x}_{i,t}^{(w)} &= O_p(1). && \text{(E-40)}
\end{aligned}$$

It then follows from (3.8), (E-35)-(E-37) that

$$\frac{N}{T}\sum_{t=1}^{T}\xi_{i,t}^{(w)}(\boldsymbol{\theta}_{0,c})\left\{\epsilon_{*,t}(\boldsymbol{\psi},\widehat{\boldsymbol{U}})+\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U}_0)\right\}=o_p\left(\sqrt{N}\right); \qquad \text{(E-41)}$$

and also from (3.8), (E-38)-(E-40), one obtains that

$$\sqrt{T}\min_{\sigma\in\sigma(\mathcal{P})}\sum_{c=1}^{G}\left|\left(\boldsymbol{\theta}_{\sigma^{(per)}(c)}-\boldsymbol{\theta}_{0,c}\right)^\top\right|\left|\frac{N}{T^{3/2}}\sum_{t=1}^{T}\boldsymbol{x}_{i,t}^{(w)}\left\{\epsilon_{*,t}(\boldsymbol{\psi},\widehat{\boldsymbol{U}})+\epsilon_{*,t}(\boldsymbol{\psi},\boldsymbol{U}_0)\right\}\right| = o_p\left(\sqrt{N}\right) \qquad \text{(E-42)}$$

for some $\boldsymbol{\psi}\in\mathcal{B}_T(\boldsymbol{\theta}_0,\eta_\theta)\times\mathcal{B}_N(\boldsymbol{\phi}_0,\eta_\phi)\times\mathcal{B}_N\left(\mu_{*,0},\eta_\mu\right)$, where these shrinking balls are defined in Theorem 9. Recall the representation (E-14) established in the proof of Theorem 7. By Theorem 9 together with (E-41) and (E-42), one then has

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{\psi}=(\boldsymbol{\theta}^\top,\boldsymbol{\phi}^\top,\mu_*)^\top \text{ s.t.}\\ \boldsymbol{\theta}\in\mathcal{B}_T(\boldsymbol{\theta}_0,\eta_\theta)\\ \boldsymbol{\phi}\in\mathcal{B}_N(\boldsymbol{\phi}_0,\eta_\phi)\\ \mu_*\in\mathcal{B}_N(\mu_{*,0},\eta_\mu)}}\left|\widehat{\mathcal{Q}}_{N,T}(\boldsymbol{\psi})-\widetilde{\mathcal{Q}}_{N,T}(\boldsymbol{\psi})\right|\\
&\qquad = O_p\left(N^{\frac{1}{2}-\frac{C_\alpha}{2}}+N^{\frac{1}{2}}T^{-\frac{C_\alpha}{2}}+N^{\gamma_M+\frac{1}{2}}\log(T)T^{\frac{\gamma_M}{2}-\frac{3}{8}\theta_\alpha}+N^{\frac{1}{2}}\exp\left(-C_{\epsilon_\eta}\frac{T^{\frac{1}{4}}}{2\log^2(T)}\right)\right).
\end{aligned}$$

The rest of this proof follows exactly the same argument in Step 3 of the proof of Theorem 7.

Appendix F. DC Decomposition of the Sum of Squared Composite Errors

In view of (2.3) the composite errors are given by

$$\epsilon_{*,t} \doteq \epsilon_{*,t}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) = \Delta y_{*,t} - \mu_* - \sum_{g=1}^{G} \frac{1}{N} \sum_{i=1}^{N} \phi_g u_{i,g} \left(y_{i,t-1} - \boldsymbol{\theta}_g^\top \boldsymbol{x}_{i,t} \right) - \sum_{g=1}^{G} \frac{1}{N} \sum_{i=1}^{N} u_{i,g} \boldsymbol{\lambda}_g^\top \boldsymbol{w}_{i,t},$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1^\top, \dots, \boldsymbol{\lambda}_G^\top)^\top$. We can show by some simple calculations that

$$\frac{N^2}{T} \sum_{t=1}^{T} \epsilon_{*,t}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) + \widetilde{\gamma} g(\boldsymbol{U}) = \widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) - \widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U})$$

