E. J. Ourso College of Business
Department of Economics

DEPARTMENT OF ECONOMICS WORKING PAPER SERIES

Identification and estimation in panel models with overspecified number of groups

Ruiqi Liu
State University of New York, Binghamton

Anton Schick
State University of New York, Binghamton

Zuofeng Shang
Indiana University-Purdue University Indianapolis

Yonghui Zhang
Renmin University of China

Qiankun Zhou
Louisiana State University

Working Paper 2018-03
http://faculty.bus.lsu.edu/papers/pap18_03.pdf

Department of Economics
Louisiana State University
Baton Rouge, LA 70803-6306
http://www.bus.lsu.edu/economics/

Identification and estimation in panel models with overspecified number of groups

Ruiqi Liu[1] **Anton Schick**[2] **Zuofeng Shang**[3] **Yonghui Zhang**[4] **Qiankun Zhou**[5]

This version **February 11, 2018**

Abstract

In this paper, we provide a simple approach to identify and estimate group structure in panel models by adapting the M-estimation method. We consider both linear and nonlinear panel models where the regression coefficients are heterogeneous across groups but homogeneous within a group and the group membership is unknown to researchers. The main result of the paper is that under certain assumptions, our approach is able to provide uniformly consistent group parameter estimator as long as the number of groups used in estimation is not smaller than the true number of groups. We also show that, with probability approaching one, our method can partition some true groups into further subgroups, but cannot mix individuals from different groups. When the true number of groups is used in estimation, all the individuals can be categorized correctly with probability approaching one, and we establish the limiting distribution for the estimates of the group parameters. In addition, we provide an information criterion to choose the number of group and established its consistency under some mild conditions. Monte Carlo simulations are conducted to examine the finite sample performance of our proposed method. Findings in the simulation confirm our theoretical results in the paper. Application to labor force participation also highlights the necessity to take into account of individual heterogeneity and group heterogeneity.

Keywords: Linear panel, Nonlinear panel, Group structure, Classification, Fixed effects, M-estimation, K-means algorithm

JEL classification: C01, C13, C23, C33

[1]Department of Mathematical Sciences, State University of New York, Binghamton, NY, 13902
[2]Department of Mathematical Sciences, State University of New York, Binghamton, NY, 13902
[3]Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, IN 46202.
[4]School of Economics, Renmin University of China, Beijing, China, 100872
[5]Department of Economics, Louisiana State University, Baton Rouge, LA 70803

1. Introduction

Panel data models are widely used in empirical research of both economics and finance. An important feature to use panel data is that it allows researchers to control individual-level heterogeneity. Unfortunately, most of these heterogeneity, however, is unobservable, e.g., willingness to pay for education, impact of economic policy, personal innate ability, etc. In practice, there are two opposite approaches to deal with this individual level heterogeneity. The first one is to completely ignore the heterogeneity among individuals by assuming common parameters across individuals, see, e.g., Lancaster (2002), Hahn and Newey (2004), Arellano and Bonhomme (2009). Indeed, this approach reduces the model complexity and facilitates statistical inference. However, this common parameters assumption might be too strong in practice and may lead to model misspecification: see, e.g., Hsiao (2014). Moreover, this assumption has also been found to be too restrictive in many empirical studies, see, for example, Hsiao and Tahmiscioglu (1997) and Lee et al. (1997), among others. The other approach is to allow cross-sectional slope heterogeneity, e.g., Hsiao and Pesaran (2008), Baltagi et al. (2008). This assumption helps avoid misspecification problem; however, it may lose latent connections between individuals and efficiency of estimation. To be more specific, if part of the individuals share a common parameter, it sacrifices this essential connection and leads to estimators with larger variance.

To allow such a possibility that part of the individuals shares a common parameter, a mild and reasonable assumption is to impose group structure in panels. Group structure in panels refers to the regression parameters that are the same within each group but differ across groups.[1] Recently, group structure in panels has received lots of attention in the literature both empirically and theoretically. To name a few, for linear model, Lin and Ng (2012) consider liner panel model with group structure on both intercept and slope. When there are only two groups and one regressor, they propose a threshold based estimation method to identify the latent group structure and show that the estimator is consistent. Under the same setup of Lin and Ng (2012), Sarafidis and Weber (2015) propose a modified k-means algorithm to determine the number of clusters and estimate parameters. Bonhomme and Manresa (2015) consider the linear panel data models with a latent group structure on the time-varying individual-specific effects and propose a group fixed effects estimator. The work of Bonhomme and Manresa (2015) has been extended to models with interactive fixed effects and nonlinear panel models by Ando and Bai (2016) and Bester and Hansen (2016), respectively. More recently, Su et al. (2016) propose a classifier Lasso (C-Lasso) penalized procedure to identify and estimate panels with group structure.

Following the work of Lin and Ng (2012) and Su et al. (2016), this paper proposes a simple and straightforward method to identify and estimate panels with group structure when the true number of groups and the membership are both unknown. The method we proposed can be

[1] There is another type of group pattern in the literature, Bonhomme and Manresa (2015) and Bester and Hansen (2016) consider the case when the individual-specific effects exhibit certain group pattern.

applied to both linear and nonlinear panels. Besides the simplicity, the proposed method has several advantages as follows.

First, the major theoretical contribution of this paper is that we show, under certain assumptions, the consistency of our proposed estimation is independent of the number of groups used as long as this number is not underestimated. The important practical implication of this result is that for estimation of the regression coefficients, one does not necessarily need to estimate the number of groups correctly as long as this number is not underestimated. The implication of this result is that a safe way in estimating the panel model with an unknown group structure is to set a slight large number of groups. This is of crucial importance to researchers since generally speaking, the number of groups in the data is usually unknown. We also show that, with probability approaching one, our method can partition some true group into further subgroups, but cannot misclassify individuals from different groups into the same group. When the true number of groups is used in estimation, all the individuals can be categorized correctly with probability approaching one.

Second, once the group membership is correctly identified and estimated, our proposed estimation performs similarly to the estimation based on true (or oracle) group membership. This oracle property allows one to combine exsiting estimation and inference technique with our method, for instance, for the classified group units, one can adapt the jackknife method in Hahn and Newey (2004) or Dhaene and Jochmans (2015) to reduce the bias for fixed effects estimation in both linear and nonlinear panels.

Finally, unlike the C-Lasso approach proposed by Su et al. (2016), which relies on the choice of tuning parameter, our approach is penalty free if the number of group is specified as a prior, which is a significant advantage for empirical application. It is well known in the literature that Lasso type methods are able to consistently select variables. However, the consistency of variable selection highly depends on the right choice of the tuning parameter (e.g., Chand (2012) and Kirkland et al. (2015)). Therefore, in empirical applications, the estimation results may be sensitive to the choice of tuning parameters, and how to choose the optimal tuning parameter in C-Lasso is still an open question. Consequently, it would be convenient to have a penalty free approach to identify the group structure in panels, and our proposed method serves this purpose.

The rest of the paper is organized as follows. In Section 2, we first introduce fixed effects model with unknown group structure, and then propose an estimation and classification procedure. Asymptotic properties of our estimator are established in Section 3. Section 4 carries out a set of Monte Carlo simulations to investigate the finite sample performance of our method. An application to labor force participation is provided in Section 5. Conclusion is made in Section 6. All mathematical derivation of main theorems and lemmas are provided in the appendix.

Notation: For any squared matrix A, let $\lambda_{\min}(A), \lambda_{\max}(A)$ be the smallest and largest eigenvalues of A. $\|A\|_2$ denotes the Frobenius norm as $\|A\|_2 = \sqrt{tr(AA')}$. For any positive integer k, define $[k] \in [N] := \{1, 2, ..., k\}$. $\xrightarrow{P}$ and $\xrightarrow{D}$ denote convergence in probability and in distribution,

respectively. Finally, $(N,T) \to \infty$ denotes N and T go to infinity jointly.

2. Panel Data Model with Misspecified Groups

Let Y_{it} be a real-valued observation and $X_{it} \in \mathbb{R}^p$ be a real vector of observed covariates, both collected on the ith individual at time t for $i \in [N] := \{1,2,\ldots,N\}$, $t \in [T] := \{1,2,\ldots,T\}$. Assume that the N individuals are actually belonging to G^0 underlying groups where G^0 is unknown. In particular, $G^0 = 1$ corresponds to the traditional fixed effect model without group structure (see Hahn and Newey (2004)). To identify group structure, a common practice is to predetermine the number of groups, denoted G, and classify the N individuals into G groups. In practice, correctly specifying G, i.e., $G = G_0$, is difficult due to the unobservability of group pattern. A more realistic way is to pick G relatively large so that $G \geq G^0$. Obviously, such misspecification brings more challenges into both theoretical study and practical applications. In this paper, we propose a method for identifying group patterns under this misspecification and investigate its asymptotic property.

For individual i, let $g_i \in [G] := \{1,2,\ldots,G\}$ denote the group membership variable, $\beta_{g_i} \in \mathbb{K} \subset \mathbb{R}^p$ denote the unobservable group-specific parameter, and $\alpha_i \in \mathbb{A} \subset \mathbb{R}$ denote the unobservable individual-specific parameter, where both $\mathbb{K}$ and $\mathbb{A}$ are compact subsets. If individuals i,j belong to the same group, then $\beta_{g_i} = \beta_{g_j}$, i.e., they share a common group parameter, but α_i and α_j might still be different due to individual-level heterogeneity. Let $\underline{\beta} = (\beta_1,\beta_2,\ldots,\beta_G) \in \mathbb{K}^G$ denote the tuple of G group-specific parameters, $\underline{\alpha} = (\alpha_1,\alpha_2,\ldots,\alpha_N) \in \mathbb{A}^N$ denote the N-vector of individual parameters, and $\gamma_N = (g_1,g_2,\ldots,g_N) \in \Gamma_N$ denote the N-vector of group membership variables, where $\Gamma_N = [G]^N$ is the class of all possible group assignments. Our aim is to estimate the triplet $\theta_N = (\underline{\beta},\underline{\alpha},\gamma_N)$ which can be performed through the following M-estimation:

$$\widehat{\theta}_N = \underset{\theta_N=(\underline{\beta},\underline{\alpha},\gamma_N)\in\Theta_N}{\arg\max} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\psi(X_{it},Y_{it},\beta_{g_i},\alpha_i), \tag{2.1}$$

where $\Theta_N = \mathbb{K}^G \times \mathbb{A}^N \times \Gamma_N$ denotes the entire parameter space, $\psi(X_{it},Y_{it},\beta_{g_i},\alpha_i)$ denotes the logarithm of the pseudo likelihood function of Y_{it} given X_{it} under parameters β_{g_i},α_i.

Unlike the C-Lasso approach proposed by Su et al. (2016), our M-estimation procedure (2.1) requires optimizing the objective function over the pre-regularized parameter space Θ_N where the parameters β_{g_i} therein naturally incorporate group constraint. This important feature avoids the delicate choice of penalty parameters as required by penalization-based methods. Various choices of the function ψ will be provided in the following Examples 1, 2 and 3.

Example 1. Linear panel model: $Y_{it} = \beta_{g_i}'X_{it} + \alpha_i + \epsilon_{it}$, $1 \leq i \leq N$, $1 \leq t \leq T$, where ϵ_{it}'s represent the idiosyncratic error. In this case, one chooses $\psi(x,y,\beta,\alpha) = -(y-\beta'x-\alpha)^2$.

Example 2. Binary choice panel model: $Y_{it} = 1(\beta_{g_i}'X_{it} + \alpha_i \geq \epsilon_{it})$, $1 \leq i \leq N$, $q \leq t \leq T$, where ϵ_{it}'s represent the idiosyncratic error with common distribution function F, and $1(\cdot)$ denotes the

indicator. In this case, we choose $\psi(x, y, \beta, \alpha) = y \log F(\beta' x + \alpha) + (1 - y) \log(1 - F(\beta' x + \alpha))$.

Example 3. Poisson panel model: Given X_{it} and under β_{g_i}, α_i, Y_{it} follows Poisson distribution with mean $\exp(\beta'_{g_i} X_{it} + \alpha_i)$. In this case, we can choose $\psi(x, y, \beta, \alpha)$ as the Poisson density function with mean $\exp(\beta' x + \alpha)$.

Due to the complex structure of the parameter space Θ_N, it is challenging to directly solve (2.1). Instead, we introduce an efficient iterative algorithm. Before that, let us introduce some notation to simplify writing. Define

$$H_i(\beta, \alpha) = E(\psi(X_{i1}, Y_{i1}, \beta, \alpha)), \quad \widehat{H}_i(\beta, \alpha) = \frac{1}{T} \sum_{t=1}^{T} \psi(X_{it}, Y_{it}, \beta, \alpha),$$

$$\Psi_N(\theta_N) = \Psi_N(\underline{\beta}, \underline{\alpha}, \gamma_N) = \frac{1}{N} \sum_{i=1}^{N} H_i(\beta_{g_i}, \alpha_i), \quad \widehat{\Psi}_N(\theta_N) = \widehat{\Psi}_N(\underline{\beta}, \underline{\alpha}, \gamma_N) = \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i}, \alpha_i).$$

Here H_i is the expected objective function for individual i, $\Psi_N(\theta_N)$ is the expected pooled objective function taking into account the group variables, $\widehat{H}_i$ and $\widehat{\Psi}_N(\theta_N)$ are their respective sample versions. Under these notation, (2.1) can be rewritten as

$$\widehat{\theta}_N = \arg\max_{\theta_N \in \Theta_N} \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i}, \alpha_i). \tag{2.2}$$

We propose the following iterative algorithm to solve (2.2):

(a) Choose initial estimators $(\underline{\beta}^{(0)}, \underline{\alpha}^{(0)})$.

(b) For each $i \in [N]$, in the sth iteration , find $g_i^{(s+1)} = \arg\max_{g \in [G]} \widehat{H}_i(\beta_g^{(s)}, \alpha_i^{(s)})$. Then set $\gamma_N^{(s+1)} = (g_1^{(s+1)}, \ldots, g_N^{(s+1)})$ and compute $(\underline{\beta}^{(s+1)}, \underline{\alpha}^{(s+1)}) = \arg\max_{(\underline{\beta}, \underline{\alpha}) \in \mathbb{K}^G \times \mathbb{A}^N} \widehat{\Psi}_N(\underline{\beta}, \underline{\alpha}, \gamma_N^{(s+1)})$.

(c) Repeat (b) until the solution converges.

The following simple procedure is recommended to choose the initial estimators. For each $i \in [N]$, let $\widehat{\beta}_i^{\mathrm{ML}}$'s and $\widehat{\alpha}_i^{\mathrm{ML}}$'s be the pseudo maximum likelihood estimators of β_i^0's and α_i^0's based on $\{X_{it}, Y_{it}\}_{t=1}^T$, i.e., $(\widehat{\beta}_i^{\mathrm{ML}}, \widehat{\alpha}_i^{\mathrm{ML}}) = \arg\max_{\beta \in \mathbb{K}, \alpha \in \mathbb{A}} \widehat{H}_i(\beta, \alpha)$. Firstly, we choose $\underline{\alpha}^{(0)} = (\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_N)$. Next, one applies the standard k-means algorithm with $k = G$ to $\widehat{\beta}_i^{\mathrm{ML}}$'s to get G clustering centers, say, $(\beta_1^{(0)}, \ldots, \beta_G^{(0)})$. Finally, let $\underline{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_G^{(0)})$ to be initial estimators for iteration. In Monte Carlo simulations, we find this initial estimator works well and leads to a very fast convergence.

3. Asymptotic theory

In this section, we prove several asymptotic results such as estimation consistency (Theorems 1 and 2) and classification consistency (Theorem 3). It is worthful to point out that such results even

hold under a misspecified G with $G \geq G^0$. As a byproduct, we provide a consistent procedure to determine the number of groups. Moreover, asymptotic normality for the estimators is established with a correctly specified G. Throughout this section, let $\theta_N^0 = (\underline{\beta}^0, \underline{\alpha}^0, \gamma_N^0)$ denote the true parameters under which the observations X_{it}, Y_{it} are generated, where $\underline{\beta}^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_{G^0}^0)$, $\underline{\alpha}^0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_N^0)$, and $\gamma_N^0 = (g_1^0, g_2^0, \ldots, g_N^0)$.

3.1. Estimation Consistency

The main result of this section is to show that the proposed M-estimation is consistent. Before stating our main theorems, let us introduce some technical conditions. To start, for each $g \in [G^0]$, we define $N_g = \sum_{i=1}^N I(g_i^0 = g)$, i.e., the true number of individuals from group g.

Assumption A1. *(a)* $\{X_{it}, Y_{it}\}_{t=1}^T$ *are mutually independent across* $i \in [N]$.

(b) For each $i \in [N]$, $\{X_{it}, Y_{it} : t \in [T]\}$ *is stationary and* α*-mixing with mixing coefficients* $\alpha_i(\cdot)$. *Moreover,* $\alpha(\tau) := \max_{1 \leq i \leq N} \alpha_i(\tau)$ *satisfies* $\alpha(\tau) \leq \exp(-C_0 \tau^{b_0})$, *where* $C_0 > 0$ *and* $b_0 > 0$ *are constants.*

(c) For each $i \geq 1$, $H_i(\beta, \alpha)$ *is uniquely maximized at* $(\beta_{g_i^0}^0, \alpha_i^0)$ *and, for any* $\epsilon > 0$,

$$\chi(\epsilon) := \inf_{i \geq 1} \inf_{\|\beta - \beta_{g_i^0}^0\|_2^2 + |\alpha - \alpha_i^0|^2 \geq \epsilon} [H_i(\beta_{g_i^0}^0, \alpha_i^0) - H_i(\beta, \alpha)] > 0.$$

(d) $d_0 \equiv \inf_{\tilde{g} \neq g} \|\beta_g^0 - \beta_{\tilde{g}}^0\|_2 > 0$.

(e) There is a non-negative function $Q(x, y)$ *such that for all* $(\beta, \alpha), (\check{\beta}, \check{\alpha}) \in \mathbb{K} \times \mathbb{A}$,

$$|\psi(x, y, \beta, \alpha) - \psi(x, y, \check{\beta}, \check{\alpha})| \leq Q(x, y)(\|\beta - \check{\beta}\|_2^2 + |\alpha - \check{\alpha}|^2)^{1/2},$$

and $|\psi(x, y, \beta, \alpha)| \leq Q(x, y)$ *for all* $(\beta, \alpha) \in \mathbb{K} \times \mathbb{A}$. *Furthermore, there exist* $b_1 \in (0, \infty]$ *and* $B_1 > 0$ *such that*

$$\sup_{i \in [N]} P(Q(X_{i1}, Y_{i1}) > v) \leq \exp\left(1 - (v/B_1)^{b_1}\right), \text{ for all } v > 0.$$

(f) For all $g \in [G^0]$, *there exists a positive constant* π_g *such that* $N_g/N \to \pi_g$ *as* $(N, T) \to \infty$.

Remark 1. *Assumption A1.(a) assumes cross-sectional independence among the individuals which is standard for panel data, e.g., Lee and Phillips (2015) and Su et al. (2016). Assumption A1.(b) imposes weak dependence for the observations along the time dimension with the level of dependence controlled by an exponential bound with parameter* b_0. *The stationarity assumption can be relaxed at cost of introducing more notation. A similar* $\alpha-$*mixing condition can be found in Su et al. (2016) and Bonhomme and Manresa (2015). Assumption A1.(c) is an identification condition, which can be verified case by case under certain mild conditions. The same condition was also assumed by Hahn and Newey (2004) and Hahn and Moon (2010). Assumption A1.(d)*

says that the pairwise differences between the group parameters are bounded from below. This condition is needed to guarantee the identification of the group parameters. Similar conditions were also assumed by Bonhomme and Manresa (2015) and Su et al. (2016). Assumption A1.(e) states that ψ is smooth satisfying certain exponential tail condition with decay rate of the tail probability characterized by b_1. When ψ is a bounded function, then we can choose $B_1 = 2\|\psi\|_\infty$ and $b_1 = \infty$. Similar tail condtions are also assumed by Bonhomme and Manresa (2015) for the error term. Compared with other conditions such as moment conditions, the exponential tail condition can lead to better convergence results and is still valid in commonly used models such as Examples 1, 2 and 3. Assumption A1.(f) excludes the groups with zero proportion. This condition is standard and necessary for panel models with finte number of groups, e.g., see Bonhomme and Manresa (2015) and Su et al. (2016).

Let $d = b_0 b_1/(b_0 b_1 + b_0 + b_1)$. Since b_0 and b_1 characterize the weak dependence of the observations and decay rate of the tail probability, respectively, as discussed in Remark 1, d can be viewed as a quantity jointly controlling both. A special case is $b_1 = \infty$, i.e., ψ is bounded, where we have $d = b_0/(1 + b_0) < 1$.

Assumption A2. $\log N = o(T^{\frac{d}{1+d}})$.

