University of Connecticut

Department of Economics Working Paper Series

High-Dimensional Weighted *K*-Means with Serial Dependence

by

Zhonghui Zhang
Nanjing Audit University

Chihwa Kao
University of Connecticut

Jungbin Hwang
University of Connecticut

Working Paper 2025-09
August 2025

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

This working paper is indexed in RePEc, http://repec.org

High-Dimensional Weighted K-Means with Serial Dependence*

Zhonghui Zhang
Joint Research Institute,
Nanjing Audit University

Chihwa Kao and Jungbin Hwang
Department of Economics,
University of Connecticut

August, 2025

Abstract

In this paper, we propose a new K-means approach for high-dimensional panel data with unknown group memberships. We highlight that the standard K-means algorithm using Euclidean distance can suffer from misclassification in finite samples due to serial correlation and heteroskedasticity in the panel data. Our proposed weighted K-means algorithm addresses this issue by weighting the Euclidean distance using the full covariance structure of idiosyncratic shocks. Assuming that both the cross-sectional and time dimensions of the panel grow large, we develop an asymptotic theory for the weighted K-means algorithm that establishes the consistency of the estimated group centroids and the oracle property for group membership estimation. For practical implementation, we propose a feasible weighted K-means method that employs a regularized estimation of the high-dimensional covariance matrix in the K-means objective function. Monte Carlo simulation results demonstrate the effectiveness of our weighted K-means algorithm in estimating grouped fixed-effects models for large panels, particularly when strong serial dependencies exist in both group-level trends and idiosyncratic components.

JEL Classification: C13, C23, C38, C63;

Keywords: Banding, Grouped fixed effects, Heteroscedasticity and autocorrelation, K-means clustering, Sample covariance matrix

*Email: zhonghui@nau.edu.cn, chih-hwa.kao@uconn.edu, jungbin.hwang@uconn.edu. We would like to thank the seminar participants at the University of Connecticut and the University of Iowa for their helpful comments. We are also grateful to Liangjun Su for his valuable feedback and suggestions on this paper.

1 Introduction

K-means (Forgy, 1965; Lloyd, 1982) is one of the most widely used unsupervised machine learning algorithms for cluster analysis in data mining. The method partitions observed data vectors into a finite number of subsets, called clusters, where each observation is assigned to the cluster with the nearest mean (or cluster centroid). Due to its simple structure and ease of implementation, K-means has also been applied to large panel data to identify and estimate unknown group structures (Bonhomme and Manresa, 2015; hereafter BM (2015)). The main contribution of this paper is to propose a new high-dimensional K-means method for the large panel data that accounts for serial dependencies and heteroskedasticity over time.

When applying K-means to panel data with n cross-sectional units, the dimension of each panel unit corresponds to the length of the time span T. Assuming T is finite, the classical asymptotic results of Pollard (1981, 1982) show that the estimated K-means centroids are consistent with the (pseudo) true population centroids and are asymptotically normal at the standard $\sqrt{n}$-rate. However, for high-dimensional K-means, which is the focus of this paper, the parameter dimension of the cluster centroids T expands, and these classical results cannot be directly applied.

More recently, BM (2015) developed an asymptotic framework for the panel group fixed-effects model using the K-means approach. In this framework, the estimated time-varying GFE parameters correspond to K-means centroid vectors stacked over time. By imposing well-separated conditions between the true group centroids, BM (2015) establish an asymptotic theory that accounts for both n and T growing to infinity. Their results show that the estimated K-means centroids remain consistent for the true cluster centroids. Notably, their framework also establishes oracle estimation of group (or cluster) memberships, which was not established in Pollard's (1981, 1982) fixed-T framework.

A key feature of the K-means algorithm is its reliance on distance-based calculations. In the standard K-means approach, as in BM (2015), the Euclidean metric is used to measure the distance between two panel units by summing the equally weighted squared differences. However, the Euclidean distance metric can be less informative when attempting to separate the true latent group structure from noise in the panel data, particularly when the noise exhibits temporal dependencies and heteroskedasticity over time. Consequently, the standard K-means approach may be prone to misclassifying true group memberships in finite samples, especially when T is not sufficiently large and when serial correlation and heteroskedasticity are pronounced.

To address this issue, this paper proposes the weighted K-means approach that employs a new distance measure accounting for serial correlation and heteroskedasticity in the idiosyncratic shock process in panel data. Our Mahalanobis-type distance measure incorporates weights based on the entire covariance structure of the time series of idiosyncratic shocks in the panel data, where the inverse of the covariance matrix is used to rescale the Euclidean distance in the weighted K-means objective, e.g., Gnanadesikan et al. (1995) and Kao et al. (2021). As a result, the weighted K-means objective effectively captures the true signal of the latent group structure by filtering out idiosyncratic noise caused by

temporal dependencies and heteroskedasticity, and is expected to improve the precision of latent group membership recovery in finite samples.

Because the weighted K-means objective relies on the inverse of a covariance matrix whose dimension T grow to infinity, incorporating such weights complicates asymptotic analysis due to the growing number of parameters and potential instability in high dimensions. To address this point, we follow the literature on high-dimensional statistics, e.g., Bickel and Levina (2008a&b) and impose well-conditioned covariance matrices. In a stationary time series, this condition applies to a Toeplitz matrix with a spectral density bounded from below and above. By leveraging the well-conditioned covariance structure of the idiosyncratic components in K-means, we show that our weighted K-means approach preserves the desirable asymptotic properties of the high-dimensional K-means method in BM (2015), such as the consistency of cluster centroids and the oracle estimation of cluster (group) memberships.

For the feasible implementation of weighted K-means, estimating the high-dimensional covariance matrix and its inverse is required. One may consider a naive approach by directly estimating the inverse of the sample covariance matrix of the idiosyncratic noise in K-means. However, the covariance estimation in high-dimensional settings poses significant challenges, often referred to as the curse of dimensionality, e.g,. Bai and Silverstein (2010) and Bai and Shi (2011). In particular, for time series data, Basak et al. (2014) show that the limiting behavior of the high-dimensional sample autocovariance matrix does not coincide with that of its population counterpart. As a result, using an unregularized sample covariance matrix leads to poor finite-sample performance, particularly in terms of misclassification under our weighted K-means objective.

To tackle the issue of the high-dimensional weight matrix estimation, our feasible weighted K-means approach adopts the tapered/banded covariance estimation methods proposed by Bickel and Levina (2008) and Xiao and Wu (2012). The concepts of banding and tapering in time series analysis have also been applied in nonparametric spectral density estimation in econometrics, e.g. Newey and West (1987) and Andrews (1991). The use of positive definite kernel functions in our banded/tapered covariance estimation ensures that the inverse of the regularized covariance matrix remains positive definite, which is crucial for formulating the weighted K-means objective function. Assuming that both the cross-sectional dimension and time dimension in panel data grow to infinity, we develop asymptotic theory that establishes the consistency of the regularized high-dimensional weight matrix in spectral norm. This consistency result allows us to control the high-dimensional estimation noise inherent in the weighted K-means procedure, ensuring asymptotic equivalence to the infeasible weighted K-means.

The practical implementation of the feasible weighted K-means is analogous to classical generalized least squares (GLS) regression. First, we run an initial unweighted K-means to estimate the inverse covariance matrix and use it to rescale the original dataset. Using this transformed dataset, the weighted K-means can then be implemented by simply running the standard unweighted K-means algorithm. This makes the weighted K-means procedure practically convenient, as we can use readily available standard K-means packages in popular statistical software such as `R` and `MATLAB`. We also propose an iterated version of

the weighted K-means, analogous to the iterated GMM estimator of Hansen et al. (1996). Starting from an initial weighted K-means, the iterated algorithms update the weight matrix in the K-means objective at each iteration using residuals from the latest membership and centroid estimates, and then reapply the algorithm. This process is repeated until membership assignments stabilize or a fixed number of iterations is reached.

Our Monte Carlo simulation results confirm the improvement in group membership estimation using the weighted K-means approach. Specifically, we find that the weighted K-means algorithm with a regularized inverse covariance matrix estimator substantially improves classification accuracy when the panel data exhibit temporal dependencies. In contrast, the standard unweighted K-means suffers from high misclassification rates in finite samples under temporal dependence. Also, the use of an unregularized inverse covariance estimator in the weighted K-means objective offers little to no improvement, with results largely mirroring those of the unweighted version. We also consider a data-generating process that incorporates a common slope parameter, as in the grouped panel fixed effects model of BM (2015), and show that incorporating the weighted K-means objective into the group fixed effects estimation can substantially improve the finite-sample performance of the slope parameter estimator.

This paper contributes to the growing literature on grouped panel data, including Hahn and Moon (2010), Lin and Ng (2012), BM (2015), Su et al. (2016), Vogt and Linton (2017), and Chetverikov and Manresa (2022), which incorporate group structures into individual-level data observed over time. Among these, BM (2015) is the first to establish a foundational asymptotic framework for the high-dimensional K-means approach to latent group analysis in panel data. The grouped fixed-effects panel data model in BM (2015) is closely related to the interactive fixed-effects model in Bai (2009), which extends the common approximate factor structure of Bai and Ng (2002) and Bai (2003) to accommodate grouped structures in high-dimensional panel data, e.g., Bian and Su (2025). Beyond K-means, alternative clustering methods such as Classifier-Lasso (C-Lasso), developed by Su et al. (2016) and extended by Huang et al. (2021), further enrich the toolkit for identifying latent group structures. See also Ando and Bai (2016), Okui and Wang (2022), Lumsdaine et al. (2023), Wang and Su (2021), and Su et al. (2023) for recent advances in grouped panel data models.

Compared with prior methods for estimating unknown group memberships, our weighted K-means approach is the first to integrate a high-dimensional weighting scheme into the group estimation objective, improving the accuracy of group membership estimation and reducing misclassification in finite samples. The key idea is to exploit the entire covariance structure of the residuals for parameter estimation, an approach analogous to the classical generalized least squares (GLS) method in linear regression. Our analysis shows that this GLS-type principle can be effectively extended to the context of group membership estimation in large panel data.

The remainder of the paper is structured as follows. Section 2 presents the formulation of the weighted K-means approach for large panel data and establishes the asymptotic theory for the infeasible weighted K-means procedure. Section 3 develops a feasible weighted K-means method that employs a regularized estimation of the high-dimensional covariance

matrix in the objective function. Section 4 provides practical guidance on implementing the weighted K-means algorithm, including a discussion on numerical methods for obtaining state-of-the-art solutions. Monte Carlo simulation results and empirical illustrations are presented in Sections 5 and Section 6, respectively, and the conclusion is provided in Section 7. Proofs, formula derivations, tables, and figures are provided in the Appendix 8.

2 Asymptotics for High-dimensional Weighted K-means

We consider the following clustered (or grouped) panel model of Bonhomme and Manresa (2015, BM hereafter):

$$y_{it} = x_{it}'\theta_0 + \alpha^0_{g_0(i),t} + v_{it} \text{ for } i \in \{1,\dots,n\} \text{ and } t \in \{1,\dots,T\}, \tag{1}$$

and $E[v_{it}] = 0$. For simplicity in presenting the idea of our weighted K-means approach, we assume that θ^0 is known and set $\theta_0 = 0$ in (1). The extension to the case with an unknown common parameter θ_0 is discussed in subsection 5.2.

The model can be estimated by the classical K-means clustering method (Forgy, 1965; Lloyd, 1982), which jointly estimates the unknown group membership mapping function $g_0(\cdot)$ as well as the true cluster centroid parameter vector $\alpha^0_g = (\alpha^0_{g1},\dots,\alpha^0_{gT})'$. Specifically, let $y_i = (y_{i1},\dots,y_{iT})'$ and $\alpha_{g(i)} = (\alpha_{g(i),1},\dots,\alpha_{g(i),T})'$ be T-dimensional vectors. Then, the objective function for the weighted K-means is defined as:

$$\min_{(\alpha,g(\cdot))\in\mathcal{A}^{GT}\times\Gamma_G} \sum_{i=1}^{n} (y_i - \alpha_{g(i)})' W_T (y_i - \alpha_{g(i)}), \tag{2}$$

where the objective function minimizes the aggregated squared differences between the observed data vectors y_i and their corresponding cluster centroids $\alpha_{g(i)}$, weighted by a $T\times T$ matrix W_T. The minimization is taken jointly over the sets $\mathcal{A}^{GT}$ and Γ_G, where Γ_G denotes the set of all possible group membership mappings $g(\cdot) : \{1,\dots,n\} \to \{1,\dots,G\}$, and $\mathcal{A}^{GT}$ represents the set of all possible values $\alpha = (\alpha_{gt})$ for $g \in \{1,\dots,G\}$ and $t \in \{1,\dots,T\}$.

When the weight matrix is $W_T = I_T$, the objective function in (2) coincides with that of the standard K-means algorithm, which employs the squared Euclidean distance $||y_i - \alpha_{g(i)}||^2 = (y_i - \alpha_{g(i)})'(y_i - \alpha_{g(i)})$. More generally, the weighted K-means algorithm considered in this paper employs a high-dimensional weight matrix $W_T = \Sigma_T^{-1}$, where $\Sigma_T = var(v_i)$ denotes the $T \times T$ population covariance matrix of $v_i = (v_{i1},\dots,v_{iT})'$. The (t,s)-th element of Σ_T is given by

$$\Sigma_T(t,s) = E[v_{it}v_{is}] \text{ for } t,s \in \{1,\dots,T\},$$

which gives rise to the weighted Euclidean distance defined as $||y_i - \alpha_{g(i)}||^2_w := (y_i - \alpha_{g(i)})'\Sigma_T^{-1}(y_i - \alpha_{g(i)})$ in (2). This so-called Mahalanobis distance (Mahalanobis, 1936) can be interpreted as the squared Euclidean distance between linearly transformed variables, $\breve{y}_i = \Sigma_T^{-1/2} y_i$ and $\tau_{g(i)} = \Sigma_T^{-1/2}\alpha_{g(i)}$. Employing a Mahalanobis-type distance metric in clustering analysis can be particularly advantageous when variables differ significantly in

scale or exhibit strong intercorrelations, e.g., Gnanadesikan et. al (1995) and Kao et al. (2021). In our high-dimensional K-means setting with panel data, the alternative distance $||\cdot||_w$ captures the full covariance structure of the idiosyncratic shocks $\{v_{it}\}_t$, thereby effectively isolating the true signal of the latent group structure by filtering out idiosyncratic noise caused by temporal dependence and heteroskedasticity.

We denote $(\tilde{\alpha}_w, \tilde{g}_w(\cdot))$ as the solution to the weighted K-means problem, where $\tilde{\alpha}_w = (\tilde{\alpha}_1^w, \ldots, \tilde{\alpha}_G^w)'$ represents the estimated group fixed effects, where each group effect $\tilde{\alpha}_g^w = (\tilde{\alpha}_{g1}^w, \ldots, \tilde{\alpha}_{gT}^w)'$ corresponds to group $g \in \{1, \ldots, G\}$. For given values of $\alpha \in \mathcal{A}^{GT}$, let $\tilde{g}_w(i;\alpha)$ denote the optimal group assignment for each individual unit i, defined as

$$\tilde{g}_w(i;\alpha) = \underset{g\in\{1,\ldots,G\}}{\arg\min} (y_i - \alpha_{g(i)})'\Sigma_T^{-1}(y_i - \alpha_{g(i)}), \tag{3}$$

where the smallest g is chosen in the case of a non-unique solution. Then, the weighted K-means problem in (1) can be reformulated as

$$\tilde{\alpha}_w = \underset{\alpha\in\mathcal{A}^{GT}}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha_{\tilde{g}_w(i;\alpha)})'\Sigma_T^{-1}(y_i - \alpha_{\tilde{g}_w(i;\alpha)}) \tag{4}$$

with the first-order condition for $\tilde{\alpha}_g^w$ given by

$$\sum_{i=1}^{n} 1(\tilde{g}_w(i) = g)(y_i - \hat{\alpha}_{\tilde{g}_w(i)}^w) = 0 \text{ for each } g \in \{1, \ldots, G\}. \tag{5}$$

The estimator for the cluster membership is defined as $\tilde{g}_w(i) = \tilde{g}_w(i;\tilde{\alpha}_w)$. Under this formulation, each $\tilde{\alpha}_g^w$ is the simple averaged values of $\{y_i\}_{i=1}^n$ across individuals assigned to group g by the weighted K-means solution $\tilde{g}_w(\cdot)$. Thus, our weighted K-means objective function with the common weight matrix Σ_T^{-1} primarily serves to enhance the estimation of the unknown group membership function $g_0(\cdot)$, which is expected to lead to subsequent improvements in the estimation of the cluster centroid parameters $\tilde{\alpha}_g^w$ as well.

Our asymptotic analysis of the weighted K-means focuses on the case where both n and T grow to infinity. As T increases, the parameter space of cluster centroids $\mathcal{A}^{GT}$ expands, and thus the classical fixed-T asymptotic techniques of Pollard (1981, 1982) cannot be directly applied to establish the asymptotic properties of $\tilde{\alpha}_w$ and $\tilde{g}_w(\cdot)$. With $W_T = I_T$, BM (2015) establishes conditions under which the high-dimensional clustering estimator in K-means is consistent. In our setting, which involves a non-identity weight matrix, $W_T = \Sigma_T^{-1}$, the growing dimensionality can potentially introduce instability into the objective function and further complicate the asymptotic analysis. To address these issues, we impose the following assumptions.

Assumption 1 *There exists $\epsilon > 0$ that does not depend on T such that $0 < \epsilon \leq \lambda_{\min}(\Sigma_T) \leq \lambda_{\max}(\Sigma_T) \leq 1/\epsilon < \infty$ for all T.*

Assumption 2 *Assume the following conditions hold: (a) $\mathcal{A}^{GT} := \mathcal{A} \times \cdots \times \mathcal{A} = \Pi_{i=1}^{GT}\mathcal{A}$, where $\mathcal{A}$ is a compact subset of $\mathbb{R}$. (b) $\{v_i\}$ with $v_i = (v_{i1}, \ldots, v_{iT})'$ is i.i.d over $i \in \{1, \ldots, n\}$ such that $E[v_{it}] = 0$ and $E[v_{it}^4] \leq M$ for some constant $M > 0$.*
(c) $|(n^2T)^{-1}\sum_{i,j=1}^{n}\sum_{t,s=1}^{T} Cov(v_{it}v_{jt}, v_{is}v_{js})| \leq M$.

Assumption 3 *Assume the following conditions hold: (a) For all $g \in \{1, \ldots, G\}$, we have that*

$$\underset{n\to\infty}{p\lim}\frac{1}{n}\sum_{i=1}^{n} 1(g_0(i) = g) = \pi_g > 0.$$

(b) For all $(g, h) \in \{1, \ldots, G\}^2$ such that $g \neq h$, we have that

$$\underset{T\to\infty}{p\lim}\frac{(\alpha_g^0 - \alpha_h^0)'\Sigma_T^{-1}(\alpha_g^0 - \alpha_h^0)}{T} = c_{gh} > 0.$$

Assumption 1 can be referred to the well-conditioned covariance matrix assumption commonly found in the high-dimensional statistics literature, e.g., Bickel and Levina (2008b). When the time series process $\{v_{it}\}_{t=1}$ is (covariance) stationary with autocovariance function $\gamma(k) = E[v_{i0}v_{ik}]$ for $k \in \{0, \pm 1, \pm 2, \ldots\}$, the matrix Σ_T is a Toeplitz matrix whose (t, s)-th element is given by $\Sigma_T(t, s) = \gamma(|t - s|)$. Due to the Toeplitz structure of Σ_T, its eigenvalues are connected to the spectral density of $\{v_{it}\}_t$, which is defined as

$$f(\omega) = \frac{1}{2\pi}\sum_{k\in\mathbb{Z}} \gamma(k)\cos(k\omega).$$

By construction, $\gamma(k) = \int_{-\pi}^{\pi} f(\omega)\exp(-ik\omega)d\omega$, and results from Grenander and Szegö (1958) can be applied to establish the following bound on the eigenvalues of Σ_T :

$$2\pi \min_{\omega\in[-\pi,\pi]} f(\omega) \leq \lambda_{\min}(\Sigma_T) \leq \lambda_{\max}(\Sigma_T) \leq 2\pi \max_{\omega\in[-\pi,\pi]} f(\omega).$$

From this result, Assumption 1 is satisfied when the spectral density $f(\omega)$ is bounded both below and above. In other words, controlling the dependence structure of the time series by bounding its spectral density ensures that the influence of the high dimensional matrix, Σ_T^{-1}, remains stable, thereby preventing overfitting and promoting convergence to the true clustering structure.

Assumptions 2-(a) and (b) impose regularity on the grouped structure, requiring a compact parameter set and finite moments. The cross-sectional independence of v_i in Assumption 2-(b) is imposed for technical convenience in establishing the results for the feasible weighted K-means in the next section. This assumption can be relaxed to allow for weak cross-sectional dependence, as in BM (2015). Assumption 2-(c) also restrict the amount of weak dependency and interactions over time and individuals. They mirror conditions imposed in BM (2015).

Assumption 3-(a) and (b) are key requirements for identifying cluster memberships, particularly when the G population groups are large and well separated. In particular, Assumption 3-(b) requires that the weighted sum of squared differences between the time effects of any two groups be bounded away from zero. This condition ensures that the weighted distance metric remains sufficiently informative to measure group differences, thereby stabilizing the cluster boundaries. Given the well-conditioned covariance matrix specified in Assumption 1-(a), it can be shown that Assumption 3-(b) holds if and only if

$$\underset{T\to\infty}{p\lim}\frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt}^0 - \alpha_{ht}^0)^2 = \tilde{c}_{gh} > 0, \tag{6}$$

for some $\tilde{c}_{gh}$, which is equivalent to a condition imposed in BM (2015). In this sense, Assumptions 1 and 3 ensures that no two clusters can become arbitrarily close due to temporal correlation, thereby avoiding the violation of the identification condition (6) for the unweighted K-means.