with

$$\begin{aligned}
\widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) &= \frac{5}{2}\rho_u \sum_{g=1}^{G}\sum_{i=1}^{N} u_{i,g}^2 + 2\rho_\phi \sum_{g=1}^{G} \phi_c^2 + 2\rho_\theta \sum_{g=1}^{G} \boldsymbol{\theta}_g^\top \boldsymbol{\theta}_g + 2\rho_\lambda \sum_{g=1}^{G} \boldsymbol{\lambda}_g^\top \boldsymbol{\lambda}_g + 2\rho_\mu \mu_*^2, \\
\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) &= \widetilde{\mathcal{G}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) - \frac{N^2}{T} \sum_{t=1}^{T} \epsilon_{*,t}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U}) - \widetilde{\gamma} g(\boldsymbol{U}),
\end{aligned}$$

and $g(\boldsymbol{U}) = \sum_{g=1}^{G} \sum_{i=1}^{N} u_{i,g}(1 - u_{i,g})$ for some $\widetilde{\gamma} > 0$.

The function $\widetilde{\mathcal{H}}_{N,T}$ is convex for some choice of $\boldsymbol{\rho} = (\rho_u, \rho_\phi, \rho_\theta, \rho_\lambda, \rho_\mu)$ similar to what was asserted in Lemmas 8-10. Some algebraic manipulations yield the following gradient vector of $\widetilde{\mathcal{H}}_{N,T}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \mu_*, \boldsymbol{U})$:

$$\begin{aligned}
\frac{\partial \widetilde{\mathcal{H}}_{N,T}}{\partial u_{i,g}} &= 5\rho_u u_{i,g} + 2\frac{N}{T} \sum_{t=1}^{T} \epsilon_{*,t} \left\{ \phi_g (y_{i,t-1} - \boldsymbol{\theta}_g^\top \boldsymbol{x}_{i,t}) + \boldsymbol{\lambda}_g^\top \boldsymbol{w}_{i,t} \right\} + \widetilde{\gamma}(2u_{i,g} - 1), \\
\frac{\partial \widetilde{\mathcal{H}}_{N,T}}{\partial \phi_g} &= 5\rho_\phi \phi_g + 2\frac{N}{T} \sum_{t=1}^{T} \epsilon_{*,t} \left\{ \sum_{i=1}^{N} u_{i,g} \left(y_{i,t-1} - \boldsymbol{\theta}_g^\top \boldsymbol{x}_{i,t} \right) \right\}, \\
\frac{\partial \widetilde{\mathcal{H}}_{N,T}}{\partial \boldsymbol{\theta}_g} &= 4\rho_\theta \boldsymbol{\theta}_g - 2\frac{N}{T} \phi_g \sum_{t=1}^{T} \epsilon_{*,t} \sum_{i=1}^{N} u_{i,g} \boldsymbol{x}_{i,t}, \\
\frac{\partial \widetilde{\mathcal{H}}_{N,T}}{\partial \boldsymbol{\lambda}_g} &= 4\rho_\lambda \boldsymbol{\lambda}_g + 2\frac{N}{T} \sum_{i=1}^{N} u_{i,g} \sum_{t=1}^{T} \epsilon_{*,t} \boldsymbol{w}_{i,t}, \\
\frac{\partial \widetilde{\mathcal{H}}_{N,T}}{\partial \mu_*} &= 4\rho_\mu \mu_* + 2\frac{N^2}{T} \sum_{t=1}^{T} \epsilon_{*,t}.
\end{aligned}$$

Appendix G. DC Programming and DCA: A Synopsis

Recall that, in the case when group memberships are known, the composite likelihood function is convex. One can then employ various algorithms in convex optimization (e.g., Newton-Raphson or Simulated Annealing) to compute solutions to the problem of the composite likelihood maximization. However, these algorithms for convex optimization are not sufficient to deal with a large-scale non-convex optimization problem that arises when one has to incorporate unknown group membership variables into the composite likelihood function. To surpass the difficulty of optimizing large-scale non-convex functions, a theory of optimization for a superclass of convex functions, so-called difference-of-convex (d.c.) functions, has been extensively developed (see, e.g., Hiriart-Urruty (1985, 1988) for precursors to the DC programming). A concise review of DC programming and global optimization of d.c. functions is also provided in Hoang (1995) and Thoai (1999).

