

**Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.**

**Revealing Cluster Structures Based on Mixed Sampling
Frequencies**

Yeonwoo Rho, Yun Liu, and Hie Joo Ahn

2020-082

Please cite this paper as:

Rho, Yeonwoo, Yun Liu, and Hie Joo Ahn (2020). “Revealing Cluster Structures Based on Mixed Sampling Frequencies,” Finance and Economics Discussion Series 2020-082. Washington: Board of Governors of the Federal Reserve System, <https://doi.org/10.17016/FEDS.2020.082>.

NOTE: Staff working papers in the Finance and Economics Discussion Series (FEDS) are preliminary materials circulated to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. References in publications to the Finance and Economics Discussion Series (other than acknowledgement) should be cleared with the author(s) to protect the tentative character of these papers.

Revealing Cluster Structures Based on Mixed Sampling Frequenciesⁱ

YEONWOO RHO¹, YUN LIU², AND HIE JOO AHN³ ⁱⁱ

¹Michigan Technological University

²Quicken Loans

³Federal Reserve Board

September 9, 2020

Abstract

This paper proposes a new nonparametric mixed data sampling (MIDAS) model and develops a framework to infer clusters in a panel regression with mixed frequency data. The nonparametric MIDAS estimation method is more flexible and substantially simpler to implement than competing approaches. We show that the proposed clustering algorithm successfully recovers true membership in the cross-section, both in theory and in simulations, without requiring prior knowledge of the number of clusters. This methodology is applied to a mixed-frequency Okun's law model for state-level data in the U.S. and uncovers four meaningful clusters based on the dynamic features of state-level labor markets.

key words: Clustering; forecasting; mixed data sampling regression model; panel data; penalized regression.

ⁱOpinions expressed herein are those of the authors alone and do not necessarily reflect the views of the Federal Reserve System. We thank Gianni Amisano, Eric Ghysels and the participants of Midwest Econometrics Group 2018 Meeting, and the joint meeting of 11th International Conference of the ERCIM WG on Computational and Methodological Statistics and 12th International Conference on Computational and Financial Econometrics. Rho and Liu are partially supported by NSF-CPS grant #1739422. A primitive version of this paper had been circulated under the title "Panel Nonparametric MIDAS Model: A Clustering Approach" by the first two authors.

ⁱⁱEmails: Y. Rho (yrho@mtu.edu), Y. Liu (yliu26@mtu.edu), and H. J. Ahn (HieJoo.Ahn@frb.gov)

1 Introduction

Following technological advances, the diffusion of social media, and the efforts of statistical agencies and private companies, new data sources have recently become available for empirical research in economics. In many cases, these data are characterized by large time series and cross sectional dimensions, with detailed information on economic agents often at a level of disaggregation more granular than that of traditional data sources. To cope with the changing data environment, new methods have been developed in the econometrics literature. For instance, to utilize the high-frequency data for forecasting, regression models concatenating low- and high-frequency variables—so called mixed data sampling (MIDAS) models (e.g., Ghysels et al. [2007])— are now widely used in practical applications. In addition, due to the increasing availability of richer cross-section data it has become particularly important to efficiently summarize and identify the most important features of subjects in the cross-section, with the aim of properly account for unit heterogeneity. In this respect, clustering algorithms developed in the machine learning literature are increasingly being used in econometric applications.

Currently, there is no statistical methodology that allows a researcher to identify distinct groups in a panel data and mixed-frequency regression setting. This paper aims at filling this gap in the literature by proposing a new nonparametric method. The key innovation of our proposed approach is the following. First, we propose a new non-parametric MIDAS model based on the Fourier flexible form and polynomials. In parametric MIDAS models, arbitrary parametric functions (e.g., exponential Almon lag function, beta function) are used to model the coefficients on high-frequency variables. As these parametric functions are highly nonlinear in general, complicated numerical optimization is required for estimation. As this estimation is numerically costly and challenging, practitioners often give MIDAS models a wide berth. Our model requires just ordinary least squares (OLS), which eschews estimation difficulties. In addition, it does not require any arbitrary choice of parametric functional forms in specifying the distributed lags structure of the coefficients on high frequency variables. To model the coefficients on high-frequency variables, the Fourier flexible form and polynomials are used, which allows the trajectory of coefficients on high-frequency variables to be flexibly determined by the data.ⁱⁱⁱ

ⁱⁱⁱA recent study by Babii et al. [2019] develops a general framework with dictionaries for machine-learning time-

Breitung and Roling [2015] have recently proposed a nonparametric MIDAS model that, like ours, does not require to specify any functional form for the coefficients on the high-frequency. Unlike our approach, however, Breitung and Roling [2015]’s methodology requires the careful setting of two tuning parameters, which is much more computationally demanding.

45 Another contribution of this paper is to extend the new MIDAS model to the panel data framework, and to devise a novel clustering method for this panel setting. This clustering algorithm uses the idea of penalized regression with penalty function as in Ma and Huang [2017], and is adapted to general panel data settings including mixed-frequency panel data. An important feature of the proposed method is that the identification of clusters is entirely data-driven, while previous studies
50 (e.g. Andreou et al. [2010]) arbitrarily divide observations into different groups. However, this approach crucially depends on prior knowledge and homogeneity within each group is not necessarily guaranteed.^{iv}

Simulations conducted in this paper show many desirable features of our method. First, the proposed non-parametric MIDAS model provides the best one-step-ahead forecast among paramet-
55 ric and non-parametric MIDAS models widely used in empirical studies. Second, our clustering method yields more precise parameter estimation and better forecasting properties than competing approaches. Third, the proposed clustering algorithm is faster than competitors (e.g. K-means clustering) when the number of underlying clusters is unknown.

As a relevant empirical application, we use our method to explore heterogeneity in labor market
60 dynamics across states in the U.S. using a mixed-frequency panel Okun’s law model. Okun’s law is an empirical relationship that relates changes in unemployment rate to GDP growth. Usually an Okun’s law model is specified at quarterly frequency, as GDP growth is available only quarterly. In

series regression models, and recommends to use Legendre polynomials for the coefficients on high-frequency variables in a MIDAS model. Our use of trigonometric functions for the MIDAS coefficients is a special case of Babii et al. [2019]’s framework, and inherits advantages similar to those of Legendre polynomials in the estimation.

^{iv}It should be noted that the application of proposed clustering algorithm is not limited to datasets with mixed sampling frequencies. It can be applied to a general panel data setting with a fixed sampling frequency, in which case the proposed method can be an appealing alternative to Su et al. [2016]’s method. There are three aspects why our approach may have advantages over Su et al. [2016]’s method. First, our penalty function is more general than theirs. Second, Su et al. [2016]’s method requires to pre-specify a possible ranges for the number of clusters in addition to the turning parameter. If there is no prior knowledge of the number of clusters and if the size of cross-section is large, then finding right clusters becomes very computationally challenging as the various possibilities need be considered. Our method requires a few user-chosen parameters, but the range of choices is well defined and does not yield intractable possibilities. Lastly, Su et al. [2016]’s method requires the number of clusters to be fixed regardless of the size of cross-section. The theory behind our clustering algorithm does flexibly allow the number of groups to adjust to a change in the size of cross-section.

our application, we include weekly initial claims of unemployment insurance (UI) benefits, which is known as the most timely indicator of job losses, as the high-frequency variable in a mixed
65 frequency Okun’s law model. By doing so, the model can better characterize the sudden rise in unemployment rate at the onset of a recession, picking up sudden bursts in layoffs. An additional desirable feature of the mixed-frequency Okun’s law model is that it can be used to nowcast the unemployment rate at the state level on a weekly basis.

The algorithm identifies four clusters of states based on the responsiveness of unemployment
70 rate to GDP growth and on the pattern of coefficients on weekly initial claims within the quarter. The coefficients on GDP growth and initial claims most likely reflect structural aspects of state-level labor markets (e.g. the industry composition) and local labor-market practices in hiring and firing. Hence, the clusters identified by the model most likely capture relevant heterogeneity in the functioning of labor markets in different states.

75 We relate the identified clusters to observable state-level attributes such as the small-firms employment share, industry composition, the relevance of oil production, and the share of long-term unemployment out of total unemployment. Each cluster exhibits multidimensional attributes, suggesting that the differences in labor-market dynamics across states cannot be determined or accurately summarized by one or two observable factors. Another way to say this is that the
80 clustering algorithm is able to capture a state’s unobserved attributes which are not fully reflected in the data but are nevertheless crucial for unemployment dynamics. In this regard, our proposed methodology can reveal similarities and differences across states in the functioning of their labor markets purely based on the data, and can provide a new understanding of regional heterogeneity in labor market dynamics.

85 The rest of this paper is organized as follows. Section 2 introduces the proposed nonparametric MIDAS approach using the Fourier flexible form and polynomials. Subsection 2.1 discusses the nonparametric MIDAS estimation in a non-panel setting. Subsection 2.2 demonstrates the proposed method’s estimation and forecasting accuracy in finite samples. Section 3 presents the clustering algorithm using the Fourier-transformed high-frequency variable and other possible covariates. The
90 proposed clustering approach delivers accurate estimates, as proved in theory and shown in the finite sample simulations. Section 4.1 provides an empirical application of the method. Technical

proofs are relegated to Section B in the appendix.

The following notation will be used throughout the paper. The p -norm of a vector $x = (x_1, \dots, x_m)'$ is $\|x\|_p = (\sum_{i=1}^m x_i^p)^{1/p}$. For an $m \times n$ matrix A with its (i, j) th element being a_{ij} , $\|A\|_p$ indicates the p -norm induced by the corresponding vector norm. That is, $\|A\|_p = \sup_{x \neq 0} \|Ax\|_p / \|x\|_p$. In particular, $\|A\|_1 = \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}|$ and $\|A\|_\infty = \max_{i=1, \dots, m} \sum_{j=1}^n |a_{ij}|$. For a symmetric and positive definite matrix A , let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ indicate the smallest and largest eigenvalues of A , respectively. It is worth noting that $\|A\|_2 = \lambda_{\max}(A)$. I_p is a $p \times p$ identity matrix and \otimes denotes the Kronecker product. For any real number x , $\lfloor x \rfloor$ denotes the largest integer that is smaller than or equal to x . The symbol $\mathbf{1}\{\cdot\}$ denotes the indicator function.

2 Nonparametric MIDAS

In this section, we introduce our nonparametric MIDAS approach using the Fourier flexible form [Gallant, 1981]. We first introduce the framework, then confirm that the proposed nonparametric MIDAS model is a good approximation of popular parametric MIDAS models in finite samples.

2.1 Nonparametric MIDAS with the Fourier flexible form and polynomials

Consider the following MIDAS model with the forecast lead $h \geq 0$:

$$y_{t+h} = \sum_{i=1}^q \alpha_i z_{t,i} + \sum_{j=0}^{m-1} \beta_{j/m}^* x_{t,j} + \varepsilon_{t+h} = \mathbf{z}_t' \boldsymbol{\alpha} + \mathbf{x}_t' \boldsymbol{\beta}^* + \varepsilon_{t+h}, \quad (1)$$

for $t = 1, \dots, T$. Here, \mathbf{z}_t is the q -vector of low-frequency covariates at time t , and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$ is the corresponding coefficient vector. The vector $\mathbf{x}_t = (x_{t,0}, \dots, x_{t,m-1})'$ is the high-frequency variable at t and $\boldsymbol{\beta}^* = (\beta_{0/m}^*, \dots, \beta_{m-1/m}^*)'$ are the coefficients that aggregate \mathbf{x}_t to the low-frequency. In a parametric MIDAS model, the coefficients $\beta_{j/m}^*$ can be written as a multiple of $\omega_j(\boldsymbol{\theta})$, where weights $\omega_j(\boldsymbol{\theta})$ are assumed to be generated by, for example, an exponential Almon lag function

$$\beta_{j/m}^* = \alpha^* \omega_j(\boldsymbol{\theta}) = \frac{\alpha^* \exp(\theta_1 j + \theta_2 j^2 + \dots + \theta_Q j^Q)}{\sum_{i=0}^{m-1} \exp(\theta_1 i + \theta_2 i^2 + \dots + \theta_Q i^Q)},$$

and α^* and $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_Q)$ are parameters that need to be estimated from data. However, the form of $\omega_j(\cdot)$ is somewhat limited, and it requires nonlinear estimation. In this paper, we propose
110 to model $\beta_{j/m}^*$ using the Fourier flexible form and polynomials. The MIDAS coefficients $\beta_{j/m}^*$ are assumed to be generated by

$$\beta_{j/m}^* = \sum_{l=0}^L \beta_l(j/m)^l + \sum_{k=1}^K \{\beta_{1,k} \sin(2\pi k \cdot j/m) + \beta_{2,k} \cos(2\pi k \cdot j/m)\}, \quad (2)$$

for some positive integers L and K . The Fourier flexible form has been frequently used in macroeconomics and finance since Gallant [1981]. It has been demonstrated that the Fourier flexible form is capable of approximating most forms of nonlinear time trends to any degree of accuracy if a suffi-
115 cient number of parameters is used, and that a small K is often enough to reasonably approximate smooth functions with finite numbers of breaks [Becker et al., 2004, 2006, Enders and Lee, 2012, Rodrigues and Robert Taylor, 2012, Güriş, 2017, Perron et al., 2017]. In addition to the Fourier flexible form, we also consider a few polynomial trends to cover wider range of nonlinear functions, following suggestions in Perron et al. [2017].

The MIDAS model (1) with the Fourier flexible form (2) can be expressed as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} = \mathbf{Z}\boldsymbol{\alpha} + \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

120 where $\mathbf{y} = (y_{1+h}, \dots, y_{T+h})'$, $\boldsymbol{\varepsilon} = (\varepsilon_{1+h}, \dots, \varepsilon_{T+h})'$, $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_T]'$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]'$, $\mathbf{W} = (\mathbf{Z}, \tilde{\mathbf{X}})$, $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{M}' = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_T]'$, and

$$M = \begin{bmatrix} (0/m)^0 & (1/m)^0 & \cdots & ((m-1)/m)^0 \\ \vdots & \vdots & & \vdots \\ (0/m)^L & (1/m)^L & \cdots & ((m-1)/m)^L \\ \sin(2\pi \cdot 1 \cdot 0/m) & \sin(2\pi \cdot 1 \cdot 1/m) & \cdots & \sin(2\pi \cdot 1 \cdot (m-1)/m) \\ \cos(2\pi \cdot 1 \cdot 0/m) & \cos(2\pi \cdot 1 \cdot 1/m) & \cdots & \cos(2\pi \cdot 1 \cdot (m-1)/m) \\ \vdots & \vdots & & \vdots \\ \sin(2\pi \cdot K \cdot 0/m) & \sin(2\pi \cdot K \cdot 1/m) & \cdots & \sin(2\pi \cdot K \cdot (m-1)/m) \\ \cos(2\pi \cdot K \cdot 0/m) & \cos(2\pi \cdot K \cdot 1/m) & \cdots & \cos(2\pi \cdot K \cdot (m-1)/m) \end{bmatrix}. \quad (3)$$

Here, the matrix M can be understood as a Fourier transform operator. This Fourier transformation summarizes the information in an m -dimensional vector \mathbf{x}_t into a $(2K + L + 1)$ -dimensional vector $\tilde{\mathbf{x}}_t = \mathbf{M}\mathbf{x}_t = (\tilde{x}_{t,0}, \tilde{x}_{t,1}, \dots, \tilde{x}_{t,L}, \tilde{x}_{t,1}^{(s)}, \tilde{x}_{t,1}^{(c)}, \dots, \tilde{x}_{t,K}^{(s)}, \tilde{x}_{t,K}^{(c)})'$, where $\tilde{x}_{t,l}$, $\tilde{x}_{t,k}^{(s)}$ and $\tilde{x}_{t,k}^{(c)}$ are transformed
125 high-frequency data for $l = 0, \dots, L$ and $k = 1, \dots, K$, and are defined as $\tilde{x}_{t,l} = (j/m)^l x_{t,j}$, $\tilde{x}_{t,k}^{(s)} = \sin(2\pi k \cdot j/m) x_{t,j}$, and $\tilde{x}_{t,k}^{(c)} = \cos(2\pi k \cdot j/m) x_{t,j}$.^v

Unlike parametric MIDAS models, this model is linear. Noting that $\beta^* = \mathbf{M}'\beta$, the ordinary least squares (OLS) estimator of β^* can be written as

$$\widehat{\beta^*} = \mathbf{M}'\mathbf{D}\widehat{\gamma} = \mathbf{M}'\mathbf{D}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} = \beta^* + \mathbf{M}'\mathbf{D}\left(\frac{1}{T}\mathbf{W}'\mathbf{W}\right)^{-1}\left(\frac{1}{T}\mathbf{W}'\varepsilon\right), \quad (4)$$

where $\mathbf{D} = [\mathbf{0}_{(L+1+2K) \times q}, \mathbf{I}_{L+1+2K}]$, and \mathbf{I}_{L+1+2K} is an identity matrix. Under some regularity
130 conditions, β can be estimated consistently by the OLS estimator $\widehat{\beta^*}$.

2.2 Simulation: Nonparametric MIDAS

This section presents simulation results to check the estimation and forecasting accuracy of the proposed nonparametric MIDAS estimation using the Fourier flexible form in finite samples. Our method is compared with the nonparametric MIDAS approach proposed by Breitung and Roling
135 [2015], where a smoothness condition was imposed on $\beta_{j/m}^*$, without involving a weight function $\omega_j(\cdot)$, similarly to our method. However, Breitung and Roling [2015] requires choosing a tuning parameter. See Section A.1 in the supplementary material for more details about Breitung and Roling [2015].

The data is generated in a similar setting as Breitung and Roling [2015]. For $j = 0, \dots, m - 1$,
140 $t = 1, \dots, T$,

$$y_{t+h} = \alpha_0 + \sum_{j=0}^{m-1} \beta_j^* x_{t,j} + \varepsilon_{t+h}, \quad x_{t,j} = c + dx_{t,j-1} + u_{t,j}, \quad (5)$$

where $\varepsilon_{t+h} \stackrel{iid}{\sim} N(0, 0.125)$, $u_{t,j} \stackrel{iid}{\sim} N(0, 1)$, $\alpha_0 = 0.5$, $\beta_j^* = \alpha_1 \omega_j(\theta)$, $\alpha_1 \in \{0.2, 0.3, 0.4\}$, $T \in \{100, 200, 400\}$, and the frequency ratio $m \in \{20, 40, 60, 150, 365\}$. For the AR(1) high-frequency

^vNote that $\tilde{\mathbf{x}}_t$ can effectively summarize the information in \mathbf{x}_t , because relatively small K and L are enough to capture main characteristics of a nonparametric trend function. For instance, Enders and Lee [2012] reported that even a single frequency $K = 1$ allows for multiple smooth breaks.

regressor, $c = 0.5$ and $d = 0.9$ are considered. Four MIDAS weight functions $\omega_j(\boldsymbol{\theta})$ are considered:

- Exponential Decline: $\omega_j(\theta_1, \theta_2) = \frac{\exp\{\theta_1 j + \theta_2 j^2\}}{\sum_{i=0}^{m-1} \exp\{\theta_1 i + \theta_2 i^2\}}$, $\theta_1 = 7 \times 10^{-4}$, $\theta_2 = -6 \times 10^{-3}$;
- 145 • Hump-Shaped: $\omega_j(\theta_1, \theta_2) = \frac{\exp\{\theta_1 j - \theta_2 j^2\}}{\sum_{i=0}^{m-1} \exp\{\theta_1 i - \theta_2 i^2\}}$, $\theta_1 = 0.08$, $\theta_2 = \theta_1/10, \theta_1/20, \theta_1/30$;
- Linear Decline: $\omega_j(\theta_1, \theta_2) = \frac{\theta_1 + \theta_2(j-1)}{\theta_1(m) + \theta_2(m)(m+1)/2}$, $\theta_1 = 1$, $\theta_2 = 0.05$;
- Cyclical: $\omega_j(\theta_1, \theta_2) = \frac{\theta_1}{m} \left\{ \sin \left(\theta_2 + 2\pi \frac{j}{m-1} \right) \right\}$, $\theta_2 = 0.01$, $\theta_1 = 5, 5/2, 5/3$;
- Discrete: $\omega_j = (0, 0, \dots, 0, 5/m, \dots, 5/m)$ where the value $5/m$ is assigned to the last one fifth elements and 0 to the rest.

150 For our method, $K = 3$ and $L = 2$ are used, and the tuning parameter in Breitung and Roling [2015] is chosen to minimize AIC. For the evaluation of the estimation accuracy, the root mean square errors (RMSE) of estimators of $\boldsymbol{\beta}^* = (\beta_0^*, \dots, \beta_{m-1}^*)'$ are considered. Our estimator $\widehat{\boldsymbol{\beta}}$ is brought back to the original scale by taking $\mathbf{M}'\widehat{\boldsymbol{\beta}}$. The RMSE of our method is calculated as $RMSE = \|\mathbf{M}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. The number of Monte-Carlo (MC) replications is 1000. For the comparison
155 of forecasting accuracy, the root mean square forecast error (RMSFE) of the one-step-ahead forecast is considered. The number of MC replication is 250. The RMSFE is calculated as following:

1. Obtain the estimated parameter $\widehat{\boldsymbol{\beta}}_{T/2}^*$ in the regression model $y_{t+h} = \mathbf{x}_t' \boldsymbol{\beta}^* + \varepsilon_{t+h}$ for $t = 1, \dots, T/2$.
2. Calculate the one-step-ahead forecast using $\widehat{\boldsymbol{\beta}}_{T/2}^*$, that is, $\widehat{y}_{T/2+h+1} = \mathbf{x}_{T/2+1}' \widehat{\boldsymbol{\beta}}_{T/2}^*$.
- 160 3. Repeat steps 1-2 and obtain $\widehat{y}_{T/2+h+k} = \mathbf{x}_{T/2+k}' \widehat{\boldsymbol{\beta}}_{T/2+k-1}^*$ for $k = 2, \dots, T/2$. Here, $\widehat{\boldsymbol{\beta}}_{T/2+k-1}^*$ is calculated using (y_{t+h}, \mathbf{x}_t') for all $t = k, \dots, T/2 + k - 1$.
4. Once the estimated responses \widehat{y}_{t+h} for $t = T/2 + 1, \dots, T$ are calculated, calculate the RMSFE of the predicted response: $RMSFE = \sqrt{(2/T) \sum_{k=1}^{T/2} (\widehat{y}_{T/2+h+k} - y_{T/2+h+k})^2}$.

Table 1 presents the medians of RMSEs of $\boldsymbol{\beta}$ estimation using Breitung and Roling [2015]
165 (B&R) and our method (Fourier). For both methods, the estimation accuracy generally increases as the frequency ratio or the sample size become larger. Also for all five shapes of MIDAS weights,

our approach substantially improves estimation accuracy compared with B&R’s method. This improvement is more substantial when the sample size T or the frequency ratio m is relatively large. This finding implies that our approach tends to capture the flexibility of various shapes of MIDAS weights more precisely than B&R’s approach. Another notable feature is that α_1 does not have much effect on the accuracy of the estimation for both methods. It seems that for these two nonparametric methods, the MIDAS shape matters, but not the magnitude of the signal. Table 2 presents the median one-step ahead RMSFEs. For both methods, the forecasts become more accurate as the sample size T , or the frequency ratio m increases for all five MIDAS shapes. In general, the Fourier flexible form tends to provide slightly more precise forecasts compared with the B&R’s method. These results show that the proposed the Fourier flexible form approach tends to deliver more accurate estimation and forecasting compared with a competing nonparametric method.