Remark 1 *While our weighted K-means objective assumes a common weighting matrix Σ_T^{-1}– reflecting the assumption that the covariance structure of each v_i is identical across different i – it is possible to extend the framework to allow for heterogeneous covariances. Specifically, we can consider a case where each i has its own covariance structure, denoted by $\Sigma_{i,T}(t,s) = E[v_{it}v_{is}]$. The corresponding inverse matrix $\Sigma_{i,T}^{-1}$ can be used in the weighted K-means objective. In this setting, the assumption on well-conditioned covariance matrices (Assumption 1) needs to be modified to ensure that for all i and T, the eigenvalues of $\Sigma_{i,T}$ are uniformly bounded both below and above. Also, the FOC condition for $\tilde{\alpha}_g^w$, given $\tilde{g}_w(i)$, is now:*

$$\sum_{\{i:\tilde{g}_w(i)=g\}} \Sigma_{i,T}^{-1}(y_i - \hat{\alpha}_g^w) = 0.$$

Solving this yields a weighted least squares-type estimator $\hat{\alpha}_g^w = \sum_{\{i:\tilde{g}_w(i)=g\}} \Psi_{i,T}^{(g)} y_i$ with $\Psi_{i,T}^{(g)} = (\sum_{\{i:\tilde{g}_w(i)=g\}} \Sigma_{i,T}^{-1})^{-1}\Sigma_{i,T}^{-1}$. In this case, implementing the feasible weighted K-means requires estimating n different high-dimensional weight matrices $\Sigma_{i,T}^{-1}$, which presents greater computational and theoretical challenges than in the current common-weight case.

Remark 2 *As noted in BM (2015), the K-means objective function in (2) are invariant to relabeling of memberships, implying that cluster membership $\tilde{g}_w(\cdot)$ is identified only up to a permutation of the true membership function $g_0(\cdot)$. Specifically, define the relabeling mapping $\sigma(\cdot) : \{1,\ldots,G\} \to \{1,\ldots,G\}$ as*

$$\sigma(g) = \underset{\breve{g}\in\{1,\ldots,G\}}{\arg\min} \frac{1}{T}(\tilde{\alpha}_{\breve{g}}^w - \alpha_g^0)'\Sigma_T^{-1}(\tilde{\alpha}_{\breve{g}}^w - \alpha_g^0). \tag{7}$$

Theorem 1 shows that $\sigma(\cdot)$ is invertible with probability approaching one, so we can always relabel the estimated cluster $\tilde{\alpha}_{\sigma(g)}^w$ to target the true cluster centroids α_g^0.

Theorem 1 *Let Assumptions 1–3 hold, and denote $\tilde{\alpha}_{\tilde{g}_w(i),t}^w$ as the estimated cluster centroids for $\alpha_{g_0(i),t}^0$ using the weighted K-means objective in (2). Then, as n and T tend to infinity, we have that*

$$\frac{1}{nT}\sum_{i=1}^{n}(\tilde{\alpha}_{\tilde{g}_w(i)}^w - \alpha_{g_0(i)}^0)'\Sigma_T^{-1}(\tilde{\alpha}_{\tilde{g}_w(i)}^w - \alpha_{g_0(i)}^0) = o_p(1). \tag{8}$$

Also, the relabeling mapping $\sigma(\cdot)$ defined in (7) is one-to-one with probability tending to one, and we have that

$$\frac{1}{T}(\tilde{\alpha}_{\sigma(g)}^w - \alpha_g^0)'\Sigma_T^{-1}(\tilde{\alpha}_{\sigma(g)}^w - \alpha_g^0) = o_p(1) \text{ for all } g \in \{1,\ldots,G\}. \tag{9}$$

The first result of (8) in Theorem 1 establishes the consistency of the estimated cluster centroids in terms of the weighted quadratic mean. The next result in (9) shows that each estimated cluster centroid vector, with growing dimension of T, converges to its true counterpart, after properly relabeling the group indices and scaling the weighted Euclidean norm by T^{-1}. Letting $\Sigma_T = I_T$, $\tilde{\alpha}^w_{\tilde{g}_w(i)}$ corresponds to the standard K-means with Euclidean distance. Since the identity matrix satisfies Assumption 1 trivially, the result in Theorem 1 can be applied, reducing to the result in BM (2015, Theorem 1 and (B.3)). Theorem 1 generalizes these results to the weighted K-means setting—provided that the weight matrix satisfies the well-separated covariance condition in Assumption 1.

While Theorem 1 establishes the consistency of the estimated cluster centroids, a natural question arises regarding the asymptotic properties of the group (or cluster) membership assignments $\tilde{g}_w(\cdot)$. Specifically, the oracle group membership property requires that the misclassification probability

$$P\left(\tilde{g}_w(i) \neq \sigma(g_0(i)) \text{ for some } i \in \{1, \ldots, n\}\right) = o(1), \tag{10}$$

as $n, T \to \infty$. In the proof of Theorem 2, we show that the above misclassification probability can be bounded (up to a constant) by the following expression:

$$n \cdot \left(\sum_{\tilde{g} \neq g} P\left(\frac{(\alpha^0_g - \alpha^0_{\tilde{g}})' \Sigma_T^{-1} (\alpha^0_g - \alpha^0_{\tilde{g}})}{T} \leq c_1 \right) + \sum_{\tilde{g} \neq g} P\left(\frac{v_i' \Sigma_T^{-1} v_i}{T} \geq c_2 \right) \right. \tag{11}$$

$$\left. + \sum_{\tilde{g} \neq g} P\left(\left| \frac{(\alpha^0_{\tilde{g}} - \alpha^0_g)' \Sigma_T^{-1} v_i}{T} \right| \geq c_3 \right) \right) + o(1) \tag{12}$$

for some positive constants c_1, c_2, and c_3 that do not depend on T. To analyze these probabilities, we define the rescaled components:

$$\tau^0_g := L_T \alpha^0_g = (\tau^0_{g1}, \ldots, \tau^0_{gT})' \text{ and } u_i := L_T v_i = (u_{i1}, \ldots, u_{iT})',$$

where L_T denotes the Cholesky factor of Σ_T^{-1}. Bounding the probabilities of misclassification requires controlling the behavior of the transformed processes, $\{\tau^0_g\}_t$ and $\{u_{it}\}_t$, as well as their product $\{(\tau^0_{\tilde{g}t} - \tau^0_{gt})u_{it}\}_t$. We formalize these conditions below.

Assumption 4 *Assume the following conditions hold for all $g, \tilde{g} \in \{1, \ldots, G\}$ such that $g \neq \tilde{g}$, and for all $i \in \{1, \ldots, n\}$: (a) The processes $\{\tau^0_{gt}\}_t$, $\{u_{it}\}_t$, and $\{(\tau^0_{\tilde{g}t} - \tau^0_{gt})u_{it}\}_t$ are strongly mixing with mixing coefficients $\alpha[t]$ satisfying $\alpha[t] \leq \exp(-at^{d_1})$ for some $a > 0$ and $d_1 > 0$. Moreover, $E[(\tau^0_{\tilde{g}t} - \tau^0_{gt})u_{it}] = 0$. (b) For all $Z_{it} \in \{\tau^0_{gt}, u_{it}, (\tau^0_{\tilde{g}t} - \tau^0_{gt})u_{it}\}$, there exist constants $b > 0$ and $d_2 > 0$ such that for all i, t, and $m > 0$, $P(|Z_{it}| > m) \leq \exp(1 - (m/b)^{d_2})$.*

Assumptions 4 (a) and (b) impose exponentially fast-decaying time series dependence, measured by the strong mixing coefficient, as well as tail conditions on the weighted processes. When $\Sigma_T = I_T$, these conditions align with Assumptions 2-(c) and 2-(d) in

BM (2015), which pertain to the original processes $\{\alpha_{gt}^0\}_t$, $\{v_{it}\}_t$, and $\{(\alpha_{\tilde{g}t}^0 - \alpha_{gt}^0)v_{it}\}_t$. These assumptions allow to apply exponential inequalities to dependent processes, e.g., BM (2015, Lemma B.5) and Merlevède et al. (2011), and bound the first two terms in (11) with $\Sigma_T = I_T$. In our weighted K-means setting, the use of non-identity, high-dimensional weight matrices Σ_T^{-1} introduces additional complexity in controlling the tail behavior of the weighted sum in (12). Specifically, consider the cross-term in that expression, where the transformed components are given by

$$\tau_{gt}^0 = \sum_{s=1}^{T} L_T(t,s)\alpha_{gs}^0 \text{ and } u_{it} = \sum_{s=1}^{T} L_T(t,s)v_{is}, \tag{13}$$

and $L_T(t,s)$ is the (t,s)-th element of L_T. These expressions show that the transformed processes are linear combinations of all time points in the original series, respectively. As a result, maintaining the strong mixing properties and tail bounds for the transformed process becomes essential for controlling the misclassification probability in (10) for the weighted K-means procedure. The conditions in Assumption 4 ensure that the probability bound terms in (11) shrink to zero at an arbitrarily fast rate as T increases, even when Σ_T is non-identity.

Theorem 2 *Let Assumptions 1–4 hold. Then, as n and T tend to infinity, and for all $\delta > 0$, we have*

$$P\left(\tilde{g}_w(i) \neq \sigma(g_0(i)) \text{ for some } i \in \{1,\ldots,n\}\right) = o(1) + o\left(\frac{n}{T^\delta}\right). \tag{14}$$

Theorem 2 shows that the estimated group membership obtained via the weighted K-means converges to the oracle group memberships as n and T tend to infinity at arbitrary relative rates. For each g and t, let $\breve{\alpha}_{gt}$ denote the infeasible cluster centroid estimate, obtained by replacing the estimated cluster membership $\tilde{g}_w(i;\alpha)$ with the true membership $g_0(i)$. Then, it is straightforward to verify–based on the result of Theorem 2– that the infeasible $\breve{\alpha}_{gt}$ is asymptotically equivalent to the weighted K-means cluster centroid estimate $\tilde{\alpha}^w_{\sigma(g),t}$.

As discussed, the key conditions for establishing the oracle group membership property are the fast-decaying dependence and tail conditions in Assumption 4 are applied to the weighted transformed processes defined in (13). When the original processes $\{\alpha_{gt}^0\}_t$ and $\{v_{it}\}$ satisfy conditions in Assumption 4 with $\Sigma_T = I_T$, one way to ensure that the same conditions also hold for the transformed processes is to impose additional structure on Σ_T. For example, suppose that v_{it} follows a finite-order autoregressive (AR) process of order p :

$$v_{it} = \sum_{\ell=1}^{p} \phi_\ell v_{it-\ell} + u_{it} \text{ for } t \in \{1,\ldots,T\}, \tag{15}$$

where u_{it} is serially uncorrelated, zero-mean time series with $var(u_{it}) = \sigma_u^2$. Without loss of generality, we set the initial values $v_{i0},\ldots,v_{i,-p+1}$ to zeros and $\sigma_u^2 = 1$. An important result from high-dimensional covariance analysis in time series, see, for example, Pourahmadi

(1999, 2013), shows that under the AR(p) structure, the inverse covariance matrix Σ_T^{-1} becomes a p-band limited matrix. That is, only the main diagonal and the p off-diagonals on each side contain nonzero entries. Specifically, we have that $\Sigma_T^{-1} = L_T' L_T$, where the Cholesky factor L_T is lower triangular and given by:

$$L_T = \begin{pmatrix} A & 0 & 0 & \cdots & 0 \\ -\psi' & 1 & 0 & \cdots & 0 \\ 0 & -\psi' & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -\psi' & 1 \end{pmatrix} \text{ with } A = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -\phi_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\phi_{p-1} & -\phi_{p-2} & \cdots & -\phi_1 & 1 \end{pmatrix},$$

and $\psi = (\phi_p, \phi_{p-1}, \ldots \phi_1)'$. The sparsity structure in L_T (and Σ_T^{-1}) helps control the behavior of the transformed processes in (13). In particular, it allows the transformed cluster centroid process to be written as $\tau_{gt}^0 = \alpha_{gt}^0 - \sum_{\ell=1}^p \phi_\ell \alpha_{g,t-\ell}^0$, which is a finite transformation of the original cluster centroid process $\{\alpha_{gt}^0\}_t$. As a result, the strong mixing property imposed on $\{\alpha_{gt}^0\}_t$ can be preserved in the transformed process $\{\tau_{gt}^0\}_t$. By similar reasoning, the strong mixing property of $\{v_{it}\}_t$ can also be naturally preserved in its decorated process $\{u_{it}\}_t$ under the AR(p) specification. The asymptotic validity of weighted K-means, in terms of misclassification rates under the AR specification for v_{it}, is numerically confirmed in our Monte Carlo simulation in subsection 5.1.

3 Feasible Weighted K-means

Building on the asymptotic theory developed for weighted K-means in the previous section, we propose a feasible weighted K-means method that employs a regularized estimation of the high-dimensional covariance matrix appearing in the objective function. Let $\hat{\alpha}$ and $\hat{g}(\cdot)$ denote the initial K-means estimates in (2), which are obtained using Euclidean distance with $W_T = I_T$. To implement the feasible version of the weighted K-means in (2), a natural approach is to consider a naive plug-in estimator of Σ_T^{-1}, given by the inverse of the sample covariance matrix, $\hat{\Sigma}_T$, where

$$\hat{\Sigma}_T = [\hat{\sigma}_{st}]_{1 \le s,t \le T} = \frac{1}{n} \sum_{i=1}^n \hat{v}_i \hat{v}_i' \text{ with } \hat{v}_i = y_i - \hat{\alpha}_{\hat{g}(i)}. \tag{16}$$

However, estimating the high-dimensional covariance matrix $\hat{\Sigma}_T$ and its inverse presents significant challenges, commonly referred to as the "curse of dimensionality" (e.g., Bai and Silverstein, 2010; Bai and Shi, 2011). In particular, the large number of parameters in the $T \times T$ covariance matrix $\hat{\Sigma}$, makes it difficult to obtain accurate estimates with limited data (Bickel and Levina, 2008a&b). The limitations of using an unregularized $\hat{\Sigma}_T$ in the feasible weighted K-means method are confirmed by our numerical experiments in Section 5.

To address this challenge, we employ the following version of the tapered/banded covariance estimator as in Bickel and Levina (2008b) and Xiao and Wu (2012) :

$$\hat{\Sigma}_{B_T} = \hat{\Sigma}_T \star \mathcal{K}_T = \left[K\left(\frac{s-t}{B_T}\right) \hat{\sigma}_{st} \right]_{1 \le s,t \le T}, \tag{17}$$

where $[\mathcal{K}_T]_{1 \le s,t \le T} = K((s-t)/B_T)$. The kernel function $K(\cdot)$ is symmetric and positive-definite with $K(0) = 1$, $|K(x)| \le 1$ for all $x \in \mathbb{R}$. Additionally, it satisfies the following condition for truncation outside the unit interval, i.e.,

$$K(x) = 0 \text{ for } |x| > 1. \tag{18}$$

The notation $\star$ denotes Schur (or Hadamard) product, which is the element-wise multiplication of matrices. Also, B_T in (17) is the bandwidth (or smoothing) parameter of the truncation lag in $\hat{\Sigma}_{B_T}$, which satisfies the standard rate condition $B_T \to \infty$ such that $B_T/T \to 0$. The positive-definite kernel function $K(\cdot)$ guarantees that $\mathcal{K}_T$ is positive definite as well. Then, by the Schur product theorem in matrix theory, e.g., Horn and Johnson (2012), since $\hat{\Sigma}$ is nonnegative definite, the Schur product $\hat{\Sigma}_{B_T} = \hat{\Sigma} \star \mathcal{K}_T$ is also nonnegative. Some possible choices for $K(\cdot)$ include:

$$\text{Bartlett: } K_{BT}(x) = \begin{cases} 1 - |x|, \text{ for } |x| \le 1 \\ 0, \text{ otherwise} \end{cases}; \tag{19}$$

$$\text{Parzen: } K_P(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 \text{ for } 0 \le |x| \le 1/2 \\ 2(1-|x|)^3 \text{ for } 1/2 \le |x| \le 1 \\ 0 \text{ otherwise} \end{cases}; \tag{20}$$

$$\text{Tukey-Hanning: } K_{TH}(x) = \begin{cases} (1 + \cos(\pi x))/2 \text{ for } |x| \le 1 \\ 0, \text{ otherwise} \end{cases}, \tag{21}$$

These kernel functions are commonly used in heteroskedasticity and autocorrelation consistent/robust (HAC/HAR) inference, e.g., Andrews (1991), Phillips (2005), and Sun (2014), which satisfy that $K((s-t)/B_T)\hat{\sigma}_{st} = 0$ for $|s-t| > B_T$. The banding effect on $\hat{\Sigma}_{B_T}$ plays a convenient role in proving the desired convergence rate for the high-dimensional covariance matrix estimate $\hat{\Sigma}_{B_T}$ in Lemma 3. The positive-definite property and condition (18) of $K(\cdot)$ guarantee that $\hat{\Sigma}_{B_T}$ is a B_T-banded positive definite matrix, and that its inverse $\hat{\Sigma}_{B_T}^{-1}$ is also positive definite. However, $\hat{\Sigma}_{B_T}^{-1}$ is generally not a banded matrix. Nevertheless, we show below that the well-conditioned covariance structure of Σ_T ensures the desirable asymptotic properties of both the inverse $\hat{\Sigma}_{B_T}^{-1}$ and the regularized estimator $\hat{\Sigma}_{B_T}$ itself.

With (17), we formulate the estimation of the feasible weighted K-means clustering as below:

$$(\hat{\alpha}_w, \hat{g}_w(\cdot)) = \underset{(\alpha, g(\cdot)) \in \mathcal{A}^{GT} \times \Gamma_G}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha_{g(i)})' \hat{\Sigma}_{B_T}^{-1} (y_i - \alpha_{g(i)}). \tag{22}$$

To establish the asymptotic properties of $(\hat{\alpha}_w, \hat{g}_w(\cdot))$, we first present the asymptotic properties of the banded covariance estimator $\hat{\Sigma}_{B_T}$, along with its convergence rates. For any symmetric real matrix A, let $\lambda(\cdot)$ denote l_2/l_2 operator norm or spectral radius defined as

$$\lambda(A) := \max_{|x|=1} |Ax|,$$

where x is a real vector and $|\cdot|$ indicates Euclidean norms. Then, $\lambda(A)$ is the maximal absolute eigenvalue of A. Also, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues (or singular values) of a matrix A, respectively. Then, it is well-known that $\lambda_{\max}^{1/2}(A'A)$ is equal to the operator norm $\lambda(A)$, and $\lambda(A) := \lambda_{\max}(A)$ when A is a real positive (semi) definite matrix. The following conditions are required to establish the asymptotic properties of the feasible weighted K-means in (22).

Assumption 5 *(a) $K(\cdot)$ is symmetric, positive-definite, piecewise smooth with $K(0) = 1$ such that (18) holds, with its Parzen characteristic exponent defined by*

$$q = \max\left\{q_0 : q_0 \in \mathbb{Z}^+, k_{q_0} = \lim_{x\to 0}\frac{1-K(x)}{|x|^{q_0}} < \infty\right\}$$

is greater than or equal to one. (b) $\sup_t \sum_{j=-\infty}^{\infty} |j|^q ||E[v_{it}v_{it-j}]|| \leq M$ for some constant $M > 0$.

Assumption 6 *The conditions in Assumptions 1–4 hold, with Σ_T is replaced with I_T.*

Assumption 5-(a) specifies standard kernel conditions in HAC/HAR approaches in time series, which are satisfied by the aforementioned kernel function examples $q = 1$ for (19) and $q = 2$ for (20) and (21). Assumption 5-(b) imposes restrictions on the decaying dependence structure of $E[v_{it}v_{it-j}]$ as j increases, which is equivalent to imposing smoothness on the spectral density of the time series $\{v_{it}\}_t$.

Assumption 6 imposes the conditions for the initial (unweighted) K-means with $W_T = I_T$ in (2), which are required to justify using the estimated residuals $\hat{v}_i$ in the inverse covariance estimator $\hat{\Sigma}_{B_T}^{-1}$ for the weighted K-means. Specifically, it requires that the conditions in Assumption 4 hold when the processes $\{\tau_{gt}^0\}_t$, $\{u_{it}\}_t$, and $\{(\tau_{\tilde{g}t}^0 - \tau_{gt}^0)u_{it}\}_t$ are replaced by the original processes $\{\alpha_{gt}^0\}_t$, $\{v_{it}\}_t$, and $\{(\alpha_{\tilde{g}t}^0 - \alpha_{gt}^0)v_{it}\}_t$. This allows us to apply the result in Theorem 2, replacing $\tilde{g}_w(\cdot)$ with $\hat{g}(\cdot)$, which implies the oracle property of the group membership estimate obtained from the initial unweighted K-means using the Euclidean distance, as in Theorem 1 of BM (2015).

Lemma 3 *Let Assumptions 5 and 6 hold. Suppose n, T, and B_T increase to infinity such that $B_T/T = o(1)$ and $B_T^2/n = o(1)$, Then, we have*

$$\lambda(\hat{\Sigma}_{B_T} - \Sigma_T) = O_p\left(B_T\sqrt{\frac{\log T}{n}} + \frac{1}{B_T^q}\right); \tag{23}$$

$$\lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) = O_p\left(B_T\sqrt{\frac{\log T}{n}} + \frac{1}{B_T^q}\right). \tag{24}$$

The result in Lemma 3 establishes the theoretical properties of the banded covariance matrix. The convergence rates in (23) and (24) reflect the variance and bias of $\hat{\Sigma}_{B_T}$. This result aligns with that of Bickel and Levina (2008a), who first proved the spectral norm consistency of the banded sample covariance estimator under (sub)-Gaussian noises.

Our settings in Lemma 3 extend their results to the high-dimensional K-means problem with sub-exponential shocks. The proof of Lemma 3 explicitly counts on the estimation uncertainties embodied in the estimated $\hat{v}_{it} = y_{it} - \hat{\alpha}_{\hat{g}(i),t}$, which is introduced by the initial (unweighted) K-means estimation. Additionally, we extend Bickel and Levina's (2008a) result with the Bartlett kernel (19) to a class of positive-definite and second-order kernel functions, such as (20) and (21), in the banded covariance estimation**.** Importantly, our estimation of $\hat{\Sigma}_{B_T}$ is nonparametric and agnostic to the specific form of the serial dependence structure in Σ_T.