A DC program can be defined as

$$(P_{dc})\ \min\{f(\boldsymbol{x}) \doteq g(\boldsymbol{x}) - h(\boldsymbol{x}) :\ \boldsymbol{x} \in \mathbb{R}^d\},$$

where $g(\cdot)$ and $h(\cdot)$ represent lower semi-continuous proper convex functions on $\mathbb{R}^d$. Such a function as $f(\cdot)$ is called a d.c. function. Class of d.c. functions is rather large so that most of functions encountered in econometric applications are d.c. functions. Note that convex constraints of type, $\boldsymbol{x} \in \mathcal{C} \subset \mathbb{R}^d$, can be taken into account using a set characteristic function, $\min\{f(\boldsymbol{x}) \doteq g(\boldsymbol{x}) - h(\boldsymbol{x}) :\ \boldsymbol{x} \in \mathcal{C}\} = \min\{\chi_{\mathcal{C}}(\boldsymbol{x}) + f(\boldsymbol{x}) :\ \boldsymbol{x} \in \mathbb{R}^d\}$, where $\chi_{\mathcal{C}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{C}$, and $= +\infty$ otherwise. Let

$$g^*(\boldsymbol{y}) = \sup\{< \boldsymbol{x}, \boldsymbol{y} > - g(\boldsymbol{x}) :\ \boldsymbol{x} \in \mathbb{R}^d\}$$

be the conjugate of $g(\boldsymbol{x})$. One then obtains the following dual program of P_{dc} :

$$(D_{dc})\ \min\{h^*(\boldsymbol{y}) - g^*(\boldsymbol{y}) :\ \boldsymbol{y} \in \mathbb{R}^d\}.$$

To see this duality, notice that, since $g(x) = \sup\{< \boldsymbol{x}, \boldsymbol{y} > - g^*(\boldsymbol{y}) :\ \boldsymbol{y} \in \mathbb{R}^d\}$ and $h(x) = \sup\{< \boldsymbol{x}, \boldsymbol{y} > - h^*(\boldsymbol{y}) :\ \boldsymbol{y} \in \mathbb{R}^d\}$, one has

$$\begin{aligned} \inf\{g(\boldsymbol{x}) - h(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d\} &= \inf\{g(\boldsymbol{x}) - \sup\{< \boldsymbol{x}, \boldsymbol{y} > - h^*(\boldsymbol{y}) :\ \boldsymbol{y} \in \mathbb{R}^d\} : \boldsymbol{x} \in \mathbb{R}^d\} \\ &= \inf\left\{\inf\{g(\boldsymbol{x}) - < \boldsymbol{x}, \boldsymbol{y} > + h^*(\boldsymbol{y}) :\ \boldsymbol{x} \in \mathbb{R}^d\} :\ \boldsymbol{y} \in \mathbb{R}^d\right\} = \inf\left\{h^*(\boldsymbol{y}) - g^*(\boldsymbol{y}) :\ \boldsymbol{y} \in \mathbb{R}^d\right\}. \end{aligned}$$

Therefore the optimal solution to the program P_{dc} is the same as the optimal solution to the program D_{dc}. The existence of these optimal solutions is warranted by the generalized Kuhn-Tucker global optimality condition (see, e.g., Hoang (2010, Proporsition 3.19)). However, in many large-scale non-convex problems, a number of algorithms searching for a point satisfying the global optimality condition - such as branch-and-bound and cutting cones - do not often compute optimal points efficiently (see Horst and Hoang (1993)). The DCA based on the duality in d.c. optimization - first introduced by Pham Dinh and Souad (1988) - is among a few algorithms which allow to solve large-scale d.c. optimization problems (see, e.g., Le Thi Hoai An and Pham Dinh Tao (2003); Pham Dinh Tao and Le Thi Hoai An (1998)). In the d.c. programming literature the DCAs converge to local solutions due to their local optimality nature; however, they often yield the global optimum, and a number of regularization and initialization methods can be used to facilitate the finding of global optimum from local ones in many different cases. A comprehensive introduction to DCA is provided in Pham Dinh Tao and Le Thi Hoai An (1997); and an incisive outline of DCA is given in Le Thi Hoai An (2014); Pham Dinh Tao and Le Thi Hoai An (2014). DCA has been successfully applied to many large-scale non-convex problems in applied science, especially, in Machine Learning where the use of the DCA often leads to global solutions and proves to be more robust than the standard methods (see, Le Thi Hoai An, Le Hoai Minh, and Pham Dinh Tao (2014) and references therein).

