Supplementary Appendix

Myungkou Shin*

November 26, 2024

A Exchangeability

Assumption 1 assumes that the cluster-level distribution contains sufficient information on the cluster heterogeneity λ_j. To motivate this assumption, let us consider a simple binary treatment model $Z_j \in \{0, 1\}$. When we consider a population distribution with a fixed number of individual per cluster and random sampling, Assumption 1 is a direct result of selection-on-observable and exchangeability. Let N_j^* denote the population number of individuals per cluster. N_j out of N_j^* individuals are randomly sampled. The observed dataset is

$$\left\{\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j\right\}_{j=1}^{J}$$

where $Y_{ij} = Y_{ij}(1) \cdot Z_j + Y_{ij}(0) \cdot (1 - Z_j)$ and the underlying population is

$$\left\{\{Y_{ij}(1)^*, Y_{ij}(0)^*, X_{ij}^*\}_{i=1}^{N_j^*}, Z_j\right\}_{j=1}^{J}$$

Clusters are independent of each other. Assume the following three assumptions:

(*random sampling*) *There is a random injective function* $\sigma_j : \{1, \cdots, N_j\} \to \{1, \cdots, N_j^*\}$

*School of Social Sciences, University of Surrey. Email: m.shin@surrey.ac.uk

such that

$$\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j} = \left\{Y_{\sigma_j(i)j}(1)^*, Y_{\sigma_j(i)j}(0)^*, X^*_{\sigma_j(i)j}\right\}_{i=1}^{N_j}.$$

σ_j is independent of $\left(\left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N_j*}, Z_j\right)$. Also, for any distinct $\left(i_1, \cdots, i_{N_j}\right)$

$$\Pr\left\{\sigma_j(1) = i_1, \cdots, \sigma_j(N_j) = i_{N_j}\right\} = \frac{\left(N_j^* - N_j\right)!}{N_j^*!}.$$

(*unconfoundedness*)

$$\{Y_{ij}(1)^*, Y_{ij}(0)^*\}_{i=1}^{N_j^*} \perp\!\!\!\perp Z_j \;\Big|\; \left\{X^*_{ij}\right\}_{i=1}^{N_j^*}.$$

(*exchangeability*) *For any permutation* σ^* *on* $\{1, \cdots, N_j^*\}$,

$$\left(\left\{Y_{ij}(1), Y_{ij}(0), X_{ij}\right\}_{i=1}^{N_j^*}, Z_j\right) \stackrel{d}{\equiv} \left(\left\{Y_{\sigma^*(i)j}(1), Y_{\sigma^*(i)j}(0), X_{\sigma^*(i)j}\right\}_{i=1}^{N_j^*}, Z_j\right).$$

Note that the *exchangeability* assumption restricts dependence structure within a given cluster in a way that the labelling of individuals should not matter. However, it still allows individual-level outcomes within a cluster to be arbitrarily correlated after conditioning on control covariates: for example, when X_{ij} includes a location variable, individuals close to each other is allowed to be more correlated than individuals further away from each other. Proposition A.1 follows immediately.

Proposition A.1. *Under random sampling, unconfoundedness and exchangeability,*

$$\{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j} \perp\!\!\!\perp Z_j \;\Big|\; \mathbf{F}_j$$

where $\mathbf{F}_j(x) = \frac{1}{N_j^*}\sum_{i=1}^{N_j^*} \mathbf{1}\{X^*_{ij} \leq x\}$.

Proof. Firstly, find that $\mathbf{E}[Z_j|\mathbf{F}_j]$ is an weighted average of $\mathbf{E}[Z_j|X^*_{\sigma^*(1)j}, \cdots, X^*_{\sigma^*(N_J)j}]$ across

all possible permutations σ^*. Thus, under the *exchangeability*,

$$\mathbf{E}[Z_j|\mathbf{F}_j] = \mathbf{E}[Z_j|X^*_{1j}, \cdots, X^*_{N_j j}] = \mathbf{E}[Z_j|X^*_{\sigma^*(1)j}, \cdots, X^*_{\sigma^*(N_j)j}]$$

for any permutation σ^*. Let $\pi(\mathbf{F}_j)$ denote $\mathbf{E}[Z_j|\mathbf{F}_j]$. Then,

$$\begin{aligned}
&\Pr\left\{Z_j = 1 \middle| \mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\right\} \\
&= \mathbf{E}\left[\mathbf{E}\left[Z_j \middle| \left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N^*_j}, \sigma_j\right] \middle| \mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[Z_j \middle| \left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N^*_j}\right] \middle| \mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[Z_j \middle| \left\{X^*_{ij}\right\}_{i=1}^{N^*_j}\right] \middle| \mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\right] \\
&= \mathbf{E}\left[\pi(\mathbf{F}_j) \middle| \mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\right] = \pi(\mathbf{F}_j) = \Pr\left\{Z_j = 1 \middle| \mathbf{F}_j\right\}.
\end{aligned}$$

The first equality holds since $\mathbf{F}_j$ is a function of $\{X^*_{ij}\}_{i=1}^{N^*_j}$ and $\{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}$ is a function of $\{Y_{ij}(1)^*, Y_{ij}(0)^*\}_{i=1}^{N^*_j}$ and σ_j. The second equality holds since *random sampling* implies that Z_j is independent of σ_j given $\left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N^*_j}$. The third equality is from *unconfoundedness*. □

Proposition A.1 suggests propensity score matching based on $\mathbf{F}_j$, the population distribution function of X_{ij} for cluster j. In this example, the population distribution is assumed to be discrete to explicitly invoke the exchangeability condition. Assumption 1 extends on this idea and assumes that the population distribution is possibly continuous and can be written as a function of a latent low-dimensional factor λ_j, which controls for the cluster-level heterogeneity, as does the propensity score $\pi(\mathbf{F}_j)$ in this example.

B Additional discussion on empirical illustration

B.1 Background

There exists a unique opportunity in research design when studying the question of whether an increase in minimum wage level leads to higher unemployment rate in the United States: the state-level variation in minimum wage. In the United States, each state has their own minimum wage level in addition to the federal minimum wage level and thus we see states with different minimum wage levels for the same time period. The state-level policy variation is helpful since it allows us to control for time heterogeneity in a flexible way, by comparing contemporaneous outcomes across states.

However, there could still be spatial heterogeneity that affects both minimum wage level and employment at the state level, which complicates the causal interpretation of a minimum wage regression. The literature has suggested several remedies for this spatial heterogeneity problem. For example, difference-in-differences (DiD) compares over-the-time difference in employment rate across states, assuming that spatial heterogeneity only exists as state heterogeneity and the state heterogeneity is cancelled out by taking the over-the-time difference (Card and Krueger, 1994). Some researchers limit their scope of analysis to counties that are located near the state border to account for spatial heterogeneity (Dube et al., 2010). Some use a more relaxed functional form assumption on state heterogeneity than DiD, such as state-specific linear trends (Allegretto et al., 2011, 2017). Some have the data construct a synthetic state that is comparable to an observed state (Neumark et al., 2014).

The clustered data setup in the paper fits the empirical context of the US minimum wage application well. Firstly, employment status, the outcome of interest, is observed at the individual level while the minimum wage level, the regressor of interest, is observed at the state level, i.e. the dataset is hierarchical. Secondly, an assumption that is shared in the minimum wage literature as a common denominator is that there is no dependence across states. In other words, it is believed that the decision of whether and how much the state

minimum wage level changes is only determined by what happens within the state. This corresponds to the clusters being independent.

Building on this observation, I apply the results of Sections 2 and 3 in the main text to control for the spatial heterogeneity in estimating the disemployment effect of the minimum wage. The key assumption in doing so is that the state-level distribution of individual-level demographic and socioeconomic characteristics sufficiently controls for the spatial heterogeneity. If the information that state legislators look at when deciding their state's minimum wage level is completely incorporated in the state-level distribution, the assumption would naturally hold. This 'distribution-as-control' approach is complementary to assuming that there exists some unrestricted and time-invariant state-level heterogeneity as in the two-way fixed-effect specification in the DiD literature. In the 'distribution-as-control' approach, the state-level heterogeneity is allowed to vary over time, but restricted in the sense that it is a function of the (near-)observable state-level distribution of individual-level characteristics.