Remark 2. *For theoretical consideration, compared to the standard assumption on the rate of N and T in the literature where the ratio of T/N being a nonnzero constant(e.g., Hahn and Newey (2004) among others), Assumption A2 is a relatively weak condition, since Assumption A2 allows N to diverge exponentially faster than T, where the ratio of T/N goes to zero. Furthermore, Assumption A2 is also quite reasonable in practice, since most microeconomic datasets are with moderate large T and very large N.*

In order to prove the consistency of $\widehat{\theta}_N$, we introduce the following pseudo metric d_N on Θ_N. For any $\theta_N = (\underline{\beta}, \underline{\alpha}, \gamma_N), \widetilde{\theta}_N = (\underline{\widetilde{\beta}}, \underline{\widetilde{\alpha}}, \widetilde{\gamma}_N) \in \Theta_N$, define

$$d_N(\theta_N, \widetilde{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N}\left(\|\beta_{g_i} - \widetilde{\beta}_{\widetilde{g}_i}\|_2 + |\alpha_i - \widetilde{\alpha}_i|\right).$$

Specifically, $d_N(\theta_N, \widetilde{\theta}_N)$ measures the average discrepancy of (β_{g_i}, α_i)'s and $(\widetilde{\beta}_{\widetilde{g}_i}, \widetilde{\alpha}_i)$'s. Theorem 1 below proves consistency for $\widehat{\theta}_N$ under this pseudo metric.

Theorem 1. *Suppose $G \geq G^0$ and Assumptions A1 and A2 hold. Then $d_N(\widehat{\theta}_N, \theta_N^0) \xrightarrow{P} 0$ as $(N, T) \to \infty$.*

Theorem 1 establishes the consistency results for the parameter set θ_N including the slope coefficients and fixed effects. If the parameters of interest are slope coefficients, then it is easy to see that

$$\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 \xrightarrow{P} 0 \text{ as } (N, T) \to \infty.$$

In other words, the estimators of the group parameters are consistent only in an "average" sense, and it is possible that a small proportion of $\widehat{\beta}_{\widehat{g}_i}$'s may still be inconsistent. In Theorem 2, we can strengthen this result by showing that $\widehat{\beta}_{\widehat{g}_i}$'s are uniformly consistent without any additional assumption.

Theorem 2. *Suppose* $G \geq G^0$ *and Assumptions A1 and A2 hold. Then* $\sup_{1\leq i\leq N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 \xrightarrow{P} 0$ *as* $(N,T) \to \infty$.

Theorem 2 states that the estimators of all group parameters uniformly converge to the true group parameters. Again, both Theorems 1 and 2 only require $G \geq G^0$. If $G < G_0$, then the above results will be invalid since in this scenario, individuals from different groups need to be classified into the same group, and this will lead to inconsistency.

3.2. Detection of Group Structure among Individuals

Detection of group structure in panel data is a fundamentally important problem. The popular C-LASSO approach recently proposed by Su et al. (2016) requires the use of penalty for effectively classifying the individuals. In this section, we study our penalty-free grouping method and investigate its asymptotic property. Our theory and method are valid under $G \geq G^0$.

Recall that $\widehat{\gamma}_N = (\widehat{g}_1, \widehat{g}_2, \ldots, \widehat{g}_N)$ is the estimator of the group membership variables obtained in (2.2). Our grouping method is simply based on $\widehat{g}_i$'s as follows. For $g \in [G]$, define $\widehat{\mathcal{C}}_g = \{i \in [N] : \widehat{g}_i = g\}$, i.e., $\widehat{\mathcal{C}}_g$ is the collection of the individuals belonging to the g-th estimated group. Also define $\mathcal{C}^0_g = \{i \in [N] : g_i^0 = g\}$ for $g \in [G^0]$, i.e., $\mathcal{C}^0_g$ is the population analogy based on the true group membership variables. It is important to provide the conditions under which such a simple grouping method is valid, that is, for any $g \in [G]$, there exists a $\widetilde{g} \in [G^0]$ such that $\widehat{\mathcal{C}}_g \subseteq \mathcal{C}^0_{\widetilde{g}}$ with probability approaching one. Formal statement of this result is provided in Theorem 3. Such property implies that the individuals are correctly grouped.

To prove this result, we need stronger assumptions on the smoothness of ψ. In order to deal with partial derivatives of a multivariate function, we introduce the following multi-index notation. Let $\mathbf{k} = (k_1, k_2, \ldots, k_{p+1})$ denote a multi-index, where k_l's are non-negative integers. For any $\beta \in \mathbb{K} \subset \mathbb{R}^p$, denote $\beta = (\beta_{[1]}, \beta_{[2]}, \ldots, \beta_{[p]})$, where $\beta_{[l]}$ is the lth coordinate of β. Define the $\mathbf{k}$th order partial derivative of $\psi(x, y, \beta, \alpha)$ with respect to β, α as follows:

$$D^{\mathbf{k}}\psi(x,y,\beta,\alpha) = \frac{\partial^{|\mathbf{k}|}\psi(x,y,\beta,\alpha)}{\partial\beta_{[1]}^{k_1}\ldots\partial\beta_{[p]}^{k_p}\partial\alpha^{k_{p+1}}},$$

where $|\mathbf{k}| = k_1 + k_2 + \cdots + k_{p+1}$. Also denote the Hessian of ψ and H_i (with respect to β, α) by

$$\ddot{\psi}(x,y,\beta,\alpha) = \begin{pmatrix} \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta\partial\beta'} & \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta'\partial\alpha} \\ \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta\partial\alpha} & \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\alpha^2} \end{pmatrix}, \quad \ddot{H}_i(\beta,\alpha) = E(\ddot{\psi}(X_{i1}, Y_{i1}, \beta, \alpha)).$$

We require the following conditions on the partial derivatives of ψ and Hessian of H_i's. Let $\mathcal{B}_i = \{(\beta, \alpha) \in \mathbb{K} \times \mathbb{A} : \|\beta - \beta^0_{g_i^0}\|_2 + |\alpha - \alpha_i^0| \leq a_0\}$ for $i \geq 1$, and $\mathcal{B} = \cup_{i \geq 1} \mathcal{B}_i$.

Assumption A3. *(a) There exist some function $J(x, y)$, constant $a_0 > 0$ and integer $q_0 \geq 4$ such that for any $\boldsymbol{k}$ with $|\boldsymbol{k}| \leq 4$ and $(\beta, \alpha) \in \mathcal{B}$,*

$$|D^{\boldsymbol{k}}\psi(x, y, \beta, \alpha)| \leq J(x, y), \quad \sup_{i \geq 1} EJ^{q_0}(X_{i1}, Y_{i1}) < \infty.$$

(b) The Hessian matrices $\{\ddot{H}_i(\beta^0_{g_i^0}, \alpha_i^0),\ i \geq 1\}$ are negative definite with the largest eigenvalues uniformly bounded by zero, i.e., $\sup_{i \geq 1} \lambda_{\max}(\ddot{H}_i(\beta^0_{g_i^0}, \alpha_i^0)) < 0$.

Remark 3. *Assumption A3.(a) requires higher-order smoothness and finite q_0th moment on the objective function ψ to guarantee correct classification. Similar assumption has been made by Hahn and Newey (2004) and Su et al. (2016). Assumption A3.(b) requires the Hessian matrices of the expected objective function to be uniformly negative definite. It can be compared to the conditions on the Hessian matrices of the profiled objective function in Su et al. (2016).*

Below is the main result of this section which provides the classification consistency of our grouping method even under $G \geq G^0$.

Theorem 3. *Suppose $G \geq G^0$ and Assumptions A1-A3 hold. Then for each $g \in [G]$, there exists a $\widetilde{g} \in [G^0]$ such that $\lim_{(N,T) \to \infty} P\left(\widehat{\mathcal{C}}_g \subseteq \mathcal{C}^0_{\widetilde{g}}\right) = 1$.*

Remark 4. *Theorem 3 demonstrates that the proposed grouping method is valid under misspecification in the sense that, with probability approaching one, any grouped individuals asymptotically belong to a population group. This implies that any population group is either identical to a selected group or is partitioned into subgroups without any misclassification, which is possibly the best result one can expect under $G \geq G^0$. In the special case $G = G^0$, Theorem 3 naturally leads to classification consistency, i.e., upto a proper relabeling, with probability approaching one, $\widehat{\mathcal{C}}_g = C^0_g$ for any $g \in [G^0]$. Classification consistency was also established by Su et al. (2016) when $G = G^0$.*

Remark 5. *Intuitively, Theorem 3 implies that under Assumptions A1-A3 and if $G > G^0$, then with probability approaching one: (i) individuals from the same group may be divided into different subgroups; (ii) individuals from different groups can not be categorized into the same group.*

Remark 6. *The implication of Theorem 3 is that, since the true number of groups is unknown in practice, it is safe to use a relative large number of groups to classify the data and to obtain consistent estimation. Otherwise, if $G < G_0$, different from both Theorems 2 and 3, neither the estimation nor the classification is consistent.*

3.3. Determination of Number of Groups

Though our estimation and classification results are valid for misspecified G, it is still of interest to estimate the number of groups. In this section, we propose an efficient approach based on

penalization to address this problem and establish its theoretical validity. Let $\widehat{\theta}_N^G$ be the estimator in (2.2) using G as the number of groups. To estimate G^0, we define a penalized criterion function

$$PC(G) = \widehat{\Psi}_N(\widehat{\theta}_N^G) - \eta_{NT} G,$$

where $\eta_{NT} > 0$ is a penalty parameter that is used to exclude the extremely large and unlikely choice of G. We estimate G^0 based on following procedure:

$$\widehat{G} = \arg\max_{G \in [G_{\max}]} PC(G), \tag{3.1}$$

where $G_{\max}$ is a predetermined upper bound for G. The following theorem shows that $\widehat{G}$ is consistent, i.e., $\widehat{G} = G_0$ with high probability.

Theorem 4. *Suppose Assumptions A1 and A3 hold. If* $\log N = o(T^{\frac{d}{2(1+d)}})$, $\eta_{NT} T^{\frac{1}{2(1+d)}} \to \infty$ *and* $\eta_{NT} \to 0$, *then* $\lim_{(N,T)\to\infty} P(\widehat{G} = G^0) = 1$.

Note that the rate condition $\log N = o(T^{\frac{d}{2(1+d)}})$ in Theorem 4 is slightly stronger than Assumption A2, though both conditions allow N to grow exponentially with T. One can check by using the fact $d < 1$ that the choice $\eta_{NT} \asymp T^{-1/4}$ fulfills the rate conditions in Theorem 4.

3.4. Asymptotic Normality

In this section, we study the asymptotic normality of $\underline{\widehat{\beta}}$ under $G = G^0$. For this, we introduce the following "oracle" estimator $\underline{\widetilde{\beta}}$ of $\underline{\beta}$ when the true group assignment γ_N^0 is known. Specifically, let

$$\underline{\widetilde{\beta}} = \arg\max_{\underline{\beta} \in \mathbb{K}^{G^0}} \max_{\alpha_i \in \mathbb{A}} \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i^0}, \alpha_i).$$

Of course, $\underline{\widetilde{\beta}}$ is infeasible since γ_N^0 is practically unavailable. Interestingly, $\underline{\widetilde{\beta}}$ and $\underline{\widehat{\beta}}$ are in fact asymptotically equivalent as summarized in the following lemma.

Lemma 1. *Suppose* $G = G^0$ *and Assumptions A1-A3 hold. Under appropriate relabeling, it holds that* $\lim_{(N,T)\to\infty} P(\underline{\widehat{\beta}} = \underline{\widetilde{\beta}}) = 1$.

It can be seen from Lemma 1 that, to derive the asymptotic normality of $\underline{\widehat{\beta}}$, it is sufficient to derive the asymptotic normality of $\underline{\widetilde{\beta}}$. To achieve the latter, we make an additional Assumption A4. Before that, let us introduce some notation. Define

$$\begin{aligned}
&\rho_i = E^{-1}\Big(\frac{\partial^2 \psi}{\partial\alpha\partial\alpha}(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)\Big) E\Big(\frac{\partial^2 \psi}{\partial\beta\partial\alpha}(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)\Big),\\
&U_i(x,y,\beta,\alpha) = \frac{\partial\psi}{\partial\beta}(x,y,\beta,\alpha) - \rho_i \frac{\partial\psi}{\partial\alpha}(x,y,\beta,\alpha), \quad R_i(x,y,\beta,\alpha) = \frac{\partial\psi}{\partial\alpha}(x,y,\beta,\alpha),\\
&V_i(x,y,\beta,\alpha) = \frac{\partial U_i}{\partial\beta'}(x,y,\beta,\alpha), \quad \mathcal{I}_i = E(V_i(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)).
\end{aligned}$$

The above notation are standard in the literature of nonlinear panel models, e.g., Hahn and Newey (2004) and Arellano and Hahn (2007). To simplify writing, we introduce the following notation: $U_i^\alpha = \partial U_i/\partial\alpha$, $U_i^{\alpha\alpha} = \partial^2 U_i/\partial\alpha^2$, $U_{it} = U_i(X_{it}, Y_{it}, \beta_{g_i^0}^0, \alpha_i^0)$ and $U_{it}^\alpha = U_i^\alpha(X_{it}, Y_{it}, \beta_{g_i^0}^0, \alpha_i^0)$. We define R_{it}, R_{it}^α analogically. For each $i \geq 1$, let Λ_i denote the asymptotic covariance matrix of $\sum_{t=1}^T U_{it}/\sqrt{T}$ as $T \to \infty$, which has an expression

$$\Lambda_i = E(U_{it}U_{it}') + 2\sum_{t=1}^{\infty} E(U_{i1}U_{i,1+t}').$$

Convergence of the above series holds uniformly for i due to Assumptions A1 and A3.

Assumption A4. *(a) There exists a constant $0 < B_3 < 1$ such that*

$$B_3 \leq \inf_{i\geq 1} \lambda_{\min}(\Lambda_i) \leq \sup_{i\geq 1} \lambda_{\max}(\Lambda_i) \leq 1/B_3.$$

Moreover, for each $g \in [G^0]$, there exist positive definite matrices D_g and W_g such that

$$\lim_{N\to\infty} \sum_{i:g_i^0=g} \Lambda_i/N_g = D_g \text{ and } \lim_{N\to\infty} \sum_{i:g_i^0=g} \mathcal{I}_i/N_g = W_g.$$

(b) For each $g \in [G^0]$, there exists a vector $\Delta_g \in \mathbb{R}^p$ such that

$$\lim_{(N,T)\to\infty} \frac{1}{N_g T} \sum_{i:g_i^0=g} E\left\{ \left(\frac{\sum_{t=1}^T R_{it}}{\sqrt{T}E(R_{i1}^\alpha)} \right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T [U_{it}^\alpha - \frac{E(U_{i1}^{\alpha\alpha})}{2E(R_{it}^\alpha)} R_{it}] \right) \right\} = \Delta_g.$$

Remark 7. *Assumption A4.(a) requires that the eigenvalues of the covariance matrices Λ_i are bounded away from zero and infinity. Assumption A4.(b) is a common condition for handling asymptotic bias (see Hahn and Newey (2004) and Arellano and Hahn (2007) for similar conditions).*

As the main result of this section, Theorem 5 shows that the elements of $\underline{\widehat{\beta}}$ are asymptotically normally distributed.

Theorem 5. *Suppose $G = G^0$ and Assumptions A1, A3, A4 hold and $N/T \to \kappa$ for some $\kappa \geq 0$. Then under appropriate relabeling, as $(N,T) \to \infty$, for each $g \in [G^0]$,*

$$\sqrt{NT}(\widehat{\beta}_g - \beta_g^0) - \sqrt{N/T} W_g^{-1}\Delta_g \xrightarrow{D} N(0, \pi_g^{-1} W_g^{-1} D_g W_g^{-1}),$$

As a consequence, under appropriate relabeling, for each $g \in [G^0]$,

$$\sqrt{NT}(\widehat{\beta}_g - \beta_g^0) \xrightarrow{D} N(\sqrt{\kappa\pi_g^{-1}} W_g^{-1}\Delta_g, \pi_g^{-1} W_g^{-1} D_g W_g^{-1}),$$

Remark 8. *Theorem 5 is closely related to a number of work on panel data models with fixed effects. First, the asymptotic bias of $\widehat{\beta}_g$ is of order $O(\sqrt{N/T})$. For fixed effects model, Hahn and Newey (2004) derived the same order for the asymptotic bias of the fixed effects estimator.*

In particular, $\widehat{\beta}$ becomes asymptotically unbiased when $N = o(T)$. Second, when $\{X_{it}, Y_{it} : i \in [N], t \in [T]\}$ are independent, the bias term has an expression:

$$\sqrt{\kappa\pi_g^{-1}}W_g^{-1}\Delta_g = \sqrt{\kappa\pi_g^{-1}}W_g^{-1}\lim_{N\to\infty}\frac{1}{N_g}\sum_{i:g_i^0=g}\left(\frac{E(R_{i1}U_{i1}^{\alpha})}{E(R_{i1}^{\alpha})} - \frac{E(U_{i1}^{\alpha\alpha})E(|R_{i1}|^2)}{2E^2(R_{i1}^{\alpha})}\right).$$

For fixed effects model without group structure, i.e., $\pi_g = N_g/N = 1$, the above expression coincides with Arellano and Hahn (2007). When $T = o(N)$, the bias of $\widehat{\beta}_g$ tends to infinity. This issue can be resolved by adapting the jackknife procedure proposed by Hahn and Newey (2004) and Dhaene and Jochmans (2015) into our M-estimation procedure.

4. Monte Carlo Simulation

In order to evaluate the finite-sample performance of the classification and estimation procedure, following Su et al. (2016), we consider three data generating processes (DGPs) that cover both linear and nonlinear panels of static and dynamic models. Throughout these DGPs, we generate the fixed effect α_i and the idiosyncratic error u_{it} are I.I.D $N(0,1)$ across i and t. Moreover u_{it} is also independent of all regressors. We set the number of groups to be three (e.g., $G_0 = 3$), and the number of elements in each group are given by $N_1 = \lfloor 0.3N \rfloor$, $N_2 = \lfloor 0.3N \rfloor$ and $N_3 = N - N_1 - N_2$, where N is the total number of cross-sectional units and $\lfloor \cdot \rfloor$ denotes the integer part of "$\cdot$".

DGP 1 (Linear panel model): The data is generated as

$$y_{it} = \alpha_i + X_{it}'\beta_{g_i} + u_{it}, \tag{4.1}$$

where $X_{it} = (0.2\alpha_i + e_{it,1}, 0.2\alpha_i + e_{it,2})'$ and $e_{it,1}, e_{it,2} \sim$ I.I.D$N(0,1)$ across i, t and are independent of α_i. The true coefficients are $(0.4, 1.6)$, $(1,1)$, $(1.6, 0.4)$ for the three groups, respectively.

DGP 2 (Linear dynamic panel model): The data is generated as

$$y_{it} = \alpha_i(1 - \gamma_{g_i}) + \gamma_{g_i}y_{it-1} + X_{it}'\beta_{g_i} + u_{it}, \tag{4.2}$$

where X_{it} is a 2×1 vector of exogenous variables following two dimensional standard normal distribution. The true coefficients are $(0.4, 1.6, 1.6)$, $(0.6, 1, 1)$, $(0.8, 0.4, 0.4)$ for the three groups, respectively.

DGP 3 (Dynamic Panel Probit model):

$$y_{it} = 1(\gamma_{g_i}y_{it-1} + x_{it}\beta_{1,g_i} + \beta_{2,g_i} + \alpha_i > u_{it}), \tag{4.3}$$

where $x_{it} = 0.1\alpha_i + e_{it}$ with $e_{it} \sim$ I.I.D $N(0,1)$ and is independent of all other variables. The true coefficients are $(1, -1, 0.5)$, $(0.5, 0, -0.25)$, and $(0, 1, 0)$. It should be noted that γ_{g_i} and β_{1,g_i} are identifiable in this model, whereas β_{2,g_i} is unidentifiable because it is absorbed into the individual specified effects α_i.

For all the three DGPs, we consider the combinations of (N,T) with $N = (100, 200)$ and $T = (15, 25, 50)$. During the replication, the group membership is held fixed. The number of replication is set to be $R = 1000$. Since the goal of this paper is consistently estimate the regression coefficients, group membership and number of groups, we follow Su et al. (2016) to consider the following three criteria to examine the finite sample performance of the proposed M-estimation.

(1) We use the algorithm to determine the number of groups and then estimate the parameters through M-estimation procedure. For the estimation of parameters, the estimators are evaluated using the root mean squared error (RMSE) for each estimated group number G defined as[2]

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2^2}.$$

When $G = G_0$, we also consider another type RMSE similar to Su et al. (2016), defined as

$$\text{Group RMSE} = \sqrt{\frac{1}{G^0}\sum_{g=1}^{G^0} \|\widehat{\beta}_g - \beta^0_g\|_2^2}.$$

(2) Frequency or empirical percentage of selecting the number of groups for a given true number of groups ($G^0 = 3$ in our designs).

(3) Classification, which is the percentage of correct classification. It is calculated as the percentage of correct classification of the N units, calculated as $\sum_{i=1}^{N} I(\widehat{g}_i = g_i^0)/N$ under appropriate relabelling, averaged over the Monte Carlo replications.

