

# SPECTRAL AND POST-SPECTRAL ESTIMATORS FOR GROUPED PANEL DATA MODELS

DENIS CHETVERIKOV AND ELENA MANRESA

ABSTRACT. In this paper, we develop spectral and post-spectral estimators for grouped panel data models. Both estimators are consistent in the asymptotics where the number of observations  $N$  and the number of time periods  $T$  simultaneously grow large. In addition, the post-spectral estimator is root- $NT$  consistent and asymptotically normal with mean zero under the assumption of well-separated groups even if  $T$  is growing much slower than  $N$ . The post-spectral estimator has, therefore, theoretical properties that are similar to those of the grouped fixed-effect estimator developed by Bonhomme and Manresa in [9]. In contrast to the grouped fixed-effect estimator, however, our post-spectral estimator is computationally straightforward.

## 1. INTRODUCTION

Consider a grouped panel data model

$$y_{it} = x'_{it}\beta + \alpha_{g_i t} + v_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (1)$$

where  $i$  denotes cross-sectional units,  $t$  denotes time periods,  $y_{it} \in \mathbb{R}$  is a dependent variable,  $x_{it} = (x_{it1}, \dots, x_{itd})' \in \mathbb{R}^d$  is a corresponding vector of covariates,  $g_i \in \{1, \dots, G\}$  is a group-membership variable,  $v_{it} \in \mathbb{R}$  is a zero-mean noise random variable,  $\beta = (\beta_1, \dots, \beta_d)' \in \mathbb{R}^d$  is a vector of parameters of interest, and  $(\alpha_{1t}, \dots, \alpha_{Gt})' \in \mathbb{R}^G$  is a vector of group-specific time effects. Here, we assume that the noise and covariates are uncorrelated,

$$\mathbb{E}[v_{it}x_{it}] = 0_d, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (2)$$

where  $0_d = (0, \dots, 0)' \in \mathbb{R}^d$ , but group-specific time effects and group-membership variables can be arbitrarily correlated with covariates.

The model (1) was originally introduced by Bonhomme and Manresa in [9], who also developed a so-called grouped fixed-effect estimator of the vector of parameters

---

*Date:* May 18, 2021.

We are grateful to Stephane Bonhomme, Victor Chernozhukov, Andrew Chesher, Jin Hahn, Arturas Juodis, Yusuke Narita, Andres Santos, Martin Weidner, Daniel Wilhelm, UCL and Chamberlain's seminar participants, and especially our discussant at Chamberlain's seminar Roger Moon for useful comments.

$\beta$  in this model. This estimator has attractive theoretical properties but, being a solution to a non-convex optimization problem, is computationally difficult. It is therefore of interest to see if there exist alternative estimators that would be easier to compute. In this paper, we answer this question affirmatively, under certain additional assumptions, and propose an estimator of  $\beta$ , which we call the *post-spectral estimator*, that also has nice theoretical properties but, in contrast to the grouped fixed-effect estimator, is computationally simple.

Like in the previous papers on grouped panel data models, we consider large  $(N, T)$ -asymptotics, i.e. we assume that  $T \rightarrow \infty$ , potentially very slowly, as  $N \rightarrow \infty$ , since otherwise  $\beta$  is in general not identified. In contrast to the previous papers, however, we impose a special structure on the data-generating process for the covariates  $x_{it}$ . In particular, we assume that

$$x_{it} = \sum_{m=1}^M \rho_{im} \alpha_{git}^m + z_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (3)$$

where  $(\alpha_{1t}^m, \dots, \alpha_{Gt}^m)' \in \mathbb{R}^G$  for  $m = 1, \dots, M$  are group-specific time effects,  $\rho_{im} \in \mathbb{R}^d$  for  $m = 1, \dots, M$  are individual-specific vectors of coefficients, and  $z_{it}$  is a zero-mean component of  $x_{it}$  that is independent of group-specific time effects, group-membership variable  $g_i$ , and vectors of coefficients  $\rho_{i1}, \dots, \rho_{iM}$ . Here, without loss of generality, we assume that  $\alpha_{\gamma t}^1 = \alpha_{\gamma t}$  for all  $\gamma = 1, \dots, G$  and  $t = 1, \dots, T$ . We believe that this factor-analytic model for the covariates  $x_{it}$  is rather flexible as it allows for individual-specific correlations between covariates and group-specific time effects.

Our post-spectral estimator consists of three steps. In the first step, we carry out preliminary estimation of  $\beta$ . To do so, we prove that as long as the data-generating process is given by equations (1) and (3), there exists a convex and quadratic function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  such that (i) its unique minimum is achieved by  $\beta$  and (ii) for each value of  $b \in \mathbb{R}^d$ , the value of  $f(b)$  can be consistently estimated by the sum of  $2GM + 2$  largest in absolute value eigenvalues of a certain matrix. We then demonstrate that this function and its consistent estimator can be used to construct an estimator of  $\beta$  that is both consistent and computationally simple. This estimator, which we call the *spectral estimator*, may have slow rate of convergence if  $T$  is growing rather slowly, and so we proceed to the second and the third steps. In the second step, using the preliminary spectral estimator of  $\beta$  obtained in the first step, we carry out classification of units  $i = 1, \dots, N$  into groups  $\gamma = 1, \dots, G$ . Importantly, our classification algorithm is fast and does not require solving any non-convex optimization problems. It is based on consistent estimation of group-specific time effects via the method of asymptotic principle components as in [6]. In the third step, we obtain the *post-spectral estimator*

of  $\beta$  by performing OLS-type estimation pooling all units within the same group together.

We prove that our post-spectral estimator is generally consistent and has particularly attractive properties under the assumption of well-separated groups, which means that the vectors  $(\alpha_{\gamma 1}, \dots, \alpha_{\gamma T})'$ ,  $\gamma = 1, \dots, G$ , are not too close to each other, and which was also used in [9]. Specifically, we show that under this assumption, the classifier constructed in the second step is consistent in the sense that with probability approaching one, any two units are getting classified into the same group if and only if they belong to the same group, and so the post-spectral estimator of  $\beta$  is asymptotically equivalent to the pooled-OLS estimator with known group memberships (i.e., the oracle estimator). In turn, the latter is root- $NT$  consistent and admits the standard OLS inference. Under the assumption of well-separated groups, our post-spectral estimator thus can be used for testing hypotheses and for constructing confidence intervals for  $\beta$  using standard panel-data methods, ignoring the preliminary estimation and classification steps. Inference without the assumption of well-separated groups, however, remains an open (and challenging) question for future work. Theoretical properties of our post-spectral estimator are thus similar to those of the grouped fixed-effect estimator, with the caveat that we impose the special structure on the data-generating process for the covariates  $x_{it}$  given in (3).

To compare computational properties of the post-spectral estimator to those of the grouped fixed-effect estimator, we note that the latter, as well as many related estimators ([3, 4, 10, 11, 36]), are built around the K-means optimization problem. This optimization problem is known to be NP-hard (see [5]), implying that finding its solution is in general computationally difficult. Any proposed *fast* implementation of the aforementioned estimators therefore must fail occasionally. For example, the grouped fixed-effect estimator is defined as the (global) minimizer of the sum of squared residuals over all parameter values and over all partitions of units into groups, and the main algorithm to calculate this minimizer in [9] proceeds by initializing randomly selected values of parameters  $\beta$  and  $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$  and then alternating between two steps: (1) optimization over the values of group memberships  $g_1, \dots, g_N$  given the values of parameters  $\beta$  and  $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$ , and (2) optimization over the values of parameters  $\beta$  and  $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$  given the values of group memberships  $g_1, \dots, g_N$ . Since this procedure at best converges to a local minimum, it is repeated over many different initial values of parameters to find the global minimum, which corresponds to the grouped fixed-effect estimator. As [9] notes, however, “a prohibitive number of initial values may be needed to obtain reliable solutions.” In addition, it seems never possible

to say whether the global minimum has been found, as this in general would require minimizing the sum of squared residuals over parameter values for *each* partition of units into groups, and the number of these partitions,  $G^N$ , is tremendously large even in small samples. [9] also proposed a few other algorithms to calculate the grouped fixed-effect estimator that tend to perform better in simulations but they are all subject to the same critique: if they are fast, they must fail occasionally. In contrast, our post-spectral estimator is easy to compute and does not suffer from the occasional failure problem.

There is also growing literature on non-linear panel data models with group structure of individual-level parameters ([18, 22, 28, 8, 27, 33, 17, 16]) originated by Hahn and Moon in [18]. This literature is conceptually related to the grouped panel data model (1) but estimation techniques developed in this literature are very different from those considered here because all aforementioned papers assume that the individual-level parameters are time-independent, and so preliminary consistent estimation of these parameters is possible by performing estimation separately for each unit. The latter is not possible in the model (1) because individual effects  $\alpha_{g_{it}}$  are varying over time, which creates one of the key challenges in estimating this model.

We note also that the grouped panel data model (1) is a special case of a panel data model with interactive fixed effects, corresponding to factor loadings with finite support in the latter model. The methods developed for estimating panel data models with interactive fixed effects can therefore be used to estimate  $\beta$  in (1) as well. To the best of our knowledge, however, most of these methods are either computationally difficult or require conditions that are substantially different from those used in our paper. For example, the estimator in [7] is based on a solution to a non-convex optimization problem, and the estimators in [21, 1, 2] require certain IV-type conditions. Like in our approach, the estimators in [25] also require restricting the data-generating process for the covariates  $x_{it}$  but the nature of imposed restrictions is very different. In particular, [25] either imposes a certain rank condition, which can only be satisfied if the dimensionality of  $x_{it}$  is sufficiently large, or requires the factor loadings in the equation for covariates to be independent of the factor loadings in the equation for the dependent variable, which in our model would correspond to assuming that  $g_i$  in (3) is different and independent of  $g_i$  in (1),<sup>1</sup> thus leading to a random, rather than fixed, effect model. In fact, the only exception in this literature we are aware of is [24], who developed a computationally simple method and used

---

<sup>1</sup>[25] claims that his estimators are consistent without requiring independence of the factors but a counter-example is given in [34], which proves that if the rank condition is not satisfied, then independence is essentially a necessary condition for consistency of the estimators in [25].

conditions that are similar (actually somewhat weaker) than those used in our paper. Their estimator can potentially replace our preliminary spectral estimator. However, the convergence rate of their estimator is only  $(T \wedge N)^{-1/2}$ , which can be particularly slow if  $T$  is growing much slower than  $N$ , a case of special interest in the grouped panel data model. In contrast, the rate of convergence of our spectral estimator is  $(T \wedge N)^{-1}$ , which seems more acceptable for the preliminary estimation.

In addition, we study three extensions of the model (1). First, we consider a dynamic version of the model, where lagged values of  $y_{it}$  appear on the right-hand side of (1). We demonstrate that our spectral and post-spectral estimators work for this model too, as long as the number of factors  $M$  is appropriately modified. We allow for both pre-determined and exogenous covariates in this model. Second, we consider a high-dimensional version of the model (1), where the number of covariates  $d$  is large, potentially much larger than  $NT$ , but the vector of coefficients  $\beta$  is sparse in the sense that it has relatively few non-zero components. We demonstrate how to modify the spectral estimator via  $\ell^1$ -penalization to obtain a computationally simple and consistent estimator of  $\beta$  in this case. Third, we consider an interactive fixed-effect panel data model, where  $\alpha_{g_i t}$  in (1) is replaced by  $\kappa_i' \phi_t$ , with  $\phi_t$  being a vector of factors and  $\kappa_i$  being a vector of factor loadings. We demonstrate that the spectral estimator, with appropriately modified parameters  $G$  and  $M$ , is consistent in this model with the convergence rate  $(T \wedge N)^{-1}$ . Being computationally simple, our spectral estimator thus can serve as an alternative to existing estimators in the literature on interactive fixed-effect panel data models. Note, however, that in all three extensions, we maintain a version of (3).

The rest of the paper is organized as follows. In the next section, we discuss details of implementation of our spectral and post-spectral estimators. In Section 3, we derive their asymptotic properties. In Section 4, we study the extensions of the baseline model. In Section 5, we provide results of a small-scale Monte Carlo simulation study that shed some light on finite-sample properties of our estimators. In Section 6, we present all the proofs. In Section 7, we collect some technical lemmas that are useful for the proofs of our main results. In Appendix A, we describe a method for calculating eigenvalues of large matrices, which may be needed for implementing our estimators. In Appendix B, we provide some details on the assumptions of the dynamic model extension.

## 2. ESTIMATION

Our proposed estimation procedure consists of three steps. The first step is preliminary consistent estimation of  $\beta$ , which is based on the spectral analysis of certain matrices and gives the spectral estimator. The second step is classification of units into groups. The third step is pooled OLS estimation of  $\beta$  on classified units, which gives the post-spectral estimator. Under the assumption of well-separated groups, the post-spectral estimator will be root- $NT$  consistent and asymptotically normal with mean zero, making inference based on this estimator straightforward.

**2.1. Spectral Estimator.** For all  $b \in \mathbb{R}^d$ , let  $A^b$  be an  $N \times N$  matrix whose  $(i, j)$ -th element is

$$A_{ij}^b = \frac{1}{NT} \sum_{t=1}^T \left\{ (y_{it} - x'_{it}b) - (y_{jt} - x'_{jt}b) \right\}^2, \quad \text{for all } i, j = 1, \dots, N, \quad (4)$$

Since  $A^b$  is an  $N \times N$  symmetric matrix, it has  $N$  real eigenvalues. Let  $\lambda_1^b, \dots, \lambda_{2GM+2}^b$  be its  $2GM + 2$  largest in absolute value eigenvalues. We will show below that under mild conditions,

$$\lambda_1^b + \dots + \lambda_{2GM+2}^b = b'\Sigma b + S'b + L + o_P(1), \quad \text{for all } b \in \mathbb{R}^d, \quad (5)$$

where  $\Sigma$  is a  $d \times d$  symmetric positive definite matrix,  $S$  is a  $d \times 1$  vector,  $L$  is a scalar, and, importantly,  $\beta$  is the unique minimizer of the function  $b \mapsto f(b) = b'\Sigma b + S'b + L$ . We therefore define our spectral estimator as

$$\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_d)' = \arg \min_{b \in \mathbb{R}^d} \left\{ b'\hat{\Sigma}b + \hat{S}'b + \hat{L} \right\}, \quad (6)$$

where  $\hat{\Sigma}$ ,  $\hat{S}$ , and  $\hat{L}$  are estimators of  $\Sigma$ ,  $S$ , and  $L$ , respectively, to be constructed below. By the first-order conditions, the estimator  $\tilde{\beta}$  can be equivalently defined as

$$\tilde{\beta} = -\hat{\Sigma}^{-1}\hat{S}/2.$$

We will prove in the next section that  $\tilde{\beta} \rightarrow_P \beta$ .

Next, we discuss estimators  $\hat{\Sigma}$ ,  $\hat{S}$ , and  $\hat{L}$ . For brevity of notation, denote

$$\hat{f}(b) = \lambda_1^b + \dots + \lambda_{2GM+2}^b, \quad \text{for all } b \in \mathbb{R}^d, \quad (7)$$

so that by (5),

$$\hat{f}(b) = b'\Sigma b + S'b + L + o_P(1), \quad \text{for all } b \in \mathbb{R}^d. \quad (8)$$

Also, for all  $k = 1, \dots, d$ , let  $e_k = (0, \dots, 0, 1, 0, \dots, 0)'$  be the  $d \times 1$  vector with 1 in the  $k$ -th position and 0 in all other positions, and let  $0_d = (0, \dots, 0)'$  be the  $d \times 1$

vector with 0 in all positions. Since (5) implies  $\hat{f}(0_d) = L + o_P(1)$ , we set  $\hat{L} = \hat{f}(0_d)$ . Further, since  $\hat{f}(e_k) - \hat{f}(-e_k) \rightarrow_P 2S_k$ , we set

$$\hat{S}_k = \frac{\hat{f}(e_k) - \hat{f}(-e_k)}{2}, \quad \text{for all } k = 1, \dots, d,$$

and  $\hat{S} = (\hat{S}_1, \dots, \hat{S}_d)'$ . Finally, since  $\hat{f}(e_k) + \hat{f}(-e_k) = 2(\Sigma_{kk} + L) + o_P(1)$ , we set

$$\hat{\Sigma}_{kk} = \frac{\hat{f}(e_k) + \hat{f}(-e_k)}{2} - \hat{L}, \quad \text{for all } k = 1, \dots, d,$$

and since  $\hat{f}(e_k + e_l) = \Sigma_{kk} + \Sigma_{ll} + 2\Sigma_{kl} + S_k + S_l + L + o_P(1)$ , we set

$$\hat{\Sigma}_{kl} = \hat{\Sigma}_{lk} = \frac{\hat{f}(e_k + e_l) - \hat{\Sigma}_{kk} - \hat{\Sigma}_{ll} - \hat{S}_k - \hat{S}_l - \hat{L}}{2},$$

for all  $k, l = 1, \dots, d$ ,  $k > l$ , and let  $\hat{\Sigma}$  be the matrix whose  $(k, l)$ -th component is  $\hat{\Sigma}_{kl}$ .

Under result (5), the estimators  $\hat{\Sigma}$  and  $\hat{S}$  are consistent for  $\Sigma$  and  $S$ , respectively, and so  $\tilde{\beta} = -\hat{\Sigma}^{-1}\hat{S}/2 \rightarrow_P -\Sigma^{-1}S/2 = \beta$ , as long as  $\Sigma$  is invertible, which is the case under mild conditions. The bulk of our derivations in the next section will thus be related to proving (5).

Before we move on, we note that the  $N \times N$  matrices  $A^b$  may be rather large, and the reader might wonder how much time it takes to calculate their eigenvalues  $\lambda_1^b, \dots, \lambda_{2GM+2}^b$ . Fortunately, there exists a class of fast randomized algorithms that allow to calculate these eigenvalues approximately. In the Appendix, we review one such algorithm, whose output is estimators  $\hat{\lambda}_1^b, \dots, \hat{\lambda}_{2GM+2}^b$  such that

$$\hat{\lambda}_1^b + \dots + \hat{\lambda}_{2GM+2}^b = \lambda_1^b + \dots + \lambda_{2GM+2}^b + o_P(1).$$

Our results below thus go through if we replace eigenvalues  $\lambda_1^b, \dots, \lambda_{2GM+2}^b$  by their estimators  $\hat{\lambda}_1^b, \dots, \hat{\lambda}_{2GM+2}^b$  in (7), and since all other steps in calculating  $\tilde{\beta}$  are computationally trivial, it follows that our preliminary spectral estimator  $\tilde{\beta}$  is easy to calculate.

**Remark 2.1** (Alternative Version of Spectral Estimator). Given that we have (5) and that  $\beta$  is the unique minimizer of the function  $b \mapsto b'\Sigma b + S'b + L$ , it seems natural to consider

$$\check{\beta} = \arg \min_{b \in \mathbb{R}^d} \lambda_1^b + \dots + \lambda_{2GM+2}^b$$

as an alternative to the spectral estimator  $\tilde{\beta}$  appearing in (6). The minimization problem here, however, is not necessarily convex, even though the criterion function is asymptotically convex. Computing  $\check{\beta}$  may therefore be difficult. In contrast, our

spectral estimator  $\tilde{\beta}$  circumvents this problem by employing the parametric structure of the limit of this criterion function.  $\blacksquare$

**2.2. Classifier.** To classify units into groups, we will use sample splitting. To this end, let  $h_1, \dots, h_N$  be i.i.d. random variables that are independent of the data and that are taking values 0 and 1, each with probability 1/2. We split all cross-sectional units  $i = 1, \dots, N$  into two subsamples,  $\mathcal{I}_0 = \{i = 1, \dots, N: h_i = 1\}$  and  $\mathcal{I}_1 = \{i = 1, \dots, N: h_i = 0\}$ . Further, for  $i = 1, \dots, N$ , denote  $y_i = (y_{i1}, \dots, y_{iT})'$  and  $x_i = (x_{i1}, \dots, x_{iT})'$ . Also, for  $h = 0, 1$ , let  $\tilde{\beta}^h$  be the spectral estimator calculated using the subsample  $\mathcal{I}_h$  and let  $\hat{B}^h$  be a  $T \times T$  matrix given by

$$\hat{B}^h = \frac{2}{NT} \sum_{i \in \mathcal{I}_h} (y_i - x_i \tilde{\beta}^h)(y_i - x_i \tilde{\beta}^h)'. \quad (9)$$

Since  $\hat{B}^h$  is a  $T \times T$  symmetric positive definite matrix, it has  $T$  non-negative eigenvalues and  $T$  corresponding orthonormal eigenvectors. Let  $\hat{F}_h$  be a  $T \times G$  matrix whose columns are orthonormal eigenvectors corresponding to  $G$  largest eigenvalues of the matrix  $\hat{B}^h$ . Moreover, for all  $i = 1, \dots, N$ , let  $\hat{A}_i$  be a  $T \times 1$  vector defined by

$$\hat{A}_i = \hat{F}_{h_i} \hat{F}_{h_i}' (y_i - x_i \tilde{\beta}^{h_i}). \quad (10)$$

Intuitively, the vectors  $\hat{A}_1, \dots, \hat{A}_N$  estimate the vectors  $\alpha_{g_1}, \dots, \alpha_{g_N}$ , where we denoted  $\alpha_\gamma = (\alpha_{\gamma 1}, \dots, \alpha_{\gamma T})'$  for all  $\gamma = 1, \dots, G$ . We therefore classify units  $i = 1, \dots, N$  into  $G$  groups using these vectors. To do so, fix a tuning parameter  $\lambda > 0$ , to be chosen below, and consider the following algorithm:

**Classification Algorithm.**

- Step 1:* set  $\mathcal{A}_1 = \{1\}$ ,  $m = 1$ , and  $i = 1$ ;
- Step 2:* replace  $i$  by  $i + 1$ ;
- Step 3:* if  $i = N + 1$ , stop the algorithm;
- Step 4:* set  $\mathcal{C}_i = \{\gamma = 1, \dots, m: \|\hat{A}_i - |\mathcal{A}_\gamma|^{-1} \sum_{l \in \mathcal{A}_\gamma} \hat{A}_l\| \leq \lambda\}$ ;
- Step 5:* if  $\mathcal{C}_i$  is empty, replace  $m$  by  $m + 1$ , set  $\mathcal{A}_m = \{i\}$ , and go to Step 2;
- Step 6:* if  $\mathcal{C}_i$  is not empty, replace  $\mathcal{A}_\gamma$  by  $\mathcal{A}_\gamma \cup \{i\}$  for  $\gamma = \min \mathcal{C}_i$  and go to Step 2.

