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 $\sqrt{NT}$-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 comparable to those of the grouped fixed-effect estimator developed by Bonhomme and Manresa in [11]. In contrast to the grouped fixed-effect estimator, however, our post-spectral estimator is computationally straightforward.

1. Introduction

Consider a grouped panel data model

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

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

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

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. Also, throughout the paper, we assume that the number of groups G is known (consistently estimated).

Date: December 29, 2022.

We are grateful to Tim Armstrong, Stephane Bonhomme, Victor Chernozhukov, Andrew Chesher, Jin Hahn, Yusuke Narita, Martin Weidner, Daniel Wilhelm, Andrei Zeleneev, participants at many seminars, and especially to Arturas Juodis, Andres Santos, and our discussant at Chamberlain's seminar Roger Moon for useful comments. We also thank Martin Weidner and Roger Moon for providing the code for their estimator.

The model (1) was originally introduced by Bonhomme and Manresa in [11], who also developed a so-called grouped fixed-effect estimator of the vector of parameters β in this model. This estimator has attractive theoretical properties but 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 to be specified in the next paragraph, and propose an estimator of β, 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 $T \to \infty$, potentially very slowly, as $N \to \infty$, since otherwise β 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 for some $M \geq 1$,

$$x_{it} = \sum_{m=1}^{M} \rho_{im} \alpha_{g_i t}^m + z_{it}, \quad \text{for all } i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{3}$$

where $(\alpha_{1t}^m, \ldots, \alpha_{Gt}^m)' \in \mathbb{R}^G$ for $m = 1, \ldots, M$ are group-specific time effects, $\rho_{im} \in \mathbb{R}^d$ for $m = 1, \ldots, M$ 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 $\rho_{i1}, \ldots \rho_{iM}$. Here, no quantity on the right-hand side of (3) is observed, except for the number of time effects M, which we assume to be known (consistently estimated). Also, without loss of generality, we assume that $\alpha_{\gamma t}^1 = \alpha_{\gamma t}$ for all $\gamma = 1, \ldots, G$ and $t = 1, \ldots, T$. 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. In Appendix E, we also provide an example in terms of agricultural production functions and environmental economics to motivate equation (3).

Our post-spectral estimator consists of three steps. In the first step, we carry out preliminary estimation of β. To do so, we prove that as long as the data-generating process is given by equations (1) and (3), there exists a convex quadratic function $f \colon \mathbb{R}^d \to \mathbb{R}$ such that (i) its unique minimum is achieved by β 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 β 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 β 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, which is a version of spectral clustering method ([38, 28, 39, 27]), is fast and does not require solving any non-convex optimization problems. In the third step, we obtain the *post-spectral estimator* of β 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 [11].[1] 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 β is asymptotically equivalent to the pooled-OLS estimator with known group memberships (i.e., the oracle estimator). In turn, the latter is $\sqrt{NT}$ 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 β 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 comparable 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, 12, 13, 43]), are built around the K-means optimization problem. This optimization problem is known to be NP-hard (see [5]), and so it is rather unlikely that there exists a fast algorithm for finding its solutions. Any proposed fast implementation of the aforementioned estimators therefore is likely to 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 [11] proceeds by initializing randomly selected values of parameters β and $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$ and then alternating between two steps: (1) optimization over the values of group memberships

[1] We make no assumptions on the distance between vectors $(\alpha_{\gamma 1}^m, \dots, \alpha_{\gamma T}^m)'$ for $m \geq 2$.

$g_1, \ldots, g_N$ given the values of parameters β and $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$, and (2) optimization over the values of parameters β and $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$ given the values of group memberships $g_1, \ldots, g_N$. 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 [11] 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. [11] 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 potential failure problem.

There is also growing literature on non-linear panel data models with group structure of individual-level parameters ([21, 26, 35, 10, 34, 40, 20, 19]) originated by Hahn and Moon in [21]. 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 $\alpha_{g_i t}$ 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 β 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 [9] is based on a solution to a non-convex optimization problem, and the estimators in [25, 1, 2] require certain IV-type conditions. Like in our approach, the estimators in [32] also require restricting the data-generating process for the covariates x_{it} but the nature of imposed restrictions is very different. In particular, [32] 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),[2] thus leading to a random, rather than fixed, effect model. In fact, the only exception in this literature we are aware of is [29], who developed a computationally relatively simple method and used conditions that are similar to (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 much more acceptable for the preliminary estimation. Indeed, we find via simulations reported below that our spectral estimator leads to much better results in reasonably large samples. Finally, [7] has recently developed a method for debiasing estimators with the slow convergence rate $(T \wedge N)^{-1/2}$ yielding estimators with the fast convergence rate $(T \wedge N)^{-1}$. An advantage of our estimator here is that we obtain the fast rate in just one step, instead of two steps, which again may be preferable when T is relatively small, so that the original estimator to be debiased is rather imprecise and is hard to debias.

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 β is sparse in the sense that it has relatively few non-zero components. We demonstrate how to modify the spectral estimator via ℓ^1-penalization to obtain a computationally simple and consistent estimator of β in this case. Third, we consider an interactive fixed-effect panel data model, where α_{g_it} in (1) is replaced by $\kappa_i'\phi_t$, with ϕ_t being a vector of factors and κ_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).

[2][32] claims that his estimators are consistent without requiring independence of the factors but a counter-example is given in [41], which proves that if the rank condition is not satisfied, then independence is essentially a necessary condition for consistency of the estimators in [32].

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 state their asymptotic properties. In Section 4, we provide the extensions of the baseline model. In Section 5, we discuss 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 main proofs. In Appendix A, we collect some technical lemmas that are useful for the proofs of our main results. In Appendix B, we present remaining proofs. In Appendix C, we describe a method for calculating eigenvalues of large matrices, which may be needed for implementing our estimators. In Appendix D, we provide some details on the assumptions of the dynamic model extension. In Appendix E, we discuss an example motivating equation (3).

2. Estimation

Our proposed estimation procedure consists of three steps. The first step is preliminary consistent estimation of β, 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 β on classified units, which gives the post-spectral estimator. Under the assumption of well-separated groups, the post-spectral estimator will be $\sqrt{NT}$-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

$$A^b_{ij} = \frac{1}{NT} \sum_{t=1}^{T} \Big\{ (y_{it} - x'_{it} b) - (y_{jt} - x'_{jt} b) \Big\}^2, \quad \text{for all } i, j = 1, \ldots, N, \tag{4}$$

Since A^b is an $N \times N$ symmetric matrix, it has N real eigenvalues. Let $\lambda^b_1, \ldots, \lambda^b_{2GM+2}$ be its $2GM+2$ largest in absolute value eigenvalues. We will show below that under mild conditions,

$$\lambda^b_1 + \cdots + \lambda^b_{2GM+2} = b'\Sigma b + S'b + L + o_P(1), \quad \text{for all } b \in \mathbb{R}^d, \tag{5}$$

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

$$\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)' = \arg\min_{b \in \mathbb{R}^d} \Big\{ b'\hat{\Sigma} b + \hat{S}'b + \hat{L} \Big\}, \tag{6}$$

where $\hat{\Sigma}$, $\hat{S}$, and $\hat{L}$ are estimators of Σ, 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 $\tilde{\beta} \to_P \beta$.

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

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

so that by (5),

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

Also, for all $k = 1, \ldots, d$, let $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)'$ be the $d \times 1$ vector with 1 in the k-th position and 0 in all other positions, and let $0_d = (0, \ldots, 0)'$ 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 $\hat{f}(e_k) - \hat{f}(-e_k) \to_P 2S_k$, we set

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

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

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

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 $k, l = 1, \ldots, d,\ k > l$, and let $\hat{\Sigma}$ be the matrix whose (k, l)-th component is $\hat{\Sigma}_{kl}$.[3]

Under result (5), the estimators $\hat{\Sigma}$ and $\hat{S}$ are consistent for Σ and S, respectively, and so $\tilde{\beta} = -\hat{\Sigma}^{-1}\hat{S}/2 \to_P -\Sigma^{-1}S/2 = \beta$, as long as Σ 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, however, 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 $\lambda_1^b, \ldots, \lambda_{2GM+2}^b$. Fortunately, there exists a class of fast randomized algorithms that allow to calculate these eigenvalues arbitrarily well; see Appendix C for details.

[3]Note that the quality of the estimators $\hat{\Sigma}$, $\hat{S}$, and $\hat{L}$ could potentially be improved by exploiting additional values of the vector b but we leave the question of optimal estimation for future work.

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

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

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. ■

Remark 2.2 (Tuning Parameters for Spectral Estimator)**.** Implementing the spectral estimator $\tilde{\beta}$ requires choosing the product GM but does not require knowing G and M separately, which means that we only need one tuning parameter instead of two of them. In addition, the proof of Theorem 3.1 below reveals that consistency of the spectral estimator holds even if we replace GM in the definition of the estimator $\tilde{\beta}$ by any number that is bigger than GM (as long as it is independent of N and T). Thus, to implement the spectral estimator, we actually only need an upper bound on the product GM. Moreover, the proof of Theorem 3.1 also shows that for any vector $b \in \mathbb{R}^d$, the matrix A^b has at most $2GM + 2$ eigenvalues that are not asymptotically vanishing. This suggests a method to estimate the product GM by counting the number of eigenvalues of the matrix A^b exceeding certain threshold, which is chosen to slowly converge to zero. This method can underestimate the product GM, which happens if the matrix A^b actually has fewer than $2GM + 2$ eigenvalues that are not asymptotically vanishing, but whenever this happens, we can lose only asymptotically vanishing eigenvalues in the sum (7), which can not break consistency of the spectral estimator. For brevity of the paper, however, we leave the question of formally deriving results with an estimated product GM to future work. ■

2.2. **Classifier.** To classify units into groups, we will use a version of the spectral clustering method.[4] For reasons to be explained in Remark 3.2 in the next section, we will also rely on sample splitting. To this end, let $h_1, \ldots, h_N$ 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 $i = 1, \ldots, N$ into two subsamples, $\mathcal{I}_0 = \{i = 1, \ldots, N \colon h_i = 1\}$ and $\mathcal{I}_1 = \{i = 1, \ldots, N \colon h_i = 0\}$. Further, for $i = 1, \ldots, N$, denote $y_i = (y_{i1}, \ldots, y_{iT})'$ and $x_i = (x_{i1}, \ldots, x_{iT})'$. Also, for $h = 0, 1$, let $\tilde{\beta}^h$