Simulation results of DPGs 1-3 are summarized in Tables 1-3. Several interesting findings can be observed in Tables 1-4. First, Table 1 provides the RMSE for the proposed M-estimation using different number of groups with $G^0 = 3$. As we show in Theorem 2, our M-estimation procedure can lead to consistent estimator as long as $G \geq G^0$. From Table 1, we can observe that the RMSE decreases rapidly with the increase of either N or T, which is evident that the M-estimation is consistent. Moreover, as shown by Table 2, the group RMSE also decreases with the increase of N and T, and performs similarly to the oracle estimator (e.g., knowing the true group membership), which is consistent with our findings in Theorems 1-2. Second, Table 3 summarizes the accuracy of determination of number of groups using the criterion $PC(G)$ proposed in Section 3.3. We note that throughout all our designs of both linear and nonlinear panels, the determination of number of groups using the proposed algorithm is very accurate in the sense that the percentage of choosing the true number of groups is quite close to 1 with the increase of either N or T. Finally, Table 4 presents the simulation results of correct classification and group RMSE. For the correctness of classification, we can observe that with the increase of N and T, the algorithm we proposed is able to provide very accurate classification for both linear and nonlinear panels, which

[2]We don't compare the performance of C-LASSO by Su et al. (2016) with ours in the simulation. The main reason is that we are unclear of the choice of tuning parameters of C-LASSO when G is different from G_0.

is evident that the classification is consistent as shown in Theorem 3. In all, we can claim that the simulation results confirm our theoretical findings in this paper regarding the identification and estimation for panels with unknown group structure under misspecification.

TABLE 1
RMSE under $G = 3, 4, 5$ with $G^0 = 3$.

		DGP1			DGP2			DGP3		
N	T	3	4	5	3	4	5	3	4	5
100	15	0.190	0.217	0.234	0.141	0.166	0.184	0.296	0.512	0.571
100	25	0.113	0.140	0.157	0.078	0.104	0.118	0.190	0.256	0.277
100	50	0.036	0.068	0.083	0.035	0.052	0.064	0.119	0.173	0.182
200	15	0.188	0.214	0.233	0.136	0.158	0.174	0.286	0.381	0.399
200	25	0.109	0.136	0.153	0.076	0.098	0.112	0.185	0.240	0.261
200	50	0.032	0.065	0.080	0.027	0.048	0.060	0.116	0.162	0.176

TABLE 2
Bias and RMSE of DGPs 1-3 with $G^0 = 3$

		DGP1				DGP2				DGP3			
		Bias		GRMSE		Bias		GRMSE		Bias		GRMSE	
N	T	estimate	oracle	estimate	oracle	estimate	oracle	estimate	oracle	estimate	oracle	estimate	oracle
100	15	0.023	0.012	0.070	0.048	0.066	0.029	0.060	0.040	0.020	0.008	0.180	0.135
100	25	0.016	0.009	0.042	0.037	0.055	0.021	0.039	0.031	0.006	0.003	0.110	0.091
100	50	0.007	0.004	0.034	0.025	0.032	0.013	0.033	0.020	0.006	0.003	0.077	0.066
200	15	0.025	0.014	0.050	0.036	0.073	0.031	0.043	0.038	0.013	0.003	0.125	0.094
200	25	0.016	0.009	0.031	0.026	0.059	0.023	0.031	0.029	0.006	0.003	0.080	0.068
200	50	0.006	0.003	0.021	0.018	0.033	0.013	0.022	0.018	0.003	0.002	0.054	0.047

Note: "estimate" refers to estimation based estimated group membership, "oracle" refers to estimation using the true group membership, i.e., g_i^0.

TABLE 3
Percentage of choosing $G = 1, 2, \ldots, 5$ with $G^0 = 3$.

		DGP1					DGP2					DGP3				
N	T	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
100	15	0	0.004	**0.976**	0.02	0	0	0	**0.484**	0.366	0.15	0	0.081	**0.612**	0.262	0.045
100	25	0	0	**0.996**	0.004	0	0	0	**0.94**	0.058	0.002	0	0.058	**0.810**	0.128	0.004
100	50	0	0	**0.988**	0.012	0	0	0	**0.984**	0.016	0	0	0.007	**0.895**	0.098	0
200	15	0	0	**0.996**	0.004	0	0	0	**0.658**	0.246	0.096	0	0.063	**0.705**	0.221	0.011
200	25	0	0	**1**	0	0	0	0	**0.942**	0.058	0	0	0.011	**0.881**	0.106	0.002
200	50	0	0	**1**	0	0	0	0	**0.996**	0.004	0	0	0.002	**0.932**	0.066	0

TABLE 4
Percentage of correct classification with $G^0 = 3$

N	T	DGP1	DGP2	DGP3
100	15	0.902	0.926	0.883
100	25	0.934	0.978	0.949
100	50	0.966	0.989	0.979
200	15	0.903	0.932	0.883
200	25	0.967	0.980	0.949
200	50	0.995	0.998	0.980

5. Empirical Application

In this section, we apply the above estimation and classification method to study the women's labor force participation. The dataset comes from Panel Study of Income Dynamics (PSID) and contains 1461 married women for 10 calendar years 1979-1988. We consider the following dynamic panel binary choice model with fixed effects

$$y_{it} = 1\left(\alpha_i + \gamma_{g_i} y_{it-1} + x_{it}'\beta_{g_i} + \varepsilon_{it} > 0\right)$$

where y_{it} takes value one if woman i participate in period t and zero otherwise, α_i and δ_t represent individual specific effects and time effects, respectively. Other independent variables are $x_{it} = (\#\text{children}_{it}, \log\textit{income}_{it}, \textit{race}\ , \textit{eduwife}, \textit{agewife}$ and $\textit{agewife}^2)$, where $\#\text{children}_{it}$ is the number of children aged between 0 and 17, log*income* is the log of husband's labor income deflated by Consumer Price Index, *race* is an indicator function and takes value 1 for black, *eduwife* is the years of education of woman, *agewife* is the age of women (divided by 10) and $\textit{agewife}^2$ is squared age. Similar variables are also considered by Hyslop (1999) and Carro (2007).

Using the classification method in the previous section, we are able to divide the original sample into two groups, i.e., $G = 2$. The summary statistics for the original sample and two groups are provided in Table 5. From Table 5, we can observe that these two groups have quite distinct observations for some variables. For example, comparatively, individuals in group 2 have more children, lower percentage of black race and younger age, while, individuals in group 1 have more years of education. The difference in these two groups make a lot of difference in the estimation. Furthermore, based on the grouping, we can note that, on average, individuals from group 2 have much higher tendency to join the labor market comparing with individuals from group 1, e.g., the mean of labor force participation rate is 0.7898 for individuals from group 2 and is 0.3982 for group 1.

For the estimated group membership, we apply the fixed effects logit regression for each group and the whole sample. The estimation results are summarized in Table 6. Several interesting findings can be observed in the above estimation. First of all, we note that the effects of variables of previous year's labor force participation, husband's income and wife's age remain the same across the whole sample and two groups, even if the effects are quite different across different groups. Second, we note that race has negative effects on the labor force participation in the whole sample and group 1, while race is no longer significant in group 2. From the summary statistics, we note that group 2 has relative low percentage of race black, which indicates that effects of race is offset by other variables in this group. Finally, we observe that education of wife is not significant in the whole sample and group 1, while it is significant in group 2, which indicates that education indeed has positive significant effect on the labor force participation for individuals in group 2. In all, we can conclude that, in order to capture the individual heterogeneity and group heterogeneity, it would be of crucial importance to classify individuals into different groups instead of pooling all individuals in the same group.

TABLE 5
Summary statistics for the original sample and two groups

	Whole sample			Group1			Group2		
Variables	min	mean	max	min	mean	max	min	mean	max
yit	0	0.5743	1	0	0.3982	1	0	0.7898	1
#children	0	1.76	7	0	1.691	6	0	1.841	7
logincome	5.806	10.471	13.846	5.806	10.483	12.995	6.64	10.46	13.85
Race	0	0.1642	1	0	0.1788	1	0	0.1471	1
eduwife	5	12.05	18	5	12.13	18	5	11.95	18
agewife	1.8	3.557	6.3	1.9	3.671	6.2	1.8	3.424	6.3
agewife2	3.24	13.41	39.69	3.61	14.25	38.44	3.24	12.43	39.69

TABLE 6
Logit estimation for the whole sample and two groups

Variable	Whole sample	Group1	Group2
yit-1	2.0504***	2.1779***	0.8835***
	(0.0683)	(0.0841)	(0.1044)
#children	0.00001	-0.0395	-0.0915**
	(0.0295)	(0.0401)	(0.047)
logincome	-0.1933***	-0.2408***	-0.1787**
	(0.0472)	(0.0615)	(0.0814)
Race	-0.1735**	-0.1938*	0.1215
	(0.0842)	(0.1135)	(0.1412)
eduwife	0.0096	0.0088	0.0257**
	(0.0082)	(0.0118)	(0.0126)
agewife	1.2635***	1.8465***	2.6434***
	(0.2977)	(0.4133)	(0.4523)
agewife2	-0.161***	-0.2297***	-0.3097***
	(0.0386)	(0.0534)	(0.0598)

Note: *, **, *** refer to significance at 10%, 5% and 1% level, respectively.

6. Conclusion

In this paper, we consider the identification and estimation of panel models with group structure when the true number of group and the group membership are unknown to researchers. We propose an M-estimation procedure to estimate the parameters of interest and a information criterion function to determine the number of groups. The method we proposed is applicable to both linear and nonlinear panels. Asymptotic properties are established for the estimation and classification as well as the determination of number of groups. As a major theoretical contribution, we show that under certain assumptions, the consistency of our proposed estimation and classification procedure is independent of the number of groups used in estimation as long as this number is not underestimated. The important practical implication of this result is that for estimation on the slope coefficients, one does not necessarily need to estimate the number of groups consistently. Monte Carlo simulations are conducted to examine the finite sample properties of the proposed method, and simulation results confirm our theoretical findings. Application to labor force participation also highlights the necessity to take into account of individual heterogeneity and group heterogeneity.

Acknowledgement: Shang's research is supported by NSF DMS-1764280. Zhang's research is supported by the National Natural Science Foundation of China (Project No.71401166).

References

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Arellano, M. and Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, 77(2):489–536.

Arellano, M. and Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs*, 43:381.

Baltagi, B. H., Bresson, G., and Pirotte, A. (2008). To pool or not to pool? In Heidelberg, S. B., editor, *The econometrics of panel data*, pages 517–546.

Bester, C. A. and Hansen, C. B. (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Carro, J. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140(2):503–528.

Chand, S. (2012). On tuning parameter selection of lasso-type methods-a monte carlo study. pages 120–129. In Applied Sciences and Technology.

Dhaene, G. and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Science Business Media, LLC.

Hahn, J. and Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(03):863–881.

Hahn, J. and Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319.

Hsiao, C. (2014). *Analysis of panel data (No. 54)*. University Press, Cambridge.

Hsiao, C. and Pesaran, H. (2008). Random coefficient panel data models. In *The Econometrics of Panel Data*, pages 187–216. Springer Berlin Heidelberg.

Hsiao, C. and Tahmiscioglu, A. K. (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association*, 92(438):455–465.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, 67(6):1255–1294.

Kirkland, L. A., Kanfer, F., and Millard, S. (2015). Lasso tuning parameter selection. pages 49–56. Proceedings of the South African Statistical Association Conference.

Lancaster, T. (2002). Orthogonal parameters and panel data. *The Review of Economic Studies*, 69(3):647–666.

Lee, K., Pesaran, M., and Smith, R. (1997). Growth and convergence in a multi-country empirical stochastic growth model. *Journal of Applied Econometrics*, 12(2):357–392.

Lee, Y. and Phillips, P. C. (2015). Model selection in the presence of incidental parameters. *Journal of Econometrics*, 188(2):474–489.

Lin, C. C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42–55.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474.

Sarafidis, V. and Weber, N. (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, 77(2):274–296.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

APPENDIX

This supplement includes the mathematical proofs that are omitted in the main paper. We first provide the proof of the main theorems and then present the relevant lemmas that are used in proving these theorems.

A.1. Proofs of main theorems

This section contains proofs of main theorems. To proceed further, we introduce some notation as follows. For fixed $\beta \in \mathbb{K}$, define

$$\widehat{\alpha}_i(\beta) \equiv \arg\max_{\alpha \in \mathbb{A}} \widehat{H}_i(\beta, \alpha),$$

and for $\underline{\beta} \equiv (\beta_1, \beta_2, \ldots, \beta_G) \in \mathbb{K}^G$, define

$$\widehat{\gamma}_N(\underline{\beta}) \equiv \arg\max_{\gamma_N \in \Gamma_N} \max_{\underline{\alpha} \in \mathbb{A}^N} \widehat{\Psi}_i(\underline{\beta}, \underline{\alpha}),$$

with $(\widehat{g}_1(\underline{\beta}), \widehat{g}_2(\underline{\beta}), \ldots, \widehat{g}_N(\underline{\beta}))$ being the elements in $\widehat{\gamma}_N(\underline{\beta})$. Let

$$S_{NT} = \sup_{1 \leq \leq N} \sup_{\beta \in \mathbb{K}, \alpha \in \mathbb{A}} |\widehat{H}_i(\beta, \alpha) - H_i(\beta, \alpha)|,$$

then under certain assumptions, Lemma A.7 shows that $S_{NT} = o_P(1)$, which plays an important role in our proof. Moreover, definition of $\widehat{\Psi}$ suggests the following inequality:

$$\begin{aligned} \sup_{\theta_n \in \Theta_N} |\widehat{\Psi}(\theta_N) - \Psi(\theta_N)| &= \sup_{(\underline{\beta}, \underline{\alpha}, \gamma_N) \in \mathbb{K} \times \mathbb{A} \times \Gamma_N} \left| \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{H}_i(\beta_{g_i}, \alpha_i) - H_i(\beta_{g_i}, \alpha_i) \right) \right| \\ &\leq S_{NT}. \end{aligned}$$

To compare $(\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_G)$ and $(\beta_1^0, \beta_2^0, \ldots, \beta_{G^0}^0)$ with possibly $G \neq G^0$, we define map $\sigma : [G^0] \to [G]$ such that:

$$\sigma(g) = \arg\min_{\widetilde{g} \in [G]} \|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2, \text{ for } g \in [G^0]. \tag{A.1}$$

Proof of Theorem 1. First by definition of $\widehat{\theta}_N$, we have

$$\Psi_N(\theta_N^0) - S_{NT} = \widehat{\Psi}_N(\theta_N^0) \leq \widehat{\Psi}_N(\widehat{\theta}_N) \leq \Psi_N(\widehat{\theta}_N) + S_{NT} \leq \Psi_N(\theta_N^0) + S_{NT}.$$

Above inequality and Lemma A.7 shows that

$$\Psi_N(\widehat{\theta}_N) - \Psi_N(\theta_N^0) = o_P(1). \tag{A.2}$$

In the following, we will use contradiction argument to prove the result. Assume $d_N(\widehat{\theta}_N, \theta_N^0) \to 0$ fails to hold. There is a sub sequence (N_k, T_k) of (N, T) and $c_0 > 0$ such that $P\{d_{N_k}(\widehat{\theta}_{N_k}, \theta_{N_k}^0) \geq$

$c_0\} \geq c_0$ for all large enough k. W.L.O.G, we can assume $P\{d_N(\widehat{\theta}_N, \theta_N^0) \geq c_0\} \geq c_0$ for all large enough (N,T). Define event $A_{NT} = \{d_N(\widehat{\theta}_N, \theta_N^0) \geq c_0\}$. By Lemma A.4 and Lemma A.7, we have

$$\begin{aligned}\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N) &\geq [\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N)]I(A_{NT}) \\ &\geq \frac{c_0}{2R}\chi(c_0^2/8)I(A_{NT}).\end{aligned}$$

So let $\epsilon_0 = c_0\chi(c_0^2/8)/(4R)$ and it holds that

$$\begin{aligned}\liminf_{(N,T)\to\infty} P(\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N) \geq \epsilon_0) &\geq \liminf_{(N,T)\to\infty} P(\frac{c_0}{2R}\chi(c_0^2/8)I(A_{NT}) \geq \epsilon_0) \\ &= \liminf_{(N,T)\to\infty} P(A_{NT}) \geq c_0 > 0,\end{aligned}$$

which is a contradiction to (A.2). Proof completed. □

Lemma A.1. *Suppose Assumptions A1, A2 hold and $G \geq G^0$, then for each $g \in [G^0]$,*

$$\|\widehat{\beta}_{\sigma(g)} - \beta_g^0\|_2 = o_P(1).$$

Proof of Lemma A.1. By Theorem 1, we have for any $g \in [G^0]$

$$\begin{aligned}\|\widehat{\beta}_{\sigma(g)} - \beta_g^0\|_2 &= \frac{1}{\sum_{i=1}^N I(g = g_i^0)} \sum_{i=1}^N I(g = g_i^0)\|\widehat{\beta}_{\sigma(g_i^0)} - \beta_{g_i^0}^0\|_2 \\ &\quad \text{(By definition of } \sigma) \\ &\leq \frac{1}{N_g}\sum_{i=1}^N I(g = g_i^0)\|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2 \\ &\leq \frac{N}{NN_g}\sum_{i=1}^N \|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2 \\ &\quad \text{(By Assumption A1.(f))} \\ &\leq [\frac{1}{\pi_g} + o(1)]d_N(\widehat{\theta}_N, \theta_N^0) = o_P(1).\end{aligned}$$

□

Proof of Theorem 2 . First notice $\|\widehat{\beta}_{\sigma(g_i^0)} - \beta_{g_i^0}^0\|_2 \leq \|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2$ and $\widehat{H}_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) \leq \widehat{H}_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))$. By definition of S_{NT}, we have

$$\begin{aligned}H_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) - S_{NT} &= \widehat{H}_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) \\ &\leq \widehat{H}_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) \\ &= H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) + S_{NT}.\end{aligned}$$

So it follows that

$$\sup_{1\leq i\leq N}\left(H_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) - H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))\right) \leq 2S_{NT}. \tag{A.3}$$

Next we will prove the convergence result by contradiction. Assume $\sup_{1\le i\le N}\|\widehat{\beta}_{\widehat{g}_i}-\beta^0_{g^0_i}\|_2\to 0$ in probability fails to hold. Then W.L.O.G, there exist $\epsilon_0>0$ such that $P(\sup_{1\le i\le N}\|\widehat{\beta}_{\widehat{g}_i}-\beta^0_{g^0_i}\|_2\ge\epsilon_0)\ge\epsilon_0$ for (N,T) is large enough, otherwise we can take subsequence. Let event $A_{NT}=\{\sup_{1\le i\le N}\|\widehat{\beta}_{\widehat{g}_i}-\beta^0_{g^0_i}\|_2\ge\epsilon_0\}$. By Assumption A1.(c), we have

$$\begin{aligned}\sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\widehat{g}_i},\widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))\right) &\ge \sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\widehat{g}_i},\widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))\right)I(A_{NT})\\ &\ge \chi(\epsilon_0^2)I(A_{NT}).\end{aligned}\tag{A.4}$$

And by Lemma A.5, it follows that,

$$\begin{aligned}&\sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\sigma(g^0_i)},\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)}))\right)\\ \le\ &\sup_{1\le i\le N}B_2\left(\|\widehat{\beta}_{\sigma(g^0_i)}-\beta^0_{g^0_i}\|_2^2+|\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)})-\alpha^0_i|^2\right)^{1/2}\\ \le\ &\sup_{1\le i\le N}B_2\left(\|\widehat{\beta}_{\sigma(g^0_i)}-\beta^0_{g^0_i}\|_2+|\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)})-\alpha^0_i|\right).\end{aligned}\tag{A.5}$$

Next we will bound two terms in above inequality respectively. For first term in (A.5), by Lemma A.1 and direct examination, we have

$$\begin{aligned}\sup_{1\le i\le N}\|\widehat{\beta}_{\sigma(g^0_i)}-\beta_{g^0_i}\|_2 &= \sup_{1\le i\le N}\sum_{g=1}^{G^0}I(g^0_i=g)\|\widehat{\beta}_{\sigma(g^0_i)}-\beta_{g^0_i}\|_2\\ &= \sup_{1\le i\le N}\sum_{g=1}^{G^0}I(g^0_i=g)\|\widehat{\beta}_{\sigma(g)}-\beta_g\|_2\\ &\le \sum_{g=1}^{G^0}\|\widehat{\beta}_{\sigma(g)}-\beta_g\|_2=o_P(1).\end{aligned}\tag{A.6}$$

For second term, combing (A.6) and Lemma A.8, we have

$$\sup_{1\le i\le N}|\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)})-\alpha^0_i|=o_P(1).\tag{A.7}$$

Hence it follows from (A.5), (A.6) and (A.7) that

$$\sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\sigma(g^0_i)},\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)}))\right)=o_P(1)\tag{A.8}$$

Combining (A.3), (A.4), (A.8) and Lemma A.7, we have

$$\begin{aligned}&\chi(\epsilon_0^2)I(A_{NT})\\ \le\ &\sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\widehat{g}_i},\widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))\right)\\ \le\ &\sup_{1\le i\le N}\left(H_i(\beta^0_{g^0_i},\alpha^0_i)-H_i(\widehat{\beta}_{\sigma(g^0_i)},\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)}))\right)+\sup_{1\le i\le N}\left(H_i(\widehat{\beta}_{\sigma(g^0_i)},\widehat{\alpha}(\widehat{\beta}_{\sigma(g^0_i)}))-H_i(\widehat{\beta}_{\widehat{g}_i},\widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))\right)\\ =\ &o_P(1),\end{aligned}$$

which leads to a contradiction of Assumption A1.(c) and $P(A_{NT}) \geq c_0$ for all large enough (N,T). □

Before proving Theorem 3, we will introduce some notation. Define a η-neighborhood of θ_N^0 by $\mathcal{N}_\eta = \{\underline{\beta} \in \mathbb{K}^G : \max_{\widetilde{g}\in[G]} \min_{g\in[G^0]} \|\beta_{\widetilde{g}} - \beta_g^0\|_2 < \eta\}$. Also for each $\underline{\beta} \in \mathcal{N}_\eta$, we define sets $\mathcal{A}_\eta(\underline{\beta}, g) = \{\widetilde{g} \in [G] : \|\beta_{\widetilde{g}} - \beta_g^0\|_2 < \eta\} \subset [G]$, for all $g \in [G^0]$. Here $\mathcal{A}_\eta(\underline{\beta}, \cdot)$ plays a role of relabelling that connect labels in $[G^0]$ with labels in $[G]$.