This algorithm creates  $m$  groups  $\mathcal{A}_1, \dots, \mathcal{A}_m$ , with the number of groups  $m$  depending on  $\lambda$ , so that  $m = m(\lambda)$ . Clearly,  $\lambda \mapsto m(\lambda)$  is a right-continuous function, and so

$$\hat{\lambda} = \min \left\{ \lambda > 0: m(\lambda) \leq G \right\}$$

is well-defined. (In practice,  $\hat{\lambda}$  can be calculated using the values of  $m(\lambda)$  on a fine grid.) We classify units  $i = 1, \dots, N$  into  $G$  groups using this algorithm with  $\lambda = \hat{\lambda}$ .



The result of the algorithm is then  $m(\hat{\lambda}) \leq G$  groups  $\mathcal{A}_1, \dots, \mathcal{A}_{m(\hat{\lambda})}$ , and for all  $i = 1, \dots, N$ , there exists a unique  $\gamma = \gamma(i) \in \{1, \dots, m(\hat{\lambda})\}$  such that  $i \in \mathcal{A}_\gamma$ . We set

$$\hat{g}_i = \gamma(i), \quad \text{for all } i = 1, \dots, N,$$

and  $\hat{g} = (\hat{g}_1, \dots, \hat{g}_N)'$ . Note that this classifier can occasionally lead to less than  $G$  groups, which happens when  $m(\hat{\lambda}) < G$ , but we will show in the next section that it is consistent in the sense that

$$P\left(\text{for all } i, j = 1, \dots, N, \hat{g}_i = \hat{g}_j \text{ if and only if } g_i = g_j\right) \rightarrow 1, \quad (11)$$

under the assumption of well-separated groups.

**Remark 2.2** (Replacing  $\tilde{\beta}$  with Other Estimators). The proof of (11) in Theorem 3.2 below will not use any properties of  $\tilde{\beta}$  except for its consistency. Therefore, the same result will hold if we replace  $\tilde{\beta}$  in (9) and (10) by any other consistent estimator of  $\beta$ . For example, once the post-spectral estimator  $\hat{\beta}$  is calculated in the next subsection, we can reiterate classification and pooled OLS estimation steps by using  $\hat{\beta}$  instead of  $\tilde{\beta}$  in the second classification step. This may potentially improve properties of the final, iterated post-spectral, estimator. ■

**Remark 2.3** (Covariate-Based Classifiers). Recall that we assume the group structure in the data-generating process for covariates  $x_{it}$ , equation (3). In principle, this structure could be used to classify units into groups as well. This seemingly sensible alternative to our procedures is interesting because it does not require estimating  $\beta$  on the first step, and so looks much easier than our procedures. However, a substantial drawback of this procedure is that it may not be consistent if the group structure in the data-generating process for  $x_{it}$ 's is coarser than the group-structure in the data-generating process for  $y_{it}$ 's. For example, suppose that  $M = 2$  and  $\rho_{i1} = 0$  for all  $i = 1, \dots, N$ . Then equation (3) becomes

$$x_{it} = \rho_{i2}\alpha_{git}^2 + z_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T.$$

Now, if we assume that  $G = 3$  but  $\alpha_{1t}^2 = \alpha_{2t}^2 \neq \alpha_{3t}^2$ , there are effectively only two groups in the data-generating process for  $x_{it}$ 's. Therefore, any reasonable classification based on this equation would merge groups 1 and 2, which would make (11) impossible. ■

**Remark 2.4** (Relation to the Literature). Our classifier  $\hat{g}$  is related to the classifier proposed in [31] in the sense that both classifiers first estimate the vectors  $\alpha_{g1}, \dots, \alpha_{gN}$

and then perform classification of units into groups based on these estimates. However, our Classification Algorithm is different from the corresponding algorithm in [31]. An advantage of our Classification Algorithm is that it does not require the assumption of weakly-isotropic noise variables, which means that  $\mathbb{E}[(\sum_{t=1}^T u_t v_{it})^2]$  is constant across all  $u = (u_1, \dots, u_T)' \in \mathbb{R}^T$  satisfying  $\sum_{t=1}^T u_t^2 = 1$ , and which is used in [31]. This assumption is clearly restrictive and, as [31] note explicitly, it is not clear how to generalize their algorithm to avoid this assumption. More generally, our classifier falls into the class of so-called spectral classifiers; see [23] for more information on these and other classifiers.  $\blacksquare$

**2.3. Post-Spectral Estimator.** Once we have classified units into groups, estimation of  $\beta$  is straightforward. In particular, we rely upon a pooled OLS estimator:

$$(\hat{\beta}, \hat{\alpha}) = \arg \min_{b \in \mathcal{B}, a \in \mathcal{A}_{G,T}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x'_{it} b - a_{\hat{g}_i t})^2, \quad (12)$$

where  $\mathcal{B}$  is a parameter space for the vector  $\beta$ , and  $\mathcal{A}_{G,T}$  is a parameter space for the matrix  $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$ . We refer to  $\hat{\beta}$  as the post-spectral estimator for grouped panel data models. We will show in the next section that under the assumption of well-separated groups, this estimator is asymptotically equivalent to the estimator based on correct classification,

$$(\hat{\beta}^0, \hat{\alpha}^0) = \arg \min_{b \in \mathcal{B}, a \in \mathcal{A}_{G,T}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x'_{it} b - a_{g_i t})^2, \quad (13)$$

and thus to the grouped fixed-effect estimator of [9]. Hence, the standard OLS inference ignoring group classification applies.

**Remark 2.5** (Estimating  $\beta$  by OLS of  $y_{it}$  on estimated  $z_{it}$ ). Observe that our data-generating process for covariates  $x_{it}$  in (3) is given by a factor-analytic model; namely, it can be written as

$$x_{it} = \sum_{m=1}^M \sum_{\gamma=1}^G \rho_{im} 1\{g_i = \gamma\} \alpha_{\gamma t}^m = \omega'_i \phi_t + z_{it}, \quad \text{for all } i = 1, \dots, N, t = 1, \dots, T,$$

where  $\phi_t = (\alpha_{1t}^1, \dots, \alpha_{Gt}^M)'$  is a  $GM \times 1$  vector of factors and  $\omega_i = (\rho_{i1} 1\{g_i = 1\}, \dots, \rho_{iM} 1\{g_i = G\})'$  is the  $GM \times d$  matrix of factor loadings. Here, factors  $\phi_t$  and the factor loadings  $\omega_i$  can be estimated by the method of asymptotic principle components as in Section 3 of [6]; see also [12, 13, 26, 15]. Denoting these estimators  $\hat{\phi}_t$  and  $\hat{\omega}_i$  and letting  $\hat{z}_{it} = x_{it} - \hat{\omega}'_i \hat{\phi}_t$ , we are then able to obtain an estimator of  $\beta$  by simply running OLS of  $y_{it}$  or  $\hat{z}_{it}$ . This estimator is easy to compute and is

consistent under weak conditions since  $z_{it}$  is uncorrelated with both  $\alpha_{g_it}$  and  $x_{it} - z_{it}$ . However, this estimator performs poorly in the case of weak factors, i.e. when the factor loadings  $\omega_i$  are close to zero, in which case our post-spectral estimator has better performance by an order of magnitude; see our simulation results in Section 5.

■

### 3. ASYMPTOTIC THEORY

In this section, we derive asymptotic properties of the procedures described above. For convenience, we do so in three separate subsections: spectral estimator, classifier, and post-spectral estimator.

Throughout the rest of the paper, we assume that membership variables  $g_i$ , group-specific time effects  $\alpha_{\gamma t}^m$  and individual specific vectors of coefficients  $\rho_{im}$  are non-stochastic, i.e. our analysis is conditional on these random quantities. Also, given that we set  $\alpha_{\gamma t}^1 = \alpha_{\gamma t}$ , group-specific time effects  $\alpha_{\gamma t}$  are non-stochastic as well. Moreover, we assume that the units  $i$  are independent.

**3.1. Spectral Estimator.** Let  $\mathcal{S}^T$  denote the unit sphere in  $\mathbb{R}^T$ , i.e.  $\mathcal{S}^T = \{u \in \mathbb{R}^T : \|u\| = 1\}$ . Also, for any random variable  $w$ , let  $\|w\|_{\psi_2}$  denote the sub-Gaussian norm of  $w$ , i.e.

$$\|w\|_{\psi_2} = \inf \left\{ \epsilon > 0 : \mathbb{E}[\exp(w^2/\epsilon^2)] \leq 2 \right\};$$

see Section 2.5.2 in [32] on properties of the sub-gaussian norm. Intuitively, a random variable has a finite sub-Gaussian norm if the tails of its distribution are not heavier than tails of the Gaussian distribution. For example, every bounded random variable has a finite sub-Gaussian norm. To prove consistency and to derive the rate of convergence of the spectral estimator  $\tilde{\beta}$ , we will use the following assumptions.

**Assumption 3.1.** *For some constant  $C_1 > 0$ , we have  $\|\sum_{t=1}^T u_t v_{it}\|_{\psi_2} \leq C_1$  for all  $i = 1, \dots, N$  and  $u = (u_1, \dots, u_T)' \in \mathcal{S}^T$ .*

**Assumption 3.2.** *For some constant  $C_2 > 0$ , we have  $\|\sum_{t=1}^T u_t z_{itk}\|_{\psi_2} \leq C_2$  for all  $i = 1, \dots, N$ ,  $u = (u_1, \dots, u_T)' \in \mathcal{S}^T$ , and  $k = 1, \dots, d$ .*

**Assumption 3.3.** *We have  $\|(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T v_{it} z_{it}\| = O_P(1/\sqrt{NT})$ .*

**Assumption 3.4.** *We have  $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T z_{it} z'_{it} = \Sigma/2 + O_P(1/\sqrt{NT})$ , where  $\Sigma$  is a positive definite  $d \times d$  matrix.*

By Hoeffding's inequality (Proposition 2.6.1 in [32]), Assumption 3.1 holds if the random variables  $v_{it}$  have finite sub-Gaussian norm and are independent across  $t$ . More generally, due to numerous versions of Hoeffding's inequality for time series

data (e.g., see [14, 30]), Assumption 3.1 holds as long as the dependence of the random variables  $v_{it}$  across  $t$  is not too strong. Assumption 3.2 is similar to Assumption 3.1 but imposes the integrability and time series dependence restrictions on  $z_{it}$  instead of  $v_{it}$ . We admit that the assumption of random variables having finite sub-Gaussian norm may be somewhat strong but we emphasize that a version of the results below can be derived under weaker integrability assumptions. We have chosen to work with Assumptions 3.1 and 3.2 in order to minimize technicalities of our analysis. Further, since  $\mathbb{E}[v_{it}] = 0$ ,  $\mathbb{E}[v_{it}x_{it}] = 0_d$  and we assume that  $\alpha_{git}^m$  and  $\rho_{im}$  are non-stochastic, it follows from (3) that  $\mathbb{E}[v_{it}z_{it}] = 0_d$ . Hence, Assumption 3.3 is a quantitative law of large numbers for the random vector  $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T v_{it}z_{it}$ . Similarly, Assumption 3.4 is a quantitative law of large numbers for the random matrix  $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T z_{it}z'_{it}$ . Assumption 3.4 also imposes the constraint that the probability limit of this matrix is positive-definite, which is an identification condition.

**Theorem 3.1** (Rate of Convergence of Spectral Estimator  $\tilde{\beta}$ ). *Under Assumptions 3.1–3.4,*

$$\tilde{\beta} = \beta + O_P\left(\frac{1}{T \wedge N}\right). \quad (14)$$

**Remark 3.1.** DETAILED COMPARISON WITH MOON AND WEIDNER HERE.

**3.2. Classifier.** To prove consistency of the classifier  $\hat{g}$ , we will use the following assumptions.

**Assumption 3.5.** *For some constant  $C_3 > 0$ , we have  $\|\rho_{im}\| \leq C_3$  for all  $m = 1, \dots, M$  and  $i = 1, \dots, N$ .*

**Assumption 3.6.** *For some constant  $C_4 > 0$ , we have  $|\alpha_{\gamma t}^m| \leq C_4$  for all  $m = 1, \dots, M$ ,  $\gamma = 1, \dots, G$ , and  $t = 1, \dots, T$ .*

**Assumption 3.7.** *For some constant  $c_1 > 0$ , we have  $T^{-1} \sum_{i=1}^T (\alpha_{\gamma_1 t} - \alpha_{\gamma_2 t})^2 \geq c_1$  for all  $\gamma_1, \gamma_2 = 1, \dots, G$  such that  $\gamma_1 \neq \gamma_2$ .*

**Assumption 3.8.** *For some constant  $c_2 > 0$ , we have  $N_\gamma \geq c_2 N$  for all  $\gamma = 1, \dots, G$ .*

**Assumption 3.9.** *We have  $\log N = o(T)$  and  $\log T = o(N)$ .*

Assumptions 3.5 and 3.6 are self-explanatory. Assumption 3.7 means that the groups are well-separated in the sense that the vectors of group-specific time effects,  $(\alpha_{\gamma 1}, \dots, \alpha_{\gamma T})$  for  $\gamma = 1, \dots, G$ , are not too close to each other. We will use this assumption to prove consistency of the classifier  $\hat{g}$  and asymptotic normality of the post-spectral estimator  $\hat{\beta}$  but we will not use it to prove consistency of the

post-spectral estimator  $\hat{\beta}$ . Assumption 3.8 requires that each group  $\gamma = 1, \dots, G$  constitutes a non-trivial fraction of all units. If we were to assume random group assignment, where each unit is assigned to group  $\gamma$  with probability  $p_\gamma > 0$ , so that  $\sum_{\gamma=1}^G p_\gamma = 1$ , and units are assigned independently, this assumption would be satisfied with probability approaching one as  $N \rightarrow \infty$ . Finally, Assumption 3.9 specifies how fast  $T$  and  $N$  are required to grow relative to each other in the asymptotics. The most important observation here is that we allow  $T$  to be much smaller than  $N$ , which is the main case of interest for grouped panel data models; see [9].

**Theorem 3.2** (Consistency of Classifier  $\hat{g}$ ). *Under Assumptions 3.1–3.9, we have*

$$P\left(\text{for all } i, j = 1, \dots, N, \text{ we have } \hat{g}_i = \hat{g}_j \text{ if and only if } g_i = g_j\right) \rightarrow 1,$$

as long as  $T \rightarrow \infty$  as  $N \rightarrow \infty$ .

**Remark 3.2** (On the Role of Sample Splitting in Theorem 3.2). Using sample splitting to construct the vectors  $\hat{A}_1, \dots, \hat{A}_N$ , which are in turn used in the Classification Algorithm to obtain the classifier  $\hat{g}$ , is important for our analysis. Specifically, sample splitting allows us to avoid some restrictive assumptions on the geometry of group-specific time effects. Indeed, suppose that we do full-sample estimation, i.e. we set  $\hat{A}_i = \hat{F}\hat{F}'(y_i - x_i\tilde{\beta})$  for all  $i = 1, \dots, N$ , where  $\tilde{\beta}$  is the full-sample spectral estimator and  $\hat{F}$  is the  $T \times G$  matrix consisting of orthonormal eigenvectors corresponding to  $G$  largest eigenvalues of the matrix

$$\hat{B} = \frac{1}{NT} \sum_{i=1}^N (y_i - x_i'\tilde{\beta})(y_i - x_i'\tilde{\beta})',$$

and consider the following example. Let  $G = 2$  and  $\alpha_1 = 2\alpha_2$ , where  $\|\alpha_2\| \geq c\sqrt{T}$  for some constant  $c > 0$ . In this example, the assumption of well-separated groups (Assumption 3.7) is satisfied but the  $T \times T$  matrix  $B = N^{-1} \sum_{i=1}^N \alpha_{g_i} \alpha_{g_i}'$  has only one non-zero eigenvalue. Therefore, given that  $\hat{B}$  consistently estimate  $B$ , the matrix  $\hat{F}$  may not have a probability limit. As a result,  $T^{-1/2}\hat{F}'(v_{i1}, \dots, v_{iT})'$  may not converge to zero in probability (which is guaranteed in the construction based on sample splitting), and the vectors  $\hat{A}_1, \dots, \hat{A}_N$  may turn out poor estimators of the vectors  $\alpha_{g_1}, \dots, \alpha_{g_N}$ , leading to inconsistency of the classifier  $\hat{g}$ . More generally, with full-sample estimation, we would have to impose in Theorem 3.2 an extra assumption that the matrix  $B = N^{-1} \sum_{i=1}^N \alpha_{g_i} \alpha_{g_i}'$  has  $G$  eigenvalues bounded away from zero, which seems difficult to justify.<sup>2</sup> ■

<sup>2</sup>As a side note, we also observe that [31] do not use sample splitting to estimate vectors  $\alpha_{g_1}, \dots, \alpha_{g_N}$  but some parts of their derivations are difficult to verify. In particular, in Section 5, they use an

**3.3. Post-Spectral Estimator.** In this subsection, we present two results on our post-spectral estimator  $\hat{\beta}$ . First, we show that this estimator is generally consistent. Second, we show that under the assumption of well-separated groups, Assumption 3.7, this estimator is root- $NT$  consistent and has simple asymptotic distribution.

For all  $\nu = (\nu_1, \dots, \nu_N)' \in \{1, \dots, G\}^N$  and  $\gamma_1, \gamma_2 = 1, \dots, G$ , denote

$$\mathcal{I}(\nu, \gamma_1, \gamma_2) = \left\{ i = 1, \dots, N : \nu_i = \gamma_1 \text{ and } g_i = \gamma_2 \right\}$$

and

$$\bar{x}_{\nu, \gamma_1, \gamma_2, t} = \frac{1}{|\mathcal{I}(\nu, \gamma_1, \gamma_2)|} \sum_{i \in \mathcal{I}(\nu, \gamma_1, \gamma_2)} x_{it}, \quad \text{for all } t = 1, \dots, T.$$

To derive consistency of the post-spectral estimator  $\hat{\beta}$ , we will use the following conditions:

**Assumption 3.10.** *For some constant  $c_3 > 0$ , the minimal eigenvalue of the matrix*

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_{\nu, \nu_i, g_i, t})(x_{it} - \bar{x}_{\nu, \nu_i, g_i, t})'$$

*is bounded from below by  $c_3$  with probability  $1 - o(1)$  uniformly over  $\nu \in \{1, \dots, G\}^N$ .*

**Assumption 3.11.** *(i) The set  $\mathcal{B}$  is compact and (ii) for some constant  $C_5 > 0$ , all elements  $\{a_{\gamma t}\}_{\gamma, t=1}^{G, T}$  of the set  $\mathcal{A}_{G, T}$  satisfy  $|a_{\gamma t}| \leq C_5$  for all  $\gamma = 1, \dots, G$  and  $t = 1, \dots, T$ .*

Assumption 3.10 requires that covariates  $x_{it}$  have sufficient within-group variation over time and across units. This assumption was used in [9] as well. Assumption 3.11 is a standard compactness condition used in the statistical analysis of non-linear models.

**Theorem 3.3** (Consistency of Post-Spectral Estimator  $\hat{\beta}$ ). *Under Assumptions 3.1–3.6 and 3.8–3.11, we have  $\hat{\beta} \rightarrow_P \beta$ .*

Finally, we prove root- $NT$  consistency and asymptotic normality of the post-spectral estimator  $\hat{\beta}$ . To do so, denote

$$\bar{x}^{\gamma, t} = \frac{1}{N_\gamma} \sum_{i: g_i = \gamma} x_{it}, \quad \text{for all } \gamma = 1, \dots, G, \quad t = 1, \dots, T$$

and

$$\check{x}_{it} = x_{it} - \bar{x}^{g_i, t}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T.$$

---

observation that the projection of a Gaussian random vector on any subspace remains Gaussian but in fact the subspace in their construction is random, in which case the projection may not be Gaussian.

We will use the following condition:

**Assumption 3.12.** We have (i)  $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it} \check{x}'_{it} \rightarrow_P \check{\Sigma}$  for some positive-definite  $d \times d$  matrix  $\check{\Sigma}$  and (ii)  $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T v_{it} \check{x}_{it} \rightarrow_D N(0, \Omega)$  for some symmetric  $d \times d$  matrix  $\Omega$ .

This assumption is similar to the corresponding assumptions in [9].