[4]Note that the special structure on the data-generating process for the covariates x_{it} given in (3) was used for the construction of the spectral estimator only. This structure has no role for our classifier and for the post-spectral estimator described below.

be the spectral estimator calculated using the subsample $\mathcal{I}_h$ and let $\hat{B}^h$ be a $T \times T$ matrix given by

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

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

$$\hat{A}_i = \hat{F}_{h_i} \hat{F}_{h_i}' (y_i - x_i \tilde{\beta}^{h_i}).^5 \tag{10}$$

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 $\mathcal{C}_i = \{\gamma = 1, \dots, m \colon \|\hat{A}_i - |\mathcal{A}_\gamma|^{-1} \sum_{l \in \mathcal{A}_\gamma} \hat{A}_l\| \leq \lambda\}$;
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 λ, so that $m = m(\lambda)$. Clearly, $\lambda \mapsto m(\lambda)$ is a right-continuous function, and so

$$\hat{\lambda} = \min \Big\{ \lambda > 0 \colon m(\lambda) \leq G \Big\}$$

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

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

[5]Thus, for all units i with $h_i = 1$, we calculate $\hat{A}_i$ using $\hat{F}_1$ and $\tilde{\beta}_1$, which are obtained from the subsample $\mathcal{I}_1$ consisting of all units j with $h_j = 0$ and, vice versa, for all units i with $h_i = 0$, we calculate $\hat{A}_i$ using $\hat{F}_0$ and $\tilde{\beta}_0$, which are obtained from the subsample $\mathcal{I}_0$ consisting of all units j with $h_j = 1$.

and $\hat{g} = (\hat{g}_1, \ldots, \hat{g}_N)'$. 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

$$\mathrm{P}\Big(\text{for all } i, j = 1, \ldots, N, \quad \hat{g}_i = \hat{g}_j \text{ if and only if } g_i = g_j\Big) \to 1, \tag{11}$$

under the assumption of well-separated groups.

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 β 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 $i = 1, \ldots, N$. Then equation (3) becomes

$$x_{it} = \rho_{i2}\alpha^2_{g_i t} + z_{it}, \quad \text{for all } i = 1, \ldots, N,\ t = 1, \ldots, T.$$

Now, if we assume that $G = 3$ but $\alpha^2_{1t} = \alpha^2_{2t} \neq \alpha^2_{3t}$, 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. ■

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

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

where $\mathcal{B}$ is a parameter space for the vector β, 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,

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

and thus to the grouped fixed-effect estimator of [11]. Hence, the standard OLS inference ignoring group classification applies.

Remark 2.4 (Estimating β 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

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

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 ϕ_t and the factor loadings ω_i can be estimated by the method of asymptotic principle components as in Section 3 of [8]; see also [15, 16, 33, 18]. 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 β 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 α_{g_it} and $x_{it} - z_{it}$. However, it performs poorly in the case of weak factors, i.e. when the factor loadings ω_i are close to zero, as factors ϕ_t can not be consistently estimated in this case; see [14, 31] for details. In particular, our simulation experience confirms that the post-spectral estimator substantially outperforms this simple estimator in the case of weak factors.

Remark 2.5 (Tuning Parameters for Post-Spectral Estimator)**.** Implementing the post-spectral estimator $\hat{\beta}$ requires choosing the number of groups G but does not require to specify the number of time effects M in the equation for covariates. Thus, like in the case of the spectral estimator, we only need one tuning parameter instead of two to implement the post-spectral estimator. In turn, estimating the number of groups G is relatively easy. In particular, we can employ penalization techniques as developed in [8], in the same fashion as discussed in [11]. However, for brevity of the paper, we leave the question of formally deriving results with an estimated number of groups G to future work. ∎

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 ρ_{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. $\mathcal{S}^T = \{u \in \mathbb{R}^T \colon \|u\| = 1\}$. Also, for any random variable w, let $\|w\|_{\psi_2}$ denote the sub-Gaussian norm of w, i.e.

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

see Section 2.5.2 in [39] on properties of the sub-Gaussian norm.[6] 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. *(i) For some constant $C_1 > 0$, we have $\|\sum_{t=1}^T u_t v_{it}\|_{\psi_2} \leq C_1$ for all $i = 1, \ldots, N$ and $u = (u_1, \ldots, u_T)' \in \mathcal{S}^T$. (ii) In addition, for some constant $C_2 > 0$, we have $\|\sum_{t=1}^T u_t z_{itk}\|_{\psi_2} \leq C_2$ for all $i = 1, \ldots, N$, $u = (u_1, \ldots, u_T)' \in \mathcal{S}^T$, and $k = 1, \ldots, d$.*

By Hoeffding's inequality (Proposition 2.6.1 in [39]), Assumption 3.1(i) 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 [17, 37]), Assumption 3.1(i) holds as long as the dependence of the random variables v_{it} across t is not too strong. Assumption 3.1(ii) is similar to Assumption 3.1(i) 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 Theorem 3.1 below, with slower rates, can be derived under weaker integrability assumptions. We have chosen to work with Assumption 3.1 in order to minimize technicalities of our analysis.

Assumption 3.2. *(i) We have $\|(NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T v_{it} z_{it}\| = O_P(1/\sqrt{NT})$. (ii) In addition, $(NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T z_{it} z_{it}' = \Sigma/2 + O_P(1/\sqrt{NT})$, where Σ is a positive definite $d \times d$ matrix.*

Since $\mathbb{E}[v_{it}] = 0$, $\mathbb{E}[v_{it} x_{it}] = 0_d$, and we assume that $\alpha_{g_{it}}^m$ and ρ_{im} are non-stochastic, it follows from (3) that $\mathbb{E}[v_{it} z_{it}] = 0_d$. Hence, Assumption 3.2(i) 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 3.2(ii) 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 3.2(ii) also imposes the constraint that the probability limit of this matrix is positive-definite, which is an identification condition.

[6] Sub-Gaussian norm also often appears in the literature under the name of Orlicz norm.

Theorem 3.1 (Rate of Convergence of Spectral Estimator $\tilde{\beta}$)**.** *Under Assumptions 3.1 and 3.2,*

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

Remark 3.1 (Relaxing Data-Generating Process for Covariates)**.** Inspecting the proof of Theorem 3.1 reveals that the theorem continues to hold even if we allow for a substantially larger class of data-generating processes for covariates instead of that specified in (3). Indeed, if we simply assume that $x_{it} = \varsigma_{it} + z_{it}$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$ and some $N \times T$ matrix ς of rank M, then Theorem 3.1 holds as long as the spectral estimator $\hat{\beta}$ uses $2(G + M + 1)$ instead of $2GM + 2$ eigenvalues of the matrices A^b. Throughout the paper, however, we prefer to work with (3) as this seems to be the most natural assumption on the data-generating process for covariates.[7] ■

3.2. **Classifier.** For all $\gamma = 1, \ldots, G$, let $N_\gamma = 1\{g_i = \gamma\}$ be the number of units i within group γ. To prove consistency of the classifier $\hat{g}$, we will use the following assumptions.

Assumption 3.3. *(i) For some constant $C_3 > 0$, we have $\|\rho_{im}\| \leq C_3$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. (ii) In addition, for some constant $C_4 > 0$, we have $|\alpha_{\gamma t}^m| \leq C_4$ for all $m = 1, \ldots, M$, $\gamma = 1, \ldots, G$, and $t = 1, \ldots, T$.*

Assumption 3.4. *For some constant $c_1 > 0$, we have $T^{-1} \sum_{t=1}^T (\alpha_{\gamma_1 t} - \alpha_{\gamma_2 t})^2 \geq c_1$ for all $\gamma_1, \gamma_2 = 1, \ldots, G$ such that $\gamma_1 \neq \gamma_2$.*

Assumption 3.3 is self-explanatory. Assumption 3.4 means that the groups are well-separated in the sense that the vectors of group-specific time effects, $(\alpha_{\gamma 1}, \ldots, \alpha_{\gamma T})'$ for $\gamma = 1, \ldots, G$, are not too close to each other. We will use this assumption to prove consistency of the classifier $\hat{g}$ and to derive asymptotic normality of the post-spectral estimator $\hat{\beta}$ but we will not use it to prove consistency of the post-spectral estimator $\hat{\beta}$. Note, however, that if groups are not well-separated, it is possible that the spectral estimator actually outperforms the post-spectral one. Also, note that Assumption 3.4 is sufficient for consistency of the classifier $\hat{g}$ but by no means necessary. In particular, using more sophisticated arguments as in [27] and stronger conditions on the noise variables v_{it} (i.e. isotropic Gaussianity, which means that the random variables v_{it} are i.i.d. centered Gaussian), one can replace Assumption 3.4 by a much weaker condition $\sum_{t=1}^T (\alpha_{\gamma_1 t} - \alpha_{\gamma_2 t})^2 \geq c \log N$ for all $\gamma_1, \gamma_2 = 1, \ldots, G$ such that $\gamma_1 \neq \gamma_2$ and a suitable constant $c > 0$.

[7]The representation $x_{it} = \varsigma_{it} + z_{it}$, however, emphasizes the fact that our methods are able to deal with the case when the equation for covariates x_{it} have more groups than the equation for the dependent variable y_{it}.

Assumption 3.5. *For some constant $c_2 > 0$, we have $N_\gamma \geq c_2 N$ for all $\gamma = 1, \ldots, G$.*

Assumption 3.5 requires that each group $\gamma = 1, \ldots, G$ constitutes a non-trivial fraction of all units. If we were to assume random group assignment, where each unit is assigned to group γ with probability $\mathrm{p}_\gamma > 0$, so that $\sum_{\gamma=1}^{G} \mathrm{p}_\gamma = 1$, and units are assigned independently, this assumption would be satisfied with probability approaching one as $N \to \infty$.

Assumption 3.6. *We have $\log N = o(T)$ and $\log T = o(N)$.*

Assumption 3.6 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 [11].

Theorem 3.2 (Consistency of Classifier $\hat{g}$)**.** *Under Assumptions 3.1–3.6, we have*

$$\mathrm{P}\Big(\textit{for all } i, j = 1, \ldots, N, \textit{ we have } \hat{g}_i = \hat{g}_j \textit{ if and only if } g_i = g_j\Big) \to 1,$$

as $N \to \infty$.