Table 1: Parameter Estimation Accuracy of Nonparametric MIDAS methods for Single Time Series

T	Method	$m = 20$						40			60			150			365			
		$\alpha_1 = 0.2$	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	
Exp Decline	100	B&R	0.9066	0.9195	0.9168	0.6568	0.6561	0.6534	0.5704	0.5618	0.5700	0.4858	0.4811	0.4820	0.4435	0.4455	0.4432	0.0618	0.0898	0.1179
		Fourier	0.5695	0.5829	0.5851	0.2340	0.2340	0.2299	0.1341	0.1306	0.1322	0.0717	0.0931	0.1188	0.0618	0.0898	0.1179			
	200	B&R	0.8630	0.8790	0.8801	0.5811	0.5806	0.5911	0.4940	0.4914	0.4941	0.3962	0.3962	0.3989	0.3814	0.3795	0.3786	0.0594	0.0872	0.1156
		Fourier	0.4162	0.4157	0.4068	0.1560	0.1560	0.1596	0.0918	0.0920	0.0917	0.0627	0.0873	0.1127	0.0594	0.0872	0.1156			
	400	B&R	0.8383	0.8441	0.8489	0.5435	0.5435	0.5378	0.4441	0.4421	0.4443	0.3410	0.3407	0.3418	0.3130	0.3143	0.3126	0.0583	0.0862	0.1146
		Fourier	0.2850	0.2818	0.2851	0.1086	0.1086	0.1093	0.0649	0.0641	0.0649	0.0583	0.0840	0.1104	0.0583	0.0862	0.1146			
Hump Shaped	100	B&R	0.9052	0.9172	0.9172	0.6563	0.6554	0.6537	0.5692	0.5696	0.5620	0.4868	0.4828	0.4820	0.4465	0.4411	0.4437	0.0227	0.0232	0.0240
		Fourier	0.5695	0.5829	0.5851	0.2339	0.2339	0.2298	0.1341	0.1341	0.1307	0.0465	0.0468	0.0459	0.0227	0.0232	0.0240			
	200	B&R	0.8639	0.8796	0.8776	0.5817	0.5804	0.5913	0.4935	0.4935	0.4915	0.4022	0.3993	0.4014	0.3779	0.3795	0.3777	0.0163	0.0164	0.0176
		Fourier	0.4162	0.4157	0.4069	0.1560	0.1560	0.1597	0.0920	0.0920	0.0921	0.0322	0.0321	0.0325	0.0163	0.0164	0.0176			
	400	B&R	0.8390	0.8439	0.8490	0.5435	0.5442	0.5378	0.4440	0.4440	0.4420	0.3407	0.3412	0.3392	0.3133	0.3152	0.3133	0.0222	0.0124	0.0136
		Fourier	0.2850	0.2818	0.2851	0.1085	0.1122	0.1092	0.0651	0.0650	0.0640	0.0222	0.0226	0.0221	0.0116	0.0124	0.0136			
Linear Decline	100	B&R	0.9064	0.9191	0.9147	0.6406	0.6511	0.6448	0.5686	0.5613	0.5661	0.4868	0.4828	0.4822	0.4465	0.4411	0.4437	0.0222	0.0225	0.0223
		Fourier	0.5694	0.5829	0.5851	0.2234	0.2201	0.2164	0.1341	0.1307	0.1350	0.0465	0.0468	0.0459	0.0222	0.0225	0.0223			
	200	B&R	0.8635	0.8786	0.8787	0.5836	0.5829	0.5854	0.4935	0.4917	0.4953	0.4023	0.3993	0.4014	0.3779	0.3795	0.3777	0.0158	0.0153	0.0155
		Fourier	0.4162	0.4157	0.4068	0.1551	0.1537	0.1498	0.0920	0.0920	0.0922	0.0321	0.0321	0.0325	0.0158	0.0153	0.0155			
	400	B&R	0.8379	0.8441	0.8483	0.5294	0.5314	0.5369	0.4433	0.4416	0.4416	0.3406	0.3412	0.3393	0.3133	0.3152	0.3133	0.1087	0.0109	0.0110
		Fourier	0.2850	0.2818	0.2851	0.1046	0.1052	0.1060	0.0651	0.0640	0.0649	0.0222	0.0226	0.0221	0.1087	0.0109	0.0110			
Cyclical	100	B&R	0.9144	0.9256	0.9257	0.6578	0.6538	0.6569	0.5698	0.5611	0.5662	0.4870	0.4828	0.4820	0.4465	0.4411	0.4437	0.0222	0.0225	0.0223
		Fourier	0.5689	0.5825	0.5694	0.2340	0.2297	0.2304	0.1341	0.1307	0.1350	0.0465	0.0468	0.0459	0.0222	0.0225	0.0223			
	200	B&R	0.8677	0.8774	0.8807	0.5796	0.5897	0.5893	0.4935	0.4915	0.4964	0.4022	0.3992	0.4014	0.3779	0.3795	0.3777	0.0158	0.0153	0.0155
		Fourier	0.4163	0.4159	0.4061	0.1560	0.1597	0.1599	0.0916	0.0920	0.0922	0.0321	0.0321	0.0325	0.0158	0.0153	0.0155			
	400	B&R	0.8426	0.8456	0.8472	0.5480	0.5390	0.5340	0.4435	0.4421	0.4405	0.3407	0.3411	0.3391	0.3133	0.3152	0.3133	0.0109	0.0109	0.0110
		Fourier	0.2848	0.2818	0.2850	0.1085	0.1092	0.1101	0.0507	0.0640	0.0649	0.0222	0.0226	0.0221	0.0109	0.0109	0.0110			
Discrete	100	B&R	1.0838	1.2615	1.4540	0.6833	0.7113	0.7555	0.5797	0.5868	0.6062	0.4876	0.4839	0.4849	0.4465	0.4411	0.4440	0.0393	0.0533	0.0683
		Fourier	0.7854	0.9965	1.2264	0.3704	0.4870	0.6196	0.2356	0.3177	0.4095	0.0908	0.1262	0.1634	0.0393	0.0533	0.0683			
	200	B&R	1.0064	1.1517	1.3293	0.6044	0.6375	0.6785	0.5019	0.5105	0.5275	0.4033	0.4008	0.4048	0.3781	0.3796	0.3781	0.0356	0.0503	0.0657
		Fourier	0.6779	0.9027	1.1426	0.3260	0.4562	0.5928	0.2132	0.3028	0.3956	0.0841	0.1209	0.1589	0.0356	0.0503	0.0657			
	400	B&R	0.9444	1.0658	1.2004	0.5636	0.5837	0.6115	0.4505	0.4584	0.4713	0.3408	0.3429	0.3414	0.3135	0.3153	0.3134	0.0336	0.0490	0.0645
		Fourier	0.6053	0.8492	1.1063	0.3055	0.4420	0.5810	0.2030	0.2958	0.3903	0.0807	0.1185	0.1569	0.0336	0.0490	0.0645			

Each cell reports the median of RMSEs of 1000 MC samples, which is further multiplied by 100.

Table 2: One-Step-Ahead Forecasting Accuracy of Nonparametric MIDAS methods for Single Time Series

T	Method	$m = 20$						40			60			150			365			
		$\alpha_1 = 0.2$	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	
Exp Decline	100	B&R	0.2106	0.2076	0.2107	0.1990	0.2001	0.2000	0.1967	0.1984	0.2021	0.2424	0.2320	0.2381	0.3187	0.3186	0.3220	0.3187	0.3186	0.3220
		Fourier	0.1358	0.1366	0.1367	0.1375	0.1380	0.1378	0.1380	0.1366	0.1409	0.1515	0.1630	0.1880	0.2598	0.3504	0.4526	0.2598	0.3504	0.4526
	200	B&R	0.2044	0.2056	0.2062	0.1888	0.1887	0.1881	0.1848	0.1842	0.1830	0.1897	0.1890	0.1895	0.2602	0.2631	0.2603	0.2602	0.2631	0.2603
		Fourier	0.1310	0.1319	0.1303	0.1317	0.1295	0.1296	0.1316	0.1309	0.1310	0.1343	0.1370	0.1421	0.2457	0.3346	0.4289	0.2457	0.3346	0.4289
	400	B&R	0.2015	0.2028	0.2032	0.1837	0.1857	0.1857	0.1760	0.1757	0.1743	0.1680	0.1684	0.1681	0.1926	0.1923	0.1924	0.1926	0.1923	0.1924
		Fourier	0.1280	0.1283	0.1277	0.1280	0.1289	0.1282	0.1276	0.1280	0.1283	0.1316	0.1349	0.1400	0.2376	0.3257	0.4220	0.2376	0.3257	0.4220
Hump Shaped	100	B&R	0.2105	0.2077	0.2107	0.1990	0.2001	0.2001	0.1968	0.1984	0.2021	0.2423	0.2320	0.2381	0.3181	0.3189	0.3228	0.3181	0.3189	0.3228
		Fourier	0.1358	0.1365	0.1367	0.1374	0.1379	0.1374	0.1376	0.1363	0.1397	0.1380	0.1360	0.1371	0.1379	0.1385	0.1416	0.1379	0.1385	0.1416
	200	B&R	0.2045	0.2057	0.2063	0.1887	0.1884	0.1880	0.1850	0.1841	0.1830	0.1897	0.1890	0.1895	0.2601	0.2631	0.2601	0.2601	0.2631	0.2601
		Fourier	0.1310	0.1320	0.1302	0.1317	0.1294	0.1296	0.1314	0.1309	0.1305	0.1311	0.1312	0.1298	0.1296	0.1324	0.1326	0.1296	0.1324	0.1326
	400	B&R	0.2017	0.2028	0.2031	0.1837	0.1859	0.1856	0.1757	0.1756	0.1743	0.1680	0.1684	0.1681	0.1926	0.1923	0.1925	0.1926	0.1923	0.1925
		Fourier	0.1279	0.1283	0.1277	0.1279	0.1287	0.1278	0.1273	0.1274	0.1273	0.1281	0.1279	0.1278	0.1275	0.1285	0.1294	0.1275	0.1285	0.1294
Linear Decline	100	B&R	0.2106	0.2076	0.2107	0.1991	0.2001	0.2002	0.1974	0.1986	0.2021	0.2423	0.2320	0.2381	0.3181	0.3189	0.3228	0.3181	0.3189	0.3228
		Fourier	0.1359	0.1366	0.1368	0.1375	0.1378	0.1377	0.1377	0.1360	0.1402	0.1380	0.1362	0.1376	0.1378	0.1377	0.1405	0.1378	0.1377	0.1405
	200	B&R	0.2045	0.2056	0.2062	0.1887	0.1883	0.1881	0.1851	0.1844	0.1830	0.1897	0.1890	0.1895	0.2601	0.2631	0.2183	0.2601	0.2631	0.2183
		Fourier	0.1311	0.1319	0.1304	0.1317	0.1296	0.1296	0.1313	0.1309	0.1306	0.1314	0.1313	0.1300	0.1292	0.1307	0.1314	0.1292	0.1307	0.1314
	400	B&R	0.2015	0.2028	0.2073	0.1838	0.1857	0.1856	0.1757	0.1757	0.1743	0.1680	0.1684	0.1681	0.1926	0.1923	0.1925	0.1926	0.1923	0.1925
		Fourier	0.1281	0.1284	0.1277	0.1280	0.1288	0.1280	0.1274	0.1276	0.1277	0.1281	0.1280	0.1281	0.1271	0.1275	0.1283	0.1271	0.1275	0.1283
Cyclical	100	B&R	0.2102	0.2078	0.2119	0.1990	0.1999	0.2000	0.1971	0.1983	0.2021	0.2423	0.2320	0.2381	0.3181	0.3189	0.3228	0.3181	0.3189	0.3228
		Fourier	0.1353	0.1367	0.1368	0.1376	0.1381	0.1371	0.1375	0.1367	0.1398	0.1380	0.1362	0.1373	0.1380	0.1374	0.1395	0.1380	0.1374	0.1395
	200	B&R	0.2035	0.2061	0.2063	0.1890	0.1890	0.1880	0.1849	0.1842	0.1832	0.1897	0.1890	0.1890	0.2601	0.2631	0.2601	0.2601	0.2631	0.2601
		Fourier	0.1310	0.1320	0.1301	0.1315	0.1295	0.1297	0.1314	0.1310	0.1305	0.1310	0.1309	0.1309	0.1293	0.1307	0.1308	0.1293	0.1307	0.1308
	400	B&R	0.2017	0.2026	0.2029	0.1839	0.1859	0.1856	0.1758	0.1756	0.1745	0.1680	0.1684	0.1684	0.1926	0.1923	0.1925	0.1926	0.1923	0.1925
		Fourier	0.1280	0.1282	0.1278	0.1278	0.1286	0.1278	0.1273	0.1273	0.1274	0.1281	0.1278	0.1278	0.1270	0.1276	0.1276	0.1270	0.1276	0.1276
Discrete	100	B&R	0.2113	0.2116	0.2140	0.1998	0.2011	0.2005	0.1967	0.1989	0.2029	0.2420	0.2325	0.2382	0.3177	0.3188	0.3236	0.3177	0.3188	0.3236
		Fourier	0.1365	0.1406	0.1427	0.1400	0.1432	0.1447	0.1409	0.1437	0.1507	0.1442	0.1479	0.1566	0.1433	0.1512	0.1626	0.1433	0.1512	0.1626
	200	B&R	0.2039	0.2083	0.2090	0.1891	0.1898	0.1890	0.1853	0.1845	0.1836	0.1899	0.1888	0.1894	0.2603	0.2633	0.2604	0.2603	0.2633	0.2604
		Fourier	0.1320	0.1349	0.1354	0.1335	0.1338	0.1368	0.1336	0.1367	0.1411	0.1361	0.1409	0.1498	0.1360	0.1452	0.1532	0.1360	0.1452	0.1532
	400	B&R	0.2023	0.2039	0.2049	0.1844	0.1861	0.1857	0.1758	0.1757	0.1750	0.1680	0.1682	0.1682	0.1927	0.1923	0.1923	0.1927	0.1923	0.1923
		Fourier	0.1290	0.1312	0.1317	0.1296	0.1331	0.1355	0.1300	0.1328	0.1371	0.1329	0.1387	0.1464	0.1320	0.1411	0.1511	0.1320	0.1411	0.1511

Each cell reports the median of RMSFEs of 250 MC samples.

3 Panel Data and Clustering

In this section, a clustering procedure of MIDAS coefficients for panel data is proposed. The high-frequency regressors are aggregated using nonparametric MIDAS coefficient functions introduced in Section 2 for each cross-section object. These coefficients are further clustered using a penalized regression approach. The linearity of our MIDAS model confers a great advantage to the proposed clustering procedure, as the clustering alone would require quite heavy computations. We first review relevant literature on clustering.

3.1 Literature Review on Clustering Based on Penalized Regression

Clustering using penalized regression based on similarity in the coefficients is a recent development. Su et al. [2016], to our best knowledge, is the first study that develops a clustering algorithm for panel data using penalized regression. Su et al. [2016] modified the traditional Lasso penalty in regression models into classifier-Lasso (C-Lasso) that penalizes the difference between the estimated parameters of each subject and the estimated average parameters of groups. C-Lasso requires a predetermined maximum for the number of groups and a choice of tuning parameter.

Ma and Huang [2017] introduced a penalized method for cross-sectional data, with clustering based on intercepts. The penalty functions used in Ma and Huang [2017] are minimax concave penalty (MCP) [Zhang, 2010] and smoothly clipped absolute deviations penalty (SCAD) [Fan and Li, 2001], which not only share the sparsity properties like Lasso but are also asymptotically unbiased. Later on, Ma and Huang [2016] extended their work, increasing the number of parameters used in clustering. However, neither Ma and Huang [2016] nor Ma and Huang [2017] can be applied to a panel data setting. Indeed, their method is based on strong assumptions that make it nontrivial to extend to panel data.

Zhu and Qu [2018] is the only study, to the best of our knowledge, that extends Ma and Huang's clustering procedure to a data environment similar to panel data. Zhu and Qu [2018] applied Ma and Huang [2017]'s algorithm to repeated cross-section data with one dependent variable and one covariate. In their model, the dependent variable is assumed to vary smoothly in response to the covariate, and this smooth function is estimated using a nonparametric B-spline. Strictly speaking,

Zhu and Qu [2018]’s method is not designed for panel data. However, if a covariate is allowed to vary over time, Zhu and Qu [2018]’s method can be applied to simple panel data. Lv et al. [2019] further extended Zhu and Qu [2018]’s approach allowing for one random effect as an additional covariate.

210 The clustering procedure we propose is based on Ma and Huang [2017] and Ma and Huang [2016]. However, it should also be noted that this extension is nontrivial. In particular, the assumption (C3) in Ma and Huang [2016] requires all variables on the right-hand side of the equation to be non-random and have length exactly 1. This assumption may be appropriate for a clinical trial setting, for which Ma and Huang [2017], Ma and Huang [2016], Zhu and Qu [2018], Lv et al. [2019],
 215 and other related papers are developed. However, this assumption is too strong for a more general panel data setting where time-varying regressors are included. The theory we present circumvent this issue.

3.2 Nonparametric MIDAS for Panel Data and Clustering

Suppose there are n subjects in the cross-section of panel data. For simplicity, all subjects are assumed to have the same sample size T and frequency ratio m . For the i -th subject, let $\mathbf{z}_{i,t}$ be the q -vector of covariates including the intercept at time t ($t = 1, \dots, T$), and let $\boldsymbol{\alpha}_i$ be the corresponding coefficient. Consider the following MIDAS model with lead $h \geq 0$:

$$y_{i,t+h} = \mathbf{z}_{i,t}' \boldsymbol{\alpha}_i + \mathbf{x}_{i,t}' \boldsymbol{\beta}_i^* + \varepsilon_{i,t+h}, \quad t = 1, \dots, T, \quad i = 1, \dots, n,$$

or equivalently,

$$\mathbf{y}_i = Z_i \boldsymbol{\alpha}_i + X_i \boldsymbol{\beta}_i^* + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n, \quad (6)$$

220 where $\mathbf{y}_i = (y_{i,1+h}, \dots, y_{i,T+h})'$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1+h}, \dots, \varepsilon_{i,T+h})'$, $\boldsymbol{\beta}_i^* = (\beta_{i,0}^*, \dots, \beta_{i,m-1}^*)'$, X_i is a $T \times m$ matrix with t -th row being $\mathbf{x}_{i,t}' = (x_{i,t,0}, x_{i,t,1}, \dots, x_{i,t,m-1})$, and Z_i is a $T \times q$ matrix with t -th row being $\mathbf{z}_{i,t}' = (z_{i,t,1}, \dots, z_{i,t,q})$.

We assume that the MIDAS coefficients $\boldsymbol{\beta}_i^*$ takes the Fourier flexible form as in (2). For each subject $i = 1, \dots, n$, $\tilde{X}_i = X_i M'$, where M is from (3). Let $W_i = (Z_i, \tilde{X}_i)$ and $\boldsymbol{\gamma}_i = (\boldsymbol{\alpha}_i', \boldsymbol{\beta}_i')'$. The

225 equation (6) can be rewritten as

$$\mathbf{y}_i = (Z_i, X_i) \begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i^* \end{pmatrix} + \boldsymbol{\varepsilon}_i = (Z_i, \tilde{X}_i) \begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{pmatrix} + \boldsymbol{\varepsilon}_i = W_i \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i \quad (7)$$

Concatenating the \mathbf{y}_i in (7) into \mathbf{y} , a vector of length nT , we have:

$$\mathbf{y} = W\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (8)$$

where $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)'$, $W = \text{diag}(W_1, \dots, W_n)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_n)'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_n)'$. Let $p = q + 2K + L + 1$. In our formulation, $\boldsymbol{\gamma}_i$ is a vector of length p and $\boldsymbol{\gamma}$ is of length np^{vi} .

Remark 1. The arguments in this section should still be valid with different sample sizes and different frequency ratios for different subjects/time periods, at the expense of more complicated notation and slight changes in the results. The major complication arises from the need of using different $M_{i,t}$ for each i and t . That is, $\tilde{\mathbf{x}}_{i,t} = \mathbf{M}_{i,t}\mathbf{x}_{i,t}$, where

$$M_{i,t} = \begin{bmatrix} (0/m_{i,t})^0 & (1/m_{i,t})^0 & \cdots & \{(m_{i,t} - 1)/m_{i,t}\}^0 \\ \vdots & \vdots & & \vdots \\ (0/m_{i,t})^L & (1/m_{i,t})^L & \cdots & \{(m_{i,t} - 1)/m_{i,t}\}^L \\ \sin(2\pi \cdot 1 \cdot 0/m_{i,t}) & \sin(2\pi \cdot 1 \cdot 1/m_{i,t}) & \cdots & \sin\{2\pi \cdot 1 \cdot (m_{i,t} - 1)/m_{i,t}\} \\ \cos(2\pi \cdot 1 \cdot 0/m_{i,t}) & \cos(2\pi \cdot 1 \cdot 1/m_{i,t}) & \cdots & \cos\{2\pi \cdot 1 \cdot (m_{i,t} - 1)/m_{i,t}\} \\ \vdots & \vdots & & \vdots \\ \sin(2\pi \cdot K \cdot 0/m_{i,t}) & \sin(2\pi \cdot K \cdot 1/m_{i,t}) & \cdots & \sin\{2\pi \cdot K \cdot (m_{i,t} - 1)/m_{i,t}\} \\ \cos(2\pi \cdot K \cdot 0/m_{i,t}) & \cos(2\pi \cdot K \cdot 1/m_{i,t}) & \cdots & \cos\{2\pi \cdot K \cdot (m_{i,t} - 1)/m_{i,t}\} \end{bmatrix}$$

should be used, and \mathbf{y} is a vector of length $\sum_{i=1}^n T_i$ rather than nT . As this makes the notation for

230 the subsequent proofs more complicated without adding fundamental differences, this generalization

is not pursued in this paper. In contrast, it is necessary to use the same L and K for all subjects

^{vi}The framework introduced in this section and in Section 2.1 considers only one high-frequency variable. However, our framework can easily extend to accommodate more than one high-frequency variables. For instance, if two high-frequency variables are considered, the length of $\boldsymbol{\gamma}_i$ will be $q + 2K + L + 1 + 2K' + L' + 1$, where $2K'$ and $L' + 1$ are the numbers of trigonometric functions and polynomials considered for the second high-frequency variable. The subsequent clustering procedure is also straightforward.

$i = 1, \dots, n$, to allow for direct comparison of coefficients γ_i .

Consider the estimation of parameters in (8) if the subjects can be separated into a small number of groups. Denote the number of groups by G . The advantage of the proposed procedure is that it does not require any prior knowledge of group information or the number of groups. The only information required is the features of cluster. For example, if we are willing to assume that a cluster has the same parameters of interest—that is, all elements in γ_i are the same within a group—the clusters are identified solely based on parameter estimates.^{vii}

An OLS solution of (8) would minimize $\frac{1}{2} \|\mathbf{y} - W\boldsymbol{\gamma}\|_2^2$, but this would not reflect the relevant group information. To reveal clusters, we propose a penalized regression method to force all elements in γ_i to have similar values within a group. Our method is based on the assumption that if two subjects i and j belong to the same group, the difference of their group-specific parameter would be zero, i.e., $\boldsymbol{\eta}_{ij} = \gamma_i - \gamma_j = \mathbf{0}$. In this case, the OLS estimator of $\boldsymbol{\eta}_{ij}$ would also be somewhat close to a zero vector, though it would not be exactly zero. However, since i and j are in the same group, $\boldsymbol{\eta}_{ij}$ should be better estimated to be exactly zero, rather than “somewhat close” to zero. This can be forced by imposing a penalty for small values of $\boldsymbol{\eta}_{ij}$. In particular, if the number of groups N is much smaller than the number of subjects n , only a small number of $\boldsymbol{\eta}_{ij}$ would be nonzero. The following penalized objective function is considered:

$$Q(\boldsymbol{\gamma}) = \frac{1}{2} \|\mathbf{y} - W\boldsymbol{\gamma}\|_2^2 + \sum_{1 \leq i < j \leq n} \rho_{\theta}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j, \lambda_1), \quad (9)$$

where $\rho_{\theta}(\cdot, \lambda_1)$ is an appropriate penalty function, and θ and λ_1 are tuning parameters that discipline clustering. Clustering using a penalized regression as in (9) has been explored in a number of papers [Ma and Huang, 2017, 2016, Zhu and Qu, 2018, Lv et al., 2019]. As illustrated in the previously-mentioned papers, this optimization problem can be solved using the alternating direction method of multipliers (ADMM) algorithm, which can also be implemented in our setting. Section A.2 in the supplementary material introduces the ADMM algorithm in our setting, proving that the proposed

^{vii}It is possible to relax this assumption by letting some of γ_i be individual-specific, rather than assuming all parameters are strongly tied with groups. If there are subject-specific coefficients, a similar argument would still work, although some rates and conditions would change. In particular, the number of coefficients that are subject-specific should be added following a similar argument to Ma and Huang [2016, 2017]. However, for brevity, this direction will be not elaborated in this paper.

algorithm is convergent. The tuning parameter λ_1 can be chosen by minimizing information criteria such as

$$BIC_{\lambda_1} = \log \left(\frac{\|\mathbf{y} - W\hat{\boldsymbol{\gamma}}\|_2^2}{n} \right) + \frac{\log(n) \cdot (\hat{G}p)}{n}.$$

The rest of this section presents theoretical properties of the estimators that solve the optimization problem in (9). Let G be the true number of groups and \mathcal{G}_g be the set of subject indices that corresponds to the g -th group, for $g = 1, \dots, G$. Assume that each subject belongs to exactly one group; that is, $\mathcal{G}_1, \dots, \mathcal{G}_G$ are mutually exclusive and $\mathcal{G}_1 \cup \dots \cup \mathcal{G}_G = \{1, \dots, n\}$. Denote $|\mathcal{G}_g|$ be the number of elements in \mathcal{G}_g , for $g = 1, \dots, G$. Define $g_{\min} = \min_{g=1, \dots, G} |\mathcal{G}_g|$ and $g_{\max} = \max_{g=1, \dots, G} |\mathcal{G}_g|$. Let $\boldsymbol{\gamma}_i^0$ be the true parameter of the i -th subject, and $\boldsymbol{\varphi}_g^0$ the true common vector for the group \mathcal{G}_g . The common value for the $\boldsymbol{\gamma}_i$ s of the group \mathcal{G}_g is denoted by $\boldsymbol{\varphi}_g$; that is, $\boldsymbol{\gamma}_i = \boldsymbol{\varphi}_g$ for all $i \in \mathcal{G}_g$ and for any $g = 1, \dots, G$. Set $\boldsymbol{\gamma}^0 = (\boldsymbol{\gamma}_1^{0'}, \dots, \boldsymbol{\gamma}_n^{0'})'$, $\boldsymbol{\varphi}^0 = (\boldsymbol{\varphi}_1^{0'}, \dots, \boldsymbol{\varphi}_G^{0'})'$, and $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1', \dots, \boldsymbol{\varphi}_G')'$. Denote the estimated group by $\hat{\mathcal{G}}_g = \{i : \hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\varphi}}_g, 1 \leq i \leq n\}$, for $g = 1, \dots, \hat{G}$, where \hat{G} is the estimated number of groups. For an estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, the corresponding estimated group parameter for the g -th group is defined as $\hat{\boldsymbol{\varphi}}_g = |\hat{\mathcal{G}}_g|^{-1} \sum_{i \in \hat{\mathcal{G}}_g} \hat{\boldsymbol{\gamma}}_i$. Note that $\hat{\boldsymbol{\varphi}}_1, \dots, \hat{\boldsymbol{\varphi}}_{\hat{G}}$ are the distinct values, since the clustering algorithm would lead to $\hat{\boldsymbol{\eta}}_{ij} = \mathbf{0}$ [Ma and Huang, 2017].