**Theorem 3.4** (Asymptotic Distribution of Post-Spectral Estimator  $\hat{\beta}$ ). *Under Assumptions 3.1–3.12, we have*

$$\sqrt{NT}(\hat{\beta} - \beta) \rightarrow N(0_d, \check{\Sigma}^{-1} \Omega \check{\Sigma}^{-1}).$$

**Remark 3.3** (On Assumptions of Theorems 3.1–3.4). Recall that we stated in the Introduction that  $\mathbb{E}[v_{it}] = 0$  and  $\mathbb{E}[z_{it}] = \mathbb{E}[v_{it}x_{it}] = 0_d$ . These identities are quite intuitive and help motivate Assumptions 3.1, 3.2, and 3.3. However, we emphasize that these identities are actually not used in the proofs of Theorems 3.1–3.4: the results of the theorems hinge only on assumptions that are explicitly mentioned in the statements of the theorems. This observation will be helpful below, when we discuss dynamic grouped panel data models. ■

## 4. EXTENSIONS

In this section, we consider three extensions of the model we studied above. The first extension is concerned with a dynamic version of the model, i.e. a model that allows for lagged values of  $y_{it}$  on the right-hand side of equation (1). The second extension is concerned with a high-dimensional version of the model, i.s. a model with high-dimensional  $\beta$ . The third extension is concerned with an interactive fixed-effect model.

**4.1. Dynamic Model.** Consider a dynamic grouped panel data model

$$y_{it} = \theta y_{it-1} + x'_{it} \beta + \alpha_{git} + v_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (15)$$

where  $x_{it}$  is a  $d \times 1$  vector of pre-determined covariates. As before, assume also that (3) is satisfied. This model is different from the model we studied above as we now allow for the lagged dependent variable on the right-hand side of (15). In this section, we explain what changes one has to carry out in the spectral and post-spectral estimators to estimate parameters  $\theta$  and  $\beta$  of this model. We will assume throughout that  $|\theta| < 1$  since otherwise the extension seems difficult. The results below can be easily extended to allow for additional lagged values of  $y_{it}$  on the right-hand side of (15), i.e.  $y_{it-2}$ ,  $y_{it-3}$ , etc.

To motivate our approach, note that iterating (15) and substituting (3) yields

$$\begin{aligned} y_{it-1} &= \theta^{t-1} y_{i0} + \sum_{r=0}^{t-2} \theta^r (x'_{it-r-1} \beta + \alpha_{g_i t-r-1} + v_{it-r-1}) \\ &= \sum_{m=1}^M \rho_{im}^y \alpha_{g_i t}^{m,y} + z_{it}^y, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \end{aligned}$$

where we denoted

$$\begin{aligned} \rho_{im}^y &= 1\{m = 1\} + \rho'_{im} \beta, \quad \text{for all } i = 1, \dots, N, \quad m = 1, \dots, M, \\ \alpha_{\gamma t}^{m,y} &= \sum_{r=0}^{t-2} \theta^r \alpha_{\gamma t-r-1}^m, \quad \text{for all } t = 1, \dots, T, \quad \gamma = 1, \dots, G, \quad m = 1, \dots, M, \\ z_{it}^y &= \theta^{t-1} y_{i0} + \sum_{r=0}^{t-2} \theta^r (z'_{it-r-1} \beta + v_{it-r-1}), \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \end{aligned}$$

where the sum  $\sum_{r=0}^{-1}$  is treated as zero. Hence, we can write

$$y_{it} = \dot{x}'_{it} \dot{\beta} + \alpha_{g_i t} + v_{it},$$

where we denoted  $\dot{x}_{it} = (y_{it-1}, x'_{it})'$  and  $\dot{\beta} = (\theta, \beta)'$  and the vector of covariates  $\dot{x}_{it}$  satisfies

$$\dot{x}_{it} = \sum_{m=1}^{2M} \dot{\rho}_{im} \dot{\alpha}_{g_i t}^m + \dot{z}_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T,$$

where

$$\begin{aligned} \dot{\rho}_{im} &= (0, \rho'_{im})' 1\{m \leq M\} + (\rho'_{im-M}, 0'_d)' 1\{m > M\}, \\ \dot{\alpha}_{g_i t}^m &= \alpha_{g_i t}^m 1\{m \leq M\} + \alpha_{g_i t}^{m-M,y} 1\{m > M\}, \quad \text{and } \dot{z}_{it} = (z_{it}^y, z'_{it})'. \end{aligned}$$

Thus, the dynamic model considered here reduces to the model studied in Sections 2 and 3 with  $x_{it}$ ,  $\beta$ ,  $\rho_{im}$ ,  $\alpha_{\gamma t}^m$ ,  $z_{it}$ , and  $M$  replaced by  $\dot{x}_{it}$ ,  $\dot{\beta}$ ,  $\dot{\rho}_{im}$ ,  $\dot{\alpha}_{\gamma t}^m$ ,  $\dot{z}_{it}$ , and  $2M$ , respectively. Therefore, the parameters of the dynamic model,  $\dot{\beta} = (\theta, \beta)'$ , can be estimated by the spectral and post-spectral estimators as in Section 2 with  $M$ ,  $x_{it}$ ,  $d$ , and  $\mathcal{B}$  replaced by  $2M$ ,  $\dot{x}_{it}$ ,  $d + 1$ , and  $\dot{\mathcal{B}}$ , respectively, where  $\dot{\mathcal{B}}$  is a parameter space for  $\dot{\beta}$ , e.g.  $\dot{\mathcal{B}} = [-1, 1] \times \mathcal{B}$ .

To state the formal results, let Assumptions 4.1–4.12 be the same as Assumptions 3.1–3.12 with  $z_{it}$ ,  $x_{it}$ ,  $d$ ,  $\Sigma$ , and  $\mathcal{B}$  replaced by  $\dot{z}_{it}$ ,  $\dot{x}_{it}$ ,  $d + 1$ ,  $\dot{\Sigma}$ , and  $\dot{\mathcal{B}}$ , respectively.

**Theorem 4.1** (Dynamic Model). *In the setting of this section, the statements of Theorems 3.1–3.4 continue to hold, with  $\beta$  replaced by  $\dot{\beta}$ , if Assumptions 3.1–3.12 are replaced by Assumptions 4.1–4.12.*



**Remark 4.1** (Relation between Assumptions 3.1–3.12 and Assumptions 4.1–4.12). Assumptions 4.1–4.12 are stronger than Assumptions 3.1–3.12 in the sense that they require Assumptions 3.1–3.12 to be supplemented by some extra conditions. We provide a detailed analysis of these conditions in Appendix B but we note here that one of these extra conditions is that the noise variables  $v_{it}$  satisfy

$$\mathbb{E}[v_{it}|y_{it-1}, \dots, y_{i0}, x_{it}, \dots, x_{i1}] = 0, \quad \text{for all } i = 1, \dots, N, t = 1, \dots, T, \quad (16)$$

which is rather standard in the literature and allows for predetermined covariates  $x_{it}$ . In addition, we note that these conditions do not require the random variables  $y_{i0}$  to have mean zero. The latter means that the random vectors  $\overset{\circ}{z}_{it}$  may not be centered but this does not contradict the results of Theorem 4.1 per our discussion in Remark 3.3 ■

**4.2. High-Dimensional Model.** Consider a high-dimensional grouped panel data model

$$y_{it} = x'_{it}\beta + \alpha_{git} + v_{it}, \quad \text{for all } i = 1, \dots, N, t = 1, \dots, T,$$

where  $x_{it}$  is a  $d \times 1$  vector of covariates and  $d$  can be large, potentially much larger than  $NT$ , but the vector of coefficients  $\beta$  is sparse, i.e.  $s = \|\beta\|_0 = \sum_{k=1}^d 1\{\beta_k \neq 0\}$  is relatively small (in the sense to be made precise later). As before, assume also that (3) is satisfied. Estimating this model requires introducing penalized versions of the spectral and post-spectral estimators. For brevity, however, we focus here on the penalized spectral estimator only, as deriving results for the penalized post-spectral estimator requires taking care of a lot of technicalities but does not seem to bring any new insight.

To define the penalized spectral estimator, let  $\hat{S}$  and  $\hat{\Sigma}$  be the same  $d \times 1$  vector and  $d \times d$  matrix as those appearing in Section 2.1. Note that calculating these quantities requires only  $O(d^2)$  operations, and thus is computationally rather simple. We then define the penalized spectral estimator as

$$\hat{\beta}_\lambda = \arg \min_{b \in \mathbb{R}^d} \left\{ b' \hat{\Sigma} b + \hat{S}' b + \lambda \|b\|_1 \right\},$$

where  $\lambda > 0$  is a penalty parameter and  $\|b\|_1 = \sum_{k=1}^d |b_k|$  denotes the  $\ell^1$ -norm of the vector  $b = (b_1, \dots, b_d)'$ . The optimization problem here is convex and can be carried out using standard software.

To analyze this estimator, we are going to rely on the triangular array asymptotics, where the model, as well as the dimension  $d$  of the vector of covariates, are allowed to depend on  $N$  and  $T$  but for brevity of notation, we keep this dependence implicit. Also, we now have to modify Assumptions 3.3 and 3.4. Indeed, Assumption 3.3 is

impossible to satisfy if  $d$  is growing together with  $N$  and  $T$  because of the  $\ell^2$ -norm appearing in the assumption. Assumption 3.4 also has to be modified as the concept  $O_P(\cdot)$  is not well-defined when applied to matrices of growing dimensions. However, since the required modification are only used in this subsection, we spell them directly in the statement of the theorem. In addition, to state the formal result, for any matrix  $A$ , we will use  $\|A\|_\infty$  to denote its  $\ell^\infty$ -norm, i.e. the maximum of absolute values of components of  $A$ .

**Theorem 4.2** (High-Dimensional Model). *In the setting of this section, suppose that Assumptions 3.1 and 3.2 are satisfied. In addition, suppose that*

$$\max_{1 \leq k \leq d} \left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} z_{itk} \right| = O_P \left( \sqrt{\frac{\log d}{NT}} \right) \quad (17)$$

and

$$\max_{1 \leq k, l \leq d} \left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T z_{itk} z_{itl} - \Sigma_{kl} \right| = O_P \left( \sqrt{\frac{\log d}{NT}} \right) \quad (18)$$

for some positive definite  $d \times d$  matrix  $\Sigma$ . Also, suppose that for some constant  $C_z > 0$ , we have  $\|\sum_{t=1}^T u_t z'_{it} \beta\|_{\psi_2} \leq C_z$  for all  $i = 1, \dots, N$  and  $u = (u_1, \dots, u_T)' \in \mathcal{S}^T$ . Moreover, suppose that  $\log d = o(N)$  and  $\|\beta\|_1 \leq C_\beta$  for some constant  $C_\beta > 0$ . Finally, let  $c_\Sigma > 0$  be the minimal eigenvalue of the matrix  $\Sigma$ ,  $c_\lambda > 1$  be some constant, and  $S = -2\Sigma\beta$  be a  $d \times 1$  vector. Then

$$\|\hat{S} - S\|_\infty \vee \|\hat{\Sigma} - \Sigma\|_\infty = O_P \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right). \quad (19)$$

Moreover, on the intersection of events

$$\lambda \geq c_\lambda \left( \|\hat{S} - S\|_\infty + 2C_\beta \|\hat{\Sigma} - \Sigma\|_\infty \right) \quad (20)$$

and

$$s \|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{c_\Sigma}{2(1 + \bar{c}_\lambda)^2}, \quad (21)$$

we have

$$\|\hat{\beta}_\lambda - \beta\|_1 \leq \frac{2(1 + c_\lambda)(1 + \bar{c}_\lambda)s\lambda}{c_\Sigma c_\lambda} \quad \text{and} \quad \|\hat{\beta}_\lambda - \beta\| \leq \frac{2(1 + c_\lambda)\sqrt{s}\lambda}{c_\Sigma c_\lambda} \quad (22)$$

where  $\bar{c}_\lambda = (c_\lambda + 1)/(c_\lambda - 1)$ .

**Remark 4.2** (Main Implication of Theorem 4.2). By the result (19), it follows that selecting a constant  $C > 0$  large enough and setting

$$\lambda = C \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right) \quad (23)$$

will ensure that the event (20) holds with probability arbitrarily close to one. Also, the result (19) ensures that the event (21) holds with probability approaching one as long as

$$s \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right) \rightarrow 0. \quad (24)$$

Thus, setting  $\lambda$  according to (23) and assuming (24) ensures that

$$\|\hat{\beta}_\lambda - \beta\|_1 \leq \bar{C}s \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right) \text{ and } \|\hat{\beta}_\lambda - \beta\|_2 \leq \bar{C}\sqrt{s} \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right)$$

with probability arbitrarily close to one, where  $\bar{C}$  is some constant. Here, (24) explains how small  $s$  has to be in order for our results to go through.

The choice of the penalty parameter  $\lambda$  in (23) is not practical, as it does not specify the constant  $C > 0$ . From the asymptotic point of view, we can of course set  $C = \log \log n$ , in which case the bounds on  $\|\hat{\beta}_\lambda - \beta\|_1$  and  $\|\hat{\beta}_\lambda - \beta\|_2$  presented above hold with probability approaching one, with  $\bar{C}$  replaced by  $\bar{C} \log \log n$ , but this choice may of course perform poorly in finite samples. However, the question of practical choices of  $\lambda$  is beyond the scope of this paper and we leave it for future work. ■

**Remark 4.3** (Conditions of Theorems 4.2). Conditions (17) and (18) used in Theorem 4.2 are a natural extension of conditions used in the literature on high-dimensional models and can be derived from appropriate maximal inequalities. For example, assuming that components of  $z_{it}$  are bounded and  $v_{it}$  is sub-Gaussian, both (17) and (18) follow from a combination of the union bound and a time-series version of Hoeffding's inequality as long as the dependence of random vectors  $(v_{it}, z'_{it})'$  across  $t$  is not too strong. Moreover, a version of Theorem 4.2 can be derived if conditions (17) and (18) are relaxed, in which case we would simply have to correspondingly modify the bound (19). ■

**4.3. Interactive Fixed-Effect Model.** Consider an interactive fixed-effect panel data model

$$y_{it} = x'_{it}\beta + \kappa'_i\phi_t + v_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (25)$$

where  $\phi_t$  is a  $J \times 1$  vector of factors and  $\kappa_i$  is a  $J \times 1$  vector of factor loadings. [7] developed an OLS-type interactive fixed-effect estimator of  $\beta$  in this model. The proposed estimator, however, is a solution to a non-convex optimization problem, and like in the case of the grouped fixed-effect estimator, it may be difficult to find (ensure that we found) the global solution to this optimization problem. To fix this issue, [24] developed an estimator of  $\beta$  based on the nuclear-norm minimization. Their estimator solves a convex optimization problem and so is computationally rather simple. However, the convergence rate of their estimator is only  $(T \wedge N)^{-1/2}$ , which can be rather slow, especially if  $T \ll N$  or  $N \ll T$ . It is therefore of interest to look for alternative estimators of  $\beta$  in this model.

In this subsection, we demonstrate that a trivial modification of our spectral estimator can be used to estimate  $\beta$  in this model with the rate  $(T \wedge N)^{-1}$  as long as we assume that the covariates  $x_{it}$  satisfy the factor model as well, namely

$$x_{it} = \omega_i' \phi_t + z_{it}, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T, \quad (26)$$

where  $\phi_t$  is a  $J \times 1$  vector of factors and  $\omega_i$  is a  $J \times d$  matrix of factor loadings. Note here that it is without loss of generality to assume that the same factors  $\phi_t$  appear both in (25) and in (26) as we can always merge the factors from two equations.

Let  $\tilde{\beta}_J$  be the same spectral estimator  $\tilde{\beta}$  as that defined in Section 2.1 with  $J$  replacing  $GM$ . We then have the following result.

**Theorem 4.3** (Interactive Fixed-Effect Model). *In the setting of this section, suppose that Assumptions 3.1–3.4 are satisfied. Then*

$$\tilde{\beta}_J = \beta + O_P\left(\frac{1}{T \wedge N}\right).$$

**Remark 4.4** (Combining  $\tilde{\beta}_J$  and Interactive Fixed-Effect Estimator). Although the convergence rate of our estimator  $\tilde{\beta}_J$  is faster than the convergence rate of the estimator proposed in [24], it is still slower than the convergence rate of the interactive fixed-effect estimator proposed in [7], which is  $(NT)^{-1/2}$ . Like in [24], we therefore consider our estimator as a starting point in the problem of finding a local minimum of the optimization problem used to define the interactive fixed-effect estimator. The resulting estimator will be asymptotically equivalent to the interactive fixed-effect estimator under certain conditions. ■

## 5. MONTE CARLO SIMULATION STUDY

In this section, we present results of a Monte Carlo simulation study. The study consists of two parts. In the first part, we demonstrate that the finite sample performance of the post-spectral estimator is comparable to that of the grouped fixed-effect estimator and also compare the performance of the post-spectral estimator to that of the spectral estimator. To be upfront, we do not expect the post-spectral estimator to outperform the grouped fixed-effect estimator if the latter is computed as the *global* minimum of the grouped fixed-effect optimization problem, and we do make effort to find the global minimum, as described below. In the second part, we quantify potential issues with the grouped fixed-effect estimator if we fail to find the global minimum of the grouped fixed-effect optimization problem. Specifically, we demonstrate that there exist data-generating processes where finding the global minimum of this optimization problem is difficult and using local minima instead may substantially undermine the performance of the grouped fixed-effect estimator. Throughout the study, for all data-generating processes considered, we also report results for the oracle estimator, which is the pooled OLS estimator based on the true partition of units into groups. Although we do not discuss these results, an interested reader can use them as a benchmark for the results on the spectral, post-spectral, and grouped fixed effect estimators.

For the first part of the simulation study, we generate the data  $(x'_{it}, y_{it})'$ ,  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , according to equations (1) and (3), where we set  $d = 2$ , so that the model contains two covariates. Depending on the experiment, we set  $N = 100, 200$ , or  $300$  and  $T = 20, 50$ , or  $100$ . We also set  $G = 2, 4$ , or  $7$  and  $M = 1$  or  $2$ . The random variables  $z_{it}$ ,  $v_{it}$ , and  $\alpha_{\gamma t}^m$  are generated from a truncated standard normal distribution,  $Z1\{|Z| \leq C\}$ , where  $Z \sim N(0, 1)$  and  $C = 5$ , independently of each other and across indices. We have checked that results are robust with respect to different specifications of the truncation constant  $C$ . Further, we set the number of observations  $i$  within each group  $\gamma$  to be the same, except for the last group, which contains more observations if  $G$  does not divide  $N$ .

Throughout all experiments, we set  $\beta = (\beta_1, \beta_2)'$ , where  $\beta_1 = -1$  and  $\beta_2 = 0.8$  but we note that the particular choice of these values does not seem to matter much for the simulation results. Also, when  $M = 1$ , we set  $\rho_{i1} = (\rho_{i11}, \rho_{i12})'$  with  $\rho_{i11} = \varrho + Z_{i1}1\{|Z_{i1}| \leq C\}$  and  $\rho_{i12} = Z_{i2}1\{|Z_{i2}| \leq C\}$ , where  $Z_{i1}$  and  $Z_{i2}$  are independent of each other, across  $i$ , and of all other random variables in the data-generating process. Depending on the experiment, we set  $\varrho = 1$  or  $3$ . Here,  $\varrho$  can be understood as a measure of endogeneity in the model as it governs the correlation between covariates

and grouped-specific time effects. Finally, when  $M = 2$ , we set  $\rho_{i11} = \varrho + Z_{i11}1\{|Z_{i11} \leq C\}$ ,  $\rho_{i21} = 1 + Z_{i21}1\{|Z_{i21} \leq C\}$ , and  $\rho_{im2} = Z_{im2}1\{|Z_{im2} \leq C\}$  for  $m = 1$  and  $2$ , where random variables  $Z_{imj}$  have standard normal distribution and are independent of each other, across indices, and of all other random variables in the data-generating process. Like above, we set  $\varrho = 1$  or  $3$  depending on the experiment.

Next, we note that calculating spectral and post-spectral estimators is straightforward but calculating the grouped fixed-effect estimator is not. The later is so because even with  $N = 100$  and  $G = 2$ , it is never possible to be sure that any given algorithm produces the global minimum of the grouped fixed-effect optimization problem since this would require separately considering  $G^N$  partitions of units  $i = 1, \dots, N$  into  $G$  groups. Having in mind this potential issue, which we elaborate on more in the second part of our simulation study, we calculate the grouped fixed-effect estimator repeating Algorithm 1 in [9] with 30 different initial values of parameters and taking the result of the algorithm corresponding to the lowest value of the criterion function to be the grouped fixed-effect estimator. We generate the first 29 initial values from the standard normal distribution in  $\mathbb{R}^{2+G \times T}$  (here, the first two components correspond to  $\beta = (\beta_1, \beta_2)'$  and the last  $G \times T$  components correspond to  $\{\alpha_{\gamma t}\}_{\gamma, t=1}^{G, T}$ ). The 30th initial value is taken to be the oracle estimator, i.e. the pooled OLS estimator based on the true partition of units into groups. This value would of course be infeasible in estimation based on a real dataset but using this value substantially increases our chances to find the global minimum of the grouped fixed-effect optimization problem as we expect the solution to be near the oracle estimator. We use a relatively small number of initial values because even with this number calculating the grouped fixed-effect estimator takes non-trivial time.