Remark 3 *As shown in Lemma 3, when n and T grow to infinity, the optimal choice of the banding parameter, B_T, can be determined by balancing the trade-off between variance $(B_T^2 \log T)/n$ and bias (B_T^{-q}), as implied in (23). The corresponding optimal rate, $(\log T/n)^{-1/2(q+1)}$, however, may offer limited practical guidance, as the optimal B_T also depends on dependence structure of $\{v_{it}\}$, which is an infinite-dimensional object of interest. In the Appendix, we propose a data-driven procedure for selecting the optimal banding size, adapting Bickel and Levina's (2008a&b) approach to our weighted K-means setting. This procedure employs a resampling scheme to estimate and minimize the L_1/L_1-based matrix norm-induced risk of the regularized (banded) covariance estimator.*

The use of spectral norm, $\lambda(\cdot)$, in Lemma 3 enables us to properly control of the high-dimensional estimation noises embodied in the weighted K-means estimation (22). To be more specific, using the result in (24), we can show that

$$\frac{1}{nT}\sum_{i=1}^{n}(y_i - \alpha_{g(i)})'\hat{\Sigma}_{B_T}^{-1}(y_i - \alpha_{g(i)}) = \frac{1}{nT}\sum_{i=1}^{n}(y_i - \alpha_{g(i)})'\Sigma_T^{-1}(y_i - \alpha_{g(i)}) + o_p(1) \tag{25}$$

holds uniformly over $(\alpha, g(\cdot)) \in \mathcal{A}^{GT} \times \Gamma_G$. Theorem 4 below formally establishes the asymptotic properties of the feasible weighted K-means estimator, including the consistency of the estimated cluster centroids and its oracle group membership estimates.

Theorem 4 *Let Assumptions 1–6 hold. Then, (25) holds uniformly for the feasible weighted K-means objective in (2), and*

$$\frac{1}{nT}\sum_{i=1}^{n}(\hat{\alpha}^w_{\hat{g}_w(i)} - \alpha^0_{g_0(i)})'\Sigma_T^{-1}(\hat{\alpha}^w_{\hat{g}_w(i)} - \alpha^0_{g_0(i)}) = o_p(1), \tag{26}$$

as n, T, and B_T increase to infinity such that $B_T/T = o(1)$ and $B_T^2/n = o(1)$. The relabeling map $\sigma(\cdot)$, defined in (7) using $\hat{\alpha}^w_{\hat{g}}$, is one-to-one with probability tending to one. Moreover, for all $g \in \{1, \ldots, G\}$, we have:

$$\frac{1}{T}(\hat{\alpha}^w_{\sigma(g)} - \alpha^0_g)'\Sigma_T^{-1}(\hat{\alpha}^w_{\sigma(g)} - \alpha^0_g) = o_p(1) \text{ for all } g \in \{1, \ldots, G\}. \tag{27}$$

Finally, for any $\delta > 0$, it holds that:

$$P\left(\hat{g}_w(i) \neq \sigma(g_0(i)) \text{ for some } i \in \{1, \ldots, n\}\right) = o(1) + o\left(\frac{n}{T^\delta}\right). \tag{28}$$

Theorem 4 parallels the results in Theorem 1 and 2, indicating that the use of the banded covariance matrix estimator in the feasible weighted K-means preserves the same properties as the infeasible weighted K-means.

While the feasible weighted K-means procedure in Theorem 4 covers the two-step approach—where the initial membership $\hat{g}(\cdot)$ is obtained from the standard K-means estimator with $\Sigma_T = I_T$ and then updated by $\hat{g}_w(\cdot)$ using the weighted K-means objective in (22)—it is also possible to accommodate an iterated procedure. This iterated version is in the same spirit as the iterated generalized method of moments (GMM) estimator in Hansen et al. (1996). Specifically, let the initial membership estimate $\hat{g}_{w,(1)}(\cdot)$ by the feasible weighted K-means with the weight $\hat{\Sigma}^{-1}_{B_T,(0)}$ using $\hat{v}_{i,(0)} := y_i - \hat{\alpha}_{\hat{g}(i)}$. Then, for each iteration $m \geq 2$, given the previous membership estimate $\hat{g}_{w,(m-1)}(\cdot)$, we recompute the weight matrix $\hat{\Sigma}^{-1}_{B_T,(m-1)}$ based on the residuals $\hat{v}_{i,(m-1)}, := y_i - \hat{\alpha}_{\hat{g}_{w,(m-1)}(i)}$ which are induced by the latest membership and cluster centroid estimates. The weighted K-means is then updated by solving:

$$(\hat{\alpha}_{w,(m)}, \hat{g}_{w,(m)}(\cdot)) = \underset{(\alpha,g(\cdot))\in\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha_{g(i)})' \hat{\Sigma}^{-1}_{B_T,(m-1)} (y_i - \alpha_{g(i)}).$$

Repeating this procedure for a fixed number of iterations or until the membership estimates converge, i.e., $\hat{g}_{w,(m)}(\cdot) = \hat{g}_{w,(m+1)}(\cdot)$, yields the iterated weighted K-means outputs $(\hat{\alpha}_{w,\infty}, \hat{g}_{w,\infty}(\cdot))$. Our numerical experiments in Section 5 confirm that the iterated weighted K-means procedure can further improve both the membership assignments and the cluster centroid estimates.

4 Computation of Weighted K-Means Algorithm

As is well known, the computation of the K-means problem is NP-hard, as there are almost G^n ways to divide n observations into G clusters. To overcome this computational challenge, practitioners typically rely on efficient heuristic algorithms that converge quickly to a local optimum. In this section, we outline a computational procedure for obtaining feasible weighted K-means solutions.

Given the estimated banded covariance matrix $\hat{\Sigma}_{B_T}$, as defined in (17), we begin by weighting the original data and computing $\breve{y}_i = \hat{\Sigma}^{-1/2}_{B_T} y_i = (\breve{y}_{i1}, \ldots, \breve{y}_{iT})'$ for $i \in \{1, \ldots, n\}$ and $g \in \{1, \ldots, G\}$. Under this transformation, the weighted K-means objective function in (22) reduces to the following unweighted K-means objective:

$$(\hat{\tau}_w, \hat{g}_w(\cdot)) = \underset{(\tau,g(\cdot))\in\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{n} \sum_{t=1}^{T} (\breve{y}_{it} - \tau_{gt})^2,$$

where $\tau_g = (\tau_{g1}, \ldots \tau_{gT})' = \hat{\Sigma}^{-1/2}_{B_T} \alpha_g$ denotes a linear transformation of the original parameter. One widely used approach is the simple iterative algorithm of Lloyd (1982) and Forgy (1965), which BM (2015) adapts for high-dimensional settings. Specifically, the algorithm proceeds as follows:

1. Set $s = 0$ and randomly draw initial starting values of centroids, $\hat{\tau}_w^{(s)} = (\hat{\tau}_1^{w,(s)}, \ldots, \hat{\tau}_G^{w,(s)})'$.

2. For each $i \in \{1, \ldots, n\}$, compute

$$\hat{g}_w(i; \hat{\tau}_w^{(s)}) = \underset{g \in \{1,\ldots,G\}}{\arg\min} \sum_{t=1}^{T} (\breve{y}_{it} - \hat{\tau}_{gt}^{w,(s)})^2,$$

   where the minimum of g is taken in case of a non-unique solution.

3. Using $\hat{g}_w(\cdot; \hat{\tau}_w^{(s)})$ from the previous step, update the centroids as follows:

$$\hat{\tau}_w^{(s+1)} = \underset{\tau \in \mathcal{A}^{GT}}{\arg\min} \sum_{i=1}^{n} \sum_{t=1}^{T} (\breve{y}_{it} - \tau_{\hat{g}_w(i;\hat{\tau}_w^{(s)}),t})^2,$$

   where the corresponding solution $\hat{\tau}_g^{w,(s+1)}$ in $\hat{\tau}_w^{(s+1)}$, for each $g \in \{1, \ldots, G\}$, is the simple average of $\breve{y}_i$ across individuals sharing the same group membership as induced by $\hat{g}_w(i; \hat{\tau}_w^{(s)})$.

4. Repeat steps 2 and 3 until numerical convergence is achieved for the group memberships, that is, $\hat{g}_w(i; \hat{\tau}_w^{(s)}) = \hat{g}_w(i; \hat{\tau}_w^{(s+1)})$ for all $i \in \{1, \ldots, n\}$. If one or more groups are empty at this final stage, we reassign the observations farthest from their current centroids to the empty clusters and repeat the previous steps.

After steps in 1–4, the cluster centroid estimates $\hat{\alpha}_g^w$ can be obtained by weighting the transformed centroid estimates: $\hat{\alpha}_g^w = \hat{\Sigma}_{B_T}^{1/2} \hat{\tau}_g^w$ for each $g \in \{1, \ldots, G\}$. The iterative procedure is computationally efficient, but it may converge only to a local optimum and does not guarantee finding the global solution. In particular, since step 1 involves randomly selecting initial centroids from the data points $\{\breve{y}_i\}_{i=1}^n$, poor initialization can often cause the algorithm to converge to suboptimal local minima. To mitigate this issue, we rely on multiple random initializations, though this comes at the cost of additional computation time. Another approach is to add a second phase to steps 1–4, using a variant of the Hartigan and Wong (1976) algorithm, which iteratively reassigns individuals to the clusters that most reduce the objective function and updates the centroids accordingly, continuing until no further improvement is possible. This refinement, known as the HK-means algorithm (Hansen and Mladenovic̆, 2001) is widely implemented as the default in statistical software such as `R` and `MATLAB`. The details of the second-phase algorithm are provided in the Appendix.

Once we obtain the output of the weighted K-means, $(\hat{\alpha}_w, \hat{g}_w(\cdot))$, we can further refine the solution by iteratively updating the cluster memberships using the recomputed weight matrix until numerical convergence is achieved. Specifically, the iterative algorithm proceeds as follows:

5. Let $\hat{g}_{w,(1)}(\cdot)$ denote the membership estimate obtained from steps 1–4. We then recompute the weight matrix $\hat{\Sigma}_{B_T,(1)}^{-1}$ based on the new residuals $\hat{v}_{i,(1)}, := y_i - \hat{\alpha}_{\hat{g}_{w,(1)}(i)}$,

which are induced by the latest membership and cluster centroid estimates. The weighted K-means is subsequently updated by solving:

$$(\hat{\alpha}_{w,(2)}, \hat{g}_{w,(2)}(\cdot)) = \underset{(\alpha,g(\cdot))\in\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha_{g(i)})' \hat{\Sigma}^{-1}_{B_T,(1)} (y_i - \alpha_{g(i)}),$$

which effectively amounts to repeating steps 1–8 with the following weighted variables:

$$\breve{y}_i := \hat{\Sigma}^{-1/2}_{B_T,(1)} y_i \text{ and } \tau_g = \hat{\Sigma}^{-1/2}_{B_T,(1)} \alpha_g,$$

In this iteration, for the step 1, we use the weighted K-means cluster centroid estimates $\hat{\tau}^w_g$ as the initial centroids instead of randomly drawing them from the data $\{\breve{y}_i\}$.

6. We repeat the step 5 until the membership estimates converge, i.e., $\hat{g}_{w,(m)}(\cdot) = \hat{g}_{w,(m+1)}(\cdot)$ for some $m \geq 2$, yields the iterated weighted K-means outputs $(\hat{\alpha}_{w,\infty}, \hat{g}_{w,\infty}(\cdot))$.

5 Numerical Illustration of the Weighted K-Means

5.1 Grouped panel data

To evaluate the finite-sample performance of weighted K-means, we conduct a Monte Carlo simulation based on the simple grouped panel data model described in (1). Specifically, we consider the following data-generating process (DGP):

$$\begin{aligned} y_{it} &= \alpha^0_{g_0(i),t} + v_{it} \\ v_{it} &= \rho_0 v_{it-1} + \ u_{it} \text{ with } \ u_{it} \overset{\text{i.i.d.}}{\sim} \chi^2(1) - 1 \end{aligned}$$

for $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$. The group fixed effects, defined on a compact (closed and bounded) parameter space as in Assumption 2-(a), are generated according to the following process:

$$\alpha^0_{gt} = \mu^0_g + \omega_{gt}; \tag{29}$$

$$\omega_g = (\omega_{g1}, \ldots, \omega_{gT})' = \Omega^{1/2} e_g \text{ with } \Omega = \left[\psi_0^{|s-t|}\right]_{1\leq s,t\leq T} \ ; \tag{30}$$

$$e_g = (e_{g1}, \ldots, e_{gT})' \text{ with } e_{gt} \overset{\text{i.i.d.}}{\sim} \text{Truncated normal } N(0,1) \text{ on the interval } [-5,5] \tag{31}$$

for $g \in \{1, \ldots, G\}$ and $t \in \{1, \ldots, T\}$. We assume that the number of groups $G = 9$ to be known, and each group has the same number of individuals. Also, we set the group specific location parameter μ_g as

$$(\mu^0_1, \mu^0_2, \ldots, \mu^0_9)' = (-0.5, -0.25, \ldots, 1.25, 1.50)'.$$

We consider $n \in \{180, 270, 450\}$ with the balanced group sizes,i.e. $n_1 = \ldots = n_G = n/G$ and consider the number of time periods $T \in \{25, 50, 100\}$. The true parameters (ρ_0, θ_0) for the DGP are set to the values in $\{0, 0.25, 0.50, 0.75\}$. Below, we summarize the different K-means procedures implemented in our simulations.

- Euclidean K-means (BM 2015) in (2) with $W = I_T$. It is denoted as `KM` in the table of results.

- The feasible weighted K-mean using no-banded covariance estimator $\hat{\Sigma}_T$ in (16), denoted as `WKM` in the table. Note that we used `KM` to construct the plugged-in estimation of the unobserved shock $\hat{v}_i = (\hat{v}_{i1}, \ldots, \hat{v}_{iT})' = y_i - \hat{\alpha}_{\hat{g}(i)}$ in $\hat{\Sigma}_T$.

- The iterated version of `WKM`, which is denoted as `I-WKM.`

- The infeasible version of the banded weighted K-mean which uses banded covariance estimator $\tilde{\Sigma}_{B_T}$ that is based on the *true* idiosyncratic shock, i.e., $\tilde{\Sigma}_{B_T^*} = \tilde{\Sigma}_T \star W_T$, where
$$\tilde{\Sigma}_T = \frac{1}{n}\sum_{i=1}^{n} v_i v_i', \text{ where } v_i = y_{it} - \alpha^0_{g_0(i),t}.$$
It is denoted as `Inf-B-WKM`. From the theory, this procedure is expected to yield the best performance out of all procedures considered here. The bandwidth size is B_T^* can be computed by the resampling-based selection rule, as presented in subsection 8.1. The number of random permutation parameter M is chosen as 200.

- The feasible banded weighted K-means using the banded covariance estimator $\hat{\Sigma}_{B_T^*} = \hat{\Sigma}_T \star W_T$, using B_T^*, as described in $\tilde{\Sigma}_{B_T^*}$. The procedure is denoted as `B-WKM`. Note that we used `KM` to construct the plugged-in estimation of the unobserved shock $\hat{v}_i = (\hat{v}_{i1}, \ldots, \hat{v}_{iT})' = y_i - \hat{\alpha}_{\hat{g}(i)}$ in $\hat{\Sigma}_{B_T^*}$.

- The iterated version of `B-WKM`, which is denoted as `I-B-WKM.`

For all banding-type K-means procedures described above, we used the second-order Parzen kernel function $K_P(\cdot)$, which is presented in (20). In all of our procedures that solve (weighted) K-means algorithms, we utilize the built-in R function `kmeans`, with the number of random initializations set to $1,000$. After running a total of $5,000$ simulation replications, we compute the missclassification frequency of the estimated group membership, $\hat{g}(\cdot)$, which is equal to
$$\frac{1}{n}\sum_{i=1}^{n} 1(\hat{g}(i) = g_0(i)). \tag{32}$$
It is important to point out that the asymptotic oracle property of any group membership estimate, $\hat{g}(\cdot)$, is subject to the permutation (relabeling) of the group membership $g_0(\cdot)$. In our numerical experiments, we set the number of groups $G = 9$, and the total number of possible relabelling ($G!$) is more than 350,000. Therefore, it is essential to deal with the randomly relabeled result $\hat{g}(\cdot)$ when computing the missclassification rate in (32) at each simulated sample. For example, BM (2015) addresses the relabeling issue by finding the minimum over many randomly generated permutations. However, BM (2015)'s approach adds a significant computational burden as it may require generating hundreds of thousands of permutations for each simulation. To alleviate the computational challenge in BM (2015)'s approach, our numerical work implements a combinatorial optimization method

called the Hungarian method. The procedure, also known as the Kuhn-Munkres algorithm (Kuhn, 1956; Munkres, 1957), is particularly powerful in our simulation setting because it can efficiently find, in polynomial time, the optimal relabeling for $\hat{g}(\cdot)$ that matches the true group membership partition.

Tables 2–10 present missclassification rates of various membership estimates, as described above, across different values of $n \in \{180, 270, 450\}$, $T \in \{25, 50, 100\}$, and $\psi_0 \in \{0.25, 0.50, 0.75\}$, and $\rho_0 \in \{0.25, 0.50, 0.75\}$. The results first indicate that the missclassification of all K-means-based procedures considered in Tables 2–10 decrease as both the time dimension (T) and the cross-sectional dimension (n) increase.

Our results also indicate that the standard unweighted K-means algorithm exhibits finite-sample missclassification when the panel data exhibit temporal dependencies over time. Specifically, our results illustrate that even with moderately large time periods, such as $T = 50$, the missclassification rates of `KM` are more than 50% with $\psi = 0.75$. In addition, we also find that the unregularized weighting method, `WKM`, does not improve the performance of the K-means methods. Across all data generations in Tables 2–10, the missclassification rates of `WKM` largely mirror those of `KM`.

The results in Tables 2–10 also confirm that weighted K-means, `BWKM`, using the regularized weight matrix, can effectively reduce missclassification rates, compared to unweighted K-means across all considered DGPs. For instance, when $T = 50$ and $n = 180$ with $\psi_0 = 0.25$, the results in Table 2 show that the missclassification rates of `BWKM` range from 0.9% to 4.6%, whereas those of `KM` very from 1.4% to 62%, across the values of ρ_0 ranging from 0.25 to 0.75. We also find that the iterated weighted K-means with a banded weighted matrix, `I-BWKM`, can further improve finite-sample performance, particularly under the higher degrees of serial correlations (ρ_0 and ψ_0) in the panel data.

5.2 Grouped panel fixed-effects model

In this subsection, we extend our DGP by incorporating a common slope parameter and covariates $x_{it}'\theta_0$, as in the grouped panel fixed effects model of BM (2015):

$$y_{it} = x_{it}'\theta_0 + \alpha^0_{g_0(i),t} + v_{it}, \tag{33}$$

where $\theta_0 \in \Theta \subset \mathbb{R}^d$ denotes the true common slope parameter. Let x_i be a $T \times d$ matrix that stacks the covariates $x_{it} \in \mathbb{R}^d$ over time, i.e., $x_i' = (x_{i1}, x_{i2}, \ldots, x_{iT})$. For estimation with a common slope parameter, we can naturally extend the framework of BM (2015) by incorporating θ directly into the weighted K-means objective function:

$$\min_{(\theta,\alpha,g(\cdot))\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G} \sum_{i=1}^{n} (y_i - x_i\theta - \alpha_{g(i)})'\hat{\Sigma}^{-1}_{B_T}(y_i - x_i\theta - \alpha_{g(i)}), \tag{34}$$

where the parameters, θ, α, and $g(\cdot)$, are jointly optimized over the parameter spaces $\Theta \subset \mathbb{R}^d$, $\mathcal{A}^{GT}$, and Γ_G, respectively.

Although the joint search for $(\theta, \alpha, g(\cdot))$ in the weighted K-means can be implemented in a straightforward manner, it is well recognized in the literature, e.g., BM (2015), Moon

and Weidner (2018), and Chertverikov and Manresa (2022, CM hereafter), that such joint minimization can require a prohibitive number of initial starting values and lead to an exponential increase in computational burden, particularly as the dimension of the covariates, d, grows. A recent contribution by CM (2022) addresses this issue by proposing spectral estimation technique, which can also be readily incorporated into our weighted K-means framework. Specifically, CM (2022) assumes that the covariates x_{it} are generated by the linear combination of R-number of group specific time effects:

$$x_{it} = \sum_{r=1}^{R} \xi_{ir} \alpha_{g(i),t}^{(r)} + z_{it}, \tag{35}$$

where $\xi_{ir} \in \mathbb{R}^d$, and $z_{it} \in \mathbb{R}^d$ is an idiosyncratic zero-mean noise term. Following CM (2022), we can let the first group component in (35) is equal to the group fixed effect parameter,i.e., $\alpha_{gt}^{(1)} = \alpha_{gt}$ for all g and t, without loss of generality. Define an $n \times n$ symmetric matrix $A(\theta) = [A_{ij}(\theta)]_{1 \leq i,j \leq n}$ that depends on $\theta \in \Theta$ as follows:

$$A_{ij}(\theta) := \frac{1}{nT} \sum_{t=1}^{T} \left\{ (y_{it} - x_{it}'\theta) - (y_{jt} - x_{jt}'\theta) \right\}^2$$

for all $i, j \in \{1, \ldots, n\}$. Additionally, let $\nu_1(\theta), \ldots, \nu_{2(GR+1)}(\theta)$ be the $2GR + 2$ largest absolute eigenvalues for $A(\theta)$, and define $F(\theta)$ as their sum:

$$F(\theta) = \nu_1(\theta) + \ldots + \nu_{2(GR+1)}(\theta) = \theta'\Omega\theta + S'\theta + L + o_p(1),$$

where $\Omega \in \mathbb{R}^{d \times d}$, $L \in \mathbb{R}$, and $S \in \mathbb{R}^{d \times 1}$. These components can be estimated as follows. For $k, l \in \{1, \ldots, d\}$, define:

$$\begin{aligned}
\hat{L} = F(0) \text{ and } \hat{S}_k &= \frac{F(e_k) - F(-e_k)}{2}; \\
\hat{\Omega}_{kk} = F(e_k) - \hat{S}_k - \hat{L} &= \frac{F(e_k) + F(-e_k)}{2} - \hat{L}; \\
\hat{\Omega}_{kl} &= \frac{F(e_k + e_l) - \hat{\Omega}_{kk} - \hat{\Omega}_{ll} - \hat{S}_k - \hat{S}_l - \hat{L}}{2} \text{ for } k \neq l,
\end{aligned}$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)'$ is the unit vector with 1 in the j-th position and zeros elsewhere. We can then define the spectral estimator as

$$\hat{\theta}_{sp} = \arg\min_{\theta \in \Theta} \left(\theta'\hat{\Omega}\theta + \hat{S}'\theta + \hat{L} \right) = -\frac{\hat{\Omega}^{-1}\hat{S}}{2}. \tag{36}$$

Using $\hat{\theta}_{sp}$, we construct the weighted K-means objective function as

$$\min_{\alpha, g(\cdot) \in \mathcal{A}^{GT} \times \Gamma_G} \sum_{i=1}^{n} (y_i - x_i\hat{\theta}_{sp} - \alpha_{g(i)})' \hat{\Sigma}_{B_T}^{-1} (y_i - x_i\hat{\theta}_{sp} - \alpha_{g(i)}). \tag{37}$$

After plugging in the spectral estimator $\hat{\theta}_{sp}$, the objective function is minimized only over $\mathcal{A}^{GT}$ and Γ_G. As a result, the spectral estimation framework becomes directly applicable to our proposed weighted K-means algorithms by simply replacing the outcome variable y_i with $y_i - x_i\hat{\theta}_{sp}$. Therefore, the computational burden of joint minimization in the weighted K-means procedure can be substantially reduced in the grouped panel data setting.