The DCA is an iterative primal-dual subgradient method consisting of two sequences, $\{\boldsymbol{x}^{(\ell)}\}$ and $\{\boldsymbol{y}^{(\ell)}\}$, chosen such that $\{g(\boldsymbol{x}^{(\ell)}) - h(\boldsymbol{x}^{(\ell)})\}$ and $\{h^*(\boldsymbol{y}^{(\ell)}) - g^*(\boldsymbol{y}^{(\ell)})\}$ are decreasing so that $\{\boldsymbol{x}^{(\ell)}\}$ and $\{\boldsymbol{y}^{(\ell)}\}$ converge to a feasible primal solution, $\boldsymbol{x}^*$, and a feasible dual solution, $\boldsymbol{y}^*$, respectively. These feasible solutions are shown to satisfy local optimality conditions and

$$\boldsymbol{x}^* \in \partial g^*(\boldsymbol{y}^*) \text{ and } \boldsymbol{y}^* \in \partial h(\boldsymbol{x}^*), \tag{G-1}$$

where $\partial h(\boldsymbol{x}^*)$ is the subdifferential of $h(\boldsymbol{x})$ at $\boldsymbol{x}^*$. Replacing $h(\cdot)$ in the program P_{dc} with its affine minorization, $h_\ell(\boldsymbol{x}) \doteq h(\boldsymbol{x}^{(\ell)}) + < \boldsymbol{x} - \boldsymbol{x}^{(\ell)}, \boldsymbol{y}^{(\ell)} >$ with $\boldsymbol{y}^{(\ell)} \in \partial h(\boldsymbol{x}^{(\ell)})$, one can then obtain a convex approximation to the primal d.c. program P_{dc}:

$$(P_\ell) \ \min\{g(\boldsymbol{x}) - h_\ell(\boldsymbol{x})\}.$$

By the following property of subdifferentials of convex functions: $\boldsymbol{y} \in \partial g(\boldsymbol{x}) \leftrightarrow \boldsymbol{x} \in \partial g^*(\boldsymbol{y}) \leftrightarrow < \boldsymbol{x}, \boldsymbol{y} >= g(\boldsymbol{x}) + g^*(\boldsymbol{y})$, the optimal solution $\boldsymbol{x}^{(\ell+1)}$ to the program P_ℓ satisfies $\boldsymbol{x}^{(\ell+1)} \in \partial g^*(\boldsymbol{y}^{(\ell)})$. This gives rise to the following generic DCA scheme:

The DCA has a linear convergence rate so that every limiting point of the sequence $\{\boldsymbol{x}^{(\ell)}\}$ or $\{\boldsymbol{y}^{(\ell)}\}$ is a generalized KKT point of $g - h$ or $h^* - g^*$ regardless of chosen starting values. It is worth mentioning that many standard methods of convex and non-convex programming are particular cases of DCA, for example, Expectation-Maximization (EM) of Dempster, Laird, and Rubin (1977), Successive Linear Approximation (SLA) of Bradley and Mangasarian (1998), and Convex-Concave Procedure (CCCP) of Yuille and Rangarajan (2003).

Efficient implementation of DCA involves an appropriate d.c. decomposition of $f(\cdot)$ and a good

Algorithm 5 Generic DCA Scheme

```
given an initial guess, x^(0) ∈ R^d, and an error tolerance level, ε
ℓ ← 0
do
    calculate y^(ℓ) ∈ ∂h(x^(0))
    calculate x^(ℓ+1) ∈ ∂g*(y^(ℓ)), which is equivalent to x^(ℓ+1) ∈ argmin{g(x) − h(x^(ℓ)) − < x − x^(ℓ), y^(ℓ) >: x ∈ R^d}.
    ℓ ← ℓ + 1
while ‖x^(ℓ+1) − x^(ℓ)‖ < ε
```

given an initial guess, $\boldsymbol{x}^{(0)} \in \mathbb{R}^d$, and an error tolerance level, ϵ
$\ell \leftarrow 0$
do
  calculate $\boldsymbol{y}^{(\ell)} \in \partial h(\boldsymbol{x}^{(0)})$
  calculate $\boldsymbol{x}^{(\ell+1)} \in \partial g^*(\boldsymbol{y}^{(\ell)})$, which is equivalent to $\boldsymbol{x}^{(\ell+1)} \in \operatorname{argmin}\{g(\boldsymbol{x}) - h(\boldsymbol{x}^{(\ell)}) - < \boldsymbol{x} - \boldsymbol{x}^{(\ell)}, \boldsymbol{y}^{(\ell)} >: \ \boldsymbol{x} \in \mathbb{R}^d\}$.
  $\ell \leftarrow \ell + 1$
while $\|\boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)}\| < \epsilon$