Let Π be an $n \times G$ matrix with (i, g) -th element being 1 if i -th subject belongs to the g -th group, and 0 otherwise. Then $\boldsymbol{\gamma} = (\Pi \otimes I_p) \boldsymbol{\varphi} = \Gamma \boldsymbol{\varphi}$, where $\Gamma = (\Pi \otimes I_p)$. An oracle estimator of $\boldsymbol{\gamma}^0$ can be defined as $\hat{\boldsymbol{\gamma}}^{or} = \Gamma \hat{\boldsymbol{\varphi}}^{or}$, where $\hat{\boldsymbol{\varphi}}^{or} = \operatorname{argmin}_{\boldsymbol{\varphi} \in \mathbb{R}^{Gp}} \frac{1}{2} \|\mathbf{y} - W\Gamma \boldsymbol{\varphi}\|_2^2 = (\Gamma' W' W \Gamma)^{-1} \Gamma' W' \mathbf{y}$. The matrix $\Gamma' W' W \Gamma$ is invertible as long as $n \gg G$. Here, the estimator $\hat{\boldsymbol{\gamma}}^{or}$ is called an oracle estimator since it utilizes the knowledge of the true group memberships in Π , which is not feasible in practice. Asymptotic properties of this oracle estimator $\hat{\boldsymbol{\gamma}}^{or}$ will be presented in Theorem 1. Then the asymptotic equivalence of our estimator $\hat{\boldsymbol{\gamma}}$ and the oracle estimator will be introduced in Theorem 2.

Assumption 1. *The number of clusters is much smaller than the number of subjects, i.e., $G \ll n$. In this paper, the case with $G \geq 2$ is considered. The smallest group size g_{\min} is smaller than n/G .*

Assumption 2. *Assume $\lambda_{\min}(\sum_{i \in \mathcal{G}_g} W_i' W_i) \geq c |\mathcal{G}_g| T$, $\lambda_{\max}(\sum_{i \in \mathcal{G}_g} W_i' W_i) \leq c' n T$, and $\max_{1 \leq i \leq n} \lambda_{\max}(W_i' W_i) \leq c'' T$ for some constants c , c' and c'' that do not depend on $g = 1, \dots, G$. Further,*

assume that for any $\epsilon > 0$, there exist $M_1, \dots, M_4 > 0$ such that

$$\begin{aligned} P\left(\sup_{i=1,\dots,n} \|Z'_i Z_i\|_\infty > \sqrt{qT}M_1\right) &< \epsilon, \quad P\left(\sup_{i=1,\dots,n} \|X'_i X_i\|_\infty > \sqrt{mT}M_2\right) < \epsilon, \\ P\left(\sup_{i=1,\dots,n} \|Z'_i X_i\|_\infty > \sqrt{mT}M_3\right) &< \epsilon, \quad P\left(\sup_{i=1,\dots,n} \|X'_i Z_i\|_\infty > \sqrt{qT}M_4\right) < \epsilon. \end{aligned}$$

Assumption 3. The penalty function $\rho_\theta(t) = \lambda^{-1}\rho_\theta(t, \lambda)$ is symmetric, nondecreasing, and concave in t , on $t \in [0, \infty)$. There exists a positive constant c_ρ such that $\rho(t)$ is constant for all $t \geq c_\rho \lambda$. Assume that $\rho(t)$ is differentiable, $\rho'(t)$ is continuous except for a finite number of t , $\rho(0) = 0$, and $\rho'(0+) = 1$.

Assumption 4. There exists a constant $\tilde{c} > 0$ such that

$$E \left\{ \exp \left(\sum_{i=1}^n \sum_{t=1}^T \nu_{i,t} \varepsilon_{i,t} \right) \right\} \leq \exp \left(\tilde{c} \sum_{i=1}^n \sum_{t=1}^T \nu_{i,t}^2 \right)$$

for any real numbers $\nu_{i,t}$, for $i = 1, \dots, n$ and $t = 1, \dots, T$.

Assumption 1 assures sparsity, which is often necessary for the validity of the penalized regression such as (9). We also limit our interest to the case with more than one cluster, but similar arguments also works for the homogeneous case^{viii}. Assumption 2 is reasonable considering the usual assumption that the smallest eigenvalue of $W'_i W_i$ is bounded by cT where T is the sample size and c is some constant. This condition can be relaxed allowing different c_g for different groups. In such a case, our results would not hold if the number of clusters G grows to infinity. It would still work as long as G is finite, by choosing $c = \min_{g=1,\dots,G} c_g$ in the statement of Theorem 1. Assumption 3 is adapted from Ma and Huang [2017] and is conventional in the literature. Popular penalty functions such as MCP and SCAD penalty satisfy this assumption. Assumption 4 holds for any independent subgaussian vector ε , which is commonly assumed in high-dimensional settings.

Remark 2. Assumption 2 is more appropriate for time series data than those in Ma and Huang [2017, 2016]. For instance, assumption (C3) in Ma and Huang [2016] requires, for a given t , $\sum_{i=1}^n z_{i,t,l}^2 = n$, for $l = 1, \dots, q$ and $\sum_{i=1}^n \tilde{x}_{i,t,j}^2 \mathbf{1}\{i \in \mathcal{G}_g\} = |\mathcal{G}_g|$ for $j = 1, \dots, 2K + L + 1$, if the clustering is solely based on $\tilde{x}_{i,t,j}$, but not on $z_{i,t,l}$. Here, $\tilde{x}_{i,t,j}$ are the elements of Fourier transformed

^{viii}The extension to a homogeneous case can be done similarly to that of Ma and Huang [2017].

high-frequency variable $\tilde{\mathbf{x}}_{i,t} = \mathbf{M}\mathbf{x}_{i,t}^{\text{ix}}$. If we were to extend these assumptions to a panel setting, one might modify them to $\sum_{t=1}^T \sum_{i=1}^n z_{i,t,l}^2 = nT$, for $l = 1, \dots, q$ and $\sum_{t=1}^T \sum_{i=1}^n \tilde{x}_{i,t,j}^2 \mathbf{1}\{i \in \mathcal{G}_g\} = |\mathcal{G}_g|T$ for $j = 1, \dots, 2K + L + 1$. These assumptions are unnecessarily strong for panel data. The former assumption, $\sum_{t=1}^T \sum_{i=1}^n z_{i,t,l}^2 = nT$, cannot be satisfied for a time series $z_{i,t,l}$ unless it is properly standardized. Standardizing relevant variables before clustering is often necessary, but only for the variables that are involved in clustering. In this case, clustering is not based on $z_{i,t,l}$, standardizing this variable would add a redundant step that would not even affect the clustering results. The latter assumption, $\sum_{t=1}^T \sum_{i=1}^n \tilde{x}_{i,t,j}^2 \mathbf{1}\{i \in \mathcal{G}_g\} = |\mathcal{G}_g|T$, is also too strong, as it requires standardizing $\tilde{x}_{i,t,j}$ within its true cluster, even before any clustering can be done. To remedy the issues in Ma and Huang [2017, 2016], we lifted these strong assumptions and replaced them with Assumption 2 above, which is more appropriate for time series. Lemmas in Section B.1 of the supplementary material address the issues in proofs due to the absence of these strong assumptions.

The following theorem provides conditions for the convergence of the oracle estimator $\hat{\gamma}^{or}$.

Theorem 1. *If Assumptions 1–4 hold, then*

$$P(\|\hat{\gamma}^{or} - \gamma^0\|_\infty \leq \phi_{n,T,G,\zeta}) \geq 1 - e^{-\iota},$$

where $\phi_{n,T,G,\zeta} = \frac{\sqrt{2\bar{c}}}{c} B_{q,m}^{1/2} \frac{(m\tilde{M}g_{\max})^{1/2}(Gp)^{3/4}}{g_{\min}T^{3/4}} (Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2}$, $B_{q,m} = [q^{1/2} + m^{1/2}(L + 1 + 2K)]^{1/2}$, $\tilde{M} = \max\{M_1, M_2, M_3, M_4\}$, $\iota = \min\{\zeta, -\log(\epsilon)\} - \log(2)$, for ϵ chosen in Assumption 2. Furthermore, if $g_{\min}^3/g_{\max} \gg n^{5/3}T^{1/3}$, for any vector $c_n \in \mathbb{R}^{Gp}$ such that $\|c_n\|_2 = 1$, the asymptotic distribution of $\hat{\gamma}^{or}$ is

$$c_n'(\hat{\gamma}^{or} - \gamma^0) \rightarrow N(0, \sigma_\gamma^2),$$

where $\sigma_\gamma^2 = \text{Var}(\hat{\gamma}^{or} - \gamma^0)$.

The proof of Theorem 1 can be found in Section B in the supplementary material. Theorem 1 implies that with an appropriate choice of $\zeta_{n,T,G}$, the oracle estimator converges to the true

^{ix}Note that this setting is slightly different from our setting, where both $z_{i,t,l}$ and $\tilde{x}_{i,t,l}$ are considered in the clustering procedure. Treatments for variables that are not included in the clustering is described in Section A.5 in supplementary material.

parameter in probability.

Corollary 1. *Under the assumptions of Theorem 1, the oracle estimator $\hat{\gamma}^{or}$ converges to the true parameter γ^0 in probability if one of the following conditions holds:*

1. *The number n is fixed, and $T \rightarrow \infty$.*
- 310 2. *The number $n \rightarrow \infty$, and G is fixed. The number T is either fixed or $T \rightarrow \infty$. Further, the size of the smallest group is large enough such that $g_{min} = O(n^{1/2+\tilde{\alpha}_4})$ for a positive constant $\tilde{\alpha}_4 < 1/2$.*
3. *The number $n \rightarrow \infty$, and $G \rightarrow \infty$. The number T is either fixed or $T \rightarrow \infty$. Further, the size of the smallest group is large enough such that $g_{min} = O(n^{5/7+\tilde{\alpha}_5})$ for a positive constant*
315 *$\tilde{\alpha}_5 < 2/7$.*

Corollary 1 states that the oracle estimator is consistent if n is fixed, or if the size of the smallest group grows somewhat comparably to the increase of n . More specifically, if n is fixed, increasing information across time is necessary for consistent estimation. On the contrary, when increasing information across panel can be obtained, T can be held fixed, as long as all the groups have
320 reasonable sizes.

Theorem 2 demonstrates that the proposed estimator of the parameter γ is equivalent to the oracle estimator with probability approaching to 1, which implies that our estimator converges to the true parameter without prior knowledge of the true group memberships. For our clustering algorithm to work properly, groups should be distinctive enough. Assumption 5 states that the
325 pairwise differences of the true parameters should be large enough for different groups.

Assumption 5. *The minimal difference of the common values between two panels is*

$$b_{n,T,G} = \min_{i \in \mathcal{G}_g, j \in \mathcal{G}_{g'}, g \neq g'} \|\gamma_i^0 - \gamma_j^0\|_2 = \min_{g \neq g'} \|\varphi_g^0 - \varphi_{g'}^0\|_2 > a\lambda_1 + 2p\phi_{n,T,G},$$

for some constant $a > 0$.

Theorem 2. *Assume the conditions of Theorem 1 and Assumption 5 hold. For $\lambda_1 \gg p\phi_{n,T,G}$, where $\phi_{n,T,G}$ is given in Theorem 1, the local minimizer $\hat{\gamma}$ of (9) is almost surely the same as the oracle estimator $\hat{\gamma}^{or}$, if one of the following conditions hold:*

- 330 1. Suppose $n \rightarrow \infty$, and T is fixed. The size of the smallest group is large enough such that
 $(p + 2\sqrt{p} + 2)^{1/2}n^{1/2} \ll g_{\min} = O(n^{7/9+\tilde{\alpha}_0})$ for a positive constant $\tilde{\alpha}_0 < 2/9$;
2. Suppose $n, T \rightarrow \infty$, and G is fixed. The size of the smallest group is large enough such that
 $g_{\min} = O(n^{1/2+\tilde{\alpha}_4})$ for a positive constant $\tilde{\alpha}_4 < 1/2$;
3. Suppose $n, T, G \rightarrow \infty$. The size of the smallest group is large enough such that one of the
335 following conditions is met:
- (a) For a positive constant $\tilde{\alpha}_3 < 2/9$, $\max \left\{ \frac{n^{7/13}}{T^{1/13}}, (p + 2\sqrt{p} + 2)^{1/2}n^{1/2} \right\} \ll g_{\min} = O(n^{7/9+\tilde{\alpha}_3})$;
or,
(b) for a positive constant $\tilde{\alpha}_5 < 2/7$, $g_{\min} = O(n^{5/7+\tilde{\alpha}_5})$.

That is, if one of the above conditions holds, as $nT \rightarrow \infty$,

$$P(\hat{\gamma} = \hat{\gamma}^{or}) \rightarrow 1.$$

Theorem 2 demonstrates that our estimator with prior knowledge of the group information
340 is, asymptotically, as good as the oracle estimator with probability 1, under the presented set of
assumptions.^x There are a couple of differences between the two estimators, $\hat{\gamma}$ and $\hat{\gamma}^{or}$. The first
note-worthy difference is that the oracle estimator converges to the true parameter with probability
1, even when n is fixed, whereas the non-oracle estimator does need $n \rightarrow \infty$. This is expected since
the oracle estimator already knows the true group membership, so that increasing the information
345 in time domain only can make the estimator precise enough. On the contrary, the non-oracle
estimator needs increasing information in cross-section to estimate the group memberships correctly.
The other difference is that the non-oracle needs stronger assumption on the minimum group size.
Again, this is natural, since the non-oracle estimator lacks the group information.

^xThe theoretical results presented in this section handle the case with $G \geq 2$. In the homogeneous panel case, similar arguments can be made, following Ma and Huang [2016]. The details are not presented in this paper, but are available upon request.

3.3 Simulation: Clustering

350 The simulation settings in this section are designed to achieve two goals. The first goal is demonstrating the clustering accuracy of the proposed method in finite samples, and providing a guidance on the choice of the tuning parameters involved in our method, in particular, θ and λ_1 . The other goal is comparing the performance of the proposed method with other clustering algorithms using the penalized regression idea. To our best knowledge, the only other such clustering algorithm 355 is Su et al. [2016] (SSP, hereafter). This method is employed in our MIDAS context by taking the Fourier transformation and applying SSP’s penalty function as in equation (A.4) in Section A.4 in the supplementary material, rather than (9). This method is labeled as “Fourier-SSP.” In addition, our penalty function (9) does not limit how the MIDAS part should be handled. Therefore, B&R’s approach can be adapted in place of the Fourier flexible form for our method. This method is labeled 360 as “B&R-clust.” Our method is labeled as “F-clust”. Sections A.3 and A.4 in the supplementary material provide algorithms and relevant details of theses two additional clustering methods.

Section 3.3.1 provides the finite sample performance of F-clust and B&R-clust for different values of θ . Section 3.3.2 provides guidance on the choice of θ and λ_1 for our method. Using the optimal θ suggested in Section 3.3.2, Section 3.3.3 compares the three methods, F-clust, B&R-clust, and 365 Fouier-SSP, in term of parameter estimation accuracy and forecasting accuracy.

In all simulation settings, two clusters with the exponential decline and the cyclical function shapes shown in Section 2.2 are considered. In each cluster, 15 independent time series are generated. That is, there are 30 coefficient vectors, and two groups are expected after clustering. Each data process follows (5) shown in Section 2.2; $\theta \in \{2, 2.5\}$, $\lambda_1 \in \{1, 1.5, \dots, 4.5\}$, $\beta_0 = 0$, 370 $T \in \{100, 200, 400\}$, $m = \{20, 40\}$, and $\alpha_1 \in \{0.2, 0.3, 0.4\}$ are considered. Notice that α_i are all null-vectors, that is, $\gamma_i = \beta_i$. The ADMM algorithm used for the optimization problem (9) requires one additional parameter, λ_2 . See Algorithm 1 in Section A.2 in the supplementary material for more details. Following the choice of Zhu and Qu [2018], $\lambda_2 = 1$ is used. The clustering algorithm was forced to stop at the 3,000-th iteration if the stopping conditions cannot be satisfied during the 375 process. For the Fourier flexible form and polynomials, $K = 3$ and $L = 2$ is used.

Remark 3. The B&R-clust method involves an additional tuning parameter (θ_{γ^*} in Section A.3

in the supplementary material) to manage the B-spline-like smoothing. According to the simulation results in Breitung and Roling [2015], the choice of θ_{γ^*} is sensitive to the sample size. In all our simulations, comparisons are made with the optimal choice of B&R-clust for each pair of θ and λ_1 . The optimal θ_{γ^*} is chosen in $[0, 100]$ that minimizes AIC. It should be noted that by involving an additional parameter, the B&R-clust method is much more time-consuming than our proposed method, F-clust. This additional parameter also tends to make it difficult to find the optimal values for other tuning parameters by increasing the dimension of the parameter space by one.

Remark 4. In general, Fourier-based methods are much faster than other methods based on B&R's nonparametric MIDAS estimation. This is because our the Fourier flexible form and polynomials reduces the number of parameters from m to $q + 2K + L + 1$. Small values of K and L are generally acceptable. This makes the Fourier-based clustering computationally fast, when m is much larger than K and L .

3.3.1 Clustering Performance

This section explores the clustering accuracy of the proposed method and B&R-clust over a range of θ and λ_1 . As measures of clustering accuracy, the Rand index [Rand, 1971], the adjusted Rand index (ARI) [Hubert and Arabie, 1985], Jaccard Index [Jaccard, 1912], the estimated number of groups \hat{G} , and the median of RMSE of $\hat{\gamma}$ are presented. In particular, the first three measures (Rand, ARI, and Jaccard) assess the similarity of the estimated clusters and the true clusters, and defined as $Rand = \frac{TP + TN}{TP + TN + FP + FN}$, $ARI = \frac{Rand - E(Rand)}{\max(Rand) - E(Rand)}$, and $Jaccard = \frac{TP}{TP + FP + FN}$. Here, TP, TN, FP, and FN indicate true positives, true negatives, false positives, and false negatives, respectively. The estimated number of clusters and median RMSE of estimated $\hat{\gamma}$ are also presented. The RMSE of F-clust is calculated as $RMSE = \sqrt{n^{-1} \sum_{i=1}^n \|\mathbf{M}'\hat{\gamma}_i - \gamma_i^*\|_2^2}$; 200 Monte Carlo samples are generated to evaluate performance.

Table 3 reports clustering indexes, the number of clusters, and medians of RMSE of estimated γ , for $T = 100$, $m = 20$, and $\alpha_1 = 0.4$. When $\theta = 2$, B&R-clust reveals much better clustering performance than our method in general. In particular, B&R-clust presents almost perfect clustering, when λ_1 exceeds 2.5. On the contrary, our method exhibits poor clustering performance; the Rand and Jaccard indexes for our method are about half of that of B&R-clust, and the ARI is almost

Table 3: The Influence of Tuning Parameters (θ and λ_1) on the Clustering Performance

θ	λ_1	Method	Rand	ARI	Jaccard	Clusters	RMSE	$\lambda_{1,BIC}$
2	1	F-clust	0.531	0.030	0.026	26.45	0.5246	3.157
		B&R-clust	0.530	0.756	0.027	27.05	0.4270	
	1.5	F-clust	0.545	0.057	0.059	23.62	0.5741	
		B&R-clust	0.950	0.899	0.899	3.57	0.5984	
	2	F-clust	0.526	0.020	0.021	26.32	0.6197	
		B&R-clust	0.950	0.899	0.899	3.63	0.7630	
	2.5	F-clust	0.483	0.000	0.483	1.00	0.6620	
		B&R-clust	0.995	0.989	0.989	2.17	0.8954	
	3	F-clust	0.517	0.007	0.517	1.07	0.6937	
		B&R-clust	0.998	0.996	0.996	2.05	1.0139	
	3.5	F-clust	0.483	0.000	0.483	1.00	0.7408	
		B&R-clust	0.999	0.998	0.998	2.01	1.1296	
	4	F-clust	0.483	0.000	0.480	1.20	0.7676	
		B&R-clust	0.995	0.989	0.989	2.12	1.2880	
	4.5	F-clust	0.483	0.000	0.483	1.00	0.8055	
		B&R-clust	0.984	0.967	0.966	2.47	1.3130	
	BIC	F-clust	0.498	0.029	0.497	1.06	0.7094	
		B&R-clust	0.951	0.905	0.931	2.51	0.8385	
2.5	1	F-clust	0.671	0.326	0.319	13.52	0.5308	2.190
		B&R-clust	0.962	0.924	0.922	3.19	0.4368	
	1.5	F-clust	0.906	0.810	0.805	5.43	0.5789	
		B&R-clust	0.985	0.983	0.966	2.13	0.6120	
	2	F-clust	0.968	0.935	0.933	3.00	0.6321	
		B&R-clust	0.998	0.996	0.995	2.06	0.7533	
	2.5	F-clust	0.999	0.998	0.998	2.01	0.6618	
		B&R-clust	1.000	1.000	1.000	2.00	0.8597	
	3	F-clust	1.000	1.000	1.000	2.00	0.6897	
		B&R-clust	1.000	1.000	1.000	2.00	0.9593	
	3.5	F-clust	1.000	1.000	1.000	2.00	0.7325	
		B&R-clust	1.000	1.000	1.000	2.00	1.0465	
	4	F-clust	1.000	1.000	1.000	2.00	0.7736	
		B&R-clust	1.000	1.000	1.000	2.00	1.1210	
	4.5	F-clust	1.000	1.000	1.000	2.00	0.8146	
		B&R-clust	1.000	1.000	1.000	2.00	1.1837	
	BIC	F-clust	0.998	0.996	0.996	2.05	0.6534	
		B&R-clust	0.994	0.987	0.987	2.18	0.4388	

200 MC samples, $T = 100$, $\alpha_1 = 0.4$, $m = 20$.

Each cell in the "RMSE" column reports the median of RMSEs of 200 MC samples, which is further multiplied by 100.

405 zero. Nonetheless, it is interesting that in terms of RMSE, our method still is comparable or sometimes better than B&R-clust. However, when $\theta = 2.5$, with a proper choice of tuning parameter λ_1 , the two clustering methods seem to have similar clustering performance. In particular, when λ_1 exceeds 1.5, both methods result in almost perfect clustering. RMSEs seem to be comparable as well.

410 Two conclusions can be drawn from this simulation exercise. One is that choosing the right range of tuning parameters affects the result greatly. The other is that upon the right choice of tuning parameters, both methods lead to reliable clusters, and they both estimate the coefficients quite precisely.

3.3.2 Selection of Tuning Parameters

415 The clustering performance shown above raises the need to carefully choose the tuning parameters, θ and λ_1 , as they affect the clustering performance considerably. In particular, when $\theta = 2$, our clustering method does not work well for the simulation settings we considered, no matter what λ_1 is. Therefore, it is important for users to make a wise choice of θ . Due to the limited knowledge about the true groups in practice, the BIC alone cannot be the right guide to choose the parameters appropriately. In this section, we shall propose a strategy to select the tuning parameters by 420 calculating the globally convex interval.

Let $c_\theta^*(\lambda_1)$ be the minimal eigenvalue of the corresponding design matrix $W(\Pi^* \otimes I_p)/n$, where Π^* contains the estimated group information with the given parameters. Note that Π^* is similar to Π , except that it is built with estimated groups from data rather than the true groups. Following 425 the arguments in Breheny and Huang [2011], it can be shown that a subset of the globally convex regions of θ and λ_1 is given by $\lambda_1 \geq \lambda_1^*$ and θ that satisfy:

$$\begin{aligned} \lambda_1^* &= \inf\{\lambda_1 : \theta > 1/c_\theta^*(\lambda_1)\} \quad \text{if the MCP penalty is used,} \\ \lambda_1^* &= \inf\{\lambda_1 : \theta > 1 + 1/c_\theta^*(\lambda_1)\} \quad \text{if the SCAD penalty is used.} \end{aligned} \tag{10}$$

Here is one strategy to find a convex region:

Step 1 For a given θ , choose λ_1 that minimizes BIC. Denote it as $\lambda_{1,BIC}$.

Table 4: Selection of λ_1 given θ

	Sample	$\theta = 2$	$\theta > 2$
$\lambda_{1,BIC}$	1	4.5	3.5
	2	5.0	4.0
$c_\theta^*(\lambda_{1,BIC})$	1	0.1452	0.0681
	2	0.1420	0.0694
Globally Convex Interval of θ	1	(6.89, ∞)	(14.69, ∞)
	2	(7.04, ∞)	(14.41, ∞)

Step 2 Find $c_\theta^*(\lambda_{1,BIC})$ and λ_1^* .

430 Step 3 Check if $\lambda_1 > \lambda_1^*$ and θ satisfy (10). If not, increase the value of θ and go back to Step 1.

Table 4 presents examples of subsets of convex intervals for an MCP penalty, determined from the simulation settings in Section 3.3.1. Two random samples are considered.

Let us take sample 1 as an example. When $\theta = 2$, the BIC-chosen λ_1 is 4.5, and the subset of the globally convex interval for θ is calculated as $(6.89, \infty)$. Since θ is not in this region, increase
435 the value of θ . Repeat the process with $\theta=2.1$. The convex interval for θ is $(14.68, \infty)$, which does not include θ . We need to increase θ again. As a matter of fact, for our simulation setting, the clustering results was the same for all $\theta = 2.1, 2.2, \dots, 16$, which successfully identify the true clusters. The design matrices are also the same as a result, which leads to the (almost) same choice of $\lambda_{1,BIC}$ and $c_\theta^*(\lambda_{1,BIC})$. Therefore, for this dataset, sample 1, as long as θ is greater 2, we would
440 have an optimal clustering results with a BIC-chosen λ_1 . This observation is consistent with our simulation results. Our method performed well when $\theta > 2$ but not when $\theta = 2$.