Proof of Theorem 3. First we will prove the following claim: for sufficient small $\eta > 0$, with probability approaching one, $\{\mathcal{A}_\eta(\widehat{\underline{\beta}}, g), g \in [G^0]\}$ is a partition of $[G]$ and each $\mathcal{A}_\eta(\widehat{\underline{\beta}}, g)$ is non empty for all $g \in [G^0]$. To see this, by Theorem 2, we have with probability approaching one, $\widehat{\underline{\beta}} \in \mathcal{N}_\eta$. Therefore, by definition, each $\mathcal{A}_\eta(\widehat{\underline{\beta}}, g)$ is not empty. Moreover, the uniformly convergence in Theorem 2 also implies $\cup_{g=1}^{G^0} \mathcal{A}_\eta(\widehat{\underline{\beta}}, g) = [G]$. Now we remain to show that with probability approaching one, $\{\mathcal{A}_\eta(\widehat{\underline{\beta}}, g), g \in [G^0]\}$ is a partition of $[G]$. Let event $A_{NT} = \{\text{Exsits } g_{12} \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_1) \cap \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_2) \text{ for some distinct } g_1, g_2 \in [G^0]\}$. Assume above claim fails to hold, then there exists $\epsilon_0 > 0$ such that $P(A_{NT}) \geq \epsilon_0$ for all (N,T) is sufficiently large. On event $A_{NT} \cap \{\widehat{\underline{\beta}} \in \mathcal{N}_\eta\}$, there exist some $g_{12} \in [G], g_1, g_2 \in [G^0]$ such that $g_{12} \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_1) \cap \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_2)$. As a consequence, for $\eta < d_0/2$ and by Assumption A1.(d), it follows that,

$$d_0 \leq \|\beta_{g_1}^0 - \beta_{g_2}^0\|_2 \leq \|\widehat{\beta}_{g_{12}} - \beta_{g_1}^0\|_2 + \|\widehat{\beta}_{g_{12}} - \beta_{g_2}^0\|_2 < 2\eta < d_0.$$

which is a contradiction, since $P(A_{NT} \cap \{\widehat{\underline{\beta}} \in \mathcal{N}_\eta\}) \geq \epsilon_0/2$ for all (N,T) that is sufficiently large. Now we prove the claim.

Next by Lemma A.16 and the fact that with probability approaching one, $\widehat{\underline{\beta}} \in \mathcal{N}_\eta$, we have

$$\lim_{(N,T)\to\infty} P\left(\widehat{g}_i \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_i^0), \forall i \in [N]\right) = 1.$$

Finally, suppose $i, j \in \widehat{\mathcal{C}}_g$ for some $g \in [G]$, then $\widehat{g}_i = \widehat{g}_j = g$. From argument above, we can see, with probability approaching one, $g \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_i^0)$ and $g \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_j^0)$. Notice with probability approaching one, $\{\mathcal{A}_\eta(\widehat{\underline{\beta}}, g), g \in [G^0]\}$ is a partition of $[G]$, so it follows that $g_i^0 = g_j^0$. Now define $\widetilde{g} = g_i^0 = g_j^0 \in [G^0]$, then $i, j \in \mathcal{C}_{\widetilde{g}}$. Therefore, with probability approaching one, for each $g \in [G]$, there exist $\widetilde{g} \in [G^0]$, such that $\widehat{\mathcal{C}}_g \subset \mathcal{C}_{\widetilde{g}}$. □

Proof of Theorem 4. It suffices to show

$$\lim_{(N,T)\to\infty} P(PC(G) \leq PC(G^0)) = 1. \tag{A.9}$$

Now we consider two cases, namely $G < G^0$ and $G > G^0$.

Under fitting case, $G < G^0$: By direct examination and Lemma A.19 , for (N,T) is large enough, it follows that

$$\begin{aligned} PC(G^0) - PC(G) &= \widehat{\Psi}_N(\widehat{\theta}_N^{G^0}) - \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^{G}) + \widehat{\Psi}_N(\theta_N^0) - \eta_{NT}(G^0 - G) \\ &\geq \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^{G}) - \eta_{NT}(G^0 - G) \\ &= \Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N^{G}) - \eta_{NT}(G^0 - G) + o_p(1) \\ &\geq B_4/2 + o_P(1). \end{aligned} \tag{A.10}$$

Since $\eta_{NT} \to 0$, it follows from (A.10) that (A.9) holds for the case $G < G^0$.

Over fitting case, $G > G^0$: By Lemma A.21, it follows that

$$\begin{aligned} PC(G^0) - PC(G) &= \widehat{\Psi}_N(\widehat{\theta}_N^{G^0}) - \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^{G}) + \widehat{\Psi}_N(\theta_N^0) + \eta_{NT}(G - G^0) \\ &= O_p(T^{-\frac{1}{2(1+d)}}) + \eta_{NT}(G - G^0). \end{aligned} \tag{A.11}$$

Since $\eta_{NT} T^{\frac{1}{2(1+d)}} \to \infty$ and $G > G^0$, so (A.9) holds for the case when $G > G^0$. □

Proof of Lemma 1. Suppose $G = G^0$, then by Theorem 3, under appropriate relabelling, it follows that for each $g \in [G^0]$,

$$\lim_{(N,T)\to\infty} P(\widehat{\mathcal{C}}_g = \mathcal{C}_g) = 1.$$

And above equation implies that

$$\lim_{(N,T)\to\infty} P(\widehat{g}_i = g_i^0, \forall i \in [N]) = 1.$$

Since on the event $\{\widehat{g}_i = g_i^0, \forall i \in [N]\}$, we have $\underline{\widehat{\beta}} = \underline{\widetilde{\beta}}$. Therefore, we finish the proof. □

Lemma A.2. *Suppose Assumptions A1, A3, A4 hold and $N = O(T)$, then for all $g \in [G^0]$*

$$\sqrt{N_g T}(\widetilde{\beta}_g - \beta_g^0) - \sqrt{N_g/T} W_g^{-1} \Delta_g \xrightarrow{D} N(0, W_g^{-1} D_g W_g^{-1}).$$

Proof of Lemma A.2. By Lemma A.9, it follows that

$$\frac{1}{N_g T} \sum_{i:g_i^0=g}^{N} \sum_{t=1}^{T} V_i(X_{it}, Y_{it}, \beta_g^0, \alpha_i^0) = \frac{1}{N_g T} \sum_{i:g_i^0=g} \mathcal{I}_i + o_P(1). \tag{A.12}$$

By definition of $\widetilde{\beta}_g$, we find that

$$\frac{1}{N_g T} \sum_{i:g_i^0=g} \sum_{t=1}^{T} U_i(X_{it}, Y_{it}, \widetilde{\beta}_g, \widehat{\alpha}_i(\widetilde{\beta}_g)) = 0. \tag{A.13}$$

Apply the same argument in Lemma A.10, we can show that the following term will uniformly converge to its expectation for all $i \geq 1$,

$$\left(\frac{\sum_{t=1}^{T} R_{it}}{\sqrt{T} E(R_{i1}^{\alpha})}\right)\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} [U_{it}^{\alpha} - \frac{E(U_{i1}^{\alpha\alpha})}{2E(R_{it}^{\alpha})} R_{it}]\right). \tag{A.14}$$

Combining Assumption A4.(b), (A.12), (A.12) and Lemma A.18, we have following equation,

$$\begin{aligned}\frac{1}{\sqrt{N_gT}}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0) &= [\frac{1}{N_g}\sum_{i:g_i^0=g}\mathcal{I}_i]\sqrt{N_gT}(\widetilde{\beta}_g-\beta_g^0)\\ &\quad+\sqrt{N_g/T}[\Delta_g+o_P(1)]\\ &\quad+o_P(\sqrt{N_gT}\|\widetilde{\beta}_g-\beta_g^0\|_2)+o_P(\sqrt{N_g/T}). \quad \text{(A.15)}\end{aligned}$$

Since $N=O(T)$ by assumption, (A.15) shows the asymptotic distribution of $\widetilde{\beta}_g$ is contributed by

$$[\frac{1}{N_g}\sum_{i:g_i^0=g}\mathcal{I}_i]^{-1}\frac{1}{\sqrt{N_gT}}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0). \quad \text{(A.16)}$$

Next we will derive the asymptotic distribution of (A.16) by Lyapunov C.L.T and Cramer-Wold device. For any $u\in\mathbb{R}^p$, define $\zeta_{Ti}=\sum_{t=1}^{T}u'U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)/\sqrt{T}$. By Lemma A.9 and Lemma A.3, for some constant $C_u\geq 0$ depending on u, we have

$$\sum_{i:g_i^0=g}E(\zeta_i^3) \leq N_gC_u. \quad \text{(A.17)}$$

Direct examination implies

$$\begin{aligned}s_{N_g}^2 &\equiv \sum_{i:g_i^0=g}E(\zeta_i^2)\\ &= \sum_{i:g_i^0=g}u'E\Big(\sum_{t=1}^{T}U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)\sum_{t=1}^{T}U_i'(X_{it},Y_{it},\beta_g^0,\alpha_i^0)\Big)u/T \quad \text{(A.18)}\end{aligned}$$

Thanks to Lemma A.10 and Assumption A4.(a), we can show that

$$\lim_{(N,T)\to\infty}s_{N_g}^2/N_g=u'D_gu. \quad \text{(A.19)}$$

Combining (A.17), (A.18), (A.19) and Assumption A4.(a), we have

$$\begin{aligned}\lim_{(N,T)\to\infty}\frac{\sum_{i:g_i^0=g}E(\zeta_i^3)}{s_{N_g}^3} &= \lim_{(N,T)\to\infty}\frac{N_gC_u}{\Big(N_gu'D_gu\Big)^{3/2}}\\ &\leq \lim_{(N,T)\to\infty}\frac{N_gC_u}{\Big(N_gB_3\|u\|_2^2\Big)^{3/2}}=0. \quad \text{(A.20)}\end{aligned}$$

By (A.19), (A.20) and Lyapunov C.L.T., for any $u\in\mathbb{R}^p$, we have

$$\frac{u'}{\sqrt{N_gT}}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)\xrightarrow{D}N(0,u'D_gu).$$

Since u is arbitrary, by above equation and Assumption A4.(a), it follows that

$$\sqrt{N_g T}(\widetilde{\beta}_g - \beta_g^0) - \sqrt{N_g/T} W_g^{-1} \Delta_g \xrightarrow{D} N(0, W_g^{-1} D_g W_g^{-1}).$$

□

Proof of Theorem 5. The asymptotic distribution follows from asymptotic equivalence in Lemma 1 and Lemma A.2. □

A.2. Proof of relevant lemmas

This section contains proofs of relevant lemmas for proving the main theorems. Set

$$R = \sup_{\beta_1, \beta_2 \in \mathbb{K}, \alpha_1, \alpha_2 \in \mathbb{A}} \|\beta_1 - \beta_2\|_2 + |\alpha_1 - \alpha_2|.$$

Lemma A.3. *Under Assumption A1 and A3, there exists a non negative function $\widetilde{J}(x,y)$ such that for all $(\beta_1, \alpha_1), (\beta_2, \alpha_2) \in \mathcal{B}_i, i \geq 1$ and all $|\boldsymbol{k}| \leq 3$,*

$$\begin{aligned} |D^{\boldsymbol{k}}\psi(x,y,\beta,\alpha) - D^{\boldsymbol{k}}\psi(x,y,\beta,\alpha)| &\leq \widetilde{J}(x,y)(\|\beta_1 - \beta_2\|_2^2 + |\alpha_1 - \alpha_2|^2)^{1/2}, \\ |D^{\boldsymbol{k}}\psi(x,y,\beta,\alpha)| &\leq \widetilde{J}(x,y), \end{aligned}$$

and

$$\sup_{i \geq 1} E(\widetilde{J}^{q_0}(X_{i1}, Y_{i1})) < \infty.$$

Proof of Lemma A.3. This is a consequence of Assumption A1.(a) and mean value theorem. □

Lemma A.4. *Under Assumption A1, the inequality*

$$\inf_{d_N(\theta_N, \theta_N^0) \geq \epsilon} [\Psi_N(\theta_N^0) - \Psi_N(\theta_N)] \geq \frac{\epsilon}{2R} \chi(\epsilon^2/8)$$

holds for every $0 < \epsilon < R$.

Proof. Fix $0 < \epsilon < R$, let θ_N and θ_N^0 satisfy

$$d_N(\theta_N, \theta_N^0) = \frac{1}{N} \sum_{i=1}^{N} [\|\beta_{g_i} - \beta_{g_i^0}^0\|_2 + |\alpha_i - \alpha_i^0|] \geq \epsilon.$$

Then the cardinality k of the set of indices $A = \{i \in [N] : \|\beta_{g_i} - \beta_{g_i^0}^0\|_2 + |\alpha_i - \alpha_i^0| \geq \epsilon/2\}$ satisfies the inequality $(N-k)\epsilon/2 + kR \geq N d_N(\theta_N, \theta_N^0) \geq N\epsilon$. From this we conclude $k \geq N\epsilon/(2R - \epsilon) \geq N\epsilon/(2R)$. The inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and Assumption A1.(c) yield

$$[\Psi_N(\theta_N^0) - \Psi_N(\theta_N)] \geq \frac{1}{N} \sum_{i \in A} [H_i(\beta_{g_i^0}^0, \alpha_i^0) - H_i(\beta_{g_i}, \alpha_i)] \geq \frac{k}{N} \chi(\epsilon^2/8) \geq \frac{\epsilon}{2R} \chi(\epsilon^2/8).$$

By taking infimum on above inequality, the desired result follows. □

Lemma A.5. *Under Assumption A1, the following Lipchitz condition holds*

$$\sup_{i\geq 1}\sup_{(\beta_1,\alpha_1)\neq(\beta_2,\alpha_2)\in\mathbb{K}\times\mathbb{A}}\frac{|H_i(\beta_1,\alpha_1)-H_i(\beta_2,\alpha_2)|}{(\|\beta_1-\beta_2\|_2^2+|\alpha_1-\alpha_2|^2)^{1/2}}\leq B_2,$$

with $B_2=\int_0^\infty \exp\left(1-(t/B_1)^{b_1}\right)dt$ *if* $0<b_1<\infty$ *and* $B_2=B_1$ *if* $b_1=\infty$.

Proof of Lemma A.5. The desired result is valid when $d_1=\infty$. Now it suffices to show the case when $d_1<\infty$. In the view of Assumption A1.(e), we have

$$\begin{aligned}\sup_{i\geq 1}E\Big(Q(X_{i1},Y_{i1})\Big) &\leq \int_0^\infty \sup_{i\geq 1}P\Big(Q(X_{i1},Y_{i1})>t\Big)dt\\ &\leq \int_0^\infty \exp\Big(1-(t/B_1)^{b_1}\Big)dt<\infty.\end{aligned}$$

Combining above inequality with Jensen's inequality and A1.(e), it follows that

$$\begin{aligned}&\sup_{i\geq 1}\sup_{(\beta_1,\alpha_1)\neq(\beta_2,\alpha_2)\in\mathbb{K}\times\mathbb{A}}\frac{|H_i(\beta_1,\alpha_1)-H_i(\beta_2,\alpha_2)|}{(\|\beta_1-\beta_2\|_2^2+|\alpha_1-\alpha_2|^2)^{1/2}}\\ \leq\ &\sup_{i\geq 1}\sup_{(\beta_1,\alpha_1)\neq(\beta_2,\alpha_2)\in\mathbb{K}\times\mathbb{A}}E\left(\frac{|\psi(X_{i1},Y_{i1},\beta_1,\alpha_1)-\psi(X_{i1},Y_{i1},\beta_1,\alpha_1)|}{(\|\beta_1-\beta_2\|_2^2+|\alpha_1-\alpha_2|^2)^{1/2}}\right)\\ \leq\ &\sup_{i\geq 1}E\Big(Q(X_{i1},Y_{i1})\Big)\\ \leq\ &B_2.\end{aligned}$$

Proof completes. □

Lemma A.6. *Let* $Z_t, t\in[T]$ *be a sequence of stationary variables with zero mean such that* $\alpha(t)\leq\exp(-C_0t^{b_0})$ *and* $P(|Z_t|>z)\leq\exp^{1-(z/B_1)^{b_1}}$. *Furthermore, if* $E(Z_1^2)+2\sum_{t=1}^\infty E(Z_1Z_{1+t})\leq M$, *then*

$$P\left(\frac{1}{T}\left|\sum_{t=1}^T Z_t\right|>z\right)\leq 4\left(1+\frac{T^{\frac{1}{1+d}}z^2}{16M}\right)^{-T^{\frac{d}{1+d}}/2}+\frac{16L_1}{z}\exp\Big(-L_2T^{\frac{d}{1+d}}z^d\Big),\ \textit{for all } z>0,$$

where L_1, L_2 *are positive constants only relying on* b_0, b_1 *and* M *and* $d=b_0b_1/(b_0b_1+b_0+b_1)$.

Proof of Lemma A.6. Evaluating equation (1.7) in Merlevède et al. (2011) at $\lambda=Tz/4$ and $r=T^{\frac{d}{1+d}}$, we will finish the proof. □

Lemma A.7. *Suppose Assumption A1 holds, then there exist positive constants* C_3, C_4, C_5 *not relying on* i, T, N, z *such that, for all* $z>0$ *and* $T^{\frac{d}{1+d}}\geq 4(p+2)$,

$$\begin{aligned}&P\left(\sup_{(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}}\left|\widehat{H}_i(\beta,\alpha)-H_i(\beta,\alpha)\right|>6z\right)\\ \leq\ &C_4\left[1+\left(\frac{1}{z^{2(p+2)}}\right)\right]\left[\left(1+\frac{T^{\frac{1}{1+d}}z^2}{C_5}\right)^{-T^{\frac{d}{1+d}}/4}+\frac{2}{d}\exp\Big(-C_3T^{\frac{d}{1+d}}z^d\Big)+\frac{\exp(-C_3dT^{\frac{d}{1+d}}z^d)}{1-\exp(-C_3dT^{\frac{d}{1+d}}z^d)}\right].\end{aligned}$$

Furthermore if $\log N = o(T^{\frac{d}{1+d}})$, *then*

$$\sup_{1\le i\le N}\sup_{(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}}\left|\widehat{H}_i(\beta,\alpha)-H_i(\beta,\alpha)\right| = o_P(1),$$

and

$$\sup_{\theta_N\in\Theta_N}\left|\widehat{\Psi}_N(\theta_N)-\Psi_N(\theta_N)\right| = o_P(1).$$

Proof of Lemma A.7. Define $\tau=(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}$. For $\tau_1=(\beta_1,\alpha_1),\tau_2=(\beta_2,\alpha_2)\in\mathbb{K}\times\mathbb{A}$, define $l_{it}=\psi(X_{it},Y_{it},\beta_1,\alpha_1)-\psi(X_{it},Y_{it},\beta_2,\alpha_2)$. By Assumption A1.(e) and Lemma A.5, we have

$$|l_{it}-E(l_{it})|\le\Big(Q(X_{it},Y_{it})+B_2\Big)\|\tau_1-\tau_2\|_2. \tag{A.21}$$

Assumption A1.(e) implies

$$P\Big(Q(X_{i1},Y_{i1})+B_2>t\Big)\le\exp\Big(1-(t/C)^{b_1}\Big),\ \text{for all } t>0, \tag{A.22}$$

with $C=B_1+B_2$. Thanks to Assumption A1.(e), we can see

$$\begin{aligned}\sup_{i\ge1}E\Big(|Q(X_{i1},Y_{i1})+B_2|^3\Big) &= \int_0^\infty P(|Q(X_{i1},Y_{i1})+B_2|^3>t)dt\\ &\le \int_0^\infty\exp\Big(1-t^{b_1/3}/C^{b_1}\Big)dt<\infty\end{aligned}$$

In the view of Fan and Yao (2003)[Proposition 2.5] and (A.21), one concludes

$$\begin{aligned}|\mathrm{Cov}(l_{it},l_{is})| &\le 8\alpha_{[i]}^{1/3}(t-s)E^{2/3}\Big(|l_{i1}-E(l_{i1})|^3\Big)\\ &\le \alpha_{[i]}^{1/3}(t-s)E^{2/3}\Big(|Q(X_{i1},Y_{i1})+B_2|^3\Big)\|\tau_1-\tau_2\|_2^2\\ &\le \exp(-\frac{C_0}{3}|t-s|^{b_0})E^{2/3}\Big(|Q(X_{i1},Y_{i1})+B_2|^3\Big)\|\tau_1-\tau_2\|_2^2,\ \text{for all } t\ge s.\end{aligned}$$

Combing above, one finds that

$$\sup_{i\ge1}\Big[\mathrm{Cov}(l_{i1},l_{i1})+2\sum_{t>1}\mathrm{Cov}(l_{i1},l_{it})\Big] \le M\|\tau_1-\tau_2\|_2^2, \tag{A.23}$$

where

$$M=2\left([\int_0^\infty\exp\Big(1-t^{b_1/3}/C^{b_1}\Big)\,dt]^{2/3}\right)\sum_{t=0}^{\infty}\exp(-\frac{C_0}{3}t^{b_0})<\infty.$$

Combining (A.22) and (A.23) and apply Lemma A.6 with $Z_t=(l_{it}-E(l_{it}))\|\tau_1-\tau_2\|_2^{-1}$, one shows that

$$\begin{aligned}&P\left(\|\tau_1-\tau_2\|_2^{-1}\left|\frac{1}{T}\sum_{t=1}^{T}(l_{it}-E(l_{it}))\right|>z\right)\\ \le\ & 4\left(1+\frac{T^{\frac{1}{1+d}}z^2}{16M}\right)^{-T^{\frac{d}{1+d}}/2}+\frac{16C_{21}}{z}\exp\Big(-C_{31}T^{\frac{d}{1+d}}z^d\Big),\end{aligned} \tag{A.24}$$

where C_{21}, C_{31} are positive constants which are free of i, z and T. In the following, we will apply chaining argument to prove the concentration inequality.