starting point. If $f(\cdot)$ is such a d.c. function that $\frac{1}{2}\rho\|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ is convex for some sufficiently large ρ, then $f(\cdot) = g(\cdot) - h(\cdot)$, where $g(\boldsymbol{x}) = \frac{1}{2}\rho\|\boldsymbol{x}\|^2$ and $h(\boldsymbol{x}) = \frac{1}{2}\rho\|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$. This special decomposition gives rise to the following algorithm:

Algorithm 6 Special-Decomposition DCA Scheme

1: given an initial guess, $\boldsymbol{x}^{(0)} \in \mathbb{R}^d$, and an error tolerance level, ϵ
2: $\ell \leftarrow 0$
3: **do**
4: calculate $\boldsymbol{y}^{(\ell)} \in \partial\left(\frac{1}{2}\rho\|\cdot\|^2 - f(\cdot)\right)(\boldsymbol{x}^{(\ell)})$
5: $\boldsymbol{x}^{(\ell+1)} \in \operatorname{argmin}\{\frac{1}{2}\rho\|\boldsymbol{x}\|^2 - < \boldsymbol{x}, \boldsymbol{y}^{(\ell)} >: \ \boldsymbol{x} \in \mathcal{C} \subset \mathbb{R}^d\}$.
6: $\ell \leftarrow \ell + 1$
7: **while** $\|\boldsymbol{x}^{(\ell+1)} - \boldsymbol{x}^{(\ell)}\| < \epsilon$

Algorithm 6 is practically convenient because the convex minimization subproblem on line 5 can easily be solved by using the orthogonal projection, i.e., $\boldsymbol{x}^{(\ell+1)} = \operatorname{Proj}_{\mathcal{C}}(\frac{\boldsymbol{y}^{(\ell)}}{\rho})$; in fact, there are many algorithms to compute the projection onto convex sets (e.g., box constraints, polyhedron, simplices) [see, e.g., Chen and Ye (2011); Júdice, Raydan, Rosa, and Santos (2008)].

References

Alexiadis, S. (2013): *Convergence Clubs and Spatial Externalities: Models and Applications of Regional Convergence in Europe*, Advances in Spatial Science. Springer-Verlag, Berlin Heidelberg.

Ando, T., and J. Bai (2016): "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, 31, 163–191.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49(4), 431434.

Bester, C. A., and C. B. Hansen (2016): "Grouped effects estimators in fixed effects models," *Journal of Econometrics*, 190(1), 197–208.

Billingsley, P. (1968): *Convergence of Probability Measures*. John Wiley & Sons, New York, Singapore, Toronto, 1 edn.

Bonhomme, S., T. Lamadon, and E. Manresa (2016): "Discritizing Unobserved Heterogeneity: Approximate Clustering Methods for Dimension Reduction," mimeo.

Bonhomme, S., and E. Manresa (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

Bonnans, J. F., and A. Shapiro (2000): *Perturbation Analysis of Optimization Problems*. Springer, New York, Berlin, Heidelberg.

Bradley, P. S., and O. L. Mangasarian (1998): "Feature Selection via Concave Minimization and Support Vector Machines," in *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, ed. by J. Shavlik, pp. 82–90, San Francisco, California. Morgan Kaufmann.

Brown, D. E., and C. L. Huntley (1992): "A practical application of simulated annealing to clustering," *Pattern Recognition*, 25(4), 401–412.

Bulinski, A., and A. Shashkin (2006): "Strong invariance principle for dependent random fields," in *Dynamics & Stochastics : Festschrift in honor of M. S. Keane*, ed. by D. Denteneer, F. den Hollander, and E. Verbitskiy, IMS Lecture Notes - Monograph Series Volume 48, pp. 128–143, Beachwood, Ohio, USA. Institute of Mathematical Statistics.

Bulinski, A., and A. Shashkin (2007): *Limit Theorems for Associated Random Fields and Related Systems*. World Scientific, Singapore, 1 edn.