To evaluate the finite-sample performance of weighted K-means using the initial spectral estimator $\hat{\theta}_{sp}$, we simulate the grouped fixed effect in (33), where the DGP for $\alpha^0_{g_0(\cdot),t}$ and v_{it} are the same as in subsection 5.1. The number of covariates $x_{it} = (x^{(1)}_{it}, \ldots, x^{(d)}_{it})'$ is set as $d = 2$ with $\theta_0 = (-1, 0.8)'$, and we have that

$$x^{(j)}_{it} = \xi^{(j)}_i \alpha^0_{g_0(i),t} + z^{(j)}_{it} \text{ for } j \in \{1, 2\} \text{ with } R = 1; \tag{38}$$

$$x^{(j)}_{it} = \xi^{(j)}_{i1} \alpha^0_{g_0(i),t} + \xi^{(j)}_{i2} \tilde{\alpha}^{(0)}_{g_0(i),t} + z^{(j)}_{it} \text{ for } j \in \{1, 2\} \text{ with } R = 2, \tag{39}$$

where the additional group specific shock $\tilde{\alpha}^{(0)}_{g(i),t}$ in (39) is drawn in the same way as in (29)–(31), with $\psi_0 = 0.25$. The idiosyncratic shock $z^{(j)}_{it}$ is generated as

$$z^{(j)}_{it} = \rho_{z0} z^{(j)}_{it-1} + \ u^{(j)}_{z,it} \text{ with } u^{(j)}_{z,it} \overset{\text{i.i.d.}}{\sim} N(0, 1)$$

with $\rho_{z0} = 0.25$. The coefficients $\xi^{(j)}_i$ and $(\xi^{(j)}_{i1}, \xi^{(j)}_{i2})'$ in (38) and (39) are generated by

$$\xi^{(j)}_i = \varrho + e^{(j)}_i 1(|e^{(j)}_i| < 5) \text{ with } e^{(j)}_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$$

for (38), and we have that

$$\xi^{(j)}_{i1} = \varrho + e^{(j)}_{i1} 1(|e^{(j)}_{i1}| < 5) \text{ with } e^{(j)}_{i1} \overset{\text{i.i.d.}}{\sim} N(0, 1) \text{ for } j \in \{1, 2\};$$
$$\xi^{(j)}_{i2} = \varrho + e^{(j)}_{i2} 1(|e^{(j)}_{i2}| < 5) \text{ with } e^{(j)}_{i2} \overset{\text{i.i.d.}}{\sim} N(0, 1) \text{ for } j \in \{1, 2\}$$

for (38). The parameter ϱ measures the degree of endogeneity in the grouped panel data, as it controls the correlation between the covariate $x^{(j)}_{it}$ and the time-varying grouped-fixed effect parmater $\alpha^0_{g_0(i),t}$. We set the value $\varrho = 3$.

Tables 11–16 present the misclassification rates for various membership estimation methods after incorporating the common parameter θ obtained from (37). Specifically, Tables 11–13 refer to the DGP with $R = 1$, while Tables 14–16 correspond to the DGP with $R = 2$. The results consistently show that our weighted K-means method, with data-dependent banding and iterations, significantly improves the accuracy of group classification across all DGP parameter combinations.

We also assess the performance of the post-estimator of the common slope parameter θ, which is based on the estimated group memberships from feasible unweighted and weighted K-means (`KM`,`B-WKM`, `I-B-WKM`), as well as the spectral estimator proposed in CM (2022), denoted by `CM`. The comparison focuses on the bias and standard deviation (efficiency) of the common parameter estimates across various DGP parameter settings. This comparison focused on the bias and standard deviations (efficiency) of the common parameter estimates

across various DGP parameter settings. To save pages, we only report results for $n = 180$, as similar qualitative implications are observed for other values of n. The results are presented in Tables 17 and 18, with Table 17 corresponding to DGP $R = 1$, and Table 18 to DGP $R = 2$. Due to its more accurate membership estimation, our weighted K-means method produces lower bias and standard deviations in the common parameter estimates compared to both the post-spectral estimator of CM (2022) and the GFE estimator of BM (2015), which relies on standard K-means. This improvement is particularly evident in cases where the data exhibit strong serial correlation.

For instance, as in Table 17, our results indicate that the standard unweighted K-means algorithm (`KM`) shows notable bias and variability when handling high levels of serial correlation in panel data, particularly when $\rho_0 = 0.75$. Specifically, even with moderately large time periods, such as $T = 50$, `KM` suffers from increased bias, as demonstrated by a bias of 0.1385 for $\hat{\theta}_1$, while the post-spectral estimation (`CM`), though slightly better, still exhibits significant bias (0.1412). Additionally, the iterated weighted K-means method with a banded weight matrix (`I-BWKM`) shows a clear advantage over both `KM` and `CM`. When $T = 50$ with $\rho_0 = 0.75$, `I-BWKM` reduces the bias of $\hat{\theta}_1$ to 0.0612, with a smaller value of the standard deviation, 0.035), outperforming the other methods. This pattern holds as the time period increases to $T = 100$, where `I-BWKM` maintains the lowest bias (0.0237) and smallest standard deviation (0.0217), highlighting its robustness in addressing strong serial correlations.

6 Empirical illustration

In this section, we present a brief empirical study on the impact of Foreign Direct Investment (FDI) on the innovation capacity of China's cities as a numerical illustration of our approach using real data. Over the past two decades, China has emerged as a leading global recipient of FDI, with its cities playing a pivotal role in fostering economic growth and technological advancement. FDI has been a crucial factor in China's transition from a manufacturing-based economy to one increasingly driven by knowledge, research, and innovation (Chen et al., 2022). However, the impact of FDI on innovation is not uniform across China's cities, which vary significantly in terms of economic development, industrial structure, human capital, and institutional frameworks. Rather than relying on pre-defined classifications, such as administrative regions or provinces, our data-driven weighted K-means approach uncovers natural groupings of cities based purely on the underlying data. Specifically, our approach addresses potential serial correlation in the data, enhancing the robustness of our findings and providing deeper insights into group patterns that standard unweighted K-means may not capture. This allows for a more nuanced analysis, revealing whether factors like geographic location (coastal vs. inland) or city size (large metropolitan areas vs. smaller cities) influence how FDI fosters innovation. By identifying these latent group patterns, we can better understand how FDI impacts different types of cities and uncover relationships that may be missed with traditional methods.

6.1 Data, Variables, and Model

The dataset used in this research is derived from the China Urban Statistical Yearbook and spans the period from 2004 to 2019. It covers more than 290 cities, including both entire municipalities and administrative districts. The dataset offers a comprehensive view of urban social and economic development in China, providing statistics across multiple sectors. These include demographics, labor force, land resources, economy, industry, transportation, telecommunications, trade, foreign economic relations, fixed asset investment, education, culture, health, living standards, social security, municipal utilities, and environmental protection.

The main variables selected for analysis are as follows: Innovation level (Inno) is measured by the number of invention patent applications in a city. FDI represents the actual amount of foreign investment utilized, measured in USD. Human capital (HumCap) is calculated as the ratio of undergraduate students in regular higher education institutions to the total population. Physical capital (PhyCap) is the ratio of gross fixed asset investment to GDP. R&D expenditure is the proportion of government spending on science and technology relative to total fiscal spending. Population density (Pop) is measured as the total population divided by the city's land area. University count (Univ) refers to the number of higher education institutions in a city, and GDP per capita (GDPpc) is the GDP divided by the total population. All variables are at the city level and have been log-transformed for better data presentation.

We estimate the following panel data regression model, incorporating latent group-specific time-varying fixed effects, to analyze the factors influencing the innovation level across cities:

$$\text{Inno}_{it} = x_{it}'\theta + \alpha_{g(i),t} + v_{it}, \tag{40}$$

where $x_{it}' = (\text{FDI}_{it}, \text{HumCap}_{it}, \text{PhyCap}_{it}, \text{RDexp}_{it}, \text{Pop}_{it}, \text{Univ}_{it}, \text{GDPpc}_{it})$ is the vector of independent variables, and $\alpha_{g(i),t}$ represents the latent group-specific time-varying effects, with group membership of each city being unknown.

6.2 Results

We first estimate the potential number of groups for all 279 data available China's cities using three commonly applied methods: the Elbow Method, the Silhouette Method, and the Gap Statistic. The results are presented in Figure 1. Based on these results, selecting $G = 4$ appears to be a reasonable choice. In the Elbow Method, a clear "elbow" is observed at $G = 4$, where further increases provide diminishing returns in variance reduction. The Silhouette Method and the Gap Statistic also show a relatively high score at $G = 4$, indicating well-separated clusters. Therefore, we set $G = 4$ for our subsequent analysis.

Next, we estimate the group memberships of the 279 cities using both the standard K-means method and our Iterated Banded Weighted K-means (I-BWKM), which applies a banded weighted matrix, all based on innovation levels. The results are visually represented on a map of China in Figure 2, where different colors distinguish the group memberships. Cities belonging to the group with the highest average group-specific effects are shaded in the darkest blue. The left panel of figure 2 shows the group memberships estimated using

the standard unweighted K-means method, though the grouping pattern is not clearly discernible. However, after addressing potential serial correlation using our weighted K-means (`I-B-WKM`) approach, the group patterns become much more distinct, as shown in the right panel of Figure 2, e.g., Beijing, Shanghai, and Shenzhen, China's largest metropolises, stand out distinctly, grouped into the top-tier category (Group 4). Similarly, cities such as Tianjin, Nanjing, Suzhou, Hangzhou, Wuhan, Xi'an, Chengdu, and Guangzhou, all large urban hubs, are assigned to Group 3. These cities become prominent only after mitigating the serial correlation issue; without this adjustment, their significance is muted, blending into the broader urban landscape.

Based on the estimated group memberships of China's cities, we conduct a standard linear regression analysis using the variables listed in (40), including interaction terms between city group memberships and years. The regression results are presented in Table 1. To save space, we omit the estimates for the intercept and group-specific time-varying fixed effects. Our results indicate a positive correlation between city innovation levels and FDI, consistent with the findings in the literature, such as Cantwell (1989) and Dunning and Lundan (2008).

One possible explanation is that FDI facilitates technology transfer and spillovers, where multinational corporations introduce advanced technologies and innovative practices to host cities. Local firms benefit from this exposure, often leading to improvements in productivity and innovation (Aitken and Harrison, 1999). Additionally, FDI intensifies competition, compelling domestic firms to innovate in order to remain competitive. As foreign firms enter local markets, domestic companies are driven to improve their products, services, and processes to maintain market share (Porter, 2011). Finally, FDI provides local firms with access to international markets, integrating them into global production networks. This exposure encourages local firms to innovate in order to meet international standards (Dunning and Lundan, 2008).

7 Conclusion

This paper proposes a weighted K-means approach with a Mahalanobis-type distance measure that accounts for serial correlation and heteroskedasticity in panel-data idiosyncratic shocks. The distance uses the inverse covariance matrix to rescale Euclidean distance, thereby incorporating the full covariance structure. As a result, our new weighted K-means objective function enables improved precision in latent group recovery in finite samples by filtering out noise from temporal dependence and heteroskedasticity.

Assuming both the cross-sectional and time dimensions grow large, we develop an asymptotic theory establishing consistency of the estimated centroids and the oracle property for group membership. For practical implementation, we propose a feasible weighted K-means method based on regularized covariance estimation and prove its asymptotic equivalence to the infeasible weighted K-means. Our Monte Carlo simulations confirm the effectiveness of the feasible weighted K-means algorithm in terms of group classification rates in finite samples, particularly under strong serial dependence in group trends and idiosyncratic components. Finally, we illustrate the method with an empirical study of the

impact of Foreign Direct Investment (FDI) on the innovation capacity of Chinese cities.

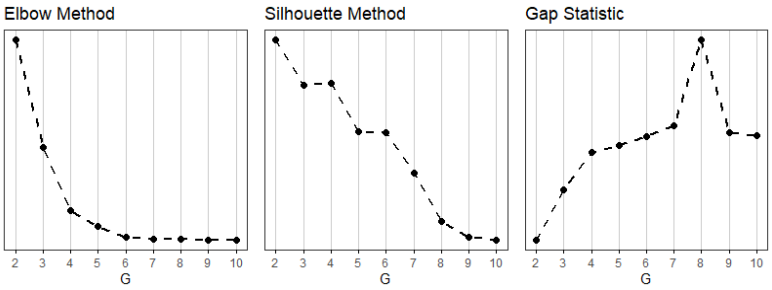

Figure 1: Estimation of the number of groups

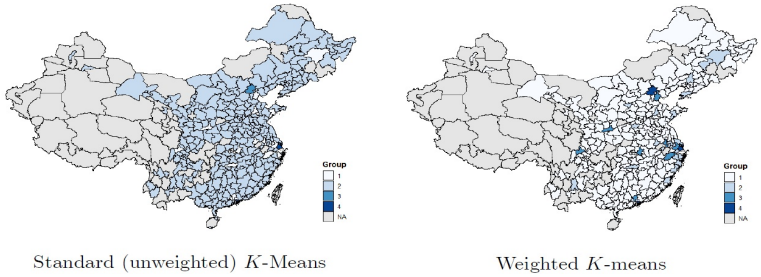

Figure 2: Group memberships of 279 China's cities based on innovation levels

Table 1: Results for grouped fixed effect regression in (40)

	Independent Variable							
Inno	FDI	RDexp	HumCap	PhyCap	RDemp	Pop	Univ	GDPpc
	0.1831***	0.3575***	0.2171***	-0.5456***	0.5827***	0.0010***	-0.0005	0.1644***
	(0.0170)	(0.0330)	(0.0359)	(0.0515)	(0.0240)	(0.0002)	(0.0067)	(0.0360)
	Observations 1,499 (Number of cities 279), Adjusted R^2= 0.8397							
	Group-specific time-varying FE Yes							

Note: $^{*}p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.001$;Clustered standard errors (by group) in parentheses.

8 Appendix

8.1 Practical guidance on banding size and the second-phase algorithm

In this subsection, we provide practical guidance and details on algorithms to implement the weighted K-means. We first provide an algorithm to choose the optimal banding parameter:

i) Run the initial unweighted K-means clustering and obtain the estimated $\{\hat{v}_{it}\}$ with $\hat{v}_{it} = y_{it} - \hat{\alpha}_{\hat{g}(i),t}$. This can be done by following steps 1–4 in Section 4 by setting $\hat{\Sigma}_{B_T} = I_T$.

ii) Randomly split the panel data $\{y_i\}_{i=1}^n$ by $n_1 = n/3$ and $n_2 = (2n)/3$ blocks up to M-times.

iii) For each $m \in \{1, \ldots, M\}$ split, we compute the two unbanded sample covariance matrices, $\hat{\Sigma}_{(I)}^{(m)}$ and $\hat{\Sigma}_{(II)}^{(m)}$ for each n_1 and n_2- numbers of splitted samples, respectively.

iv) Let $\hat{\Sigma}_{(II)}^{(m)}$ be the the "target" to compare its distance with the regularized estimator, $\hat{\Sigma}_{(I),B_T}^{(m)} = \hat{\Sigma}_{(I)}^{(m)} \star \mathcal{K}_T$, which depends on B_T via $\mathcal{K}_T$.

v) Choose B_T^* that minimizes the following empirical l_1 to l_1 matrix norm:

$$B_T^* = \underset{B_T \in \{0,1,\ldots,T\}}{\arg\min} \hat{R}(B_T), \tag{41}$$

where

$$\hat{R}(B_T) = \frac{1}{M} \sum_{m=1}^{M} ||\hat{\Sigma}_{(I),B_T}^{(m)} - \hat{\Sigma}_{(II)}^{(m)}||_{(1,1)},$$

and $||A||_{(1,1)} = \max_{1 \le j \le T} \sum_{i=1}^{T} |a_{ij}|$ for $A = [a_{ij}]_{1 \le i,j \le T}$.

In v), we employ the unregularized (unbanded) sample covariance $\hat{\Sigma}_{(II)}^{(m)}$ as the target object in $\hat{R}(B_T)$. As noted in Section 2, the point estimation of $\hat{\Sigma}_{(II)}^{(m)}$ possesses high noise, especially in high-dimensional cases. Thus, the use of $\hat{\Sigma}_{(II)}^{(m)}$ instead of the true Σ_T tends to overestimate the actual value of the targeted risk, $R(B_T) = E[||\hat{\Sigma}_{(I),B_T} - \Sigma_T||_{(1,1)}]$. Nevertheless, the unbiased property of the unbanded sample covariance can deliver favorable finite-sample properties for the feasible optimal banding parameter, as also noted in Bickel and Levina (2008a&b).

Regarding the choice of kernel weights $K(\cdot)$, the asymptotic theory developed in Section 2 indicates that any kernel weight function $K(\cdot)$ in Assumption 5 guarantees the consistency properties of estimated $\hat{\Sigma}_{B_T}$ and group membership $\hat{g}_w(\cdot)$. When the spectral density of $\{v_{it}\}$ satisfies the smoothness condition in Assumption 2-(c) with $q = 2$, second-order kernels such as Parzen and Tukey-Hanning (see (20)) can outperform the first-order Bartlett

kernel in our weighted K-means problem. This can be easily verified from our proof of Lemma 3, which demonstrates that the bound on the bias of $\hat{\Sigma}_{B_T}$ is

$$\left(\frac{k_q}{B_T^q}\right)\sup_t \sum_{h=-\infty}^{\infty} |h|^q |E[v_{it}v_{it-h}]|.$$

Thus, we recommend to use the second order kernel functions, such as $K_P(\cdot)$ and $K_{TH}(\cdot)$, in our weighted K-means problem. Among the second-order kernels, results from Andrews (1991) and Sun and Yang (2020) show that the following QS-kernel has the optimal property, minimizing the term k_q with $q = 2$:

$$K_{QS}(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\right),$$

and one may consider a modified version of QS kernel function, $\tilde{K}_{QS}(x) = 1(x \leq 1)K_{QS}(x)$ and utilize it to construct $\hat{\Sigma}_{B_T}$ to benefit from its optimal property. However, this approach can sacrifice the positive definite property of the high-dimensional matrix $\hat{\Sigma}_{B_T}$, and thus requiring an additional scheme to ensure $\hat{\Sigma}_{B_T}$ is positive definite. In this regard, developing more accurate bias-corrected weights with positive definiteness in the weighted K-means problem is an interesting open question, which we leave for future exploration.

Finally, we introduce the detailed second-phase algorithm that can be implemented on top of the basic iterative steps 1–4 in Section 4. Let $(\hat{\tau}_w^{(s)}, \hat{g}_w(i; \hat{\tau}_w^{(s)}))$ be the last output of the algorithm in the first phase algorithms in steps 1–4. Also, denote

$$Q_{n,T}^{(s)} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\breve{y}_{it} - \hat{\tau}^{(s)}_{\hat{g}_w(i;\hat{\tau}_w^{(s)}),t})^2, \tag{42}$$

as the value of the K-means objective function plugged by the current s-th solution. The key process of the second phase involves computing the change in $Q_{n,T}^{(s)}$ resulting from only moving a single individual i from its current cluster $g_i := \hat{g}_w(i; \hat{\tau}_w^{(s)})$ to a new cluster $h \neq g_i$. We then compute the new objective value of (42) after accounting for the loss of i from group g_i and its addition to the new group h, respectively. Let $\Delta^{(s)}(g_i, h)$ denote the change in the objective function. The amount of change $\Delta^{(s)}(g_i, h)$ can be simply calculated as below (Steinley and Brusco, 2007):

$$\Delta^{(s)}(g_i, h) := \frac{1}{nT} \times \left(\frac{\hat{n}_{g_i}}{\hat{n}_{g_i} - 1}\sum_{t=1}^{T}(\breve{y}_{it} - \hat{\tau}^{w,(s)}_{g_i,t})^2 - \frac{\hat{n}_h}{\hat{n}_h + 1}\sum_{t=1}^{T}(\breve{y}_{it} - \hat{\tau}^{w,(s)}_{ht})^2\right), \tag{43}$$

where $\hat{n}_{g_i}$ and $\hat{n}_h$ indicate the estimated group sizes for g_i and h, respectively, which are implied by $(\hat{\tau}^s, \hat{g}_w(i; \hat{\tau}_w^{(s)}))$. The individual i is reassigned to the new cluster that results in the largest decrease of the objective, and the group-fixed centroids are updated to their newly calculated positions. This process of evaluating moves and updating centroids continues iteratively for the next individual and stops when no more moves result in a decrease, indicating convergence. The detailed algorithms for this second phase, which follow steps 1–4 in Section 4, are outlined as follows:

1. Given $\hat{\tau}_w^{(s)}$ and $\hat{g}_w(i;\hat{\tau}_w^{(s)})$ which are estimated in the last stage s in the step 4, and the first individual $i=1$, we compute $\Delta^{(s)}(g_i,h)$ for all $h \neq g_i$, and get

$$h^* = \arg\min_{h \neq g_i} \Delta^{(s)}(g_i,h).$$

2. If $\Delta^{(s)}(g_i,h^*) > 0$, then reassign i to the new group h^*, and save the new group fixed effect estimates. We denote the updated results as $(\hat{\tau}_w^{(s+1)}, \hat{g}(i;\hat{\tau}_w^{(s+1)}))$. If $\Delta^{(s)}(g_i,h^*) \leq 0$, the individual i stays in its current group g_i, and there are no-updates on the $(\hat{\tau}_w^{(s)}, \hat{g}_w(i;\hat{\tau}_w^{(s)}))$.