B.2 Estimation

Following Allegretto et al. (2011); Neumark et al. (2014); Allegretto et al. (2017), I focus on the teen employment since it is likely that teenagers work at jobs that pay near the minimum wage level compared to adults, thus being more responsive to a change in the minimum wage level. I constructed a dataset by pooling the Current Population Survey (CPS) data from 2000 to 2021, collecting the same demographic control covariates on teenagers as Allegretto et al. (2011), and additional control covariates on all individuals. The additional variables were collected for every individual to construct state-level distributions, since information only from teenagers may not accurately reflect the state-level labor market status. Let $\mathcal{I}_{jt}$ denote the set of teens in state j at time t and $\tilde{\mathcal{I}}_{jt}$ denote the set of all individuals in state j at time t, from the CPS: $\mathcal{I}_{jt} \subset \tilde{\mathcal{I}}_{jt}$. Since the CPS is collected every month, the dataset contains $264 = 12 \cdot 22$ time periods in total.

The main regression specification I use is motivated from Allegretto et al. (2011). As

one of the two main regression specifications, Allegretto et al. (2011) estimates the following linear model: for teen i in state j at time t,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{1}$$

There are two noteworthy observations to be made here. Firstly, the regressor of interest $MinWage_{jt}$ varies on the state-by-time level, making state-by-time fixed-effects infeasible. This is exactly the same type of multicollinearity problem discussed in Section 2 of the main text. When treatment is assigned at the cluster level, treatment effects cannot be identified under a model with fully flexibly cluster heterogeneity. Thus, Allegretto et al. (2011) uses census-division-by-time fixed-effects by grouping 50 states and Washington D.C. into 9 census divisions: $\delta_{cd(j)t}$. Secondly, Equation (1) already implements the idea of aggregating individual-level information: the state-by-time employment rate $EmpRate_{jt}$ computed from Y_{ijt}. In using $EmpRate_{jt}$, a conscious choice was made by the researcher to use the mean to summarize the individual-level information for each state.

In this paper, I build upon the two observations above and develop a more flexible regression model:

$$Y_{ijt} = \alpha_j + \lambda_{jt}^{\intercal}\delta_t + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{2}$$

In the regression model, λ_{jt} is a time-varying state-level latent factor that I assume to be one-to-one with state-level distributions of individual-level characteristics. Specifically, I use the following two variables: $EmpHistory_{ijt}$ and $WageInc_{ijt}$. By using λ_{jt} as a control, I implement the 'distribution-as-control' approach. The latent factor λ_{jt} allows us to control for the spatial heterogeneity while not subsuming the variation in $MinWage_{jt}$. In doing so, λ_{jt} summarizes the available information at the individual level, in a more flexible way than the simple mean as in $EmpRate_{jt}$. In the next two paragraphs, I provide more detail on how I estimate the latent factor λ_{jt}, from the two state-level distributions.

Firstly, I apply the K-means clustering algorithm to the distribution of $EmpHistory_{ijt}$, an individual-level employment history variable:

$$\begin{aligned} EmpHistory_{ijt} &= \left(EmpStatus_{ijt-1}, \cdots, EmpStatus_{ijt-4}\right) \\ &\in \{Emp, Unemp, NotInLaborForce\}^4 =: \mathcal{X}. \end{aligned}$$

$EmpStatus_{ijt}$ is an employment status variable for individual i in state j at time t. It is a categorical variable with three possible values: being employed, being unemployed, and not being in the labor force. $EmpHistory_{ijt}$ concatenates $EmpStatus_{ij\tau}$ for $\tau = t-4, \cdots, t-1$; $EmpHistory_{ijt}$ is a four-month-long history of employment status. Since $EmpStatus_{ijt}$ is a categorical variable with a finite support of three elements, $EmpHistory_{ijt}$ has a finite support of 81 elements. Note that $Y_{ijt} = 1 \Leftrightarrow EmpStatus_{ijt} = Emp$ and thus $EmpHistory_{ijt}$ can be understood as a vector of lagged outcome variables, but defined for both teenagers and adults. To aggregate the information from $EmpHistory_{ijt}$ to learn about the labor market fundamental of a given state, I compute the empirical distribution function: for $x \in \mathcal{X}$,

$$\hat{\mathbf{F}}_{jt}(x) = \frac{1}{\sum_{i \in \tilde{\mathcal{I}}_{jt}} \tilde{\omega}_i} \sum_{i \in \tilde{\mathcal{I}}_{jt}} \mathbf{1}\{EmpHistory_{ijt} = x\}\tilde{\omega}_i$$

$\{\tilde{\omega}_i\}_i$ are the longitudinal weights provided by the IPUMS-CPS to construct a four-month-long panel using the CPS sample. Note that $\tilde{\mathcal{I}}_{jt}$ is used instead of $\mathcal{I}_{jt}$; information from adults' employment history is included. When evaluating the distance between states measured in terms of $\hat{\mathbf{F}}_{jt}$, I use the uniform weighting function since $\mathcal{X}$ is a finite set. By applying the K-means algorithm to $\left\{\hat{\mathbf{F}}_{jt}\right\}_{j=1}^{J}$, I get $\{\hat{\lambda}_{jt,EmpHistory}\}_{j=1}^{J}$.

Secondly, I apply the functional PCA to the distribution of $WageInc_{ijt}$. $WageInc_{ijt}$ is a wage income variable for individual i in state j at time t. Since the current and past unemployment rates are already controlled with $EmpRate_{jt}$ and the distribution of $EmpHistory_{ijt}$, I consider a truncated distribution of $WageInc_{ijt}$ by focusing on individuals

whose wage income is strictly positive. The wage income variable comes from the March Annual Social and Economic Supplement (ASEC). The ASEC sample is collected only once a year in March and is different from the basic monthly CPS sample. Let $\breve{\mathcal{I}}_{jt}$ denote the set of all individuals with positive wage income in state j, from the most recent ASEC sample at time t. Then, $\breve{\mathcal{I}}_{jt} = \breve{\mathcal{I}}_{jt+1}$ except when t corresponds to a month of March and $\breve{\mathcal{I}}_{jt} \neq \tilde{\mathcal{I}}_{jt}$ in general. To aggregate the information from $WageInc_{ijt}$, I compute the product of the state-level conditional densities of $\log WageInc_{ijt}$. The j-th row and k-th column component of the estimated product matrix $\hat{M}_t$ is

$$\hat{M}_{jkt} = \begin{cases} \frac{\sum_{i\in\breve{\mathcal{I}}_{jt}, i'\in\breve{\mathcal{I}}_{kt}} \breve{\omega}_i\breve{\omega}_{i'}}{\sum_{i\in\breve{\mathcal{I}}_{jt}, i'\in\breve{\mathcal{I}}_{kt}} \breve{\omega}_i\breve{\omega}_{i'}} \int_{\mathbb{R}} \frac{\breve{\omega}_i\breve{\omega}_{i'}}{h^2} K\left(\frac{x-\log WageInc_{ijt}}{h}\right) \cdot K\left(\frac{x-\log WageInc_{i'kt}}{h}\right) w(x)dx, & \text{if } j \neq k \\ \frac{\sum_{i,i'\in\breve{\mathcal{I}}_{jt}, i\neq i'} \breve{\omega}_i\breve{\omega}_{i'}}{\sum_{i,i'\in\breve{\mathcal{I}}_{jt}, i\neq i'} \breve{\omega}_i\breve{\omega}_{i'}} \int_{\mathbb{R}} \frac{\breve{\omega}_i\breve{\omega}_{i'}}{h^2} K\left(\frac{x-\log WageInc_{ijt}}{h}\right) \cdot K\left(\frac{x-\log WageInc_{i'jt}}{h}\right) w(x)dx, & \text{if } j = k \end{cases}.$$

$\{\breve{\omega}_i\}_i$ are the cross-sectional weights provided by the IPUMS-CPS to construct a cross-section with the ASEC sample. For the weighting function w, I use the uniform weighting on [0,15]: $w(x) = \frac{1}{1001}\mathbf{1}\{x \in \{0, 15/1000, \cdots, 15\}\}$. By applying the eigenvalue decomposition to $\hat{M}_t$, I get $\{\hat{\lambda}_{jt,WageInc}\}_{j=1}^J$. An estimate for the entire latent factor λ_{jt} is obtained from concatenating $\hat{\lambda}_{jt,EmpHistory}$ and $\hat{\lambda}_{jt,WageInc}$.

B.2.1 Cross-validation on the dimension of the latent factor

Both of the latent factor estimation methodologies introduced in the paper involve an unknown parameter: ρ, the dimension of the latent factor. To decide on ρ, I conduct a 5-fold cross-validation exercise for a given time t.