Carbon, M., L. T. Tran, and B. Wu (1997): "Kernel density estimation for random fields: the L_1 theory," *Journal of Nonparametric Statistics*, 6, 157–170.

Carnicer, J. M., T. N. T. Goodman, and J. M. Pena (1999): "Linear conditions for positive determinants," *Linear Algebra and Its Applications*, 292, 39–59.

Chang, Y., and R. T. Smith (2014): "Feldstein-Horioka puzzles," *European Economic Review*, 72, 98–112.

Chen, Y., and X. Ye (2011): "Projection onto a simplex," memeo.

Coakley, J., A.-M. Fuertes, and F. Spagnolo (2004): "Is the Feldstein-Horioka Puzzle History?," *Manchester School*, 72(5), 569–590.

Coakley, J., F. Kulasi, and R. Smith (1996): "Current account solvency and the feldstein-horioka puzzle," *Economic Journal*, 106, 620–627.

Conley, T. G. (1999): "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 92, 1–45.

Corrado, L., R. Martin, and M. Weeks (2005): "Identifying and interpreting regional convergence clusters across Europe," *Economic Journal*, 115(502), C133–C160.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society B*, 39(1), 1–38.

Deo, C. M. (1975): "A Functional Central Limit Theorem for Stationary Random Fields," *Annals of Probability*, 3(4), 573–739.

Dhaene, G., and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *Review of Economic Studies*, 82(3), 991–1030.

Driscoll, J. C., and A. C. Kraay (1998): "Consistent covariance matrix estimation with spatially dependent panel data," *The Review of Economics and Statistics*, 80(4), 549–560.

Druedahl, J., T. H. Jørgensen, and D. Kristensen (2016): "Estimating Dynamic Economic Models with Non-Parametric Heterogeneity," mimeo.

Durlauf, S. N., P. A. Johnson, and J. R. Temple (2005): "Growth econometrics," in *Handbook of Economic Growth*, ed. by P. Aghion, and S. N. Durlauf, vol. 1A, pp. 555–677, Amsterdam. North Holland.

Durlauf, S. N., and D. T. Quah (1999): "The new empirics of economic growth," in *Handbook of Macroeconomics*, ed. by J. B. Taylor, and M. Woodford, vol. 1A, pp. 235–308, New York. Elsevier, Chap. 4.

Feldstein, M., and C. Horioka (1980): "Domestic Savings and International Capital Flows," *Economic Journal*, 90, 314–329.

Forgy, E. W. (1965): "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, 21, 768–769.

Fotheringham, A. S., M. Charlton, and C. Brunsdon. (1997): "Measuring spatial variations in relationships with geographically weighted regression," in *Recent developments in spatial analysis*, ed. by M. M. Fischer, and A. Getis, pp. 60–82, Berlin. Springer-Verlag.

Garey, M., and D. Johnson (1979): *Computers and Intractability - A Guide to the Theory of NP-completeness*. W H Freeman & Co, first edn.

Guyon, X. (1995): *Random Fields on a Network*. Springer-Verlag, New York, Berlin, Heidelberg, second edn.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., and H. R. Moon (2010): "Panel data models with finite number of multiple equilibria," *Econometric Theory*, 26, 863–881.

Hallin, M., Z. Lu, and L. T. Tran (2004): "Local linear spatial regression," *Annals of Statistics*, 32(6), 2469–2500.

Hansen, P., and N. Mladenović (1997): "Variable neighborhood search," *Computers & Operational Research*, 24, 1097–1100.

Hiriart-Urruty, J. B. (1985): *Generalized differentiability, duality and optimization for problems dealing with differences of convex functions*vol. 256 of *Lecture Note in Economics and Mathematical Systems*, pp. 37–70. Springer-Verlag.

——— (1988): "From convex optimization to non-convex optimization. Part: Necessary and sufficient conditions for global optimality," in *Nonsmooth Optimization and Related Topics*, vol. 43 of *Ettore Majorana International Sciences*, pp. 219–239. Plenum Press.

Hoang, T. (1995): "D.C. Optimization: Theory, Methods and Algorithms," in *Handbook of Global Optimization*, ed. by R. Horst, and P. M. Pardalos, Nonconvex Optimization and Its Applications, pp. 149–209, Dordrecht, Boston, London. Kluwer Academic Publishers.

——— (2010): *Convex Analysis and Global Optimization*. Kluwer Academic Publishers, Dordrecht, Boston, London.