For $\tau = (\beta, \alpha)$, define process $X_i(\tau) = \sum_{t=1}^{T} [\psi(X_{it}, Y_{it}, \beta, \alpha) - E(\psi(X_{it}, Y_{it}, \beta, \alpha))] / T$. For simplicity and without causing confusion, we write $X(\tau) = X_i(\tau)$ in rest of the proof. We construct a sequence of nested sets $T_0 \subset T_1 \subset \cdots \subset \mathbb{K} \times \mathbb{A}$ such that

$$\|\tau - \widetilde{\tau}\|_2 > 4^{-j},$$

for every distinct points $\tau, \widetilde{\tau} \in T_j$, and that each T_j is "maximal" in the sense that no additional points can be added to T_j without violating above inequality. Therefore, by construction, the cardinality $|T_j| \leq D4^{j(p+1)}$, where $D = \max(\text{diam}(\mathbb{K} \times \mathbb{A}), 2^{1/d})$. Now we link every element $\tau_{j+1} \in T_{j+1}$ to one and only one $\tau_j \in T_j$ such that

$$\|\tau_{j+1} - \tau_j\|_2 \leq 4^{-j}, \tag{A.25}$$

which can be done by the construction of T_{j+1} and T_j. Continue this process to link all points in T_j with points in T_{j-1}, and so on, to obtain for every $\tau_{j+1} \in T_{j+1}$ a chain $t\tau_{j+1}, \tau_j, \ldots, \tau_0$ that connects to a point in T_0. For each integer $k \geq 0$ and points $\tau_{k+1}, \widetilde{\tau}_{k+1}\ in T_{k+1}$, they link to elements $\tau_0, \widetilde{\tau} \in T_0$. Therefore, by triangular inequality, we have

$$\begin{aligned}
\left|[X(\tau_{k+1}) - X(\tau_0)] - [X(\widetilde{\tau}_{k+1}) - X(\widetilde{\tau}_0)]\right| &= \left|\sum_{j=0}^{k}[X(\tau_{j+1}) - X(\tau_j)] - \sum_{j=0}^{k}[X(\widetilde{\tau}_{j+1}) - X(\widetilde{\tau}_j)]\right| \\
&\leq 2\sum_{j=0}^{k}\max\left|X(\tau) - X(\widetilde{\tau})\right|,
\end{aligned}$$

where for each fixed j, the maximum is take over all links $(\tau, \widetilde{\tau})$ from T_{j+1} to T_j. Therefore, for fixed j, the maximum is take over at most $|T_{j+1}| \leq D4^{(j+1)(p+1)}$ elements, with each links satisfying $\|\tau, \widetilde{\tau}\|_2 \leq 4^{-j}$. Combing above inequality with (A.24) and (A.25), we have for every

$z > 0$

$$
\begin{aligned}
& P\Big(\sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}} \Big| [X(\tau_{k+1}) - X(\tau_0)] - [X(\widetilde{\tau}_{k+1}) - X(\widetilde{\tau}_0)] \Big| > 4z \Big) \\
\leq\ & P\Big(\sum_{j=0}^{k} \max \Big| X(\tau) - X(\widetilde{\tau}) \Big| > 2z \Big) \\
\leq\ & P\Big(\sum_{j=0}^{k} \max \Big| X(\tau) - X(\widetilde{\tau}) \Big| > \sum_{j=0}^{k} 2^{-j} z \Big) \\
\leq\ & \sum_{j=0}^{k} P\Big(\max \frac{\Big| X(\tau) - X(\widetilde{\tau}) \Big| \|\tau - \widetilde{\tau}\|_2}{\|\tau - \widetilde{\tau}\|_2} > 2^{-j} z \Big) \\
\leq\ & \sum_{j=0}^{k} P\Big(\max \frac{\Big| X(\tau) - X(\widetilde{\tau}) \Big| 4^{-j}}{\|\tau - \widetilde{\tau}\|_2} > 2^{-j} z \Big) \\
\leq\ & \sum_{j=0}^{k} P\Big(\max \frac{\Big| X(\tau) - X(\widetilde{\tau}) \Big|}{\|\tau - \widetilde{\tau}\|_2} > 2^{j} z \Big) \\
\leq\ & \sum_{j=0}^{k} P\Big(\max \frac{\Big| X(\tau) - X(\widetilde{\tau}) \Big|}{\|\tau - \widetilde{\tau}\|_2} > 2^{j} z \Big) D 4^{(j+1)(p+1)} \\
\leq\ & \sum_{j=0}^{\infty} 4^{p+1} D \Big[4 \Big(1 + \frac{T^{\frac{1}{1+d}} 4^j z^2}{16M} \Big)^{-T^{\frac{1}{1+d}}/2} + \frac{16 C_{21}}{2^j z} \exp\Big(- C_{31} 2^{jd} T^{\frac{d}{1+d}} z^d \Big) \Big] 4^{j(p+1)}. \\
\equiv\ & K_1 + K_2.
\end{aligned}
\tag{A.26}
$$

Now we are ready to establish a bound for two terms in (A.26). Direct examination shows

$$
\begin{aligned}
K_1 &= 4^{p+2}D\sum_{j=0}^{\infty}\left(1+\frac{T^{\frac{1}{1+d}}4^j z^2}{16M}\right)^{-T^{\frac{d}{1+d}}/2}4^{j(p+1)} \\
&= 4^{p+2}D\left(\frac{16M}{T^{\frac{1}{1+d}}z^2}\right)^{p+1}\sum_{j=0}^{\infty}\left(1+\frac{T^{\frac{1}{1+d}}4^j z^2}{16M}\right)^{-T^{\frac{d}{1+d}}/2}\left(\frac{T^{\frac{1}{1+d}}4^j z^2}{16M}\right)^{p+1} \\
&\quad (\text{Let } a_z = \frac{T^{\frac{1}{1+d}}z^2}{16M}) \\
&= 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\sum_{j=0}^{\infty}\left(1+4^j a_z\right)^{-T^{\frac{d}{1+d}}/2}\left(4^j a_z\right)^{p+1} \\
&\le 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\sum_{j=0}^{\infty}\left(1+4^j a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{2}+p+1} \\
&\quad (\text{Notice } T^{\frac{d}{1+d}} \ge 4(p+2)) \\
&\le 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\sum_{j=0}^{\infty}\left(1+4^j a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1} \\
&= 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\left[\left(1+4a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1}+\left(1+4a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1}+\sum_{j=2}^{\infty}\left(1+4^j a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1}\right] \\
&\quad (\text{Notice } 4^j \ge j \text{ for } j \ge 2) \\
&\le 2\times 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\left(1+a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1}+4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\sum_{j=2}^{\infty}\left(1+ja_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1} \\
&\le 2\times 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\left(1+a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}}+4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\int_1^{\infty}\left(1+xa_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}-1}dx \\
&= 2\times 4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\left(1+a_z\right)^{-\frac{T^{\frac{d}{1+d}}}{4}}+4^{p+2}D\left(\frac{1}{a_z}\right)^{p+1}\frac{4}{a_z T^{\frac{d}{1+d}}}\left(1+a_z\right)^{-\frac{-T^{\frac{d}{1+d}}}{4}} \\
&\le 4^{p+4}D\left[\left(\frac{1}{a_z}\right)^{p+1}+\left(\frac{1}{a_z}\right)^{p+2}\right]\left(1+a_z\right)^{-\frac{-T^{\frac{d}{1+d}}}{4}}. \qquad (A.27)
\end{aligned}
$$

A bound for K_2 can be derived as follows:

$$
\begin{aligned}
K_2 &= \sum_{j=0}^{\infty} \frac{4^{p+3}C_{21}D}{z} \exp\Big(- C_{31}2^{jd}T^{\frac{d}{1+d}}z^d\Big)4^{j(p+1)}2^{-j} \\
&\le \sum_{j=0}^{\infty} \frac{4^{p+3}C_{21}D}{z} \exp\Big(- C_{31}2^{jd}T^{\frac{d}{1+d}}z^d\Big)2^{2j(p+1)} \\
&= \sum_{j=0}^{\infty} \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \exp\Big(- C_{31}2^{jd}T^{\frac{d}{1+d}}z^d\Big)\Big(\frac{dC_{31}2^{jd}T^{\frac{d}{1+d}}z^d}{4(p+1)}\Big)^{2(p+1)/d} \\
&\quad \text{(Let } b_z = C_{31}T^{\frac{d}{1+d}}z^d\text{)} \\
&= \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \sum_{j=0}^{\infty} \exp\Big(- 2^{jd}b_z\Big)\Big(\frac{d2^{jd}b_z}{4(p+1)}\Big)^{2(p+1)/d} \\
&\quad \text{(Notice } x^y \le \exp(xy) \text{ for all } x>0, y>0\text{)} \\
&\le \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \sum_{j=0}^{\infty} \exp\Big(- 2^{jd}b_z/2\Big) \\
&\le \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \Big[\sum_{j:0\le j\le 1/d} \exp\Big(- 2^{jd}b_z/2\Big) + \sum_{j:j\ge 1/d} \exp\Big(- 2^{jd}b_z/2\Big)\Big] \\
&\quad \text{(Notice } 2^{jd} \ge jd \text{ for all } j \ge 1/d\text{)} \\
&\le \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \Big[\Big(\frac{1}{d}+1\Big)\exp\Big(- b_z/2\Big) + \sum_{j:j\ge 1/d} \exp\Big(- jdb_z/2\Big)\Big] \\
&\le \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \Big[\frac{2}{d}\exp\Big(- b_z/2\Big) + \sum_{j=1}^{\infty} \exp\Big(- jdb_z/2\Big)\Big] \\
&= \frac{4^{p+3}C_{21}D}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}} \Big(\frac{4(p+1)}{dC_{31}}\Big)^{2(p+1)/d} \Big[\frac{2}{d}\exp\Big(- b_z/2\Big) + \frac{\exp(-db_z/2)}{1-\exp(-db_z/2)}\Big]. \qquad \text{(A.28)}
\end{aligned}
$$

Therefore, in the view of (A.26), (A.27) and (A.28), we conclude that, when $T^{\frac{d}{1+d}} \ge 4(p+2)$, it follows that for all $z>0$

$$
\begin{aligned}
&P\Big(\sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}} \Big|[X(\tau_{k+1})-X(\tau_0)] - [X(\widetilde{\tau}_{k+1})-X(\widetilde{\tau}_0)]\Big| > 4z\Big) \\
\le\ & C_{41}\Big[\Big(\frac{1}{T^{\frac{1}{1+d}}z^2}\Big)^{p+1} + \Big(\frac{1}{T^{\frac{d}{1+d}}z^2}\Big)^{p+2} + \frac{1}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}}\Big] \\
&\times\Big[\Big(1+\frac{T^{\frac{1}{1+d}}z^2}{16M}\Big)^{-T^{\frac{d}{1+d}}/4} + \frac{2}{d}\exp\Big(- C_{31}T^{\frac{d}{1+d}}z^d/2\Big) + \frac{\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/2)}{1-\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/2)}\Big], \qquad \text{(A.29)}
\end{aligned}
$$

where $C_{41} = 4^{p+4}D(16M)^{p+1} + 4^{p+4}D(16M)^{p+2} + 4^{p+3}C_{21}D[4(p+1)d^{-1}C_{31}^{-1}]^{2(p+1)/d} < \infty$. By

triangle inequality, it follows that

$$\sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}}\Big|X(\tau_{k+1})-X(\widetilde{\tau}_{k+1})\Big| \leq \sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}}\Big|[X(\tau_{k+1})-X(\tau_0)]-[X(\widetilde{\tau}_{k+1})-X(\widetilde{\tau}_0)]\Big| + \sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}}\Big|X(\tau_0)-X(\widetilde{\tau}_0)\Big|, \tag{A.30}$$

where $\tau_0,\widetilde{\tau}_0$ are linked with $\tau_{k+1},\widetilde{\tau}_{k+1}$. The first term of the right side in (A.30) can be bounded in probability by (A.29). For the second term of the right side in (A.30), we notice, the maximum is taken over at most $|T_0|^2\leq D^2$ terms. Applying (A.24) to $X(\tau_0)-X(\widetilde{\tau}_0)$ and noticing $\|\tau_0-\widetilde{\tau}_0\|\leq D$, it follows that for all $z>0$

$$\begin{aligned}
&P\bigg(\sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}}\Big|X(\tau_0)-X(\widetilde{\tau}_0)\Big|>z\bigg)\\
\leq\ & D^2P\bigg(\frac{\Big|X(\tau_0)-X(\widetilde{\tau}_0)\Big|\|\tau_0-\widetilde{\tau}_0\|_2}{\|\tau_0-\widetilde{\tau}_0\|_2}>z\bigg)\\
\leq\ & D^2P\bigg(\frac{\Big|X(\tau_0)-X(\widetilde{\tau}_0)\Big|D}{\|\tau_0-\widetilde{\tau}_0\|_2}>z\bigg)\\
\leq\ & 4D^2\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16MD^2}\bigg)^{-T^{\frac{d}{1+d}}/2}+\frac{16C_{21}D^3}{z}\exp\bigg(-\frac{C_{31}T^{\frac{d}{1+d}}z^d}{D^d}\bigg)\\
\leq\ & 4D^2\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16MD^2}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{16C_{21}D^3}{z}\exp\bigg(-\frac{C_{31}T^{\frac{d}{1+d}}z^d}{D^d}\bigg).
\end{aligned}\tag{A.31}$$

Combining (A.29), (A.30), (A.31) and the fact $D^d\geq 2$, we can conclude that

$$\begin{aligned}
&P\bigg(\sup_{\tau,\widetilde{\tau}\in T_{k+1}}\Big|X(\tau)-X(\widetilde{\tau})\Big|>5z\bigg)\\
\leq\ & P\bigg(\sup_{\tau,\widetilde{\tau}\in T_{k+1}}\Big|[X(\tau_{k+1})-X(\tau_0)]-[X(\widetilde{\tau}_{k+1})-X(\widetilde{\tau}_0)]\Big|>4z\bigg)\\
&+P\bigg(\sup_{\tau_{k+1},\widetilde{\tau}_{k+1}\in T_{k+1}}\Big|X(\tau_0)-X(\widetilde{\tau}_0)\Big|>z\bigg)\\
\leq\ & C_{42}\bigg[1+\bigg(\frac{1}{T^{\frac{1}{1+d}}z^2}\bigg)^{p+1}+\bigg(\frac{1}{T^{\frac{d}{1+d}}z^2}\bigg)^{p+2}+\frac{1}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}}+\frac{1}{z}\bigg]\\
&\times\bigg[\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16MD^2}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{2}{d}\exp\bigg(-C_{31}T^{\frac{d}{1+d}}z^d/D^d\bigg)+\frac{\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/D^d)}{1-\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/D^d)}\bigg],
\end{aligned}\tag{A.32}$$

where $C_{42}=C_{41}+4D^2+16C_{21}D^3<\infty$. By continuity of process $X(\tau)$, we can show that with probability one

$$\sup_{\tau,\widetilde{\tau}\in\cup_{k=1}^{\infty}T_k}\Big|X(\tau)-X(\widetilde{\tau})\Big|=\sup_{\tau,\widetilde{\tau}\in\mathbb{K}\times\mathbb{A}}\Big|X(\tau)-X(\widetilde{\tau})\Big|.$$

Furthermore notice the right side of (A.32) is independent of k. Therefore, by monotone convergence theorem, it follows that

$$\begin{aligned} & P\bigg(\sup_{\tau,\widetilde{\tau}\in\mathbb{K}\times\mathbb{A}}\Big|X(\tau)-X(\widetilde{\tau})\Big|>5z\bigg) \\ = \; & P\bigg(\sup_{\tau,\widetilde{\tau}\in\cup_{k=1}^{\infty}T_k}\Big|X(\tau)-X(\widetilde{\tau})\Big|>5z\bigg) \\ = \; & \lim_{k\to\infty}P\bigg(\sup_{\tau,\widetilde{\tau}\in T_{k+1}}\Big|X(\tau)-X(\widetilde{\tau})\Big|>5z\bigg). \end{aligned} \tag{A.33}$$

Using similar argument in proving (A.24) and Assumption A1.(e), we have for any $\tau_0=(\beta_0,\alpha_0)\in\mathbb{K}\times\mathbb{A}$, it follows that

$$\begin{aligned} P\bigg(\Big|X(\tau_0)\Big|>z\bigg) \; &\le \; 4\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16M}\bigg)^{-T^{\frac{d}{1+d}}/2}+\frac{16C_{21}}{z}\exp\bigg(-C_{31}T^{\frac{d}{1+d}}z^d\bigg) \\ &\le \; 4\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16M}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{16C_{21}}{z}\exp\bigg(-C_{31}T^{\frac{d}{1+d}}z^d/D^d\bigg). \end{aligned} \tag{A.34}$$

As a result, in view of (A.32), (A.33) and (A.34), we conclude that for all $z>0$

$$\begin{aligned} & P\bigg(\sup_{\tau\in\mathbb{K}\times\mathbb{A}}\Big|X(\tau)\Big|>6z\bigg) \\ \le \; & P\bigg(\sup_{\tau\in\mathbb{K}\times\mathbb{A}}\Big|X(\tau)-X(\tau_0)\Big|>5z\bigg)+P\bigg(\Big|X(\tau_0)\Big|>z\bigg) \\ \le \; & P\bigg(\sup_{\tau,\widetilde{\tau}\in T_{k+1}}\Big|X(\tau)-X(\widetilde{\tau})\Big|>5z\bigg)+P\bigg(\Big|X(\tau_0)\Big|>z\bigg) \\ \le \; & C_{43}\bigg[1+\bigg(\frac{1}{T^{\frac{1}{1+d}}z^2}\bigg)^{p+1}+\bigg(\frac{1}{T^{\frac{d}{1+d}}z^2}\bigg)^{p+2}+\frac{1}{T^{\frac{2(p+1)}{1+d}}z^{2(p+1)+1}}+\frac{1}{z}\bigg] \\ & \times\bigg[\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{16MD^2}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{2}{d}\exp\bigg(-C_{31}T^{\frac{d}{1+d}}z^d/D^d\bigg)+\frac{\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/D^d)}{1-\exp(-C_{31}dT^{\frac{d}{1+d}}z^d/D^d)}\bigg] \\ \le \; & C_{43}\bigg[1+\frac{1}{z^{2(p+1)}}+\frac{1}{z^{2(p+2)}}+\frac{1}{z^{2(p+1)+1}}+\frac{1}{z}\bigg] \\ & \times\bigg[\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{C_5}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{2}{d}\exp\bigg(-C_3T^{\frac{d}{1+d}}z^d\bigg)+\frac{\exp(-C_3dT^{\frac{d}{1+d}}z^d)}{1-\exp(-C_3dT^{\frac{d}{1+d}}z^d)}\bigg] \\ \le \; & C_4\bigg[1+\frac{1}{z^{2(p+2)}}\bigg] \\ & \times\bigg[\bigg(1+\frac{T^{\frac{1}{1+d}}z^2}{C_5}\bigg)^{-T^{\frac{d}{1+d}}/4}+\frac{2}{d}\exp\bigg(-C_3T^{\frac{d}{1+d}}z^d\bigg)+\frac{\exp(-C_3dT^{\frac{d}{1+d}}z^d)}{1-\exp(-C_3dT^{\frac{d}{1+d}}z^d)}\bigg] \end{aligned} \tag{A.35}$$

where $C_{43} = C_{42}+4+16C_{21} < \infty$, $C_3 = C_{31}/D^d$, $C_4 = 5C_{43}$ and $C_5 = 16MD^2$. As a consequence, under conditions $\log(N) = o(T^{\frac{d}{1+d}})$, it follows that

$$\begin{aligned}
& P\left(\sup_{\theta_N \in \Theta_N} \left| \widehat{\Psi}_N(\theta_N) - \Psi_N(\theta_N) \right| > 6z \right) \\
\leq & \left(\sup_{\theta_N \in \Theta_N} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{H}_i(\beta_{g_i}, \alpha_i) - H_i(\beta_{g_i}, \alpha_i) \right] \right| > 6z \right) \\
\leq & P\left(\sup_{1 \leq i \leq N} \sup_{(\beta,\alpha) \in \mathbb{K} \times \mathbb{A}} \left| \widehat{H}_i(\beta, \alpha) - H_i(\beta, \alpha) \right| > 6z \right) \\
\leq & N \max_{1 \leq i \leq N} P\left(\sup_{(\beta,\alpha) \in \mathbb{K} \times \mathbb{A}} \left| \widehat{H}_i(\beta, \alpha) - H_i(\beta, \alpha) \right| > 6z \right) \\
= & N \max_{1 \leq i \leq N} P\left(\sup_{\tau \in \mathbb{K} \times \mathbb{A}} \left| X_i(\tau) \right| > 6z \right) \to 0.
\end{aligned}$$

□

Lemma A.8. *Suppose Assumptions A1 and A2 hold, then*

$$\sup_{1 \leq i \leq N} |\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha_i^0| = o_P(1).$$

Furthermore, let $\{\beta_{Ti}\}, i \in [N]$ *be a random sequence such that* $\sup_{1 \leq i \leq N} \|\beta_{Ti} - \beta^0_{g_i^0}\|_2 = o_P(1)$, *then*

$$\sup_{1 \leq i \leq N} |\widehat{\alpha}_i(\beta_{Ti}) - \alpha_i^0| = o_P(1).$$

Proof of Lemma A.8. For first convergence, by definition of S_{NT} and Assumption A1.(c)

$$\begin{aligned}
0 \geq & \widehat{H}_i(\beta^0_{g_i^0}, \alpha_i^0) - \widehat{H}_i(\beta^0_{g_i^0}, \widehat{\alpha}_i(\beta^0_{g_i^0})) \\
\geq & H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta^0_{g_i^0}, \widehat{\alpha}_i(\beta^0_{g_i^0})) - 2S_{NT} \\
\geq & \chi(|\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha_i^0|^2) - 2S_{NT}.
\end{aligned}$$

So by above inequality and the fact $\chi(\epsilon)$ is non decreasing, it follows from Lemma A.7 that

$$\begin{aligned}
\chi(\sup_{1 \leq i \leq N} |\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha_i^0|^2) = & \sup_{1 \leq i \leq N} \chi(|\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha_i^0|^2) \\
\leq & 2S_{NT} = o_P(1).
\end{aligned}$$

Noticing $\chi(0) = 0$ and $\chi(\epsilon) > 0$ for all $\epsilon > 0$, above inequality implies

$$\sup_{1 \leq i \leq N} |\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha_i^0|^2 = o_P(1), \tag{A.36}$$

which is the first convergence.