1. Fix ρ and randomly split the individual indices for a given state into five subsets: $\tilde{\mathcal{I}}_{jt} = \cup_{k=1}^5 \tilde{\mathcal{I}}_{jt,k}$ and $\breve{\mathcal{I}}_{jt} = \cup_{k=1}^5 \breve{\mathcal{I}}_{jt,k}$, respectively for $EmpHistory_{ijt}$ and $WageInc_{ijt}$. For each k, define the train sets $\left\{\tilde{\mathcal{I}}_{jt,-k} = \tilde{\mathcal{I}}_{jt} \setminus \tilde{\mathcal{I}}_{jt,k}\right\}_{j=1}^J$ and $\left\{\breve{\mathcal{I}}_{jt,-k} = \breve{\mathcal{I}}_{jt} \setminus \breve{\mathcal{I}}_{jt,k}\right\}_{j=1}^J$.

2. For each k, construct $\left\{\hat{\mathbf{F}}_{jt,-k}(x)\right\}_{j=1}^J$ and $\hat{M}_{t,-k}$ from their respective train sets and estimate $\lambda_{jt,EmpHistory}$ and $\lambda_{jt,WageInc}$ with the predetermined value of ρ.

3. Evaluate the out-of-sample performance of the estimated models from Step 2, using the test sets. For each k, construct $\left\{\hat{\mathbf{F}}_{jt,k}\right\}_{j=1}^{J}$ and $\hat{M}_{t,k}$ with their respective test sets $\left\{\tilde{\mathcal{I}}_{jt,k}\right\}_{j=1}^{J}$ and $\left\{\breve{\mathcal{I}}_{jt,k}\right\}_{j=1}^{J}$ and let

$$\text{SSFE}_{t,EmpHistory}(\rho) = \frac{1}{5}\sum_{k=1}^{5}\sum_{j=1}^{J}\sum_{x\in\mathcal{X}}\left(\hat{\mathbf{F}}_{jt,k}(x) - \hat{G}_{-k,EmpHistory}\left(\hat{\lambda}_{jt,-k,EmpHistory}\right)\right)^2,$$

$$\text{SSFE}_{t,WageInc}(\rho) = \frac{1}{5}\sum_{k=1}^{5}\left\|\hat{M}_{t,k} - \tilde{M}_{t,-k}\right\|_F^2.$$

$\hat{G}_{-k,EmpHistory}\left(\hat{\lambda}_{jt,-k,EmpHistory}\right)$ is the fitted value of the empirical distribution function $\mathbf{F}_{jt}$ from applying the K-means algorithm with ρ groups to $\left\{\hat{\mathbf{F}}_{jt,-k}\right\}_{j=1}^{J}$ from the train set. $\tilde{M}_{t,-k}$ is the fitted value of the product matrix M from applying the eigenvalue decomposition to the estimated product matrix $\hat{M}_{t,-k}$ from the train set and suppressing the $J-\rho$ smallest eigenvalues to zero.

The random splitting is a valid strategy in constructing a test set and a train set since the individuals are assumed to be iid within a cluster. To evaluate the performance of a latent factor model with the dimension ρ, I use the same criteria used in estimating the latent factor model. To see if the cross-validation result is stable across t, I consider the first and the last months of the timeframe—January 1990 and December 2021—and a month in the middle—January 2007—, which is used for a cross-sectional regression in the main text.

	ρ			
t	2	3	5	7
January 1990	0.5861	0.5854	0.5616	0.5558
January 2007	0.6139	0.5891	0.5970	0.5991
December 2021	0.7690	0.7141	0.7581	0.7592
average	0.6503	0.6295	0.6389	0.6380

Table 1: $\text{SSFE}_{t,EmpHistory}(\rho)$

t	ρ				
	1	2	3	5	7
March 1989	0.7018	0.6847	0.6852	0.6852	0.6855
March 2006	1.0001	0.9826	0.9834	0.9842	0.9842
March 2021	0.8711	0.8006	0.8010	0.8018	0.8019
average	0.9011	0.8833	0.8840	0.8845	0.8846

Table 2: $\text{SSFE}_{t,WageInc}(\rho)$

The distribution of $WageInc_{ijt}$ is only observed once a year, in March. Thus, the distributions of $WageInc_{ijt}$ at the time periods above are used as control for the three months I consider: January 1990, January 2007 and December 2021.
The triangular kernel is used and the tuning parameter h is selected by the *density* function in *R*.

Table 1 contains the cross-validation results for the K-means algorithm on the distribution of $EmpHistory_{ijt}$ and Table 2 contains the cross-validation results for the functional PCA on the distribution of $WageInc_{ijt}$. Table 1 shows that the cross-validation result is not stable across t for the K-means algorithm on the distribution of $EmpHistory_{ijt}$. The cross-validation result from January 1990 suggests using a latent factor model with larger dimension while the other two cross-validation results suggest using a latent factor model with $\rho_{Kmeans} = 3$. I take the average of the three cross-validation results and let $\rho_{Kmeans} = 3$. As a robustness check, I also present the estimation results from $\rho_{Kmeans} = 5$ below: Section B.3.2. On the other hand, the cross-validation result is stable across t for the functional PCA on the distribution of $WageInc_{ijt}$. I let $\rho_{fPCA} = 2$.

When the sole purpose of estimating the cluster-level latent factors λ_{jt} is to use the factors as controls for the spatial/state-level heterogeneity, one could repeat the cross-validation exercise for every t and let ρ vary across t. However, when the state-level heterogeneity is an object of interest on this own, letting ρ time-invariant can be helpful since then we could connect the support of the latent factor $\mathcal{S}_\lambda$ across different time periods and obtain a pooled estimate on the equilibrium/contextual effect that the state-level distribution $\mathbf{F}_{jt}$ has on

individual-level outcomes. This point on the aggregate-level heterogeneity will be reiterated in Section B.3.2.

B.3 Empirical results

B.3.1 Latent factor estimation for January 2007

Before discussing the estimation results from the regression model (2), here I illustrate how the two latent factor estimation methods are implemented on an actual dataset, by looking at a snapshot of the dataset. As for the timing of the snapshot, I choose January 2007 as I did in the main text, since January 2007 was when the most states raised their minimum wage levels without a federal minimum wage raise.

The outcome of the K-means latent factor estimation is a grouping structure on states. Since $EmpHistory_{ijt}$ captures the latest four month history of individual employment status, the latent factor estimation for January 2007 assigns 50 states and Washington D.C. into one of the $\rho_{Kmeans} = 3$ groups based on the state-level distribution of employment history from September 2006 to December 2006. Figure 1 contains the grouping result and below is the list of states in each group:

Group 1: **Arizona***, Arkansas, **California***, DC, Louisiana, Michigan, Mississippi, New Mexico, **New York***, Oklahoma, **Oregon***, South Carolina, Tennessee, West Virginia

Group 2: Alabama, **Connecticut***, **Delaware***, **Florida***, Georgia, **Hawaii***, Idaho, Illinois, Indiana, Kentucky, Maine, Maryland, **Massachusetts***, **Missouri***, Nevada, New Jersey, **North Carolina***, **Ohio***, **Pennsylvania***, **Rhode Island***, Texas, Utah, Virginia

Group 3: Alaska, **Colorado***, Iowa, Kansas, Minnesota, **Montana***, Nebraska, New Hampshire, North Dakota, South Dakota, **Vermont***, **Washington***, Wisconsin, Wyoming

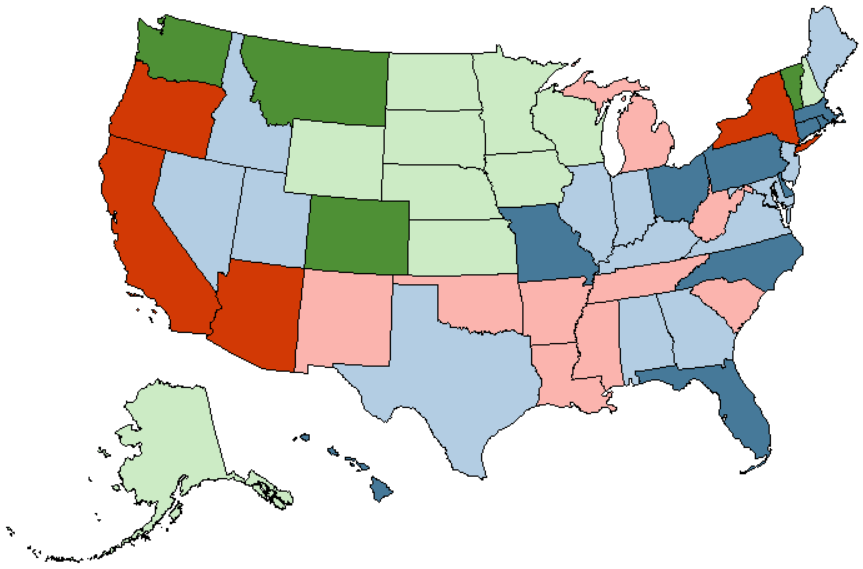

Figure 1: Grouping of states from the distribution of $EmpHistory_{ijt}$, January 2007

50 states and Washing D.C. are grouped into three groups based on the state-level distribution of individual-level employment history from September 2006 to December 2006, which tracks employment, unemployment, and labor force participation. Colors—red, blue, green—denote different groups and darker shades denote an increase in the minimum wage level in January 2007.