3.3.3 Comparison of the three clustering methods

This section compares the three clustering methods (F-clust, B&R-clust, and Fourier-SSP) and the subject-wise nonparametric MIDAS using the Fourier flexible form and polynomials (F-noclust).
445 These methods are compared in terms of the accuracy for parameter estimation in RMSE and for forecasting in RMSFE. For F-clust and B&R-clust, $\theta = 2.5$ is considered, following the suggestion in Section 3.3.2. The frequency ratios m selected in Table 5 are 20 and 40 to save workload on B&R's method. 250 samples are generated in MC simulation. Other than that, the sample size T and the scale α_1 of weights are the same as those considered in Sections 2.2 and 3.3.1. In Fourier-SSP

method, the maximum number of groups is fixed as two for the grid search to save the calculation load, taking advantage of prior knowledge of the true number of clusters. However, in practice, it could be a problem if this number is improperly chosen.

Table 5: Parameter Estimation Accuracy in a Panel Setting

α_1	Method	m	20			40		
		T	100	200	400	100	200	400
0.2	F-clust		0.4031	0.3466	0.3059	0.1587	0.1442	0.1304
	B&R-clust		0.3945	0.3279	0.2261	0.1487	0.1103	0.1005
	Fourier-SSP		9.6804	8.4929	7.9253	8.8707	7.5578	6.2652
	F-noclust		8.2571	5.5480	3.7683	13.7938	8.4324	5.5765
0.3	F-clust		0.5163	0.4691	0.4315	0.2152	0.2012	0.1828
	B&R-clust		0.4306	0.3531	0.2404	0.1670	0.1221	0.0922
	Fourier-SSP		7.4175	6.8505	6.2241	7.2194	5.8779	4.3612
	F-noclust		8.2573	5.5478	3.7685	13.7938	8.3948	5.5765
0.4	F-clust		0.6392	0.5966	0.5558	0.2744	0.2603	0.2145
	B&R-clust		0.4496	0.3663	0.2482	0.1789	0.1364	0.0959
	Fourier-SSP		5.8207	5.4699	5.1157	5.8455	4.8590	4.6615
	F-noclust		8.2573	5.5478	3.7685	13.7938	8.3948	5.5765

Each cell reports the median of RMSEs of 250 MC samples, which is further multiplied by 100.

Table 5 presents median RMSEs of $\hat{\gamma}$. In particular, the RMSE of all Fourier-based methods are calculated as $RMSE = \sqrt{n^{-1} \sum_{i=1}^n \|\mathbf{M}'\hat{\beta}_i - \beta_i^*\|_2^2}$. Measures of clustering accuracy are not presented in this table, because all three clustering methods have perfectly identified the true clusters using BIC. In terms of estimation accuracy, F-clust and B&R-clust tend to outperform Fourier-SSP and the subject-level linear regression. Fourier-SSP and the subject-level linear regression do become more accurate as the sample size increases, but not to the extent that they exceed the accuracy of the other two methods based on the penalized regression with (9). The B&R-clust seems to have the best performance for all settings, while our approach is quite close to the B&R-clust. All three cluster-based method tend to improve as the scale α_1 increases, whereas F-noclust is not affected. This is consistent with the results in Table 5, where Fourier method is not affected much by a different α_1 . In contrast, the three cluster-based methods tend to perform better if the signal is stronger.

Computation is the fastest in F-noclust since it does not involve the penalized optimization. F-clust is the next fastest method, followed by Fourier-SSP^{xi}. B&R-clust is the slowest, taking

^{xi}In our simulations, these two methods have similar computation time. This is because we limit the maximum

at least three times the computation time of our method. F-noclust does not utilize the group information, and parameter estimation tends to be less accurate than in other methods, especially when the sample size T is smaller or the frequency ratio m is larger. The quality does get better at a faster rate than Fourier-SSP as T increases, but to achieve the same amount of accuracy as F-clust or B&R cluster, one would need $T \gg 400$, which is often not possible in practice. When T is relatively small for a given m , using the neighbor information in the same cluster can be one way to improve the quality of the parameter estimation. Therefore, our method (F-clust) successfully identifies true clusters and save computation time substantially, without losing too much accuracy in parameter estimation.

Table 6: One-Step-Ahead Forecasting Accuracy in a Panel Setting

α_1	Method	m	20			40		
		T	100	200	400	100	200	400
0.2	F-clust		0.7700	0.7437	0.7164	0.7591	0.7173	0.7081
	B&R-clust		0.9942	0.7925	0.7214	0.7781	0.7336	0.7139
	Fourier-SSP		2.5779	2.6228	2.7425	2.6582	2.5276	2.4786
	F-noclust		0.1619	0.1401	0.1319	0.2916	0.1617	0.1398
0.3	F-clust		0.7911	0.7591	0.7192	0.7836	0.7150	0.7051
	B&R-clust		0.9937	0.8214	0.7197	0.8010	0.7243	0.7144
	Fourier-SSP		2.4774	2.4952	2.5131	2.5103	2.3803	2.3066
	F-noclust		0.1619	0.1401	0.1319	0.2916	0.1621	0.1398
0.4	F-clust		0.8072	0.7722	0.7290	0.8058	0.7252	0.7176
	B&R-clust		1.0281	0.8336	0.7289	0.8166	0.7277	0.7257
	Fourier-SSP		2.2377	2.2315	2.2493	2.2844	2.1698	2.0982
	F-noclust		0.1619	0.1401	0.1319	0.2916	0.1621	0.1398

Each cell reports the median of RMSFEs of 250 MC samples.

Table 6 presents the median RMSFEs of the one-step-ahead forecast. The RMSFEs are computed in a similar way as presented in Section 2.2, replacing $\widehat{\beta}^*$ with the one obtained from the penalized regression (9), and $RMSFE = \sqrt{(nT/2)^{-1} \sum_{k=1}^{T/2} \sum_{j=1}^n (\widehat{y}_{j,T/2+h+k} - y_{j,T/2+h+k})^2}$. It is worth noting that F-noclust outperforms all the cluster-based approaches. This is somewhat expected, as $\widehat{\beta}^*$ from the subject-level regression is supposed to be the most efficient estimator of β^* among all unbiased estimators under our set of assumptions. Nonetheless, F-clust and B&R-clust provide reasonably accurate forecast compared to the Fourier-SSP method. This observation demonstrates that our penalty functions in (9) may perform better than SSP's penalty functions, number of groups of Fourier-SSP to 2, utilizing the true group information, which saves the computation time considerably. In reality, our method is faster when the true group information cannot be used.

both in terms of estimation and forecast accuracy in a setting similar to ours.

485 Overall, if one is interested in identifying clusters in a panel MIDAS data without prior knowledge on group structures, it seems that our method performs reasonably well without requiring too heavy computations.

4 Heterogeneity in Labor Market Dynamics across States: Through the Lens of a Mixed-Frequency Okun’s Law Model

490 With the new method, we explore the heterogeneity in labor market dynamics across states through the lens of a mixed-frequency Okun’s law model.

4.1 Panel Data of State-Level Labor Markets and Model Description

Okun’s law refers to the empirical negative correlation between output growth and unemployment rate. A popular specification often adopted in the literature (e.g., Knotek II [2007]) is the following. Let u_t be the first-differenced unemployment rate and y_t be the growth rates of GDP. 495 Okun’s law is a linear relationship between these two variables

$$u_t = \delta + \alpha y_t + \varepsilon_t,$$

where δ is a constant, ε_t is an error term and the coefficient α has a negative sign.^{xii}

It has been observed that an Okun’s law model might encounter difficulties in dealing with a sudden and abrupt rise in the unemployment rate due to a burst of job losses at the inception of an economic downturn (e.g., Lee [2000], Moazzami and Dadgostar [2011], Kargı [2016]). In other 500 words, an Okun’s law model with GDP growth as the sole explanatory variable is likely to have difficulties in explaining the nonlinear feature of unemployment dynamics. Weekly initial claims have the highest frequency among the publicly available labor market indicators, and thus can capture the magnitude of job loss in a timely manner. In this regard, the Okun’s law model with 505 weekly initial claims can better capture the non-linearity in unemployment dynamics and also can be used to nowcast the unemployment rate on a weekly basis in real time.

^{xii}This specification is often referred to as the differenced version of Okun’s law.

The variables that we use for the mixed-frequency Okun’s law model are the quarterly growth rate of log GDP in state i in quarter t ($y_{i,t}$), the first-differenced unemployment rate of state i in quarter t ($u_{i,t}$), and the log of initial claims in week j of quarter t in state i ($x_{i,t,j}$).^{xiii}

510 We consider 50 states and the District of Columbia, a total 51 cross-section units (or subjects).^{xiv} The sample period is from 2005 to the second quarter of 2018, as the quarterly real GDP at the state level is available from 2005.

The mixed-frequency Okun’s law model is specified as follows:

$$u_{i,t} = \delta_i + \alpha_i y_{i,t} + \mathbf{x}_{i,t}' \boldsymbol{\beta}_i^* + \varepsilon_{i,t},$$

where $\mathbf{x}_{i,t} = (x_{i,t,1}, \dots, x_{i,t,m_t})'$ is the collection of weekly initial claims of the corresponding quarter. One complication of the mixed-frequency Okun’s law model is that the distributed lag structure of weekly initial claims coefficient $\boldsymbol{\beta}_i^*$ is not well defined, as a quarter has a different number of weeks ranging from 12 to 14. In this case, the construction of a MIDAS model usually requires a more complicated parameterization to cope with these irregular frequencies. Notably, our method does not require such a procedure. The proposed MIDAS model can flexibly handle the changing number of MIDAS parameters, as the algorithm allows the Fourier transformation matrix $M_{i,t}$ to vary over time as noted in Remark 1. The Fourier-transformed log initial claims can be written as $\tilde{\mathbf{x}}_{i,t} = \mathbf{M}_{i,t} \mathbf{x}_{i,t}$, and now the model is re-specified as follows:

$$u_{i,t} = \delta_i + \alpha_i y_{i,t} + \tilde{\mathbf{x}}_{i,t}' \boldsymbol{\beta}_i + \varepsilon_{i,t},$$

where $\boldsymbol{\beta}_i = (\beta_{i,1}, \dots, \beta_{i,2K+L+1})'$ and L and K are the number of parameters in Fourier approximation. We cluster states based on α_i and β_i , as these parameters capture the dynamic features
515 of unemployment in state i . States that share similar values for $(\alpha_i, \boldsymbol{\beta}_i')$ are allocated to the same group.

^{xiii}The state-level GDP growth is from Bureau of Economic Analysis, the state-level unemployment rate is from Local Area Unemployment Statistics by Bureau of Labor Statistics, and the state-level initial claims are from Department of Labor. We seasonally adjust initial claims using seasonal-trend decomposition using LOESS (STL), and use the seasonally adjusted claims for the estimation of Okun’s law model.

^{xiv}In some states, there is a small number of weeks when the initial claims data are not released, due, for instance, to the shutdown of a local agency collecting the data.

In our clustering algorithm, $L = 1$ and $K = 2$ are chosen to effectively summarize the high-frequency information.^{xv} This selection ensures that the total number of parameters is smaller than in the conventional MIDAS model. For this dataset, we could not find a global convex area for θ and λ_1 . Instead, we conducted a grid search on a range of λ_1 for each $\theta = 2, 3, 5, 8, 10$. Among the values on the grid, we found that $\theta = 2$ and $\lambda_1 = 2.6$ minimize both AIC and BIC criteria and therefore our reported results are based on these values.

4.2 Clustering Analysis for the State-Level Labor Markets in the United States

Table 7 summarizes estimation results. The algorithm identifies four clusters. There are 24 states in cluster 1, 19 states in cluster 2, 7 states in cluster 3, and 1 state in cluster 4. Clusters 1, 2, and 3 account for 47.0%, 36.4%, and 15.3% of national payroll employment, respectively. Cluster 4 consists of a single state—Louisiana—and constitutes 1.4% of aggregate employment. The clusters are determined jointly by the coefficients on GDP growth and the coefficients on log weekly initial claims. Based on the absolute size of coefficients on GDP growth and log initial claims (columns 5 and 6 of Table 7), the labor markets of cluster 3 are the most cyclically sensitive, while those in clusters 2 and 1 are moderately and weakly cyclical, respectively. Quite differently, the coefficient on GDP growth in cluster 4 is close to zero and statistically insignificant, but the sum of coefficients on log initial claims is positive and statistically significant: hence, in cluster 4 GDP growth rate does not affect the unemployment rate, while initial claims do. The clusters are further distinguished by the pattern of coefficients on log weekly initial claims. The estimated trajectories within the quarter are plotted in Figure 1, and summarized in Column 7 of Table 7. These trajectories are quite distinct across clusters as shown in Figure 1. The coefficients exhibit an uptrend in cluster 1, while those in cluster 2 and 3 show “W ” and “M ” shapes, respectively. The coefficients of Cluster 4 show an “N ” shape. This result suggests that both the trajectory and the size of these coefficients are important in distinguishing clusters.

The large coefficients on certain weeks’ initial claims suggest that those who file for UI benefits during these weeks are more likely to raise the state’s unemployment rate than others. This might

^{xv}The clustering results were similar to unreported results with $L = 2$ and $K = 3$.

Table 7: Summary of identified clusters

	Number of states	Member states	Emp. share	GDP Coeff.	Sum of IC Coeff.	IC coeff.'s Shape
Cluster 1	24	South Carolina, North Carolina, Florida, Wisconsin, Colorado, Rhode Island, Iowa, South Dakota, Kansas, North Dakota, Hawaii, Indiana, Wyoming, Oklahoma, New Hampshire, New Jersey, Maine, Michigan, Vermont, Nebraska, California, Delaware, New York, Alaska	47.0%	-0.141 (0.00925)	0.862	Upward sloping
Cluster 2	19	Georgia, Oregon, Ohio, Utah, Tennessee, Texas, New Mexico, West Virginia, Missouri, Mississippi, Arkansas, Massachusetts, Kentucky, District of Columbia, Massachusetts, Idaho, Pennsylvania, Montana, Connecticut	36.4%	-0.203 (0.0150)	0.972	W-shape
Cluster 3	7	Alabama, Arizona, Illinois, Washington, Nevada, Minnesota, Virginia	15.3%	-0.313 (0.0280)	1.468	M-shape
Cluster 4	1	Louisiana	1.4%	0.0250 (0.0625)	1.244	N-shape

Note to Table 7: The abbreviation "Emp. share" refers to the share out of aggregate payroll employment; "GDP Coeff." refers to the coefficient on GDP growth; "IC coeff.'s shape" refers to the shape of coefficients on the weekly initial claims through the corresponding quarter. Numbers in the parentheses are the standard errors.

be related hirings and layoffs practices and the timing of regular employment turnovers in different states or regions. Layoffs related to temporary hirings might be concentrated in particular weeks in some states. Workers previously hired by the firms who periodically lay off and recall their workers typically file for UI claims during specific weeks of quarter and are pretty quickly re-employed. On the other hand, those who file for UI benefits outside these weeks might be more likely to be permanent job losers who tend to stay unemployed for a longer period. Therefore, the initial claims filed by these workers tend to be more strongly correlated with the unemployment rate than those filed by temporary job losers. As an example, in cluster 2 more permanent job losers might file for UI claims during weeks 1, 5, 6, 11 and 12 than in other weeks. Quite differently, in cluster 1 where the coefficients on initial claims exhibit an upward trend, temporary layoffs might be concentrated early in the quarter. In synthesis, each cluster's coefficients pattern might reflect these institutional factors. Hence, the different shapes of coefficients can be interpreted as the outcomes of labor market conventions that differ across clusters.^{xvi}

^{xvi}The large positive coefficients observed on week 5, 6, 11, and 12 might also be related to the reference week of Current Population Survey that usually falls in the second week of each month. The number of those file for the

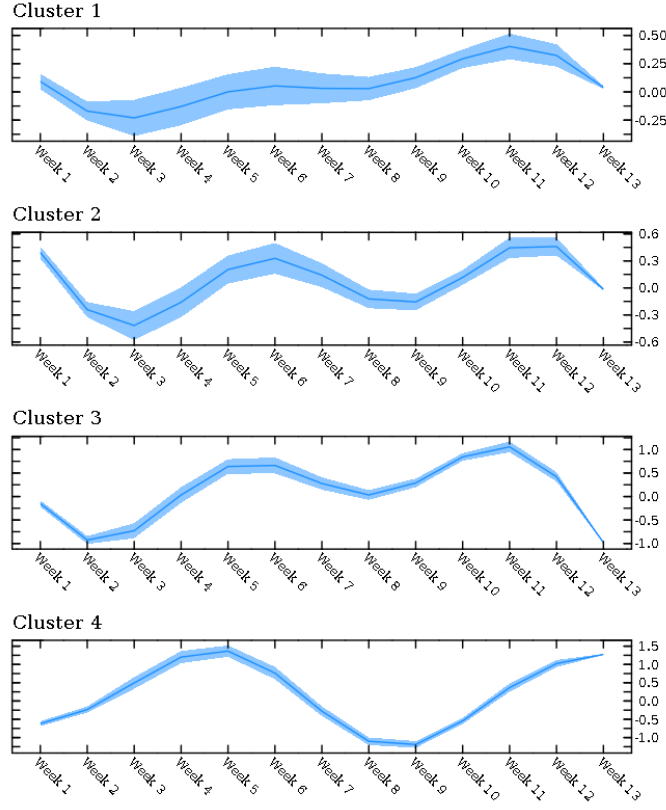


Figure 1: Coefficients on log initial claims by cluster

Note to Figure 1: Authors' calculations. The shaded area denotes the 95% confidence intervals.

Figure 2 displays the geographical locations of clusters. Cluster 1, denoted as light blue, is composed of (1) agricultural states in the Midwest region, (2) manufacturing states in the East-north central region, and (3) states in the Northeast. Far from the states in cluster 1, however, California, Alaska, Florida, and North and South Carolina also belong to Cluster 1.^{xvii} Cluster 2 denoted as pink is broadly composed of (1) agricultural states in the West (mountain region), (2) states in the central South region, and (3) manufacturing states in the middle Atlantic region of the Northeast. Overall, states that belong to the same cluster are adjacent with each other in cluster 1 and 2. Quite differently, states in cluster 3 denoted as orange are widely dispersed.

UI claims in the first half of month might be highly correlated with a change in the unemployment rate captured by the survey, if the recent filing of UI claims make the survey respondent more likely to report unemployment as their labor force status. However, this feature is not clearly observed in other clusters. Therefore, the pattern of coefficients on initial claims is less likely to be the outcome of reference-week effect.

^{xvii}We follow the Census Bureau's division of regions.

565 This observation suggests that the geographical proximity is not a necessary condition for the identification of a cluster, but is only partially correlated with cluster membership. This might be because adjacent states often share similar structural characteristics, such as available natural resources, oil production, and industrial structure.

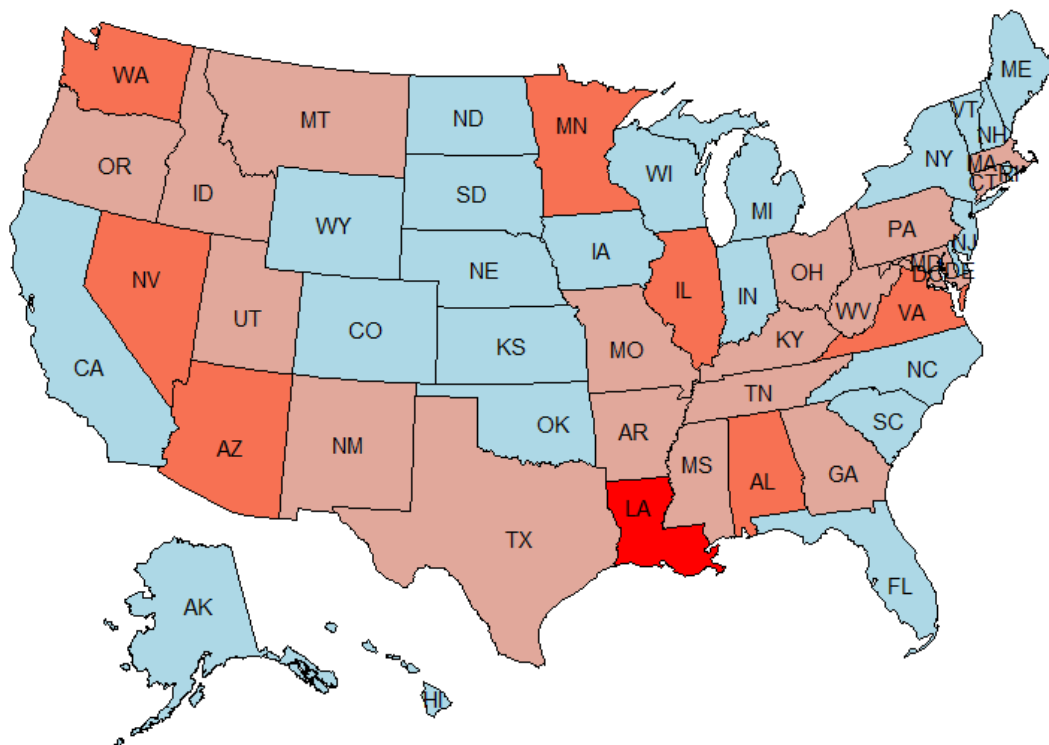


Figure 2: States by cluster (cluster 1=light blue, cluster 2=pink, cluster 3 = orange, cluster 4=red)
 Note to Figure 2: Authors' calculations.

We further relate the clusters to observable state characteristics in order to find an economic interpretation. To this end, we consider five variables: (1) the small firms share, (2) the employment share of manufacturing, (3) the employment share of finance industry, (4) the GDP share of oil production, and (5) the fraction of long-term unemployment on total unemployment. The first four characteristics are considered in Hamilton and Owyang [2012] as possible explanations for heterogeneous regional business cycles.^{xviii} We also include the share of long-term unemployment,

^{xviii}Following this study, we also analyze the clusters based on these attributes. The state-level data of four variables

Table 8: Features of each cluster

	Small-firm share	Manufacturing intensive	Finance intensive	Oil- producing	Long-term unemployment
Cluster 1	0.54	0.46	0.46	0.17	0.38
Cluster 2	0.26	0.53	0.37	0.26	0.63
Cluster 3	0.14	0.43	0.43	0	0.43
Cluster 4	1	0	0	1	0

Note to Table 8: Authors' calculations. Numbers in red are larger than 0.5; those in blue are between 0.4 and 0.5.

575 a component that is likely to reflect the structural unemployment.^{xix} The unemployment rate of a state where the share of long-term unemployment is high might be less responsive to changes in labor demand.

We find that the four clusters are moderately distinct in these five observable dimensions. Table 8 and Figure 3 summarize the observable features of each cluster. The feature of each cluster is
580 computed from the fraction of states in the cluster whose particular observable characteristic is more prominent than the average of all states. For example, according to the second column of Table 8, the fraction of states in cluster 1 whose small firms share is larger than the average of all states is 0.54, and that in cluster 2 is 0.26.

Cluster 1 is summarily described as small-firm/manufacturing/finance intensive. More than half
585 of the states in this cluster have the share of small firms higher than the average of all states. At the same time, a little less than a half of the states in this cluster have above average employment share of manufacturing.

Cluster 2 is characterized as long-term-unemployment prone and manufacturing intensive. About 60% of states in this cluster have a higher than average share of long-term unemployment, and a
590 little more than half of the states have manufacturing shares in employment above the average.

Cluster 3 is characterized as manufacturing-finance intensive and long-term unemployment prone. Three out of seven states have larger than average fraction of employment in manufacturing and finance. In addition, the three states have above-average long-term unemployment shares.

Louisiana (cluster 4) is an oil-producing state, whose share of small firms is larger than average.

595 Summing up, clusters are heterogeneous in multiple dimensions, characterized by differences

are from Hamilton and Owyang [2012].

^{xix}The fraction is calculated based on the micro data from the Current Population Survey (CPS).

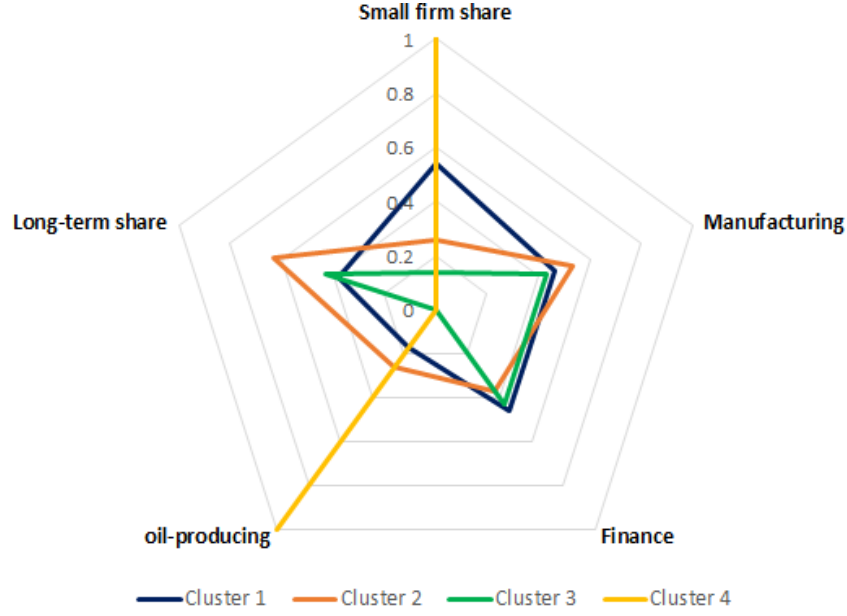


Figure 3: Features of each cluster
Note to Figure 3: Authors' calculations.

in several observable attributes, as shown in Figure 3. In synthesis, the empirical application demonstrates that our algorithm is able to reveal meaningful heterogeneity in labor-market dynamics across states without requiring prior knowledge, which in many cases derives from data limitations or from theories lacking empirical support.