By Lemma A.5 and Lemma A.7, it follows that

$$\begin{aligned}
\sup_{1 \leq i \leq N} \sup_{\alpha \in \mathbb{A}} |\widehat{H}_i(\beta_{Ti}, \alpha) - \widehat{H}_i(\beta^0_{g_i^0}, \alpha)| \leq & \sup_{1 \leq i \leq N} \sup_{\alpha \in \mathbb{A}} |H_i(\beta_{Ti}, \alpha) - H_i(\beta^0_{g_i^0}, \alpha)| + 2S_{NT} \\
\leq & B_2 \sup_{1 \leq i \leq N} \|\beta_{Ti} - \beta^0_{g_i^0}\|_2 + 2S_{NT}.
\end{aligned} \tag{A.37}$$

By (A.37), one finds that

$$\begin{aligned}
& B_2 \sup_{1\le i\le N} \|\beta_{Ti}-\beta^0_{g_i^0}\|_2 + 2S_{NT} \\
\ge\ & \sup_{\alpha\in\mathbb{A}} |\widehat{H}_i(\beta_{Ti},\alpha)-\widehat{H}_i(\beta^0_{g_i^0},\alpha)| \\
\ge\ & |\widehat{H}_i(\beta_{Ti},\widehat{\alpha}_i(\beta_{Ti}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))| \\
\ge\ & |\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| - |\widehat{H}_i(\beta_{Ti},\widehat{\alpha}_i(\beta_{Ti}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| \\
& \text{(By definition of } S_{NT} \text{ and definition of } \widehat{\alpha}_i(\beta)) \\
\ge\ & |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| - 2S_{NT} - |\sup_{\alpha\in\mathbb{A}}\widehat{H}_i(\beta_{Ti},\alpha)-\sup_{\alpha\in\mathbb{A}}\widehat{H}_i(\beta^0_{g_i^0},\alpha)| \\
\ge\ & |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| - 2S_{NT} - \sup_{\alpha\in\mathbb{A}}|\widehat{H}_i(\beta_T,\alpha)-\widehat{H}_i(\beta^0_{g_i^0},\alpha)| \\
& \text{(By } (A.37)) \\
\ge\ & |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| - 4S_{NT} - B_2\sup_{1\le i\le N}\|\beta_{Ti}-\beta^0_{g_i^0}\|_2. \qquad \text{(A.38)}
\end{aligned}$$

Taking supremum on both sides of (A.38), we find that

$$\sup_{1\le i\le N} |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| \le 2B_2\sup_{1\le i\le N}\|\beta_{Ti}-\beta^0_{g_i^0}\|_2 + 6S_{NT}. \qquad \text{(A.39)}$$

In the view of (A.36), (A.39), Lemma A.5 and Lemma A.7, we have

$$\begin{aligned}
& \sup_{1\le i\le N} |H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))| \\
\le\ & \sup_{1\le i\le N} |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{Ti}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| + \sup_{1\le i\le N}|H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))| \\
\le\ & 2B_2\sup_{1\le i\le N}\|\beta_{Ti}-\beta^0_{g_i^0}\|_2 + 6S_{NT}. + B_2\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0| \\
=\ & o_P(1).
\end{aligned}$$

By above inequality and using similar argument in proving (A.36), it follows that

$$\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{Ti})-\alpha_i^0| = o_P(1),$$

which is the second result. □

Lemma A.9. *For each $i \ge 1$, let $\{\zeta_{it}, t\in[T]\}$ be a stationary process with mean 0 and α-mixing coefficient $\alpha(t)\le \exp(-C_0 t^{b_0})$ for all $t\ge 1$. Further more if $\sup_i E(|\zeta_{it}|^{q_0})\le K$, for some positive constant K, then*

$$E(|\sum_{t=1}^{T}\zeta_{it}|^{q_0}) \le CT^{q_0/2}, \text{ for all } i\ge 1,$$

and

$$P(\sup_{1\le i\le N}|\sum_{t=1}^{T}\zeta_{it}/T|>\epsilon) \le C\epsilon^{-q_0}NT^{-q_0/2},$$

where $C > 0$ is a constant only relying on c, K. As a consequence, if $N = o(T^{q_0/2})$, then

$$\sup_{1\le i\le N} |\sum_{t=1}^{T} \zeta_{it}/T| = o_P(1)$$

Proof of Lemma A.9. By Fan and Yao (2003, Theorem 2.17 and Proposition 2.7), we have

$$E(|\sum_{t=1}^{T} \zeta_{it}|^{q_0}) \le CT^{q_0/2}, \text{ for all } i \ge 1, \tag{A.40}$$

where $C > 0$ is a constant only relying on c, K, q_0. By Chebyshev's inequality and (A.40), for any $\epsilon > 0$, it follows that

$$\begin{aligned} P(\sup_{1\le i\le N} |\sum_{t=1}^{T} \zeta_{it}/T| > \epsilon) &\le N\frac{E(|\sum_{t=1}^{T} \zeta_{it}|^{q_0})}{\epsilon_0^q T_0^q} \\ &\le \frac{CN}{\epsilon_0^q T^{q_0/2}}. \end{aligned}$$

□

Recall following terms defined in Section 3.4:

$$\begin{aligned} \rho_i &= E(\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{i1}, Y_{i1}, \beta^0_{g_i^0}, \alpha_i^0))^{-1} E(\frac{\partial^2\psi}{\partial\beta\partial\alpha}(X_{i1}, Y_{i1}, \beta^0_{g_i^0}, \alpha_i^0)), \\ U_i(x, y, \beta, \alpha) &= \frac{\partial\psi}{\partial\beta}(x, y, \beta, \alpha) - \rho_i\frac{\partial\psi}{\partial\alpha}(x, y, \beta, \alpha), \\ \Lambda_i &= E(U_{it}U_{it}') + 2\sum_{t=1}^{\infty} E(U_{i1}U_{i,1+t}'), \text{ with } U_{it} = U_i(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0). \end{aligned}$$

Lemma A.10. *Under Assumption A1 and Assumption A3, we have*

$$\lim_{T\to\infty}\sup_{i\ge1}\left\| \frac{E\left(\sum_{t=1}^{T} U_i(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)\sum_{t=1}^{T} U_i'(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)\right)}{T} - \Lambda_i\right\|_2 = 0.$$

Proof of Lemma A.10. For $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$, define $\zeta_{it} = u'U_i(X_{it}, Y_{it}, \beta_g^0, \alpha_i^0)$ and autocovariance function $r_i(\tau) = \mathrm{Cov}(\zeta_{it}, \zeta_{i,t+\tau})$, for $\tau \ge 0$. Assumption A3.(b) implies that

$$\inf_{i\ge1}\left|E(\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{i1}, Y_{i1}, \beta^0_{g_i^0}, \alpha_i^0))\right| > 0.$$

Above inequality and Lemma A.3 yields

$$\begin{aligned} \lambda \equiv \sup_{i\ge1}\|\rho_i\|_2 &\le \frac{\sqrt{p}E(\widetilde{J}(X_{i1}, Y_{i1}))}{\inf_{i\ge1}\left|E(\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{i1}, Y_{i1}, \beta^0_{g_i^0}, \alpha_i^0))\right|} \\ &\le \frac{\sqrt{p}E^{1/q_0}(\widetilde{J}^{q_0}(X_{i1}, Y_{i1}))}{\inf_{i\ge1}\left|E(\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{i1}, Y_{i1}, \beta^0_{g_i^0}, \alpha_i^0))\right|} < \infty, \end{aligned}$$

and

$$|\zeta_{it}| = \left|u'U_i(x,y,\beta,\alpha)\right| \quad \leq \quad \|U_i(x,y,\beta,\alpha)\|_2 \leq \sqrt{p}\widetilde{J}(x,y) + \lambda\widetilde{J}(x,y).$$

By Fan and Yao (2003)[Proposition 2.5] and Assumption A1.(b), we have

$$\begin{aligned} |r_i(\tau)| &= |\mathrm{Cov}(\zeta_{it},\zeta_{i,t+\tau})| \\ &\leq 8\alpha_{[i]}^{1/3}(\tau)E^{1/3}(|\zeta_{it}|^3)E^{1/3}(|\zeta_{i,t+\tau}|^3) \\ &= 8\alpha_{[i]}^{1/3}(\tau)E^{2/3}(|\zeta_{it}|^3) \\ &\leq 8\alpha_{[i]}^{1/3}(\tau)\sup_{i\geq 1}E^{2/3}(\widetilde{J}^3(X_{i1},Y_{i1})) \equiv C\alpha_{[i]}^{1/3}(\tau), \end{aligned}$$

with $C = 8\sup_{i\geq 1}E^{2/3}(\widetilde{J}^3(X_{i1},Y_{i1}))$ being finite due to Lemma A.3. Therefore, above inequality and Assumption A1.(b) imply

$$\sup_{i\geq 1}\sum_{\tau=1}^{\infty}|r_i(\tau)| \leq C\sum_{\tau=1}^{\infty}\sup_{i\geq 1}\alpha_{[i]}^{1/3}(\tau) \leq C\sum_{\tau=1}^{\infty}\exp(-C_0\tau^{b_0}) < \infty. \tag{A.41}$$

Direct examination yields

$$\frac{1}{T}\mathrm{Var}(\sum_{t=1}^{T}\zeta_{it}) \quad = \quad r_i(0) + 2\sum_{\tau=1}^{T-1}(1-\frac{\tau}{T})r_i(\tau). \tag{A.42}$$

Thus it follows from (A.41), (A.42), Assumption A1.(b) and dominated convergence theorem that

$$\begin{aligned} &\lim_{T\to\infty}\sup_{i\geq 1}\left|\frac{1}{T}\mathrm{Var}(\sum_{t=1}^{T}\zeta_{it}) - \left(r_i(0) + 2\sum_{\tau=1}^{\infty}r_i(\tau)\right)\right| \\ \leq\ &\lim_{T\to\infty}\sup_{i\geq 1}\left(2\sum_{\tau=T}^{\infty}|r_i(\tau)| + 2\sum_{\tau=1}^{T-1}\frac{\tau}{T}|r_i(\tau)|\right) \\ \leq\ &2\lim_{T\to\infty}\sum_{\tau=T}^{\infty}\exp(-C_0\tau^{b_0}) + 2\lim_{T\to\infty}\sum_{\tau=1}^{\infty}\frac{\tau}{T}\exp(-C_0\tau^{b_0})I(\tau\leq T-1) = 0, \end{aligned} \tag{A.43}$$

where second limit in above inequality is 0, since each summand is bounded by $\exp(-C_0\tau^{b_0})$, which is summable. So that we can change the order of summation and limit. Direct calculation yields

$$r_i(0) + 2\sum_{\tau=1}^{\infty}r_i(\tau) \quad = \quad u'\Lambda_i u. \tag{A.44}$$

Finally combining (A.43), (A.44) and noticing u is arbitrary, we proof the desired result. □

Lemma A.11. *Suppose Assumption A1 and A3 hold. Furthermore, if $N = o(T^{q_0/2})$, then for any random sequence $\{\beta_{Ti}, i\in[N]\}$ such that $\sup_{1\leq i\leq N}\|\beta_{Ti} - \beta_{g_i^0}^0\|_2 = o_P(1)$ and for all $|\boldsymbol{k}|\leq 3$, it follows that*

$$\sup_{1\leq i\leq N}|\sum_{t=1}^{T}D^{\boldsymbol{k}}(X_{it},Y_{it},\beta_{Ti},\widehat{\alpha}_i(\beta_{Ti}))/T - E(D^{\boldsymbol{k}}(X_{it},Y_{it},\beta_{g_i^0}^0,\alpha_i^0))| = o_P(1).$$

Proof of Lemma A.11. Fix $|\mathbf{k}| \leq 3$, let $K(x, y, \beta, \alpha) = D^{\mathbf{k}}\psi(x, y, \beta, \alpha)$. By triangle inequality, we have

$$\sup_{1\leq i\leq N} |\sum_{t=1}^{T} K(X_{it}, Y_{it}, \beta_{Ti}, \widehat{\alpha}_i(\beta_{Ti}))/T - E(K(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0))| \leq T_1 + T_2, \quad \text{(A.45)}$$

$$\text{with } T_1 = \sup_{1\leq i\leq N} \left| \sum_{t=1}^{T} [K(X_{it}, Y_{it}, \beta_{Ti}, \widehat{\alpha}_i(\beta_{Ti})) - K(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)]/T \right|,$$

$$T_2 = \sup_{1\leq i\leq N} \left| \sum_{t=1}^{T} K(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)/T - E(K(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)) \right|.$$

In the following we will bound T_1 and T_2 respectively. To bound T_1, for any $a > 0$, we define

$$h_{ai}(x, y) = \sup_{\{(\beta,\alpha):\|\beta-\beta^0_{g_i^0}\|_2+|\alpha-\alpha_i^0|\leq a\}} |K(x, y, \beta, \alpha) - K(x, y, \beta^0_{g_i^0}, \alpha_i^0)|.$$

By Lemma A.15, we have

$$R_{NT} \equiv \sup_{1\leq i\leq N} \|\beta_{Ti} - \beta^0_{g_i^0}\|_2 + \sup_{1\leq i\leq N} |\widehat{\alpha}_i(\beta_{Ti}) - \alpha_i^0| = o_P(1).$$

So for $a > 0$, it follows that

$$\begin{aligned} T_1 \quad &\leq \quad I(R_{NT} \leq a) \sup_{1\leq i\leq N} \sum_{t=1}^{T} h_{ai}(X_{it}, Y_{it})/T \\ &\quad + I(R_{NT} > a) \sup_{1\leq i\leq N} |\sum_{t=1}^{T} [K(X_{it}, Y_{it}, \beta_{Ti}, \widehat{\alpha}_i(\beta_{Ti})) - K(X_{it}, Y_{it}, \beta^0_{g_i^0}, \alpha_i^0)]/T| \\ &\equiv \quad T_{11} + T_{12}. \end{aligned} \quad \text{(A.46)}$$

By Lemma A.3, for $a \leq a_0$, we have $\sup_{1\leq i\leq N} h_{ai}(x, y) \leq \widetilde{J}(x, y)a$ and $\sup_{i\geq 1} E(\widetilde{J}^{q_0}(X_{i1}, Y_{i1})) < \infty$. Hence Lemma A.9 yields

$$\sup_{1\leq i\leq N} |\sum_{t=1}^{T} h_{ai}(X_{it}, Y_{it})/T - E(h_{ai}(X_{it}, Y_{it}))| = o_P(1). \quad \text{(A.47)}$$

By (A.47) and Lemma A.3, it holds that

$$\begin{aligned} T_{11} \quad &\leq \quad \sup_{1\leq i\leq N} \sum_{t=1}^{T} h_{ai}(X_{it}, Y_{it})/T \\ &= \quad \sup_{1\leq i\leq N} E(h_{ai}(X_{it}, Y_{it})) + o_P(1) \\ &\leq \quad \sup_{i\geq 1} E(\widetilde{J}(X_{it}, Y_{it}))a + o_P(1) \\ &= \quad a \sup_{i\geq 1} E^{1/(q_0)}(\widetilde{J}^{q_0}(X_{i1}, Y_{i1})) + o_P(1). \end{aligned}$$

Since a can be arbitrary small, it follows that $T_{11} = o_P(1)$. The fact that $R_{NT} = o_P(1)$ implies $T_{12} = o_P(1)$. Combining (A.46) and above, we have

$$T_1 = o_P(1). \tag{A.48}$$

By Lemma A.9 and Lemma A.3, a bound for T_2 can be obtained as follows,

$$T_2 = o_P(1). \tag{A.49}$$

The result follows when combing (A.45), (A.48) and (A.49). □

Lemma A.12. *Let x_0 be a point in $\mathbb{R}^m$ and h be a function defined on the ball $B = \{x \in \mathbb{R}^m : \|x - x_0\| < r)$ with derivative $\dot{h}$ and Hessian matrix $\ddot{h}$ which satisfies $\|\ddot{h}(x) - \ddot{h}(x_0)\| \leq L\|x - x_0\|$ for $x \in B$. If h is maximized at x_0, then*

$$h(x_0) - h(x) \geq (1/2)\|x - x_0\|^2(\lambda - L\delta), \quad \|x - x_0\| < \delta,$$

holds for all $0 < \delta < r$ and with λ the smallest eigenvalue of the matrix $-\ddot{h}(x_0)$.

Proof. The desired result follows from the Taylor expansion

$$h(x_0 + t) - h(x_0) = t'\dot{h}(x_0) + \int_0^1 (1 - s)t'\ddot{h}(x_0 + st)t\, ds$$

the fact that x_0 is a stationary point and that the integral is bounded by $[t'\ddot{h}(x_0)t + L\|t\|^3]/2$. □

Lemma A.13. *Suppose Assumption A1, A3 hold. Then there exist constant $C_6, C_7 > 0$ such that for any $(\beta_i, \alpha_i) \in \mathbb{K} \times \mathbb{A}, 1 \leq i \leq N$ satisfying $\sup_{1 \leq i \leq N}(\|\beta_i - \beta^0_{g_i^0}\|_2^2 + |\alpha_i - \alpha_i^0|^2)^{1/2} < C_7$, it follows that for all $i \geq 1$*

$$H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta_i, \alpha_i) \quad \geq \quad C_6(\|\beta_i - \beta^0_{g_i^0}\|_2^2 + |\alpha_i - \alpha_i^0|^2),$$

and

$$\frac{1}{N}\sum_{i=1}^N \left(H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta_i, \alpha_i) \right) \geq \frac{C_6}{N}\sum_{i=1}^N (\|\beta_i - \beta^0_{g_i^0}\|_2^2 + |\alpha_i - \alpha_i^0|^2).$$

Proof of Lemma A.13. This follows from Lemma A.12 applied with x_0 equal to $(\beta^0_{g_i^0}, \alpha_i)$ and h equal to the restriction of H_i to $\mathcal{B}_i$. By Lemma A.3, it follows that that $\ddot{H}_i$ is Lipschitz in $\mathcal{B}_i$ with Lipschitz constant $L_i = (p+1)(E(\tilde{J}^2(X_{i1}, Y_{i1})))^{1/2}$ and noting that $M = \inf_{i \geq 1} \lambda_{\min}(-\ddot{H}_i(\beta^0_{g_i^0} \alpha_i^0))$ is positive and $\Lambda = \sup_{i \geq 1} L_i$ is finite. Thus the choices $C_6 = M/4$ and $C_7 = \min(a_0, M/(2\Lambda))$ work. □

Lemma A.14. *Under Assumption A1, A3, there exist constant $C_8, C_9 > 0$ such that for $\epsilon > 0$ small enough,*

$$\inf_{i \geq 1} \inf_{\|\beta - \beta^0_{g_i^0}\|_2^2 + |\alpha - \alpha_i^0|^2 \geq \epsilon} [H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta, \alpha)] \geq \min(C_8, C_9\epsilon).$$

Proof of Lemma A.14. Assumption A1.(c) and Lemma A.13 imply

$$\begin{aligned}
&\inf_{1\le i\le N}\inf_{\|\beta-\beta^0_{g_i^0}\|_2^2+|\alpha-\alpha_i^0|^2\ge\epsilon}[H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta,\alpha)]\\
=\;&\min\Bigg\{\inf_{1\le i\le N}\inf_{\|\beta-\beta^0_{g_i^0}\|_2^2+|\alpha-\alpha_i^0|^2\ge C_7^2}[H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta,\alpha)],\\
&\inf_{1\le i\le N}\inf_{C_7^2>\|\beta-\beta^0_{g_i^0}\|_2^2+|\alpha-\alpha_i^0|^2\ge\epsilon}[H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta,\alpha)]\Bigg\}\\
\ge\;&\min\Big\{\chi(C_7^2),C_6\epsilon\Big\}.
\end{aligned}$$

Therefore, we proof the result with $C_8=\chi(C_7^2)$ and $C_9=C_6$. □

Lemma A.15. *Suppose Assumptions A1-A3 hold. Let $\{\beta_{g_i}, i\in[N]\}$ be a random sequence satisfying $\sup_{1\le i\le N}\|\beta_{g_i}-\beta^0_{g_i^0}\|_2\le\eta$ for some small enough $\eta>0$ with probability approaching one, then there exists a constant $C_{10}>0$ such that with probability approaching one*

$$\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{g_i})-\alpha_i^0|\le C_{10}\sqrt{\eta}.$$

Proof of Lemma A.15. Since with probability approaching one, $\sup_{1\le i\le N}\|\beta_{g_i}-\beta^0_{g_i^0}\|_2\le\eta$, we may proceed our proof assuming $\sup_{1\le i\le N}\|\beta_{g_i}-\beta^0_{g_i^0}\|_2\le\eta$.