We present results of the first part of the simulation study in Tables 1–4. These tables give the mean absolute error (MAE) for spectral, post-spectral, grouped fixed-effect, and oracle estimators. In addition, the tables give the fraction of misclassified units based on our classification algorithm (which is based on the spectral estimator, and which is the second step in constructing our post-spectral estimator) and the fraction of misclassified units in the construction of the grouped fixed-effect estimator. Tables 1 and 2 correspond to  $M = 1$  and Tables 3 and 4 correspond to  $M = 2$ . Both the MAE and the fraction of misclassified units are calculated as the average over 50 simulations.

Tables 1–4 reveal the following patterns. First, post-spectral and grouped fixed-effect estimators do not dominate each other. For example, when  $M = 1$ ,  $\varrho = 1$ ,  $G = 7$ ,  $N = 100$  and  $T = 20$  (Table 1), the MAE of the grouped fixed-effect estimator

is 6.5 times smaller than the MAE of the post-spectral estimator, whereas when  $\varrho = 3$ ,  $M = 1$ ,  $G = 7$ ,  $N = 200$ , and  $T = 20$  (Table 2), the MAE of the grouped fixed-effect estimator is 3.3 times larger than the MAE of the post-spectral estimator. However, it seems fair to say that the performance of the grouped fixed-effect estimator is better than that of the post-spectral estimator when  $M = 2$  (Tables 3 and 4). On the other hand, two estimators perform equally well when good classification of units into groups is possible. Second, the post-spectral estimator tends to outperform the spectral estimator when  $G$  is small and/or  $T$  is large, which is expected from the theory because this is exactly the case when good classification of units into groups is possible. On the other hand, two estimators do not dominate each other when  $G$  is large or  $T$  is small, so that good classification of units into groups is not possible.

Next, we turn to the second part of the simulation study. For this part, we use the same data-generating processes as those in the first part corresponding to  $M = 1$  (Tables 1 and 2) with the following changes. First, we set  $\varrho = 6$  and  $G = 5$ . Second, depending on the experiment, we consider  $N = 100, 250, \text{ or } 500$  and  $T = 20, 40, 80, \text{ or } 160$ . Third, we change the distribution of group-specific time effects  $\alpha_{\gamma t}$  (remember that  $M = 1$ , and  $\alpha_{\gamma t}^1 = \alpha_{\gamma t}$ ). To describe the new distributions, let  $Z_{\gamma t}$  for  $\gamma = 1, \dots, G$  and  $t = 1, \dots, T$  be independent standard normal random variables and let  $s = 0.5$ . We set  $\alpha_{1t} = Z_{1t}1\{|Z_{1t}| \leq C\}$  and  $\alpha_{\gamma t} = \tilde{Z}_{\gamma t}1\{\tilde{Z}_{\gamma t} \leq C\}$  for all  $\gamma = 2, \dots, G$ , where  $\tilde{Z}_{\gamma t} = sZ_{\gamma t} + \sqrt{1 - s^2}Z_{1t}$ . This allows us to generate correlated over groups time effects. For the described data-generating processes, we compute spectral, post-spectral, oracle, and two types of group fixed-effect estimators, GFE1 and GFE2. The GFE1 estimator is computed using the same algorithm as in the first part of the study but with 50 initial values, where 49 initial values are random and the 50th initial value is the oracle estimator. The GFE2 estimator is computed by the same algorithm but all 50 initial values are random. We now use 50 instead of 30 initial values because we would like to demonstrate that even 50 initial values may not be sufficient to find the global minimum of the grouped fixed-effect optimization problem. Note that the GFE2 estimator is feasible but GFE1 is not. In Table 5, we present the mean absolute error (MAE) for these five estimators calculated as the average over 50 simulations.

The key observation following from Table 5 is that the performance of the *infeasible* GFE1 estimator is comparable to that of the post-spectral estimator but the performance of the *feasible* GFE2 estimator is substantially worse. In the most extreme case,  $N = 500$  and  $T = 160$ , the performance of the GFE2 estimator is 41 times worse than that of the post-spectral estimator. This is because the GFE2 estimator

fails to find the global minimum of the grouped fixed-effect optimization problem. In fact, we find that in all 50 repetitions and for all data-generating processes, i.e. in  $50 \times 4 \times 3 = 600$  simulations, the GFE2 estimator chooses the minimum that is much further away from the true values of parameters than the minimum chosen by the GFE1 estimator. This suggests that using even more initial values for implementation of the GFE2 estimator will not help its performance.

The second part of the simulation study thus illustrates the main motivation for the post-spectral estimator as an alternative to the grouped fixed-effect estimator: finding the global minimum of the grouped fixed-effect optimization problem may sometimes be difficult, and if we fail to find the global minimum, the post-spectral estimator can substantially outperform the grouped fixed-effect estimator.

## 6. PROOFS OF MAIN RESULTS

*Proof of Theorem 3.1.* Throughout this proof, we use  $c$  and  $C$  to denote strictly positive constants that can change from place to place but can be chosen to depend on  $C_1$  and  $C_2$  only.

We will prove that

$$\lambda_1^b + \cdots + \lambda_{2GM+2}^b = b' \Sigma b + S' b + L + O_P \left( \frac{1}{T \wedge N} \right), \quad \text{for all } b \in \mathbb{R}^d, \quad (27)$$

where  $\Sigma$  is a positive definite matrix appearing in Assumption 3.4,  $S = -2\Sigma\beta$ , and  $L = L_N = \beta' \Sigma \beta + 2(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T v_{it}^2$ . Then  $\beta = -\Sigma^{-1}S/2$ , and so  $\tilde{\beta} = -\hat{\Sigma}^{-1}\hat{S}/2$  satisfies (14) by the delta method since (27) implies that

$$\|\hat{S} - S\| \vee \|\hat{\Sigma} - \Sigma\| = O_P \left( \frac{1}{T \wedge N} \right)$$

by construction in Section 2.1.

To prove (27), fix any  $b \in \mathbb{R}^d$ . For brevity of notation, we will write  $A$  instead of  $A^b$ . Then

$$A_{ij} = \frac{1}{NT} \sum_{t=1}^T \left\{ f_{it} - f_{jt} + z'_{it}(\beta - b) - z'_{jt}(\beta - b) + v_{it} - v_{jt} \right\}^2, \quad i, j = 1, \dots, N,$$

where

$$f_{it} = \alpha_{git}^1 (1 + \rho'_{i1}(\beta - b)) + \sum_{m=2}^M \alpha_{git}^m \rho'_{im}(\beta - b), \quad i = 1, \dots, N, t = 1, \dots, T.$$



Also, let  $R$  be an  $N \times N$  matrix whose  $(i, j)$ -th element is

$$R_{ij} = -\frac{2}{NT} \sum_{t=1}^T \left\{ (\beta-b)' z_{it} z_{jt} (\beta-b) + v_{it} v_{jt} + v_{it} z'_{jt} (\beta-b) + v_{jt} z'_{it} (\beta-b) \right\}, \quad i, j = 1, \dots, N.$$

As we will see below, this matrix represents the asymptotically negligible component of  $A$  in the sense that

$$\|R\| = O_P \left( \frac{1}{T \wedge N} \right). \quad (28)$$

On the other hand,

$$\begin{aligned} \text{tr}(R) &= -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta-b)' z_{it} z_{it} (\beta-b) + v_{it}^2 + 2v_{it} z'_{it} (\beta-b) \right\} \\ &= -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta-b)' z_{it} z_{it} (\beta-b) + v_{it}^2 \right\} + O_P \left( \frac{1}{\sqrt{NT}} \right) \end{aligned}$$

by Assumption 3.3. Thus, given that  $\text{tr}(A) = 0$ , the matrix  $A_0 = A - R$  satisfies

$$\begin{aligned} \text{tr}(A_0) &= \frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta-b)' z_{it} z_{it} (\beta-b) + v_{it}^2 \right\} + O_P \left( \frac{1}{\sqrt{NT}} \right) \\ &= b' \Sigma b + S' b + L + O_P \left( \frac{1}{\sqrt{NT}} \right) \end{aligned}$$

by Assumption 3.4. In addition, we will show below that  $A_0$  has at most  $2GM + 2$  non-zero eigenvalues. Hence,

$$\lambda_1 + \dots + \lambda_{2GM+2} = b' \Sigma b + S' b + L + O_P \left( \frac{1}{\sqrt{NT}} \right) + O_P \left( \frac{1}{T \wedge N} \right)$$

by Lemma 7.1 and (28). This gives (27) since  $\sqrt{NT} \geq T \wedge N$ . Therefore, it remains to prove (28) and to show that  $A_0$  has at most  $2GM + 2$  non-zero eigenvalues.

To derive the bound on the number of non-zero eigenvalues of  $A_0$ , let us first introduce some notation. For all  $t = 1, \dots, T$ , denote

$$F_t = (f_{1t}, \dots, f_{Nt})', \quad V_t = (v_{1t}, \dots, v_{Nt})', \quad Z_t = (z'_{1t} (\beta-b), \dots, z'_{Nt} (\beta-b))'.$$

Also, denote  $1_N = (1, \dots, 1)' \in \mathbb{R}^N$ . In addition, for any vectors  $a = (a_1, \dots, a_N)'$  and  $b = (b_1, \dots, b_N)'$ , write  $(ab) = (a_1 b_1, \dots, a_N b_N)'$  and  $a^2 = (a_1^2, \dots, a_N^2)'$ . Then

$$\begin{aligned} A &= \frac{1}{NT} \sum_{t=1}^T \left( F_t^2 1'_N + 1_N (F_t^2)' + Z_t^2 1'_N + 1_N (Z_t^2)' + V_t^2 1'_N + 1_N (V_t^2)' \right) \\ &\quad + \frac{2}{NT} \sum_{t=1}^T \left( (F_t Z_t) 1'_N + 1_N (F_t Z_t)' + (F_t V_t) 1'_N + 1_N (F_t V_t)' + (Z_t V_t) 1'_N + 1_N (Z_t V_t)' \right) \end{aligned}$$

$$-\frac{2}{NT} \sum_{t=1}^T \left( F_t Z_t' + Z_t F_t' + F_t V_t' + V_t F_t' + Z_t V_t' + V_t Z_t' + F_t F_t' + Z_t Z_t' + V_t V_t' \right).$$

Also,

$$R = -\frac{2}{NT} \sum_{t=1}^T \left( Z_t Z_t' + V_t V_t' + V_t Z_t' + Z_t V_t' \right).$$

Hence, denoting

$$Q = \frac{1}{NT} \sum_{t=1}^T \left( F_t^2 + Z_t^2 + V_t^2 + 2(F_t Z_t) + 2(F_t V_t) + 2(Z_t V_t) \right)$$

and recalling  $A_0 = A - R$ , we have

$$\begin{aligned} A_0 &= Q1'_N + 1_N Q' - \frac{2}{NT} \sum_{t=1}^T \left( F_t F_t' + F_t Z_t' + Z_t F_t' + F_t V_t' + V_t F_t' \right) \\ &= Q1'_N + 1_N Q' - \frac{2}{NT} \sum_{t=1}^T F_t (F_t + Z_t + V_t)' - \frac{2}{NT} \sum_{t=1}^T (Z_t + V_t) F_t'. \end{aligned}$$

The last expression makes bounding the number of non-zero eigenvalues of the matrix  $A_0$  straightforward. Indeed,  $\text{rank}(Q1'_N) = \text{rank}(1_N Q') = 1$ . Also, for each  $m = 1, \dots, M$  and  $\gamma = 1, \dots, G$ , define an  $N \times 1$  vector  $\varrho_{m\gamma}$  whose  $i$ -th element is

$$\varrho_{m\gamma i} = \begin{cases} \mathbb{I}\{m=1\} + \rho'_{im}(\beta - b) & \text{if } g_i = \gamma, \\ 0 & \text{if } g_i \neq \gamma, \end{cases} \quad i = 1, \dots, N.$$

Then

$$F_t = \sum_{m=1}^M \sum_{\gamma=1}^G \alpha_{\gamma t}^m \varrho_{m\gamma}, \quad t = 1, \dots, T,$$

and so, for any vector  $a = (a_1, \dots, a_N)'$ , the vector

$$\begin{aligned} \sum_{t=1}^T F_t (F_t + Z_t + V_t)' a &= \sum_{m=1}^M \sum_{\gamma=1}^G \sum_{t=1}^T \alpha_{\gamma t}^m \varrho_{m\gamma} (F_t + Z_t + V_t)' a \\ &= \sum_{m=1}^M \sum_{\gamma=1}^G \varrho_{m\gamma} \sum_{t=1}^T \alpha_{\gamma t}^m (F_t + Z_t + V_t)' a \end{aligned}$$

belongs to the linear subspace of  $\mathbb{R}^N$  spanned by the vectors  $\varrho_{11}, \dots, \varrho_{MG}$ . Hence,

$$\text{rank} \left( \sum_{t=1}^T F_t (F_t + Z_t + V_t)' \right) \leq MG.$$

Similarly, using the fact that the row rank is equal to the column rank,

$$\text{rank} \left( \sum_{t=1}^T (Z_t + V_t) F_t' \right) \leq MG.$$

Thus, given that the rank operator is sub-additive, we conclude that

$$\text{rank}(A_0) \leq 2MG + 2,$$

and so  $A_0$  has at most  $2MG + 2$  non-zero eigenvalues, as claimed.

It remains to prove (28). To do so, we will show that

$$\left\| \frac{1}{NT} \sum_{t=1}^T V_t V_t' \right\| = O_P \left( \frac{1}{T \wedge N} \right) \quad (29)$$

and note that  $\|(NT)^{-1} \sum_{t=1}^T Z_t Z_t'\| = O_P(1/(T \wedge N))$  follows from the exactly same argument (the former is obtained from Assumption 3.1 whereas the latter is from Assumption 3.2). Then

$$\left\| \frac{1}{NT} \sum_{t=1}^T V_t Z_t' \right\| \leq \sqrt{\left\| \frac{1}{NT} \sum_{t=1}^T V_t V_t' \right\|} \sqrt{\left\| \frac{1}{NT} \sum_{t=1}^T Z_t Z_t' \right\|} = O_P \left( \frac{1}{T \wedge N} \right)$$

by Lemma 7.2. This gives (28).

To prove (29), we proceed by appropriately modifying the proof of Theorem 4.7.1 in [32]. Denote  $\mathcal{V} = (v_1, \dots, v_N)'$ , where  $v_i = (v_{i1}, \dots, v_{iT})'$  for all  $i = 1, \dots, N$ . Then uniformly over  $u \in \mathcal{S}^T$ ,

$$u' \mathbb{E} \left[ \frac{\mathcal{V}' \mathcal{V}}{N} \right] u = \frac{1}{N} \sum_{i=1}^N \mathbb{E} [u' v_i v_i' u] = \frac{1}{N} \sum_{i=1}^N \mathbb{E} [(v_i' u)^2] \leq C$$

by Assumption 3.1. Thus,

$$\left\| \mathbb{E} \left[ \frac{\mathcal{V}' \mathcal{V}}{N} \right] \right\| \leq C. \quad (30)$$

Further, by Exercise 4.4.3, part 2, in [32],

$$\left\| \frac{\mathcal{V}' \mathcal{V}}{N} - \mathbb{E} \left[ \frac{\mathcal{V}' \mathcal{V}}{N} \right] \right\| \leq 2 \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{V}' \mathcal{V}}{N} - \mathbb{E} \left[ \frac{\mathcal{V}' \mathcal{V}}{N} \right] \right) u \right|, \quad (31)$$

where  $\mathcal{N}$  is any  $(1/4)$ -net in  $\mathcal{S}^T$ . Moreover, by Corollary 4.2.13 in [32], the net  $\mathcal{N}$  can be chosen so that  $|\mathcal{N}_\epsilon| \leq 9^T$ , which we are going to use below.

Next, fix any  $u \in \mathcal{N}$ . Then by Assumption 3.1 and Lemma 2.7.6 in [32], for all  $i = 1, \dots, N$ , we have  $\|(u' v_i)^2\|_{\psi_1} \leq C$ , and so by Exercise 2.7.10 in [32],  $\|(u' v_i)^2 - \mathbb{E}[(u' v_i)^2]\|_{\psi_1} \leq C$ . Hence, by Bernstein's inequality (Corollary 2.8.3 in [32]), for any

$\epsilon > 0$ ,

$$\begin{aligned} & \mathbb{P} \left( \left| u' \left( \frac{\mathcal{V}'\mathcal{V}}{N} - \mathbb{E} \left[ \frac{\mathcal{V}'\mathcal{V}}{N} \right] \right) u \right| \geq \epsilon \right) \\ &= \mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^N \left( (v'_i u)^2 - \mathbb{E}[(v'_i u)^2] \right) \right| \geq \epsilon \right) \leq 2 \exp[-c(\epsilon \wedge \epsilon^2)N], \end{aligned}$$

and so by the union bound,

$$\mathbb{P} \left( \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{V}'\mathcal{V}}{N} - \mathbb{E} \left[ \frac{\mathcal{V}'\mathcal{V}}{N} \right] \right) u \right| \geq \epsilon \right) \leq 2 \times 9^T \times \exp[-c(\epsilon \wedge \epsilon^2)N].$$

Therefore, setting  $\epsilon = c^{-1}(\log 9)(T/N) + 1$ , we obtain

$$\mathbb{P} \left( \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{V}'\mathcal{V}}{N} - \mathbb{E} \left[ \frac{\mathcal{V}'\mathcal{V}}{N} \right] \right) u \right| \geq \frac{CT}{N} + 1 \right) \leq 2 \exp(-cN).$$

Combining this bound with (30) and (31) gives

$$\left\| \frac{\mathcal{V}'\mathcal{V}}{N} \right\| = O_P \left( \frac{T}{N} + 1 \right). \quad (32)$$

Hence,

$$\|\mathcal{V}\| = O_P \left( \sqrt{T} + \sqrt{N} \right),$$

and so

$$\left\| \frac{1}{T} \sum_{t=1}^T V_t V_t' \right\| = \left\| \frac{\mathcal{V}'\mathcal{V}}{T} \right\| = \frac{\|\mathcal{V}\|^2}{T} = O_P \left( 1 + \frac{N}{T} \right).$$

Conclude that

$$\left\| \frac{1}{NT} \sum_{t=1}^T V_t V_t' \right\| = O_P \left( \frac{1}{T} + \frac{1}{N} \right) = O_P \left( \frac{1}{T \wedge N} \right),$$

which gives (29) and completes the proof.  $\blacksquare$

*Proof of Theorem 3.2.* First, we introduce some notations. For all  $i = 1, \dots, N$ , denote  $v_i = (v_{i1}, \dots, v_{iT})'$  and for all  $\gamma = 1, \dots, G$ , denote  $\alpha_\gamma = (\alpha_{\gamma 1}, \dots, \alpha_{\gamma T})'$ . Then the model (1) can be written as

$$y_i = x_i \beta + \alpha_{g_i} + v_i, \quad i = 1, \dots, N.$$

Further, let  $F$  be a  $T \times G$  matrix whose columns are orthonormal and span the vectors  $\alpha_1, \dots, \alpha_G$ . Then for each  $\gamma = 1, \dots, G$ , there exists a  $G \times 1$  vector  $p_\gamma$  such that  $\alpha_\gamma = \sqrt{T} F p_\gamma$ . Here, without loss of generality, we assume that  $\Lambda = N^{-1} \sum_{i=1}^N p_{g_i} p_{g_i}'$  is a diagonal matrix  $\text{diag}(\lambda_1, \dots, \lambda_G)$  with  $\lambda_1 \geq \dots \geq \lambda_G \geq 0$  since otherwise we can consider the eigenvalue decomposition  $S \Lambda S'$  of the matrix  $N^{-1} \sum_{i=1}^N p_{g_i} p_{g_i}'$  and

replace  $F$  and  $p_1, \dots, p_G$  by  $FS$  and  $S^{-1}p_1, \dots, S^{-1}p_G$ , respectively. Also, throughout this proof, we use  $c$  and  $C$  to denote strictly positive constants that can change from place to place but can be chosen to depend on  $C_1, C_2, C_3, C_4, c_1$ , and  $c_2$  only.

We will show below that

$$\|\hat{A}_i - \alpha_{g_i}\| = o_P(\sqrt{T}) \quad \text{uniformly over } i = 1, \dots, N. \quad (33)$$

Therefore, by Assumption 3.7, there exists a constant  $\bar{C} > 0$  such that with probability approaching one,

$$\|\hat{A}_i - \hat{A}_j\| \leq \bar{C}\sqrt{T} \quad \text{for all } i, j = 1, \dots, N \text{ such that } g_i = g_j \quad (34)$$

and

$$\|\hat{A}_i - \hat{A}_j\| > 2\bar{C}\sqrt{T} \quad \text{for all } i, j = 1, \dots, N \text{ such that } g_i \neq g_j. \quad (35)$$

Now, assume that (34) and (35) are satisfied and let  $\mathcal{A}_1(\lambda), \dots, \mathcal{A}_{m(\lambda)}(\lambda)$  be the partition of units generated by the Classification Algorithm from Section 2.2 for any given value of the tuning parameter  $\lambda$  and let  $\tilde{g}(\lambda) = (\tilde{g}_1(\lambda), \dots, \tilde{g}_N(\lambda))' \in \{1, \dots, m(\lambda)\}^N$  be the corresponding vector of group assignments, so that  $\hat{g} = \tilde{g}(\hat{\lambda})$ . Then observe that for any  $\lambda \leq \bar{C}\sqrt{T}$ , the vector  $\tilde{g}(\lambda)$  has the following property:

$$\text{for all } i, j = 1, \dots, N, \text{ we have } \tilde{g}_i(\lambda) \neq \tilde{g}_j(\lambda) \text{ if } g_i \neq g_j, \quad (36)$$

i.e. units from different groups can not be classified into the same group. To see why this is so, suppose to the contrary that for some  $\lambda \leq \bar{C}\sqrt{T}$ , there exist  $i, j = 1, \dots, N$  such that  $g_i \neq g_j$  but  $\tilde{g}_i(\lambda) = \tilde{g}_j(\lambda)$ . In this case, we can consider the first misclassified unit, say unit  $i_0$ , in the Classification Algorithm. This unit satisfies the following inequality for some  $\mathcal{A} \subset \{1, \dots, N\}$ :

$$\left\| \hat{A}_{i_0} - \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l \right\| \leq \lambda, \quad (37)$$

where all units  $l \in \mathcal{A}$  are coming from the same group and  $i_0$  is coming from another group, i.e.  $g_{i_0} \neq \gamma = g_l$  for all  $l \in \mathcal{A}$ . But then, for any  $j \in \mathcal{A}$ ,

$$\begin{aligned} \|\hat{A}_{i_0} - \hat{A}_j\| &\leq \left\| \hat{A}_{i_0} - \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l \right\| + \left\| \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l - \hat{A}_j \right\| \\ &\leq \lambda + \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \|\hat{A}_l - \hat{A}_j\| \leq 2\bar{C}\sqrt{T}, \end{aligned}$$

where we used the triangle inequality as well as (34) and (37). This contradicts (35), and so (36) indeed holds for all  $\lambda \leq \bar{C}\sqrt{T}$ .