Remark 3.2 (On the Role of Sample Splitting in Theorem 3.2)**.** Using sample splitting to construct the vectors $\hat{A}_1, \ldots, \hat{A}_N$, 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 $i = 1, \ldots, N$, 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 3.4) 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, $T^{-1/2}\hat{F}'(v_{i1}, \ldots, v_{iT})'$ may not converge to zero in probability (which is guaranteed in the construction based on sample splitting), and the vectors $\hat{A}_1, \ldots, \hat{A}_N$ may turn out poor estimators of the vectors $\alpha_{g_1}, \ldots, \alpha_{g_N}$, 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. See, however, [27], who are able to avoid such conditions without using sample splitting under the isotropic Gaussianity condition mentioned above.[8] ∎

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 3.4, this estimator is $\sqrt{NT}$-consistent and has simple asymptotic distribution.

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

$$\mathcal{I}(\nu, \gamma_1, \gamma_2) = \Big\{ i = 1, \ldots, N \colon \nu_i = \gamma_1 \text{ and } g_i = \gamma_2 \Big\}$$

and

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

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

Assumption 3.7. *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 for all $\nu \in \{1, \ldots, G\}^N$ with probability $1 - o(1)$.

Assumption 3.8. *(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 $\gamma = 1, \ldots, G$ and $t = 1, \ldots, T$.*

Assumption 3.7 requires that covariates x_{it} have sufficient within-group variation over time and across units. This assumption was used in [11] as well. Assumption 3.8 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 3.1–3.3 and 3.5–3.8, we have $\hat{\beta} \to_P \beta$.*

[8] As a side note, we also observe that [38] do not use sample splitting to estimate vectors $\alpha_{g_1}, \ldots, \alpha_{g_N}$ but some parts of their derivations are difficult to verify. In particular, in Section 5, they use an 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.

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

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

and

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

We will use the following condition:

Assumption 3.9. *We have (i)* $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it}\check{x}_{it}' \to_P \check{\Sigma}$ *for some positive-definite* $d \times d$ *matrix* $\check{\Sigma}$ *and (ii)* $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T v_{it}\check{x}_{it} \to_D N(0, \Omega)$ *for some symmetric* $d \times d$ *matrix* Ω.

This assumption is similar to the corresponding assumptions in [11].

Theorem 3.4 (Asymptotic Distribution of Post-Spectral Estimator $\hat{\beta}$)**.** *Under Assumptions 3.1–3.9, we have*

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

Remark 3.3 (Variance-Covariance Matrix Estimation)**.** Theorem 3.4 leads to standard inference on the vector of parameters β as long as we can consistently estimate the asymptotic variance-covariance matrix $\check{\Sigma}^{-1}\Omega\check{\Sigma}^{-1}$. In turn, the latter is simple. Indeed, Assumption 3.9(i) implies that we can consistently estimate $\check{\Sigma}$ by $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \check{x}_{it}\check{x}_{it}'$. Also, to estimate Ω, we can use a formula from [6]: $(NT)^{-1} \sum_{i=1}^N \sum_{t_1=1}^T \sum_{t_2=1}^T \hat{v}_{it_1}\hat{v}_{it_2}\check{x}_{it_1}\check{x}_{it_2}'$, where $\hat{v}_{it} = y_{it} - x_{it}'\hat{\beta} - \hat{\alpha}_{\hat{g}_i t}$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Conditions for consistency of this formula are proven in [24]. For more detailed discussion of variance-covariance matrix estimation, please refer to [11]. ∎

Remark 3.4 (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 and 3.2(i). 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, as well as (3). 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.e. a model with high-dimensional β. The third extension is concerned with an interactive fixed-effect model.

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

$$y_{it} = \theta y_{it-1} + x_{it}'\beta + \alpha_{g_i t} + v_{it}, \quad \text{for all } i = 1, \dots, N,\ t = 1, \dots, T, \tag{15}$$

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 θ and β 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

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

where we denoted

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

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

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

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

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

where

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

Thus, the dynamic model considered here reduces to the model studied in Sections 2 and 3 with x_{it}, β, ρ_{im}, $\alpha_{\gamma t}^{m}$, z_{it}, and M replaced by $\mathring{x}_{it}$, $\mathring{\beta}$, $\mathring{\rho}_{im}$, $\mathring{\alpha}_{\gamma t}^{m}$, $\mathring{z}_{it}$, and $2M$, respectively. Therefore, the parameters of the dynamic model, $\mathring{\beta} = (\theta, \beta')'$, can be estimated by the spectral and post-spectral estimators as in Section 2 with M, x_{it}, d, and $\mathcal{B}$ replaced by $2M$, $\mathring{x}_{it}$, $d+1$, and $\mathring{\mathcal{B}}$, respectively, where $\mathring{\mathcal{B}}$ is a parameter space for $\mathring{\beta}$, e.g. $\mathring{\mathcal{B}} = [-1+\delta, 1-\delta] \times \mathcal{B}$ for some $\delta \in (0,1)$.

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, Σ, and $\mathcal{B}$ replaced by $\mathring{z}_{it}$, $\mathring{x}_{it}$, $d+1$, $\mathring{\Sigma}$, and $\mathring{\mathcal{B}}$, respectively.

Theorem 4.1 (Dynamic Model)**.** *In the setting of this section, the statements of Theorems 3.1-3.4 continue to hold, with β replaced by $\mathring{\beta}$, if Assumptions 3.1–3.9 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 D but we note here that one of these extra conditions is that the noise variables v_{it} satisfy

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

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 $\mathring{z}_{it}$ may not be centered but this does not contradict the results of Theorem 4.1 per our discussion in Remark 3.4 ∎

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

$$y_{it} = x_{it}'\beta + \alpha_{g_i t} + v_{it}, \quad \text{for all } i = 1, \ldots, N,\ t = 1, \ldots, T,$$

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 β 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.[9]

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

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

where $\lambda > 0$ is a penalty parameter and $\|b\|_1 = \sum_{k=1}^d |b_k|$ denotes the ℓ^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 Assumption 3.2. Indeed, Assumption 3.2(i) is impossible to satisfy if d is growing together with N and T because of the ℓ^2-norm appearing in the assumption. Assumption 3.2(ii) 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 ℓ^∞-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 Assumption 3.1 is satisfied. In addition, suppose that*

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

and

$$\max_{1\le k,l\le d}\left|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{itk}z_{itl} - \frac{\Sigma_{kl}}{2}\right| = O_P\left(\sqrt{\frac{\log d}{NT}}\right) \tag{18}$$

for some positive definite $d\times d$ matrix Σ. Also, suppose that for some constant $C_z > 0$, we have $\|\sum_{t=1}^T u_t z_{it}'\beta\|_{\psi_2} \le C_z$ 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 $\|\beta\|_1 \le C_\beta$ for some constant $C_\beta > 0$. Finally, let $c_\Sigma > 0$ be the minimal eigenvalue of the matrix Σ, $c_\lambda > 1$ be some

[9]Under consistent classification (11), the ℓ^1-penalized post-spectral estimator will be similar to the usual Lasso estimator, with comparable properties.

constant, and $S = -2\Sigma\beta$ *be a* $d \times 1$ *vector. Then*

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

Moreover, on the intersection of events

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

and

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

we have

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

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

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

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

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

Thus, setting λ 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 λ 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 λ 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

$$y_{it} = x_{it}'\beta + \kappa_i'\phi_t + v_{it}, \quad \text{for all } i = 1, \dots, N,\ t = 1, \dots, T, \tag{25}$$

where ϕ_t is a $J \times 1$ vector of factors and κ_i is a $J \times 1$ vector of factor loadings. [9] developed an OLS-type interactive fixed-effect estimator of β 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, [29] developed an estimator of β 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 β in this model.

In this subsection, we demonstrate that a trivial modification of our spectral estimator can be used to estimate β 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

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

where ϕ_t is a $J \times 1$ vector of factors and ω_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 ϕ_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 and 3.2 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 [29], it is still slower than the convergence rate of the interactive fixed-effect estimator proposed in [9], which is $(NT)^{-1/2}$. Like in [29], 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. We compare the performance in finite samples of the spectral (S) and post-spectral (P-S) estimators to several natural alternatives. Specifically, we consider the grouped fixed effect (GFE) estimator proposed in [11], the least squares (LS) estimator proposed in [9], and the penalized nuclear norm (Pen NN) estimator proposed in [29]. The comparison with LS and Pen NN estimators is pertinent because equation (1) can be seen as a particular case of an interactive fixed effects model where the factors are the group trends, α_{gt}, and the loadings are restricted to be of the form $(0, \ldots, 1, \ldots, 0)$, with 1 in the g_i^{th} position.

Computation of the GFE estimator is not trivial. Even with $N = 100$ and $G = 2$, it is hard to be sure that any given algorithm produces the global minimum of the GFE optimization problem since this would require separately considering G^N partitions of units $i = 1, \ldots, N$ into G groups. Having in mind this issue, we calculate the GFE estimator using Algorithm 1 in [11] with 500 random initial values, which is in a nutshell the Lloyd algorithm with covariates. Given a classification, this algorithm obtains the regression coefficients and the group-specific trends by minimizing OLS, and given the parameters, units are classified into groups according to the smallest individual-specific residual. We generate the initial values from the standard normal distribution in $\mathbb{R}^{2+G\times T}$, where the first two and the last $G \times T$ components correspond to $\beta = (\beta_1, \beta_2)'$ and $\{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$, respectively.

We also consider an infeasible GFE estimator, which uses only one initial value but this value is chosen to be the oracle estimator, i.e. the pooled OLS estimator based

on the true partition of units into groups. Using this value substantially increases the chances to find the global minimum of the GFE optimization problem as we expect the solution to be near the oracle estimator. In what follows, we refer to this infeasible GFE estimator as I-GFE.

Computation of the LS estimator can also be problematic, as its objective function is prone to local minima as well. To deal with this problem, like in the case of the GFE estimator, we use 500 random initial values and choose the one that gives the best value of the LS criterion function.