By Lemma A.5 and definition of S_{NT}, it follows that

$$\begin{aligned}
\sup_{1\le i\le N}\sup_{\alpha\in\mathbb{A}}|\widehat{H}_i(\beta_{g_i},\alpha)-\widehat{H}_i(\beta^0_{g_i^0},\alpha)| &\le \sup_{1\le i\le N}\sup_{\alpha\in\mathbb{A}}|H_i(\beta_{g_i},\alpha)-H_i(\beta^0_{g_i^0},\alpha)|+2S_{NT}\\
&\quad\text{(By Lemma A.5)}\\
&\le B_2\sup_{1\le i\le N}\|\beta_{g_i}-\beta^0_{g_i^0}\|_2+2S_{NT}\\
&\le B_2\eta+2S_{NT} \qquad\text{(A.50)}
\end{aligned}$$

By direct examination and (A.50), one concludes that

$$\begin{aligned}
B_2\eta+2S_{NT} &\ge \sup_{\alpha\in\mathbb{A}}|\widehat{H}_i(\beta_{g_i},\alpha)-\widehat{H}_i(\beta^0_{g_i^0},\alpha)|\\
&\ge |\widehat{H}_i(\beta_{g_i},\widehat{\alpha}_i(\beta_{g_i}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))|\\
&\ge |\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|-|\widehat{H}_i(\beta_{g_i},\widehat{\alpha}_i(\beta_{g_i}))-\widehat{H}_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|\\
&\quad\text{(By definition of } S_{NT} \text{ and definition of } \widehat{\alpha}_i(\beta))\\
&\ge |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|-2S_{NT}-|\sup_{\alpha\in\mathbb{A}}\widehat{H}_i(\beta_{g_i},\alpha)-\sup_{\alpha\in\mathbb{A}}\widehat{H}_i(\beta^0_{g_i^0},\alpha)|\\
&\ge |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|-2S_{NT}-\sup_{\alpha\in\mathbb{A}}|\widehat{H}_i(\beta_{g_i},\alpha)-\widehat{H}_i(\beta^0_{g_i^0},\alpha)|,\\
&\quad\text{(By (A.50))}\\
&\ge |H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|-4S_{NT}-B_2\eta. \qquad\text{(A.51)}
\end{aligned}$$

taking supremum on both sides of (A.51), we have

$$\sup_{1\le i\le N}|H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|\le 2B_2\eta+6S_{NT}. \tag{A.52}$$

In the view of (A.52), Lemma A.5, Lemma A.7 and Lemma A.8 , we have

$$\begin{aligned}\sup_{1\le i\le N}|H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))| &\le \sup_{1\le i\le N}|H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta_{g_i}))-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|\\ &\quad+\sup_{1\le i\le N}|H_i(\beta^0_{g_i^0},\alpha_i^0)-H_i(\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))|\\ &\le 2B_2\eta+6S_{NT}+B_2\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0|\\ &= 2B_2\eta+o_P(1).\end{aligned}$$

Combining above inequality and Lemma A.14, when $C_8>2B_2\eta$, it follows that

$$\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{g_i})-\alpha_i^0|\le\sqrt{2B_2/C_9\eta}+o_P(1).$$

Therefore, the first result holds with $C_{10}=\sqrt{2B_2/C_9}$. The proof of second result is almost the same as that of first one. □

Lemma A.16. *Suppose Assumption A1-A3 hold. Furthermore, if $G\ge G^0$, then for $\eta>0$ small enough , we have the following:*

(i) For all $\underline{\beta}\in\mathcal{N}_\eta$, $\{\mathcal{A}_\eta(\underline{\beta},g),g\in[G^0]\}$ is a partition of $[G]$ and each $\mathcal{A}_\eta(\underline{\beta},g)$ is non empty for all $g\in[G^0]$.

(ii) $\lim_{(N,T)\to\infty}P\Big(\sup_{\underline{\beta}\in\mathcal{N}_\eta}\sup_{1\le i\le N}I(\widehat{g}_i(\underline{\beta})\notin\mathcal{A}_\eta(\underline{\beta},g_i^0))>0\Big)=0.$

(iii) If $G=G^0$, then each $\mathcal{A}_\eta(\underline{\beta},g)$ contains exactly one element for all $g\in[G^0]$ and thus $\mathcal{A}_\eta(\underline{\beta},\cdot)$ is a permutation of $[G^0]$. Under this permutation,

$$\lim_{(N,T)\to\infty}P\Big(\sup_{\underline{\beta}\in\mathcal{N}_\eta}\sup_{1\le i\le N}I(\widehat{g}_i(\underline{\beta})\ne g_i^0)>0\Big)=0.$$

Proof of Lemma A.16. (i) For $\underline{\beta}\in\mathcal{N}_\eta$, by definition, each $\mathcal{A}_\eta(\underline{\beta},g)$ is not empty. Moreover, definition of $\mathcal{N}_\eta$ and $\mathcal{A}_\eta$ shows that $\cup_{g=1}^{G^0}\mathcal{A}_\eta(\underline{\beta},g)=[G]$. Now we remain to show that $\{\mathcal{A}_\eta(\underline{\beta},g),g\in[G^0]\}$ is a partition of $[G]$. Assume there exist some $g_{12}\in[G],g_1,g_2\in[G^0]$ such that $g_{12}\in\mathcal{A}_\eta(\underline{\beta},g_1)\cap\mathcal{A}_\eta(\underline{\beta},g_2)$, then by Assumption A1.(d) and for $\eta<d_0/2$, it follows that

$$d_0\le\|\beta^0_{g_1}-\beta^0_{g_2}\|_2\le\|\beta_{g_{12}}-\beta^0_{g_1}\|_2+\|\beta_{g_{12}}-\beta^0_{g_2}\|_2<2\eta<d_0,$$

which is a contradiction.

(ii) By definition of $\widehat{g}_i(\underline{\beta})$, we have for all $g\in[G]$ and $\widetilde{g}\in[G]$:

$$I(\widehat{g}_i(\underline{\beta})=g)\le I(\widehat{H}_i(\beta_{\widetilde{g}},\widehat{\alpha}_i(\beta_{\widetilde{g}}))\le\widehat{H}_i(\beta_g,\widehat{\alpha}_i(\beta_g))),$$

Therefore, for any $\widetilde{g}_i \in \mathcal{A}_\eta(\underline{\beta}, g_i^0)$, it implies that

$$\begin{aligned} I(\widehat{g}_i(\underline{\beta}) \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) &= \sum_{g=1}^{G} I(g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) I(\widehat{g}_i(\underline{\beta}) = g) \\ &\leq \sum_{g=1}^{G} I(g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) I(\widehat{H}_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i})) \leq \widehat{H}_i(\beta_g, \widehat{\alpha}_i(\beta_g))) \equiv \sum_{g=1}^{G} W_{ig}(\underline{\beta}). \end{aligned} \tag{A.53}$$

Since for all $\underline{\beta} \in \mathcal{N}_\eta$, $\{\mathcal{A}_\eta(\underline{\beta}, g), g \in [G^0]\}$ is a partition of $[G]$. Therefore, for all $g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)$, we have $g \in \mathcal{A}_\eta(\underline{\beta}, g_j^0)$ with $g_j^0 \in [G^0]$ and $g_j^0 \neq g_i^0$. By definition, for small enough η and by Assumption A1.(d), we have

$$\|\beta^0_{g_i^0} - \beta_g\|_2 \geq \|\beta^0_{g_i^0} - \beta^0_{g_j^0}\|_2 - \|\beta_g - \beta^0_{g_j^0}\| \geq d_0 - \eta > 0.$$

According to Lemma A.7 and above inequality, it follows that for all $\underline{\beta} \in \mathcal{N}_\eta$ and $g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)$,

$$H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta_g, \widehat{\alpha}_i(\beta_g)) \geq \chi(|d_0 - \eta|^2) \tag{A.54}$$

On the other hand, by Lemma A.5, for all $\widetilde{g}_i \in \mathcal{A}_\eta(\underline{\beta}, g_i^0)$, we have

$$\begin{aligned} H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i})) &\leq B_2 \|\beta^0_{g_i^0} - \beta_{\widetilde{g}_i}\|_2 + B_2 |\alpha_i^0 - \widehat{\alpha}_i(\beta_{\widetilde{g}_i})| \\ &\leq B_2 \eta + B_2 |\alpha_i^0 - \widehat{\alpha}_i(\beta_{\widetilde{g}_i})| \\ &\leq B_2 \eta + B_2 \sup_{1 \leq i \leq N} |\alpha_i^0 - \widehat{\alpha}_i(\beta_{\widetilde{g}_i})|. \end{aligned} \tag{A.55}$$

Next define event $A_{NT} = \{\sup_{1 \leq i \leq N} |\alpha_i^0 - \widehat{\alpha}_i(\beta_{\widetilde{g}_i})| \leq C_{10}\sqrt{\eta}\}$, then $P(A_{NT}^c) = o(1)$ by Lemma A.15. To proceed further, choose sufficiently small η such that $\epsilon_\eta \equiv \chi(|d_0 - \eta|^2) - B_2\eta - C_{10}\sqrt{\eta} > 0$ for all (N, T) is large enough and this can done by Assumption A1.(c). Hence by (A.54) and (A.55), for all $\underline{\beta} \in \mathcal{N}_\eta$ and $i \in [N]$, on the event A_{NT}, it holds that

$$H_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i})) - H_i(\beta_g, \widehat{\alpha}_i(\beta_g)) \geq \epsilon_\eta. \tag{A.56}$$

As a consequence of (A.56), it yields that

$$\begin{aligned} W_{ig}(\underline{\beta}) &= I(g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) I(\widehat{H}_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i})) \leq \widehat{H}_i(\beta_g, \widehat{\alpha}_i(\beta_g))) \\ &\leq I(g \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) I(\epsilon_\eta \leq \widehat{H}_i(\beta_g, \widehat{\alpha}_i(\beta_g)) - \widehat{H}_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i})) - H_i(\beta_g, \widehat{\alpha}_i(\beta_g)) + H_i(\beta_{\widetilde{g}_i}, \widehat{\alpha}_i(\beta_{\widetilde{g}_i}))) \\ &\quad \times I(A_{NT}) + I(A_{NT}^c) \\ &\leq 2I(\epsilon_\eta/2 \leq \sup_{\beta \in \mathbb{K}, \alpha \in \mathbb{A}} |\widehat{H}_i(\beta, \alpha) - H_i(\beta, \alpha)|) + I(A_{NT}^c). \end{aligned} \tag{A.57}$$

By Lemma A.7, (A.53) and (A.57), it yields that

$$\begin{aligned}
& P(\sup_{\underline{\beta}\in\mathcal{N}_\eta}\sup_{1\le i\le N} I(\widehat{g}_i(\underline{\beta})\notin\mathcal{A}_\eta(\underline{\beta},g_i^0))>0)\\
= \;& P(\sup_{\underline{\beta}\in\mathcal{N}_\eta}\sup_{1\le i\le N} I(\widehat{g}_i(\underline{\beta})\notin\mathcal{A}_\eta(\underline{\beta},g_i^0))>0.5)\\
\le \;& \sum_{g=1}^{G} P\{\sup_{\underline{\beta}\in\mathcal{N}_\eta}\sup_{1\le i\le N} W_{i,g}(\underline{\beta})>0.5)\\
\le \;& \sum_{g=1}^{G} P(\sup_{1\le i\le N} 2I(\epsilon_\eta/2\le \sup_{\beta\in\mathbb{K},\alpha\in\mathbb{A}}|\widehat{H}_i(\beta,\alpha)-H_i(\beta,\alpha)|)>0.25)\\
& +P(I(A_{NT}^c)>0.25)\\
= \;& \sum_{g=1}^{G} P(\sup_{1\le i\le N}\sup_{\beta\in\mathbb{K},\alpha\in\mathbb{A}}|\widehat{H}_i(\beta,\alpha)-H_i(\beta,\alpha)|\ge\epsilon_\eta/2)+P(A_{NT}^c)=o(1).
\end{aligned}$$

(iii) If $G=G^0$, since $\{\mathcal{A}_\eta(\underline{\beta},g),g\in[G^0]\}$ is a partition of $[G]=[G^0]$ and each $\mathcal{A}_\eta(\underline{\beta},g)$ is non empty, so $\mathcal{A}_\eta(\underline{\beta},g)$ only contains exactly on elements in $[G]=[G^0]$. We are able to define the permutation $\mathcal{A}_\eta(\underline{\beta},\cdot):[G^0]\to[G^0]$. Therefore, under this permutation, the second result is a specially case of first one. Proof completed. □

Lemma A.17. *Suppose Assumption A1 and A3 hold. Furthermore, if $N=o(T^{q_0/2})$, then for any random sequence $\{\beta_{Ti},i\in[N]\}$ satisfying $\sup_{1\le i\le N}\|\beta_{Ti}-\beta_{g_i^0}^0\|_2=o_P(1)$, the following holds:*

$$\sum_{i=1}^{N}|\widehat{\alpha}_i(\beta_{Ti})-\alpha_i^0|^2/N=O_p(\sup_{1\le i\le N}\|\beta_{Ti}-\beta_{g_i^0}^0\|_2^2+T^{-1}).$$

Proof of Lemma A.17. By triangular inequality, it can be seen that

$$\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{Ti})-\alpha_i^0|\le\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{Ti})-\widehat{\alpha}_i(\beta_{g_i^0}^0)|+\sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{g_i^0}^0)-\alpha_i^0|. \tag{A.58}$$

In the following, we will bound two terms in (A.58) respectively. For first term, by the definition of $\widehat{\alpha}_i(\beta)$, we have $\sum_{t=1}^{T}\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta,\widehat{\alpha}_i(\beta))=0$. By implicit function differential theorem, we have

$$\frac{\partial\widehat{\alpha}_i}{\partial\beta}(\beta)=[\sum_{t=1}^{T}\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta,\widehat{\alpha}_i(\beta))/T]^{-1}\sum_{t=1}^{T}\frac{\partial^2\psi}{\partial\beta\partial\alpha}(X_{it},Y_{it},\beta,\widehat{\alpha}_i(\beta))/T.$$

Therefore, by mean value theorem and Lemma A.3, we have for some $s_i\in[0,1]$, it holds that

$$\widehat{\alpha}_i(\beta_{Ti})-\widehat{\alpha}_i(\beta_{g_i^0}^0)=\frac{\partial\widehat{\alpha}_i}{\partial\beta}(\beta_{g_i^0}^0+s_i(\beta_{Ti}-\beta_{g_i^0}^0))(\beta_{Ti}-\beta_{g_i^0}^0), \tag{A.59}$$

According to Lemma A.11 and Assumption A3.(b), it yields that

$$\sup_{1\le i\le N}|\frac{\partial\widehat{\alpha}_i}{\partial\beta}(\beta_{g_i^0}^0+s_i(\beta_{Ti}-\beta_{g_i^0}^0))-E^{-1}[\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta_{g_i^0}^0,\alpha_i^0)]E[\frac{\partial^2\psi}{\partial\beta\partial\alpha}(X_{it},Y_{it},\beta_{g_i^0}^0,\alpha_i^0)]|=o_P(1). \tag{A.60}$$

Combining (A.59), (A.60) and Lemma A.3, we have

$$\sum_{i=1}^{N}|\widehat{\alpha}_i(\beta_{Ti})-\widehat{\alpha}_i(\beta^0_{g_i^0})|^2/N \le \sup_{1\le i\le N}|\widehat{\alpha}_i(\beta_{Ti})-\widehat{\alpha}_i(\beta^0_{g_i^0})|^2 = O_p(\sup_{1\le i\le N}\|\beta_{Ti}-\beta^0_{g_i^0}\|_2^2). \quad \text{(A.61)}$$

For second term, by Taylor expansion, for some $s_i\in[0,1]$, we have

$$\begin{aligned} & -\sum_{t=1}^{T}\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)/T \\ = & \sum_{t=1}^{T}[\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\widehat{\alpha}_i(\beta^0_{g_i^0}))-\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)]/T \\ = & \sum_{t=1}^{T}\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0+s_i(\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0))/T(\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0). \quad \text{(A.62)}\end{aligned}$$

Combing (A.62) and Lemma A.11, it follows that

$$-\sum_{t=1}^{T}\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)/T = (E[\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)]+o_P(1))(\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0),$$

where the $o_P(1)$ is free of i. By Assumption A3.(b), it yields that

$$\begin{aligned}\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0 &= (E^{-1}[\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)]\sum_{t=1}^{T}\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)/T+o_P(1)(\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0), \\ &= E^{-1}(R^\alpha_{i1})\sum_{t=1}^{T}R_{it}/T+o_P(1)(\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0), \quad \text{(A.63)}\end{aligned}$$

where the $o_P(1)$ is free of i. By Assumption A3.(b), Lemma A.9 and Lemma A.3 , it follows that

$$\begin{aligned}E(\frac{1}{N}\sum_{i=1}^{N}|E^{-1}(R^\alpha_{i1})\sum_{t=1}^{T}R_{it}/T|^2) &= \frac{1}{N}\sum_{i=1}^{N}E^{-2}[\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)]E([\frac{\sum_{t=1}^{T}\frac{\partial\psi}{\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)}{T}]^2) \\ &\le \inf_{i\ge1}E^{-2}[\frac{\partial^2\psi}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta^0_{g_i^0},\alpha_i^0)]CT^{-1}, \quad \text{(A.64)}\end{aligned}$$

where $C>0$ is an universal constant free of i,N,T. Combining (A.63) and (A.64), we have

$$\sum_{i=1}^{N}|\widehat{\alpha}_i(\beta^0_{g_i^0})-\alpha_i^0|^2/N = O_p(T^{-1}). \quad \text{(A.65)}$$

Therefore, the desired result follows from (A.61) and (A.65). Proof completed. □

Lemma A.18. *Suppose Assumption A1, A3 hold and $N = O(T^{q_0/2})$., then for any $g \in [G^0]$, we have following stochastic expansion:*

$$\begin{aligned}
&\frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}[U_i(X_{it},Y_{it},\widetilde{\beta}_g,\widehat{\alpha}_i(\widetilde{\beta}_g)) - U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)] \\
= &\frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}V_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)(\widetilde{\beta}_g-\beta_g^0) \\
&+\frac{1}{N_g T}\sum_{i:g_i^0=g}\left(\frac{\sum_{t=1}^{T}R_{it}}{\sqrt{T}E(R_{i1}^{\alpha})}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[U_{it}^{\alpha}-\frac{E(U_{i1}^{\alpha\alpha})}{2E(R_{it}^{\alpha})}R_{it}]\right) \\
&+o_P(\|\widetilde{\beta}_g-\beta_g^0\|_2)+o_P(T^{-1}).
\end{aligned}$$

Proof. For notational simplicity, we assume $\widetilde{\beta}_g$ is a scalar. Extension to the case when $\widetilde{\beta}_g$ is multi-dimensional can be easily done by similar technique. By Theorem 2 and Theorem 5, we can see $\|\widetilde{\beta}_g-\beta_g^0\|_2 = o_p(1)$. By mean value theorem, it follows that

$$\begin{aligned}
&\frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}[U_i(X_{it},Y_{it},\widetilde{\beta}_g,\widehat{\alpha}_i(\widetilde{\beta}_g)) - U_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0) - V_i(X_{it},Y_{it},\beta_g^0,\alpha_i^0)(\widetilde{\beta}_g-\beta_g^0)] \\
= &\frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}\frac{\partial U_i}{\partial\alpha}(X_{it},Y_{it},\beta_g^0,\alpha_i^0)(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0) \\
&+\frac{1}{2N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}\frac{\partial^2 U_i}{\partial\beta\partial\beta}(X_{it},Y_{it},\beta_g^0+s_i(\widetilde{\beta}_g-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))(\widetilde{\beta}_g-\beta_g^0)^2 \\
&+\frac{1}{2N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}\frac{\partial^2 U_i}{\partial\beta\partial\alpha}(X_{it},Y_{it},\beta_g^0+s_i(\widetilde{\beta}_g-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))(\widetilde{\beta}_g-\beta_g^0)(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0) \\
&+\frac{1}{2N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}\frac{\partial^2 U_i}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta_g^0+s_i(\widetilde{\beta}_g-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0)^2, \\
&\equiv T_1+T_2/2+T_3/2+T_4/2,
\end{aligned} \tag{A.66}$$

where s_i are some numbers in $[0,1]$. Next we will bound T_1, T_2, T_3, T_4 respectively. Firstly, by