The states that raised their minimum wage level starting January 2007 are denoted with boldface and asterisk in the list and with darker shade in the figure. We can estimate a 'treatment effect,' by interpreting the increase in the minimum wage level as a binary treatment. The within-group comparison is free of the potential treatment endogeneity problem when the distribution of $EmpHistory_{ijt}$ gives us unconfoundedness.

Table 3 shows how the groups estimated using the distribution of $EmpHistory_{ijt}$ differ from one another. Table 3 takes three subsets of $\mathcal{X}$ and computes the proportion of each subset across groups, putting equal weights over states. The three subsets are:

- Always-employed: $\{Emp\}^4$
- Ever-unemployed: $\left\{(EmpStatus_{-1}, \cdots) : EmpStatus_\tau = Unemp \text{ for some } \tau\right\}$
- Never-in-the-labor-force: $\{NotInLaborForce\}^4$

group	1	2	3
Always-employed	0.520	0.588	0.645
Ever-unemployed	0.076	0.060	0.060
Never-in-the-labor-force	0.337	0.281	0.227

Table 3: Heterogeneity across states, January 2007

The table reports proportions of three types of employment history, across 50 states and Washington D.C. The proportions of each employment history are firstly computed within states and then the group mean is computed by putting equal weights on states.

'Always-employed' is the proportion of individuals who have been continuously employed from September 2006 to December 2006, 'Ever-unemployed' is the proportion of individuals who was unemployed for at least one month, and 'Never-in-the-labor-force' is the proportion of individuals who have never been in the labor force from September 2006 to December 2006. Group 1 states have the lowest employment rate and Group 3 states have the highest.

Secondly, to illustrate how the functional PCA is applied to a real dataset, I look at March 2006 ASEC sample; this sample is used for the latent factor estimation on the distribution of $WageInc_{ijt}$ for January 2006, due to the ASEC sample being observed only once a year. After applying the eigenvalue decomposition to the product matrix computed from the conditional densities of $\log WageInc_{ijt}$ given $WageInc_{ijt} > 0$ across 50 states and Washington D.C., the second to the fourteenth largest eigenvalues are plotted in Figure 2. The biggest eigenvalue is much bigger than the rest of the eigenvalues, with the associated eigenvectors being mostly constant across states, and is therefore omitted. We can see that the second biggest eigenvalue is much bigger than the third to the fourteenth eigenvalues. This means that the additional gain in explaining the variation across the state-level conditional densities of $\log WageInc_{ijt}$ is much bigger when we increase ρ_{fPCA} from one to two, than when we further increase ρ_{fPCA} from two. This observation is coherent with the cross-validation results from the previous subsection that choose $\rho_{fPCA} = 2$.

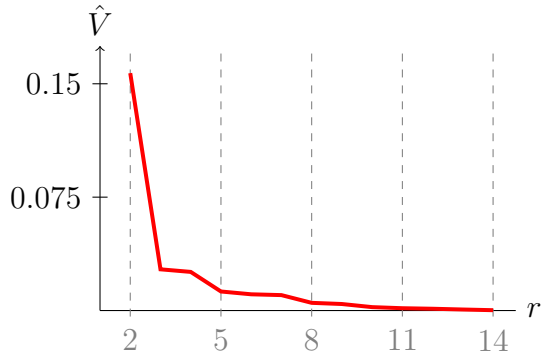

Figure 2: The scree plot of eigenvalues from the distribution of $WageInc_{ijt}$, March 2006

March 2006 ASEC sample is used in constructing the wage income densities. $WageInc_{ijt}$ is truncated at $WageInc_{ijt} > 0$ and logged. The biggest eigenvalue is not included in the plot. Its value is 15.37.

In addition, I plotted the second component of the estimated latent factor in Figure 3. Several northeastern states and Alaska have higher values of the second component of $\hat{\lambda}_{jt,WageInc}$ while some southern states such as Arkansas and Mississippi and mountain states such as Montana have lower values. We do not have an interpretation for the second

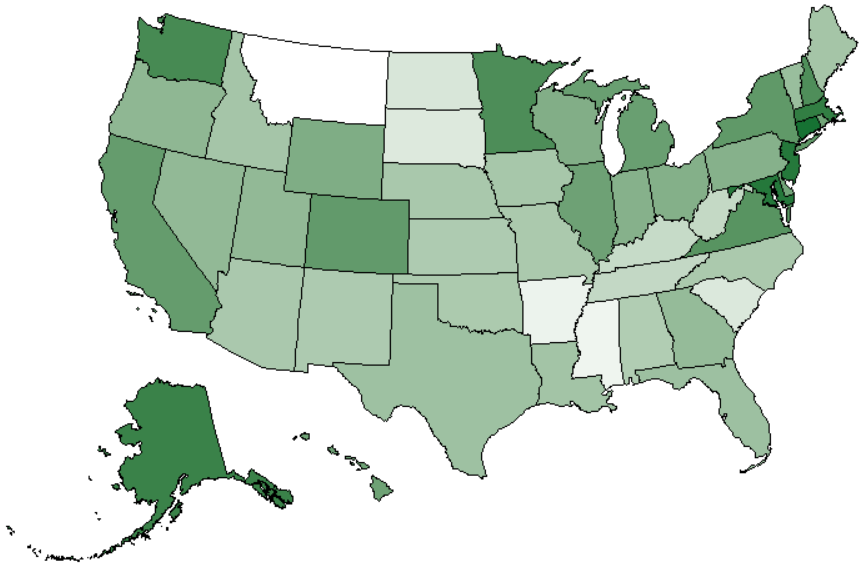

Figure 3: The second component of $\hat{\lambda}_{jt,WageInc}$ across states, March 2006

March 2006 ASEC sample is used in constructing the wage income densities. $WageInc_{ijt}$ is truncated at $WageInc_{ijt} > 0$ and logged. The darker shade corresponds to a higher value of $\hat{\lambda}_{2jt,WageInc}$ and the lighter shade correspond to a lower value.

component of $\hat{\lambda}_{jt,WageInc}$; Figure 3 only tells us which states are similar in that regard. Not being able to interpret the value of $\hat{\lambda}_{jt,WageInc}$ is due to the rotation on the latent factor and is a definite caveat of the latent factor models suggested in the paper. However, as discussed in the main text, not being able to interpret the estimates does not stop us from having an interpretable model and we can still conduct comparative statistics in terms of the distribution of $WageInc_{ijt}$.

B.3.2 Pooled regression on disemployment effect

Now, I discuss the regression results from (2). Table 4 expands Table 4 of the main text and includes estimation results when $\rho_{Kmeans} = 5$. As in the main text, I use time-specific coefficients for $\lambda_{jt,EmpHistory}$ and time-invariant coefficients for $\lambda_{jt,WageInc}$. Columns (2) and (5) contain the estimation results when $\rho_{Kmeans} = 3$ and columns (3) and (6) contain the estimation results when $\rho_{Kmeans} = 5$. The estimation results are stable across the choice of ρ_{Kmeans}.

$\hat{\beta}$	(1)	(2)	(3)	(4)	(5)	(6)
	-0.109*	-0.052	-0.061	-0.029*	-0.030*	-0.033**
	(0.061)	(0.078)	(0.086)	(0.017)	(0.017)	(0.016)
time FE	X	X	X	O	O	O
$EmpHistory$	X	O ($\rho = 3$)	O ($\rho = 5$)	X	O ($\rho = 3$)	O ($\rho = 5$)
$WageInc$	X	O	O	X	O	O
T	1 (January 2007)			264 (2000-2021)		

Table 4: Disemployment effect estimates across specifications

The categorical latent factors from the distribution of $EmpHistory_{ijt}$ are given time-varying loadings while the continuous latent factors from the distribution of $WageInc_{ijt}$ are given time-invariant loadings:

$$\lambda_{jt}{}^{\intercal}\delta_t = \lambda_{jt,EmpHistory}{}^{\intercal}\delta_{t,EmpHistory} + \lambda_{jt,WageInc}{}^{\intercal}\delta_{WageInc}.$$

The standard errors are clustered at the state level.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

In the regression model (2), the state minimum wage level $MinWage_{jt}$ enters after taking logarithm, following the convention in the literature. Thus, by dividing the slope coefficient on $\log MinWage_{jt}$ with the average teen employment rate from the dataset, which is 0.328, we get the elasticity interpretation. Based on columns (4)-(6) of Table 4, an one percentage point increase in the minimum wage level reduces teen employment by 0.087-0.099 percentage point. Neumark and Shirley (2022) provides a meta-analysis of studies on teen employment and minimum wage and find that the mean of the estimates across studies is -0.170 and the median is -0.122. By controlling for the state-level heterogeneity in a more rigorous manner using the state-level distributions, I find that the existing literature slightly overestimates the wage elasticity of teen employment.