3. Repeat the previous step 7 for the next $i=2$, 3, and so on. The procedure terminates at the stage $\tilde{s}$ once $\Delta^{(\tilde{s})}(g_i,h_i) \leq 0$ for all $i \in \{1,\ldots,n\}$.

8.2 Proofs of Main Results

To prove main results in Sections 2 and 3, we first introduce the following technical lemmas.

Lemma 5 (Lemma B.5 in BM (2015)) *Let Z_t be a strongly mixing process with* zero mean*, with strong mixing coefficients $\alpha_z[t] \leq \exp(-at^{d_1})$ for some $a, d_1 > 0$, and with tail probabilities $P(|Z_t| > z) \leq \exp(1-(z/b)^{d_2})$ for some $b, d_2 > 0$. Then, for all $z > 0$, we have,*

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| \geq t\right) = o\left(\frac{1}{T^{\delta}}\right) \text{ for all } \delta > 0.$$

Lemma 6 (Bernstein's Inequality, Corollary 2.8.3 in Vershynmin (2018)) *Let $X_1,\ldots,X_n$ be independent, mean-zero, sub-exponential random variables. Then for every $t \geq 0$, we have that*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right|\right\} \leq 2\exp\left(-cn\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right),$$

where $K = \max_i ||X_i||_{\psi_1} < \infty$, and $||X_i||_{\psi_1} = \inf\{s > 0 : E[\exp(|X_i|/s)] \leq 2\}$ denotes the sub-exponential norm.

Lemma 7 (Geršgorin's theorem, Corollary 6.15 in Horn and Johnson (2012)) *For any $p \times p$ square matrix, $A = [a_{ij}]_{1 \leq i,j \leq p}$, let $\{\lambda_k(A)\}_{k=1}^p$ be a set of eigenvalues of A. Then, we have that*

$$\max_{1 \leq k \leq p} |\lambda_k(A)| \leq \min\left\{\max_{1 \leq i \leq p} \sum_{j=1}^{p} |a_{ij}|, \ \max_{1 \leq j \leq p} \sum_{i=1}^{p} |a_{ij}|\right\}.$$

Lemma 8 (Lemma A.2. in Fan et al. (2011)) *Suppose that the two random variables Z_1 and Z_2 satisfy the following exponential tail condition: There exists $r_1, r_2 \in (0,1)$ and $b_1, b_2 > 0$, such that for all $t > 0$,*

$$P(|Z_i| > t) \leq \exp\left(1 - \left(\frac{t}{b_i}\right)^{r_i}\right) \text{ for } i \in \{1,2\}.$$

Then, for some $r_3 \in (0,1)$ and $b_3 > 0$, and any $t > 0$, the following inequality holds:

$$P(|Z_1 Z_2| > t) \leq \exp\left(1 - \left(\frac{t}{b_3}\right)^{r_3}\right).$$

Proof of Theorem 1. Let $\gamma_0 = \{g_0(1), \ldots, g_0(n)\}$ denote the true population memberships, and let $\gamma = \{g(1), \ldots, g(n)\}$ denote an arbitrary grouping of the cross-sectional units into G groups. Also, let us define

$$\begin{aligned}\hat{Q}_w(\alpha, \gamma) &= \frac{1}{nT}\sum_{i=1}^{n}(y_i - \alpha_{g(i)})'\Sigma_T^{-1}(y_i - \alpha_{g(i)}) \\ &= \frac{1}{nT}\sum_{i=1}^{n}(\alpha^0_{g_0(i)} - \alpha_{g(i)} + v_i)'\Sigma_T^{-1}(\alpha^0_{g_0(i)} - \alpha_{g(i)} + v_i).\end{aligned}$$

We also define the following auxiliary objective function:

$$\tilde{Q}_w(\alpha, \gamma) = \frac{1}{nT}\sum_{i=1}^{n}(\alpha^0_{g_0(i)} - \alpha_{g(i)})'\Sigma_T^{-1}(\alpha^0_{g_0(i)} - \alpha_{g(i)}) + \frac{1}{nT}\sum_{i=1}^{n} v_i'\Sigma_T^{-1}v_i,$$

and want to show that

$$\sup_{(\alpha,\gamma)\in\mathcal{A}^{GT}\times\Gamma_G}\left|\hat{Q}_w(\alpha,\gamma) - \tilde{Q}_w(\alpha,\gamma)\right| = o_p(1), \tag{44}$$

as $n, T \to \infty$. To prove this result, note that

$$\hat{Q}_w(\alpha,\gamma) - \tilde{Q}_w(\alpha,\gamma) = \frac{2}{nT}\sum_{i=1}^{n} v_i'\Sigma_T^{-1}\alpha^0_{g_0(i)} - \frac{2}{nT}\sum_{i=1}^{n} v_i'\Sigma_T^{-1}\alpha_{g_0(i)}.$$

Thus, we need to show that $(nT)^{-1}\sum_{i=1}^{n} v_i'\Sigma_T^{-1}\alpha_{g(i)}$ is $o_p(1)$, uniformly over the parameter space $\mathcal{A}^{GT}\times\Gamma_G$. Let $\bar{\tau}_g = \Sigma_T^{-1}\alpha_g = (\bar{\tau}_{g1}, \ldots, \bar{\tau}_{gT})'$ for $g \in \{1, \ldots, G\}$. Then, we can write that

$$\begin{aligned}\underbrace{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} v_{it}\bar{\tau}_{g_0(i),t}}_{=\frac{1}{nT}\sum_{i=1}^{n} v_i'\Sigma_T^{-1}\alpha_{g(i)}} &= \sum_{g=1}^{G}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} 1\{g_0(i) = g\}v_{it}\bar{\tau}_{gt}\right) \\ &= \sum_{g=1}^{G}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\bar{\tau}_{gt}\cdot\frac{1}{n}\sum_{i=1}^{n} 1\{g(i) = g\}v_{it}\right)\right]. \end{aligned}\tag{45}$$

Applying Cauchy-Schwarz inequality for all $g \in \{1, \ldots, G\}$, we obtain that

$$\begin{aligned}\left(\frac{1}{T}\sum_{t=1}^{T}\bar{\tau}_{gt}\cdot\frac{1}{n}\sum_{i=1}^{n} 1\{g_0(i) = g\}v_{it}\right)^2 &\leq \underbrace{\frac{1}{T}\sum_{t=1}^{T}\bar{\tau}_{gt}^2}_{=\frac{1}{T}\alpha_g'\Sigma_T^{-2}\alpha_g} \times \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n} 1\{g_0(i) = g\}v_{it}\right)^2 \\ &\leq \left(\lambda_{\min}^{-2}(\Sigma_T)\cdot\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^2\right) \times \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n} 1\{g_0(i) = g\}v_{it}\right)^2.\end{aligned}$$

In view of Assumptions 1 and 2-(a), the term $\lambda_{\min}^{-2}(\Sigma_T)\cdot T^{-1}\sum_{t=1}^{T}\alpha_{gt}^2$ is bounded. Also, note that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n}1\{g_0(i)=g\}v_{it}\right)^2 &= \frac{1}{n^2}\sum_{i,j=1}^{n}1\{g_0(i)=g\}1\{g_0(j)=g\}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}\right)\\ &\leq \frac{1}{n^2}\sum_{i,j=1}^{n}\left|\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}\right|\\ &\leq \frac{1}{n^2}\sum_{i,j=1}^{n}\left|\frac{1}{T}\sum_{t=1}^{T}E[v_{it}v_{jt}]\right| + \frac{1}{n^2}\sum_{i,j=1}^{n}\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}v_{jt}-E[v_{it}v_{jt}])\right|\end{aligned} \tag{46}$$

From Assumption 2-(b), we have that

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\left|\frac{1}{T}\sum_{t=1}^{T}E[v_{it}v_{jt}]\right| \leq \frac{M}{n}. \tag{47}$$

Also, by Cauchy-Schwarz inequality,

$$\begin{aligned}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}v_{jt}-E[v_{it}v_{jt}])\right|\right)^2 &\leq \frac{1}{n^2T}\sum_{i,j=1}^{n}\frac{1}{T}\left(\sum_{t=1}^{T}(v_{it}v_{jt}-E[v_{it}v_{jt}])\right)^2\\ &\leq \frac{M}{T\epsilon}=o(1),\end{aligned}$$

where the last inequality follows from Markov inequality and Assumption 2-(c), i.e.,

$$P\left(\frac{1}{n^2T}\sum_{i,j=1}^{n}\frac{1}{T}\left(\sum_{t=1}^{T}(v_{it}v_{jt}-E[v_{it}v_{jt}])\right)^2>\epsilon\right)\leq\frac{1}{T\epsilon}\left|\frac{1}{n^2T}\sum_{i,j=1}^{n}\sum_{t,s=1}^{T}Cov(v_{it}v_{jt},v_{is}v_{js})\right|=\frac{M}{T\epsilon}.$$

These results show that the upper bound in (46) is $o_p(1)$, which, together with $\lambda_{\min}^{-2}(\Sigma_T)\cdot T^{-1}\sum_{t=1}^{T}\alpha_{gt}^2 = O(1)$, implies that the right-hand side of (45) is $o_p(1)$ uniformly over $\mathcal{A}^{GT}\times\Gamma_G$. Consequently,
$(nT)^{-1}\sum_{i=1}^{n}v_i'\Sigma_T^{-1}\alpha_{g(i)} = o_p(1)$ holds uniformly. Now, to show the result in (8), note that by the definition of $\tilde{Q}_w(\alpha,\gamma)$,

$$P\left(\underbrace{\frac{1}{nT}\sum_{i=1}^{n}(\tilde{\alpha}^w_{\tilde{g}_w(i)}-\alpha^0_{g_0(i)})'\Sigma_T^{-1}(\tilde{\alpha}^w_{\tilde{g}_w(i)}-\alpha^0_{g_0(i)})}_{=\tilde{Q}_w(\tilde{\alpha}_w,\tilde{\gamma}_w)-\tilde{Q}_w(\alpha_0,\gamma_0)}>\epsilon\right)\leq P\left(\left|\tilde{Q}_w(\tilde{\alpha}_w,\tilde{\gamma}_w)-\tilde{Q}_w(\alpha_0,\gamma_0)\right|>\epsilon\right). \tag{48}$$

Also, by virtue of (44), we have that

$$\tilde{Q}_w(\tilde{\alpha}_w, \tilde{\gamma}_w) = \hat{Q}_w(\tilde{\alpha}_w, \tilde{\gamma}_w) + o_p(1) \leq \hat{Q}_w(\alpha_0, \gamma_0) + o_p(1) = \tilde{Q}_w(\alpha_0, \gamma_0) + o_p(1),$$

which shows that the upper bound in (48) shrinks to zero, as desired.

For the proof of the result in (9), define

$$\tilde{\tau}_g^w := \Sigma_T^{-1/2} \tilde{\alpha}_g^w = (\tilde{\tau}_{g1}^w, \ldots, \tilde{\tau}_{gT}^w)' \text{ and } \tau_g^0 := \Sigma_T^{-1/2} \alpha_g^0 = (\tau_{g1}^0, \ldots, \tau_{gT}^0)'$$

for each $g \in \{1, \ldots, G\}$, as the rescaled variables of $\tilde{\alpha}_g^w$ and α_g^0, respectively. Then, (9) can be equivalently written as

$$\min_{\breve{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\tilde{\tau}_{\breve{g}t}^w - \tau_{gt}^0)^2 = o_p(1) \text{ for all } g \in \{1, \ldots, G\}. \tag{49}$$

We have that

$$\begin{aligned} &\frac{1}{nT} \sum_{i=1}^{n} \left(\min_{\breve{g} \in \{1,\ldots,G\}} \sum_{t=1}^{T} 1(g_0(i) = g)(\tilde{\tau}_{\breve{g}t}^w - \tau_{gt}^0)^2 \right) \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} 1(g_0(i) = g) \right) \left(\min_{\breve{g} \in \{1,\ldots,G\}} \frac{1}{T} \sum_{t=1}^{T} (\tilde{\tau}_{\breve{g}t}^w - \tau_{gt}^0)^2 \right), \end{aligned} \tag{50}$$

which, together with Assumption 3-(b), implies that (49) holds if the term in (50) is $o_p(1)$ for all $g \in \{1, \ldots, G\}$. This is indeed the case because

$$\begin{aligned} &\frac{1}{nT} \sum_{i=1}^{n} \left(\min_{\breve{g} \in \{1,\ldots,G\}} \sum_{t=1}^{T} 1(g_0(i) = g)(\tilde{\tau}_{\breve{g}t}^w - \tau_{gt}^0)^2 \right) \\ &\leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} 1(g_0(i) = g)(\tilde{\tau}_{\tilde{g}_w(i),t}^w - \tau_{g_0(i),t}^0)^2 \\ &\leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (\tilde{\tau}_{\tilde{g}_w(i),t}^w - \tau_{g_0(i)t}^0)^2 = \frac{1}{nT} \sum_{i=1}^{n} (\tilde{\alpha}_{\tilde{g}_w(i)}^w - \alpha_{g_0(i)}^0)' \Sigma_T^{-1} (\tilde{\alpha}_{\tilde{g}_w(i)}^w - \alpha_{g_0(i)}^0) \\ &= o_p(1) \end{aligned}$$

for all $g \in \{1, \ldots, G\}$, where the last equation follows from the result in (8). We next show that the permutation (relabeling) of the estimated group membership $\sigma(\cdot)$, as defined in (7), is the one-to-one function with probability approaching one as $n, T \to \infty$. By triangle

inequality, we have that for any $g, h \in \{1, \ldots, G\}$ such that $g \neq h$,

$$\begin{aligned}\sqrt{\frac{1}{T}(\alpha_g^0 - \alpha_h^0)'\Sigma_T^{-1}(\alpha_g^0 - \alpha_h^0)} &= \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{gt}^0 - \tau_{ht}^0)^2\right)^{1/2} \\ &\leq \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{gt}^0 - \tilde{\tau}_{\sigma(g),t}^w)^2\right)^{1/2} + \left(\frac{1}{T}\sum_{t=1}^{T}(\tilde{\tau}_{\sigma(g),t}^w - \tilde{\tau}_{\sigma(h),t}^w)^2\right)^{1/2} \\ &\quad + \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{ht}^0 - \tilde{\tau}_{\sigma(h),t}^w)^2\right)^{1/2},\end{aligned}$$

which is equal to

$$\begin{aligned}\left(\frac{1}{T}\sum_{t=1}^{T}(\tilde{\tau}_{\sigma(g),t}^w - \tilde{\tau}_{\sigma(h),t}^w)^2\right)^{1/2} &\geq \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{gt}^0 - \tau_{ht}^0)^2\right)^{1/2} \qquad (51) \\ &\quad - \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{gt}^0 - \tilde{\tau}_{\sigma(g),t}^w)^2\right)^{1/2} - \left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{ht}^0 - \tilde{\tau}_{\sigma(h),t}^w)^2\right)^{1/2}, \qquad (52)\end{aligned}$$

where the right-hand side of the inequality converges in probability to $(c_{gh})^{1/2}$ by Assumption 3-(b). Also, the two terms in (49) are $o_p(1)$ from the result in (9). As a result, we have that

$$\frac{1}{T}\sum_{t=1}^{T}(\tilde{\tau}_{\sigma(g),t}^w - \tilde{\tau}_{\sigma(h),t}^w)^2 = \frac{1}{T}(\tilde{\alpha}_{\sigma(g)}^w - \tilde{\alpha}_{\sigma(h)}^w)'\Sigma_T^{-1}(\tilde{\alpha}_{\sigma(g)}^w - \tilde{\alpha}_{\sigma(h)}^w) > c_{gh} + o_p(1),$$

for some $c_{gh} > 0$ and $g \neq h$, which implies that the mapping $\sigma(\cdot)$ is one-to-one with probability approaching one. ■

Proof of Theorem 2. We first note that the result in Theorem 1 implies that there exists a well-defined inverse mapping $\sigma^{-1}(\cdot)$ with probability approaching one. For this reason, without loss of generality, we can assume that $\sigma(g) = g$, which is equivalent to relabeling the elements of the true parameter (α_{gt}) in the original parameter space $\mathcal{A}^{GT}$ using the values of $\sigma(g)$. Now, let $\mathcal{N}_\eta$ denote the set of parameters in $\mathcal{A}^{GT}$ that satisfy

$$\frac{1}{T}(\alpha_g - \alpha_g^0)'\Sigma_T^{-1}(\alpha_g - \alpha_g^0) = \frac{1}{T}\sum_{t=1}^{T}(\tau_{gt} - \tau_{gt}^0)^2 < \eta \text{ for all } g \in \{1, \ldots, G\},$$

where $\eta > 0$ is arbitrarily small number, $\tau_g = \Sigma_T^{-1/2}\alpha_g = (\tau_{g1}, \ldots, \tau_{gT})'$, and $\tau_g^0 = \Sigma_T^{-1/2}\alpha_g^0 = (\tau_{g1}^0, \ldots, \tau_{gT}^0)'$. We then have that

$$\begin{aligned}&P\left(\tilde{g}_w(i) \neq g_0(i) \text{ for some } i \in \{1, \ldots, n\}\right) \\ &\leq P\left(\tilde{\alpha}_w \notin \mathcal{N}_\eta\right) + n \sup_{i \in \{1,\ldots,n\}} P\left(\tilde{\alpha}_w \in \mathcal{N}_\eta \text{ and } \tilde{g}_w(i) \neq g_0(i)\right),\end{aligned}$$

and $P\left(\tilde{\alpha}_w \notin \mathcal{N}_\eta\right) = o(1)$ from the result (9) in Theorem 1. Also, with $\tilde{g}_w(i) = \tilde{g}_w(i; \tilde{\alpha}^w)$, we have that

$$n \sup_{i \in \{1,\ldots,n\}} P\left(\tilde{\alpha}_w \in \mathcal{N}_\eta \text{ and } \tilde{g}_w(i) \neq g_0(i)\right) \leq n \sup_{i \in \{1,\ldots,n\}} E\left[1\left(\tilde{\alpha}_w \in \mathcal{N}_\eta\right) \cdot 1\left(\tilde{g}_w(i; \tilde{\alpha}^w) \neq g_0(i)\right)\right].$$

For any $\alpha \in \mathcal{N}$, we can reexpress the term $1\left(\tilde{g}_w(i;\alpha) \neq g_0(i)\right)$ by

$$1\{\tilde{g}_w(i;\alpha) \neq g_0(i)\} = \underbrace{\sum_{g=1}^{G} 1\{g_0(i) \neq g\} \cdot 1\{\tilde{g}_w(i;\alpha) = g\}}_{\text{whether } i \text{ is misclassified from its true membership in group } g}.$$

Also, from the definition of $\tilde{g}_w(i;\alpha)$ in (3), the following inequality holds for all $g \in \{1,\ldots,G\}$:

$$1\{\tilde{g}_w(i;\alpha) = g\} \leq 1\left\{(y_i - \alpha_g)'\Sigma_T^{-1}(y_i - \alpha_g) \leq (y_i - \alpha_{g_0(i)})'\Sigma_T^{-1}(y_i - \alpha_{g_0(i)})\right\}.$$

After some algebraic manipulation, the upper bound in the above inequality can be further bounded by

$$\begin{aligned}
&\leq 1\left\{(\alpha_{g_0(i)} - \alpha_g)'\Sigma_T^{-1}(\alpha_{g_0(i)} - \alpha_g) + 2(y_i - \alpha_{g_0(i)})'\Sigma_T^{-1}(\alpha_{g_0(i)} - \alpha_g) \leq 0\right\} \\
&\leq 1\left\{(\alpha_{g_0(i)} - \alpha_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_{g_0(i)} - \frac{(\alpha_{g_0(i)} + \alpha_g)}{2}\right) \leq 0\right\},
\end{aligned}$$

which implies that

$$1\{\tilde{g}_w(i;\alpha) \neq g_0(i)\} \leq \sum_{g=1}^{G} Z_{ig}(\alpha),$$

where

$$Z_{ig}(\alpha) = 1\{g_0(i) \neq g\} \cdot 1\left\{(\alpha_{g_0(i)} - \alpha_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_{g_0(i)} - \frac{(\alpha_{g_0(i)} + \alpha_g)}{2}\right) \leq 0\right\}.$$

We aim to bound $Z_{ig}(\alpha)$ by a quantity that does not depend on $\alpha \in \mathcal{N}_\eta$. For each $i \in \{1,\ldots,n\}$ and $g \in \{1,\ldots,G\}$, we have that

$$\begin{aligned}
\sup_{\alpha \in \mathcal{N}_\eta} Z_{ig}(\alpha) &= 1\{g_0(i) = h, h \neq g\} \cdot 1\left\{(\alpha_h - \alpha_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_h - \frac{(\alpha_h + \alpha_g)}{2}\right) \leq 0\right\} \\
&\leq \max_{h \neq g} 1\left\{(\alpha_h - \alpha_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_h - \frac{(\alpha_h + \alpha_g)}{2}\right) \leq 0\right\}. \qquad (53)
\end{aligned}$$

For $\tilde{g} \neq g$, let us define

$$A_T = (\alpha_h - \alpha_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_h - \frac{(\alpha_h + \alpha_g)}{2}\right) - (\alpha^0_h - \alpha^0_g)'\Sigma_T^{-1}\left(v_i + \alpha^0_h - \frac{(\alpha^0_h + \alpha^0_g)}{2}\right).$$

Then, we have that

$$
\begin{aligned}
&1\left\{(\alpha_h-\alpha_g)'\Sigma_T^{-1}\left(v_i+\alpha_h^0-\frac{(\alpha_h+\alpha_g)}{2}\right)\le 0\right\}\\
&\le 1\left\{(\alpha_h-\alpha_g)'\Sigma_T^{-1}\left(v_i+\alpha_h^0-\frac{(\alpha_h+\alpha_g)}{2}\right)-A_T\le |A_T|\right\}\\
&= 1\left\{(\alpha_h^0-\alpha_g^0)'\Sigma_T^{-1}\left(v_i+\alpha_h^0-\frac{(\alpha_h^0+\alpha_g^0)}{2}\right)\le |A_T|\right\}. \qquad (54)
\end{aligned}
$$