Computation of the Pen-NN estimator is relatively straightforward: it minimizes a convex objective function. To make the comparison with the other estimators fair, however, we choose a penalty parameter for this estimator so that the total number of generated groups is equal to the true value, G, instead of using the data-driven procedure proposed by the authors. Also, instead of presenting results for the plain Pen-NN estimator, following the original paper [29], we actually present results for the LS estimator that uses the Pen-NN estimator as initial value, as adding an extra step of least squares minimization is supposed to make the estimator more precise. With some abuse of notation, we still refer to this estimator as Pen-NN.[10]

Finally, for all data-generating processes, 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 other estimators.

Next, we discuss the data-generating processes. 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 400 and $T = 20$, 50, or 100. We also set $G = 2$ or 7 and $M = 1$ or 2. The random variables v_{it} and components of the random vectors z_{it} are generated from a truncated standard normal distribution, $Z1\{|Z| \leq C\}$, where $Z \sim N(0, 1)$ and $C = 20$, independently of each other and across indices. The random variables $\alpha_{\gamma t}^m$ are generated from a truncated normal distribution as well but with different variance values, $Z1\{|Z| \leq C\}$, where $Z \sim N(0, \sigma^2)$, independently across indices and of all other random variables. Depending on the experiment, we set σ^2 equal to either 1 or 4. 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 γ to be the same, except for the last group, which contains more observations

[10][29] proposed several other estimators as well but we have chosen to present results for the Pen-NN estimator only, as it dominates the other estimators in our simulations.

if G does not divide N. We have also checked with uneven units across groups and results are unchanged.

Throughout all experiments, we set $\beta = (\beta_1, \beta_2)'$, where $\beta_1 = -1$ and $\beta_2 = .8$ 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 $\rho_{i12} = \varrho + Z_{i2}1\{|Z_{i2}| \leq C\}$, where Z_{i1} and Z_{i2} are independent of each other, across i, and of all other random variables in the data-generating process, and where $\varrho = 3$. Here, ϱ can be understood as a measure of endogeneity in the model as it govern the correlation between covariates and grouped-specific time effects. 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 $\rho_{i12} = \varrho + Z_{i12}1\{|Z_{i12}| \leq C\}$ and $\rho_{i22} = Z_{i22}1\{|Z_{i22}| \leq C\}$, where random variables Z_{imj} for $m = 1, 2$ have standard normal distribution and are independent of each other, across indices, and of all other random variables in the data-generating process.

We present results in Tables 1–4 at the end of Supplementary Materials. For each combination of N and T, these tables give the mean absolute error (MAE) for S, P-S, LS, Pen-NN, I-GFE, GFE, 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 GFE estimator. Each table presents results in the upper panel for $G = 2$ and in the lower panel for $G = 7$. Tables 1 and 2 correspond to $M = 1$ and Tables 3 and 4 correspond to $M = 2$. In addition, Tables 1 and 3 correspond to $\sigma^2 = 1$, and Tables 2 and 4 correspond to $\sigma^2 = 4$. When increasing the variance σ^2, we are effectively making the groups more separated, and hence group classification improves. Both the MAE and the fraction of misclassified units are calculated as the average over 50 simulations.

We now describe the results. We start with the behavior of the S and P-S estimators. In most cases, the P-S estimator outperforms the S estimator. Hence, it seems that doing the "post" step delivers a better estimator. Also, the P-S estimator tends to be comparable to the oracle estimator across all tables as long as N and T increase sufficiently. In addition, while the P-S estimator seems to have more trouble when the separation of the groups is less, i.e. in Tables 1 and 3, it still quickly achieves the oracle for moderate values of N and T. Moreover, when $G = 2$, the P-S estimator seems to achieve the oracle faster than when $G = 7$, but even for small values of N and T, the P-S estimator is a robust reliable estimator.

Let us now turn to the comparison with the grouped fixed effect estimators. Not surprisingly, the infeasible I-GFE estimator is often comparable to the oracle. However, the performance of the feasible GFE estimator varies substantially depending on the number and separation of the groups. For instance, in the upper panel of Table 1 ($G = 2$), where separation of groups is moderate, and for moderate values of T, the GFE estimator is close to the I-GFE estimator, indicating that the minimum might have been achieved. However, in the lower panel of Table 1 ($G = 7$), it is unlikely that 500 initial values cover a significant part of the space of all possible partitions of N units into G groups, and the performance of the GFE estimator deteriorates. In fact, missclassification for the GFE estimator is on the level of 70% for moderate values of T when $G = 7$. In this case, the P-S estimator has a comparative advantage relative to the GFE estimator. This same pattern is reproduced in Table 2. However, since the groups are more separated, and the r-squared of the model is higher, the GFE estimator improves. Interestingly, the behavior of the GFE estimator when $M = 2$ improves relative to $M = 1$.

As far the LS estimator is concerned, when the number of groups increases, the finite sample performance of this estimator deteriorates, although it improves as both N and T increase. However, it is only for large values of T that the behavior is comparable to the oracle. At least in this exercise, the P-S estimator outperforms the LS estimator.

Finally, the pen NN estimator is, in general, dominated by the other estimators for our data-generating processes. However, like with the LS estimator, its performance improves as we increase N and T. For example, in Table 4, when $G = 2$, the estimator is actually comparable to the oracle for large values of N and T.

6. Proofs of Theorems 3.1, 3.2, and 3.3

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

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

where Σ is a positive definite matrix appearing in Assumption 3.2(ii), $S = -2\Sigma\beta$, 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

$$A_{ij} = \frac{1}{NT}\sum_{t=1}^{T}\Big\{f_{it} - f_{jt} + z_{it}'(\beta - b) - z_{jt}'(\beta - b) + v_{it} - v_{jt}\Big\}^2, \quad i, j = 1, \ldots, N,$$

where

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

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

$$R_{ij} = -\frac{2}{NT}\sum_{t=1}^{T}\Big\{(\beta - b)'z_{it}z_{jt}(\beta - b) + v_{it}v_{jt} + v_{it}z_{jt}'(\beta - b) + v_{jt}z_{it}'(\beta - b)\Big\},\ i, j = 1, \ldots, N.$$

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

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

On the other hand,

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

by Assumption 3.2(i). Thus, given that $\mathrm{tr}(A) = 0$, the matrix $A_0 = A - R$ satisfies

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

by Assumption 3.2(ii). In addition, we will show below that A_0 has at most $2GM + 2$ non-zero eigenvalues. Hence,

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

by Lemma A.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 $t = 1, \ldots, T$, denote

$$F_t = (f_{1t}, \ldots, f_{Nt})', \; V_t = (v_{1t}, \ldots, v_{Nt})', \; Z_t = (z_{1t}'(\beta - b), \ldots, z_{Nt}'(\beta - b))'.$$

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

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

Also,

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

Hence, denoting

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

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

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

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

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

Then $F_t = \sum_{m=1}^{M} \sum_{\gamma=1}^{G} \alpha_{\gamma t}^m \varrho_{m\gamma}$ for all $t = 1, \ldots, T$, and so, for any vector $a = (a_1, \ldots, a_N)'$, the vector

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

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

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

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

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

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

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

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

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

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(i) whereas the latter is from Assumption 3.1(ii)). 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 A.2. This gives (28).

To prove (29), we proceed by appropriately modifying the proof of Theorem 4.7.1 in [39]. Denote $\mathcal{V} = (v_1, \ldots, v_N)'$, where $v_i = (v_{i1}, \ldots, v_{iT})'$ for all $i = 1, \ldots, N$. Then uniformly over $u \in \mathcal{S}^T$,

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

by Assumption 3.1(i). Thus,

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

Further, by Exercise 4.4.3, part 2, in [39],

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

where $\mathcal{N}$ is any (1/4)-net in $\mathcal{S}^T$. Moreover, by Corollary 4.2.13 in [39], 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(i) and Lemma 2.7.6 in [39], for all $i = 1, \ldots, N$, we have $\|(u'v_i)^2\|_{\psi_1} \leq C$, and so by Exercise 2.7.10 in [39], $\|(u'v_i)^2 - \mathbb{E}[(u'v_i)^2]\|_{\psi_1} \leq C$. Hence, by Bernstein's inequality (Corollary 2.8.3 in [39]), for any $\epsilon > 0$,

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

and so by the union bound,

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

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

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

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

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

Hence, $\|\mathcal{V}\| = O_P(\sqrt{T} + \sqrt{N})$, 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. ∎

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_γ such that $\alpha_\gamma = \sqrt{T}Fp_\gamma$. 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 $\mathrm{diag}(\lambda_1, \dots, \lambda_G)$ with $\lambda_1 \geq \cdots \geq \lambda_G \geq 0$ 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

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

Therefore, by Assumption 3.4, there exists a constant $\bar{C} > 0$ such that with probability approaching one,

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

and

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

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 λ 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:

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

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 $\mathcal{A} \subset \{1, \dots, N\}$:

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

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 $i, j = 1, \ldots, N$ with $g_i = g_j$, then $m(\lambda) > G$. Therefore, given than $m(\hat{\lambda}) = G$, 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 $i, j = 1, \ldots, N$, 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

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

To do so, note that for all $i = 1, \ldots, N$, the random vectors v_i and $\hat{F}_{h_i}$ are independent conditional on $h_1, \ldots, h_N$. Therefore, by Assumption 3.1(i) and (2.14) in [39], for all $\epsilon > 0$, $i = 1, \ldots, N$, and $\gamma = 1, \ldots, G$,

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

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

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

Combining this bound with Assumption 3.6 gives the asserted claim of this step.