In turn, (36) implies that if  $\tilde{g}_i(\lambda) \neq \tilde{g}_j(\lambda)$  for some  $i, j = 1, \dots, N$  with  $g_i = g_j$ , then  $m(\lambda) > G$ . Therefore, given that  $m(\hat{\lambda}) = G$ , it follows that if  $\hat{\lambda} \leq \bar{C}\sqrt{T}$ , then the vector  $\hat{g} = \tilde{g}(\hat{\lambda})$  has the following property: for all  $i, j = 1, \dots, N$ , we have  $\hat{g}_i = \hat{g}_j$  if and only if  $g_i = g_j$ . But we claim that  $\hat{\lambda}$  indeed satisfies the inequality  $\hat{\lambda} \leq \bar{C}\sqrt{T}$ . To see why this is so, observe that  $m(\bar{C}\sqrt{T}) \geq G$  by (36) and  $m(\bar{C}\sqrt{T}) \leq G$  by (34). Hence,  $m(\bar{C}\sqrt{T}) = G$ , and so  $\hat{\lambda} \leq \bar{C}\sqrt{T}$ , as required. Thus, the asserted claim of the theorem follows since (34) and (35) hold with probability approaching one. It thus remains to prove (33). We do so in eight steps.

**Step 1.** Here we show that

$$\|\hat{F}'_{h_i} v_i\| = o_P(\sqrt{T}) \quad \text{uniformly over } i = 1, \dots, N.$$

To do so, note that for all  $i = 1, \dots, N$ , the random vectors  $v_i$  and  $\hat{F}_{h_i}$  are independent conditional on  $h_1, \dots, h_N$ . Therefore, by Assumption 3.1 and (2.14) in [32], for all  $\epsilon > 0$ ,  $i = 1, \dots, N$ , and  $\gamma = 1, \dots, G$ ,

$$P(|\hat{F}'_{h_i \gamma} v_i| > \epsilon) \leq 2 \exp(-c\epsilon^2),$$

where  $\hat{F}_{h_i \gamma}$  denotes the  $\gamma$ -th column of the matrix  $\hat{F}_{h_i}$ . Hence, by the union bound,

$$P\left(\max_{1 \leq i \leq N} |\hat{F}'_{h_i \gamma} v_i| > \epsilon\right) \leq 2 \exp(\log N - c\epsilon^2).$$

Combining this bound with Assumption 3.9 gives the asserted claim of this step.

**Step 2.** Here we show that

$$\|\hat{F}'_{h_i} x_i (\tilde{\beta}^{h_i} - \beta)\| = o_P(\sqrt{T}) \quad \text{uniformly over } i = 1, \dots, N.$$

To do so, for all  $i = 1, \dots, N$ , denote  $z_i = (z_{i1}, \dots, z_{iT})'$  and

$$\bar{x}_i = \left( \sum_{m=1}^M \rho_{im} \alpha_{g_i 1}^m, \dots, \sum_{m=1}^M \rho_{im} \alpha_{g_i T}^m \right)',$$

so that  $x_i = \bar{x}_i + z_i$ . Then observe that

$$\|\hat{F}'_{h_i} z_i\| = o_P(\sqrt{T}) \quad \text{uniformly over } i = 1, \dots, N$$

by the same argument as that in Step 1, with Assumption 3.2 playing the role of Assumption 3.1. Also,

$$\|\hat{F}'_{h_i} \bar{x}_i\| = O(\sqrt{T}) \quad \text{uniformly over } i = 1, \dots, N$$

by Assumptions 3.5 and 3.6. Moreover,  $\|\tilde{\beta}^0 - \beta\| \vee \|\tilde{\beta}^1 - \beta\| = o_P(1)$  by Theorem 3.1. Combining these bounds gives the asserted claim of this step.

**Step 3.** Here we show that

$$\left( \frac{1}{N} \sum_{i=1}^N v_{it}^2 \right) \vee \left( \frac{1}{N} \sum_{i=1}^N \|z_{it}\|^2 \right) = O_P(1) \quad \text{uniformly over } t = 1, \dots, T.$$

To do so, note that by Assumption 3.1 and Lemma 2.7.6 in [32], for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , we have  $\|v_{it}^2\|_{\psi_1} \leq C$ , and so by Exercise 2.7.10 in [32],  $\|v_{it}^2 - \mathbb{E}[v_{it}^2]\|_{\psi_1} \leq C$ . Hence, by Bernstein's inequality (Corollary 2.8.3 in [32]), for any  $\epsilon > 0$ ,

$$\mathbb{P} \left( \frac{1}{N} \sum_{i=1}^N (v_{it}^2 - \mathbb{E}[v_{it}^2]) > \epsilon \right) \leq \exp[-c(\epsilon \wedge \epsilon^2)N],$$

and so by the union bound,

$$\mathbb{P} \left( \max_{1 \leq t \leq T} \frac{1}{N} \sum_{i=1}^N (v_{it}^2 - \mathbb{E}[v_{it}^2]) > \epsilon \right) \leq T \exp[-c(\epsilon \wedge \epsilon^2)N].$$

Hence, given that  $\mathbb{E}[v_{it}^2] \leq C$  for all  $i = 1, \dots, N$  and  $t = 1, \dots, T$  by Assumption 3.1, it follows from Assumption 3.9 that

$$\frac{1}{N} \sum_{i=1}^N v_{it}^2 = O_P(1) \quad \text{uniformly over } t = 1, \dots, T.$$

Thus, given that

$$\frac{1}{N} \sum_{i=1}^N \|z_{it}^2\| = O_P(1) \quad \text{uniformly over } t = 1, \dots, T$$

can be proven using the same argument, with Assumption 3.1 replaced by 3.2, the asserted claim of this step follows.

**Step 4.** Here we show that

$$\|\hat{B}^0 - \bar{B}^0\| \vee \|\hat{B}^1 - \bar{B}^1\| = o_P(1),$$

where

$$\bar{B}^h = \frac{2}{NT} \sum_{i \in \mathcal{I}_h} (y_i - x_i \beta)(y_i - x_i \beta)', \quad h = 0, 1.$$

To do so, fix  $h = 0, 1$ . Then

$$\begin{aligned} \hat{B}^h - \bar{B}^h &= \frac{2}{NT} \sum_{i \in \mathcal{I}_h} x_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' x_i' \\ &\quad - \frac{2}{NT} \sum_{i \in \mathcal{I}_h} (y_i - x_i \beta) (\tilde{\beta}^h - \beta)' x_i' \end{aligned}$$

$$-\frac{2}{NT} \sum_{i \in \mathcal{I}_h} x_i (\tilde{\beta}^h - \beta) (y_i - x_i \beta)'$$

All three terms here have the spectral norm  $o_P(1)$  but for brevity, we only consider the first term, i.e. we prove that

$$\left\| \frac{1}{NT} \sum_{i \in \mathcal{I}_h} x_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' x_i' \right\| = o_P(1) \quad (38)$$

and note that the other two terms can be bounded similarly. To prove (38), recall that  $x_i = \bar{x}_i + z_i$  as in Step 2. Therefore, by Lemma 7.2, the left-hand side of (38) is bounded from above by

$$2 \left\| \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \bar{x}_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' \bar{x}_i' \right\| + 2 \left\| \frac{1}{NT} \sum_{i \in \mathcal{I}_h} z_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' z_i' \right\|.$$

Here,

$$\left\| \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \bar{x}_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' \bar{x}_i' \right\| \leq \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \|\bar{x}_i\|^2 \|\tilde{\beta}^h - \beta\|^2 = o_P(1)$$

by Theorem 3.1 and Assumptions 3.5 and 3.6. Also, given that  $(NT)^{-1} \sum_{i \in \mathcal{I}_h} z_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' z_i'$  is a positive-definite matrix,

$$\begin{aligned} & \left\| \frac{1}{NT} \sum_{i \in \mathcal{I}_h} z_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' z_i' \right\| \\ & \leq \text{tr} \left\{ \frac{1}{NT} \sum_{i \in \mathcal{I}_h} z_i (\tilde{\beta}^h - \beta) (\tilde{\beta}^h - \beta)' z_i' \right\} = \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \text{tr} \left\{ (\tilde{\beta}^h - \beta)' z_i' z_i (\tilde{\beta}^h - \beta) \right\} \\ & = \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \sum_{t=1}^T |z_{it}' (\tilde{\beta}^h - \beta)|^2 \leq \frac{1}{NT} \sum_{i \in \mathcal{I}_h} \sum_{t=1}^T \|z_{it}\|^2 \|\tilde{\beta}^h - \beta\|^2 = o_P(1) \end{aligned}$$

by Step 3 and Theorem 3.1. Combining presented bounds, we obtain the asserted claim of this step.

**Step 5.** Here we show that

$$\|\hat{B}^0 - F \Lambda F'\| \vee \|\hat{B}^1 - F \Lambda F'\| = o_P(1),$$

To do so, fix  $h = 0, 1$ . Then, denoting

$$\check{B} = \frac{1}{NT} \sum_{i=1}^N (y_i - x_i \beta) (y_i - x_i \beta)',$$



we have

$$\|\bar{B}^h - \check{B}\| = o_P(1), \quad (39)$$

where the matrix  $B^h$  is defined in the previous step. To see why this is so, observe that

$$\bar{B}^h - \check{B} = \frac{2}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) (y_i - x_i\beta)(y_i - x_i\beta)'$$

Thus, recalling that  $y_i = x_i\beta + \alpha_{g_i} + v_i$ , we have

$$\begin{aligned} \|\bar{B}^h - \check{B}\| &\leq \left\| \frac{2}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) \alpha_{g_i} \alpha_{g_i}' \right\| \\ &\quad + \left\| \frac{2}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) v_i v_i' \right\| \\ &\quad + \left\| \frac{4}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) \alpha_{g_i} v_i' \right\|. \end{aligned}$$

Here,  $\|\alpha_{g_i} \alpha_{g_i}'\| = \|\alpha_{g_i}\|^2 \leq CT$  for all  $i = 1, \dots, N$ , and so applying the expectation version of Bernstein's matrix inequality (Exercise 5.4.11 in [32]) conditionally on  $(x_1, y_1), \dots, (x_N, y_N)$ ,

$$\begin{aligned} \mathbb{E} \left[ \left\| \frac{1}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) \alpha_{g_i} \alpha_{g_i}' \right\| \right] \\ \leq C \left( \frac{\sqrt{NT^2 \log T}}{NT} + \frac{T \log T}{NT} \right) = o(1), \end{aligned}$$

where the last bound follows from Assumption 3.9. Hence,

$$\left\| \frac{1}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) \alpha_{g_i} \alpha_{g_i}' \right\| = o_P(1)$$

by Markov's inequality. In addition,

$$\left\| \frac{1}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) v_i v_i' \right\| \leq \left\| \frac{1}{NT} \sum_{i=1}^N v_i v_i' \right\| = o_P(1)$$

by (32) in the proof of Theorem 3.1. Moreover, by Lemma 7.2,

$$\left\| \frac{1}{NT} \sum_{i=1}^N \left( 1\{h_i = 1 - h\} - \mathbb{E}[1\{h_i = 1 - h\}] \right) \alpha_{g_i} v_i' \right\|$$

$$\leq \sqrt{\left\| \frac{1}{NT} \sum_{i=1}^N \alpha_{g_i} \alpha'_{g_i} \right\|} \sqrt{\left\| \frac{1}{NT} \sum_{i=1}^N v_i v'_i \right\|} = o_P(1).$$

Combining these bounds, we obtain (39).

Now, given that  $\check{B} = (NT)^{-1} \sum_{i=1}^N (\alpha_{g_i} + v_i)(\alpha_{g_i} + v_i)'$ ,  $\alpha_{g_i} = \sqrt{T} F p_{g_i}$ , and  $N^{-1} \sum_{i=1}^N p_{g_i} p'_{g_i} = \Lambda$ , we have

$$\|\check{B} - F \Lambda F'\| \leq \left\| \frac{2}{NT} \sum_{i=1}^N \alpha_{g_i} v'_i \right\| + \left\| \frac{1}{NT} \sum_{i=1}^N v_i v'_i \right\| = o_P(1)$$

by the arguments above. Combining this bound with (39) and Step 4 and using the triangle inequality gives the asserted claim of this step.

**Step 6.** Here we show that there exist orthogonal  $G \times G$  matrices  $O_0$  and  $O_1$  such that for all  $\gamma = 1, \dots, G$  and  $h = 0, 1$ ,

$$\|(\hat{F}_h O_h - F)' F p_\gamma\| = o_P(1).$$

To do so, fix  $h = 0, 1$  and note that by Step 5, there exists a sequence  $\{\psi_n\}_{n \geq 1}$  of positive numbers such that  $\psi_n \rightarrow 0$  as  $n \rightarrow \infty$  and

$$P(\|\hat{B}^h - F \Lambda F'\| > \psi_n) \leq \psi_n. \quad (40)$$

Also, recall that  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_G)$ , where  $\lambda_1 \geq \dots \geq \lambda_G \geq 0$ . In addition, set  $\lambda_0 = +\infty$  and  $\lambda_{G+1} = 0$  and let  $\bar{G}$  be the largest integer  $\gamma = 0, \dots, G$  such that  $\lambda_\gamma - \lambda_{\gamma+1} > \sqrt{\psi_n}$ . Then  $\lambda_{\bar{G}+1} \leq G \sqrt{\psi_n}$ . In addition, by the Davis-Kahan theorem (see Theorem 2 in [35]), there exists an orthogonal  $G \times G$  matrix  $O_h$  such that, denoting  $\tilde{F} = \hat{F}_h O_h$  and letting  $\tilde{F}_\gamma$  and  $F_\gamma$  denote the  $\gamma$ -th columns of  $\tilde{F}$  and  $F$ , respectively, we have

$$\max_{1 \leq \gamma \leq \bar{G}} \|\tilde{F}_\gamma - F_\gamma\| \leq 4\sqrt{G} \|\hat{B}^h - F \Lambda F'\| / \sqrt{\psi_n},$$

and so by (40),

$$\max_{1 \leq \gamma \leq \bar{G}} \|\tilde{F}_\gamma - F_\gamma\| \leq 4\sqrt{G} \sqrt{\psi_n}$$

with probability at least  $1 - \psi_n$ . Therefore,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \|(\tilde{F} - F)' F p_{g_i}\|^2 &= \text{tr} \left( \frac{1}{N} \sum_{i=1}^N (\tilde{F} - F)' F p_{g_i} p'_{g_i} F' (\tilde{F} - F) \right) \\ &= \text{tr} \left( (\tilde{F} - F)' F \Lambda F' (\tilde{F} - F) \right) = \text{tr} \left( (\tilde{F} - F)' \sum_{\gamma_1=1}^G \lambda_{\gamma_1} f_{\gamma_1} f'_{\gamma_1} (\tilde{F} - F) \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{\gamma_1, \gamma_2=1}^G \lambda_{\gamma_1} \{F'_{\gamma_1} (\tilde{F}_{\gamma_2} - F_{\gamma_2})\}^2 \leq \sum_{\gamma_1=1}^{\bar{G}} \sum_{\gamma_2=1}^G \lambda_{\gamma_1} \{F'_{\gamma_1} (\tilde{F}_{\gamma_2} - F_{\gamma_2})\}^2 + G^3 \sqrt{\psi_n} \\
&\leq \sum_{\gamma_1=1}^{\bar{G}} \lambda_{\gamma_1} \{F'_{\gamma_1} (\tilde{F}_{\gamma_1} - F_{\gamma_1})\}^2 + \sum_{\gamma_1=1}^{\bar{G}} \sum_{\gamma_2 \neq \gamma_1} \lambda_{\gamma_1} \{F'_{\gamma_1} \tilde{F}_{\gamma_2}\}^2 + G^3 \sqrt{\psi_n} = o_P(1),
\end{aligned}$$

where the middle term is bounded by Lemma 7.3 and all  $\lambda_\gamma$  are bounded by

$$\max_{1 \leq \gamma \leq G} \lambda_\gamma = \|\Lambda\| = \|F\Lambda F'\| = \left\| \frac{1}{NT} \sum_{i=1}^N \alpha_{g_i} \alpha'_{g_i} \right\| \leq C,$$

where the last inequality follows from Assumption 3.6. Combining this bound with Assumption 3.8, we conclude that for all  $\gamma = 1, \dots, G$ ,

$$\begin{aligned}
\|(\hat{F}_h O_h - F)' F p_\gamma\|^2 &\leq \frac{C}{N} \sum_{i=1}^N \|(\hat{F}_h O_h - F)' F p_{g_i}\|^2 \\
&= \frac{C}{N} \sum_{i=1}^N \|(\tilde{F} - F)' F p_{g_i}\|^2 = o_P(1),
\end{aligned}$$

which gives the asserted claim of this step. By Lemma 7.3.

**Step 7.** Here we show that for all  $\gamma = 1, \dots, G$  and  $h = 0, 1$ ,

$$\|(\hat{F}_h O_h - F) p_\gamma\| = o_P(1).$$

To do so, fix  $h = 0, 1$  and note that

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \|(\hat{F}_h O_h - F) p_{g_i}\|^2 &= \text{tr} \left( \frac{1}{N} \sum_{i=1}^N (\hat{F}_h O_h - F) p_{g_i} p'_{g_i} (\hat{F}_h O_h - F)' \right) \\
&= \text{tr} \left( (\hat{F}_h O_h - F) \Lambda (\hat{F}_h O_h - F)' \right) \\
&= \text{tr} \left( \Lambda (\hat{F}_h O_h - F)' (\hat{F}_h O_h - F) \right) \\
&= \sum_{\gamma=1}^G \lambda_\gamma \|(\hat{F}_h O_h - F)_\gamma\|^2 = o_P(1)
\end{aligned}$$

by the same argument as that used in Step 6. Combining this bound with Assumption 3.8, we conclude that for all  $\gamma = 1, \dots, G$ ,

$$\|(\hat{F}_h O_h - F) p_\gamma\|^2 \leq \frac{C}{N} \sum_{i=1}^N \|(\hat{F}_h O_h - F) p_{g_i}\|^2 = o_P(1),$$

which gives the asserted claim of this step.

**Step 8.** Here we complete the proof. To do so, we have

$$\begin{aligned}
\|\hat{A}_i - \alpha_{g_i}\| &= \|\hat{F}_{h_i} \hat{F}'_{h_i} (y_i - x_i \tilde{\beta}^{h_i}) - \sqrt{T} F p_{g_i}\| \\
&= \sqrt{T} \|\hat{F}_{h_i} \hat{F}'_{h_i} F p_{g_i} - F p_{g_i}\| + o_P(\sqrt{T}) \\
&= \sqrt{T} \|\hat{F}_{h_i} O_{h_i} (\hat{F}'_{h_i} O_{h_i})' F p_{g_i} - F p_{g_i}\| + o_P(\sqrt{T}) \\
&= \sqrt{T} \|\hat{F}_{h_i} O_{h_i} p_{g_i} - F p_{g_i}\| + o_P(\sqrt{T}) = o_P(\sqrt{T})
\end{aligned}$$

uniformly over  $i = 1, \dots, N$ , where the second line follows from Steps 1 and 2, the third from  $O_{h_i}$  being orthogonal, and the fourth from Steps 6 and 7. This gives the asserted claim of this step and completes the proof of the theorem.  $\blacksquare$

*Proof of Theorem 3.3.* Throughout the proof, for any  $a \in \mathbb{R}^T$  and  $R > 0$ , we use  $B(a, R)$  to denote the ball in  $\mathbb{R}^T$  with center  $a$  and radius  $R$ , i.e.  $B(a, R) = \{u \in \mathbb{R}^T : \|u - a\| \leq R\}$ . Also, we use  $c$  and  $C$  to denote strictly positive constants that can change from place to place but can be chosen to depend on  $C_1, C_2, C_3, C_4, C_5, c_1, c_2$ , and  $c_3$  only.

We proceed in five steps. The second and third steps follow closely the arguments in the proof of Theorem 1 in [9] but we provide all the details for reader's convenience.