5 Conclusion

This paper proposed a new clustering method in a panel MIDAS setting, grouping subjects with similar MIDAS coefficients. The clustering is purely data driven, and relies on appropriate adaptation of an existing penalized regression approach. The major advantage of our method is that it does not require prior knowledge of true group membership, not even of the number of groups. A penalized regression already requires at least two tuning parameters, which are often difficult to choose, yet this choice is crucial, as we show in our simulations. In our nonparametric MIDAS specification, based on the Fourier flexible form and polynomials, only a single tuning parameter is required. We provide a strategy for choosing the tuning parameter based on a convex region approach. We show that our proposed clustering method works well both asymptotically

and in finite samples. In addition, our non-parametric approach is very simple and fast, and leads to parameter estimates that are as accurate as competing methods, in many cases even more accurate. As an empirical example, we provide an application to labor market dynamics at state level in the United States. The application, based on a mixed frequency Okun’s law model, allows us to categorize the states in four meaningful clusters that correspond to relevant and measurable differences along different dimensions.

References

- E. Andreou, E. Ghysels, and A. Kourtellis. Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158:246–261, 2010.
- A. Babii, E. Ghysels, and J. Striaukas. Estimation and HAC-based Inference for Machine Learning Time Series Regressions. *mimeo*, 2019.
- R. Becker, W. Enders, and S. Hurn. A general test for time dependence in parameters. *Journal of Applied Econometrics*, 19(7):899–906, 2004.
- R. Becker, W. Enders, and J. Lee. A stationarity test in the presence of an unknown number of smooth breaks. *Journal of Time Series Analysis*, 27(3):381–409, 2006.
- P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011.
- J. Breitung and C. Roling. Forecasting inflation rates using daily data: A nonparametric midas approach. *Journal of Forecasting*, 34(7):588–603, 2015.
- W. Enders and J. Lee. A unit root test using a fourier series to approximate smooth breaks. *Oxford Bulletin of Economics and Statistics*, 74(4):574–599, 2012.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- A. R. Gallant. On the bias in flexible functional forms and an essentially unbiased form: the fourier flexible form. *Journal of Econometrics*, 15(2):211–245, 1981.

- 635 E. Ghysels, A. Sinko, and R. Valkanov. Midas regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90, 2007.
- B. Güriş. A New Nonlinear Unit Root Test with Fourier Function. MPRA Paper 82260, University Library of Munich, Germany, Oct. 2017.
- J. D. Hamilton and M. T. Owyang. The propagation of regional recessions. *Review of Economics and Statistics*, 94(4):935–947, 2012.
- 640 L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.
- P. Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.
- B. Kargı. Okun’s law and long term co-integration analysis for oecd countries (1987-2012). *EMAJ: Emerging Markets Journal*, 6(1), 2016.
- 645 E. S. Knotek II. How useful is okun’s law? *Economic Review-Federal Reserve Bank of Kansas City*, 92(4):73, 2007.
- J. Lee. The robustness of okun’s law: Evidence from oecd countries. *Journal of macroeconomics*, 22(2):331–356, 2000.
- Y. Lv, X. Zhu, Z. Zhu, and A. Qu. Nonparametric cluster analysis on multiple outcomes of longitudinal data. *Statistica Sinica*, page Accepted, 2019.
- 650 S. Ma and J. Huang. Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*, 2016.
- S. Ma and J. Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.
- 655 B. Moazzami and B. Dadgostar. Okun’s law revisited: evidence from oecd countries. *Int. Bus. Econ. Res. J.*, 8:21–24, 02 2011. doi: 10.19030/iber.v8i8.3156.
- P. Perron, M. Shintani, and T. Yabu. Testing for flexible nonlinear trends with an integrated or stationary noise component. *Oxford Bulletin of Economics and Statistics*, 79(5):822–850, 2017.

- W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American*
660 *Statistical Association*, 66(336):846–850, 1971.
- P. M. M. Rodrigues and A. M. Robert Taylor. The flexible fourier form and local generalised least
squares de-trended unit root tests. *Oxford Bulletin of Economics and Statistics*, 74(5):736–759,
2012.
- L. Su, Z. Shi, and P. C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84
665 (6):2215–2264, 2016.
- C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):
894–942, 04 2010.
- X. Zhu and A. Qu. Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of*
Statistics, 12:171–193, 01 2018.

Supplementary Material to “Revealing Cluster Structures Based on Mixed Sampling Frequencies”

YEONWOO RHO^{1*}, YUN LIU², AND HIE JOO AHN³

¹Michigan Technological University

²Quicken Loans

³Federal Reserve Board

September 9, 2020

This supplementary material consists of two parts: Section A consists of details of algorithms used in the main paper; and Section B presents all proofs.

A Algorithms

This section contains details of algorithms introduced in the main paper. Section A.1 introduce details of B&R’s nonparametric MIDAS in our setting in Section 2. Section A.2 present our clustering algorithm (F-clust). Details on how to solve the optimization problem in (9) in our setting is presented using the alternating direction method of multipliers (ADMM) algorithm. The proposed algorithm is also shown to be convergent. Sections A.3 and A.4 present the details of the two competing clustering methods. In particular, Section A.3 introduces how to combine the penalized regression approach with objective function (9) and the B&R’s method (B&R-clust). Section A.4 present the algorithm combining Su’s penalty function and the Fourier transformation for MIDAS (Fourier-Su). Section A.5 presents the algorithm to exclude a part of parameters from clustering.

A.1 Breitung and Roling [2015]’s Nonparametric MIDAS

The nonparametric MIDAS in Breitung and Roling [2015] is based on the discrete form of the cubic smooth spline. The least-squares objective function is penalized by the sum of the second

*Address of correspondence: Yeonwoo Rho, Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA. (yrho@mtu.edu)

Emails: Y. Rho (yrho@mtu.edu), Y. Liu (yliu26@mtu.edu), and H. J. Ahn (HieJoo.Ahn@frb.gov)

difference of weights to balance the goodness of fit and the smoothness of weights. Assume that the MIDAS model is shown in (1). The penalized least-squares objective function is

$$Q_{BR} = \sum_{t=1}^T \left(y_{t+h} - \alpha_0 - \sum_{i=0}^{m-1} x_{t,i} \beta_i^* \right)^2 + \lambda_{BR} \sum_{i=2}^m (\nabla^2 \beta_i^*)^2,$$

where $\nabla^2 \beta_i^* = (\beta_i^* - 2\beta_{i-1}^* + \beta_{i-2}^*)$ indicates the second difference of weights. The smoothed least-squares (SLS) estimator [Breitung and Roling, 2015] becomes

$$\hat{\beta}_{BR}^* = \arg \min_{\beta^*} \{(\mathbf{y} - \mathbf{X}\beta^*)'(\mathbf{y} - \mathbf{X}\beta^*) + \lambda_{BR}(\mathbf{D}\beta^*)'\mathbf{D}\beta^*\},$$

where

$$D_{(m-2) \times (m+1)} = \begin{pmatrix} 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}.$$

The tuning parameter λ_{BR} can be chosen using an information criteria. For example, Breitung and Roling [2015] proposed to use the modified Akaike information criterion (AIC),

$$AIC_{\lambda_{BR}} = \log \{(\mathbf{y} - \hat{\mathbf{y}}_{BR})'(\mathbf{y} - \hat{\mathbf{y}}_{BR})\} + \frac{2(s_{\lambda_{BR}} + 1)}{T - s_{\lambda_{BR}} + 2},$$

where $\hat{\mathbf{y}}_{BR} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda_{BR}\mathbf{D}'\mathbf{D})^{-1}\mathbf{X}'\mathbf{y}$.

A.2 Clustering algorithm for the Fourier Transformed data

The optimization problem in (9) is not trivial. The alternating direction method of multipliers (ADMM) algorithm by Boyd et al. [2011] has been successfully employed solving this optimization problem [Ma and Huang, 2017, Zhu and Qu, 2018]. This section introduces the ADMM algorithm in our setting and proves that it is convergent.

By introducing $\boldsymbol{\eta}_{ij} = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j$, minimizing (9) is equivalent to minimizing

$$Q(\boldsymbol{\gamma}, \boldsymbol{\eta}) = \frac{1}{2} \|\mathbf{y} - W\boldsymbol{\gamma}\|_2^2 + \sum_{1 \leq i < j \leq n} \rho(\boldsymbol{\eta}_{ij}, \lambda_1) \quad \text{subject to} \quad \boldsymbol{\eta}_{ij} = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}'_{12}, \dots, \boldsymbol{\eta}'_{n-1,n})'$. Following Boyd et al. [2011], this constrained optimization problem can be solved using a variant of the augmented Lagrangian

$$Q_{\lambda_2}(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{y} - W\boldsymbol{\gamma}\|_2^2 + \sum_{i < j} \rho(\boldsymbol{\eta}_{ij}, \lambda_1) + \frac{\lambda_2}{2} \sum_{i < j} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j - \boldsymbol{\eta}_{ij}\|_2^2 + \sum_{i < j} \boldsymbol{\xi}'_{ij}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j - \boldsymbol{\eta}_{ij}), \quad (\text{A.1})$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}'_{12}, \boldsymbol{\xi}'_{13}, \dots, \boldsymbol{\xi}'_{n-1,n})'$ and $\boldsymbol{\xi}_{ij}$ are p -vectors of Lagrangian multipliers. As proposed by Boyd et al. [2011], the optimization problem in (A.1) can be solved using the alternating direction method of multipliers (ADMM) algorithm. At the $(s+1)$ -th step of the ADMM algorithm, estimated parameters $\boldsymbol{\gamma}^{s+1}$, $\boldsymbol{\eta}^{s+1}$ and $\boldsymbol{\xi}^{s+1}$ are updated as

$$\begin{aligned} \boldsymbol{\gamma}^{s+1} &= \arg \min_{\boldsymbol{\gamma}} Q_{\lambda_2}(\boldsymbol{\gamma}, \boldsymbol{\eta}^s, \boldsymbol{\xi}^s), \\ \boldsymbol{\eta}^{s+1} &= \arg \min_{\boldsymbol{\eta}} Q_{\lambda_2}(\boldsymbol{\gamma}^{s+1}, \boldsymbol{\eta}, \boldsymbol{\xi}^s), \\ \boldsymbol{\xi}^{s+1}_{ij} &= \boldsymbol{\xi}^s_{ij} + \lambda_2(\boldsymbol{\eta}^{s+1}_{ij} - \boldsymbol{\gamma}_i^{s+1} + \boldsymbol{\gamma}_j^{s+1}), \end{aligned} \quad (\text{A.2})$$

where $\boldsymbol{\eta}^s$ and $\boldsymbol{\xi}^s$ are the estimates in the s -th iteration. By collecting terms related to $\boldsymbol{\gamma}$, the first function in (A.2) is equivalent to minimizing

$$Q_{\lambda_2}^{\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{y} - W\boldsymbol{\gamma}\|_2^2 + \frac{\lambda_2}{2} \|D\boldsymbol{\gamma} - (\boldsymbol{\eta} + \boldsymbol{\xi}/\lambda_2)\|_2^2,$$

where $D_{ij} = (\mathbf{e}_i - \mathbf{e}_j)' \otimes I_p$, $D = (D'_{12}, D'_{13}, \dots, D'_{n-1,n})'$, \mathbf{e}_i is an n -dimension vector with the i -th element as one and the rest as zeros, and I_p is an identity matrix with rank p . Therefore, $\boldsymbol{\gamma}^{s+1} = (W'W + \lambda_2 D'D)^{-1} \{W'\mathbf{y} + \lambda_2 D'(\boldsymbol{\eta}^s + \boldsymbol{\xi}^s/\lambda_2)\}$.

The MCP is shown to be nearly unbiased and is applicable here to update $\boldsymbol{\eta}^{s+1}$ [Zhu and Qu, 2018]. The penalty function of MCP is $\rho(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j, \lambda_1) = \rho_{\theta}(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2, \lambda_1)$ where $\rho_{\theta}(x,$

$t) = t \int_0^x (1 - \frac{u}{\theta t})_+ du$. As a consequence, when the MCP is selected, $\boldsymbol{\eta}_{ij}^{s+1}$ can be updated by

$$\boldsymbol{\eta}_{ij}^{s+1} = \begin{cases} \tilde{\boldsymbol{\eta}}_{ij}^{s+1} & \text{if } \|\tilde{\boldsymbol{\eta}}_{ij}^{s+1}\|_2 \geq \theta \lambda_1, \\ \frac{\theta \lambda_2}{\theta \lambda_2 - 1} \left(1 - \frac{\lambda_1 / \lambda_2}{\|\tilde{\boldsymbol{\eta}}_{ij}^{s+1}\|_2} \right) \tilde{\boldsymbol{\eta}}_{ij}^{s+1} & \text{if } \|\tilde{\boldsymbol{\eta}}_{ij}^{s+1}\|_2 < \theta \lambda_1, \end{cases}_+$$

where $\tilde{\boldsymbol{\eta}}_{ij}^{s+1} = \boldsymbol{\gamma}_i^{s+1} - \boldsymbol{\gamma}_j^{s+1} - \boldsymbol{\xi}_{ij}^s / \lambda_2$ and $\theta > 1/\lambda_2$ for the global convexity of the second minimization function in (A.2) [Wang et al., 2018, Forthcoming].

If the minimization function of $\boldsymbol{\eta}^{s+1}$ is non-convex, assigning appropriate initial values becomes essential. A proper start will lead to an ideal solution. Inspired by Zhu and Qu [2018], the clustering method can be initialized as shown in the following algorithm.

Algorithm 1: F-clust Algorithm

Initialization:

$\boldsymbol{\xi}^0 = \mathbf{0}$, $\boldsymbol{\gamma}^0 = (W'W)^{-1} (W'\mathbf{y})$, $\boldsymbol{\eta}^0 = \arg \min_{\boldsymbol{\eta}} Q_{\lambda_2}(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\xi})$, where λ_2 and $\theta > 1/\lambda_2$ are fixed.

for $s = 0, 1, 2, \dots$ **do**

$\boldsymbol{\gamma}^{s+1} = (W'W + \lambda_2 D'D)^{-1} \{W'\mathbf{y} + \lambda_2 D'(\boldsymbol{\eta}^s + \boldsymbol{\xi}^s / \lambda_2)\}.$
 $\boldsymbol{\eta}^{s+1} = \arg \min_{\boldsymbol{\eta}} Q_{\lambda_2}(\boldsymbol{\gamma}^{s+1}, \boldsymbol{\eta}, \boldsymbol{\xi}^s),$
 $\boldsymbol{\xi}_{ij}^{s+1} = \boldsymbol{\xi}_{ij}^s + \lambda_2 (\boldsymbol{\eta}_{ij}^{s+1} - \boldsymbol{\gamma}_i^{s+1} + \boldsymbol{\gamma}_j^{s+1}),$ for all $1 \leq i < j \leq n.$
if *the stopping criteria are true* **then**
 | Break
end

end

The estimated number of groups, \hat{G} , can be obtained by $\boldsymbol{\eta}$. If $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\gamma}}_j$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ are expected to be in the same cluster. However, as a penalty $\boldsymbol{\eta}_{ij}$ has been imposed in the clustering algorithm, the equality of two estimated parameters is not achievable. As a result, the MCP penalty is utilized on $\hat{\boldsymbol{\eta}}_{ij}$. Two parameters $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ are clustered in the same group if $\hat{\boldsymbol{\eta}}_{ij} = \mathbf{0}$.

In Algorithm 1, the stopping criteria are defined as the following. Let $\boldsymbol{\kappa}_{ij}^{s+1} = \boldsymbol{\gamma}_i^{s+1} - \boldsymbol{\gamma}_j^{s+1} - \boldsymbol{\eta}_{ij}^{s+1}$, $\boldsymbol{\kappa} = (\boldsymbol{\kappa}'_{12}, \dots, \boldsymbol{\kappa}'_{n-1,n})'$ and $\boldsymbol{\tau}_k^{s+1} = -\lambda_2 \{\sum_{i=k} (\boldsymbol{\eta}_{ij}^{s+1} - \boldsymbol{\eta}_{ij}^s) - \sum_{j=k} (\boldsymbol{\eta}_{ij}^{s+1} - \boldsymbol{\eta}_{ij}^s)\}$, $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_n)'$. At any step s^* , if for some small values ϵ^κ and ϵ^τ , $\|\boldsymbol{\kappa}^{s^*}\|_2 \leq \epsilon^\kappa$ and $\|\boldsymbol{\tau}^{s^*}\|_2 \leq \epsilon^\tau$, the algorithm

stops. Following Zhu and Qu [2018], define ϵ^κ and ϵ^τ as

$$\epsilon^\kappa = \sqrt{np}\epsilon^{abs} + \epsilon^{rel}\|D'\boldsymbol{\xi}^{s*}\|_2, \quad \epsilon^\tau = \sqrt{|\mathcal{I}|p}\epsilon^{abs} + \epsilon^{rel}\max\{\|D\boldsymbol{\eta}^{s*}\|_2, \|\boldsymbol{\eta}^{s*}\|_2\},$$

where $\mathcal{I} = \{(i, j) : 1 \leq i < j \leq n\}$, $|\mathcal{I}|$ indicates the cardinality of \mathcal{I} . Here, ϵ^{abs} and ϵ^{rel} are predetermined small values.

Proposition 1. *The above clustering algorithm ensures convergence, that is, $\|\boldsymbol{\kappa}^{s+1}\|_2^2 \rightarrow 0$ and $\|\boldsymbol{\tau}^{s+1}\|_2^2 \rightarrow 0$, as $s \rightarrow \infty$.*

Proof of Proposition 1. $\|\boldsymbol{\kappa}^{s+1}\|_2^2 \xrightarrow{s \rightarrow \infty} 0$ can be shown similarly to the proof of Proposition 1 in Ma and Huang [2017]. The proof of $\|\boldsymbol{\tau}^{s+1}\|_2^2 \xrightarrow{s \rightarrow \infty} 0$ can be done by ignoring the penalty term in the objective function in the proof of Theorem 3.1 in Zhu and Qu [2018]. \square

Proposition 1 demonstrates that the clustering algorithm is convergent as the number of iteration, s , approaches infinity. The stopping criteria can be satisfied at some step eventually.

A.3 Comparable Clustering Methods 1: B&R-clust

Recall that in (6), the MIDAS regression model without Fourier transformation of each subject is

$$\mathbf{y}_i = Z_i\boldsymbol{\alpha}_i + X_i\boldsymbol{\beta}_i^* + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n.$$

For more than one subject, the penal MIDAS model can be written as

$$\mathbf{y}_i = (Z_i, X_i) \begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i^* \end{pmatrix} = \widetilde{W}_i\boldsymbol{\gamma}_i^*, \quad \text{or} \quad \mathbf{y} = \widetilde{W}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon},$$

where $\widetilde{W}_i = (Z_i, X_i)$ is the raw observations, $\boldsymbol{\gamma}_i^* = (\boldsymbol{\alpha}_i', \boldsymbol{\beta}_i^{*'})'$, $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_n^*)'$.

Refer to the main idea of Breitung and Roling [2015], the cubic smoothing spline penalty is considered. The penalized objective function will be given as

$$Q(\boldsymbol{\gamma}^*) = \frac{1}{2}\|\mathbf{y} - W\boldsymbol{\gamma}^*\|_2^2 + \frac{1}{2}\theta_{\boldsymbol{\gamma}^*}\boldsymbol{\gamma}^{*'}\mathbf{A}\boldsymbol{\gamma}^*,$$

where θ_{γ^*} is the pre-determined smoothing parameter, $\mathbf{A} = I_n \otimes (A'A)$. A is defined as

$$A_{(m-2) \times m} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}.$$

According to Zhu and Qu [2018], our goal is to solve the constrained optimization function

$$Q_{\lambda_2}(\gamma^*, \boldsymbol{\eta}, \boldsymbol{\xi}) = Q(\gamma^*) + \sum_{i < j} \rho(\boldsymbol{\eta}_{ij}, \lambda_1) + \frac{\lambda_2}{2} \sum_{i < j} \|\gamma_i^* - \gamma_j^* - \boldsymbol{\eta}_{ij}\|_2^2 + \sum_{i < j} \boldsymbol{\xi}'_{ij}(\gamma_i^* - \gamma_j^* - \boldsymbol{\eta}_{ij}). \quad (\text{A.3})$$

The clustering algorithm of (A.3) is similar to Algorithm 1.

Algorithm 2: B&R-clust Algorithm

Initialization:

$\boldsymbol{\xi}^0 = \mathbf{0}$, $\boldsymbol{\gamma}^0 = (\widetilde{W}'\widetilde{W} + \theta_{\gamma^*}\mathbf{A})^{-1}(\widetilde{W}'\mathbf{y})$, $\boldsymbol{\eta}^0 = \arg \min_{\boldsymbol{\eta}} Q_{\lambda_2}(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\xi})$, where λ_2 and $\theta > 1/\lambda_2$ are fixed.

for $s = 0, 1, 2, \dots$ **do**

$\boldsymbol{\gamma}^{s+1} = (W'W + \lambda_2 D'D + \theta_{\gamma^*}\mathbf{A})^{-1} \{W'\mathbf{y} + \lambda_2 D'(\boldsymbol{\eta}^s + \boldsymbol{\xi}^s/\lambda_2)\}.$

$\boldsymbol{\eta}^{s+1} = \arg \min_{\boldsymbol{\eta}} Q_{\lambda_2}(\boldsymbol{\gamma}^{s+1}, \boldsymbol{\eta}, \boldsymbol{\xi}^s),$

$\boldsymbol{\xi}^{s+1} = \boldsymbol{\xi}^s + \lambda_2(\boldsymbol{\eta}_{ij}^{s+1} - \boldsymbol{\gamma}_i^{s+1} + \boldsymbol{\gamma}_j^{s+1}),$ for all $1 \leq i < j \leq n.$

if the stopping criteria are true **then**

| Break

end

end

Note that Algorithm 2 follows the same main idea of Zhu and Qu [2018]. However, in Zhu and Qu [2018], the model introduces B-splines to approximate observations, while Algorithm 2 simply uses all high-frequency regressors. Moreover, an additional tuning parameter, θ_{γ^*} , is required to be predetermined. Refer to Breitung and Roling [2015], Zhu and Qu [2018], the selection of θ_{γ^*} is

based on the minimum of AIC given by

$$AIC_{\theta_{\gamma^*}} = \sum_{i=1}^n \left\{ \log \left(\frac{\|\mathbf{y}_i - W_i \hat{\boldsymbol{\gamma}}_i\|_2^2}{T} \right) + \frac{2 \cdot df_i}{T} \right\},$$

where $df_i = \text{tr}\{W_i(W_i'W_i + \theta_{\gamma^*}A'A)^{-1}W_i'\}$. The selection of λ_1 here, is by minimizing

$$BIC_{\lambda_1} = \log \left(\frac{\|\mathbf{y} - W \hat{\boldsymbol{\gamma}}\|_2^2}{n} \right) + \frac{\log(n) \left\{ \hat{G}(\frac{1}{n} \sum_{i=1}^n df_i) \right\}}{n}.$$

With fixed λ_1 , $AIC_{\theta_{\gamma^*}}$ can be obtained for different values of θ_{γ^*} . Then, fix θ_{γ^*} with minimum BIC, BIC_{λ_1} can be calculated based on the determined θ_{γ^*} .

A.4 Comparable Clustering Methods 2: Fourier-SSP

Su et al. [2016] introduced C-Lasso for clusters to identify relatively large differences between parameters and group averages rather than the traditional Lasso for each subject to select relevant covariates. The penalized profile likelihood (PPL) function mentioned in Su et al. [2016] is

$$Q(\boldsymbol{\gamma}^*) = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \phi(w_{it}; \boldsymbol{\gamma}_i^*, \hat{\mu}_i(\boldsymbol{\gamma}_i^*)).$$

By introducing the group Lasso penalty, the PPL criterion function becomes

$$Q_{G, \lambda_{PPL}} = Q(\boldsymbol{\gamma}^*) + \frac{\lambda_{PPL}}{N} \sum_{i=1}^N \prod_{g=1}^{G_0} \|\boldsymbol{\beta}_i - \boldsymbol{\alpha}_g\|_2,$$

where λ_{PPL} is a tuning parameter. The C-Lasso estimation $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$, respectively. Without any prior knowledge of the true clusters, PPL C-Lasso estimation requires a predetermination of a reasonable maximum value, G_0 , of groups. An appropriate choice of (λ_{PPL}, G_0) can be found by minimizing IC based on all possible values of clusters less than G_0 as long as predetermined values of λ_{PPL} . To start the algorithm, Su et al. [2016] suggested a natural initial value as $\hat{\boldsymbol{\alpha}}_g^{(0)} = \mathbf{0}$ for all $g = 1, \dots, G_0$ and $\hat{\boldsymbol{\gamma}}^{*(0)}$ as the quasi-maximum likelihood estimation (QMLE) of $\boldsymbol{\gamma}_i^*$ in each subjects. More details can be found in Su et al. [2016].

Algorithm 3: SSP – PPL Algorithm Given G_0 and λ_{PPL}

Initialization: $\hat{\alpha}^{(0)} = (\hat{\alpha}_1^{(0)}, \dots, \hat{\alpha}_{G_0}^{(0)})'$, $\hat{\gamma}^{*(0)} = (\hat{\gamma}_1^{*(0)}, \dots, \hat{\gamma}_n^{*(0)})'$ s.t.