Step 2. Here we show that

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

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

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

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

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

by the same argument as that in Step 1, with Assumption 3.1(ii) playing the role of Assumption 3.1(i). Also,

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

by Assumption 3.3. 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

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

To do so, note that by Assumption 3.1(i) and Lemma 2.7.6 in [39], for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$, we have $\|v_{it}^2\|_{\psi_1} \leq C$, and so by Exercise 2.7.10 in [39], $\|v_{it}^2 - \mathbb{E}[v_{it}^2]\|_{\psi_1} \leq C$. Hence, by Bernstein's inequality (Corollary 2.8.3 in [39]), for any $\epsilon > 0$,

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

and so by the union bound,

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

Hence, given that $\mathbb{E}[v_{it}^2] \leq C$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$ by Assumption 3.1(i), it follows from Assumption 3.6 that

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

Thus, given that

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

can be proven using the same argument, with Assumption 3.1(i) replaced by 3.1(ii), 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

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

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

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

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 A.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,

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

by Theorem 3.1 and Assumption 3.3. 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,

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

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

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

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

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

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

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

Here, $\|\alpha_{g_i}\alpha_{g_i}'\| = \|\alpha_{g_i}\|^2 \leq CT$ for all $i = 1, \ldots, N$, and so applying the expectation version of Bernstein's matrix inequality (Exercise 5.4.11 in [39]) conditionally on $(x_1, y_1), \ldots, (x_N, y_N)$,

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

where the last bound follows from Assumption 3.6. Hence,

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

by Markov's inequality. In addition,

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

by Lemma A.2 and (32) in the proof of Theorem 3.1. Moreover, by Lemma A.2,

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

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 $\gamma = 1, \ldots, G$ 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 $\psi_n \to 0$ as $n \to \infty$ and

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

Also, recall that $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_G)$, where $\lambda_1 \geq \ldots \lambda_G \geq 0$. In addition, set $\lambda_0 = +\infty$ and $\lambda_{G+1} = 0$ and let $\bar{G}$ be the largest integer $\gamma = 0, \ldots, G$ 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 [42]), 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_γ denote the γ-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),

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

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

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \|(\tilde{F}-F)'Fp_{g_i}\|^2 &= \mathrm{tr}\left(\frac{1}{N}\sum_{i=1}^N (\tilde{F}-F)'Fp_{g_i}p_{g_i}'F'(\tilde{F}-F)\right)\\
&= \mathrm{tr}\Big((\tilde{F}-F)'F\Lambda F'(\tilde{F}-F)\Big) = \mathrm{tr}\left((\tilde{F}-F)'\sum_{\gamma_1=1}^G \lambda_{\gamma_1}F_{\gamma_1}F_{\gamma_1}'(\tilde{F}-F)\right)\\
&= \sum_{\gamma_1,\gamma_2=1}^G \lambda_{\gamma_1}\{F_{\gamma_1}'(\tilde{F}_{\gamma_2}-F_{\gamma_2})\}^2 \le \sum_{\gamma_1=1}^{\bar{G}}\sum_{\gamma_2=1}^G \lambda_{\gamma_1}\{F_{\gamma_1}'(\tilde{F}_{\gamma_2}-F_{\gamma_2})\}^2 + G^3\sqrt{\psi_n}\\
&\le \sum_{\gamma_1=1}^{\bar{G}} \lambda_{\gamma_1}\{F_{\gamma_1}'(\tilde{F}_{\gamma_1}-F_{\gamma_1})\}^2 + \sum_{\gamma_1=1}^{\bar{G}}\sum_{\gamma_2\neq\gamma_1} \lambda_{\gamma_1}\{F_{\gamma_1}'\tilde{F}_{\gamma_2}\}^2 + G^3\sqrt{\psi_n} = o_P(1),
\end{aligned}$$

where the middle term is bounded by Lemma A.3 and all λ_γ are bounded by

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

where the last inequality follows from Assumption 3.3(ii). Combining this bound with Assumption 3.5, we conclude that for all $\gamma = 1,\dots,G$,

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

which gives the asserted claim of this step.

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

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

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

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \|(\hat{F}_hO_h-F)p_{g_i}\|^2 &= \mathrm{tr}\left(\frac{1}{N}\sum_{i=1}^N (\hat{F}_hO_h-F)p_{g_i}p_{g_i}'(\hat{F}_hO_h-F)'\right)\\
&= \mathrm{tr}\Big((\hat{F}_hO_h-F)\Lambda(\hat{F}_hO_h-F)'\Big)\\
&= \mathrm{tr}\Big(\Lambda(\hat{F}_hO_h-F)'(\hat{F}_hO_h-F)\Big)
\end{aligned}$$

$$= \sum_{\gamma=1}^{G} \lambda_\gamma \|(\hat{F}_h O_h - F)_\gamma\|^2 = o_P(1)$$

by the same argument as that used in Step 6. Combining this bound with Assumption 3.5, we conclude that for all $\gamma = 1, \ldots, G$,

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

which gives the asserted claim of this step.

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

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

uniformly over $i = 1, \ldots, N$, 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. ∎

Proof of Theorem 3.3. Throughout the proof, for any $\mathrm{a} \in \mathbb{R}^T$ and $R > 0$, we use $B(\mathrm{a}, R)$ to denote the ball in $\mathbb{R}^T$ with center a and radius R, i.e. $B(\mathrm{a}, R) = \{u \in \mathbb{R}^T \colon \|u - \mathrm{a}\| \leq R\}$. 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 [11] but we provide all the details for reader's convenience.

Step 1. Here we show that

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

To prove the first inequality, note that for all $i = 1, \ldots, N$, we have $\mathbb{E}[v_{it}^2] \leq C\|v_{it}\|_{\psi_2}^2 \leq C$ by (2.15) in [39] and Assumption 3.1(i). To prove the second inequality note that for all $i, j = 1, \ldots, N$ with $i \neq j$, we have

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

for all $\epsilon > 0$ by (2.14) in [39] and Assumption 3.1(i). Therefore, given that the function $r \mapsto \exp(-c/r)$ is concave on $\mathbb{R}_+$ for any $c > 0$, we have

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

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

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

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 $\nu = \{\nu_i\}_{i=1}^{N} \in \{1,\ldots,G\}^N$, 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

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

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 $\nu \in \{1,\ldots,G\}^N$. To do so, note that

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

Here, we have

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

where the first inequality follows from Assumption 3.1(i) and (2.15) in [39] and the second from Assumption 3.3(ii). Thus, by Markov's inequality,

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

Further, denoting $\bar{x}_{it} = \sum_{m=1}^{M} \rho_{im} \alpha_{g_i t}^m$, 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 3.3 and 3.8. 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 3.2(i) and 3.8. In addition,

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

where the second line follows from the Cauchy-Schwarz inequality and Assumption 3.8. 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,

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

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

Step 3. Here we show that

$$\bar{Q}(b, a, \nu) - \bar{Q}(\beta, \alpha, g) \geq c_3 \|b - \beta\|^2$$

for all $b \in \mathcal{B}$, $a \in \mathcal{A}_{G,T}$, and $\nu \in \{1, \ldots, G\}^N$ with probability $1 - o(1)$, where $\alpha = \{\alpha_{\gamma t}\}_{\gamma,t=1}^{G,T}$. To do so, note that

$$\bar{Q}(b, a, \nu) - \bar{Q}(\beta, \alpha, g) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it}'(b - \beta) + a_{\nu_i t} - \alpha_{g_i t} \right)^2$$

$$\geq \min_{\tilde{a}\in\mathcal{A}_{G,T}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(x_{it}'(b-\beta)+\tilde{a}_{\nu_i t}-\alpha_{g_i t}\Big)^2$$

$$\geq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big((x_{it}-\bar{x}_{\nu,\nu_i,g_i,t})'(b-\beta)\Big)^2 \geq c_3\|b-\beta\|^2$$

for all $b\in\mathcal{B}$, $a\in\mathcal{A}_{G,T}$, and $\nu\in\{1,\ldots,G\}^N$ with probability $1-o(1)$ by Assumption 3.7. The asserted claim of this step follows.

Step 4. Denote

$$\mathcal{G}_\gamma = \Big\{i=1,\ldots,N\colon \hat{g}_i=\gamma\Big\}, \quad \text{for all } \gamma=1,\ldots,G$$

and

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

In this step, we show that

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

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 $\lambda_N/\sqrt{T}\to 0$ and $\delta_N\to 0$ as $N\to\infty$ and the event

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

holds with probability at least $1-\delta_N$ (note that the proof of (33) did not use Assumption 3.4, 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 $\{\alpha_1,\ldots,\alpha_G\}$ and a constant $R_0>0$ defined by the following algorithm:

Step 1: set $\gamma=1$, $\mathrm{A}_1=\{\alpha_1,\ldots,\alpha_G\}$, and $R_1=\lambda_N$;
Step 2: if

$$\Big\{\cup_{\mathrm{a}\in\mathrm{A}_\gamma}\big(B(\mathrm{a},8R_\gamma+6\lambda_N)\setminus B(\mathrm{a},R_\gamma)\big)\Big\}\cap\mathrm{A}_\gamma=\emptyset, \tag{44}$$

then set $\mathrm{A}_0=\mathrm{A}_\gamma$ and $R_0=R_\gamma$ and stop;
Step 3: replace γ by $\gamma+1$ and set $R_\gamma=9R_{\gamma-1}+6\lambda_N$;
Step 4: let A_γ be the smallest subset of $A_{\gamma-1}$ such that

$$\{\alpha_1,\ldots,\alpha_G\}\subset\cup_{\mathrm{a}\in\mathrm{A}_\gamma}B(\mathrm{a},R_\gamma);$$ [11]

Step 5: go to step 2.

[11] If there are several smallest sets, choose one of them at random.

Observe that on Step 4 of this algorithm, we have $|\mathrm{A}_\gamma| \leq |\mathrm{A}_{\gamma-1}| - 1$. Indeed, if Step 4 is performed for given γ, it follows from Step 2 that there exist $\mathrm{a}_1, \mathrm{a}_2 \in \mathrm{A}_{\gamma-1}$ such that $\mathrm{a}_2 \in B(\mathrm{a}_1, 8R_{\gamma-1} + 6\lambda_N)$, and so $B(\mathrm{a}_2, R_{\gamma-1}) \subset B(\mathrm{a}_1, 9R_{\gamma-1} + 6\lambda_N) = B(\mathrm{a}_1, R_\gamma)$, letting us drop a_2 from $\mathrm{A}_{\gamma-1}$ while constructing A_γ and yielding $|\mathrm{A}_\gamma| \leq |\mathrm{A}_{\gamma-1}| - 1$. In turn, the latter implies that the algorithm will stop in a finite number of steps and, in fact, 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,

$$\{\alpha_1, \ldots, \alpha_G\} \subset \cup_{\mathrm{a} \in \mathrm{A}_0} B(\mathrm{a}, R_0) \tag{45}$$

by construction.