Lemma A.9 and Lemma A.17, it shows that

$$\begin{aligned} T_1 &= \frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}\frac{\partial U_i}{\partial \alpha}(X_{it},Y_{it},\beta_g^0,\alpha_i^0)(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0) \\ &= \frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_{it}^{\alpha}(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0) \\ &= \frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_{it}^{\alpha}(\widehat{\alpha}_i(\widetilde{\beta}_g)-\widehat{\alpha}_i(\beta_g^0))+\frac{1}{N_g T}\sum_{i:g_i^0=g}\sum_{t=1}^{T}U_{it}^{\alpha}(\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0) \\ &\equiv T_{11}+T_{12}. \end{aligned}$$

Concerning T_{11}, by Cauchy's inequality and notice

$$\begin{aligned} |T_{11}|^2 &\le \frac{1}{N_g^2}\sum_{i:g_i^0=g}|\sum_{t=1}^{T}U_{it}^{\alpha}/T|^2\sum_{i:g_i^0=g}|\widehat{\alpha}_i(\widetilde{\beta}_g)-\widehat{\alpha}_i(\beta_g^0)|^2 \\ &= O_p(T^{-1})O_p(\|\widetilde{\beta}_g-\beta_g^0\|_2^2), \end{aligned}$$

where the last equality comes from Lemma A.9 and (A.61) in Lemma A.17. For T_{12}, by (A.63) in Lemma A.17, it holds that

$$T_{12} = \frac{1}{N_g}E^{-1}(R_{i1}^{\alpha})\sum_{i:g_i^0=g}\Big(\sum_{t=1}^{T}U_{it}^{\alpha}/T\Big)\Big(\sum_{t=1}^{T}R_{it}/T\Big)+o_P(1)\times\frac{1}{N_g}\sum_{i:g_i^0=g}\Big(\sum_{t=1}^{T}U_{it}^{\alpha}/T\Big)(\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0). \tag{A.67}$$

Since by Cauchy's inequality, Lemma A.9 and (A.65) in Lemma A.17, we have

$$\begin{aligned} |\frac{1}{N_g}\sum_{i:g_i^0=g}\Big(\sum_{t=1}^{T}U_{it}^{\alpha}/T\Big)(\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0)|^2 &\le \frac{1}{N_g^2}\sum_{i:g_i^0=g}|\sum_{t=1}^{T}U_{it}^{\alpha}/T|^2\sum_{i:g_i^0=g}|\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0|^2 \\ &= O_p(T^{-1})O_p(T^{-1}). \end{aligned}$$

As a consequence, the second term in right side of (A.67) is $o_P(T^{-1})$. Therefore, combining above, it follows that

$$T_1 = \frac{1}{N_g}E^{-1}(R_{i1}^{\alpha})\sum_{i:g_i^0=g}\Big(\sum_{t=1}^{T}U_{it}^{\alpha}/T\Big)\Big(\sum_{t=1}^{T}R_{it}/T\Big)+O_p(\|\widetilde{\beta}_g-\beta_g^0\|_2T^{-1/2})+o_p(T^{-1}). \tag{A.68}$$

Secondly, by Lemma A.11, we have

$$\sup_{1\le i\le N}|\sum_{t=1}^{T}\frac{\partial^2 U_i}{\partial\beta\partial\beta}(X_{it},Y_{it},\beta_g^0+s_i(\widetilde{\beta}_g-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))/T-E[\frac{\partial^2 U_i}{\partial\beta\partial\beta}(X_{it},Y_{it},\beta_g^0,\alpha_i^0)]|=o_P(1).$$

Therefore, by Cauchy's inequality and Lemma A.3, it follows that

$$\begin{aligned} |T_2| &= |\frac{1}{N_g}\sum_{i:g_i^0=g} c_{iT}|(\widetilde{\beta}_g-\beta_g^0)^2 \\ &\leq |\frac{1}{N_g}\sum_{i:g_i^0=g} E[\frac{\partial^2 U_i}{\partial\beta\partial\beta}(X_{it},Y_{it},\beta_g^0,\alpha_i^0)] + o_P(1)|(\widetilde{\beta}_g-\beta_g^0)^2 \\ &= o_P(\|\widetilde{\beta}_g-\beta_g^0\|_2), \end{aligned} \tag{A.69}$$

where $c_{iT}=\sum_{t=1}^T \frac{\partial^2 U_i}{\partial\beta\partial\beta}(X_{it},Y_{it},\beta_g^0+s_i(\beta-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\beta)-\alpha_i^0))$. Similar proof can show that

$$T_3 = o_P(\|\widetilde{\beta}_g-\beta_g^0\|_2). \tag{A.70}$$

Lastly, we will deal with T_4. Let

$$d_{iT}=\sum_{t=1}^T \frac{\partial^2 U_i}{\partial\alpha\partial\alpha}(X_{it},Y_{it},\beta_g^0+s_i(\widetilde{\beta}_g-\beta_g^0),\alpha_i^0+s_i(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))/T.$$

By Lemma A.11 , we have

$$\sup_{1\leq i\leq N}|d_{iT}-E(U_{i1}^{\alpha\alpha})| = o_P(1).$$

As a consequence, by Lemma A.17 we have

$$\begin{aligned} T_4 &= \frac{1}{N_g}\sum_{i:g_i^0=g} d_{iT}(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))^2 \\ &= \frac{1}{N_g}\sum_{i:g_i^0=g} (E(U_{i1}^{\alpha\alpha})+o_P(1))(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))^2 \\ &= \frac{1}{N_g}\sum_{i:g_i^0=g} E(U_{i1}^{\alpha\alpha})(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))^2 + o_P(\|\widetilde{\beta}_g-\beta_g^0\|_2^2+T^{-1}). \end{aligned} \tag{A.71}$$

Now we will establish a bound of (A.71). By Assumption A3.(b) and Lemma A.3, it follows that $\sup_{i\geq 1}|E(U_{i1}^{\alpha\alpha})|<\infty$. Furthermore, in the view of (A.61) in Lemma A.17, we have

$$\begin{aligned} &|\frac{1}{N_g}\sum_{i:g_i^0=g} E(U_{i1}^{\alpha\alpha})(\widehat{\alpha}_i(\widetilde{\beta}_g)-\alpha_i^0))^2 - \frac{1}{N_g}\sum_{i:g_i^0=g} E(U_{i1}^{\alpha\alpha})(\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0))^2| \\ \leq &\sup_{i:g_i^0=g} E(U_{i1}^{\alpha\alpha})\Big(|\widehat{\alpha}_i(\widetilde{\beta}_g)-\widehat{\alpha}_i(\beta_g^0)|^2+2|\widehat{\alpha}_i(\widetilde{\beta}_g)-\widehat{\alpha}_i(\beta_g^0)||\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0)|\Big) \\ = &\ o_p(\|\widetilde{\beta}_g-\beta_g^0\|_2). \end{aligned} \tag{A.72}$$

Meanwhile, by (A.63) in Lemma A.17, we can show

$$\frac{1}{N_g}\sum_{i:g_i^0=g} E(U_{i1}^{\alpha\alpha})(\widehat{\alpha}_i(\beta_g^0)-\alpha_i^0))^2 = \frac{1}{N_g}\sum_{i:g_i^0=g}\frac{E(U_{i1}^{\alpha\alpha})}{E^2(R_{i1}^{\alpha})}(\sum_{t=1}^T R_{it}/T)^2 + o_P(T^{-1}). \tag{A.73}$$

Combining (A.71), (A.73) and (A.73), we have

$$T_4 = \frac{1}{N_g} \sum_{i:g_i^0=g} \frac{E(U_{i1}^{\alpha\alpha})}{E^2(R_{i1}^{\alpha})} (\sum_{t=1}^{T} R_{it}/T)^2 + o_p(\|\widetilde{\beta}_g - \beta_g^0\|_2) + o_P(T^{-1}). \tag{A.74}$$

The desired result follows from (A.66), (A.68), (A.69), (A.70) and (A.74). Proof completed. □

Lemma A.19. *Under Assumption A1, A3 and $G < G^0$, there exists a constant B_4 such that* $\liminf_{(N,T)\to\infty}[\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N)] \geq B_4 > 0$.

Proof of Lemma A.19. First we will show that for fixed $G < G^0$, the following holds:

$$\max_{g\in[G^0]} \min_{\widetilde{g}\in[G]} \|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2 \geq d_0/2. \tag{A.75}$$

Assume (A.75) fails to hold, then for each $g \in [G^0]$, there is a $\widetilde{g} = \sigma(g) \in [G]$ such that $\|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2 < d_0/2$, where $\sigma : [G^0] \to [G]$ is the map defined in (A.1). Since $G < G^0$, so for some $\widetilde{g}_0 \in [G]$, there are $g_1, g_2 \in [G^0]$ such that $\sigma(g_1) = \sigma(g_2) = \widetilde{g}_0$. Hence it follows that

$$\|\widehat{\beta}_{\widetilde{g}_0} - \beta_{g_i}^0\|_2 < d_0/2, \text{ for } i = 1, 2. \tag{A.76}$$

By triangular inequality and (A.76), we have $\|\beta_{g_1}^0 - \beta_{g_2}^0\|_2 < d_0$, which leads a contradiction to Assumption A1.(d). Hence we verify (A.75). By direct examination and (A.75), it follows that

$$\begin{aligned}
\frac{\sum_{i=1}^N \|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2}{N} &= \frac{\sum_{g=1}^{G^0}\sum_{i=1}^N I(g_i^0 = g)\|\widehat{\beta}_{\widehat{g}_i} - \beta_g^0\|_2}{N} \\
&\geq \frac{\sum_{g=1}^{G^0}\sum_{i=1}^N I(g_i^0 = g)\min_{\widetilde{g}\in[G]}\|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2}{N} \\
&\geq \frac{\sum_{g=1}^{G^0} N_g \min_{\widetilde{g}\in[G]}\|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2}{N} \\
&\geq \frac{\max_{g\in[G^0]} N_g \min_{\widetilde{g}\in[G]}\|\widehat{\beta}_{\widetilde{g}} - \beta_g^0\|_2}{N} \\
&\geq \frac{\min_{g\in[G^0]} N_g d_0}{N}.
\end{aligned} \tag{A.77}$$

By (A.77) and Assumption A1.(f), for sufficient large N, it follows that

$$\frac{\sum_{i=1}^N \|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2}{N} \geq \min_{g\in[G^0]} \pi_g d_0/2 > 0. \tag{A.78}$$

For notational simplicity, let $\epsilon_0 = \min_{g\in[G^0]} \pi_g d_0/2 > 0$. By (A.78) and Lemma A.4, we can see that

$$\liminf_{(N,T)\to\infty} [\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N)] \geq \frac{\epsilon_0}{2R}\chi(\epsilon_0^2/8) > 0.$$

Therefore, the results follow with $B_4 = \epsilon_0\chi(\epsilon_0^2/8)/(2R)$. □

Lemma A.20. *Let ν be a constant such that $0 < \nu \leq \frac{1}{2(1+d)}$. Suppose Assumptions A1-A3 hold and $G \geq G^0$. Furthermore, if $\log N = o(T^{(\frac{1}{1+d}-\nu)d})$, then*

$$\sum_{i=1}^{N}(\|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g^0_i}\|_2^2 + |\widehat{\alpha}_i - \alpha^0_i|^2)/N = O_p\left(T^{-\nu}\right).$$

Proof of Lemma A.20. For notational simplicity, define $\widetilde{d}_N(\theta_N, \theta^0_N) = \sum_{i=1}^{N}(\|\beta_{g_i} - \beta^0_{g^0_i}\|_2^2 + |\alpha_i - \alpha^0_i|^2)/N$ and $A_N = \{\theta_N \in \Theta_N : \sup_{1\leq i\leq N}(\|\beta_{g_i} - \beta^0_{g^0_i}\|_2^2 + |\alpha_i - \alpha^0_i|^2) \leq C_7^2\}$. By Lemma A.15 and Theorem 2, we have $\widetilde{d}_N(\widehat{\theta}_N, \theta^0_N) = o_P(1)$ and $\lim_{(N,T)\to\infty} P(\widehat{\theta}_N \in A_N) = 1$. Let r_{NT} be an increasing sequence of positive number and define $S_{N,j} = \{\theta_N \in \Theta_N : 2^{j-1} \leq r_{NT}\widetilde{d}_N(\theta_N, \theta^0_N) \leq 2^j\} \cap A_N$. If $r_{NT}\widetilde{d}_N(\widehat{\theta}_N, \theta^0_N) > 2^k$ for a some sufficient large integer k, then $\widehat{\theta}_N$ is in one of the set $S_{N,j}$ for some $j > k$. So it follows that, for all $\eta > 0$

$$\begin{aligned} P(r_{NT}\widetilde{d}_N(\widehat{\theta}_N, \theta^0_N) > 2^k, \widehat{\theta}_N \in A_N) \quad \leq \quad & \sum_{j\geq k, 2^j\leq \eta r_{NT}} P\left(\sup_{\theta_N\in S_{N,j}}\left[\widehat{\Psi}_N(\theta_N) - \widehat{\Psi}_N(\theta^0_N)\right] \geq 0\right) \\ & + P(\widetilde{d}_N(\widehat{\theta}_N, \theta^0_N) > \eta). \end{aligned} \tag{A.79}$$

The second term in right side of (A.79) is $o(1)$ by argument above. Now we will handle the first summation. Notice if $\theta_N \in S_{N,j}$, and choosing sufficiently small η , we have $\widetilde{d}_N(\theta_N, \theta^0_N) \leq \eta$. As a consequence, by Lemma A.13, it holds for all $\theta_N \in S_{N,j}$ that

$$\begin{aligned} \Psi_N(\theta^0_N) - \Psi_N(\theta_N) \quad &\geq \quad \frac{C_6}{N}\sum_{i=1}^{N}(\|\beta_{g_i} - \beta^0_{g^0_i}\|_2^2 + |\alpha_i - \alpha^0_i|^2) \\ &= \quad C_6\widetilde{d}_N(\theta_N, \theta^0_N) \\ &\geq \quad C_6 2^{j-1} r_{NT}^{-1}. \end{aligned} \tag{A.80}$$

In the view of (A.80), it follows that

$$\begin{aligned} & P(\sup_{\theta_N\in S_{N,j}}[\widehat{\Psi}_N(\theta_N) - \widehat{\Psi}_N(\theta^0_N)] \geq 0) \\ \leq \quad & P(\sup_{\theta_N\in S_{N,j}}[\widehat{\Psi}_N(\theta_N) - \Psi_N(\theta_N) - \widehat{\Psi}_N(\theta^0_N) + \Psi_N(\theta^0_N)] \geq C_6 2^{j-1} r_{NT}^{-1}) \\ \leq \quad & 2P(\sup_{\theta_N\in S_{N,j}}|\widehat{\Psi}_N(\theta_N) - \Psi_N(\theta_N)| \geq C_6 2^{j-1} r_{NT}^{-1}) \\ \leq \quad & 2P(\sup_{\theta_N\in \Theta_N}|\widehat{\Psi}_N(\theta_N) - \Psi_N(\theta_N)| \geq C_6 2^{j-1} r_{NT}^{-1}) \\ \leq \quad & 2P\left(\sup_{1\leq i\leq N}\sup_{(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}}\left|\widehat{H}_i(\beta,\alpha) - H_i(\beta,\alpha)\right| > C_6 2^{j-1} r_{NT}^{-1}\right). \end{aligned} \tag{A.81}$$

Next we will establish a bound of (A.81). By Lemma A.7, (A.81) and replacing r_{NT} by $C_6 T^{\nu}/6$,

then for large k that $4^{k-1}/C_5 \geq e-1$, we have

$$
\begin{aligned}
&\sum_{j\geq k, 2^j \leq \eta r_{NT}} P\bigg(\sup_{\theta_N \in S_{N,j}} \Big[\widehat{\Psi}_N(\theta_N) - \widehat{\Psi}_N(\theta_N^0) \Big] \geq 0 \bigg) \\
\leq\ & 2 \sum_{j\geq k, 2^j \leq \eta T^\nu} P\bigg(\sup_{1\leq i\leq N} \sup_{(\beta,\alpha)\in \mathbb{K}\times\mathbb{A}} \Big| \widehat{H}_i(\beta,\alpha) - H_i(\beta,\alpha) \Big| > C_6 2^{j-1} r_{NT}^{-1} \bigg) \\
\leq\ & 2 \sum_{j\geq k, 2^j \leq \eta T^\nu} P\bigg(\sup_{1\leq i\leq N} \sup_{(\beta,\alpha)\in \mathbb{K}\times\mathbb{A}} \Big| \widehat{H}_i(\beta,\alpha) - H_i(\beta,\alpha) \Big| > 6\times 2^{k-1} T^{-\nu} \bigg) \\
\leq\ & 2C_4 N \log(\eta T^\nu) \bigg[1 + \bigg(\frac{1}{T^{-2(p+2)\nu}} \bigg) \bigg] \\
& \times \bigg[\bigg(1 + \frac{T^{1-d-2\nu}}{C_5} \bigg)^{-T^d/4} + \frac{2}{d} \exp\bigg(- C_3 T^{d(1-\nu)} \bigg) + \frac{\exp(-C_3 d T^{d(1-\nu)})}{1-\exp(-C_3 d T^{d(1-\nu)})} \bigg] \\
\leq\ & 2C_4 N \log(\eta T^\nu) \bigg[1 + \bigg(\frac{T^{2(p+2)\nu}}{2^{2(p+2)(k-1)}} \bigg) \bigg] \\
& \times \bigg[\bigg(1 + \frac{T^{\frac{1}{1+d}-2\nu} 4^{k-1}}{C_5} \bigg)^{-T^{\frac{d}{1+d}}/4} + \frac{2}{d} \exp\bigg(- C_3 T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d} \bigg) + \frac{\exp(-C_3 d T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d})}{1-\exp(-C_3 d T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d})} \bigg] \\
\leq\ & 2C_4 N \log(\eta T^\nu) \bigg[1 + \bigg(\frac{T^{2(p+2)\nu}}{2^{2(p+2)(k-1)}} \bigg) \bigg] \\
& \times \bigg[\exp\Big(-T^{\frac{d}{1+d}}/4 \Big) + \frac{2}{d} \exp\bigg(- C_3 T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d} \bigg) + \frac{\exp(-C_3 d T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d})}{1-\exp(-C_3 d T^{(\frac{1}{1+d}-\nu)d} 2^{(k-1)d})} \bigg] \\
=\ & o(1). \qquad \text{(A.82)}
\end{aligned}
$$

Therefore, the desired result follows from (A.79), (A.81), (A.82) and the fact that $P(\widehat{\theta}_N \in A_N) \to 1$. □

Lemma A.21. *Suppose Assumption A1, A3 hold and $G > G^0$. Furthermore, if $\log N = o(T^{\frac{d}{2(1+d)}})$, then*

$$|\widehat{\Psi}_N(\widehat{\theta}_N) - \widehat{\Psi}_N(\theta_N^0)| = O_P(T^{-\frac{1}{2(1+d)}}).$$

Proof of Lemma A.21. By direct examination and Assumption A1.(e), it yields that

$$
\begin{aligned}
|\widehat{\Psi}_N(\widehat{\theta}_N) - \widehat{\Psi}_N(\theta_N^0)| &\leq \sum_{i=1}^N \frac{\sum_{t=1}^T Q(X_{it}, Y_{it})}{TN} (\|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i}^0\|_2^2 + |\widehat{\alpha}_i - \alpha_i^0|^2)^{1/2} \\
&\leq \sqrt{\frac{1}{N}\sum_{i=1}^N \bigg(\frac{\sum_{t=1}^T Q(X_{it}, Y_{it})}{T} \bigg)^2} \sqrt{\frac{1}{N}\sum_{i=1}^N (\|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i}^0\|_2^2 + |\widehat{\alpha}_i - \alpha_i^0|^2)}. \qquad \text{(A.83)}
\end{aligned}
$$

The first second factor in right side of (A.83) is $O_P(T^{-\frac{1}{2(1+d)}})$ by Lemma A.20 with $\nu = \frac{1}{2(1+d)}$.

Next we will bound the first factor. By direct examination, we have

$$E\left[\left(\frac{\sum_{t=1}^{T} Q(X_{it}, Y_{it})}{T}\right)^2\right] \leq 2E\left[\left(\frac{\sum_{t=1}^{T}[Q(X_{it}, Y_{it}) - E(Q(X_{it}, Y_{it}))]}{T}\right)^2\right] + 2E^2[Q(X_{i1}, Y_{it})]. \quad \text{(A.84)}$$

Utilizing Assumption A1.(b), A1.(e) and Lemma A.9, the right side of (A.84) is uniformly bounded for all $i \geq 1$. As a consequence, we obtaine the desired result. □