To formulate the upper bound for the quantity $|A_T|$, note that $|A_T|\le |A_{1T}|+|A_{2T}|$, where

$$
\begin{aligned}
A_{1T} &= (\alpha_h-\alpha_g)'\Sigma_T^{-1}v_i-(\alpha_h^0-\alpha_g^0)'\Sigma_T^{-1}v_i\\
A_{2T} &= (\alpha_h-\alpha_g)'\Sigma_T^{-1}\left(\alpha_h^0-\frac{(\alpha_h+\alpha_g)}{2}\right)-(\alpha_h^0-\alpha_g^0)'\Sigma_T^{-1}\left(\alpha_h^0-\frac{(\alpha_h^0+\alpha_g^0)}{2}\right).
\end{aligned}
$$

For $|A_T|$, we use the Cauchy–Schwarz inequality and obtain that for all $\alpha\in\mathcal{N}_\eta$,

$$
\begin{aligned}
|A_{1T}| &\le \left(\{(\alpha_h-\alpha_h^0)'\Sigma_T^{-1}(\alpha_h-\alpha_h^0)\}^{1/2}+\{(\alpha_g-\alpha_g^0)'\Sigma_T^{-1}(\alpha_g-\alpha_g^0)\}^{1/2}\right)\cdot\left(v_i'\Sigma_T^{-1}v_i\right)^{1/2}\\
&\le (2T\sqrt{\eta})\cdot\left(\frac{v_i'\Sigma_T^{-1}v_i}{T}\right)^{1/2}.
\end{aligned}
$$

Also, we have that

$$
\begin{aligned}
|A_{2T}| &\le \left|(\alpha_h-\alpha_g)'\Sigma_T^{-1}\left(\alpha_h^0-\frac{(\alpha_h+\alpha_g)}{2}\right)-(\alpha_h^0-\alpha_g^0)'\Sigma_T^{-1}\left(\alpha_h^0-\frac{(\alpha_h^0+\alpha_g^0)}{2}\right)\right|\\
&= \left|(\alpha_h-\alpha_g-(\alpha_h^0-\alpha_g^0))'\Sigma_T^{-1}\left(\alpha_h^0-\frac{(\alpha_h^0+\alpha_g^0)}{2}\right)+(\alpha_h-\alpha_g)'\Sigma_T^{-1}\left(\frac{(\alpha_h^0+\alpha_g^0)}{2}-\frac{(\alpha_h+\alpha_g)}{2}\right)\right|\\
&\le \frac{1}{2}(\alpha_h-\alpha_h^0)'\Sigma_T^{-1}(\alpha_h^0-\alpha_g^0)+\frac{1}{2}(\alpha_g-\alpha_g^0)'\Sigma_T^{-1}(\alpha_h^0-\alpha_g^0)+\frac{1}{2}(\alpha_h-\alpha_g)'\Sigma_T^{-1}(\alpha_h^0-\alpha_h) \qquad (55)\\
&\quad+\frac{1}{2}(\alpha_h-\alpha_g)'\Sigma_T^{-1}(\alpha_g^0-\alpha_g). \qquad (56)
\end{aligned}
$$

Applying the Cauchy–Schwarz inequality along with Assumptions 1 and 2, we can bound each term in (55) and (56) for $\alpha\in\mathcal{N}_\eta$ by $TC\sqrt{\eta}/2$, where the constant C does not depend on η. Combining these findings to the inequalities shown in (53) and (54), we obtain that

$$
\begin{aligned}
\sup_{\alpha\in\mathcal{N}_\eta} Z_{ig}(\alpha) &\le \max_{h\neq g} 1\left\{(\alpha_h^0-\alpha_g^0)'\Sigma_T^{-1}\left(v_i+\alpha_h^0-\frac{(\alpha_h^0+\alpha_g^0)}{2}\right)\le |A_{1T}|+|A_{2T}|\right\}\\
&\le \tilde{Z}_{ig},
\end{aligned}
$$

where $\tilde{Z}_{ig} \in \{0,1\}$ is defined as

$$\tilde{Z}_{ig} = \max_{h \neq g} 1\left\{ (\alpha_h^0 - \alpha_g^0)'\Sigma_T^{-1} v_i \leq -\frac{1}{2}(\alpha_h^0 - \alpha_g^0)'\Sigma_T^{-1}(\alpha_h^0 - \alpha_g^0) + 2T\sqrt{\eta} \cdot \left(\frac{v_i'\Sigma_T^{-1} v_i}{T} \right)^{1/2} + 2TC\sqrt{\eta} \right\}.$$

In summary so far, we have that

$$\begin{aligned}
&P\left(\tilde{g}_w(i) \neq g_0(i) \text{ for some } i \in \{1,\ldots,n\}\right) \\
&\leq P\left(\tilde{\alpha}_w \notin \mathcal{N}_\eta\right) + n \sup_{i \in \{1,\ldots,n\}} P\left(\tilde{\alpha}_w \in \mathcal{N}_\eta \text{ and } \tilde{g}_w(i) \neq g_0(i)\right) \\
&\leq n \sup_{i \in \{1,\ldots,n\}} E\left[1\left(\tilde{\alpha}_w \in \mathcal{N}_\eta\right) \cdot 1\left(\tilde{g}_w(i;\tilde{\alpha}_w) \neq g_0(i)\right)\right] + o(1) \\
&\leq n \cdot \left(\sup_{i \in \{1,\ldots,n\}} \sum_{g=1}^{G} P(\tilde{Z}_{ig} = 1) \right) + o(1). \qquad (57)
\end{aligned}$$

For each $g \in \{1,\ldots,G\}$, we have that

$$\begin{aligned}
P(\tilde{Z}_{ig} = 1) &\leq \sum_{\tilde{g} \neq g} P\left((\alpha_{\tilde{g}}^0 - \alpha_g^0)'\Sigma_T^{-1} v_i \leq -\frac{1}{2}(\alpha_{\tilde{g}}^0 - \alpha_g^0)'\Sigma_T^{-1}(\alpha_{\tilde{g}}^0 - \alpha_g^0) + 2T\sqrt{\eta} \cdot \left(\frac{v_i'\Sigma_T^{-1} v_i}{T} \right)^{1/2} + 2TC\sqrt{\eta} \right) \\
&\leq \sum_{\tilde{g} \neq g} P\left(\frac{1}{T}(\alpha_{\tilde{g}}^0 - \alpha_g^0)'\Sigma_T^{-1}(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \frac{c_{g\tilde{g}}}{2} \right) + \sum_{\tilde{g} \neq g} P\left(\frac{v_i'\Sigma_T^{-1} v_i}{T} \geq \tilde{M} \right) \qquad (58) \\
&\quad + \sum_{\tilde{g} \neq g} P\left((\alpha_{\tilde{g}}^0 - \alpha_g^0)'\Sigma_T^{-1} v_i \leq -T\frac{c_{g\tilde{g}}}{4} + 2T\tilde{M}^{1/2}\sqrt{\eta} + 2TC\sqrt{\eta} \right), \qquad (59)
\end{aligned}$$

where $c_{g\tilde{g}}$ are positive constants defined in Assumptions 1 and 3-(b), respectively, and $\tilde{M} > 1$ is a constant that does not depend on η and T. We derive probability bounds for each of the three quantities in (58) and (59).

For the first probability bound component in (58),

$$P\left(\frac{1}{T}(\alpha_{\tilde{g}}^0 - \alpha_g^0)'\Sigma_T^{-1}(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \frac{c_{g\tilde{g}}}{2} \right) = P\left(\frac{1}{T}\sum_{t=1}^{T} (\tau_{\tilde{g}t}^0 - \tau_{gt}^0)^2 \leq \frac{c_{g\tilde{g}}}{2} \right),$$

note that $\{\tau_{gt}^0\}_t$ satisfies the exponentially decaying mixing rate specified in Assumption 4-(a). By Assumption 3-(b), we have

$$\frac{1}{T}E[\sum_{t=1}^{T}(\tau_{\tilde{g}t}^0 - \tau_{gt}^0)^2] > \frac{2}{3}c_{g\tilde{g}}$$

holds for sufficiently large T. As a result, we have for sufficiently large T :

$$
\begin{aligned}
P\left(\frac{1}{T}\sum_{t=1}^{T}(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2\leq\frac{c_{g\tilde{g}}}{2}\right) &\leq P\left(\frac{1}{T}\sum_{t=1}^{T}\left\{(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2-E[(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2]\right\}\leq-\frac{c_{g\tilde{g}}}{6}\right) \\
&\leq P\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left\{(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2-E[(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2]\right\}\right|\geq\frac{c_{g\tilde{g}}}{6}\right) \\
&= o\left(\frac{1}{T^\delta}\right), \qquad (60)
\end{aligned}
$$

where the last result follows from applying Lemma 5 with $Z_t=(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2-E[(\tau_{\tilde{g}t}^0-\tau_{gt}^0)^2]$. Similarly, by the construction of u_{it} from $u_i=\Sigma_T^{-1/2}v_i=(u_{i1},\ldots,u_{iT})'$, it is straightforward to verify that $E[(u_{it})^2]=1$, and that $(u_{it})^2-1$ is a zero-mean mixing process satisfying the conditions in Assumption 4-(a) and (b). These conditions allow us to apply Lemma 5 with $Z_t=(u_{it})^2-E[(u_{it})^2]$ to conclude that

$$
P\left(\frac{1}{T}\sum_{t=1}^{T}\left((u_{it})^2-E[(u_{it})^2]\right)\geq(M-1)\right)=o\left(\frac{1}{T^\delta}\right).
$$

Finally, let c denote the minimum of $c_{g\tilde{g}}$ over $g\neq\tilde{g}$, and take

$$
\eta\leq\left(\frac{c}{32((\tilde{M})^{1/2}+C)}\right)^2,
$$

which leads to

$$
\begin{aligned}
&P\left((\alpha_{\tilde{g}}^0-\alpha_g^0)'\Sigma_T^{-1}v_i\leq-T\frac{c_{g\tilde{g}}}{4}+2T\tilde{M}^{1/2}\sqrt{\eta}+2TC\sqrt{\eta}\right) \\
&=P\left(\frac{(\alpha_{\tilde{g}}^0-\alpha_g^0)'\Sigma_T^{-1}v_i}{T}\leq-\frac{c_{g\tilde{g}}}{4}+2\sqrt{\tilde{M}}\sqrt{\eta}+2C\sqrt{\eta}\right) \\
&\leq P\left(\left|\frac{\sum_{t=1}^{T}(\tau_{gt}^0-\tau_{\tilde{g}t}^0)u_{it}}{T}\right|\geq\frac{c}{8}\right). \qquad (61)
\end{aligned}
$$

Using the same reasoning and the conditions in Assumption 4 as before, we can apply Lemma 5 with $Z_t=(\tau_{gt}^0-\tau_{\tilde{g}t}^0)u_{it}$, which implies that the upper bound in (61) also shrinks to zero at the rate $o(T^{-\delta})$. Combining these findings with (58) and (59), we conclude that

$$
\sup_{i\in\{1,\ldots,n\}}\sum_{g=1}^{G}P(\tilde{Z}_{ig}=1)\leq G(G-1)\times o\left(\frac{1}{T^\delta}\right)=o\left(\frac{1}{T^\delta}\right),
$$

which, together with (57), implies that

$$
P\left(\tilde{g}_w(i)\neq g_0(i)\text{ for some }i\in\{1,\ldots,n\}\right)=o(1)+o\left(\frac{n}{T^\delta}\right)
$$

for all $\delta > 0$, as desired. ■

Proof of Lemma 3. We first prove the result in (23). Define $\tilde{\Sigma}_{B_T} = \tilde{\Sigma}_T \star W_T$ and $\Sigma_{B_T} = \Sigma_T \star W_T$, where $\tilde{\Sigma}_T = n^{-1}\sum_{i=1}^n v_i v_i'$ is the sample covariance of unobserved $v_i = (v_{i1}, \dots v_{iT})'$. From triangle inequality, we have that

$$\lambda(\hat{\Sigma}_{B_T} - \Sigma_T) \leq \underbrace{\lambda(\hat{\Sigma}_{B_T} - \tilde{\Sigma}_{B_T})}_{:=(\text{I})} + \underbrace{\lambda(\tilde{\Sigma}_{B_T} - \Sigma_{B_T})}_{:=(\text{II})} + \underbrace{\lambda(\Sigma_{B_T} - \Sigma_T)}_{:=(\text{III})}.$$

For the term (III), we note that

$$\Sigma_{B_T} - \Sigma_T = \left[\left\{K\left(\frac{s-t}{B_T}\right) - 1\right\} E[v_{is}v_{it}]\right]_{1\leq s,t\leq T}.$$

Applying Lemma 7, we obtain that

$$\begin{aligned}\lambda(\Sigma_{B_T} - \Sigma) &\leq \max_{1\leq t\leq T}\sum_{s=1}^{T} |K\left((s-t)/B_T\right) - 1||E[v_{is}v_{it}]| \qquad (62)\\ &\leq \max_{1\leq t\leq T}\sum_{h=-T+1}^{T-1} |K\left(h/B_T\right) - 1||E[v_t v_{it-h}]|,\end{aligned}$$

where the second inequality follows from the equation

$$\max_{1\leq t\leq T}\sum_{s=1}^{T} |K\left((s-t)/B_T\right)-1||E[v_{is}v_{it}]| = \max_{1\leq t\leq T}\sum_{h=-T+1}^{T-1} |K\left(h/B_T\right)-1||E[v_t v_{it-h}]|\cdot 1(h+1\leq t).$$

For the upper bound in (62), we use Assumption 5 and obtain that

$$\begin{aligned}\max_{1\leq t\leq T}\sum_{h=-T+1}^{T-1} |K\left(h/B_T\right) - 1||E[v_t v_{it-h}]| &\leq \frac{1}{B_T^q}\max_{1\leq t\leq T}\sum_{h=-T+1}^{T-1}\left(\underbrace{\frac{|K\left(h/B_T\right)-1|}{(h/B_T)^q}}_{\to k_{q0} \text{ as } T\to\infty}\right)|E[v_{it}v_{it-h}]||h|^q\\ &\leq \left(\frac{k_{q0}}{B_T^q}\right)\sup_t\sum_{h=-\infty}^{\infty}|h|^q|E[v_{it}v_{it-h}]| = O\left(\frac{1}{B_T^q}\right),\end{aligned}$$

and this leads to conclude that the term (III) is $O_p(B_T^{-q})$. For the term (II), we again apply Lemma 7 and obtain that

$$\begin{aligned}\lambda(\tilde{\Sigma}_{B_T} - \Sigma_{B_T}) &\leq \max_{1\leq t\leq T}\sum_{s\in\{|s-t|\leq B_T\}} |K\left((s-t)/B_T\right)| \times \left|\frac{1}{n}\sum_{i=1}^n v_{is}v_{it} - E[v_{is}v_{it}]\right|\\ &= \max_{1\leq t\leq T}\sum_{h=-B_T}^{B_T} |K\left(h/B_T\right)| \cdot \left|\frac{1}{n}\sum_{i=1}^n v_{it}v_{it-h} - E[v_{it}v_{it-h}]\right|\cdot 1(h\leq t-1)\\ &\leq B_T \max_{|h|\leq B_T}\max_{h+1\leq t\leq T}|K\left(h/B_T\right)| \cdot \left|\frac{1}{n}\sum_{i=1}^n v_{it}v_{it-h} - E[v_{it}v_{it-h}]\right|.\end{aligned}$$

Using Bonferroni's method, we have that

$$P\left(\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}v_{it}v_{it-h}-E[v_{it}v_{it-h}]\right|>\epsilon\right)$$
$$\leq TB_T\cdot\max_{|h|\leq B_T}\max_{1\leq t\leq T}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(v_{it}v_{it-h}-E[v_{it}v_{it-h}])\right|>\epsilon\right)$$

for any $\epsilon>0$. Also, under Assumption 4-(b) with $\Sigma_T=I_T$, which imposes exponentially decaying tail probabilities for $\{v_{it}\}$ and $\{v_{it-h}\}$, we can apply the result in Lemma 8 and conclude that

$$P(|v_{it}v_{it-h}|>\epsilon)\leq\exp\left(1-\left(\frac{\epsilon}{b_3}\right)^{r_3}\right)$$

for some $r_3\in(0,1)$ and $b_3>0$. Next, applying Bernstein's inequality from Lemma 6 to $\{v_{it}v_{it-h}-E[v_{it}v_{it-h}]\}$ by setting $c^2=2C$ for some $C>0$, and $\alpha_{n,T}=\sqrt{\log(B_TT)/n}$, we obtain

$$TB_T\cdot\max_{|h|\leq B_T}\max_{1\leq t\leq T}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(v_{it}v_{it-h}-E[v_{it}v_{it-h}])\right|>c\cdot\alpha_{n,T}\right)$$
$$\leq TB_T\cdot\max_{|h|\leq B_T}\max_{1\leq t\leq T}\exp\left(-\frac{nc^2(\alpha_{n,T})^2}{C}\right)$$
$$\leq TB_T\cdot\exp\left(-\frac{nc^2(\alpha_{n,T})^2}{C}\right)\leq\exp(\log(TB_T))\cdot\exp\left(-\frac{nc^2(\alpha_{n,T})^2}{C}\right)$$
$$\preceq\exp(\log(TB_T))\cdot\exp(-2\log(TB_T))=O\left(\frac{1}{TB_T}\right),$$

which leads to

$$\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}(v_{it}v_{it-h}-E[v_{it}v_{it-h}])\right|=O_p\left(\sqrt{\frac{\log(B_TT)}{n}}\right).\tag{63}$$

From this result, we conclude that

$$\lambda(\tilde{\Sigma}_{B_T}-\Sigma_{B_T})=O_p\left(B_T\sqrt{\frac{\log(B_TT)}{n}}\right).$$

For the term (I), we employ similar arguments to those in the previous two terms, we have that

$$\lambda(\hat{\Sigma}_{B_T}-\tilde{\Sigma}_{B_T})\leq\max_{1\leq t\leq T}\sum_{s\in\{|s-t|\leq B_T\}}|K\left((s-t)/B_T\right)|\times\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{is}\hat{v}_{it}-v_{is}v_{it})\right|$$
$$\leq B_T\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}|K\left(h/B_T\right)|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}\hat{v}_{it-h}-v_{it}v_{it-h}\right|.\tag{64}$$

For each h,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}\hat{v}_{it-h}-v_{it}v_{it-h}\right| = \left|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{it}\hat{v}_{it-h}-\hat{v}_{it}v_{it-h}+\hat{v}_{it}v_{it-h}-v_{it}v_{it-h})\right|$$
$$\leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}(\hat{v}_{it-h}-v_{it-h})\right|}_{:=I_{11}} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_{it}-v_{it})v_{it-h}\right|}_{:=I_{12}}.$$

Using $\hat{v}_{is}-v_{is} = y_{is}-\hat{\alpha}_{\hat{g}(i),s}-(y_{is}-\alpha^0_{g_0(i),s}) = \alpha^0_{g_0(i),s}-\hat{\alpha}_{\hat{g}(i),s}$, and we can write the term I_{11} as

$$\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}(\hat{v}_{it-h}-v_{it-h})\right| = \max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}\hat{v}_{it-h}\right| + \max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{it}v_{it-h}\right|$$
$$= \underbrace{\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t}-\hat{\alpha}_{\hat{g}(i),t})(\alpha^0_{g_0(i),t-h}-\hat{\alpha}_{\hat{g}(i),t-h})\right|}_{:=I_{11,(a)}}$$
$$+ \underbrace{\max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t}-\hat{\alpha}_{\hat{g}(i),t})v_{it-h}\right|}_{:=I_{11,(b)}}.$$

Without loss of generality, we let the relabeling mapping $\sigma(\cdot)$ be the identify, i.e., $\sigma(g) = g$. Then, for any $\epsilon > 0$, the term $I_{11,(a)}$ satisfies

$$P\left(I_{11,(a)} > \epsilon\right) \leq P\left(I_{11,(a)} > \epsilon \text{ and } \hat{g}(i) = g_0(i) \text{ for all } i \in \{1,\ldots,n\}\right)$$
$$+ P(\hat{g}(i) \neq g_0(i) \text{ for some } i \in \{1,\ldots,n\}). \tag{65}$$

Under Assumption 6, we can apply the result in Theorem 2, by replacing $\tilde{g}_w(i)$ with $\hat{g}(i)$, which implies that

$$P(\hat{g}(i) \neq g_0(i) \text{ for some } i \in \{1,\ldots,n\}) = o(1) + o\left(\frac{n}{T^\delta}\right). \tag{66}$$

When $\hat{g}(i) = g_0(i)$ holds for all $i \in \{1,\ldots,n\}$, $\hat{\alpha}_{gt}$ can be expressed as

$$\hat{\alpha}_{gt} = \frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} y_{it} = \alpha^0_{gt} + \frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it},$$

where $n_g = \sum_{i=1}^{n} 1(g_0(i) = g)$. Also,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t} - \hat{\alpha}_{\hat{g}(i),t})(\alpha^0_{g_0(i),t-h} - \hat{\alpha}_{\hat{g}(i),t-h}) &= \frac{1}{n}\sum_{g=1}^{G}\sum_{\{i:g_0(i)=g\}}(\alpha^0_{gt} - \hat{\alpha}_{gt})(\alpha^0_{g,t-h} - \hat{\alpha}_{g,t-h}) \\ &= \sum_{g=1}^{G}\frac{n_g}{n}(\alpha^0_{gt} - \hat{\alpha}_{gt})(\alpha^0_{g,t-h} - \hat{\alpha}_{g,t-h}), \\ &= \sum_{g=1}^{G}\frac{n_g}{n}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it}\right)\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it-h}\right).\end{aligned}$$

Applying the Cauchy–Schwarz inequality, we obtain:

$$\begin{aligned}&\left(\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t} - \hat{\alpha}_{\hat{g}(i),t})(\alpha^0_{g_0(i),t-h} - \hat{\alpha}_{\hat{g}(i),t-h})\right)^2 \\ &\leq \sum_{g=1}^{G}\frac{n_g}{n}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it}\right)^2 \times \sum_{g=1}^{G}\frac{n_g}{n}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it-h}\right)^2 \\ &\leq G \cdot \max_{1\leq g\leq G}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it}\right)^2 \cdot \max_{1\leq g\leq G}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it}\right)^2.\end{aligned}$$