$\hat{\beta}$	(1)	(2)	(3)
$\{\lambda_{EmpHistory} = e_1\}$	-0.027	-0.027*	-0.027
	(0.017)	(0.016)	(0.017)
$\{\lambda_{EmpHistory} = e_2\}$	-0.029*	-0.028	-0.028
	(0.017)	(0.017)	(0.017)
$\{\lambda_{EmpHistory} = e_3\}$	-0.030*	-0.042*	-0.042*
	(0.017)	(0.024)	(0.023)
time FE	O	O	O
$EmpHistory$	X	O	O
$WageInc$	X	X	O
T	264 (2000-2021)		

Table 5: Aggregate-level heterogeneity in disemployment effect

The categorical latent factors from the distribution of $EmpHistory_{ijt}$ are given time-varying loadings while the continuous latent factors from the distribution of $WageInc_{ijt}$ are given time-invariant loadings:

$$\lambda_{jt}{}^{\intercal}\delta_t = \lambda_{jt,EmpHistory}{}^{\intercal}\delta_{t,EmpHistory} + \lambda_{jt,WageInc}{}^{\intercal}\delta_{WageInc}.$$

The standard errors are clustered at the state level.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

Table 5 discuss the aggregate heterogeneity in disemployment effect:

$$Y_{ijt} = \log MinWage_{jt} \cdot \left(\sum_r \beta_r \mathbf{1}\{\lambda_{jt,EmpHistory} = e_r\} \right) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (3)$$

The coefficient on the minimum wage is a function of the distribution of $EmpHistory_{ijt}$. To connect the 'labels' of the grouping structure across different time periods, I reordered $\lambda_{jt,EmpHistory}$ across t so that Group 1 (i.e. $\lambda_{jt,EmpHistory} = e_1$) is always the group of states with lower employment rate and lower labor force participation rate and Group 3 (i.e. $\lambda_{jt,EmpHistory} = e_3$) is always the group of states with higher employment rate and higher labor force participation rate. Columns (3)-(4) show us that teens in Group 3 states are more affected by the minimum wage than teens in Group 1 states. This may happen for a variety of reasons; e.g., Group 3 states may have thicker labor supply on lower end of the wage distribution and thus low-skilled teenagers get replaced more easily.

Lastly, I study the interaction between the aggregate-level heterogeneity and the individual-level heterogeneity in terms of race. The left panel of Table 6 estimates

$$Y_{ijt} = \log MinWage_{jt} \cdot \Big(\beta_1 \mathbf{1}\{White_{ij} = 1\} + \beta_0 \mathbf{1}\{White_{ij} = 0\} \Big) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (4)$$

Disemployment effect is modeled to be heterogeneous at the individual level in terms of race. β_1 is the disemployment effect coefficient on white teenagers and β_0 is the disemployment effect coefficient on non-white teenagers. The right panel of Table 6 estimates

$$Y_{ijt} = \log MinWage_{jt} \cdot \Big(\sum_{w=0,1} \sum_r \beta_{w,r} \mathbf{1}\{White_{ij} = w, \lambda_{jt,EmpHistory} = e_r\} \Big) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (5)$$

The aggregate-level heterogeneity in terms of the distribution of $EmpHistory_{ijt}$ is introduced, in addition to the individual-level heterogeneity in terms of race. $\beta_{1,r}$ is the disemployment effect coefficient on white teenagers in Group r states while $\beta_{0,r}$ is the disemployment effect coefficient on non-white teenagers in Group r states.

From the left panel of Table 6, we see that a raise in the minimum wage level decreases the employment rate of white teens and increases the employment rate of non-white teens. The racial disparity interacts with the labor market fundamentals. The right panel of Table 6 shows us that the racial disparity persists across groups and interact with the aggregate heterogeneity in a way that the employment effect for non-white teenagers is mitigated in Group 3. Figure 4 contains confidence intervals for interactive disemployment effect coefficients from Column (4) of Table 5 in the main text and Column (4) of Table 6.

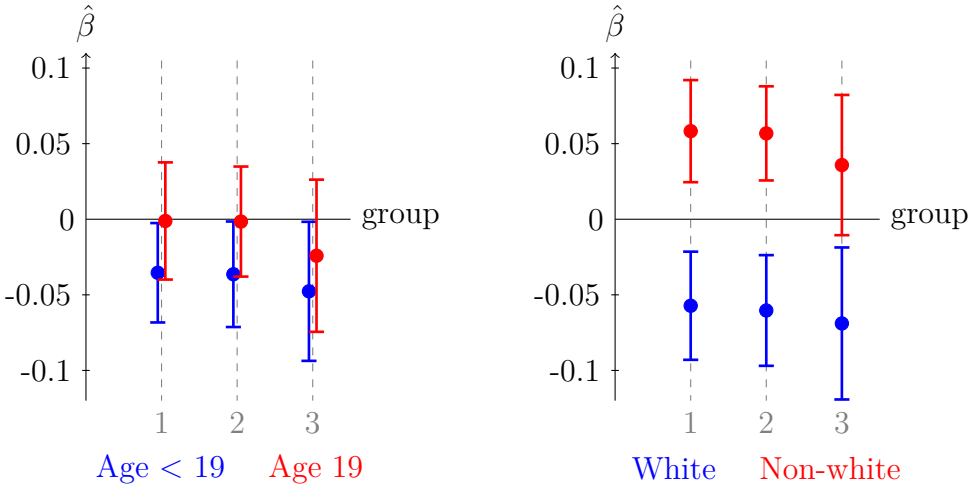

Figure 4: 95% confidence intervals on disemployment effect coefficient

The x-axis denotes the group. The color denotes the individual-level control covariate.
The y-axis is estimates and confidence interval.
Comparison across colors at each point of the x-axis relates to individual heterogeneity and comparison across x-axis for the same color relates to aggregate heterogeneity.

$\hat{\beta}$	(1)	(2)	(3)	(4)
$\{White = 1\}$	-0.061***	-0.061***		
	(0.018)	(0.018)		
$\times\{\lambda_{EmpHistory} = e_1\}$			-0.057***	-0.057***
			(0.018)	(0.018)
$\times\{\lambda_{EmpHistory} = e_2\}$			-0.060***	-0.060***
			(0.019)	(0.019)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.069**	-0.069**
			(0.026)	(0.026)
$\{White = 0\}$	0.054***	0.054***		
	(0.016)	(0.016)		
$\times\{\lambda_{EmpHistory} = e_1\}$			0.058***	0.058***
			(0.017)	(0.017)
$\times\{\lambda_{EmpHistory} = e_2\}$			0.057***	0.057***
			(0.016)	(0.016)
$\times\{\lambda_{EmpHistory} = e_3\}$			0.036	0.036
			(0.024)	(0.024)
$EmpHistory$	O	O	O	O
$WageInc$	X	O	X	O
T	264 (2000-2021)			

Table 6: Individual-level and interactive heterogeneity in disemployment effect

The categorical latent factors from the distribution of $EmpHistory_{ijt}$ are given time-varying loadings while the continuous latent factors from the distribution of $WageInc_{ijt}$ are given time-invariant loadings:

$$\lambda_{jt}{}^{\intercal}\delta_t = \lambda_{jt,EmpHistory}{}^{\intercal}\delta_{t,EmpHistory} + \lambda_{jt,WageInc}{}^{\intercal}\delta_{WageInc}.$$

The standard errors are clustered at the state level.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

C Proofs

C.1 Theorem 1

Firstly, we want to show that the objective function constructed with the estimated latent factors is close to the infeasible objective function with the rotated true latent factors: for any $\theta \in \tilde{A}\Theta$,

$$\left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) \right\|_2 = \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 + \frac{1}{\sqrt{J}} \cdot o_p(1).$$

By taking the first-order Taylor's expansion of m with regard to $\hat{\lambda}_j$ around $A\lambda_j$,

$$\begin{aligned} \left| \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) \right\|_2 - \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \right| &\leq \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) - \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \\ &= \left\| \frac{1}{J} \sum_{j=1}^{J} R_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right) \right\|_2. \end{aligned}$$