Horst, R., and T. Hoang (1993): *Global Optimization: Deterministic Approaches*. Springer-Verlag, Berlin, Heidelberg, New York, second edn.

Jansen, W. J. (1996): "Estimating saving-investment correlations: evidence for OECD countries based on an error correction model," *Journal of International Money and Finance*, 15(5), 749–781.

——— (1998): "Interpreting saving-investment correlations," *Open Economies Review*, 9, 205–217.

Jenish, N., and I. R. Prucha (2012): "On spatial processes and asymptotic inference under near-epoch dependence," *Journal of Econometrics*, 170(1), 178–190.

Júdice, J. J., M. Raydan, S. S. Rosa, and S. A. Santos (2008): "On the solution of the symmetric eigenvalue complementarity problem by the spectral projected gradient algorithm," *Numerical Algorithms*, 47, 391–407.

Kalai, A. T., A. Moitra, and G. Valiant (2010): "Efficiently learning mixtures of two Gaussians," in *Proceeding STOC '10 Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 553–562, New York. ACM.

Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

Ke, Z., J. Fan, and Y. Wu (2015): "Homogeneity pursuit," *Journal of the American Statistical Association*, 110, 175–194.

Kelejian, H. H., and I. R. Prucha (2007): "HAC estimation in a spatial framework," *Journal of Econometrics*, 140, 131154.

Klein, R. W., and R. C. Dubes (1989): "Experiments in projection and clustering by simulated annealing," *Pattern Recognition*, 22(2), 213–220.

Le Thi Hoai An (2014): "DC Programming and DCA," http://www.lita.univ-lorraine.fr/~lethi/index.php/dca.html.

Le Thi Hoai An, M. T. Belghiti, and Pham Dinh Tao (2007): "A new efficient algorithm based on DC programming and DCA for clustering," *Journal of Global Optimization*, 37, 593–608.

Le Thi Hoai An, Huynh Van Ngai, and Pham Dinh Tao (2012): "Exact penalty and error bounds in DC programming," *Journal of Global Optimization*, 52(3), 509–535.

Le Thi Hoai An, Le Hoai Minh, and Pham Dinh Tao (2014): "New and efficient DCA based algorithms for minimum sum-of-squares clustering," *Pattern Recognition*, 47, 388–401.

Le Thi Hoai An, and Pham Dinh Tao (2003): "Large-scale molecular optimization from distance matrices by a d.c. optimization approach," SIAM *Journal of Optimization*, 14(1), 77–114.

Lin, C., and S. Ng (2012): "Estimation of panel data models with parameter heterogeneity when group membership is unknown," *Journal of Econometric Methods*, 1(1), 42–55.

Lindsay, B. (1988): "Composite likelihood methods," *Contemporary Mathematics*, 80, 220–239.

Liu, Y., and X. Shen (2006): "Multicategory psi-Learning," *Journal of the American Statistical Association (Theory & Methods)*, 101(474), 500–509.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparamtric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40(2), 1132–1170.

Meliciani, V., and F. Peracchi (2006): "Convergence in per-capita GDP across European regions: A reappraisal," *Empirical Economics*, 31(3), 549–568.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953): "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21(6), 1087 – 1092.

Miniane, J. (2004): "A New Set of Measures on Capital Account Restrictions," *IMF Staff Papers*, 51(2), 276–308.

Nakhapetyan, B. S. (1988): "An approach to proving limit theorems for dependent random variables," *Theory of Probability and Its Applications*, 32(3), 535539.

Neaderhouser, C. C. (1980): "Convergence of blocks spins defined on random fields," *Journal of Statistical Physics*, 22, 673–684.

Nickell, S. (1981): "Biases in dynamic models with fixed effects," *Econometrica*, 49(6), 1417–1426.

Obstfeld, M., and K. Rogoff (2001): "The Six Major Puzzles in International Macroeconomics: Is There a Common Cause?," in *NBER Macroeconomics Annual 2000*, ed. by B. S. Bernanke, and K. Rogoff, pp. 339–412. MIT Press.

Pelgrin, F., and S. Schich (2008): "International capital mobility: What do national saving-investment dynamics tell us?," *Journal of International Money and Finance*, 27.

Pesaran, M. H. (2006): "Estimation and inference in large heterogeneous panels with a multifactor error structure," *Econometrica*, 74(4), 967–1012.