Using similar arguments as in (63), we can show that

$$\max_{|h|\leq B_T}\max_{1\leq t\leq T}\max_{1\leq g\leq G}\left(\frac{1}{n_g}\sum_{\{i:g_0(i)=g\}} v_{it-h}\right)^2 = O_p\left(\frac{\log(B_T T)}{n}\right),$$

as n_g increases to infinity. Together with (65) and (66), this result implies that

$$\begin{aligned}I_{11,(a)} &= \max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t} - \hat{\alpha}_{\hat{g}(i),t})(\alpha^0_{g_0(i),t-h} - \hat{\alpha}_{\hat{g}(i),t-h})\right| \\ &= O_p\left(\frac{\log(B_T T)}{n}\right).\end{aligned}$$

Employing the same reasoning, we can also prove that

$$\begin{aligned}I_{11,(b)} &= \max_{|h|\leq B_T}\max_{h+1\leq t\leq T}\left|\frac{1}{n}\sum_{i=1}^{n}(\alpha^0_{g_0(i),t} - \hat{\alpha}_{\hat{g}(i),t})v_{it-h}\right| \\ &= O_p\left(\frac{\log(B_T T)}{n}\right),\end{aligned}$$

and

$$I_{12} = \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{v}_{it} - v_{it}) v_{it-h} \right| = O_p \left(\frac{\log(B_T T)}{n} \right),$$

which, together with (64), implies that

$$\begin{aligned} \lambda(\hat{\Sigma}_{B_T} - \tilde{\Sigma}_{B_T}) &\preceq B_T \max_{|h| \leq B_T} \max_{1 \leq t \leq T} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{v}_{it} \hat{v}_{it-h} - v_{it} v_{it-h} \right| \\ &= O_p \left(\frac{B_T \log(B_T T)}{n} \right). \end{aligned}$$

Summing up results for each of the three terms in (I)-(III), we have:

$$\begin{aligned} \lambda(\hat{\Sigma}_{B_T} - \Sigma_T) &\leq \underbrace{\lambda(\hat{\Sigma}_{B_T} - \tilde{\Sigma}_{B_T})}_{:=(\mathrm{I})} + \underbrace{\lambda(\tilde{\Sigma}_{B_T} - \Sigma_{B_T})}_{:=(\mathrm{II})} + \underbrace{\lambda(\Sigma_{B_T} - \Sigma_T)}_{:=(\mathrm{III})} \\ &= O_p \left(\frac{B_T \log(T B_T)}{n} \right) + O_p \left(B_T \sqrt{\frac{\log(B_T T)}{n}} \right) + O_p \left(\frac{1}{B_T^q} \right) \\ &= O_p \left(B_T \sqrt{\frac{\log T}{n}} + \frac{1}{B_T^q} \right) = o_p(1), \end{aligned}$$

which is the desired result. To show the second result in (24), we use the triangle inequality and obtain:

$$\begin{aligned} \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) &\leq \lambda(\Sigma_T^{-1} - \hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1} \hat{\Sigma}_{B_T} \Sigma_T^{-1} + \Sigma_T^{-1}) + \lambda(\Sigma_T^{-1} \hat{\Sigma}_{B_T} \Sigma_T^{-1} - \Sigma_T^{-1}) \\ &= \lambda((\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})(\hat{\Sigma}_{B_T} - \Sigma_T)\Sigma_T^{-1}) + \lambda(\Sigma_T^{-1}(\hat{\Sigma}_{B_T} - \Sigma_T)\Sigma_T^{-1}) \\ &\leq \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) \lambda(\hat{\Sigma}_{B_T} - \Sigma_T) \lambda(\Sigma_T^{-1}) + \lambda(\Sigma_T^{-1})^2 \lambda(\hat{\Sigma}_{B_T} - \Sigma_T)) \\ &= \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) O_p \left(B_T \sqrt{\frac{\log T}{n}} + \frac{1}{B_T^q} \right) + O_p \left(B_T \sqrt{\frac{\log T}{n}} + \frac{1}{B_T^q} \right), \end{aligned}$$

which implies the desired result:

$$\begin{aligned} \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})(1 + o_p(1)) &= O_p \left(B_T \sqrt{\frac{\log(B_T T)}{n}} + \frac{1}{B_T^q} \right) \\ &= o_p(1). \end{aligned}$$

■

Proof of Theorem 4. We first show that (25) holds uniformly over uniformly over $(\alpha, g(\cdot)) \in \mathcal{A}^G \times \Gamma_G$. To this end, we express the feasible K-means objective function as

$$\begin{aligned}\frac{1}{nT}\sum_{i=1}^{n}(y_i - \alpha_{g(i)})'\hat{\Sigma}_{B_T}^{-1}(y_i - \alpha_{g(i)}) &= \frac{1}{nT}\sum_{i=1}^{n}(v_i + (\alpha_{g(i)}^0 - \alpha_{g(i)}))'(v_i + (\alpha_{g(i)}^0 - \alpha_{g(i)})) \\ &= \frac{1}{nT}\sum_{i=1}^{n} v_i'\hat{\Sigma}_{B_T}^{-1}v_i + \frac{2}{nT}\sum_{i=1}^{n}(\alpha_{g(i)}^0 - \alpha_{g(i)})'\hat{\Sigma}_{B_T}^{-1}v_i \\ &\quad + \frac{1}{nT}\sum_{i=1}^{n}(\alpha_{g(i)}^0 - \alpha_{g(i)})'\hat{\Sigma}_{B_T}^{-1}(\alpha_{g(i)}^0 - \alpha_{g(i)}).\end{aligned}$$

The result then follows once we prove that the following hold uniformly over $(\alpha, g(\cdot)) \in \mathcal{A}^G \times \Gamma_G$:

$$\frac{1}{nT}\sum_{i=1}^{n} v_i'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i = o_p(1); \tag{67}$$

$$\frac{1}{nT}\sum_{i=1}^{n}(\alpha_{g(i)}^0 - \alpha_{g(i)})'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i = o_p(1); \tag{68}$$

$$\frac{1}{nT}\sum_{i=1}^{n}(\alpha_{g(i)}^0 - \alpha_{g(i)})'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})(\alpha_{g(i)}^0 - \alpha_{g(i)}) = o_p(1). \tag{69}$$

To show (67), let $\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1} = Q_T R_T Q_T'$, where $Q_T Q_T' = I_T$ and $R_T = \text{Diag}(r_t)_{1\le t\le T}$, with the eigenvalues ordered so that $r_1 \ge \cdots \ge r_T$. Then, the spectral norm of the symmetric matrix $\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}$ satisfies $\lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) = \max_{1\le t\le T}|r_t|$. This implies that

$$\begin{aligned}v_i'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i = v_i'Q_T R_T Q_T' v_i &= (Q_T'v_i)'R(Q_T'v_i) \\ &\le \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) \cdot \left(\sum_{t=1}^{T} v_{it}^2\right),\end{aligned}$$

which in turn implies

$$\frac{1}{nT}\sum_{i=1}^{n} v_i'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i \le \lambda(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) \cdot \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} v_{it}^2\right) = o_p(1),$$

where the final equality follows from Lemma 3 and Assumption 2. To verify (68), we apply the Cauchy–Schwarz inequality and have:

$$\begin{aligned}\left(\frac{1}{nT}\sum_{i=1}^{n}(\alpha_{g(i)}^0 - \alpha_{g(i)})'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i\right)^2 &\le \left(\frac{1}{nT}\sum_{i=1}^{n}||\alpha_{g(i)}^0 - \alpha_{g(i)})||^2\right) \cdot \left(\frac{1}{nT}\sum_{i=1}^{n}||(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i||^2\right). \\ &\le \frac{C}{nT}\sum_{i=1}^{n}||(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i||^2, \end{aligned} \tag{70}$$

uniformly over $(\alpha, g(\cdot)) \in \mathcal{A}^G \times \Gamma_G$, for some constant $C > 0$. Here, the final inequality follows from Assumption 2-(a). To bound the right-hand side of (70), observe that

$$||(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i||^2 = (Q_T'v_i)'R_T^2(Q_T'v_i) \leq \lambda^2(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) \cdot \left(\sum_{t=1}^{T} v_{it}^2\right).$$

Therefore, the right-hand side of (70) satisfies:

$$\frac{C}{nT}\sum_{i=1}^{n} ||(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})v_i||^2 \leq \lambda^2(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) \cdot \left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} v_{it}^2\right) = o_p(1).$$

The proof of (69) follows by a similar argument and is omitted.

To show the result in (26), note that Lemma 3 and Assumption 1 imply that there exists $\tilde{\epsilon} > 0$ such that

$$0 < \tilde{\epsilon} \leq \lambda_{\min}(\hat{\Sigma}_{B_T}) \leq \lambda_{\max}(\hat{\Sigma}_{B_T}) \leq 1/\tilde{\epsilon}$$

holds with probability approaching one. This allows us to follow the same line of reasoning as in the proof of the Theorem in (1) and conclude that

$$\frac{1}{nT}\sum_{i=1}^{n}(\hat{\alpha}_{\hat{g}_w(i)}^w - \alpha_{g_0(i)}^0)'\hat{\Sigma}_{B_T}^{-1}(\hat{\alpha}_{\hat{g}_w(i)}^w - \alpha_{g_0(i)}^0) = o_p(1).$$

Moreover, by applying the same argument used to show (50), the result above also holds when $\hat{\Sigma}_{B_T}^{-1}$ is replaced with Σ_T^{-1}, which yields the result in (26). The proof of (27) can be established in a similar manner by repeating the steps in the proof of (9) in Theorem 1, which yields

$$\frac{1}{T}(\hat{\alpha}_{\sigma(g)}^w - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1}(\hat{\alpha}_{\sigma(g)}^w - \alpha_g^0) = o_p(1) \text{ for all } g \in \{1, \ldots, G\}. \tag{71}$$

Applying Lemma 3 together with Assumption 2 then gives the desired result. The one-to-one property of $\sigma(\cdot)$ can be established in a similar manner to the proof of Theorem 1, and we omit the details for brevity.

To show the result in (28), we let $\sigma(g) = g$ without loss of generality. For arbitrarily small $\eta > 0$, we let $\bar{\mathcal{N}}_\eta$ denote the set of parameters in $\mathcal{A}^{GT}$ that satisfy

$$\frac{1}{T}(\alpha_g - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1}(\alpha_g - \alpha_g^0) < \eta \text{ for all } g \in \{1, \ldots, G\}.$$

We then have that

$$\begin{aligned} &P\left(\hat{g}_w(i) \neq g_0(i) \text{ for some } i \in \{1, \ldots, n\}\right) \\ &\leq P\left(\hat{\alpha}^w \notin \bar{\mathcal{N}}_\eta\right) + n \sup_{i \in \{1,\ldots,n\}} P\left(\hat{\alpha}^w \in \bar{\mathcal{N}}_\eta \text{ and } \hat{g}_w(i) \neq g_0(i)\right), \end{aligned} \tag{72}$$

and $P\left(\hat{\alpha}^w \notin \bar{\mathcal{N}}_\eta\right) = o(1)$ from (71). By repeating the same arguments and algebraic steps used in the proof of Theorem 2, we obtain that the second term in the upper-bound probability in (72) satisfies

$$\sup_{i\in\{1,\ldots,n\}} P\left(\hat{\alpha}^w \in \bar{\mathcal{N}}_\eta \text{ and } \hat{g}_w(i) \neq g_0(i)\right) \leq n \cdot \left(\sup_{i\in\{1,\ldots,n\}} \sum_{g=1}^{G} P(\hat{Z}_{ig} = 1)\right) + o(1),$$

where $\hat{Z}_{ig} \in \{0,1\}$ is defined as

$$\hat{Z}_{ig} = \max_{h\neq g} 1\left\{(\alpha_h^0 - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1} v_i \leq -\frac{1}{2}(\alpha_h^0 - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1}(\alpha_h^0 - \alpha_g^0) + 2T\sqrt{\eta}\cdot\left(\frac{v_i'\hat{\Sigma}_{B_T}^{-1} v_i}{T}\right)^{1/2} + 2TC\sqrt{\eta}\right\}.$$

For each $g \in \{1,\ldots,G\}$, we have

$$P(\hat{Z}_{ig} = 1) \leq \sum_{h\neq g} P\left(\frac{1}{T}(\alpha_h^0 - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1}(\alpha_h^0 - \alpha_g^0) \leq \frac{c_{g\tilde{g}}}{2}\right) + \sum_{\tilde{g}\neq g} P\left(\frac{v_i'\hat{\Sigma}_{B_T}^{-1} v_i}{T} \geq \tilde{M}\right) \quad (73)$$

$$+ \sum_{h\neq g} P\left((\alpha_h^0 - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1} v_i \leq -T\frac{c_{g\tilde{g}}}{4} + 2T\tilde{M}^{1/2}\sqrt{\eta} + 2TC\sqrt{\eta}\right), \quad (74)$$

where $c_{g\tilde{g}}$ are positive constants defined in Assumptions 1 and 3-(b), respectively, and $\tilde{M} > 1$ is a constant that does not depend on η and T. We derive probability bounds for each of the three quantities in (73) and (74). The first term on the right-hand side of (73) satisfies

$$\begin{aligned}
&\sum_{h\neq g} P\left(\frac{1}{T}(\alpha_h^0 - \alpha_g^0)'\hat{\Sigma}_{B_T}^{-1}(\alpha_h^0 - \alpha_g^0) \leq \frac{c_{g\tilde{g}}}{2}\right) \\
&\quad \leq \sum_{h\neq g} P\left(\frac{1}{T}(\alpha_h^0 - \alpha_g^0)'\Sigma_T^{-1}(\alpha_h^0 - \alpha_g^0) \leq \frac{c_{g\tilde{g}}}{2} + \eta\right) + \sum_{h\neq g} P\left(\frac{1}{T}(\alpha_h^0 - \alpha_g^0)'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})(\alpha_h^0 - \alpha_g^0) > \eta\right) \\
&\quad = o\left(\frac{1}{T^\delta}\right) + o(1),
\end{aligned}$$

where the final equality follows from (60) in the proof of Theorem 2 and Lemma 3. Similarly, for the second term in (73), we have:

$$\begin{aligned}
\sum_{\tilde{g}\neq g} P\left(\frac{v_i'\hat{\Sigma}_{B_T}^{-1} v_i}{T} \geq \tilde{M}\right) &\leq \sum_{\tilde{g}\neq g} P\left(\frac{v_i'\Sigma_T^{-1} v_i}{T} \geq \frac{\tilde{M}}{2}\right) + \sum_{\tilde{g}\neq g} P\left(\frac{v_i'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1}) v_i}{T} \geq \frac{\tilde{M}}{2}\right) \\
&= o\left(\frac{1}{T^\delta}\right) + o(1).
\end{aligned}$$

Finally, we apply similar arguments to bound the last term in (74), and obtain:

$$\sum_{h \neq g} P\left((\alpha_h^0 - \alpha_g^0)' \Sigma_T^{-1} v_i \leq -T\frac{c_{g\tilde{g}}}{4} + 2T\tilde{M}^{1/2}\sqrt{\eta} + 2TC\sqrt{\eta} + T\sqrt{\eta}\right)$$
$$+ \sum_{h \neq g} P\left(\frac{1}{T}(\alpha_h^0 - \alpha_g^0)'(\hat{\Sigma}_{B_T}^{-1} - \Sigma_T^{-1})(\alpha_h^0 - \alpha_g^0) > \sqrt{\eta}\right)$$
$$= o\left(\frac{1}{T^\delta}\right) + o(1).$$

Combining these bounds with (72), we conclude that

$$P\left(\hat{g}_w(i) \neq g_0(i) \text{ for some } i \in \{1, \ldots, n\}\right) = o(1) + o\left(\frac{n}{T^\delta}\right),$$

which is the desired result. ■

8.3 Tables and Figures

Table 2: Misclassification rates with $n = 180$ and $\psi = 0.25$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0881	0.0899	0.0907	0.0479	0.0806	0.0803
0.25	0.1322	0.1310	0.1321	0.0601	0.1134	0.1120
0.50	0.3405	0.2933	0.2833	0.0743	0.1813	0.1530
0.75	0.6709	0.5694	0.5272	0.0961	0.4287	0.2147
			$T = 50$			
0.00	0.0075	0.0078	0.0078	0.0041	0.0055	0.0055
0.25	0.0141	0.0142	0.0143	0.0057	0.0093	0.0092
0.50	0.0587	0.0485	0.0489	0.0064	0.0130	0.0123
0.75	0.6178	0.4316	0.4287	0.0072	0.0460	0.0241
			$T = 100$			
0.00	0.0001	0.0007	0.0008	0.0000	0.0000	0.0000
0.25	0.0004	0.0022	0.0025	0.0000	0.0001	0.0001
0.50	0.0039	0.0130	0.0197	0.0001	0.0003	0.0003
0.75	0.3438	0.3643	0.3978	0.0001	0.0012	0.0008

Table 3: Misclassification rates with $n = 180$ and $\psi = 0.50$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1143	0.1163	0.1171	0.0603	0.1048	0.1042
0.25	0.1726	0.1729	0.1752	0.0906	0.1544	0.1530
0.50	0.3866	0.3566	0.3562	0.1429	0.2648	0.2461
0.75	0.6664	0.6054	0.5936	0.2201	0.5027	0.3761
			$T = 50$			
0.00	0.0091	0.0094	0.0094	0.0051	0.0071	0.0071
0.25	0.0208	0.0214	0.0216	0.0097	0.0153	0.0152
0.50	0.0921	0.0829	0.0872	0.0165	0.0301	0.0287
0.75	0.6096	0.5154	0.5337	0.0260	0.1129	0.0842
			$T = 100$			
0.00	0.0001	0.0009	0.0009	0.0000	0.0001	0.0001
0.25	0.0009	0.0036	0.0042	0.0001	0.0004	0.0004
0.50	0.0085	0.0283	0.0468	0.0005	0.0013	0.0013
0.75	0.3749	0.4494	0.4811	0.0011	0.0056	0.0044

Table 4: Misclassification rates with $n = 180$ and $\psi = 0.75$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1892	0.1918	0.1932	0.1124	0.1790	0.1779
0.25	0.2614	0.2654	0.2715	0.1854	0.2498	0.2502
0.50	0.4552	0.4493	0.4678	0.3227	0.4027	0.3994
0.75	0.6673	0.6440	0.6512	0.4841	0.6049	0.5815
			$T = 50$			
0.00	0.0187	0.0193	0.0194	0.0099	0.0150	0.0150
0.25	0.0429	0.0453	0.0463	0.0235	0.0361	0.0361
0.50	0.1768	0.1830	0.2053	0.0626	0.1050	0.1036
0.75	0.6081	0.5927	0.6120	0.1815	0.3766	0.3656
			$T = 100$			
0.00	0.0004	0.0015	0.0016	0.0001	0.0003	0.0003
0.25	0.0022	0.0084	0.0103	0.0009	0.0015	0.0015
0.50	0.0205	0.0782	0.1403	0.0045	0.0082	0.0081
0.75	0.4253	0.5351	0.5583	0.0186	0.0504	0.0495

Table 5: Misclassification rates with $n = 270$ and $\psi = 0.25$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0636	0.0654	0.0663	0.0440	0.0588	0.0587
0.25	0.0987	0.0951	0.0959	0.0542	0.0811	0.0804
0.50	0.3072	0.2325	0.2150	0.0610	0.1254	0.1035
0.75	0.6861	0.5449	0.4702	0.0694	0.3895	0.1450
			$T = 50$			
0.00	0.0060	0.0061	0.0061	0.0042	0.0049	0.0050
0.25	0.0118	0.0114	0.0114	0.0058	0.0076	0.0076
0.50	0.0430	0.0288	0.0277	0.0060	0.0092	0.0088
0.75	0.6279	0.3121	0.2735	0.0060	0.0259	0.0117
			$T = 100$			
0.00	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.25	0.0004	0.0005	0.0005	0.0001	0.0002	0.0002
0.50	0.0033	0.0031	0.0031	0.0001	0.0002	0.0002
0.75	0.2708	0.1166	0.1238	0.0001	0.0007	0.0005

Table 6: Misclassification rates with $n = 270$ and $\psi = 0.50$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0836	0.0856	0.0866	0.0542	0.0774	0.0775
0.25	0.1325	0.1303	0.1325	0.0785	0.1166	0.1157
0.50	0.3606	0.3071	0.3004	0.1127	0.2063	0.1854
0.75	0.6805	0.5956	0.5776	0.1680	0.4710	0.3061
			$T = 50$			
0.00	0.0073	0.0075	0.0075	0.0052	0.0061	0.0061
0.25	0.0172	0.0171	0.0171	0.0097	0.0122	0.0122
0.50	0.0684	0.0506	0.0500	0.0149	0.0208	0.0202
0.75	0.6190	0.4425	0.4468	0.0195	0.0655	0.0390
			$T = 100$			
0.00	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001
0.25	0.0007	0.0008	0.0008	0.0002	0.0004	0.0004
0.50	0.0066	0.0069	0.0071	0.0005	0.0009	0.0009
0.75	0.3207	0.2362	0.2725	0.0009	0.0031	0.0023

Table 7: Misclassification rates with $n = 270$ and $\psi = 0.75$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1493	0.1518	0.1534	0.0928	0.1422	0.1415
0.25	0.2224	0.2240	0.2296	0.1562	0.2100	0.2101
0.50	0.4420	0.4232	0.4391	0.2843	0.3658	0.3596
0.75	0.6809	0.6485	0.6589	0.4523	0.6000	0.5637
			$T = 50$			
0.00	0.0128	0.0133	0.0133	0.0091	0.0108	0.0108
0.25	0.0327	0.0338	0.0343	0.0216	0.0270	0.0269
0.50	0.1368	0.1231	0.1287	0.0491	0.0673	0.0664
0.75	0.6134	0.5678	0.5995	0.1198	0.2898	0.2689
			$T = 100$			
0.00	0.0003	0.0003	0.0003	0.0001	0.0002	0.0002
0.25	0.0019	0.0021	0.0022	0.0008	0.0013	0.0013
0.50	0.0171	0.0205	0.0224	0.0037	0.0057	0.0056
0.75	0.3905	0.4291	0.4858	0.0128	0.0262	0.0236