We can apply the Taylor's expansion since the mapping $\lambda \mapsto m\big(W_j(\lambda); \theta\big)$ is continuously differentiable on $A\mathcal{S}_\lambda$ for each $\theta \in \tilde{A}\Theta$: for any $\theta \in \tilde{A}\Theta$ and any λ' in the interior of $A\mathcal{S}_\lambda$,

$$\begin{aligned} m(W_j(\lambda'); \theta) &= m\big(W_j(A^{-1}\lambda'); \tilde{A}^{-1}\theta\big) \\ \frac{\partial}{\partial \lambda} m(W_j(\lambda); \theta)\Big|_{\lambda=\lambda'} &= \frac{\partial}{\partial \lambda} m\big(W_j(A^{-1}\lambda); \tilde{A}^{-1}\theta\big)\Big|_{\lambda=\lambda'} \\ &= \frac{\partial}{\partial \lambda} m\big(W_j(\lambda); \tilde{A}^{-1}\theta\big)\Big|_{\lambda=A^{-1}\lambda'} \cdot A^{-1} \end{aligned}$$

The first two equalities hold from Assumption 2.d. The last equality holds from the chain rule and the differentiability of the mapping $\lambda \mapsto m\big(W_j(\lambda); \theta\big)$ at $A^{-1}\lambda' \in \mathcal{S}_\lambda$ for $\tilde{A}^{-1}\theta \in \Theta$ from Assumption 2.e.

Note that $R_j(\cdot, \cdot)$ in the remainder term is a $l \times \rho$ matrix; if λ_j is a scalar, R_j would be a first-order derivative of m with regard to λ_j, evaluated at some point between $A\lambda_j$ and $\hat{\lambda}_j$.

Let $\tilde{R}_j$ denote an arbitrary row of R_j. By applying the Cauchy-Schwarz inequality to the j-th cluster in the summation,

$$\left|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right| \leq \left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2 \left\|\hat{\lambda}_j - A\lambda_j\right\|_2.$$

By applying the Cauchy-Schwarz inequality again,

$$\begin{aligned}\left|\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right| &\leq \frac{1}{J}\sum_{j=1}^{J}\left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2 \left\|\left(\hat{\lambda}_j - A\lambda_j\right)\right\|_2 \\ &\leq \left(\frac{1}{J}\sum_{j=1}^{J}\left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2^{\,2}\right)^{\frac{1}{2}}\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}\right)^{\frac{1}{2}}.\end{aligned}$$

Then, by summing over the rows of R_j, we get

$$\left\|\frac{1}{J}\sum_{j=1}^{J}R_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right\|_2^{\,2} \leq \left(\frac{1}{J}\sum_{j=1}^{J}\left\|R_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_F^{\,2}\right)\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}\right).$$

$\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}$ is $\frac{1}{J}\cdot o_p(1)$ from the conditions of Theorem 1.

It remains to show that $\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}$ is $O_p(1)$. From the Taylor's theorem, the matrix R_j can be written as an integral as follows:

$$\begin{aligned}R_j &= \int_0^1 \frac{\partial}{\partial\lambda} m\left(W_j(\lambda);\theta\right)\Big|_{\lambda = A\lambda_j + t\left(\hat{\lambda}_j - A\lambda_j\right)} dt \\ &= \int_0^1 \frac{\partial}{\partial\lambda} m\left(W_j(\lambda);\tilde{A}^{-1}\theta\right)\Big|_{\lambda = \lambda_j + t\left(A^{-1}\hat{\lambda}_j - \lambda_j\right)} \cdot A^{-1} dt.\end{aligned}$$

Find that

$$\|R_j\|_F^{\,2} \leq l\rho\left(\rho \sup_{t\in[0,1]}\left\|m\big(W_j(\lambda);\tilde{A}^{-1}\theta\big)\Big|_{\lambda=\lambda_j+t\left(A^{-1}\hat{\lambda}_j-\lambda_j\right)}\right\|_\infty \cdot \left\|A^{-1}\right\|_\infty\right)^2$$

by finding the components of $\frac{\partial}{\partial\lambda}m$ and A^{-1} with the biggest absolute values. Then,

$$\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2} \leq \frac{l\rho^3}{J}\sum_{j=1}^{J}\sup_{t\in[0,1]}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\tilde{A}^{-1}\theta\right)\Big|_{\lambda=\lambda_j+t\left(A^{-1}\hat{\lambda}_j-\lambda_j\right)}\right\|_F^{\,2}\cdot\left\|A^{-1}\right\|_F^{\,2}.$$

Lastly, from Assumption 2.f, $\left\|A^{-1}\right\|_F$ is bounded with probability going to one. Thus, $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2 \leq \left\|A^{-1}\right\|_F\cdot\max_j\left\|\hat{\lambda}_j-A\lambda_j\right\|_2 \leq \eta$ holds with probability going to one, from the condition of Theorem 1. In addition, conditioning on the event that $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\eta$ and $\left\|A^{-1}\right\|_F\leq\tilde{M}$, we have

$$\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2} \leq \frac{\tilde{M}^2l\rho^3}{J}\sum_{j=1}^{J}\sup_{\|\lambda'-\lambda_j\|_2\leq\eta}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\theta\right)\Big|_{\lambda=\lambda'}\right\|_F^{\,2}.$$

From Assumption 2.e, the RHS of the inequality above is bounded in expectation by $M\tilde{M}^2l\rho^3$.

Consequently, we have that $\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}$ is $O_p(1)$: for any $\varepsilon>0$, find large enough J^* such that the probability that $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\frac{\eta}{\tilde{M}}$ and $\left\|A^{-1}\right\|_F\leq\tilde{M}$ holds is bigger than $1-\frac{\varepsilon}{3}$ and large enough M^* such that the probability of the RHS of the inequality above being bigger than M^* is smaller than $\frac{\varepsilon}{3}$. Then, for $J\geq J^*$,

$$\begin{aligned}
&\Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}\geq M^*\right\}\\
&\leq \Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}\geq M^*,\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\eta,\left\|A^{-1}\right\|_F\leq\tilde{M}\right\}+\frac{\varepsilon}{3}\\
&\leq \Pr\left\{\frac{\tilde{M}^2l\rho^3}{J}\sum_{j=1}^{J}\sup_{\|\lambda'-\lambda_j\|_2\leq\eta}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\theta\right)\Big|_{\lambda=\lambda'}\right\|_F^{\,2}\geq M^*\right\}+\frac{\varepsilon}{3}\leq\frac{2\varepsilon}{3}.
\end{aligned}$$

Note that the stochastic boundedness is uniform across $\theta\in\tilde{A}\Theta$ since the quantity in the last probability involves a supremum over Θ.

Having shown that the feasible objective function is close to the infeasible objective function, I now show that the consistency of the infeasible GMM estimator leads to the

consistency of the feasible GMM estimator. Find that

$$\begin{aligned}
\left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) \right\|_2 &= \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \hat{\theta}\right) \right\|_2 + \frac{1}{\sqrt{J}} \cdot o_p(1) \\
&\leq \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \tilde{A}\theta^0\right) \right\|_2 + \frac{1}{\sqrt{J}} \cdot o_p(1) \\
&= \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \tilde{A}\theta^0\right) \right\|_2 + \frac{1}{\sqrt{J}} \cdot o_p(1) \\
&= \left\| \mathbf{E}\left[m\left(W_j^*; \theta^0\right) \right] \right\|_2 + o_p(1) = o_p(1).
\end{aligned}$$

The inequality is from the definition of the GMM estimator. The first equality holds for a random object $\hat{\theta}$ since the stochastic boundedness of $\frac{1}{J}\sum_{j=1}^{J} \|R_j\|_F{}^2$ does not depend on the choice of θ. The second to the last equality is from Assumption 2.c-d. Again, from Assumption 2.c-d, we get

$$\begin{aligned}
&\left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) \right\|_2 - \left\| \mathbf{E}\left[m\left(W_j; \hat{\theta}\right) \right] \right\|_2 \\
&= \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \right\|_2 - \left\| \mathbf{E}\left[m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \right] \right\|_2 = o_p(1).
\end{aligned}$$

The first equality holds from Assumption 2.d and the second equality holds from Assumption 2.c. Then,

$$\left\| \mathbf{E}\left[m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \right] \right\|_2 = \left\| \mathbf{E}\left[m\left(W_j; \hat{\theta}\right) \right] \right\|_2 = \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) \right\|_2 + o_p(1) = o_p(1).$$

We get the consistency of $\tilde{A}^{-1}\hat{\theta}$ to θ^0 from Assumption 2.b and thus the consistency of $\hat{\theta}$ to