$\sum_{i=1}^n \|\hat{\gamma}_i^{*(0)} - \hat{\alpha}_g^{(0)}\| \neq 0$ for all $g = 2, \dots, G_0$.

for $s = 1, 2, \dots$ **do**

for $g = 1, 2, \dots, G_0$ **do**

 Obtain the estimator $(\hat{\gamma}^{*(s,G)}, \hat{\alpha}_g^{(s)})$ of (γ^*, α_g) by minimizing the following
 objective function $Q_{G,\lambda_{PPL}}^{(s,g)}(\gamma^*, \alpha_g)$.

if $g = 1$ **then**

$$Q_{G,\lambda_{PPL}}^{(s,g)}(\gamma^*, \alpha_g) = Q(\gamma^*) + \frac{\lambda_{PPL}}{N} \sum_{i=1}^N \|\gamma_i^* - \alpha_g\| \prod_{k=2}^G \|\gamma_i^{*(s-1,k)} - \alpha_k^{(s-1)}\| ;$$

else if $g \neq G$ **then**

$$Q_{G,\lambda_{PPL}}^{(s,g)}(\gamma^*, \alpha_g) = Q(\gamma^*) + \frac{\lambda_{PPL}}{N} \sum_{i=1}^N \|\gamma_i^* - \alpha_g\| \prod_{j=1}^{g-1} \|\hat{\gamma}_i^{*(s,j)} - \alpha_j^{(s)}\| \prod_{k=g+1}^G \|\gamma_i^{*(s-1,k)} - \alpha_k^{(s-1)}\| ;$$

else

$$Q_{G,\lambda_{PPL}}^{(s,g)}(\gamma^*, \alpha_g) = Q(\gamma^*) + \frac{\lambda_{PPL}}{N} \sum_{i=1}^N \|\gamma_i^* - \alpha_g\| \prod_{k=1}^{G-1} \|\hat{\gamma}_i^{*(s,k)} - \alpha_k^{(s)}\| ;$$

end

end

if the stopping criteria are true then

 | Break

end

end

Su et al. [2016] provided a stopping criteria for this algorithm:

$$\hat{Q}_{G,\lambda_{PPL}}^{(s-1)} - \hat{Q}_{G,\lambda_{PPL}}^{(s)} \leq \epsilon_{tl} \text{ and } \frac{\sum_{g=1}^G \left\| \hat{\alpha}_g^{(s)} - \hat{\alpha}_g^{(s-1)} \right\|^2}{\sum_{g=1}^G \left\| \hat{\alpha}_g^{(s-1)} \right\|^2 + 0.0001} \leq \epsilon_{tl},$$

where ϵ_{tl} is a predetermined small value indicating the tolerance level.

A.5 Algorithm for dropping a part of regressors in clustering

In the framework shown in Section 3, the procedure concentrates on clustering weights of Z_i and \tilde{X}_i at the same time. To cluster part of weights, a selection matrix C_s is introduced*. The modified penalized objective function:

$$Q(\gamma) = \frac{1}{2} \|\mathbf{y} - W\gamma\|_2^2 + \sum_{1 \leq i < j \leq n} \rho(C_s \gamma_i - C_s \gamma_j, \lambda_1),$$

where C_s is a matrix of 1s and 0s that picks up the coefficient of interest. For example, if one is interested in clustering Fourier transformed weights only, the matrix C_s is the same as \mathbf{D} in (4). The group-specified parameter is $\tilde{\eta}_{ij} = C_s \gamma_i - C_s \gamma_j$, and the constrained optimization problem is

$$\begin{aligned} Q_{\lambda_2}(\gamma, \eta, \xi) &= \frac{1}{2} \|\mathbf{y} - W\gamma\|_2^2 + \sum_{i < j} \rho(\eta_{ij}, \lambda_1) \\ &\quad + \frac{\lambda_2}{2} \sum_{i < j} \|C_s \gamma_i - C_s \gamma_j - \eta_{ij}\|_2^2 + \sum_{i < j} \xi'_{ij} (C_s \gamma_i - C_s \gamma_j - \eta_{ij}). \end{aligned}$$

Equivalently,

$$Q_{\lambda_2}^\gamma(\gamma, \eta, \xi) = \frac{1}{2} \|\mathbf{y} - W\gamma\|_2^2 + \frac{\lambda_2}{2} \|\tilde{D}\gamma - (\eta + \xi/\lambda_2)\|_2^2,$$

where $D_{ij} = (\mathbf{e}_i - \mathbf{e}_j)' \otimes I_p$ and $\tilde{D} = (D'_{12}C'_s, D'_{13}C'_s, \dots, D'_{n-1,n}C'_s)'$. The corresponding algorithm can be summarized as Algorithm 4.

B Proofs

B.1 Lemmas

Assumptions on regressors (in our setting, W) made in Ma and Huang [2016] and related papers can be somewhat too strong for our panel setting. For example, (C3) in Ma and Huang [2016] assumes that each column of W , taking only the rows that correspond to the k -th group, should be nonrandom, and the sum of squares of all its elements is assumed to be equal to the size of k -th

*Although we do not provide a formal proof for this argument, the validity of this algorithm can be proved in a similar manner, following Ma and Huang [2016]'s argument. To keep the paper concise, we do not present the detail in this paper.

Algorithm 4: F-clust excluding some coefficients from clustering

Initialization:

$\xi^0 = \mathbf{0}$, $\gamma^0 = (W'W)^{-1}(W'\mathbf{y})$, $\eta^0 = \arg \min_{\eta} Q_{\lambda_2}(\gamma, \eta, \xi)$, where λ_2 and $\theta > 1/\lambda_2$ are fixed.

for $s = 0, 1, 2, \dots$ **do**

$$\gamma^{s+1} = \left(W'W + \lambda_2 \tilde{D}'\tilde{D}\right)^{-1} \left\{W'\mathbf{y} + \lambda_2 \tilde{D}'(\eta^s + \xi^s/\lambda_2)\right\}.$$

$$\eta^{s+1} = \arg \min_{\eta} Q_{\lambda_2}(\gamma^{s+1}, \eta, \xi^s),$$

$$\xi_{ij}^{s+1} = \xi_{ij}^s + \lambda_2(\eta_{ij}^{s+1} - C\gamma_i^{s+1} + C\gamma_j^{s+1}), \text{ for all } 1 \leq i < j \leq n.$$

if *the stopping criteria are true* **then**

 | Break

end

end

group, i.e., $|\mathcal{G}_k|$. This type of assumption could be realistic for data involved with an experimental design, but not suitable for panel data setting, where columns of W generally consists of random variables. In this proof, we circumvent this issue by using the following lemmas.

Lemma 1. Suppose a random vector $\varepsilon = (\varepsilon_{1,1}, \varepsilon_{1,2}, \dots, \varepsilon_{n,T})'$ of length nT as in (8) satisfies Assumption 4. Let $A \in \mathbf{R}^{a \times nT}$ be a nonrandom matrix with a positive integer a . Let $\Sigma = A'A$. For any $\zeta > 0$,

$$P \left[\|A\varepsilon\|_2^2 > 2\tilde{c}\{\text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)}\zeta + 2\|\Sigma\|_2\zeta\} \right] \leq e^{-\zeta}.$$

Proof of Lemma 1. When $a = nT$, this lemma is a special case of Theorem 2.1 in Hsu et al. [2012]. This can be easily seen by recognizing their μ , σ^2 , and α are 0, $2\tilde{c}$, and $(\nu_{1,1}, \nu_{1,2}, \dots, \nu_{n,T})'$, respectively.

If $a < nT$, a similar argument can still be used. Consider a singular value decomposition of $A = USV'$, where U and V are $a \times a$ and $nT \times nT$ orthogonal matrices, respectively. Let $\rho = (\rho_1, \dots, \rho_a)'$ denote the nonzero eigenvalues of $A'A$ and AA' . S is an $a \times nT$ matrix, where its diagonal elements are equal to $\sqrt{\rho_i}$ for $i = 1, \dots, a$ and all other entries are zero. Let z be a vector of a independent standard Gaussian random variables. Since U is orthogonal, $y = U'z$ is also an $a \times 1$ vector of a independent standard Gaussian random variables. Let $y = (y_1, \dots, y_a)'$. Applying Lemma 2.4 of Hsu et al. [2012] on $\|A'z\|^2 = Z'AA'z = z'USV'VS'U'z = ySS'y' = \sum_{i=1}^a \rho_i y_i^2$, then

$$E \left\{ \exp(\gamma \|A'z\|^2) \right\} \leq \exp \left(\|\rho\|_1 \gamma + \frac{\|\rho\|_2^2 \gamma^2}{1 - 2\|\rho\|_\infty \gamma} \right) \quad (\text{B.1})$$

for any $0 \leq \gamma < 1/(2\|\rho\|_\infty)$. For any $\lambda \in \mathbf{R}$ and $\delta \geq 0$, using similar arguments as in (2.3) and (2.4) of Hsu et al. [2012], Assumption 4, and (B.1),

$$P(\|A\epsilon\|^2 > \delta) \leq \exp\left(-\frac{\lambda^2\delta}{2}\right) \exp\left\{\|\rho\|_1(\lambda^2\tilde{c}) + \frac{\|\rho\|_2^2(\lambda^2\tilde{c})^2}{1-2\|\rho\|_\infty(\lambda^2\tilde{c})}\right\}.$$

Let $\delta = 2\tilde{c}(\|\rho\|_1 + \tau)$, $\lambda^2 = \frac{1}{\tilde{c}} \frac{1}{2\|\rho\|_\infty} \left(1 - \sqrt{\frac{\|\rho\|_2^2}{\|\rho\|_2^2 + 2\|\rho\|_\infty\tau}}\right)$, and $\tau = 2\sqrt{\|\rho\|_2^2\zeta} + 2\|\rho\|_\infty\zeta$. The desired proof is concluded by using similar arguments as Hsu et al. [2012] and observing $\|\rho\|_1 = \sum_{i=1}^a \rho_i = \text{tr}(\Sigma)$, $\|\rho\|_2^2 = \sum_{i=1}^a \rho_i^2 = \text{tr}(\Sigma^2)$, and $\|\rho\|_\infty = \max_i \rho_i = \|\Sigma\|_2$.

A similar proof works for $a > nT$. In this case, without loss of generality, the only nonzero element in S are the first nT diagonal elements of S . Let $\sqrt{\rho_i}$, $i = 1, \dots, nT$, be the nonzero diagonal elements of S . Then $\|A'z\|^2 = \sum_{i=1}^{nT} \rho_i y_i^2$, where y_i are independent standard Gaussian random variables. The rest of the proof is the same. \square

Lemma 2. *Suppose conditions of Lemma 1 hold. For any $nT \times np$ matrix W satisfying Assumption 2,*

$$\begin{aligned} P\left[\|W'\epsilon\|_2^2 > 2\tilde{c}(np + 2\sqrt{np\zeta^*} + 2\zeta^*)\|W'W\|_2 \mid W\right] &\leq e^{-\zeta^*} \quad \text{and} \\ P\left[\|\Gamma'W'\epsilon\|_2^2 > 2\tilde{c}(Gp + 2\sqrt{Gp\zeta} + 2\zeta)\|\Gamma'W'W\Gamma\|_2 \mid W\right] &\leq e^{-\zeta} \end{aligned}$$

hold for any $\zeta^* > 0$ and $\zeta > 0$.

Proof of Lemma 2. Fix a $nT \times nP$ matrix W that satisfies Assumption 2. Using Lemma 1, for any $\zeta^* > 0$,

$$\begin{aligned} P\left[\|W'\epsilon\|_2^2 > 2\tilde{c}(\text{tr}(WW') + 2\sqrt{\text{tr}((WW')^2)\zeta^*} + 2\|WW'\|_2\zeta^*) \mid W\right] &\leq e^{-\zeta^*}, \quad \text{and} \\ P\left[\|\Gamma'W'\epsilon\|_2^2 > 2\tilde{c}(\text{tr}(\Gamma WW'\Gamma') + 2\sqrt{\text{tr}((\Gamma WW'\Gamma')^2)\zeta} + 2\|\Gamma WW'\Gamma'\|_2\zeta) \mid W\right] &\leq e^{-\zeta}. \end{aligned}$$

Since $\|WW'\|_2$ is the maximum eigenvalue of WW' , using the fact that WW' is symmetric and positive definite with rank np , it can be easily seen that $\lambda_{\max}(WW') = \lambda_{\max}(W'W)$,

$$\|WW'\|_2 = \|W'W\|_2 = \|\text{diag}(W'_1W_1, \dots, W'_nW_n)\|_2 \leq \max_i \|W'_iW_i\|_2, \quad \text{and}$$

$$\text{tr}(WW') = \text{tr}(W'W) \leq np\|W'W\|_2, \quad \text{tr}((WW')^2) = \text{tr}((W'W)^2) \leq np\|W'W\|_2^2.$$

Therefore

$$\text{tr}(WW') + 2\sqrt{\text{tr}[(WW')^2]\zeta^*} + 2\|WW'\|_2\zeta^* \leq (np + 2\sqrt{np\zeta^*} + 2\zeta^*)\|W'W\|_2.$$

Similarly, $\|W\Gamma\Gamma'W'\|_2 = \|\Gamma'W'W\Gamma\|_2$,

$$\text{tr}(W\Gamma\Gamma'W') = \text{tr}(\Gamma'W'W\Gamma) \leq Gp\lambda_{\max}(\Gamma'W'W\Gamma) = Gp\|\Gamma'W'W\Gamma\|_2, \quad \text{and}$$

$$\text{tr}\{(W\Gamma\Gamma'W')^2\} = \text{tr}\{(\Gamma'W'W\Gamma)^2\} \leq Gp\{\lambda_{\max}(\Gamma'W'W\Gamma)\}^2 = Gp\|\Gamma'W'W\Gamma\|_2^2.$$

Therefore for any $\zeta > 0$,

$$\text{tr}(\Gamma'W'W\Gamma) + 2\sqrt{\text{tr}\{(\Gamma'W'W\Gamma)^2\}}\sqrt{\zeta} + 2\|\Gamma'W'W\Gamma\|_2\zeta \leq (Gp + 2\sqrt{Gp\zeta} + 2\zeta)\|\Gamma'W'W\Gamma\|_2.$$

As a result, given any matrix W , the inequalities in the statement have been validated. \square

Lemma 3. Suppose Assumptions 2 and 4 hold for W and ε . Define

$$S_\zeta = 2\tilde{c}(Gp + 2\sqrt{Gp\zeta} + 2\zeta)g_{\max}m\tilde{M}\sqrt{GpT}B_{q,m},$$

$$S_{\zeta^*} = 2\tilde{c}(np + 2\sqrt{np\zeta^*} + 2\zeta^*)m\tilde{M}\sqrt{T}B_{q,m}\sqrt{p},$$

where $B_{q,m} = (q^{1/2} + m^{1/2}(L + 1 + 2K))$, $p = q + L + 1 + 2K$, $\tilde{M} = \max(M_1, M_2, M_3, M_4)$ and \tilde{c} given in Assumption 2 and 4, then $P[\|W'\varepsilon\|_2^2 > S_{\zeta^*}] \leq e^{-\iota^*}$ and $P[\|\Gamma'W'\varepsilon\|_2^2 > S_\zeta] \leq e^{-\iota}$ where $\iota = \min(\zeta, -\log(\epsilon)) - \log(2)$ and $\iota^* = \min(\zeta^*, -\log(\epsilon)) - \log(2)$ for any ζ and ζ^* in Lemma 2.

Proof of Lemma 3. Using the law of iterated expectations,

$$\begin{aligned} E[P(\|W'\varepsilon\|_2^2 > S_{\zeta^*} \mid W)] &= P[\|W'\varepsilon\|_2 > S_{\zeta^*}] \\ &= E\left[I_{\{\|W'\varepsilon\|_2^2 > S_{\zeta^*}\}} \mid \|WW'\|_2 \leq M^*\right] P(\|WW'\|_2 \leq M^*) \\ &\quad + E\left[I_{\{\|W'\varepsilon\|_2^2 > S_{\zeta^*}\}} \mid \|WW'\|_2 > M^*\right] P(\|WW'\|_2 > M^*) \\ &= P[\|W'\varepsilon\|_2^2 > S_{\zeta^*} \mid \|WW'\|_2 \leq M^*] P(\|WW'\|_2 \leq M^*) \\ &\quad + P[\|W'\varepsilon\|_2^2 > S_{\zeta^*} \mid \|WW'\|_2 > M^*] P(\|WW'\|_2 > M^*). \end{aligned}$$

Since $\|M\|_\infty \leq m$ and $\|M'\|_\infty \leq L+1+2K$ as all elements of M in (3) smaller than 1 in magnitude,

$$\begin{aligned}
\left\| \sum_{i \in \mathcal{G}_g} Z'_i Z_i \right\|_\infty &= \sum_{i \in \mathcal{G}_g} \|Z'_i Z_i\|_\infty \leq M_1 |\mathcal{G}_g| \sqrt{qT}, \\
\left\| \sum_{i \in \mathcal{G}_g} Z'_i \tilde{X}_i \right\|_\infty &\leq \sum_{i \in \mathcal{G}_g} \|Z'_i X_i\|_\infty \|M'\|_\infty \leq M_3 |\mathcal{G}_g| \sqrt{mT} (L+1+2K), \\
\left\| \sum_{i \in \mathcal{G}_g} \tilde{X}'_i Z_i \right\|_\infty &\leq \|M\|_\infty \sum_{i \in \mathcal{G}_g} \|Z'_i X_i\|_\infty \leq M_4 |\mathcal{G}_g| m \sqrt{qT}, \quad \text{and} \\
\left\| \sum_{i \in \mathcal{G}_g} \tilde{X}'_i \tilde{X}_i \right\|_\infty &\leq \|M\|_\infty \sum_{i \in \mathcal{G}_g} \|X'_i X_i\|_\infty \|M'\|_\infty \leq M_2 |\mathcal{G}_g| m \sqrt{mT} (L+1+2K)
\end{aligned}$$

hold with probability at least $1-\epsilon$ for any $\epsilon > 0$ defined in Assumption 2. Therefore, with probability at most $1-\epsilon$,

$$\begin{aligned}
\|WW'\|_2 &= \|W'W\|_2 = \|\text{diag}(W'_1 W_1, \dots, W'_n W_n)\|_2 \leq \sup_i \|W'_i W_i\|_2 \\
&\leq \sqrt{p} \sup_i \|W'_i W_i\|_\infty = \sqrt{p} \sup_i \left\| \begin{array}{cc} Z'_i Z_i & Z'_i \tilde{X}_i \\ \tilde{X}'_i Z_i & \tilde{X}'_i \tilde{X}_i \end{array} \right\|_\infty \\
&\leq \tilde{M} m \sqrt{T} B_{q,m} \sqrt{p}.
\end{aligned}$$

Since $\text{tr}(WW') = \text{tr}(W'W) \leq np \|W'W\|_2$ and $\text{tr}((WW')^2) = \text{tr}((W'W)^2) \leq np \|W'W\|_2^2$,

$$\text{tr}(WW') + 2\sqrt{\text{tr}[(WW')^2] \zeta^*} + 2\|WW'\|_2 \zeta^* \leq (np + 2\sqrt{np \zeta^*} + 2\zeta^*) \|WW'\|_2.$$

Since $\|WW'\|_2$ is bounded in probability, for any $\epsilon > 0$, there exists some $M^* = \tilde{M} m \sqrt{T} B_{q,m} \sqrt{p}$ such that $P[\|WW'\|_2 > M^*] \leq \epsilon$. Therefore

$$P[\|W'\epsilon\|_2^2 > S_{\zeta^*} \mid W, \|WW'\|_2 \leq M^*] \leq e^{-\zeta^*}, \quad 1-\epsilon < P(\|WW'\|_2 \leq M^*) \leq 1,$$

$$P[\|W'\epsilon\|_2^2 > S_{\zeta^*} \mid W, \|WW'\|_2 > M^*] \leq 1, \quad P(\|WW'\|_2 > M^*) \leq \epsilon,$$

and $P[\|W'\epsilon\|_2^2 > S_{\zeta^*}] \leq e^{-\zeta^*} + \epsilon$ where $S_{\zeta^*} = 2\tilde{c}(np + 2\sqrt{np \zeta^*} + 2\zeta^*)M^*$.

Without loss of generality, let $\tilde{\zeta}^* = \min\{\zeta^*, -\log(\epsilon)\}$ for a large constant $\zeta^* > 1$ and for small

positive constant $\epsilon \leq 1$, then $e^{-\zeta^*} + \epsilon = e^{-\zeta^*} + e^{\log(\epsilon)} = e^{-\tilde{\zeta}^*} (1 + e^{-|\zeta^* + \log(\epsilon)|}) \leq 2e^{-\tilde{\zeta}^*} = e^{\log(2) - \tilde{\zeta}^*}$.

Take $\iota^* = \tilde{\zeta}^* - \log(2)$, then $P[\|W'\epsilon\|_2^2 > S_{\zeta^*}] \leq e^{-\iota^*}$. For large enough $\tilde{\zeta}^*$, $\log(2)$ is negligible.

Similarly, S_ζ in $P[\|\Gamma'W'\epsilon\|_2^2 > S_\zeta] \leq e^{-\iota}$ can be found as the following.

A straightforward calculation derives that

$$\Gamma'W'W\Gamma = \text{diag} \left(\sum_{i \in \mathcal{G}_1} W_i'W_i, \dots, \sum_{i \in \mathcal{G}_G} W_i'W_i \right).$$

It follows that, with probability $1 - \epsilon$,

$$\begin{aligned} \|\Gamma'W'W\Gamma\|_\infty &= \max_{1 \leq g \leq G} \left\| \sum_{i \in \mathcal{G}_g} W_i'W_i \right\|_\infty \leq \max_{1 \leq g \leq G} \sum_{i \in \mathcal{G}_g} \|W_i'W_i\|_\infty \leq g_{\max} \sup_{1 \leq i \leq n} \|W_i'W_i\|_\infty \\ &\leq g_{\max} m \tilde{M} \sqrt{T} B_{q,m}, \end{aligned}$$

and therefore,

$$\|\Gamma'W'W\Gamma\|_2 \leq \sqrt{Gp} \|\Gamma'W'W\Gamma\|_\infty \leq g_{\max} m \tilde{M} \sqrt{GpT} B_{q,m}.$$

For any $\epsilon > 0$, there exists some $M = g_{\max} m \tilde{M} \sqrt{T} B_{q,m}$, such that $P[\|W\Gamma\Gamma'W'\|_2 > M] \leq \epsilon$, then

$$P[\|\Gamma'W'\epsilon\|_2^2 > S_\zeta \mid W, \|W\Gamma\Gamma'W'\|_2^2 \leq M] \leq e^{-\zeta}, \quad 1 - \epsilon < P(\|W\Gamma\Gamma'W'\|_2 \leq M) \leq 1,$$

$$P[\|W'\epsilon\|_2 > S_\zeta \mid W, \|W\Gamma\Gamma'W'\|_2 > M] \leq 1, \quad P(\|W\Gamma\Gamma'W'\|_2 > M) \leq \epsilon.$$

Therefore, $P[\|\Gamma'W'\epsilon\|_2^2 > S_\zeta] \leq e^{-\zeta} + \epsilon$ where $S_\zeta = 2\tilde{c}(Gp + 2\sqrt{Gp\zeta} + 2\zeta)M$. Similarly, take $\iota = \min\{\zeta, -\log(\epsilon)\} - \log(2)$, then $P[\|\Gamma'W'\epsilon\|_2^2 > S_\iota] \leq e^{-\iota}$. \square

B.2 Convergence of the Oracle Estimator

Theorem 1 and Corollary 1 are proved in this section.