**Step 1.** Here we show that

$$\max_{1 \leq i \leq N} \mathbb{E} \left[ \sum_{t=1}^T v_{it}^2 \right] \leq CT \quad \text{and} \quad \max_{1 \leq i \leq N} \max_{j \neq i} \mathbb{E} \left[ \left| \sum_{t=1}^T v_{it} v_{jt} \right| \right] \leq C\sqrt{T}. \quad (41)$$

To prove the first inequality, note that for all  $i = 1, \dots, N$ , we have  $\mathbb{E}[v_{it}^2] \leq C \|v_{it}\|_{\psi_2}^2 \leq C$  by (2.15) in [32] and Assumption 3.1. To prove the second inequality note that for all  $i, j = 1, \dots, N$  with  $i \neq j$ , we have

$$\mathbb{P} \left( \left| \sum_{t=1}^T v_{it} v_{jt} \right| > \epsilon \mid v_{j1}, \dots, v_{jT} \right) \leq 2 \exp \left( - \frac{c\epsilon^2}{\sum_{t=1}^T v_{jt}^2} \right)$$

for all  $\epsilon > 0$  by (2.14) in [32] and Assumption 3.1. Therefore, given that the function  $r \mapsto \exp(-c/r)$  is concave on  $\mathbb{R}_+$  for any  $c > 0$ , we have

$$\mathbb{P} \left( \left| \sum_{t=1}^T v_{it} v_{jt} \right| > \epsilon \right) \leq 2 \exp \left( - \frac{c\epsilon^2}{\sum_{t=1}^T \mathbb{E}[v_{jt}^2]} \right) \leq 2 \exp \left( - \frac{c\epsilon^2}{T} \right)$$

by Jensen's inequality and the first inequality in (41). Hence,

$$\mathbb{E} \left[ \left| \sum_{t=1}^T v_{it} v_{jt} \right| \right] = \int_0^\infty \mathbb{P} \left( \left| \sum_{t=1}^T v_{it} v_{jt} \right| > \epsilon \right) d\epsilon \leq C\sqrt{T},$$

which gives the second inequality in (41).

**Step 2.** For all  $b \in \mathcal{B}$ ,  $a = \{a_{\gamma t}\}_{\gamma,t=1}^{G,T} \in \mathcal{A}_{G,T}$ , and  $\nu = \{\nu_i\}_{i=1}^N \in \{1, \dots, G\}^N$ , denote

$$Q(b, a, \nu) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x'_{it}b - a_{\nu_i t})^2$$

and

$$\bar{Q}(b, a, \nu) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (x'_{it}(\beta - b) + \alpha_{g_{it}} - a_{\nu_i t})^2 + \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it}^2.$$

In this step, we show that

$$Q(b, a, \nu) - \bar{Q}(b, a, \nu) = o_P(1)$$

uniformly over  $b \in \mathcal{B}$ ,  $a \in \mathcal{A}_{G,T}$ , and  $\nu \in \{1, \dots, G\}^N$ . To do so, note that

$$Q(b, a, \nu) - \bar{Q}(b, a, \nu) = -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} (x'_{it}(b - \beta) + a_{\nu_i t} - \alpha_{g_{it}}).$$

Here, we have

$$\begin{aligned} \mathbb{E} \left[ \left( \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} \alpha_{g_{it}} \right)^2 \right] &= \frac{1}{(NT)^2} \sum_{i=1}^N \mathbb{E} \left[ \left( \sum_{t=1}^T v_{it} \alpha_{g_{it}} \right)^2 \right] \\ &\leq \frac{C}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \alpha_{g_{it}}^2 \leq \frac{C}{NT} \rightarrow 0, \end{aligned}$$

where the first inequality follows from Assumption 3.1 and (2.15) in [32] and the second from Assumption 3.6. Thus, by Markov's inequality,

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} \alpha_{g_{it}} = o_P(1).$$

Further, denoting  $\bar{x}_{it} = \sum_{m=1}^M \rho_{im} \alpha_{g_{it}}^m$ , so that  $x_{it} = \bar{x}_{it} + z_{it}$ , we have

$$\left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} \bar{x}'_{it} (b - \beta) \right\| \leq C \left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} \bar{x}_{it} \right\| = o_P(1)$$

uniformly over  $b \in \mathcal{B}$  by the same argument as that we have just used and Assumptions 3.5, 3.6, and 3.11. Moreover,

$$\left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} z'_{it} (b - \beta) \right\| \leq C \left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} z_{it} \right\| = o_P(1)$$

uniformly over  $b \in \mathcal{B}$  by Assumptions 3.3 and 3.11. In addition,

$$\begin{aligned} \left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} a_{\nu_{it}} \right| &= \left| \frac{1}{NT} \sum_{\gamma=1}^G \sum_{t=1}^T a_{\gamma t} \sum_{i: \nu_i=\gamma} v_{it} \right| \\ &\leq \frac{1}{NT} \sum_{\gamma=1}^G \sqrt{\sum_{t=1}^T a_{\gamma t}^2} \sqrt{\sum_{t=1}^T \left( \sum_{i: \nu_i=\gamma} v_{it} \right)^2} \leq \frac{C}{N\sqrt{T}} \sum_{\gamma=1}^G \sqrt{\sum_{t=1}^T \left( \sum_{i: \nu_i=\gamma} v_{it} \right)^2} \\ &\leq \frac{C}{N\sqrt{T}} \sum_{\gamma=1}^G \sqrt{\sum_{i,j=1}^N \left| \sum_{t=1}^T v_{it} v_{jt} \right|}, \end{aligned}$$

where the second line follows from the Cauchy-Schwarz inequality and Assumption 3.11. In turn,

$$\mathbb{E} \left[ \sum_{i,j=1}^N \left| \sum_{t=1}^T v_{it} v_{jt} \right| \right] \leq C(N^2\sqrt{T} + NT)$$

by Step 1. Hence, by Markov's inequality,

$$\left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} a_{\nu_{it}} \right| = o_P(1)$$

uniformly over  $a \in \mathcal{A}_{G,T}$  and  $\nu \in \{1, \dots, G\}^N$  since  $T \rightarrow \infty$  as  $N \rightarrow \infty$ . Combining presented bounds gives the asserted claim of this step.

**Step 3.** Here we show that

$$\bar{Q}(b, a, \nu) - \bar{Q}(\beta, \alpha, g) \geq c\|b - \beta\|^2$$

with probability  $1 - o(1)$  for some constant  $c > 0$  uniformly over  $b \in \mathcal{B}$ ,  $a \in \mathcal{A}_{G,T}$ , and  $\nu \in \{1, \dots, G\}^N$ , where  $\alpha = \{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$ . To do so, note that

$$\begin{aligned} \bar{Q}(b, a, \nu) - \bar{Q}(\beta, \alpha, g) &= \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left( x'_{it}(b - \beta) + a_{\nu_{it}} - \alpha_{g_{it}} \right)^2 \\ &\geq \min_{\tilde{a} \in \mathcal{A}_{G,T}} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left( x'_{it}(b - \beta) + \tilde{a}_{\nu_{it}} - \alpha_{g_{it}} \right)^2 \\ &\geq \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left( (x_{it} - \bar{x}_{\nu, \nu_i, g_{i,t}})'(b - \beta) \right)^2 \geq c\|b - \beta\|^2 \end{aligned}$$

with probability  $1 - o(1)$  for some constant  $c > 0$  uniformly over  $b \in \mathcal{B}$ ,  $a \in \mathcal{A}_{G,T}$ , and  $\nu \in \{1, \dots, G\}^N$  by Assumption 3.10. The asserted claim of this step follows.

**Step 4.** Denote

$$\mathcal{G}_\gamma = \left\{ i = 1, \dots, N : \hat{g}_i = \gamma \right\}, \quad \text{for all } \gamma = 1, \dots, G$$

and

$$\check{\alpha}_{\gamma t} = \frac{1}{|\mathcal{G}_\gamma|} \sum_{i \in \mathcal{G}_\gamma} \alpha_{g_i t}, \quad \text{for all } \gamma = 1, \dots, G, \quad t = 1, \dots, T.$$

In this step, we show that

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\check{\alpha}_{\hat{g}_i t} - \alpha_{g_i t})^2 = o_P(1). \quad (42)$$

To do so, note that by (33) in the proof of Theorem 3.2, there exist sequences  $\{\lambda_N\}_{N \geq 1}$  and  $\{\delta_N\}_{N \geq 1}$  of positive numbers such that  $\lambda_N/\sqrt{T} \rightarrow 0$  and  $\delta_N \rightarrow 0$  as  $N \rightarrow \infty$  and the event

$$\max_{1 \leq i \leq N} |\hat{A}_i - \alpha_{g_i}| \leq \lambda_N \quad (43)$$

holds with probability at least  $1 - \delta_N$  (note that the proof of (33) did not use Assumption 3.7, which is not imposed here). We claim that on the event (43),  $\hat{\lambda} \leq C_G \lambda_N$ , where  $C_G > 0$  is a constant depending only on  $G$ . To see why this is so, suppose that (43) is satisfied and consider the subset  $A_0$  of the set  $\{\alpha_1, \dots, \alpha_G\}$  and a constant  $R_0 > 0$  defined by the following algorithm:

*Step 1:* set  $\gamma = 1$ ,  $A_1 = \{\alpha_1, \dots, \alpha_G\}$ , and  $R_1 = \lambda_N$ ;

*Step 2:* if

$$\left\{ \bigcup_{a \in A_\gamma} (B(a, 4R_\gamma + 4\lambda_N) \setminus B(a, R_\gamma)) \right\} \cap A_\gamma = \emptyset,$$

then set  $A_0 = A_\gamma$  and  $R_0 = R_\gamma$  and stop;

*Step 3:* replace  $\gamma$  by  $\gamma + 1$  and set  $R_\gamma = 5R_{\gamma-1} + 4\lambda_N$ ;

*Step 4:* let  $A_\gamma$  be the smallest subset of  $A_{\gamma-1}$  such that

$$\{\alpha_1, \dots, \alpha_G\} \subset \bigcup_{a \in A_\gamma} B(a, R_\gamma);$$

*Step 5:* go to step 2.

Intuitively, this algorithm generate balls in  $\mathbb{R}^T$  of radius  $R_0$  that cover  $\{\alpha_1, \dots, \alpha_G\}$  and are sufficiently well separated. Formally, we have the following. On Step 4 of this algorithm, we have  $|A_\gamma| \leq |A_{\gamma-1}| - 1$ . Indeed, if Step 4 is performed for given  $\gamma$ , then it follows from Step 2 that there exist  $a_1, a_2 \in A_{\gamma-1}$  such that  $a_2 \in B(a_1, 4R_{\gamma-1} + 4\lambda_N)$ , and so  $B(a_2, R_{\gamma-1}) \subset B(a_1, R_\gamma)$ , letting us drop  $a_2$  from  $A_{\gamma-1}$  while constructing  $A_\gamma$  and yielding  $|A_\gamma| \leq |A_{\gamma-1}| - 1$ . In turn, the latter implies that Step 3 will be performed at most  $G - 1$  times. Hence,  $R_0$  satisfies  $R_0 \leq \bar{C}_G \lambda_N$ , where  $\bar{C}_G > 0$  is a constant depending only on  $G$ . In addition, because of the condition on Step 2, we

have  $m(2R_0 + 2\lambda_N) = |A_0| \leq G$ . But the latter implies that  $\hat{\lambda} \leq 2R_0 + 2\lambda_N \leq C_G \lambda_N$ , as claimed.

Next, we claim that (43) implies even more: it implies that there exists a constant  $\tilde{C}_G > 0$  depending only on  $G$  such that for any  $\gamma = 1, \dots, G$  and any  $i, j \in \mathcal{G}_\gamma$ , we have

$$\|\alpha_{g_i} - \alpha_{g_j}\| \leq \tilde{C}_G \lambda_N. \quad (44)$$

To see why this is so, suppose again that (43) is satisfied and consider any  $\gamma = 1, \dots, G$ . Denote by  $\hat{G}_\gamma = |\{g_i : i \in \mathcal{G}_\gamma\}|$  the number of groups whose items appear in  $\mathcal{G}_\gamma$ . If  $\hat{G}_\gamma = 1$ , we trivially get  $\|\alpha_{g_i} - \alpha_{g_j}\| = 0$  for all  $i, j \in \mathcal{G}_\gamma$ , yielding (44). If  $\hat{G}_\gamma = 2$ , let  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_{|\mathcal{G}_\gamma|}$  be the order in which units are added to  $\mathcal{G}_\gamma$  according to the Classification Algorithm of Section 2, with  $i_1$  and  $i_{|\mathcal{G}_\gamma|}$  being the first and the last units added to  $\mathcal{G}_\gamma$ . Let  $k$  be the smallest number in the set  $\{2, \dots, |\mathcal{G}_\gamma|\}$  such that  $g_{i_k} \neq g_{i_{k-1}}$ . The unit  $i_k$  then satisfies the following inequality for  $\mathcal{A} = \{i_1, \dots, i_{k-1}\}$ :

$$\left\| \hat{A}_{i_k} - \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l \right\| \leq \hat{\lambda}.$$

But then, given that  $g_{i_1} = \dots = g_{i_{k-1}}$ , we have for any  $r = 1, \dots, k-1$  that

$$\begin{aligned} \|\hat{A}_{i_k} - \hat{A}_{i_r}\| &\leq \left\| \hat{A}_{i_k} - \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l \right\| + \left\| \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \hat{A}_l - \hat{A}_{i_r} \right\| \\ &\leq \hat{\lambda} + \frac{1}{|\mathcal{A}|} \sum_{l, j \in \mathcal{A}} \|\hat{A}_l - \hat{A}_j\| \leq \hat{\lambda} + 2\lambda_N \leq (C_G + 2)\lambda_N, \end{aligned}$$

where we used the triangle inequality, (43), and the fact that (43) implies  $\hat{\lambda} \leq C_G \lambda_N$ , as proven above. Hence,  $\|\alpha_{g_{i_k}} - \alpha_{g_{i_r}}\| \leq (C_G + 4)\lambda_N$ , which in turn yields (44). When  $\hat{G}_\gamma = 3$  or larger, we apply the same argument by looking at the first item from the last group to be added to  $\mathcal{G}_\gamma$  and note that whenever we increase  $\hat{G}_\gamma$  by one, the maximal value of  $\|\alpha_{g_i} - \alpha_{g_j}\|$  increases at most by  $(C_G + 4)\lambda_N$ . Since the largest possible value of  $\hat{G}_\gamma$  is  $G$ , this implies that (44) holds with  $\tilde{C}_G = (G - 1)(C_G + 4)\lambda_N$ , as claimed.

We are now ready to finish this step. On (43), by the triangle inequality and (44), we have

$$\begin{aligned} \left( \frac{1}{T} \sum_{t=1}^T (\tilde{\alpha}_{\hat{g}_i t} - \alpha_{g_i t})^2 \right)^{1/2} &\leq \frac{1}{|\mathcal{G}_{\hat{g}_i}|} \sum_{j \in \mathcal{G}_{\hat{g}_i}} \left( \frac{1}{T} \sum_{t=1}^T (\alpha_{g_j t} - \alpha_{g_i t})^2 \right)^{1/2} \\ &= \frac{1}{\sqrt{T} |\mathcal{G}_{\hat{g}_i}|} \sum_{j \in \mathcal{G}_{\hat{g}_i}} \|\alpha_{g_j} - \alpha_{g_i}\| \leq \frac{1}{\sqrt{T} |\mathcal{G}_{\hat{g}_i}|} \sum_{j \in \mathcal{G}_{\hat{g}_i}} \tilde{C}_G \lambda_N = o(1) \end{aligned}$$



uniformly over  $i = 1, \dots, N$ . This gives the asserted claim of this step because (43) holds with probability approaching one.

**Step 5.** Here we finish the proof. We have

$$\begin{aligned}\bar{Q}(\hat{\beta}, \hat{\alpha}, \hat{g}) &= Q(\hat{\beta}, \hat{\alpha}, \hat{g}) + o_P(1) \\ &\leq Q(\beta, \check{\alpha}, \hat{g}) + o_P(1) = \bar{Q}(\beta, \check{\alpha}, \hat{g}) + o_P(1),\end{aligned}$$

where the equalities follow from Step 2 and the inequality from (12). Thus, for some constant  $c > 0$ ,

$$\begin{aligned}c\|\hat{\beta} - \beta\|^2 &\leq \bar{Q}(\hat{\beta}, \hat{\alpha}, \hat{g}) - \bar{Q}(\beta, \alpha, g) \\ &= \bar{Q}(\hat{\beta}, \hat{\alpha}, \hat{g}) - \bar{Q}(\beta, \check{\alpha}, \hat{g}) + \bar{Q}(\beta, \check{\alpha}, \hat{g}) - \bar{Q}(\beta, \alpha, g) \leq o_P(1)\end{aligned}$$

by Steps 3 and 4. The asserted claim of the theorem follows.  $\blacksquare$

*Proof of Theorem 3.4.* By Theorem 3.2,  $\hat{\beta} = \hat{\beta}^0$  with probability  $1 - o(1)$  for  $\hat{\beta}^0$  appearing in (13). In turn, by the Frisch-Waugh-Lovell theorem,

$$\hat{\beta}^0 = \left( \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it} \check{x}'_{it} \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it} y_{it} \right),$$

and so

$$\sqrt{NT}(\hat{\beta}^0 - \beta) = \left( \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it} \check{x}'_{it} \right)^{-1} \left( \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it} v_{it} \right) \rightarrow_D N(\check{\Sigma}^{-1} \Omega \check{\Sigma}^{-1})$$

by Slutsky's lemma and Assumption 3.12. The asserted claim follows.  $\blacksquare$

*Proof of Theorem 4.1.* The asserted claim follows from the same arguments as those in the proofs of Theorems 3.1–3.4 as long as we can show that there exists a constant  $C > 0$  such that  $\|\hat{\rho}_{im}\| \vee |\hat{\alpha}_{\gamma t}^m| \leq C$  for all  $i = 1, \dots, N$ ,  $\gamma = 1, \dots, G$ ,  $t = 1, \dots, T$ , and  $m = 1, \dots, 2M$ , which would correspond to Assumptions 3.5 and 3.6 in the context of dynamic model. To do so, observe that

$$\begin{aligned}\max_{1 \leq m \leq 2M} \|\hat{\rho}_{im}\| &\leq \max_{1 \leq m \leq M} (\|\rho_{im}\| + |\rho_{im}^y|) \\ &\leq \max_{1 \leq m \leq M} (\|\rho_{im}\| + 1 + |\rho_{im}' \beta|) \leq 1 + C_3(1 + \|\beta\|)\end{aligned}$$

for all  $i = 1, \dots, N$  by Assumption 4.5. Also,

$$\max_{1 \leq m \leq 2M} |\hat{\alpha}_{\gamma t}^m| \leq \max_{1 \leq m \leq M} \left( |\alpha_{\gamma t}^m| + \left| \sum_{r=0}^{t-2} \theta^r \alpha_{\gamma t-r-1}^m \right| \right) \leq C_4 \left( 1 + \frac{1}{1-\theta} \right)$$

by Assumption 4.6. Therefore, the asserted claim follows if we set

$$C = 1 + C_3(1 + \|\beta\|) + C_4 \left(1 + \frac{1}{1 - \theta}\right).$$

This completes the proof of the theorem.  $\blacksquare$

*Proof of Theorem 4.2.* We first prove (19). By construction of  $\hat{\Sigma}$  and  $\hat{S}$ , it suffices to show that

$$\lambda_1^b + \cdots + \lambda_{2GM+2}^b = b' \Sigma b + S' b + L + O_P \left( \frac{1}{T \wedge N} + \sqrt{\frac{\log d}{NT}} \right)$$

uniformly over  $b \in \bar{\mathcal{B}} = \{0_d\} \cup \{e_k : k = 1, \dots, d\} \cup \{e_k + e_l : k, l = 1, \dots, d\}$ . To do so, we proceed by appropriately modifying the proof of Theorem 3.1. Throughout, we use the same notations as in the proof of Theorem 3.1. We have

$$\begin{aligned} \text{tr}(R) &= -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta - b)' z_{it} z_{it} (\beta - b) + v_{it}^2 + 2v_{it} z'_{it} (\beta - b) \right\} \\ &= -\frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta - b)' z_{it} z_{it} (\beta - b) + v_{it}^2 \right\} + O_P \left( \sqrt{\frac{\log d}{NT}} \right) \end{aligned}$$

uniformly over  $b \in \bar{\mathcal{B}}$  since

$$\left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} z'_{it} (\beta - b) \right| \leq \left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T v_{it} z_{it} \right\|_{\infty} \|\beta - b\|_1 \leq (C_{\beta} + 2) O_P \left( \sqrt{\frac{\log d}{NT}} \right)$$

uniformly over  $b \in \bar{\mathcal{B}}$ . Thus,

$$\begin{aligned} \text{tr}(A_0) &= \frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (\beta - b)' z_{it} z_{it} (\beta - b) + v_{it}^2 \right\} + O_P \left( \sqrt{\frac{\log d}{NT}} \right) \\ &= b' \Sigma b + S' b + L + O_P \left( \sqrt{\frac{\log d}{NT}} \right) \end{aligned}$$

uniformly over  $b \in \bar{\mathcal{B}}$  since

$$\begin{aligned} &\left| \frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^T (\beta - b)' z_{it} z'_{it} (\beta - b) - (\beta - b)' \Sigma (\beta - b) \right| \\ &\leq \left\| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T z_{it} z'_{it} - \Sigma \right\|_{\infty} \|\beta - b\|_1^2 \leq (C_{\beta} + 2)^2 O_P \left( \sqrt{\frac{\log d}{NT}} \right) \end{aligned}$$

uniformly over  $b \in \bar{\mathcal{B}}$ . Also,

$$\left\| \frac{1}{NT} \sum_{t=1}^T V_t V_t' \right\| = O_P \left( \frac{1}{T \wedge N} \right)$$

as in the proof of Theorem 3.1. Further, to prove that

$$\left\| \frac{1}{NT} \sum_{t=1}^T Z_{bt} Z_{bt}' \right\| = O_P \left( \frac{1}{T \wedge N} \right) \quad (45)$$

uniformly over  $b \in \bar{\mathcal{B}}$ , where we denoted  $Z_{bt} = (z'_{1t}(\beta - b), \dots, z'_{Nt}(\beta - b))'$ , we have  $\|u' z_i(\beta - b)\|_{\psi_2} \leq \|u' z_i \beta\|_{\psi_2} + \|u' z_i b\|_{\psi_2} \leq C_z + 2C_2$  uniformly over  $u = (u_1, \dots, u_T)' \in \mathcal{S}^T$  and  $b \in \bar{\mathcal{B}}$  by our assumptions, where we denoted  $z_i = (z_{i1}, \dots, z_{iT})'$  for all  $i = 1, \dots, N$ . Thus, denoting  $\mathcal{Z}_b = (z_1(\beta - b), \dots, z_N(\beta - b))'$ , we have

$$\mathbb{P} \left( \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} - \mathbb{E} \left[ \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} \right] \right) u \right| \geq \epsilon \right) \leq 2 \times 9^T \times \exp(-c(\epsilon \wedge \epsilon^2)N)$$

for all  $b \in \bar{\mathcal{B}}$ , where  $\mathcal{N}$  is the same as in the proof of Theorem 3.1 and  $c$  is a constant depending only on  $C_z$  and  $C_2$ . Hence, by the union bound,

$$\mathbb{P} \left( \max_{b \in \bar{\mathcal{B}}} \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} - \mathbb{E} \left[ \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} \right] \right) u \right| \geq \epsilon \right) \leq (1+d+d^2) \times 2 \times 9^T \times \exp(-c(\epsilon \wedge \epsilon^2)N).$$

Setting here  $\epsilon = c^{-1}(\log 9)(T/N) + 1$ , we obtain

$$\begin{aligned} \mathbb{P} \left( \max_{b \in \bar{\mathcal{B}}} \max_{u \in \mathcal{N}} \left| u' \left( \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} - \mathbb{E} \left[ \frac{\mathcal{Z}_b' \mathcal{Z}_b}{N} \right] \right) u \right| \geq c^{-1}(\log 9)(T/N) + 1 \right) \\ \leq 2(1+d+d^2) \exp(-cN) \rightarrow 0 \end{aligned}$$

because  $\log d = o(N)$ . We thus obtain (45) by the same arguments as those in the proof of Theorem 3.1. Repeating the remaining arguments of the proof of Theorem 3.1, we obtain (19).