$\tilde{A}\theta^0$ from Assumption 2.f: for any $\varepsilon > 0$,

$$\begin{aligned}\Pr\left\{\left\|\hat{\theta} - \tilde{A}\theta^0\right\|_2 \geq \varepsilon\right\} &\leq \Pr\left\{\left\|\tilde{A}\right\|_F \cdot \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \varepsilon\right\} \\ &\leq \Pr\left\{\left\|\tilde{A}\right\|_F \cdot \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \varepsilon, \left\|\tilde{A}\right\|_F \leq \tilde{M}\right\} + \Pr\left\{\left\|\tilde{A}\right\|_F > \tilde{M}\right\} \\ &\leq \Pr\left\{\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \frac{\varepsilon}{\tilde{M}}\right\} + \Pr\left\{\left\|\tilde{A}\right\|_F > \tilde{M}\right\} = o(1).\end{aligned}$$

C.2 Theorem 2

Recall that

$$\begin{aligned}\left\|\frac{1}{J}\sum_{j=1}^{J} m\left(\widehat{W}_j;\theta\right) - \frac{1}{J}\sum_{j=1}^{J} m\left(W_j;\theta\right)\right\|_2 &= O_p(1)\cdot\frac{1}{\sqrt{J}}\cdot\left(\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^2\right)^{\frac{1}{2}} \\ &= O_p(1)o_p(1)\frac{1}{\sqrt{J}}\end{aligned}$$

from the proof of Theorem 1 and therefore

$$\begin{aligned}&\left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right)\right\|_2 \\ &\geq \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 - \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right) - \frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 \\ &= \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 + o_p(1).\end{aligned}$$

From the condition of Theorem 2, we get

$$\begin{aligned}o_p(1) = \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right)\right\|_2 &\geq \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 + o_p(1) \\ o_p(1) &= \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2.\end{aligned}$$

Step 1.

For asymptotic normality result, we need a stronger consistency result for $\tilde{A}^{-1}\hat{\theta}$ than Theorem 1. For that, let us apply the first-order Taylor's expansion to the objective function with regard to the parameter of interest θ^0:

$$\begin{aligned} o_p(1) &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \\ &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j^*; \theta^0\right) + \frac{1}{J} \sum_{j=1}^{J} R_{1j}\left(\tilde{A}^{-1}\hat{\theta}, \theta^0\right) \cdot \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right). \end{aligned}$$

We can apply the Taylor's expansion since Assumption 3.a assumes twice-differentiability of m.

R_{1j} is a $l \times k$ matrix for the first-order remainder term in the expansion. The remainder term coefficient matrix R_{1j} can be rewritten as

$$R_{1j}\left(\tilde{A}^{-1}\hat{\theta}, \theta^0\right) = \int_0^1 \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt.$$

Find that

$$\begin{aligned} &\left\| \frac{1}{J}\sum_{j=1}^{J} \int_0^1 \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt - \int_0^1 \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] dt \right\|_F \\ &= \left\| \int_0^1 \left(\frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] \right) dt \right\|_F \\ &\leq \sqrt{lk} \cdot \sup_{t\in[0,1]} \left\| \frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] \right\|_F \\ &= o_p(1). \end{aligned}$$

The first equality holds from Fubini's theorem since the integral and the summation are both defined with σ-finite measures on $\{1, \cdots, J\}$ and $[0, 1]$. The inequality holds from finding that any component of the $l \times k$ matrix $\frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\right]$ for a given $t \in [0, 1]$ and therefore its integral over $[0, 1]$ are bounded by the supremum in the Frobenius norm. The

second equality holds from Assumption 3.b. Lastly, find that

$$\left\| \int_0^1 \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} \right] dt - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] \right\|_F$$
$$\leq \sqrt{lk} \cdot \sup_{t\in[0,1]} \left\| \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} \right] - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] \right\|_F = o_p(1).$$

The inquality holds since any component of the $l\times k$ matrix $\mathbf{E}\left[\frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta^0=t(\tilde{A}^{-1}\hat{\theta}-\theta^0)}\right] dt$ for a given $t\in[0,1]$ is bounded by the supremum over the Frobenius norm. $\theta \mapsto \frac{\partial}{\partial\theta} m(W_j^*;\theta)$ is continuously differentiable from Assumption 3.a. From the Leibniz's rule, its expectation is also differentiable and thus continuous; the equality holds. $\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right)$ converges to a full rank matrix $\mathbf{E}\left[\frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta^0}\right]$.

Lastly, since $\frac{1}{\sqrt{J}}\sum_{j=1}^J m\left(W_j^*;\theta^0\right)$ is $O_p(1)$ from the CLT,

$$\left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right) \cdot \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) = O_p(1).$$

Therefore $\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) = O_p(1)$ by finding a small neighborhood around $\mathbf{E}\left[\frac{\partial}{\partial\theta} m\left(W_j^*;\theta\right)\big|_{\theta=\theta^0}\right]$ such that $\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right)$ is full rank and the Frobenius norm of its left inverse is bounded: for any M^* and M_R,

$$\begin{aligned}
&\Pr\left\{ \left\| \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) \right\|_2 \geq M^* \right\} \\
&\leq \Pr\left\{ \left\| \left(\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right) \right)^{-1} \frac{1}{\sqrt{J}} \sum_{j=1}^J \left(m\left(W_j^*;\tilde{A}^{-1}\hat{\theta}\right) - m\left(W_j^*;\theta^0\right) \right) \right\|_2 \geq M^* \right\} \\
&\quad + \Pr\left\{ \frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right) \text{ is not full rank} \right\} \\
&= \Pr\left\{ \left\| \left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right)^{-1} \right\|_F \cdot \left\| O_p(1) \right\|_F \geq M^* \right\} + o(1) \\
&\leq \Pr\left\{ \left\| O_p(1) \right\|_F \geq \frac{M^*}{M_R} \right\} + \Pr\left\{ \left\| \left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right)^{-1} \right\|_F > M_R \right\} + o(1).
\end{aligned}$$

The second probability in the last inequality goes to zero for large enough M_R from the

continuous mapping theorem since each component of the inverse matrix is a continuous function of the original matrix. Given some $\varepsilon > 0$, first choose large enough M_R so that the second probability in the last inequality is arbitrarily small for large J and then choose large enough M^* so that the first probability is arbitrarily small as well. Then, we can find some J^* such that $\Pr\left\{\left\|\sqrt{J}(\tilde{A}^{-1}\hat{\theta} - \theta)\right\|_F \leq M^*\right\} \leq \varepsilon$ for $J \geq J^*$.

Step 2.

Again, let $\tilde{m}$ denote an arbitrary component of m. From the component-wise second-order Taylor's expansion,

$$\begin{aligned} o_p(1) &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \tilde{m}\left(W_j; \hat{\theta}\right) \\ &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \tilde{m}\left(W_j; \tilde{A}\theta^0\right) + \frac{1}{J} \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \tilde{m}\left(W_j; \theta\right)\big|_{\theta = \tilde{A}\theta^0} \cdot \sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) \\ &\qquad + \left(\hat{\theta} - \tilde{A}\theta^0\right)^{\intercal} \cdot \frac{1}{J} \sum_{j=1}^{J} \tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right) \cdot \sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) \end{aligned}$$

where $\tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right)$ is a $k \times k$ matrix for the second-order remainder term. Find that

$$\tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right) = \left(\tilde{A}^{\intercal}\right)^{-1} \cdot \int_0^1 (1-t) \frac{\partial^2}{\partial \theta \partial \theta^{\intercal}} \tilde{m}\left(W_j^*; \theta\right)\Big|_{\theta = \theta^0 + t\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)} dt \cdot \tilde{A}^{-1}$$

from the Taylor's theorem and by applying the chain rule. For any $M^{**} > 0$,

$$\begin{aligned} &\Pr\left\{\left\|\frac{1}{J}\sum_{j=1}^{J}\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt\right\|_F \geq M^{**}\right\} \\ &\leq \Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\left\|\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt\right\|_F \geq M^{**}, \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \leq \eta\right\} \\ &\quad + \Pr\left\{\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 > \eta\right\} \\ &\leq \Pr\left\{\frac{k}{J}\sum_{j=1}^{J}\sup_{\|\theta' - \theta^0\|_2 \leq \eta}\left\|\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta'}\right\|_F \geq M^{**}\right\} + o(1). \end{aligned}$$

The last inequality holds from the fact that any component of $\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}$ for a given $t \in [0,1]$ and therefore any component of the integral $\int(1-t)\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}dt$ are bounded by the supremum over the Frobenius norm, when $\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \leq \eta$. Given some $\varepsilon > 0$, we can find large enough M^{**} and J^{**} such that

$$\Pr\left\{\left\|\frac{1}{J}\sum_{j=1}^{J}\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}dt\right\|_F \geq M^{**}\right\} \leq \varepsilon$$

for any $J \geq J^{**}$, from Assumption 3.a. $\tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}$ is $O_p(1)$.