Proof of Theorem 1. The definition of Γ and $\mathbf{y} = W\boldsymbol{\gamma}^{or} + \boldsymbol{\varepsilon}$ lead to

$$\begin{aligned}\hat{\boldsymbol{\gamma}}^{or} - \boldsymbol{\gamma}^0 &= \Gamma(\Gamma'W'W\Gamma)^{-1}\Gamma'W'\boldsymbol{\varepsilon} \\ &= \Gamma \left\{ \text{diag} \left(\sum_{i \in \mathcal{G}_1} W_i'W_i, \dots, \sum_{i \in \mathcal{G}_G} W_i'W_i \right) \right\}^{-1} \begin{pmatrix} \sum_{i \in \mathcal{G}_1} W_i'\boldsymbol{\varepsilon}_i \\ \vdots \\ \sum_{i \in \mathcal{G}_G} W_i'\boldsymbol{\varepsilon}_i \end{pmatrix},\end{aligned}$$

where for any $g \in \{1, \dots, G\}$,

$$\sum_{i \in \mathcal{G}_g} W_i'W_i = \begin{pmatrix} \sum_{i \in \mathcal{G}_g} Z_i'Z_i & (\sum_{i \in \mathcal{G}_g} Z_i'X_i)\mathbf{M}' \\ \mathbf{M}(\sum_{i \in \mathcal{G}_g} X_i'Z_i) & \mathbf{M}(\sum_{i \in \mathcal{G}_g} X_i'X_i)\mathbf{M}' \end{pmatrix} \quad \text{and} \quad \sum_{i \in \mathcal{G}_g} W_i'\boldsymbol{\varepsilon}_i = \begin{pmatrix} \sum_{i \in \mathcal{G}_g} Z_i'\boldsymbol{\varepsilon}_i \\ \mathbf{M}(\sum_{i \in \mathcal{G}_g} X_i'\boldsymbol{\varepsilon}_i) \end{pmatrix}.$$

Assumption 2 implies that

$$\lambda_{\min}(\Gamma'W'W\Gamma) \geq cg_{\min}T,$$

so that

$$\|(\Gamma'W'W\Gamma)^{-1}\|_{\infty} \leq \sqrt{Gp}\|(\Gamma'W'W\Gamma)^{-1}\|_2 \leq \sqrt{Gp}(cg_{\min}T)^{-1}. \quad (\text{B.2})$$

For all p -norms, $\|A \otimes B\| = \|A\|\|B\|$ holds (for example, see p. 433 of Langville and Stewart [2004]),

$$\|\Gamma\|_{\infty} \leq \|\Pi\|_{\infty}\|I_p\|_{\infty} = 1. \quad (\text{B.3})$$

Lemma 3, equations (B.2) and (B.3), and the triangle inequality imply that for any $\iota > 0$,

$$\begin{aligned}\|\hat{\boldsymbol{\gamma}}^{or} - \boldsymbol{\gamma}^0\|_{\infty} &\leq \|\Gamma\|_{\infty}\|(\Gamma'W'W\Gamma)^{-1}\|_{\infty}\|\Gamma'W'\boldsymbol{\varepsilon}\|_{\infty} \\ &\leq (Gp)^{1/2}(cg_{\min}T)^{-1}\|\Gamma'W'\boldsymbol{\varepsilon}\|_2 \leq (Gp)^{1/2}(cg_{\min}T)^{-1}S_{\zeta}^{1/2},\end{aligned}$$

with probability at least $1 - e^{-\iota}$. Therefore,

$$\phi_{n,T,G,\zeta} := \frac{\sqrt{2\bar{c}}(m\tilde{M}g_{\max})^{1/2}(Gp)^{3/4}}{c g_{\min}T^{3/4}} B_{q,m}^{1/2}(Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2}, \quad (\text{B.4})$$

where $B_{q,m}$ is defined in Lemma 3. Therefore, with probability at least $1 - e^{-\iota}$,

$$\|\hat{\gamma}^{or} - \gamma^0\|_\infty \leq \phi_{n,T,G,\zeta}.$$

This proves the first part of Theorem 1. The remaining proof is for the asymptotic normality of $\hat{\gamma}^{or}$. Let $V_i = W_i(\Pi_i \otimes I_p)$ be a $T \times Gp$ matrix, where Π_i is the i -th row of the matrix Π , $V = W\Gamma = (V_1', \dots, V_n)'$. Then, for any $c_n \in \mathbb{R}^{Gp}$ with $\|c_n\|_2 = 1$,

$$c_n'(\hat{\gamma}^{or} - \gamma^0) = \sum_{i=1}^n c_n'(V'V)^{-1}V_i'\varepsilon_i = \sum_{i=1}^n c_n'(V'V)^{-1} \sum_{t=1}^T \mathbf{v}_{it}'\varepsilon_{it}.$$

Since $\{\varepsilon_i\}$ is assumed to be an i.i.d. subgaussian distributed sequence with mean 0 and variance proxy $2\tilde{c}$, then $E(\varepsilon_i) = \mathbf{0}$. Hence,

$$E[c_n'(\hat{\gamma}^{or} - \gamma^0)] = 0.$$

Suppose that Assumption 2 and 4 hold where $\lambda_{\max}(V'V) = \lambda_{\max}(\Gamma'W'W\Gamma) \leq c^*|\mathcal{G}_g|T \leq c^*g_{\max}T$ and $Var(\varepsilon_{it}) = O(2\tilde{c})$, then

$$\sigma_\gamma^2 := Var[c_n'(\hat{\gamma}^{or} - \gamma^0)] \geq \frac{Var(\varepsilon_{it})}{c^*g_{\max}T}.$$

Moreover, for any $\epsilon > 0$, applying Cauchy-Schwarz inequality,

$$\begin{aligned} & \sum_{i=1}^n E((c_n'(V'V)^{-1}V_i'\varepsilon_i)^2 \mathbb{1}_{\{|c_n'(V'V)^{-1}V_i'\varepsilon_i| > \epsilon\sigma_\gamma\}}) \\ & \leq \sum_{i=1}^n \{E(c_n'(V'V)^{-1}V_i'\varepsilon_i)^4\}^{1/2} \{E(\mathbb{1}_{\{|c_n'(V'V)^{-1}V_i'\varepsilon_i| > \epsilon\sigma_\gamma\}}^2)\}^{1/2} \\ & = \sum_{i=1}^n \{E(c_n'(V'V)^{-1}V_i'\varepsilon_i)^4\}^{1/2} \{E(\mathbb{1}_{\{|c_n'(V'V)^{-1}V_i'\varepsilon_i| > \epsilon\sigma_\gamma\}})\}^{1/2} \\ & = \sum_{i=1}^n \{E(c_n'(V'V)^{-1}V_i'\varepsilon_i)^4\}^{1/2} \{P(|c_n'(V'V)^{-1}V_i'\varepsilon_i| > \epsilon\sigma_\gamma)\}^{1/2}. \end{aligned} \tag{B.5}$$

The first term can be derived as

$$\begin{aligned}
[E(c'_n(V'V)^{-1}V'_i\epsilon_i)^4]^{1/2} &= [E(c'_n(V'V)^{-1}V'_i\epsilon_i\epsilon'_iV'_i(V'V)^{-1}c_n)^2]^{1/2} \\
&= [\{c'_n(V'V)^{-1}V_i\}^2 E(\epsilon_i\epsilon'_i)^2 \{V'_i(V'V)^{-1}c_n\}^2]^{1/2} \\
&= c'_n(V'V)^{-1}V_i [E(\epsilon_i\epsilon'_i)^2]^{1/2} V'_i(V'V)^{-1}c_n \\
&\leq \|c'_n(V'V)^{-1}V_i\|_2^2 \|E(\epsilon_i\epsilon'_i)^2\|_2^{1/2}.
\end{aligned}$$

For any $n \times n$ matrix A , $\|A\|_2 \leq \sqrt{n}\|A\|_\infty$. Since $E(\epsilon_{it}^k) \leq (2\sigma^2)^{k/2} k\Gamma(k/2)$ for $k \geq 1$, then

$$\|E(\epsilon_i\epsilon'_i)^2\|_2 \leq \sqrt{T} \|E(\epsilon_i\epsilon'_i)^2\|_\infty = \sqrt{T} \max_{\tau=1,\dots,T} E\left(\epsilon_{i\tau} \sum_{t=1}^T \epsilon_{it} \sum_{t=1}^T \epsilon_{it}^2\right) \leq \sqrt{T}(16+T)4\tilde{c}^2.$$

According to Assumption 2, $\|V_i\|_\infty$ is bounded and let the upper bound be some constant c_2 , then $\|V_i\|_2 \leq \sqrt{Gp}c_2$. Following 2, $\|(V'V)^{-1}\|_2 \geq (cg_{\min}T)^{-1}$,

$$\begin{aligned}
\{E(c'_n(V'V)^{-1}V_i\epsilon_{it})^4\}^{1/2} &\leq \|c'_n\|_2^2 \|(V'V)^{-1}\|_2^2 \|V_i\|_2^2 T^{1/4} (16+T)^{1/2} 2\tilde{c}^2 \\
&\leq \frac{c_2^2 Gp(16+T)^{1/2} 2\tilde{c}}{c^2 g_{\min}^2 T^{3/4}}.
\end{aligned}$$

Then, by Chebyshev's inequality, the second term of (B.5) can be derived as

$$P(|c'_n(V'V)^{-1}V_i\epsilon_i| > \epsilon\sigma_\gamma) \leq \frac{E[c'_n(V'V)^{-1}V_i\epsilon_i]^2}{\epsilon^2\sigma_\gamma^2}, \quad (\text{B.6})$$

where

$$\begin{aligned}
E(c'_n(V'V)^{-1}V_i\epsilon_i)^2 &= E(c'_n(V'V)^{-1}V_i\epsilon_i\epsilon'_iV'_i(V'V)^{-1}c_n) \\
&\leq \|c_n\|_2^2 \|(V'V)^{-1}\|_2^2 \|V_i\|_2^2 \|E(\epsilon_i\epsilon'_i)\|_2 \leq \frac{c_2^2 Gp2\tilde{c}}{c^2 g_{\min}^2 T^2},
\end{aligned}$$

then, (B.6) becomes

$$P(|c'_n(V'V)^{-1}V_i\epsilon_i| > \epsilon\sigma_\gamma) \leq \frac{c_2^2 Gp2\tilde{c}}{c^2 g_{\min}^2 T^2 \epsilon^2 \sigma_\gamma^2}.$$

Therefore, the following inequality can be derived.

$$\begin{aligned}
& \sigma_\gamma^{-2} \sum_{i=1}^n E \left((c'_n (V'V)^{-1} V_i \epsilon_i)^2 \mathbb{1}_{\{|c'_n (V'V)^{-1} V_i \epsilon_i| > \epsilon \sigma_\gamma\}} \right) \\
& \leq \sigma_\gamma^{-2} \sum_{i=1}^n \frac{c_2^2 G p (16+T)^{1/2} 2\tilde{c}}{c^2 g_{\min}^2 T^{3/4}} \frac{c_2 (Gp)^{1/2} \sqrt{2\tilde{c}}}{c g_{\min} T \epsilon \sigma_\varphi} = \frac{c_2^3 p^{3/2} (2\tilde{c})^{3/2} G^{3/2} (16+T)^{1/2} n}{c^3 \epsilon g_{\min}^3 T^{7/4} \sigma_\gamma^3} \\
& \leq C \frac{(2\tilde{c})^{3/2} (n/g_{\min})^{3/2} n (16+T)^{1/2}}{\sigma_\gamma^3 g_{\min}^3 T^{7/4}} = C \frac{\tilde{c}^3 n^{5/2} (16+T)^{1/2}}{\sigma_\varphi^3 g_{\min}^{9/2} T^{7/4}} \\
& = C \frac{n^{5/2} (16+T)^{1/2} c^{*3/2} g_{\max}^{3/2} T^{3/2}}{g_{\min}^{9/2} T^{7/4}} = O \left(\frac{g_{\max}^{3/2} n^{5/2} T^{1/4}}{g_{\min}^{9/2}} \right). \tag{B.7}
\end{aligned}$$

Suppose that $\frac{g_{\min}^3}{g_{\max}} \gg n^{5/3} T^{1/6}$, then (B.7) further implies that

$$\sigma_\gamma^{-2} \sum_{i=1}^n E \left((c'_n (V'V)^{-1} V_i \epsilon_i)^2 \mathbb{1}_{\{|c'_n (V'V)^{-1} V_i \epsilon_i| > \epsilon \sigma_\gamma\}} \right) = O(1).$$

By the Lindeberg-Feller Central Limit Theorem, $c'_n (\hat{\gamma}^{or} - \gamma^0) \rightarrow N(0, \sigma_\gamma^2)$. \square

Proof of Corollary 1. In the following proof, let m and q be fixed for simplification. It further indicates that p is fixed. Let $C_{q,m} = \frac{\sqrt{2\tilde{c}}}{c} m^{1/2} p^{3/4} B_{q,m}^{1/2}$, (B.4) can be simplified as

$$\phi_{n,T,G} = C_{q,m} \frac{g_{\max}^{1/2} G^{3/4}}{g_{\min} T^{3/4}} (Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2}.$$

The rest of the proof suggests a large enough ζ for each situation that allows $\phi_{n,T,G,\zeta}$ and ι to approach infinity. We often use these somewhat trivial inequalities $g_{\max} \leq n$ and $G \leq n/g_{\min}$ in the following proofs, particularly when $n \rightarrow \infty$.

1. Consider $T \rightarrow \infty$ with n fixed. Let $\zeta \rightarrow \infty$ and $\zeta = o(T^{3/2})$. Since $G \leq n \ll \zeta$, then $(Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2} = O(2\zeta^{1/2})$. Therefore,

$$\phi_{n,T,G} = C_1 T^{-3/4} O(\zeta^{1/2}) \xrightarrow{T \rightarrow \infty} 0,$$

where $C_1 = 2C_{q,m} \frac{g_{\max}^{1/2} G^{3/4}}{g_{\min}}$, which is free of T .

2. Consider $n \rightarrow \infty$ with T fixed.

(a) Consider $G \ll \zeta \rightarrow \infty$.

i. When G is fixed, then $(Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2} = O(2\zeta^{1/2})$. For some constant $\tilde{\alpha}_0 < 1/2$, let $g_{min} = O(n^{1/2+\tilde{\alpha}_0})$, $\zeta = o(n^{2\tilde{\alpha}_0})$ and $\zeta \rightarrow \infty$, then

$$\phi_{n,T,G} \leq C_3 \frac{n^{1/2}}{g_{min}} O(\zeta^{1/2}) \xrightarrow{n \rightarrow \infty} 0,$$

where $C_3 = 2C_{q,m} \frac{G^{3/4}}{T^{3/4}}$, which is free of n .

ii. When $G \rightarrow \infty$, for some constant $\tilde{\alpha}_2 < 2/7$, let $g_{min} = O(n^{5/7+\tilde{\alpha}_2})$, $\zeta = o(n^{7\tilde{\alpha}_2/2})$ and $\zeta \rightarrow \infty$, then $(Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2} = O((p + 2\sqrt{p} + 2)^{1/2} \zeta^{1/2})$. Since $G \leq n/g_{min}$, then

$$\phi_{n,T,G} \leq C_4 \frac{n^{1/2} G^{3/4}}{g_{min}} O(\zeta^{1/2}) \leq C_4 \frac{n^{5/4}}{g_{min}^{7/4}} O(\zeta^{1/2}) \xrightarrow{n, G \rightarrow \infty} 0,$$

where $C_4 = C_{q,m} \frac{1}{T^{3/4}} (p + 2\sqrt{p} + 2)^{1/2}$, which is free of n and G .

(b) Consider $G \rightarrow \infty$. Let $g_{min} = O(n^{7/9+\tilde{\alpha}_1})$ for some $\tilde{\alpha}_1 < 2/9$, $\zeta = O(G)$ and $\zeta \rightarrow \infty$, then $Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta = O((p + 2\sqrt{p} + 2)G) = O(G)$. Therefore,

$$\phi_{n,T,G} \leq C_2 \frac{n^{1/2} G^{3/4}}{g_{min}} O(G^{1/2}) \xrightarrow{n \rightarrow \infty} 0,$$

where $C_2 = C_{q,m} \frac{1}{T^{3/4}} (p + 2\sqrt{p} + 2)^{1/2}$, which is free of n .

3. Consider $T, n \rightarrow \infty$.

(a) Consider $G \ll \zeta \rightarrow \infty$,

i. When G is fixed, then $(Gp + 2\sqrt{Gp}\sqrt{\zeta} + 2\zeta)^{1/2} = O(2\zeta^{1/2})$. Let $g_{min} = O(n^{1/2+\tilde{\alpha}_0})$ for some positive constant $\tilde{\alpha}_0 < 1/2$ and $\zeta = o(n^{2\tilde{\alpha}_0} T^{3/2})$, $\zeta \rightarrow \infty$, then

$$\phi_{n,T,G} \leq C_6 \frac{n^{1/2}}{g_{min} T^{3/4}} O(\zeta^{1/2}) \xrightarrow{n, T \rightarrow \infty} 0,$$

where $C_6 = 2C_{q,m} G^{3/4}$.

- ii. When $G \rightarrow \infty$, for some positive constant $\tilde{\alpha}_2 < 2/7$, let $g_{min} = O(n^{5/7+\tilde{\alpha}_2})$ and $G \leq n/g_{min}$, $\zeta = o(n^{7\tilde{\alpha}_2/2}T^{3/2})$ and $\zeta \rightarrow \infty$, then $(Gp + 2\sqrt{Gp\zeta} + 2\zeta)^{1/2} = O((p + 2\sqrt{p} + 2)^{1/2}\zeta^{1/2})$. Since $G \leq n/g_{min}$, then

$$\phi_{n,T,G} \leq C_7 \frac{n^{1/2}G^{3/4}}{g_{min}T^{3/4}} O(\zeta^{1/2}) \leq C_7 \frac{n^{5/4}}{g_{min}^{7/4}T^{3/4}} O(\zeta^{1/2}) \xrightarrow{n,T,G \rightarrow \infty} 0,$$

where $C_7 = C_{q,m}(p + 2\sqrt{p} + 2)^{1/2}$, which is free of n, T and G .

- (b) Consider $G \rightarrow \infty$. Let $g_{min} = O(n^{7/9+\tilde{\alpha}_1})$ for some constant $\tilde{\alpha}_1 < 2/9$, $\zeta = O(G)$ and $\zeta \rightarrow \infty$, then $Gp + 2\sqrt{Gp\zeta} + 2\zeta = O((p + 2\sqrt{p} + 2)G) = O(G)$. Since $G \leq n/g_{min}$,

$$\phi_{n,T,G} \leq C_5 \frac{n^{1/2}G^{3/4}}{g_{min}T^{3/4}} O(G^{1/2}) \leq C_5 \frac{n^{7/4}}{g_{min}^{9/4}T^{3/4}} O(1) \xrightarrow{n,T,G \rightarrow \infty} 0,$$

where $C_5 = C_{q,m}(p + 2\sqrt{p} + 2p)^{1/2}$, which is free from n, T and G .

Combining items 2 and 3 above, we can summarize the choice of ζ as follows:

Case 1. The number n is fixed. Let $\zeta = o(T^{3/2})$ as $T \rightarrow \infty$;

Case 2. The number $n \rightarrow \infty$. Whether T is fixed or $T \rightarrow \infty$,

- (a) when G is fixed, and $g_{min} = O(n^{1/2+\tilde{\alpha}_4})$ for some constant $\tilde{\alpha}_4 < 1/2$. Let $\zeta = o(n^{2\tilde{\alpha}_4}T^{3/2})$ approaching infinity;
- (b) when $G \rightarrow \infty$,
 - i. suppose $g_{min} = O(n^{7/9+\tilde{\alpha}_3})$ for some constant $\tilde{\alpha}_3 < 2/9$. Let $\zeta = O(G)$ approaching infinity;
 - ii. suppose $g_{min} = O(n^{5/7+\tilde{\alpha}_5})$ for some constant $\tilde{\alpha}_5 < 2/7$. Let $\zeta = o(n^{7\tilde{\alpha}_5/2}T^{3/2}) \gg G$ approaching infinity.

□

B.3 Convergence of the Calculated Estimator ($G \geq 2$)

Proof of Theorem 2. This can be done similarly to the proof of Theorem 4.2 in Ma and Huang [2016].

Define $\mathcal{M}_G := \{\gamma \in \mathbb{R}^{np} : \gamma_i = \gamma_j, \forall i, j \in \mathcal{G}_g, g = 1, \dots, G\}$ and the scaled penalty function as $\tilde{\rho}_\theta(\|\gamma_i - \gamma_j\|) = \lambda_1^{-1} \rho_\theta(\|\gamma_i - \gamma_j\|, \lambda_1)$. Let the least-squares objective function and the penalty function be

$$\begin{aligned} L(\gamma) &= \frac{1}{2} \|\mathbf{y} - W\gamma\|_2^2, \quad P(\gamma) = \lambda_1 \sum_{i < j} \tilde{\rho}_\theta(\|\gamma_i - \gamma_j\|_2) \\ L^G(\varphi) &= \frac{1}{2} \|\mathbf{y} - \mathbf{W}\Gamma\varphi\|_2^2, \quad P^G(\varphi) = \lambda_1 \sum_{g < g'} |\mathcal{G}_g| |\mathcal{G}_{g'}| \tilde{\rho}_\theta(\|\varphi_g - \varphi_{g'}\|_2). \end{aligned} \quad (\text{B.8})$$

Let $Q(\gamma) = L(\gamma) + P(\gamma)$, $Q(\gamma)^G(\varphi) = L^G(\varphi) + P^G(\varphi)$ and define

- ◇ $F : \mathcal{M}_G \rightarrow \mathbb{R}^{Gp}$. The g -th vector component of $F(\gamma)$ equals to the common value of γ_i for $i \in \mathcal{G}_g$.
- ◇ $F^* : \mathbb{R}^{np} \rightarrow \mathbb{R}^{Gp}$. $F^*(\gamma) = \{|\mathcal{G}_g|^{-1} \sum_{i \in \mathcal{G}_g} \gamma'_i, g = 1, \dots, G\}'$, which implies the average of each cluster vectors.

It results in that $F(\gamma) = F^*(\gamma)$ if $\gamma \in \mathcal{M}_G$. Hence, for every $\gamma \in \mathcal{M}_G$, $P(\gamma) = P^G(F(\gamma))$, and for every $\varphi \in \mathbb{R}^{Gp}$, $P(F^{-1}(\varphi)) = P^G(\varphi)$. Hence,

$$Q(\gamma) = Q^G(F(\gamma)), \quad Q^G(\varphi) = Q(F^{-1}(\varphi)). \quad (\text{B.9})$$

Theorem 1 results in that for some $\iota > 0$,

$$P(\sup_i \|\hat{\gamma}_i^{or} - \gamma_i^0\|_2 \leq p \sup_i \|\hat{\gamma}_i^{or} - \gamma_i^0\|_\infty = p \|\hat{\gamma}^{or} - \gamma^0\|_\infty \leq p \phi_{n,T,G,\zeta}) \geq 1 - e^{-\iota},$$

there exists an event E_1 in which $\sup_i \|\hat{\gamma}_i^{or} - \gamma_i^0\|_2 \leq p \phi_{n,T,G} = \tilde{\phi}_{n,T,G}$, such that $P(E_1^C) \leq e^{-\iota}$.

Consider the neighborhood of the true parameter γ^0 ,

$$\Theta := \{\gamma \in \mathbb{R}^{np} : \sup_i \|\gamma_i - \gamma_i^0\|_2 \leq \tilde{\phi}_{n,T,G}\}.$$

It implies that $\hat{\gamma}^{or} \in \Theta$ on the event E_1 . For any $\gamma \in \mathbb{R}^{np}$, let $\gamma^* = F^{-1}(F^*(\gamma))$, then $\gamma_i^* = \frac{1}{|\mathcal{G}_g|} \sum_{i \in \mathcal{G}_g} \gamma_i$ which implies that γ^* is a vector with duplicated group average of γ_i . Through two steps as the following, the statement can be proved that with probability approximating to 1, $\hat{\gamma}^{or}$

is a strictly local minimizer of $Q(\gamma)$.

- i. In E_1 , $Q(\gamma^*) > Q(\hat{\gamma}^{or})$ for any $\gamma \in \Theta$ and $\gamma^* \neq \hat{\gamma}^{or}$. This indicates that the oracle estimator $\hat{\gamma}^{or}$ is the minimizer over all duplicated group average γ^* .
- ii. There exists an event E_2 such that for large enough ι^* , $P(E_2^C) \leq e^{-\iota^*}$. In $E_1 \cap E_2$, there exists a neighborhood Θ_n of $\hat{\gamma}^{or}$ such that $Q(\gamma) \geq Q(\gamma^*)$ for all $\gamma^* \in \Theta_n \cap \Theta$ for sufficiently large n . It means that for all γ , the duplicated group average γ^* is the minimizer.

Then, it results in $Q(\gamma) > Q(\hat{\gamma}^{or})$ for any $\gamma \in \Theta_n \cap \Theta$ and $\gamma \neq \hat{\gamma}^{or}$ in $E_1 \cap E_2$. Hence, over $E_1 \cap E_2$, for large enough ι and ι^* , $\hat{\gamma}^{or}$ is a strictly local minimizer of $Q(\gamma)$ with the probability $P(E_1 \cap E_2) \geq 1 - e^{-\iota} - e^{-\iota^*}$.

First, show $P^G(F^*(\gamma)) = C$ for any $\gamma \in \Theta$, where C is a constant which does not depend on γ . It implies that when γ is close enough to the true parameter γ^0 , the penalty term would not affect the objective function with respect to different values of γ . Let $F^*(\gamma) = \varphi$. Consider the triangle inequality $\|\varphi_g - \varphi_{g'}\|_2 \geq \|\varphi_g^0 - \varphi_{g'}^0\|_2 - 2 \sup_g \|\varphi_g - \varphi_g^0\|_2$. Since $\gamma \in \Theta$, then

$$\begin{aligned}
\sup_g \|\varphi_g - \varphi_g^0\|_2^2 &= \sup_g \left\| |\mathcal{G}_g|^{-1} \sum_{i \in \mathcal{G}_g} \gamma_i - \varphi_g^0 \right\|_2^2 \\
&= \sup_g \left\| |\mathcal{G}_g|^{-1} \sum_{i \in \mathcal{G}_g} (\gamma_i - \gamma_i^0) \right\|_2^2 = \sup_g |\mathcal{G}_g|^{-2} \left\| \sum_{i \in \mathcal{G}_g} (\gamma_i - \gamma_i^0) \right\|_2^2 \\
&\leq |\mathcal{G}_g|^{-1} \sup_g \sum_{i \in \mathcal{G}_g} \|(\gamma_i - \gamma_i^0)\|_2^2 \leq \sup_i \|(\gamma_i - \gamma_i^0)\|_2^2 \leq \tilde{\phi}_{n,T,G}^2,
\end{aligned} \tag{B.10}$$

Since $b_{n,T,G} := \min_{g \neq g'} \|\varphi_g^0 - \varphi_{g'}^0\|$, then for all $g \neq g'$ and $b_{n,T,G} > a\lambda + 2\tilde{\phi}_{n,T,G}$,

$$\|\varphi_g^0 - \varphi_{g'}^0\|_2 \geq \|\varphi_g^0 - \varphi_{g'}^0\|_2 - 2 \sup_g \|\varphi_g - \varphi_g^0\|_2 \geq b_{n,T,G} - 2\tilde{\phi}_{n,T,G} > a\lambda_1,$$

for some $a > 0$. Then by Assumption 6, $\rho(\|\varphi_g - \varphi_{g'}\|_2)$ is a constant, and furthermore, $P^G(F^*(\varphi))$ is a constant. Therefore, $P^G(F^*(\gamma)) = C$, and $Q^G(F^*(\gamma)) = L^G(T^*(\gamma)) + C$ for all $\gamma \in \Theta$. Since $\hat{\varphi}^{or}$ is the unique global minimizer of $L_n^G(\varphi)$, then $L^G(T^*(\gamma)) > L^G(\hat{\varphi}^{or})$ for all $T^*(\gamma) \neq \hat{\varphi}^{or}$ and hence $Q^G(T^*(\gamma)) > Q^G(\hat{\varphi}^{or})$ for all $T^*(\gamma) \neq \hat{\varphi}^{or}$. By the property of the clustering algorithm, for

the g -th group, $\hat{\varphi}_g^{or} = |\mathcal{G}_g|^{-1} \sum_{i \in \mathcal{G}_g} \hat{\gamma}_i^{or}$, which implies that, along with the definition of operation F , $\hat{\varphi}_g^{or}$ equals to the g -th component of $F(\hat{\gamma}^{or})$ for all $i \leq g \leq G$. Then, by (B.9),

$$Q^{\mathcal{G}}(\hat{\varphi}^{or}) = Q^{\mathcal{G}}(T(\hat{\gamma}^{or})) = Q(\hat{\gamma}^{or}).$$

Furthermore, $Q_n^{\mathcal{G}}(T^*(\gamma)) = Q(T^{-1}(T^*(\gamma))) = Q(\gamma^*)$. Therefore, $Q(\gamma^*) > Q(\hat{\gamma}^{or})$ for all $\gamma^* \neq \hat{\gamma}^{or}$.