Further, since (44) is satisfied with $\gamma = 0$ by construction, it follows that

$$\mathrm{a}_1, \mathrm{a}_2 \in \mathrm{A}_0, \text{ implies that } \|\mathrm{a}_2 - \mathrm{a}_1\| \leq R_0 \text{ or } \|\mathrm{a}_2 - \mathrm{a}_1\| > 8R_0 + 6\lambda_N. \tag{46}$$

This allows us to partition A_0 into equivalence subclasses as follows. Say that $\mathrm{a}_1 \sim \mathrm{a}_2$ if $\|\mathrm{a}_2 - \mathrm{a}_1\| \leq R_0$. Then for any $\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3 \in \mathrm{A}_0$, we have that $\mathrm{a}_1 \sim \mathrm{a}_2$ and $\mathrm{a}_2 \sim \mathrm{a}_3$ imply $\mathrm{a}_1 \sim \mathrm{a}_3$, meaning that the relation $\sim$ is actually an equivalence relation. Thus, we can partition A_0 into $k \leq G$ equivalence subclasses $\mathrm{A}_{0,1}, \ldots, \mathrm{A}_{0,k}$ such that for any $\mathrm{a}_1, \mathrm{a}_2 \in \mathrm{A}_0$ we have $\mathrm{a}_1 \sim \mathrm{a}_2$ if and only if a_1 and a_2 belong to the same subclass. Then it follows from (43), (45), and (46) that for each $i = 1, \ldots, N$, there exists a unique $\gamma(i) \in \{1, \ldots, k\}$ such that $\|\hat{A}_i - \mathrm{a}\| \leq R_0 + \lambda_N$ for some $\mathrm{a} \in \mathrm{A}_{0,\gamma(i)}$. Thus, for any $i_1, i_2 = 1, \ldots, N$, we have that $\|\hat{A}_{i_1} - \hat{A}_{i_2}\| \leq 3R_0 + 2\lambda_N$ if $\gamma(i_1) = \gamma(i_2)$ and that $\|\hat{A}_{i_1} - \hat{A}_{i_2}\| > 6R_0 + 4\lambda_N$ if $\gamma(i_1) \neq \gamma(i_2)$. In turn, the latter implies that if we run the Classification Algorithm from Section 2 with $\lambda = 3R_0 + 2\lambda_N$, we obtain $m(\lambda) = k$ groups $\mathcal{A}_1, \ldots, \mathcal{A}_k$ such that any two units $i_1, i_2 = 1, \ldots, N$ are classified to the same group if and only if $\gamma(i_1) = \gamma(i_2)$. To see why this is so, suppose that for some $\gamma = 1, \ldots, k$, we have two units i_1, i_2 that are classified to the same group $\mathcal{A}_\gamma$ but are such that $\gamma(i_1) \neq \gamma(i_2)$. For this γ, let $i_1 \to i_2 \to \cdots \to i_{|\mathcal{A}_\gamma|}$ be the order in which units are added to $\mathcal{A}_\gamma$ by the Classification Algorithm, and let r be the smallest number in the set $\{2, \ldots, |\mathcal{A}_\gamma|\}$ such that $\gamma(i_r) \neq \gamma(i_{r-1})$. Then $\|\hat{A}_{i_r} - \hat{A}_{i_1}\| > 6R_0 + 4\lambda_N$ and $\|\hat{A}_{i_l} - \hat{A}_{i_1}\| \leq 3R_0 + 2\lambda_N$ for all $l = 1, \ldots, r-1$. This implies that

$$\left\| \hat{A}_{i_r} - \frac{1}{r-1} \sum_{l=1}^{r-1} \hat{A}_{i_l} \right\| > (6R_0 + 4\lambda_N) - (3R_0 + 2\lambda_N) = 3R_0 + 2\lambda_N,$$

yielding a contradiction. Now suppose that there are two units i_1, i_2 that are classified to different groups $\mathcal{A}_{\gamma_1}$ and $\mathcal{A}_{\gamma_2}$ but are such that $\gamma(i_1) = \gamma(i_2)$. For these γ_1 and γ_2, let $i_1 \to i_2 \to \cdots \to i_{|\mathcal{A}_{\gamma_1}| + |\mathcal{A}_{\gamma_2}|}$ be the order in which units are added to $\mathcal{A}_{\gamma_1}$ and $\mathcal{A}_{\gamma_2}$

by the Classification Algorithm. Assume, without loss of generality, that the first unit, i_1, is added to $\mathcal{A}_{\gamma_1}$, and let r be the smallest number in the set $\{2, \ldots, |\mathcal{A}_{\gamma_1}| + |\mathcal{A}_{\gamma_2}|\}$ such that the unit i_r is added to the set $\mathcal{A}_{\gamma_2}$. Then $\|\hat{A}_{i_r} - \hat{A}_{i_l}\| \leq 3R_0 + 2\lambda_N$ for all $l = 1, \ldots, r-1$, and so

$$\left\| \hat{A}_{i_r} - \frac{1}{r-1} \sum_{l=1}^{r-1} \hat{A}_{i_l} \right\| \leq 3R_0 + 2\lambda_N,$$

yielding a contradiction, and finishing the proof of our claim that the Classification Algorithm from Section 2 with $\lambda = 3R_0 + 2\lambda_N$ yields $m(\lambda) = k$ groups $\mathcal{A}_1, \ldots, \mathcal{A}_k$ such that any two units $i_1, i_2 = 1, \ldots, N$ are classified to the same group if and only if $\gamma(i_1) = \gamma(i_2)$. The latter than implies that $m(3R_0 + 2\lambda_N) = k \leq G$, and so $\hat{\lambda} \leq 3R_0 + 2\lambda_N \leq C_G \lambda_N$ for some $C_G > 0$ depending only on G, as desired.

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 $\gamma = 1, \ldots, G$ and any $i_1, i_2 \in \mathcal{G}_\gamma$, we have

$$\|\alpha_{g_{i_2}} - \alpha_{g_{i_1}}\| \leq \tilde{C}_G \lambda_N. \tag{47}$$

To see why this is so, suppose again that (43) is satisfied. As we have proven above, it is then not possible that the Classification Algorithm with $\lambda = \hat{\lambda} \leq 3R_0 + 2\lambda_N$ generates groups $\mathcal{A}_1, \ldots, \mathcal{A}_{m(\lambda)}$ such that there are two units i_1, i_2 the are classified to the same group $\mathcal{A}_\gamma$ but are such that $\gamma(i_1) \neq \gamma(i_2)$. In turn, for any i_1, i_2 such that $\gamma(i_1) = \gamma(i_2)$, we have $\|\hat{A}_{i_2} - \hat{A}_{i_1}\| \leq 3R_0 + 2\lambda_N$, and so $\|\alpha_{g_{i_2}} - \alpha_{g_{i_1}}\| \leq 3R_0 + 4\lambda_N \leq \tilde{C}_G \lambda_N$ for some $\tilde{C}_G > 0$, as desired.

We are now ready to finish this step. On (43), by the triangle inequality and (47), we have

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

uniformly over $i = 1, \ldots, N$. 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. ∎

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

[7] Armstrong, T., Weidner, M., and Zeleneev, A. (2022). Robust estimation and inference in panels with interactive fixed effects. *Working paper*.

[8] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191-221.

[9] Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229-1279.

[10] Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing unobserved heterogeneity. *Working paper*.

[11] Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* **83** 1147-1184.

[12] Chen, H., Leng, X., and Wang, W. (2019). Latent group structures with heterogeneous distributions: identification and estimation. *Working paper*.

[13] Cheng, X., Schorfheide, F., and Shao, P. (2019). Clustering for multi-dimensional heterogeneity. *Working paper*.

[14] Chudik, A., Pesaran, H., and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal*, **14** C45-C90.

[15] Connor, G. and Korajzcyk, R. (1986). Performance measurement with the arbitrage pricing theory: a new framework for analysis. *Journal of Financial Economics* **15** 373-394.

[16] Connor, G. and Korajzcyk, R. (1988). Risk and return in an equilibrium APT: application to a new test methodology. *Journal of Financial Economics* **21** 255-289.

[17] Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. *Electronic Journal of Probability*.

[18] 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.

[19] Gao, Z. and Shi, Z. (2020). Implementing convex optimization in R: two econometric examples. *Computational Economics*.

[20] Gu, J. and Volgushev, S. (2019). Panel data quantile regression with grouped fixed effects. *Journal of Econometrics* **213** 68-91.

[21] Hahn, J. and Moon, R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory* **26** 863-881.

[22] Halko, N., Martinsson, P., and Tropp, J. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* **53** 217-288.

[23] Han, F. and Li, Y. (2020). Moment bounds for large autocovariance matrices under dependence. *Journal of Theoretical Probability* **33** 1445-1492.

[24] Hansen, C. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics* **141** 597-620.

[25] Holtz-Eakin, D., Newey, W., and Rosen, H. (1988). Estimating vector autoregressions with panel data. *Econometrica* **56** 1371-1395.

[26] 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.

[27] Löffler, M., Zhang, A., and Zhou, H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *Annals of Statistics* **49** 2506-2530.

[28] Moitra, A. (2018). Algorithmic aspects of machine learning. *Cambridge University Press*.

[29] Moon, R. and Weidner, M. (2019). Nuclear norm regularized estimation of panel regression models. *Working paper*.

[30] Nicholls, P. (1988). Factors influencing entry of pesticides into soil water. *Pesticide Science* **22** 123-137.

[31] Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168** 244-258.

[32] Pesaran, H. (2006). Estimation and inference in large heterogenous panels with a multifactor error structure. *Econometrica* **74** 967-1012.

[33] Stock, J. and Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics* **44** 293-335.

[34] Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* **206** 554-573.

[35] Su, L., Shi, Z., and Phillips, P. (2016). Identifying latent structures in panel data. *Econometrica* **84** 2215-2264.

[36] Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society. Series B* **58** 267-288.

[37] Van der Geer, S. (2002). On Hoeffding's inequality for dependent random variables. *In Empirical Process Techniques for Dependent Data*.

[38] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68** 841-860.

[39] Vershynin, R. (2018). High-dimensional probability: an introduction with applications in data science. *Cambridge Series in Statistical and Probabilistic Mathematics*.

[40] Wang, W., Phillips, P., and Su, L. (2018). Homogeneity pursuit in panel data models: theory and application. *Journal of Applied Econometrics* **33** 797-815.

[41] 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.

[42] Yu, Y., Wang, T., and Samworth, R. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrica* **102** 315-323.

[43] Zhang, Y., Wang, H., and Zhu, Z. (2017). Quantile-regression-based clustering for panel data. *Working paper*.