Next, we prove that (20) and (21) imply (22). To do so, we assume for the rest of the proof that both (20) and (21) are satisfied. Then, by the definition of  $\hat{\beta}_\lambda$ ,

$$\hat{\beta}'_\lambda \hat{\Sigma} \hat{\beta}_\lambda + \hat{S}' \hat{\beta}_\lambda + \lambda \|\hat{\beta}_\lambda\|_1 \leq \beta' \hat{\Sigma} \beta + \hat{S}' \beta + \lambda \|\beta\|_1.$$

Also,

$$(\hat{\beta}_\lambda - \beta)' \hat{\Sigma} (\hat{\beta}_\lambda - \beta) = \hat{\beta}'_\lambda \hat{\Sigma} \hat{\beta}_\lambda - \beta' \hat{\Sigma} \beta + 2(\beta - \hat{\beta}_\lambda)' \hat{\Sigma} \beta.$$

Taking the sum of these two displays, we obtain

$$\begin{aligned} (\hat{\beta}_\lambda - \beta)' \hat{\Sigma} (\hat{\beta}_\lambda - \beta) &= (\hat{S} + 2\hat{\Sigma} \beta)' (\beta - \hat{\beta}_\lambda) + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_\lambda\|_1 \\ &\leq \|\hat{S} + 2\hat{\Sigma} \beta\|_\infty \|\hat{\beta}_\lambda - \beta\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_\lambda\|_1 \end{aligned}$$

$$\leq \left( \|\hat{S} - S\|_\infty + 2C_\beta \|\hat{\Sigma} - \Sigma\|_\infty \right) \|\hat{\beta}_\lambda - \beta\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_\lambda\|_1,$$

where the third line follows by recalling that  $S + 2\Sigma\beta = 0$  and  $\|\beta\|_1 \leq C_\beta$ . Therefore, by (20), we have

$$(\hat{\beta}_\lambda - \beta)' \hat{\Sigma} (\hat{\beta}_\lambda - \beta) \leq (\lambda/c_\lambda) \|\hat{\beta}_\lambda - \beta\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_\lambda\|_1. \quad (46)$$

Further, denote  $\delta = \hat{\beta}_\lambda - \beta$ ,  $\mathcal{T} = \{k = 1, \dots, d: \beta_k \neq 0\}$ , and  $\mathcal{T}^c = \{1, \dots, d\} \setminus \mathcal{T}$ . Also, let  $\delta_{\mathcal{T}} = (\delta_{\mathcal{T}1}, \dots, \delta_{\mathcal{T}d})'$  be a  $d \times 1$  vector such that  $\delta_{\mathcal{T}k} = \delta_k 1\{k \in \mathcal{T}\}$  and, similarly, let  $\delta_{\mathcal{T}^c} = (\delta_{\mathcal{T}^c1}, \dots, \delta_{\mathcal{T}^cd})'$  be a  $d \times 1$  vector such that  $\delta_{\mathcal{T}^ck} = \delta_k 1\{k \in \mathcal{T}^c\}$ . Then, given that  $(\hat{\beta}_\lambda - \beta)' \hat{\Sigma} (\hat{\beta}_\lambda - \beta) \geq 0$ , it follows from (46) that

$$\begin{aligned} 0 &\leq (1/c_\lambda) \|\hat{\beta}_\lambda - \beta\|_1 + \|\beta\|_1 - \|\hat{\beta}_\lambda\|_1 \\ &\leq (1/c_\lambda) (\|\delta_{\mathcal{T}}\|_1 + \|\delta_{\mathcal{T}^c}\|_1) + \|\delta_{\mathcal{T}}\|_1 - \|\delta_{\mathcal{T}^c}\|_1, \end{aligned}$$

where the second inequality follows from observing that  $\beta_k = \beta_k 1\{k \in \mathcal{T}\}$  for all  $k = 1, \dots, d$ . Therefore,  $(1 - 1/c_\lambda) \|\delta_{\mathcal{T}^c}\|_1 \leq (1 + 1/c_\lambda) \|\delta_{\mathcal{T}}\|_1$ , and so  $\|\delta_{\mathcal{T}^c}\|_1 \leq \bar{c}_\lambda \|\delta_{\mathcal{T}}\|_1$ . Thus,

$$\begin{aligned} \delta' \hat{\Sigma} \delta &= \delta' \Sigma \delta - \delta' (\Sigma - \hat{\Sigma}) \delta \geq c_\Sigma \|\delta\|^2 - \|\delta\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty \\ &\geq c_\Sigma \|\delta\|^2 - (1 + \bar{c}_\lambda)^2 \|\delta_{\mathcal{T}}\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty \\ &\geq c_\Sigma \|\delta\|^2 - s(1 + \bar{c}_\lambda)^2 \|\delta\|^2 \|\hat{\Sigma} - \Sigma\|_\infty \geq c_\Sigma \|\delta\|^2 / 2 \end{aligned}$$

by (21). Substituting this bound into (46), we obtain

$$c_\Sigma \|\delta\|^2 / 2 \leq \lambda(1 + 1/c_\lambda) \|\delta_{\mathcal{T}}\|_1 \leq \sqrt{s} \lambda (1 + 1/c_\lambda) \|\delta\|.$$

Rearranging this bound gives the second inequality in (22). To obtain the first inequality in (22), observe that

$$\|\delta\|_1 = \|\delta_{\mathcal{T}}\|_1 + \|\delta_{\mathcal{T}^c}\|_1 \leq (1 + \bar{c}_\lambda) \|\delta_{\mathcal{T}}\|_1 \leq \sqrt{s} (1 + \bar{c}_\lambda) \|\delta\|.$$

This completes the proof of the theorem.  $\blacksquare$

*Proof of Theorem 4.3.* The proof is closely related to that of Theorem 3.1, with the main difference is that we now have

$$f_{it} = (\kappa_i + \omega_i(\beta - b))' \phi_t, \quad \text{for all } i = 1, \dots, N, \quad t = 1, \dots, T.$$

This difference in turn requires us to change the calculation of the bound on the number of non-zero eigenvalues of the matrix  $A_0$ . Recalling that  $F_t = (f_{1t}, \dots, f_{Nt})'$

for all  $t = 1, \dots, T$ , we now have

$$\sum_{t=1}^T F_t(F_t + Z_t + V_t)' = \sum_{t=1}^T \mathcal{K} \phi_t(F_t + Z_t + V_t)' = \mathcal{K} \sum_{t=1}^T \phi_t(F_t + Z_t + V_t)',$$

where  $\mathcal{K} = (\kappa_1 + \omega_1(\beta - b), \dots, \kappa_N + \omega_N(\beta - b))'$  and  $Z_t$  and  $V_t$  are the same as in the proof of Theorem 3.1. Thus, given that  $\mathcal{K}$  is an  $N \times J$  matrix, it follows that the rank of the matrix  $\sum_{t=1}^T F_t(F_t + Z_t + V_t)'$  is at most  $J$ . Similarly, the rank of the matrix  $\sum_{t=1}^T (Z_t + V_t)F_t'$  is also at most  $J$ , as the column rank coincides with the row rank. We have thus replaced  $GM$  in the proof of Theorem 3.1 by  $J$ . The rest of the proof is the same as that of Theorem 3.1.  $\blacksquare$

## 7. TECHNICAL LEMMAS

**Lemma 7.1.** *Let  $A$  and  $B$  be two symmetric  $N \times N$  matrices, and suppose that the matrix  $B$  has only  $K \leq N$  non-zero eigenvalues  $\lambda_1^B, \dots, \lambda_K^B$ . Further, let  $\lambda_1^A, \dots, \lambda_K^A$  be the  $K$  largest in absolute value eigenvalues of the matrix  $A$ . Then*

$$\left| \sum_{k=1}^K \lambda_k^A - \sum_{k=1}^K \lambda_k^B \right| \leq 3K \|A - B\|.$$

**Remark 7.1.** This lemma does not follow from Weyl's inequality immediately because  $\lambda_1^A, \dots, \lambda_K^A$  are the largest *in absolute value* eigenvalues of  $A$ .  $\blacksquare$

*Proof.* Let  $\lambda_{K+1}^B, \dots, \lambda_N^B$  be the remaining eigenvalues of  $B$ , so that

$$\lambda_{K+1}^B = \dots = \lambda_N^B = 0, \quad (47)$$

and let  $\lambda_{K+1}^A, \dots, \lambda_N^A$  be the remaining eigenvalues of  $A$ , so that

$$\min(|\lambda_1^A|, \dots, |\lambda_K^A|) \geq \max(|\lambda_{K+1}^A|, \dots, |\lambda_N^A|). \quad (48)$$

By Weyl's inequality ([32], Theorem 4.5.3), one can construct a one-to-one function  $f: \{1, \dots, N\} \rightarrow \{1, \dots, N\}$  such that

$$|\lambda_k^A - \lambda_{f(k)}^B| \leq \|A - B\|, \quad \text{for all } k = 1, \dots, N. \quad (49)$$

Using this function, define  $\mathcal{K} = \{j = 1, \dots, K: f^{-1}(j) > K\}$  and note that if this set is non-empty, then there exists  $k = 1, \dots, K$  such that  $f(k) > K$ , and so for any  $j \in \mathcal{K}$ ,

$$|\lambda_j^B| \leq |\lambda_{f^{-1}(j)}^A| + \|A - B\| \leq |\lambda_k^A| + \|A - B\| \leq |\lambda_{f(k)}^B| + 2\|A - B\| = 2\|A - B\|,$$

where the first inequality follows from (49) and the triangle inequality, the second from (48), the third from (49) and the triangle inequality, and the fourth from (47).

Hence,

$$\left| \sum_{k=1}^K \lambda_k^B - \sum_{k=1}^K \lambda_{f(k)}^B \right| = \left| \sum_{j \in \mathcal{K}} \lambda_j^B \right| \leq 2K \|A - B\|.$$

In addition,

$$\left| \sum_{k=1}^K \lambda_k^A - \sum_{k=1}^K \lambda_{f(k)}^B \right| \leq K \|A - B\|$$

by (49). Combining the last two inequalities gives the asserted claim.  $\blacksquare$

**Lemma 7.2.** *Let  $\mu_1, \dots, \mu_N$  and  $\nu_1, \dots, \nu_N$  be two sequences of vectors in  $\mathbb{R}^T$ . Define*

$$A = \sum_{i=1}^N \mu_i \mu_i', \quad B = \sum_{i=1}^N \nu_i \nu_i', \quad C = \sum_{i=1}^N \mu_i \nu_i', \quad \text{and} \quad D = \sum_{i=1}^N (\mu_i + \nu_i)(\mu_i + \nu_i)'$$

*Then  $\|C\| \leq \sqrt{\|A\|} \sqrt{\|B\|}$  and  $\|D\| \leq 2(\|A\| + \|B\|)$ .*

*Proof.* For any  $x \in \mathbb{R}^T$  such that  $\|x\| = 1$ ,

$$\begin{aligned} x' C x &= \sum_{i=1}^N (x' \mu_i)(\nu_i' x) \leq \sqrt{\sum_{i=1}^N (x' \mu_i)^2} \sqrt{\sum_{i=1}^N (\nu_i' x)^2} \\ &= \sqrt{x' A x} \sqrt{x' B x} \leq \sqrt{\|A\|} \sqrt{\|B\|}, \end{aligned}$$

by the Cauchy-Schwarz inequality. Taking the supremum over all  $x \in \mathbb{R}^T$  with  $\|x\| = 1$  of the left- and right-hand sides of this chain of inequalities gives the first asserted claim.

To prove the second asserted claim, note that

$$\begin{aligned} \|D\| &\leq \left\| \sum_{i=1}^N \mu_i \mu_i' \right\| + \left\| \sum_{i=1}^N \mu_i \nu_i' \right\| + \left\| \sum_{i=1}^N \nu_i \mu_i' \right\| + \left\| \sum_{i=1}^N \nu_i \nu_i' \right\| \\ &\leq \|A\| + \sqrt{\|A\|} \sqrt{\|B\|} + \sqrt{\|B\|} \sqrt{\|A\|} + \|B\| \\ &\leq \|A\| + (\|A\| + \|B\|)/2 + (\|B\| + \|A\|)/2 + \|B\| = 2(\|A\| + \|B\|). \end{aligned}$$

This completes the proof of the lemma.  $\blacksquare$

**Lemma 7.3.** *Let  $F_1, F_2$ , and  $F_3$  be vectors in  $\mathbb{R}^T$  and suppose that (i)  $\|F_2 - F_1\| \leq \sqrt{2}$ , (ii)  $F_2' F_3 = 0$ , and (iii)  $\|F_1\| = \|F_2\| = \|F_3\| = 1$ . Then  $|F_1' F_3| \leq 3\|F_2 - F_1\|$ .*

*Proof.* By (iii),

$$F_1' F_3 = 1 - \frac{\|F_3 - F_1\|^2}{2}. \quad (50)$$

By the triangle inequality,

$$\|F_3 - F_1\| \geq \|F_3 - F_2\| - \|F_2 - F_1\|,$$

and so, given that  $\|F_3 - F_2\| = \sqrt{2}$  by (ii), it follows from (i) that

$$\|F_3 - F_1\|^2 \geq \left( \|F_3 - F_2\| - \|F_2 - F_1\| \right)^2 \geq 2 - 2\sqrt{2}\|F_2 - F_1\|. \quad (51)$$

Combining (50) and (51),

$$F_1'F_3 \leq 1 - 1 + \sqrt{2}\|F_2 - F_1\| = \sqrt{2}\|F_2 - F_1\|. \quad (52)$$

Also, again by the triangle inequality,

$$\|F_3 - F_1\| \leq \|F_3 - F_2\| + \|F_2 - F_1\|,$$

and so, by (i),

$$\begin{aligned} \|F_3 - F_1\|^2 &\leq \left( \|F_3 - F_2\| + \|F_2 - F_1\| \right)^2 \\ &\leq 2 + 2\sqrt{2}\|F_2 - F_1\| + \sqrt{2}\|F_2 - F_1\| = 2 + 3\sqrt{2}\|F_2 - F_1\|. \end{aligned} \quad (53)$$

Combining (50) and (53),  $F_1'F_3 \geq -3\|F_2 - F_1\|$ . Combining this inequality with (52) gives the asserted claim.  $\blacksquare$

**Lemma 7.4.** *Let  $\theta \in (-1, 1)$  be a real number and  $T \geq 2$  be an integer. Then for all  $u = (u_1, \dots, u_T)' \in \mathbb{R}^T$  such that  $\sum_{t=1}^T u_t^2 \leq 1$ , we have*

$$\sum_{t=1}^{T-1} \left( \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right)^2 \leq \frac{1}{(1-\theta)^2}.$$

*Proof.* We proceed by induction. When  $T = 2$ , we have

$$\sum_{t=1}^{T-1} \left( \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right)^2 = u_2^2 \leq 1 \leq \frac{1}{(1-\theta)^2},$$

so that the claim holds. Now suppose that the claim holds for all  $T = 2, \dots, k$ . We will prove that the claim holds for  $T = k+1$ . To do so, fix any  $u = (u_1, \dots, u_T)' \in \mathbb{R}^T$  such that  $\sum_{t=1}^T u_t^2 \leq 1$  and observe that

$$\begin{aligned} \sum_{t=1}^{T-1} \left( \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right)^2 &= \sum_{t=1}^{T-2} \left( u_{t+1} + \sum_{r=t+1}^{T-1} u_{r+1} \theta^{r-t} \right)^2 + u_T^2 \\ &= \sum_{t=1}^{T-2} \left( u_{t+1}^2 + 2u_{t+1} \sum_{r=t+1}^{T-1} u_{r+1} \theta^{r-t} + \left( \sum_{r=t+1}^{T-1} u_{r+1} \theta^{r-t} \right)^2 \right) + u_T^2 \\ &\leq 1 + \sum_{t=1}^{T-2} \sum_{r=t+1}^{T-1} (u_{t+1}^2 + u_{r+1}^2) \theta^{r-t} + \theta^2 \sum_{t=1}^{T-2} \left( \sum_{r=t}^{T-2} u_{r+2} \theta^{r-t} \right)^2, \end{aligned}$$

where the third line follows from  $\sum_{t=1}^T u_t^2 \leq 1$  and an elementary inequality  $2ab \leq a^2 + b^2$ . Also, by the induction hypothesis,

$$\sum_{t=1}^{T-2} \left( \sum_{r=t}^{T-2} u_{r+2} \theta^{r-t} \right)^2 \leq \frac{1}{(1-\theta)^2}.$$

In addition,

$$\sum_{t=1}^{T-2} \sum_{r=t+1}^{T-1} (u_{t+1}^2 + u_{r+1}^2) \theta^{r-t} \leq 2 \sum_{l=1}^{T-2} \sum_{t=1}^T \theta^l u_t^2 \leq \frac{2\theta}{1-\theta}.$$

Hence,

$$\sum_{t=1}^{T-1} \left( \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right)^2 \leq 1 + \frac{2\theta}{1-\theta} + \frac{\theta^2}{(1-\theta)^2} = \frac{1}{(1-\theta)^2},$$

which completes the induction argument and thus gives the asserted claim for all  $T \geq 2$ .  $\blacksquare$

#### APPENDIX A: RANDOMIZED ALGORITHM FOR CALCULATING EIGENVALUES OF LARGE MATRICES

To calculate the spectral estimator  $\tilde{\beta}$  in Section 2, we had to calculate  $2GM + 2$  largest in absolute value eigenvalues of the  $N \times N$  matrix  $A^b$ . When  $N$  is large, calculating these eigenvalues exactly may be difficult. Fortunately, there exists a class of randomized algorithms that allow to calculate these eigenvalues approximately with minimal efforts. In this section, we describe one such algorithm. Our discussion here mostly follows [19], where an interested reader can find several other related algorithms.