Table 8: Misclassification rates with $n = 450$ and $\psi = 0.25$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0463	0.0474	0.0480	0.0400	0.0431	0.0432
0.25	0.0635	0.0567	0.0569	0.0418	0.0488	0.0481
0.50	0.2330	0.1169	0.0958	0.0382	0.0562	0.0470
0.75	0.7073	0.4502	0.2364	0.0340	0.3426	0.0514
			$T = 50$			
0.00	0.0050	0.0051	0.0051	0.0038	0.0043	0.0043
0.25	0.0079	0.0071	0.0070	0.0038	0.0048	0.0047
0.50	0.0249	0.0112	0.0104	0.0030	0.0041	0.0040
0.75	0.6446	0.0887	0.0335	0.0022	0.0172	0.0037
			$T = 100$			
0.00	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000
0.25	0.0002	0.0002	0.0002	0.0000	0.0001	0.0001
0.50	0.0014	0.0010	0.0010	0.0000	0.0001	0.0001
0.75	0.1317	0.0103	0.0080	0.0000	0.0001	0.0001

Table 9: Misclassification rates with $n = 450$ and $\psi = 0.50$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0602	0.0618	0.0627	0.0499	0.0564	0.0564
0.25	0.1029	0.0972	0.0987	0.0707	0.0872	0.0866
0.50	0.3292	0.2421	0.2287	0.0926	0.1520	0.1354
0.75	0.6918	0.5699	0.5299	0.1276	0.4400	0.2327
			$T = 50$			
0.00	0.0064	0.0066	0.0066	0.0050	0.0055	0.0055
0.25	0.0155	0.0145	0.0145	0.0092	0.0109	0.0108
0.50	0.0576	0.0330	0.0315	0.0140	0.0172	0.0169
0.75	0.6225	0.3006	0.2529	0.0172	0.0446	0.0241
			$T = 100$			
0.00	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
0.25	0.0006	0.0006	0.0006	0.0002	0.0003	0.0003
0.50	0.0060	0.0042	0.0042	0.0004	0.0007	0.0007
0.75	0.2443	0.0579	0.0489	0.0006	0.0019	0.0013

Table 10: Misclassification rates with $n = 450$ and $\psi = 0.75$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1088	0.1108	0.1124	0.0773	0.1040	0.1041
0.25	0.1803	0.1783	0.1834	0.1285	0.1662	0.1664
0.50	0.4246	0.3832	0.3958	0.2390	0.3174	0.3119
0.75	0.6900	0.6452	0.6621	0.4085	0.5869	0.5384
			$T = 50$			
0.00	0.0104	0.0108	0.0109	0.0086	0.0093	0.0093
0.25	0.0284	0.0281	0.0284	0.0204	0.0229	0.0229
0.50	0.1098	0.0816	0.0822	0.0430	0.0506	0.0500
0.75	0.6176	0.5160	0.5644	0.0815	0.1867	0.1518
			$T = 100$			
0.00	0.0003	0.0003	0.0003	0.0001	0.0002	0.0002
0.25	0.0016	0.0016	0.0016	0.0008	0.0011	0.0011
0.50	0.0157	0.0130	0.0131	0.0035	0.0047	0.0047
0.75	0.3403	0.2412	0.2784	0.0100	0.0177	0.0150

Table 11: Misclassification rates with with $R = 1$ and $n = 180$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1472	0.1499	0.2112	0.1022	0.1411	0.1425
0.25	0.2243	0.2229	0.2695	0.1404	0.2030	0.1958
0.50	0.4028	0.3686	0.3702	0.1576	0.2804	0.2449
0.75	0.6750	0.5866	0.5559	0.1803	0.4918	0.3348
			$T = 50$			
0.00	0.0094	0.0096	0.0420	0.0057	0.0075	0.0081
0.25	0.0238	0.0242	0.0610	0.0113	0.0168	0.0171
0.50	0.1091	0.0941	0.1302	0.0183	0.0316	0.0276
0.75	0.6326	0.4857	0.4892	0.0351	0.1660	0.0816
			$T = 100$			
0.00	0.0001	0.0008	0.0029	0.0000	0.0001	0.0001
0.25	0.0006	0.0027	0.0072	0.0001	0.0003	0.0003
0.50	0.0056	0.0169	0.0363	0.0002	0.0005	0.0005
0.75	0.3669	0.3888	0.4169	0.0037	0.0137	0.0088

Table 12: Misclassification rates with with $R = 1$ and $n = 270$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0904	0.0926	0.1480	0.0654	0.0863	0.0897
0.25	0.1620	0.1578	0.2076	0.0982	0.1408	0.1390
0.50	0.3696	0.3128	0.3024	0.1140	0.2129	0.1793
0.75	0.6926	0.5703	0.5120	0.1273	0.4623	0.2482
			$T = 50$			
0.00	0.0059	0.0061	0.0161	0.0041	0.0048	0.0048
0.25	0.0148	0.0143	0.0288	0.0077	0.0096	0.0095
0.50	0.0646	0.0443	0.0561	0.0112	0.0156	0.0148
0.75	0.6404	0.3815	0.3597	0.0210	0.0819	0.0340
			$T = 100$			
0.00	0.0001	0.0001	0.0007	0.0001	0.0001	0.0001
0.25	0.0005	0.0006	0.0015	0.0001	0.0002	0.0002
0.50	0.0036	0.0034	0.0054	0.0001	0.0003	0.0003
0.75	0.3301	0.1662	0.1873	0.0017	0.0051	0.0031

Table 13: Misclassification rates with with $R = 1$ and $n = 450$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			T=25			
0.00	0.0526	0.0539	0.0889	0.0434	0.0497	0.0523
0.25	0.1097	0.1023	0.1427	0.0702	0.0914	0.0917
0.50	0.3286	0.2350	0.2119	0.0830	0.1490	0.1207
0.75	0.7067	0.5396	0.4298	0.0920	0.4346	0.1744
			T=50			
0.00	0.0045	0.0046	0.0068	0.0035	0.0039	0.0039
0.25	0.0116	0.0107	0.0134	0.0065	0.0077	0.0077
0.50	0.0458	0.0232	0.0229	0.0087	0.0111	0.0107
0.75	0.6519	0.2336	0.1652	0.0148	0.0445	0.0194
			T=100			
0.00	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001
0.25	0.0003	0.0003	0.0007	0.0001	0.0001	0.0001
0.50	0.0027	0.0018	0.0021	0.0001	0.0002	0.0002
0.75	0.2181	0.0346	0.0274	0.0007	0.0019	0.0012

Table 14: Misclassification rates with with $R = 2$ and $n = 180$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1944	0.1967	0.2346	0.1594	0.1903	0.1881
0.25	0.2563	0.2564	0.2861	0.1954	0.2436	0.2386
0.50	0.3801	0.3542	0.3545	0.2018	0.2919	0.2701
0.75	0.6372	0.5352	0.5028	0.2151	0.4503	0.3365
			$T = 50$			
0.00	0.0179	0.0183	0.0513	0.0107	0.0157	0.0164
0.25	0.0411	0.0419	0.0807	0.0201	0.0322	0.0319
0.50	0.1208	0.1088	0.1368	0.0244	0.0457	0.0394
0.75	0.5779	0.4242	0.4196	0.0378	0.1570	0.0839
			$T = 100$			
0.00	0.0002	0.0010	0.0044	0.0001	0.0002	0.0002
0.25	0.0010	0.0038	0.0099	0.0003	0.0005	0.0005
0.50	0.0070	0.0222	0.0473	0.0003	0.0008	0.0008
0.75	0.2448	0.2860	0.3170	0.0015	0.0068	0.0038

Table 15: Misclassification rates with with $R = 2$ and $n = 270$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.1299	0.1320	0.1709	0.1034	0.1267	0.1293
0.25	0.1971	0.1943	0.2314	0.1394	0.1803	0.1805
0.50	0.3487	0.3037	0.2986	0.1486	0.2341	0.2081
0.75	0.6613	0.5162	0.4588	0.1563	0.4198	0.2566
			$T = 50$			
0.00	0.0065	0.0066	0.0186	0.0047	0.0054	0.0057
0.25	0.0181	0.0176	0.0363	0.0095	0.0124	0.0125
0.50	0.0689	0.0511	0.0626	0.0115	0.0176	0.0160
0.75	0.5923	0.3231	0.2942	0.0177	0.0768	0.0319
			$T = 100$			
0.00	0.0001	0.0002	0.0010	0.0001	0.0001	0.0001
0.25	0.0007	0.0007	0.0019	0.0002	0.0003	0.0003
0.50	0.0040	0.0038	0.0058	0.0002	0.0003	0.0003
0.75	0.2300	0.1148	0.1320	0.0005	0.0021	0.0011

Table 16: Misclassification rates with with $R = 2$ and $n = 450$

ρ_0	KM	WKM	I-WKM	Inf-B-WKM	B-WKM	I-B-WKM
			$T = 25$			
0.00	0.0683	0.0697	0.1011	0.0559	0.0658	0.0711
0.25	0.1411	0.1351	0.1677	0.0960	0.1250	0.1258
0.50	0.3205	0.2482	0.2302	0.1079	0.1798	0.1515
0.75	0.6844	0.4976	0.4025	0.1142	0.4065	0.1996
			$T = 50$			
0.00	0.0045	0.0047	0.0067	0.0035	0.0038	0.0040
0.25	0.0124	0.0115	0.0153	0.0071	0.0083	0.0085
0.50	0.0445	0.0230	0.0235	0.0080	0.0102	0.0097
0.75	0.6087	0.1921	0.1373	0.0107	0.0381	0.0151
			$T = 100$			
0.00	0.0001	0.0001	0.0004	0.0001	0.0001	0.0001
0.25	0.0004	0.0004	0.0009	0.0001	0.0002	0.0002
0.50	0.0027	0.0018	0.0022	0.0001	0.0002	0.0002
0.75	0.1435	0.0226	0.0176	0.0001	0.0006	0.0003

Table 17: Performances of various estimators with $R = 1$ and $n = 180$

	CM		KM		B-WKM		I-B-WKM	
ρ_0	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
				$T = 25$				
0.00	0.0951	0.0966	0.0582	0.0433	0.0567	0.0410	0.0570	0.0411
	(0.0650)	(0.0644)	(0.0427)	(0.0338)	(0.0430)	(0.0339)	(0.0431)	(0.0334)
0.25	0.1192	0.1091	0.0830	0.0484	0.0789	0.0459	0.0774	0.0441
	(0.0690)	(0.0659)	(0.0450)	(0.0292)	(0.0457)	(0.0302)	(0.0454)	(0.0304)
0.50	0.1297	0.1156	0.1120	0.0619	0.0942	0.0555	0.0889	0.0503
	(0.0677)	(0.0646)	(0.0381)	(0.0273)	(0.0433)	(0.0302)	(0.0437)	(0.0296)
0.75	0.1309	0.1156	0.1353	0.0714	0.1221	0.0817	0.1013	0.0612
	(0.0655)	(0.0608)	(0.0276)	(0.0263)	(0.0294)	(0.0236)	(0.0405)	(0.0295)
				$T = 50$				
0.00	0.0434	0.0457	0.0127	0.0112	0.0119	0.0098	0.0120	0.0100
	(0.0302)	(0.0302)	(0.0119)	(0.0098)	(0.0109)	(0.0085)	(0.0116)	(0.0091)
0.25	0.0683	0.0657	0.0238	0.0176	0.0207	0.0143	0.0206	0.0141
	(0.0375)	(0.0356)	(0.0201)	(0.0144)	(0.0184)	(0.0125)	(0.0185)	(0.0124)
0.50	0.1000	0.0944	0.0631	0.0388	0.0337	0.0188	0.0321	0.0172
	(0.0430)	(0.0407)	(0.0326)	(0.0228)	(0.0265)	(0.0168)	(0.0249)	(0.0145)
0.75	0.1412	0.1270	0.1385	0.0828	0.0828	0.0465	0.0612	0.0265
	(0.0447)	(0.0423)	(0.0180)	(0.0183)	(0.0371)	(0.0271)	(0.0359)	(0.0215)
				$T = 100$				
0.00	0.0218	0.0222	0.0051	0.0053	0.0050	0.0051	0.0050	0.0051
	(0.0157)	(0.0163)	(0.0040)	(0.0039)	(0.0038)	(0.0037)	(0.0039)	(0.0037)
0.25	0.0300	0.0297	0.0063	0.0065	0.0058	0.0056	0.0058	0.0056
	(0.0198)	(0.0199)	(0.0052)	(0.0048)	(0.0046)	(0.0041)	(0.0046)	(0.0041)
0.50	0.0461	0.0469	0.0118	0.0109	0.0066	0.0060	0.0066	0.0060
	(0.0263)	(0.0258)	(0.0097)	(0.0084)	(0.0054)	(0.0045)	(0.0053)	(0.0045)
0.75	0.1256	0.1235	0.1167	0.0776	0.0264	0.0104	0.0237	0.0095
	(0.0407)	(0.0378)	(0.0256)	(0.0222)	(0.0227)	(0.0086)	(0.0217)	(0.0070)

Table 18: Performances of various estimators with $R = 2$ and $n = 180$

	CM		KM		B-WKM		I-B-WKM	
ρ_0	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
				$T = 25$				
0.00	0.0813	0.0848	0.0381	0.0297	0.0376	0.0288	0.0377	0.0284
	(0.0605)	(0.0601)	(0.0306)	(0.0252)	(0.0308)	(0.0255)	(0.0311)	(0.0253)
0.25	0.0955	0.0951	0.0520	0.0277	0.0509	0.0272	0.0504	0.0267
	(0.0628)	(0.0617)	(0.0340)	(0.0215)	(0.0339)	(0.0219)	(0.0342)	(0.0218)
0.50	0.1026	0.1011	0.0653	0.0281	0.0572	0.0283	0.0557	0.0268
	(0.0638)	(0.0614)	(0.0329)	(0.0203)	(0.0332)	(0.0207)	(0.0335)	(0.0206)
0.75	0.1068	0.1039	0.0822	0.0293	0.0698	0.0375	0.0615	0.0298
	(0.0616)	(0.0583)	(0.0274)	(0.0200)	(0.0273)	(0.0194)	(0.0319)	(0.0198)
				$T = 50$				
0.00	0.0332	0.0340	0.0129	0.0117	0.0122	0.0108	0.0126	0.0110
	(0.0252)	(0.0249)	(0.0139)	(0.0121)	(0.0135)	(0.0115)	(0.0137)	(0.0118)
0.25	0.0431	0.0434	0.0221	0.0168	0.0197	0.0147	0.0195	0.0146
	(0.0302)	(0.0294)	(0.0192)	(0.0142)	(0.0187)	(0.0136)	(0.0187)	(0.0137)
0.50	0.0565	0.0579	0.0382	0.0248	0.0250	0.0164	0.0240	0.0152
	(0.0357)	(0.0350)	(0.0240)	(0.0164)	(0.0218)	(0.0145)	(0.0212)	(0.0134)
0.75	0.0798	0.0801	0.0756	0.0393	0.0463	0.0265	0.0363	0.0180
	(0.0416)	(0.0385)	(0.0169)	(0.0148)	(0.0271)	(0.0181)	(0.0262)	(0.0147)
				$T = 100$				
0.00	0.0199	0.0195	0.0042	0.0044	0.0041	0.0041	0.0041	0.0042
	(0.0154)	(0.0144)	(0.0037)	(0.0034)	(0.0034)	(0.0033)	(0.0034)	(0.0033)
0.25	0.0222	0.0224	0.0056	0.0057	0.0050	0.0049	0.0050	0.0049
	(0.0164)	(0.0162)	(0.0050)	(0.0047)	(0.0047)	(0.0040)	(0.0046)	(0.0040)
0.50	0.0259	0.0274	0.0110	0.0104	0.0058	0.0054	0.0057	0.0053
	(0.0188)	(0.0193)	(0.0102)	(0.0088)	(0.0054)	(0.0046)	(0.0052)	(0.0045)
0.75	0.0488	0.0581	0.0510	0.0365	0.0122	0.0080	0.0106	0.0067
	(0.0336)	(0.0337)	(0.0230)	(0.0181)	(0.0138)	(0.0070)	(0.0123)	(0.0056)

References

[1] Aitken, B. J., & Harrison, A. E. (1999). Do domestic firms benefit from direct foreign investment? Evidence from Venezuela. American Economic Review, 89(3), 605-618.

[2] Ando, T., & Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. Journal of Applied Econometrics, 31(1), 163-191.

[3] Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica: Journal of the Econometric Society, 817-858.

[4] Bai, J. (2009). Panel data models with interactive fixed effects. Econometrica, 77(4), 1229-1279.

[5] Bai, Z., & Silverstein, J. W. (2010). Spectral analysis of large dimensional random matrices (Vol. 20). New York: Springer.

[6] Bai, J., & Shi, S. (2011). Estimating high dimensional covariance matrices and its applications.

[7] Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191-221.

[8] Bai, J. (2003). Inferential theory for factor models of large dimensions. Econometrica, 71(1), 135-171.

[9] Basak, A., Bose, A., & Sen, S. (2014). Limiting spectral distribution of sample autocovariance matrices. Bernoulli, 20(3), 1234–1259. http://www.jstor.org/stable/42919430

[10] Bian, Y., & Su, L. (2025). A note on factor models with latent group structures. Economics Letters, 112357.

[11] Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y. A., Klaassen, J., Wellner, J. A., & Ritov, Y. A. (1993). Efficient and adaptive estimation for semiparametric models (Vol. 4). Baltimore: Johns Hopkins University Press.

[12] Bickel, P. J., & Levina, E. (2008a). Regularized estimation of large covariance matrices. Ann. Statist. 36 (1) 199-227, February 2008. https://doi.org/10.1214/009053607000000758

[13] Bickel, P. J., & Levina, E. (2008b). Covariance Regularization Thresholding. Ann. Statist, 36(6), 2577-2604.

[14] Bonhomme, S., & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. Econometrica, 83(3), 1147-1184.

[15] Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika, 73(1), 125-144.

[16] Cantwell, J. (1989). Technological innovation and multinational corporations. Basil Blackwell.

[17] Chen, Y., Jiang, H., Liang, Y., & Pan, S. (2022). The impact of foreign direct investment on innovation: Evidence from patent filings and citations in China. Journal of Comparative Economics, 50(4), 917-945.

[18] Chetverikov, D., & Manresa, E. (2022). Spectral and post-spectral estimators for grouped panel data models. arXiv preprint arXiv:2212.13324.

[19] Dunning, J. H., & Lundan, S. M. (2008). Multinational enterprises and the global economy. Edward Elgar Publishing.

[20] Fan, J., Liao, Y., & Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. Annals of statistics, 39(6), 3320.

[21] Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics, 21, 768-769.

[22] Gnanadesikan, R., Harvey, J. W., & Kettenring, J. R. (1993). Mahalanobis metrics for cluster analysis. Sankhyā: The Indian Journal of Statistics, Series A, 494-505.

[23] Grenander, U. and Szegö, G. (1958). Toeplitz Forms and Their Applications. Univ. California Press, Berkeley.

[24] Hahn, J., & Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. Econometric Theory, 26(3), 863-881.

[25] Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. Journal of Business & Economic Statistics, 14(3), 262-280.

[26] Hansen, P., & Mladenovi$\hat{c}$, N. (2001). J-means: a new local search heuristic for minimum sum of squares clustering. Pattern recognition, 34(2), 405-413.

[27] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

[28] Horn, R. A., & Johnson, C. R. (2012). Matrix analysis. Cambridge university press.

[29] Huang, W., Jin, S., & Su, L. (2020). Identifying latent grouped patterns in cointegrated panels. Econometric Theory, 36(3), 410-456.

[30] Kao, C., Kim, M. S., & Zhang, Z. (2021). Mahalanobis Metric Based Clustering for Fixed Effects Model. Sankhya B, 83, 493-506.

[31] Kuhn, H. W. (1956). Variants of the Hungarian method for assignment problems. Naval research logistics quarterly, 3(4), 253-258.

[32] Lin, C. C., & Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. Journal of Econometric Methods, 1(1), 42-55.

[33] Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.

[34] Lumsdaine, R. L., Okui, R., & Wang, W. (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. Journal of Econometrics, 233(1), 45-65.

[35] Mahalanobis, P. C. (1936). A note on the statistical and biometric writings of Karl Pearson. Sankhyā: The Indian Journal of Statistics (1933-1960), 2(4), 411-422.

[36] Merlevède, F., Peligrad, M., & Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. Probability Theory and Related Fields, 151(3), 435-474.

[37] Moon, H. R., & Weidner, M. (2018). Nuclear norm regularized estimation of panel regression models. arXiv preprint arXiv:1810.10987.

[38] Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics, 5(1), 32-38.

[39] Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55(3), 703–708.

[40] Okui, R., & Wang, W. (2021). Heterogeneous structural breaks in panel data models. Journal of Econometrics, 220(2), 447-473.

[41] Phillips, P. C. (2005). HAC estimation by automated regression. Econometric Theory, 21(1), 116-142.

[42] Pollard, D. (1981). Strong consistency of K-means clustering. The Annals of Statistics, 135-140.

[43] Pollard, D. (1982). A central limit theorem for K-means clustering. The Annals of Probability, 10(4), 919-926.

[44] Porter, M. E. (2011). Competitive advantage of nations: creating and sustaining superior performance. Simon and Schuster

[45] Su, L., Shi, Z., & Phillips, P. C. (2016). Identifying latent structures in panel data. Econometrica, 84(6), 2215-2264.

[46] Su, L., Wang, W., & Xu, X. (2023). Identifying latent group structures in spatial dynamic panels. Journal of Econometrics, 235(2), 1955-1980.

[47] Sun, Y. (2014). Fixed-smoothing asymptotics in a two-step generalized method of moments framework. Econometrica, 82(6), 2327-2370.

[48] Sun, Y., & Yang, J. (2020). Testing-optimal kernel choice in HAR inference. Journal of Econometrics, 219(1), 123-136.

[49] Vogt, M., & Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(1), 5-27.

[50] Wang, W., & Su, L. (2021). Identifying latent group structures in nonlinear panels. Journal of Econometrics, 220(2), 272-295.

[51] Xiao, H., & Wu, W. B. (2012). Covariance matrix estimation for stationary time series. Ann. Statist. 40 (1) 466 - 493, February 2012. https://doi.org/10.1214/11-AOS967

[52] Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science (Vol. 47). Cambridge university press.