Supplementary Materials for "Spectral and Post-Spectral Estimators for Grouped Panel Data Models" by D. Chetverikov and E. Manresa

Appendix A. Technical Lemmas

Lemma A.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 $\lambda_1^B, \ldots, \lambda_K^B$. Further, let $\lambda_1^A, \ldots, \lambda_K^A$ 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 A.1. This lemma does not follow from Weyl's inequality immediately because $\lambda_1^A, \ldots, \lambda_K^A$ are the largest *in absolute value* eigenvalues of A. ∎

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

$$\lambda_{K+1}^B = \ldots \lambda_N^B = 0, \tag{48}$$

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

$$\min(|\lambda_1^A|, \ldots, |\lambda_K^A|) \geq \max(|\lambda_{K+1}^A|, \ldots, |\lambda_N^A|). \tag{49}$$

By Weyl's inequality ([39], Theorem 4.5.3), one can construct a one-to-one function $f \colon \{1, \ldots, N\} \to \{1, \ldots, N\}$ such that

$$|\lambda_k^A - \lambda_{f(k)}^B| \leq \|A - B\|, \quad \text{for all } k = 1, \ldots, N. \tag{50}$$

Using this function, define $\mathcal{K} = \{j = 1 \ldots, K \colon f^{-1}(j) > K\}$ and note that if this set is non-empty, then there exists $k = 1, \ldots, K$ 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 (50) and the triangle inequality, the second from (49), the third from (50) and the triangle inequality, and the fourth from (48). 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 (50). Combining the last two inequalities gives the asserted claim. ∎

Lemma A.2. *Let $\mu_1, \ldots, \mu_N$ and $\nu_1, \ldots, \nu_N$ 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 $\|x\| = 1$,

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

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. ∎

Lemma A.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),

$$F_1'F_3 = 1 - \frac{\|F_3 - F_1\|^2}{2}. \tag{51}$$

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

$$\|F_3 - F_1\|^2 \geq \Big(\|F_3 - F_2\| - \|F_2 - F_1\|\Big)^2 \geq 2 - 2\sqrt{2}\|F_2 - F_1\|. \tag{52}$$

Combining (51) and (52),

$$F_1'F_3 \leq 1 - 1 + \sqrt{2}\|F_2 - F_1\| = \sqrt{2}\|F_2 - F_1\|. \tag{53}$$

Also, again by the triangle inequality,

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

and so, by (i),

$$\begin{aligned}\|F_3 - F_1\|^2 &\leq \Big(\|F_3 - F_2\| + \|F_2 - F_1\|\Big)^2 \\ &\leq 2 + 2\sqrt{2}\|F_2 - F_1\| + \sqrt{2}\|F_2 - F_1\| = 2 + 3\sqrt{2}\|F_2 - F_1\|. \qquad (54)\end{aligned}$$

Combining (51) and (54), $F_1'F_3 \geq -3\|F_2 - F_1\|$. Combining this inequality with (53) gives the asserted claim. ∎

Lemma A.4. *Let $\theta \in (-1, 1)$ be a real number and $T \geq 2$ be an integer. Then for all $u = (u_1, \ldots, u_T)' \in \mathbb{R}^T$ 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 $T = 2, \ldots, k$. We will prove that the claim holds for $T = k+1$. To do so, fix any $u = (u_1, \ldots, u_T)' \in \mathbb{R}^T$ 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 $T\geq 2$. ∎

Appendix B. Proofs for Remaining Results from Main Text

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

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

by Slutsky's lemma and Assumption 3.9. The asserted claim follows. ∎

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 $\|\mathring{\rho}_{im}\|\vee|\mathring{\alpha}_{\gamma t}^m|\leq C$ for all $i=1,\ldots,N$, $\gamma=1,\ldots,G$, $t=1,\ldots,T$, and $m=1,\ldots,2M$, which would correspond to Assumption 3.3 in the context of dynamic model. To do so, observe that

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

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

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

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. ∎

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 $b \in \overline{\mathcal{B}} = \{0_d\} \cup \{e_k \colon k = 1, \ldots, d\} \cup \{e_k + e_l \colon k, l = 1, \ldots, d\}$. 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

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

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

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

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

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

uniformly over $b \in \overline{\mathcal{B}}$. 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

$$\left\| \frac{1}{NT} \sum_{t=1}^{T} Z_{bt} Z_{bt}' \right\| = O_P\left(\frac{1}{T \wedge N} \right) \tag{55}$$

uniformly over $b \in \overline{\mathcal{B}}$, where we denoted $Z_{bt} = (z_{1t}'(\beta - b), \ldots, z_{Nt}'(\beta - b))'$, 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 $u = (u_1, \ldots, u_T)' \in \mathcal{S}^T$ and $b \in \overline{\mathcal{B}}$ by our assumptions, where we denoted $z_i = (z_{i1}, \ldots, z_{iT})'$ for all $i = 1, \ldots, N$. Thus, denoting $\mathcal{Z}_b = (z_1(\beta - b), \ldots, z_N(\beta - b))'$, we have

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

for all $b \in \overline{\mathcal{B}}$, 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,

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

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

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

because $\log d = o(N)$. We thus obtain (55) 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

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

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

$$(\hat{\beta}_\lambda - \beta)'\hat{\Sigma}(\hat{\beta}_\lambda - \beta) \leq (\lambda/c_\lambda)\|\hat{\beta}_\lambda - \beta\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}_\lambda\|_1. \tag{56}$$

Further, denote $\delta = \hat{\beta}_\lambda - \beta$, $\mathcal{T} = \{k = 1, \ldots, d\colon \beta_k \neq 0\}$, and $\mathcal{T}^c = \{1, \ldots, d\} \setminus \mathcal{T}$. Also, let $\delta_\mathcal{T} = (\delta_{\mathcal{T}1}, \ldots, \delta_{\mathcal{T}d})'$ be a $d \times 1$ vector such that $\delta_{\mathcal{T}k} = \delta_k 1\{k \in \mathcal{T}\}$ and, similarly, let $\delta_{\mathcal{T}^c} = (\delta_{\mathcal{T}^c 1}, \ldots, \delta_{\mathcal{T}^c d})'$ be a $d \times 1$ vector such that $\delta_{\mathcal{T}^c k} = \delta_k 1\{k \in \mathcal{T}^c\}$. Then, given that $(\hat{\beta}_\lambda - \beta)'\hat{\Sigma}(\hat{\beta}_\lambda - \beta) \geq 0$, it follows from (56) 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 $k = 1, \ldots, d$. 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 (56), 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 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. ∎

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

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

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 $F_t = (f_{1t}, \ldots, f_{Nt})'$ for all $t = 1, \ldots, T$, 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 $\mathcal{K} = (\kappa_1 + \omega_1(\beta - b), \ldots, \kappa_N + \omega_N(\beta - b))'$ 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. ∎

Appendix C. 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 [22], 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, $\lambda_{(1)}, \ldots, \lambda_{(k)}$, ordered so that $|\lambda_{(1)}| \geq \cdots \geq |\lambda_{(k)}|$. 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 $q = [\log N]$;
Step 3: draw an $N \times (k + p)$ random matrix $\Omega = \{\Omega_{ij}\}_{i,j=1}^{N,k+p} \overset{iid}{\sim} N(0, 1)$;
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 $\tilde{s}_1, \ldots, \tilde{s}_{k+p}$ of the $(k + p) \times (k + p)$ matrix BB';
Step 8: compute $N \times 1$ vectors $s_j = B'\tilde{s}_j$ for $j = 1, \ldots, k + p$;
Step 9: compute $\hat{\lambda}_j = \text{sign}(s_j'As_j)(\|B'Bs_j\|/\|s_j\|)^{1/2}$ for $j = 1, \ldots, k + p$;
Step 10: order values $\hat{\lambda}_1, \ldots, \hat{\lambda}_{k+p}$ into $\hat{\lambda}_{(1)}, \ldots, \hat{\lambda}_{(k+p)}$ so that $|\hat{\lambda}_{(1)}| \geq \cdots \geq |\hat{\lambda}_{(k+p)}|$;
Step 11: return $\hat{\lambda}_{(1)}, \ldots, \hat{\lambda}_{(k)}$.

The result of this algorithm is k values $\hat{\lambda}_{(1)}, \ldots, \hat{\lambda}_{(k)}$. These are estimators of k largest in absolute value eigenvalues $\lambda_{(1)}, \ldots, \lambda_{(k)}$ of the matrix A. As follows from results in [22], these estimators are consistent as $N \to \infty$ under conditions to be discussed below even though they are based on a realization of the random matrix Ω.

Specifically, Corollary 10.10 in [22] 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,

$$\|A - QQ'A\| = O_P(|\lambda_{(k+1)}|). \tag{57}$$

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 $\hat{\lambda}^2_{(1)}, \ldots, \hat{\lambda}^2_{(k)}$ 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 $\hat{\lambda}^2_{(1)}, \ldots, \hat{\lambda}^2_{(k)}$ are consistent estimators of $\lambda^2_{(1)}, \ldots, \lambda^2_{(k)}$ as long as $\|A\| = O_P(1)$ and $|\lambda_{(k+1)}| = o_P(1)$ as $N \to \infty$. Hence, $\hat{\lambda}_{(j)} \to \lambda_{(j)}$ for all $j = 1, \ldots, k$ 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 and 3.2 by the proof of Theorem 3.1.

Appendix D. 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 [11], 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 $i = 1, \ldots, N$. 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(i) and Lemma A.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 3.1(ii) and Lemma A.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,

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

for some constant $C > 0$ by (16), the triangle inequality, Assumption 3.1, the fact that $\|y_{i0}\|_{\psi_2} \leq C_y$ for all $i = 1, \dots, N$, and (2.15) in [39]. 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 auto-covariance 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.

Appendix E. Motivating Example

To motivate equation (3), we provide a specific example in the context of agricultural production and environmental economics. Suppose a production process has an

input that is potentially polluting, x, such as a pesticide. In addition, suppose we want to assess the incidence of the use of this substance in the environment's pollution, such as water or soil. See, for example, [30]. The incidence of polluting substances on the environment depends on the intensity of use, but also on the characteristics of the environment, such as soil permeability, rainfall, etc.