Pesaran, M. H., Y. Shin, and R. P. Smith (1999): "Pooled mean group estimation of dynamic heterogeneous panels," *Journal of the American Statistical Association*, 94(446), 621–634.

Pesaran, M. H., and R. P. Smith (1995): "Estimating long-run relationships from dynamic heterogenous panels," *Journal of Econometrics*, 68, 79–113.

Pesaran, M. H., R. P. Smith, and K. S. Im (1996): "Dynamic linear models for heterogenous panels," in *The Econometrics of Panel Data*, ed. by L. Mátyás, and P. Sevestre, Advanced Studies in Theoretical and Applied Econometrics, pp. 145–195, Dordrecht/Boston/London. Kluwer Academic Publishers.

Pham Dinh, T., and E. B. Souad (1988): "Duality in d.c. (difference of convex functions) optimization. Subgradient methods," in *Trends in Mathematical Optimization*, International Series of Numerical Mathematics 84, pp. 277–293, Basel. Birkhäuser.

Pham Dinh Tao, and Le Thi Hoai An (1997): "Convex analysis approach to d.c. programming: theory, algorithms and applications," *Acta Mathematica Vietnamica*, 22(1), 289–355.

——— (1998): "A d.c. optimization algorithm for solving the trust-region subproblem," SIAM *Journal of Optimization*, 8(2), 476–505.

——— (2014): "Recent advances in d.c. programming and DCA," in *Lecture Notes in Computer Science*, vol. 8342 of *Transactions on Computational Intelligence XIII*, pp. 1–37.

Phillips, P. C. B., and D. Sul (2007): "Transition modeling and econometric convergence tests," *Econometrica*, 75(6), 1771–1855.

——— (2009): "Economic transition and growth," *Journal of Applied Econometrics*, 24(7), 1153–1185.

Rand, W. M. (1971): "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66(336), 846–850.

Rao, B. L. S. P. (1987): *Asymptotic Theory of Statistical inference*. Wiley, New York.

Reid, N. (2013): "Aspects of likelihood inference," *Bernoulli*, 19(4), 1404–1418.

Rio, E. (1995): "The Functional Law of the Iterated Logarithm for Stationary Strongly Mixing Sequences," *The Annals of Probability*, 23(3), 1188–1203.

Rosenblatt, M. (1985): *Stationary Sequences and Random Fields*. Birkhauser, Boston.

Selim, S. Z., and K. Alsultan (1991): "A simulated annealing algorithm for the clustering problem," *Pattern Recognition*, 24(10), 1003–1008.

Su, L., Z. Shi, and P. C. B. Phillips (2016): "Identifying latent structures in panel data," *Econometrica*, 84(6), 2215–2264.

Sun, Y. (2005): "Estimation and inference in panel structure models," UCSD memeo.

Sunklodas, J. (2008): "On normal approximation for strongly mixing random fields," *Theory of Probability and Its Applications*, 52(1), 125–132.

Takahata, H. (1983): "On the rates in the central limit theorem for weakly dependent random fields," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64, 445–456.

Tan, P. N., M. Steinbach, and V. Kumar (2005): *Introduction to Data Mining*. Addison-Wesley, Upper Saddle River.

Taniguchi, M., J. Hirukawa, and K. Tamaki (2008): *Optimal Statistical Inference in Financial Engineering*. Chapman & Hall/CRC, Florida.

Thoai, N. V. (1999): "DC Programming: An overview," *Journal of Optimization Theory and Application*, 193(1), 1–43.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

Truong, Y. K., and C. J. Stone (1992): "Nonparametric Function Estimation Involving Time Series," *The Annals of Statistics*, 20(1), 77–97.

Varin, C., N. Reid, and D. Firth (2011): "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.

Vogt, M., and O. Linton (2017): "Classification of non-parametric regression functions in longitudinal data models," *Journal of the Royal Statistical Society: Series B*, 79(1), 5–27.

Wang, W., P. C. B. Phillips, and L. Su (2016): "Homogeneity pursuit in panel data models: theory and applications," Cowles Foundation Discussion Paper No. 2063.

Wu, J. (2012): *Advances in K-means Clustering: A Data Mining Thinking*. Springer, Heidelberg, New York, London.

Yuille, A. L., and A. Rangarajan (2003): "The concave-convex procedure," *Neural Computation*, 15(4), 915–936.