For brevity of notation, suppose that we have an  $N \times N$  symmetric matrix  $A$  and we would like to calculate its  $k$  largest in absolute value eigenvalues,  $\lambda_{(1)}, \dots, \lambda_{(k)}$ , ordered so that  $|\lambda_{(1)}| \geq \dots \geq |\lambda_{(k)}|$ . Consider the following algorithm:

##### Randomized Algorithm for Calculating Eigenvalues.

- Step 1:* choose an oversampling parameter  $p > 0$ , e.g.  $p = 5$  or  $10$ ;
- Step 2:* set a multiplication parameter  $q = \lceil \log N \rceil$ ;
- Step 3:* draw an  $N \times (k+p)$  random matrix  $\Omega = \{\Omega_{ij}\}_{i,j=1}^{N,k+p} \stackrel{iid}{\sim} N(0, 1)$ ;
- Step 4:* compute the  $N \times (k+p)$  matrix  $Y = A^{q+1}\Omega$ ;
- Step 5:* compute QR decomposition  $Y = QR$  with  $Q$  having orthonormal columns;
- Step 6:* compute the  $(k+p) \times N$  matrix  $B = Q'A$ ;
- Step 7:* compute eigenvectors  $\tilde{s}_1, \dots, \tilde{s}_{k+p}$  of the  $(k+p) \times (k+p)$  matrix  $BB'$ ;
- Step 8:* compute  $N \times 1$  vectors  $s_j = B'\tilde{s}_j$  for  $j = 1, \dots, k+p$ ;
- Step 9:* compute  $\hat{\lambda}_j = \text{sign}(s_j' A s_j) (\|B' B s_j\| / \|s_j\|)^{1/2}$  for  $j = 1, \dots, k+p$ ;



*Step 10:* order values  $\hat{\lambda}_1, \dots, \hat{\lambda}_{k+p}$  into  $\hat{\lambda}_{(1)}, \dots, \hat{\lambda}_{(k+p)}$  so that  $|\hat{\lambda}_{(1)}| \geq \dots \geq |\hat{\lambda}_{(k+p)}|$ ;  
*Step 11:* return  $\hat{\lambda}_{(1)}, \dots, \hat{\lambda}_{(k)}$ .

The result of this algorithm is  $k$  values  $\hat{\lambda}_{(1)}, \dots, \hat{\lambda}_{(k)}$ . These are estimators of  $k$  largest in absolute value eigenvalues  $\lambda_{(1)}, \dots, \lambda_{(k)}$  of the matrix  $A$ . As follows from results in [19], these estimators are consistent as  $N \rightarrow \infty$  under conditions to be discussed below even though they are based on a realization of the random matrix  $\Omega$ . Specifically, Corollary 10.10 in [19] shows that

$$\mathbb{E}[\|A - QQ'A\|] \leq \left(1 + \sqrt{\frac{k}{p-1}} + \frac{e\sqrt{N(k+p)}}{p}\right)^{\frac{1}{q+p+1}} |\lambda_{(k+1)}|,$$

where  $\lambda_{(k+1)}$  is the  $(k+1)$ th largest in absolute value eigenvalue of  $A$ . Therefore, by Markov's inequality,

$$\|A - QQ'A\| = O_P(|\lambda_{(k+1)}|). \quad (54)$$

In turn, by the triangle inequality, the fact that  $Q'Q$  is the identity matrix, and the definition  $B = Q'A$ ,

$$\begin{aligned} \|A'A - B'B\| &= \|A'A - A'QQ'QQ'A\| \\ &\leq \|A\| \|A - QQ'A\| + \|QQ'A\| \|A - QQ'A\| \leq 2\|A\| \|A - QQ'A\| \end{aligned}$$

Thus, given that  $\hat{\lambda}_{(1)}^2, \dots, \hat{\lambda}_{(k)}^2$  are  $k$  largest eigenvalues of the matrix  $B'B$  by construction (see Steps 8,9, and 10 in the algorithm above), it follows from Weyl's inequality that  $\hat{\lambda}_{(1)}^2, \dots, \hat{\lambda}_{(k)}^2$  are consistent estimators of  $\lambda_{(1)}^2, \dots, \lambda_{(k)}^2$  as long as  $\|A\| = O_P(1)$  and  $|\lambda_{(k+1)}| = o_P(1)$  as  $N \rightarrow \infty$ . Hence,  $\hat{\lambda}_{(j)} \rightarrow \lambda_{(j)}$  for all  $j = 1, \dots, k$  under the same conditions by the Davis-Kahane theorem.

The described algorithm can be applied in Section 2 to calculate the spectral estimator  $\tilde{\beta}$  with  $A = A^b$  and  $k = 2GM + 2$ . In this case, the aforementioned conditions  $\|A\| = O_P(1)$  and  $|\lambda_{(k+1)}| = o_P(1)$  are satisfied under Assumptions 3.1–3.4 by the proof of Theorem 3.1.

## APPENDIX B: RELATION BETWEEN ASSUMPTIONS 3.1–3.12 AND ASSUMPTIONS 4.1–4.12

In this section, we explain what conditions one has to impose on top of Assumptions 3.1–3.12 to obtain Assumptions 4.1–4.12. To start with, note that Assumptions 4.1 and 4.5–4.9 coincide with Assumptions 3.1 and 3.5–3.9. Also, Assumption 4.11 follows immediately from Assumption 3.11 if we define  $\mathring{\mathcal{B}} = [-1, 1] \times \mathcal{B}$ , as proposed in the main text. In addition, Assumptions 3.10 and 3.12 depend on  $x_{it}$  but their versions

corresponding to the dynamic model, i.e. Assumptions 4.10 and 4.12, are discussed in [9], where some interpretations as well as low-level conditions are provided. We thus only need to discuss Assumptions 4.2, 4.3, and 4.4.

To this end, suppose that Assumptions 3.1, 3.2, 3.3, and 3.4 are satisfied and, in addition, suppose that the noise variables  $v_{it}$  satisfy (16). Moreover, suppose that there exists a constant  $C_y > 0$  such that  $\|y_{i0}\|_{\psi_2} \leq C_y$  for all  $i = 1, \dots, N$ . Then for all  $u = (u_1, \dots, u_T)' \in \mathcal{S}^T$ , we have

$$\left\| \sum_{t=1}^T u_t \sum_{r=0}^{t-2} \theta^r v_{it-r-1} \right\|_{\psi_2} = \left\| \sum_{t=1}^{T-1} v_{it} \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right\|_{\psi_2} \leq \frac{C_1}{1-\theta}$$

by Assumption 3.1 and Lemma 7.4 and, similarly,

$$\left\| \sum_{t=1}^T u_t \sum_{r=0}^{t-2} \theta^r z'_{it-r-1} \beta \right\|_{\psi_2} = \left\| \sum_{t=1}^{T-1} z'_{it} \beta \sum_{r=t}^{T-1} u_{r+1} \theta^{r-t} \right\|_{\psi_2} \leq \frac{d_z C_2 \|\beta\|_\infty}{1-\theta}$$

by Assumption 3.2 and Lemma 7.4, where  $d_z$  is the dimension of  $z_{it}$ . Also,

$$\left\| \sum_{t=1}^T u_t \theta^{t-1} y_{i0} \right\|_{\psi_2} \leq \frac{\|y_{i0}\|_{\psi_2}}{1-\theta} \leq \frac{C_y}{1-\theta}.$$

Hence, by the triangle inequality,

$$\left\| \sum_{t=1}^T u_t z_{it}^y \right\|_{\psi_2} \leq \frac{C_y + C_1 + d_z C_2 \|\beta\|_\infty}{1-\theta},$$

and so Assumption 4.2 is satisfied.

Next,

$$\mathbb{E} \left[ \left( \sum_{i=1}^N \sum_{t=1}^T v_{it} z_{it}^y \right)^2 \right] = \sum_{i=1}^N \sum_{t=1}^T \mathbb{E} [(v_{it} z_{it}^y)^2] \leq \sum_{i=1}^N \sum_{t=1}^T \sqrt{\mathbb{E}[v_{it}^4] \mathbb{E}[(z_{it}^y)^4]} \leq CNT$$

for some constant  $C > 0$  by (16), the triangle inequality, Assumptions 3.1 and 3.2, the fact that  $\|y_{i0}\|_{\psi_2} \leq C_y$  for all  $i = 1, \dots, N$ , and (2.15) in [32]. Thus, Assumption 4.3 is satisfied as well. Finally, regarding Assumption 4.4, observe that the convergence  $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \mathring{z}_{it} \mathring{z}'_{it} = \mathring{\Sigma} + O_P(1/\sqrt{NT})$  holds by a law of large numbers as long as the dependence of  $(z_{it}, v_{it})$ 's across  $t$  is not too strong. Also, by Assumption 3.4, the matrix  $\mathring{\Sigma}$  has at most one zero eigenvalue, whereas Assumption 4.4 requires that  $\mathring{\Sigma}$  has no zero eigenvalues, so that  $\mathring{\Sigma}$  is invertible. Although it seems difficult to provide low-level conditions implying invertibility of  $\mathring{\Sigma}$ , as it depends on the autocovariance structure of the random processes  $(z'_{it}, v_{it})'$ ,  $t = 1, \dots, T$ , we note that  $\mathring{\Sigma}$

can be consistently estimated (as it is done in the process of constructing the spectral estimator), and so the invertibility condition is testable.

## REFERENCES

- [1] Ahn, S., Lee, Y., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics* **101** 219-255.
- [2] Ahn, S., Lee, Y., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics* **174** 1-14.
- [3] Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics* **31** 163-191.
- [4] Ando, T. and Bai, J. (2017). Clustering huge number of financial time series: a panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association* **112** 1182-1198.
- [5] Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Mach Learn* **75** 245-248.
- [6] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191-221.
- [7] Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229-1279.
- [8] Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing unobserved heterogeneity. *Working paper*.
- [9] Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* **83** 1147-1184.
- [10] Chen, H., Leng, X., and Wang, W. (2019). Latent group structures with heterogeneous distributions: identification and estimation. *Working paper*.
- [11] Cheng, X., Schorfheide, F., and Shao, P. (2019). Clustering for multi-dimensional heterogeneity. *Working paper*.
- [12] Connor, G. and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: a new framework for analysis. *Journal of Financial Economics* **15** 373-394.
- [13] Connor, G. and Korajczyk, R. (1988). Risk and return in an equilibrium APT: application to a new test methodology. *Journal of Financial Economics* **21** 255-289.
- [14] Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electronic Journal of Probability*.

- [15] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* **82** 540-554.
- [16] Gao, Z. and Shi, Z. (2020). Implementing convex optimization in R: two econometric examples. *Computational Economics*.
- [17] Gu, J. and Volgushev, S. (2019). Panel data quantile regression with grouped fixed effects. *Journal of Econometrics* **213** 68-91.
- [18] Hahn, J. and Moon, R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory* **26** 863-881.
- [19] Halko, N., Martinsson, P., and Tropp, J. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* **53** 217-288.
- [20] Han, F. and Li, Y. (2020). Moment bounds for large autocovariance matrices under dependence. *Journal of Theoretical Probability* **33** 1445-1492.
- [21] Holtz-Eakin, D., Newey, W., and Rosen, H. (1988). Estimating vector autoregressions with panel data. *Econometrica* **56** 1371-1395.
- [22] Lin, C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods* **1** 42-55.
- [23] Moitra, A. (2018). Algorithmic aspects of machine learning. *Cambridge University Press*.
- [24] Moon, R. and Weidner, M. (2019). Nuclear norm regularized estimation of panel regression models. *Working paper*.
- [25] Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967-1012.
- [26] Stock, J. and Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics* **44** 293-335.
- [27] Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* **206** 554-573.
- [28] Su, L., Shi, Z., and Phillips, P. (2016). Identifying latent structures in panel data. *Econometrica* **84** 2215-2264.
- [29] Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society. Series B* **58** 267-288.
- [30] Van der Geer, S. (2002). On Hoeffding's inequality for dependent random variables. *In Empirical Process Techniques for Dependent Data*.

- [31] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68** 841-860.
- [32] Vershynin, R. (2018). High-dimensional probability: an introduction with applications in data science. *Cambridge Series in Statistical and Probabilistic Mathematics*.
- [33] Wang, W., Phillips, P., and Su, L. (2018). Homogeneity pursuit in panel data models: theory and application. *Journal of Applied Econometrics* **33** 797-815.
- [34] Westerlund, J. and Urbain, J.-P. (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters* **119** 247-250.
- [35] Yu, Y., Wang, T., and Samworth, R. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102** 315-323.
- [36] Zhang, Y., Wang, H., and Zhu, Z. (2017). Quantile-regression-based clustering for panel data. *Working paper*.

TABLE 1. Mean Absolute Error (MAE) of spectral, post-spectral, grouped fixed-effect, and oracle estimators of  $\beta_1$  and fraction of misclassified units for spectral and grouped fixed-effect estimators. Case  $\rho = 1$  and  $M = 1$ .

$G = 2$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0330	.0126	.0122	.0121	.003	.002
	200	.0188	.0096	.0095	.0094	.003	.002
	300	.0148	.0074	.0075	.0074	.004	.003
50	100	.0129	.0089	.0089	.0089	.000	.000
	200	.0105	.0056	.0056	.0055	.000	.000
	300	.0091	.0052	.0052	.0052	.000	.000
100	100	.0095	.0042	.0042	.0042	.000	.000
	200	.0064	.0043	.0043	.0043	.000	.000
	300	.0051	.0044	.0044	.0044	.000	.000
$G = 4$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0392	.0278	.0136	.0121	.065	.014
	200	.0211	.0145	.0089	.0089	.027	.008
	300	.0175	.0107	.0090	.0084	.014	.007
50	100	.0225	.0070	.0070	.0070	.000	.000
	200	.0109	.0052	.0052	.0052	.000	.000
	300	.0090	.0057	.0057	.0057	.000	.000
100	100	.0110	.0055	.0055	.0055	.000	.000
	200	.0075	.0049	.0049	.0049	.000	.000
	300	.0057	.0030	.0030	.0030	.000	.000
$G = 7$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0603	.0840	.0129	.0094	.222	.010
	200	.0304	.0480	.0122	.0081	.137	.017
	300	.0170	.0359	.0080	.0071	.111	.016
50	100	.0235	.0087	.0085	.0085	.001	.000
	200	.0126	.0059	.0055	.0055	.001	.000
	300	.0095	.0046	.0045	.0045	.000	.000
100	100	.0136	.0061	.0061	.0061	.000	.000
	200	.0066	.0032	.0032	.0032	.000	.000
	300	.0060	.0041	.0041	.0041	.000	.000

TABLE 2. Mean Absolute Error (MAE) of spectral, post-spectral, grouped fixed-effect, and oracle estimators of  $\beta_1$  and fraction of misclassified units for spectral and grouped fixed-effect estimators. Case  $\rho = 3$  and  $M = 1$ .

$G = 2$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0355	.0139	.0139	.0121	.002	.002
	200	.0192	.0101	.0115	.0094	.003	.004
	300	.0142	.0086	.0085	.0074	.003	.003
50	100	.0131	.0089	.0089	.0089	.000	.000
	200	.0106	.0056	.0056	.0055	.000	.000
	300	.0089	.0052	.0052	.0052	.000	.000
100	100	.0091	.0042	.0042	.0042	.000	.000
	200	.0066	.0043	.0043	.0043	.000	.000
	300	.0049	.0044	.0044	.0044	.000	.000

  

$G = 4$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0376	.0388	.0743	.0121	.050	.159
	200	.0239	.0256	.0234	.0089	.026	.035
	300	.0180	.0152	.0113	.0084	.013	.007
50	100	.0184	.0070	.0070	.0070	.000	.000
	200	.0111	.0052	.0052	.0052	.000	.000
	300	.0094	.0057	.0057	.0057	.000	.000
100	100	.0121	.0055	.0055	.0055	.000	.000
	200	.0082	.0049	.0049	.0049	.000	.000
	300	.0054	.0030	.0030	.0030	.000	.000

  

$G = 7$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.043	.092	.248	.009	.170	.648
	200	.025	.070	.233	.008	.112	.625
	300	.020	.060	.196	.007	.099	.531
50	100	.027	.010	.008	.008	.004	.000
	200	.012	.005	.005	.005	.000	.000
	300	.010	.005	.004	.004	.000	.000
100	100	.015	.006	.006	.006	.000	.000
	200	.007	.003	.003	.003	.000	.000
	300	.006	.004	.004	.004	.000	.000

TABLE 3. Mean Absolute Error (MAE) of spectral, post-spectral, grouped fixed-effect, and oracle estimators of  $\beta_1$  and fraction of misclassified units for spectral and grouped fixed-effect estimators. Case  $\rho = 1$  and  $M = 2$ .

$G = 2$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0328	.0159	.0137	.0118	.280	.164
	200	.0172	.0122	.0090	.0062	.246	.152
	300	.0124	.0103	.0066	.0052	.218	.126
50	100	.0194	.0076	.0060	.0064	.126	.054
	200	.0117	.0067	.0055	.0051	.084	.051
	300	.0056	.0052	.0040	.0035	.090	.040
100	100	.0105	.0048	.0050	.0048	.013	.009
	200	.0063	.0034	.0035	.0035	.012	.009
	300	.0041	.0025	.0023	.0022	.016	.008
$G = 4$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0452	.0315	.0228	.0116	.515	.336
	200	.0263	.0222	.0138	.0062	.478	.296
	300	.0160	.0216	.0124	.0057	.462	.275
50	100	.0233	.0198	.0078	.0070	.342	.058
	200	.0140	.0140	.0055	.0039	.254	.073
	300	.0081	.0118	.0044	.0034	.225	.073
100	100	.0129	.0098	.0056	.0055	.156	.006
	200	.0069	.0048	.0031	.0032	.075	.015
	300	.0057	.0031	.0023	.0022	.039	.011
$G = 7$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0773	.0429	.0277	.0116	.661	.400
	200	.0466	.0441	.0205	.0077	.669	.364
	300	.0291	.0366	.0166	.0048	.634	.333
50	100	.0269	.0344	.0068	.0067	.507	.031
	200	.0151	.0228	.0058	.0051	.404	.060
	300	.0110	.0184	.0043	.0037	.336	.074
100	100	.0178	.0178	.0058	.0058	.266	.001
	200	.0097	.0090	.0033	.0034	.185	.004
	300	.0061	.0076	.0024	.0022	.141	.014



TABLE 4. Mean Absolute Error (MAE) of spectral, post-spectral, grouped fixed-effect, and oracle estimators of  $\beta_1$  and fraction of misclassified units for spectral and grouped fixed-effect estimators. Case  $\rho = 3$  and  $M = 2$ .

$G = 2$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0297	.0289	.0212	.0118	.270	.182
	200	.0178	.0274	.0190	.0062	.248	.172
	300	.0125	.0230	.0148	.0052	.200	.136
50	100	.0197	.0134	.0076	.0064	.117	.053
	200	.0120	.0114	.0078	.0051	.084	.054
	300	.0055	.0116	.0056	.0035	.090	.042
100	100	.0100	.0055	.0053	.0048	.013	.009
	200	.0063	.0037	.0037	.0035	.011	.009
	300	.0042	.0034	.0024	.0022	.016	.008

  

$G = 4$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0401	.0657	.0693	.0116	.510	.451
	200	.0288	.0504	.0484	.0062	.464	.371
	300	.0175	.0498	.0365	.0057	.461	.325
50	100	.0205	.0425	.0089	.0070	.320	.038
	200	.0127	.0295	.0095	.0039	.219	.065
	300	.0077	.0278	.0094	.0034	.211	.075
100	100	.0138	.0214	.0057	.0055	.151	.002
	200	.0096	.0105	.0032	.0032	.074	.010
	300	.0057	.0063	.0028	.0022	.041	.012

  

$G = 7$							
$T$	$N$	Mean Absolute Error				Misclassification	
		Spectral	P-Spectral	GFE	Oracle	Spectral	GFE
20	100	.0783	.0815	.1018	.0116	.637	.608
	200	.0452	.0822	.0911	.0077	.646	.605
	300	.0321	.0755	.0619	.0048	.612	.449
50	100	.0270	.0689	.0107	.0067	.481	.045
	200	.0161	.0516	.0080	.0051	.376	.056
	300	.0138	.0450	.0090	.0037	.330	.076
100	100	.0170	.0348	.0058	.0058	.226	.001
	200	.0128	.0226	.0034	.0034	.166	.003
	300	.0070	.0200	.0026	.0022	.138	.008

TABLE 5. Mean Absolute Error (MAE) of spectral, post-spectral, grouped fixed-effect 1, grouped fixed-effect 2, and oracle estimators of  $\beta_1$ . Case  $\varrho = 6$ ,  $G = 5$ , and  $M = 1$ . Grouped fixed-effect 1 (GFE1) estimator is the grouped fixed-effect estimator obtained by applying Algorithm 1 of [9] with 50 initial values where the first 49 initial values are random and the 50th initial value is an oracle estimator. Grouped fixed-effect 2 (GFE2) estimator is the grouped fixed-effect estimator obtained by applying Algorithm 1 of [9] with 50 random initial values.

$T$	$N$	Mean Absolute Error				
		Spectral	P-Spectral	GFE1	GFE2	Oracle
20	100	.0745	.0682	.1229	.1234	.0121
	250	.0393	.0699	.1260	.1263	.0096
	500	.0216	.0711	.1250	.1252	.0051
40	100	.0391	.0679	.1256	.1212	.0102
	250	.0205	.0544	.1259	.1259	.0065
	500	.0104	.0483	.1261	.1244	.0037
80	100	.0244	.0422	.0914	.1206	.0078
	250	.0099	.0276	.0973	.1264	.0042
	500	.0065	.0164	.0510	.1217	.0029
160	100	.0112	.0247	.0053	.1243	.0052
	250	.0059	.0056	.0029	.1181	.0028
	500	.0037	.0028	.0021	.1152	.0022

(D. Chetverikov) DEPARTMENT OF ECONOMICS, UCLA, BUNCHE HALL, 8283, 315 PORTOLA PLAZA, LOS ANGELES, CA 90095, USA.

*Email address:* `chetverikov@econ.ucla.edu`

(E. Manresa) DEPARTMENT OF ECONOMICS, NEW YORK UNIVERSITY, 19 WEST 4TH STREET, NEW YORK, NY 10003, USA.

*Email address:* `em1849@nyu.edu`