Suppose we have the following model to quantify the effect of the use of input x on some measure of pollution:

$$y_{it} = x_{it}\beta + \alpha_{g_i t} + \varepsilon_{it}$$

where y_{it} is a measure of pollution, x_{it} is the quantity of pesticide, $\alpha_{g_i t}$ are unobserved characteristics of the region, g_i, such as soil permeability, rain incidence, and other characteristics that influence the presence of chemicals in the environment, and ε is an unobserved shock. The farmer chooses to purchase pesticide in the local market as to maximize expected profits:

$$x_{it} = \arg\max_x \mathbb{E}\left[\pi_i(x, \alpha_{g_i t}, z_{it}, \nu_{it}) | \alpha_{g_i 1}, \ldots \alpha_{g_i T}, z_{i1}, \ldots, z_{iT}\right] - C(x, p^x_{g_i t})/p_{g_i t} \quad (58)$$

where $\pi_i(x, \alpha_{gt}, z_{it}, \varepsilon_{it})$ is the production function of farmer i, and $C(x, p^x_{g_i t})$ is the cost associated to purchase x when the price of the input is $p^x_{g_i t}$, and $p_{g_i t}$ is the price of the produced good in the local market of farmer i. The production process depends on the use of observable inputs, such as x, but also soil and rain characteristics specific to the region, that affect both the prevalence of pesticides in the environment but also the agricultural process, and z_{it}, a variable that affects production, such as quality of inputs, observed to the farmer but not to the econometrician, which does not affect the environment. Finally ν is an idiosyncratic shock in the production process, unforeseeable by both the farmer and the econometrician.

Assume $\pi_i(x, \alpha_{g_i t}, z_{it}, \nu_{it}) = \tilde{\pi}_i(x, \alpha_{g_i t}, z_{it}) + \nu_{it}$, and $\mathbb{E}\left[\nu_{it} | \alpha_{g_i 1}, \ldots \alpha_{g_i T}, z_{i1}, \ldots, z_{iT}\right] = 0$. In addition, let $\tilde{\pi}'_i(x, \alpha_{g_i t}, z_{it}) = \beta_i x + \alpha_{g_i t} + z_{it}$, and $C'(x, p^x_{g_i t}) = p^x_{g_i t}$ be the partial derivative with respect to the first argument, respectively for $\tilde{\pi}_i$ and C. The F.O.C. in (58) implies that x_{it} is defined as:

$$\beta_i x_{it} + \alpha_{gt} + z_{it} - \frac{p^x_{gt}}{p_{gt}} = 0,$$

which implies: $x_{it} = \frac{1}{\beta_i}\left(\frac{p^x_{gt}}{p_{gt}} - \alpha_{gt}\right) - \frac{1}{\beta_i} z_{it}$, which takes the form of equation (3) with $M = 2$.

TABLE 1. Mean Absolute Error (MAE) and misclassification results: $\sigma^2 = 1$, $M = 1$

$G = 2$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.035	0.018	0.059	0.151	0.016	0.034	0.014	0.009	0.058
	200	0.024	0.009	0.019	0.150	0.009	0.009	0.008	0.002	0.003
	400	0.016	0.007	0.010	0.151	0.007	0.008	0.006	0.003	0.006
50	100	0.024	0.007	0.012	0.149	0.007	0.007	0.007	0.000	0.000
	200	0.015	0.006	0.008	0.150	0.006	0.006	0.006	0.000	0.000
	400	0.010	0.004	0.006	0.149	0.004	0.004	0.004	0.000	0.000
100	100	0.014	0.006	0.010	0.153	0.006	0.006	0.006	0.000	0.000
	200	0.008	0.004	0.006	0.151	0.004	0.004	0.004	0.000	0.000
	400	0.004	0.002	0.004	0.152	0.002	0.002	0.002	0.000	0.000

$G = 7$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.116	0.109	0.144	0.151	0.023	0.144	0.011	0.346	0.705
	200	0.057	0.085	0.149	0.152	0.029	0.149	0.010	0.234	0.744
	400	0.025	0.062	0.142	0.148	0.055	0.147	0.006	0.139	0.764
50	100	0.034	0.014	0.142	0.147	0.007	0.147	0.007	0.006	0.651
	200	0.015	0.008	0.143	0.148	0.006	0.137	0.006	0.001	0.645
	400	0.009	0.004	0.062	0.150	0.004	0.137	0.004	0.000	0.622
100	100	0.022	0.006	0.148	0.151	0.006	0.110	0.006	0.000	0.364
	200	0.009	0.003	0.083	0.152	0.003	0.112	0.003	0.000	0.289
	400	0.006	0.003	0.006	0.150	0.003	0.078	0.003	0.000	0.203

TABLE 2. Mean Absolute Error (MAE) and misclassification results: $\sigma^2 = 4$, $M = 1$

$G = 2$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.026	0.005	0.018	0.153	0.005	0.005	0.005	0.000	0.000
	200	0.015	0.003	0.014	0.159	0.003	0.003	0.003	0.000	0.000
	400	0.009	0.002	0.009	0.157	0.002	0.002	0.002	0.000	0.000
50	100	0.015	0.003	0.014	0.156	0.003	0.003	0.003	0.000	0.000
	200	0.009	0.002	0.009	0.157	0.002	0.002	0.002	0.000	0.000
	400	0.007	0.001	0.007	0.144	0.001	0.001	0.001	0.000	0.000
100	100	0.010	0.002	0.007	0.140	0.002	0.002	0.002	0.000	0.000
	200	0.007	0.001	0.006	0.063	0.001	0.001	0.001	0.000	0.000
	400	0.004	0.001	0.005	0.011	0.001	0.001	0.001	0.000	0.000

$G = 7$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.055	0.005	0.099	0.155	0.004	0.055	0.004	0.000	0.154
	200	0.029	0.003	0.035	0.158	0.003	0.073	0.003	0.000	0.216
	400	0.016	0.002	0.013	0.158	0.002	0.060	0.002	0.000	0.195
50	100	0.026	0.003	0.014	0.157	0.003	0.049	0.003	0.000	0.135
	200	0.012	0.002	0.009	0.157	0.002	0.048	0.002	0.000	0.138
	400	0.007	0.002	0.006	0.157	0.002	0.071	0.002	0.000	0.219
100	100	0.023	0.002	0.011	0.156	0.002	0.048	0.002	0.000	0.122
	200	0.009	0.001	0.007	0.161	0.001	0.062	0.001	0.000	0.165
	400	0.005	0.001	0.004	0.158	0.001	0.036	0.001	0.000	0.101

TABLE 3. Mean Absolute Error (MAE) and misclassification results: $\sigma^2 = 1$, $M = 2$

$G = 2$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.048	0.019	0.016	0.171	0.015	0.016	0.014	0.015	0.004
	200	0.028	0.009	0.011	0.169	0.008	0.008	0.008	0.004	0.005
	400	0.019	0.007	0.008	0.169	0.007	0.007	0.006	0.003	0.003
50	100	0.022	0.008	0.010	0.168	0.008	0.008	0.008	0.000	0.000
	200	0.016	0.007	0.007	0.168	0.007	0.007	0.007	0.000	0.000
	400	0.009	0.003	0.006	0.168	0.003	0.003	0.003	0.000	0.000
100	100	0.013	0.005	0.008	0.168	0.005	0.005	0.005	0.000	0.000
	200	0.011	0.004	0.005	0.169	0.004	0.004	0.004	0.000	0.000
	400	0.006	0.003	0.004	0.167	0.003	0.003	0.003	0.000	0.000

$G = 7$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.104	0.069	0.146	0.167	0.015	0.166	0.012	0.169	0.639
	200	0.075	0.057	0.108	0.169	0.012	0.159	0.009	0.144	0.641
	400	0.038	0.046	0.040	0.170	0.014	0.153	0.006	0.103	0.608
50	100	0.050	0.013	0.094	0.169	0.009	0.081	0.009	0.004	0.204
	200	0.024	0.006	0.011	0.168	0.005	0.014	0.005	0.002	0.019
	400	0.014	0.004	0.006	0.167	0.004	0.004	0.004	0.000	0.000
100	100	0.025	0.006	0.011	0.166	0.006	0.017	0.006	0.000	0.019
	200	0.012	0.004	0.006	0.168	0.004	0.004	0.004	0.000	0.000
	400	0.007	0.003	0.003	0.168	0.003	0.003	0.003	0.000	0.000

TABLE 4. Mean Absolute Error (MAE) and misclassification results: $\sigma^2 = 4$, $M = 2$

$G = 2$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.036	0.005	0.006	0.167	0.005	0.005	0.005	0.000	0.000
	200	0.017	0.003	0.004	0.169	0.003	0.003	0.003	0.000	0.000
	400	0.012	0.003	0.004	0.167	0.003	0.003	0.003	0.000	0.000
50	100	0.016	0.003	0.004	0.122	0.003	0.003	0.003	0.000	0.000
	200	0.012	0.002	0.003	0.087	0.002	0.002	0.002	0.000	0.000
	400	0.006	0.001	0.002	0.038	0.001	0.001	0.001	0.000	0.000
100	100	0.009	0.002	0.003	0.030	0.002	0.002	0.002	0.000	0.000
	200	0.007	0.002	0.002	0.009	0.002	0.002	0.002	0.000	0.000
	400	0.005	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000

$G = 7$

T	N	Mean Absolute Error							Misclassification	
		S	P-S	LS	Pen NN	I-GFE	GFE	Oracle	S	GFE
20	100	0.103	0.006	0.007	0.164	0.004	0.004	0.004	0.003	0.000
	200	0.063	0.003	0.005	0.169	0.003	0.003	0.003	0.000	0.000
	400	0.028	0.002	0.004	0.169	0.002	0.002	0.002	0.000	0.000
50	100	0.036	0.003	0.005	0.171	0.003	0.003	0.003	0.000	0.000
	200	0.015	0.002	0.003	0.169	0.002	0.002	0.002	0.000	0.000
	400	0.011	0.001	0.002	0.168	0.001	0.001	0.001	0.000	0.000
100	100	0.021	0.002	0.003	0.173	0.002	0.002	0.002	0.000	0.000
	200	0.010	0.001	0.002	0.167	0.001	0.001	0.001	0.000	0.000
	400	0.005	0.001	0.001	0.166	0.001	0.001	0.001	0.000	0.000

(D. Chetverikov) Department of Economics, UCLA, Bunche Hall, 8283, 315 Portola Plaza, Los Angeles, CA 90095, USA.

E-mail address: chetverikov@econ.ucla.edu

(E. Manresa) Department of Economics, New York University, 19 West 4th Street, New York, NY 10003, USA.

E-mail address: em1849@nyu.edu