Second, for a positive sequence r_n , let $\Theta_n := \{\gamma_i : \sup_i \|\gamma_i - \hat{\gamma}_i^{or}\|_2 \leq r_n\}$. For any $\gamma \in \Theta_n \cap \Theta$, by the first order Taylor's expansion,

$$Q(\gamma) - Q(\gamma^*) = \frac{dQ(\gamma^m)}{d\gamma'}(\gamma - \gamma^*) = \frac{dL(\gamma^m)}{d\gamma'}(\gamma - \gamma^*) + \sum_{i=1}^n \frac{\partial P(\gamma^m)}{\partial \gamma'_i}(\gamma - \gamma^*),$$

and let $S_1 = \frac{dL(\gamma^m)}{d\gamma'_i}(\gamma - \gamma_i^*)$ and $S_2 = \sum_{i=1}^n \frac{\partial P(\gamma^m)}{\partial \gamma'_i}(\gamma_i - \gamma_i^*)$. Since

$$\begin{aligned} \frac{dL(\gamma)}{d\gamma_i} &= \frac{1}{2}(-2\mathbf{y}'W + 2\gamma'W'W) = -(\mathbf{y}' - \gamma'W)W \quad \text{and} \\ \frac{\partial P(\gamma)}{\partial \gamma_i} &= \lambda_1 \sum_{i=1}^n \tilde{\rho}'_{\theta}(\|\gamma_i - \gamma_j\|_2) \frac{1}{2\|\gamma_i - \gamma_j\|_2} 2(\gamma_i - \gamma_j) \\ &= \lambda_1 \sum_{i=1}^n \tilde{\rho}'_{\theta}(\|\gamma_i - \gamma_j\|_2) \frac{\gamma_i - \gamma_j}{\|\gamma_i - \gamma_j\|_2}, \end{aligned}$$

we have

$$S_1 = -(\mathbf{y}' - \gamma^m W)W(\gamma - \gamma^*) \quad \text{and} \quad S_2 = \sum_{i=1}^n \frac{\partial P(\gamma^m)}{\partial \gamma'_i}(\gamma_i - \gamma_i^*).$$

Let $\gamma^m = \vartheta\gamma + (1 - \vartheta)\gamma^*$ for some constant $\vartheta \in (0, 1)$. Then,

$$\begin{aligned}
S_2 &= \lambda_1 \sum_{i < j} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_i^*) \\
&\quad + \lambda_1 \sum_{i > j} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_i^*) \\
&= \lambda_1 \sum_{i < j} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_i^*) \\
&\quad + \lambda_1 \sum_{i < j} \tilde{\rho}'_{\theta}(\|\gamma_j^m - \gamma_i^m\|_2) \|\gamma_j^m - \gamma_i^m\|_2^{-1} (\gamma_j^m - \gamma_i^m)' (\gamma_j - \gamma_j^*) \\
&= \lambda_1 \sum_{i < j} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' [(\gamma_i - \gamma_i^*) - (\gamma_j - \gamma_j^*)].
\end{aligned} \tag{B.11}$$

Consider separating S_2 into two parts, $i, j \in \mathcal{G}_g$, and $i \in \mathcal{G}_g, j \in \mathcal{G}_{g'}$ for $g \neq g'$. When $i, j \in \mathcal{G}_g$, since $\gamma^* = T^{-1}(T^*(\gamma)) \in \mathcal{M}_{\mathcal{G}}$, then $\gamma_i^* = \gamma_j^*$. Thus, the RHS of (B.11) becomes

$$\begin{aligned}
S_2 &= \lambda_1 \sum_{g=1}^G \sum_{i, j \in \mathcal{G}_g, i < j} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_j) \\
&\quad + \lambda_1 \sum_{g < g'} \sum_{i \in \mathcal{G}_g, j \in \mathcal{G}_{g'}} \tilde{\rho}'_{\theta}(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i^m - \gamma_j^m\|_2^{-1} (\gamma_i^m - \gamma_j^m)' [(\gamma_i - \gamma_i^*) - (\gamma_j - \gamma_j^*)].
\end{aligned} \tag{B.12}$$

Furthermore, by (B.10), for any $\gamma \in \Theta_n \cap \Theta$, $F^*(\gamma) = \varphi$, and therefore, for all $i \in \mathcal{G}_g$, $\gamma_i^* = \varphi_g$.

This lead to

$$\sup_i \|\gamma_i^* - \gamma_i^0\|_2^2 = \sup_g \|\varphi_g - \varphi_g^0\|_2^2 \leq \tilde{\phi}_{n,T,G}^2, \tag{B.13}$$

where the inequality in (B.13) is obtained by (B.10). Since $\gamma_i^m = \vartheta\gamma_i + (1 - \vartheta)\gamma_i^*$, by the triangle inequality,

$$\begin{aligned}
\sup_i \|\gamma_i^m - \gamma_i^0\|_2 &= \sup_i \|\vartheta\gamma_i + (1 - \vartheta)\gamma_i^* - \gamma_i^0\|_2 \\
&= \sup_i \|\vartheta\gamma_i + (1 - \vartheta)\gamma_i^* - (\vartheta + 1 - \vartheta)\gamma_i^0\|_2 \\
&\leq \vartheta \sup_i \|\gamma_i - \gamma_i^0\|_2 + (1 - \vartheta) \sup_i \|\gamma_i^* - \gamma_i^0\|_2 \\
&\leq \vartheta \tilde{\phi}_{n,T,G} + (1 - \vartheta) \tilde{\phi}_{n,T,G} = \tilde{\phi}_{n,T,G}.
\end{aligned}$$

Hence, for $g \neq g', i \in \mathcal{G}_g, j \in \mathcal{G}_{g'}$,

$$\begin{aligned} \|\gamma_i^m - \gamma_j^m\|_2 &= \|\gamma_i^m - \gamma_i^0 - \gamma_j^m + \gamma_j^0\|_2 \geq \|\gamma_i^0 - \gamma_j^0\|_2 - 2 \max_{1 \leq k \leq n} \|\gamma_k^m - \gamma_k^0\|_2 \\ &\geq \min_{i \in \mathcal{G}_g, j' \in \mathcal{G}_{g'}} \|\gamma_i^0 - \gamma_{j'}^0\|_2 - 2 \max_{1 \leq k \leq n} \|\gamma_k^m - \gamma_k^0\|_2 \geq b_{n,T,G} - 2\tilde{\phi}_{n,T,G} > a\lambda_1. \end{aligned}$$

Since $\tilde{\rho}_\theta(x)$ is constant for all $x \geq a\lambda_1$, then $\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2) = 0$. Therefore, following $\gamma_i^m - \gamma_j^m = \vartheta(\gamma_i - \gamma_j)$ for $i, j \in \mathcal{G}_g$, (B.12) becomes

$$\begin{aligned} S_2 &= \lambda_1 \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \frac{\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2)}{\|\gamma_i^m - \gamma_j^m\|_2} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_j) \\ &\quad + \lambda_1 \sum_{g < g'} \sum_{i \in \mathcal{G}_g, j \in \mathcal{G}_{g'}} \frac{\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2)}{\|\gamma_i^m - \gamma_j^m\|_2} (\gamma_i^m - \gamma_j^m)' [(\gamma_i - \gamma_i^*) - (\gamma_j - \gamma_j^*)] \\ &= \lambda_1 \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \frac{\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2)}{\|\gamma_i^m - \gamma_j^m\|_2} (\gamma_i^m - \gamma_j^m)' (\gamma_i - \gamma_j) \\ &= \lambda_1 \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \frac{\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2)}{\|\vartheta(\gamma_i - \gamma_j)\|_2} \vartheta(\gamma_i - \gamma_j)' (\gamma_i - \gamma_j) \\ &= \lambda_1 \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2) \|\gamma_i - \gamma_j\|_2. \end{aligned}$$

Furthermore, similarly to (B.10), for all $i \in \mathcal{G}_g$, $\gamma_i^* = \varphi_g$, $\sup_i \|\gamma_i^* - \hat{\gamma}_i^{or}\|_2^2 = \sup_g \|\varphi_g - \hat{\varphi}_g^{or}\|_2^2 \leq \sup_i \|\gamma_i - \hat{\gamma}_i^{or}\|_2^2$. Then, since $\gamma_i^* = \gamma_j^*$,

$$\begin{aligned} \sup_i \|\gamma_i^m - \gamma_j^m\|_2 &= \sup_i \|\gamma_i^m - \gamma_i^* - \gamma_j^m + \gamma_j^*\|_2 \\ &\leq \|\gamma_i^* - \gamma_j^*\|_2 + 2 \sup_i \|\gamma_i^m - \gamma_i^*\|_2 \leq 2 \sup_i \|\gamma_i^m - \gamma_i^*\|_2 \\ &= 2 \sup_i \|\vartheta \gamma_i + (1 - \vartheta) \gamma_i^* - \gamma_i^*\|_2 \\ &= 2\vartheta \sup_i \|\gamma_i - \gamma_i^*\|_2 \leq 2 \sup_i \|\gamma_i - \gamma_i^*\|_2 \\ &\leq 2(\sup_i \|\gamma_i - \hat{\gamma}_i^{or}\|_2 + \sup_i \|\gamma_i^* - \hat{\gamma}_i^{or}\|_2) \\ &\leq 4 \sup_i \|\gamma_i - \hat{\gamma}_i^{or}\|_2 \leq 4r_n. \end{aligned}$$

Hence, $\tilde{\rho}'_\theta(\|\gamma_i^m - \gamma_j^m\|_2) \geq \tilde{\rho}'_\theta(4r_n)$, because $\rho(x)$ is nondecreasing and concave as assumed in Assumption 3. Then,

$$S_2 \geq \lambda_1 \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \tilde{\rho}'_\theta(4r_n) \|\gamma_i - \gamma_j\|_2. \quad (\text{B.14})$$

Let $U = (U'_1, \dots, U'_n)' = [(\mathbf{y} - W\gamma^m)'W]'$, then

$$\begin{aligned} S_1 &= -U'(\gamma - \gamma^*) = -(U'_1, \dots, U'_n)' \begin{pmatrix} \gamma_1 - \gamma_1^* \\ \gamma_2 - \gamma_2^* \\ \vdots \\ \gamma_n - \gamma_n^* \end{pmatrix} = -\sum_{i=1}^n U'_i(\gamma_i - \gamma_i^*) \\ &= -\sum_{g=1}^G \sum_{i \in \mathcal{G}_g} \frac{1}{|\mathcal{G}_g|} U'_i \left(|\mathcal{G}_g| \gamma_i - \sum_{j \in \mathcal{G}_g} \gamma_j \right) \\ &= -\sum_{g=1}^G \sum_{i \in \mathcal{G}_g} \frac{1}{|\mathcal{G}_g|} U'_i \sum_{j \in \mathcal{G}_g} (\gamma_i - \gamma_j) = -\sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g} \frac{U'_i(\gamma_i - \gamma_j)}{|\mathcal{G}_g|} \\ &= -\sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g} \frac{U'_i(\gamma_i - \gamma_j)}{2|\mathcal{G}_g|} + \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g} \frac{U'_j(\gamma_i - \gamma_j)}{2|\mathcal{G}_g|} \\ &= -\sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g} \frac{(U_j - U_i)'(\gamma_j - \gamma_i)}{2|\mathcal{G}_g|} \\ &= -\sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \frac{(U_j - U_i)'(\gamma_j - \gamma_i)}{|\mathcal{G}_g|}. \end{aligned} \quad (\text{B.15})$$

In addition, $U_i = W'_i(\mathbf{y}_i - W_i\gamma_i^m) = W'_i(W_i\gamma_i^0 + \boldsymbol{\varepsilon}_i - W_i\gamma_i^m) = W'_i(\boldsymbol{\varepsilon}_i + W_i(\gamma_i^0 - \gamma_i^m))$, and then,

$$\begin{aligned} \sup_i \|U_i\|_2 &\leq \sup_i \{\|W'_i\boldsymbol{\varepsilon}_i\|_2 + \|W'_i W_i(\gamma_i^0 - \gamma_i^m)\|_2\} \\ &\leq \sup_i \|W'_i\boldsymbol{\varepsilon}_i\|_2 + \sup_i \sqrt{p} \|W'_i W_i\|_\infty \tilde{\phi}_{n,T,G} \\ &\leq \sup_i \|W'_i\boldsymbol{\varepsilon}_i\|_2 + m\sqrt{pT}(q^{1/2} + m^{1/2}(L + 1 + 2K))\tilde{\phi}_{n,T,G} \\ &\leq \sup_i \sqrt{p} \|W'_i\boldsymbol{\varepsilon}_i\|_\infty + m\sqrt{pT}(q^{1/2} + m^{1/2}(L + 1 + 2K))\tilde{\phi}_{n,T,G} \\ &\leq \sqrt{p} \|W'\boldsymbol{\varepsilon}\|_2 + m\sqrt{pT}(q^{1/2} + m^{1/2}(L + 1 + 2K))\tilde{\phi}_{n,T,G} \\ &= \sqrt{p} \|W'\boldsymbol{\varepsilon}\|_2 + m\sqrt{pT} B_{q,m} \tilde{\phi}_{n,T,G}, \end{aligned}$$

where $B_{q,m} = q^{1/2} + m^{1/2}(L + 1 + 2K)$. By Lemma 3, $P \left[\|W' \varepsilon\|_2^2 > 2\tilde{c}(np + 2\sqrt{np\zeta^*} + 2\zeta^*)m\tilde{M}\sqrt{T}B_{q,m}\sqrt{p} \right] \leq e^{-\iota^*}$, where $B_{q,m} = (q^{1/2} + m^{1/2}(L + 1 + 2K))$, $p = q + L + 1 + 2K$, $\tilde{M} = \max(M_1, M_2, M_3, M_4)$ and \tilde{c} given in Assumption 2 and 4. ι^* is defined in Lemma 3. Then, over the event E_2 ,

$$\begin{aligned} & \left| \frac{(U_j - U_i)'(\gamma_j - \gamma_i)}{|\mathcal{G}_g|} \right| \leq g_{min}^{-1} \|U_j - U_i\|_2 \|\gamma_j - \gamma_i\|_2 \leq g_{min}^{-1} 2 \sup_i \|U_i\|_2 \|\gamma_i - \gamma_j\|_2 \\ & \leq 2g_{min}^{-1} T^{1/4} (mp)^{1/2} \|\gamma_i - \gamma_j\|_2 \\ & \left(p^{1/4} \tilde{B}_{q,m}^{1/2} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} + T^{1/4} m^{1/2} B_{q,m} \tilde{\phi}_{n,T,G} \right). \end{aligned} \quad (\text{B.16})$$

Therefore, by (B.14), (B.15) and (B.16),

$$\begin{aligned} & Q(\gamma) - Q(\gamma^*) \\ & \geq \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \|\gamma_i - \gamma_j\|_2 \\ & \quad \left\{ \lambda_1 \tilde{\rho}'_\theta(4r_n) - 2g_{min}^{-1} T^{1/4} (mp)^{1/2} (p^{1/4} \tilde{B}_{q,m}^{1/2} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} \right. \\ & \quad \left. + T^{1/4} m^{1/2} B_{q,m} \tilde{\phi}_{n,T,G}) \right\} \\ & \geq \sum_{g=1}^G \sum_{i,j \in \mathcal{G}_g, i < j} \|\gamma_i - \gamma_j\|_2 \\ & \quad \left\{ \lambda_1 \tilde{\rho}'_\theta(4r_n) - B_1 g_{min}^{-1} T^{1/4} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} - B_2 g_{min}^{-1} T^{1/2} \tilde{\phi}_{n,T,G} \right\}, \end{aligned}$$

where $B_1 = 2(mp\tilde{B}_{q,m})^{1/2} p^{1/4}$ and $B_2 = 2mp^{1/2} B_{q,m}$.

Let $r_n = o(1)$, then $\tilde{\rho}'_\theta(4r_n) \rightarrow 1$. Suppose that the following condition is true over the event $E_1 \cap E_2$,

$$B_1 g_{min}^{-1} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} T^{1/4} \rightarrow 0, \quad B_2 p g_{min}^{-1} T^{1/2} \tilde{\phi}_{n,T,G} \rightarrow 0, \quad (\text{B.17})$$

then $P(Q(\gamma) - Q(\gamma^*) \geq 0) \geq 1 - e^\iota - e^{\iota^*}$. Once (B.17) holds, $Q(\gamma) - Q(\gamma^*) \geq 0$ with probability approaching to 1 as $\iota, \iota^* \rightarrow \infty$.

Note that $\zeta^* = \zeta_{n,T,G}^*$ can be chosen as any sequence of numbers, as long as $\zeta^* \rightarrow \infty$ to ensure $\iota^* \rightarrow \infty$. In the following argument, conditions on n, T, G , and other numbers that satisfies (B.17) are spelled out:

1. As $T \rightarrow \infty$ with n fixed, the proposed estimator does not converge to the oracle estimator.
2. As $n \rightarrow \infty$ with T fixed, if conditions in Corollary 1 are satisfied, the second part of (B.17) is true. It is enough discuss the conditions for first part of (B.17). Choose ζ^* such that $\zeta^* \leq n$ and $\zeta^* \rightarrow \infty$ as $n \rightarrow \infty$. Let $g_{\min} \gg (p + 2\sqrt{p} + 2)^{1/2} n^{1/2}$. Since $(np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} = (p + 2\sqrt{p} + 2)^{1/2} O(n^{1/2})$,

$$B_1 g_{\min}^{-1} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} T^{1/4} \leq B_1 T^{1/4} g_{\min}^{-1} (p + 2\sqrt{p} + 2)^{1/2} O(n^{1/2}) \rightarrow 0.$$

- 3-1. Let $T, n \rightarrow \infty$. Consider the first part of (B.17). Choose ζ^* such that $\zeta^* \leq n$ and $\zeta^* \rightarrow \infty$ as $n \rightarrow \infty$. Let $g_{\min} \gg (p + 2\sqrt{p} + 2)^{1/2} n^{1/2} T^{1/4}$. Then

$$B_1 g_{\min}^{-1} (np + 2\sqrt{np\zeta^*} + 2\zeta^*)^{1/2} T^{1/4} \leq B_1 g_{\min}^{-1} (p + 2\sqrt{p} + 2)^{1/2} n^{1/2} T^{1/4} \rightarrow 0.$$

- 3-2. Let $T, n \rightarrow \infty$. Consider the second part of (B.17).

- (a) Suppose G is fixed. Choose ζ such that $\zeta = o(n^{4\tilde{\alpha}_1} T^{1/2})$ and $\zeta \rightarrow \infty$ as $n, T \rightarrow \infty$. Let $g_{\min} = O(n^{1/4 + \tilde{\alpha}_1})$ for some positive constant $\tilde{\alpha}_1 < 3/4$. Then, $(Gp + 2\sqrt{Gp\zeta} + 2\zeta)^{1/2} = O(2\zeta^{1/2})$, and

$$B_2 p g_{\min}^{-1} T^{1/2} \phi_{n,T,G} \leq B_2 p C_6 \frac{n^{1/2}}{g_{\min}^2 T^{1/4}} O(\zeta^{1/2}) \xrightarrow{n,T \rightarrow \infty} 0,$$

where $C_6 = 2C_{q,m} G^{3/4}$.

- (b) Suppose $G \rightarrow \infty$. Choose ζ such that $\zeta \leq G$ and $\zeta \rightarrow \infty$ as $n, T, G \rightarrow \infty$. Let $\frac{n^{7/13}}{T^{1/13}} \ll g_{\min} < n/G$. Then, $G \ll \frac{T^{1/13}}{n^{6/13}}$ and $Gp + 2\sqrt{Gp\zeta} + 2\zeta \leq (p + 2\sqrt{p} + 2)G = O(G)$. Further, since $G \leq n/g_{\min}$,

$$\begin{aligned} B_2 p g_{\min}^{-1} T^{1/2} \phi_{n,T,G} &\leq B_2 p C_5 \frac{n^{1/2} G^{3/4} T^{1/2}}{g_{\min}^2 T^{3/4}} O(G^{1/2}) \\ &\leq B_2 p C_5 \frac{n^{7/4}}{g_{\min}^{13/4} T^{1/4}} O(1) \xrightarrow{n,T,G \rightarrow \infty} 0, \end{aligned}$$

where $C_5 = C_{q,m} (p + 2\sqrt{p} + 2p)^{1/2}$, which is free from n, T and G .

- (c) Suppose $G \rightarrow \infty$. Let $g_{\min} = O(n^{5/11+\tilde{\alpha}_7})$ for a positive constant $\tilde{\alpha}_7 < 6/11$. Choose ζ such that $G \ll \zeta$ and $\zeta = o(n^{11\tilde{\alpha}_7/2}T^{1/2})$. Then, $(Gp + 2\sqrt{Gp\zeta} + 2\zeta)^{1/2} = o((p + 2\sqrt{p} + 2)^{1/2}\zeta^{1/2})$. Since $G \leq n/g_{\min}$,

$$B_2 p g_{\min}^{-1} T^{1/2} \phi_{n,T,G} \leq B_2 p C_7 \frac{n^{5/4}}{g_{\min}^{11/4} T^{1/4}} O(\zeta^{1/2}) \xrightarrow{n,T,G \rightarrow \infty} 0,$$

where $C_7 = C_{q,m}(p + 2\sqrt{p} + 2)^{1/2}$, which is free of n, T and G .

Combining the above calculations and the proof of Corollary 1, the conditions for (B.17) can be summarized as follows:

1. Suppose $n \rightarrow \infty$ with T fixed. Let $(p + 2\sqrt{p} + 2)^{1/2}n^{1/2} \ll g_{\min} = O(n^{7/9+\tilde{\alpha}_0}) \leq n/2$, then (B.17) holds;
2. Suppose $n, T \rightarrow \infty$ and G is fixed. Let $g_{\min} = O(n^{1/2+\tilde{\alpha}_4})$ for some constant $\tilde{\alpha}_4 < 1/2$. Then, (B.17) holds by choosing ζ and ζ^* such that $\zeta = o(\min(n^{1+4\tilde{\alpha}_4}T^{1/2}, n^{2\tilde{\alpha}_4}T^{3/2}))$ approaching infinity and $\zeta^* \leq n$ approaching infinity;
3. Suppose $n, T, G \rightarrow \infty$.
 - (a) Let $\max\left\{\frac{n^{7/13}}{T^{1/13}}, (p + 2\sqrt{p} + 2)^{1/2}n^{1/2}\right\} \ll g_{\min} = O(n^{7/9+\tilde{\alpha}_3})$ for some constant $\tilde{\alpha}_3 < 2/9$. Then, (B.17) holds by choosing $\zeta = O(G)$ and $\zeta^* \leq n$ approaching infinity;
 - (b) Let $g_{\min} = O(n^{5/7+\tilde{\alpha}_5})$ for some constant $\tilde{\alpha}_5 < 2/7$. Then, (B.17) holds by choosing $\zeta = o(\min\{n^{10/7+11/2\tilde{\alpha}_5}T^{1/2}, n^{7\tilde{\alpha}_5/2}T^{3/2}\})$.

□

References

- S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

- J. Breitung and C. Roling. Forecasting inflation rates using daily data: A nonparametric midas approach. *Journal of Forecasting*, 34(7):588–603, 2015.
- D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:1–6, 2012. doi: 10.1214/ECP.v17-2079.
- A. N. Langville and W. J. Stewart. The kronecker product and stochastic automata networks. *Journal of Computational and Applied Mathematics*, 167(2):429 – 447, 2004. ISSN 0377-0427.
- S. Ma and J. Huang. Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*, 2016.
- S. Ma and J. Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.
- L. Su, Z. Shi, and P. C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264, 2016.
- H. Wang, Z. Shi, and C.-S. Leung. Admm-mcp framework for sparse recovery with global convergence. *IEEE Transactions on Signal Processing*, Sep 2018, Forthcoming.
- X. Zhu and A. Qu. Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12:171–193, 01 2018.