Lastly, since $\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right) = O_p(1)$, the second-order remainder term in the second-order approximation is $o_p(1)$:

$$\begin{aligned}
&\left|\left(\hat{\theta} - \tilde{A}\theta^0\right)^\intercal \cdot \frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\cdot\sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right)\right| \\
&\left|\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)^\intercal \cdot \tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}\cdot\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)\right| \\
&\leq \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \cdot \left\|\tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}\right\|_F \cdot \left\|\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)\right\|_2 = o_p(1).
\end{aligned}$$

Thus,

$$\sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{\partial}{\partial\theta}m\left(W_j;\tilde{A}\theta^0\right)\right)^{-1}\frac{1}{\sqrt{J}}\sum_{j=1}^{J}m\left(W_j;\tilde{A}\theta^0\right) + o_p(1).$$

C.3 Proposition 1

For the convenience of notation, let $\lambda_j \in \{1,\cdots,\rho\}$ for true latent factor λ_j as well.

Step 1

From the within-cluster iidness,

$$\begin{aligned}
&\mathbf{E}\left[N_j\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2\right] \\
&= \mathbf{E}\left[N_j\mathbf{E}\left[\int\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{1}\{X_{ij}\le x\} - \Big(G(\lambda_j)\Big)(x)\right)^2 w(x)dx\,\middle|\,N_j, Z_j, \lambda_j\right]\right] \\
&= \mathbf{E}\left[\int \mathrm{Var}\left(\mathbf{1}\{X_{ij}\le x\}\middle|N_j, Z_j, \lambda_j\right) w(x)dx\right] \le \frac{1}{4}.
\end{aligned}$$

The second equality holds from exchanging the order of integration and expectation. Thus,

$$\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2 = O_p\left(\frac{1}{N_{\min}}\right)$$

Step 2

Let us connect $\hat{G}(1), \cdots, \hat{G}(\rho)$ to $G(1), \cdots, G(\rho)$. Define $\sigma(r)$ such that

$$\sigma(r) = \arg\min_{\tilde{r}}\left\|\hat{G}(\tilde{r}) - G(r)\right\|_{w,2}.$$

We can think of $\sigma(r)$ as the 'oracle' estimate that cluster j would have been assigned to, when $\mathbf{F}_j = G(r)$ is directly observed and $\hat{G}(1), \cdots, \hat{G}(\rho)$ are given. Then,

$$\begin{aligned}
&\left\|\hat{G}(\sigma(r)) - G(r)\right\|_{w,2}^2 \\
&= \frac{J}{\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}}\cdot\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{G}(\sigma(r)) - G(\lambda_j)\right\|_{w,2}^2\mathbf{1}\{\lambda_j = r\} \\
&\le \frac{J}{\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}}\cdot\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{G}(\hat{\lambda}_j) - G(\lambda_j)\right\|_{w,2}^2 \\
&\le \frac{2J}{\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}}\cdot\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{G}(\hat{\lambda}_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2 + \frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2\right) \\
&\le \frac{4J}{\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}}\cdot\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2.
\end{aligned}$$

The last inequality holds since $\sum_{j=1}^{J}\left\|\hat{G}(\hat{\lambda}_j)-\hat{\mathbf{F}}_j\right\|_{w,2}^{2} \leq \sum_{j=1}^{J}\left\|G(\lambda_j)-\hat{\mathbf{F}}_j\right\|_{w,2}^{2}$ from the definition of $\hat{G}$ and $\hat{\lambda}$. From Assumption 4.a, $\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}/J \xrightarrow{p} \mu(r) > 0$ as $J \to \infty$. Thus,

$$\left\|\hat{G}(\sigma(r))-G(r)\right\|_{w,2}^{2} \xrightarrow{p} 0$$

as $J \to \infty$ from Assumption 4.c and Step 1.

Note that for any $r' \neq r$,

$$\left\|\hat{G}(\sigma(r))-G(r')\right\|_{w,2}^{2} \geq \frac{1}{2}\|G(r)-G(r')\|_{w,2}^{2} - \left\|\hat{G}\left(\sigma\left(r\right)\right)-G(r)\right\|_{w,2}^{2} = \frac{1}{2}c(r,r') + o_p(1)$$

as $J \to \infty$ from the same argument from above and Assumption 4.c.

Find that σ is bijective with probability converging to one: with $\varepsilon^* = \min_{k\neq k'}\frac{1}{8}c(r,r')$,

$$\begin{aligned}
\Pr\{\sigma \text{ is not bijective.}\} &\leq \sum_{r\neq r'}\Pr\{\sigma(r)=\sigma(r')\} \\
&\leq \sum_{r\neq r'}\Pr\left\{\left\|\hat{G}(\sigma(r))-\hat{G}(\sigma(r'))\right\|_{w,2}^{2} < \varepsilon^*\right\} \\
&\leq \sum_{r\neq r'}\Pr\left\{\frac{1}{2}\left\|\hat{G}(\sigma(r))-G(r')\right\|_{w,2}^{2} - \left\|\hat{G}(\sigma(r'))-G(r')\right\|_{w,2}^{2} < \varepsilon^*\right\} \\
&\leq \sum_{r\neq r'}\Pr\left\{\frac{1}{4}\|G(r)-G(r')\|_{w,2}^{2} + o_p(1) < \varepsilon^*\right\} \to 0
\end{aligned}$$

as $J \to \infty$. When σ is bijective, relabel $\hat{G}(1),\cdots,\hat{G}(\rho)$ so that $\sigma(r) = r$.

Step 3

Let us put a bound on $\Pr\left\{\hat{\lambda}_j \neq \sigma(\lambda_j)\right\}$, the probability of estimated group being different from 'oracle' group; this means that there is at least one $r \neq \sigma(\lambda_j)$ such that that $\hat{\mathbf{F}}_j$ is closer to $\hat{G}(r)$ than $\hat{G}(\sigma(\lambda_j))$:

$$\Pr\left\{\hat{\lambda}_j \neq \sigma(\lambda_j)\right\} \leq \Pr\left\{\exists\ r \text{ s.t. } \left\|\hat{G}(r)-\hat{\mathbf{F}}_j\right\|_{w,2} \leq \left\|\hat{G}(\sigma(\lambda_j))-\hat{\mathbf{F}}_j\right\|_{w,2}\right\}.$$

The discussion on the probability is much more convenient when σ is bijective and $\hat{G}(\sigma(r))$ is close to $G(r)$ for every r. Thus, let us instead focus on the joint probability:

$$\Pr\left\{\hat{\lambda}_j \neq \lambda_j, \sum_{r=1}^{\rho} \left\|\hat{G}(r) - G(r)\right\|_{w,2}^2 < \varepsilon, \text{ and } \sigma \text{ is bijective.}\right\}.$$

Note that in the probability, $\sigma(r)$ is replaced with r and $\sigma(\lambda_j)$ with λ_j since we are conditioning on the event that σ is bijective: relabeling is applied and $\hat{G}(r)$ can be thought of as a direct estimate for $G(r)$. For notational brevity, let A_ε denote the event of σ being bijective and $\sum_{r=1}^{\rho} \left\|\hat{G}(r) - G(r)\right\|_{w,2}^2 < \varepsilon$. From Step 2, we have that $\Pr\{A_\varepsilon\} \to 1$ as $J \to \infty$ for any $\varepsilon > 0$.

Then, with $c^* = \min_{r \neq r'} c(r, r') > 0$,

$$\begin{aligned}
\Pr\left\{\hat{\lambda}_j \neq \lambda_j, A_\varepsilon\right\} &\leq \Pr\left\{\exists\, r \neq \lambda_j \text{ s.t. } \left\|\hat{G}(r) - \hat{\mathbf{F}}_j\right\|_{w,2} \leq \left\|\hat{G}(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}, A_\varepsilon\right\} \\
&\leq \Pr\left\{\exists\, r \neq \lambda_j \text{ s.t. } \frac{1}{2}\left\|\hat{G}(r) - G(\lambda_j)\right\|_{w,2}^2 - \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2 \right. \\
&\qquad \left. \leq 2\left\|\hat{G}(\lambda_j) - G(\lambda_j)\right\|_{w,2}^2 + 2\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\} \\
&\leq \Pr\left\{\exists\, r \neq \lambda_j \text{ s.t. } \frac{1}{4}\|G(r) - G(\lambda_j)\|_{w,2}^2 - \frac{1}{2}\left\|\hat{G}(r) - G(r)\right\|_{w,2}^2 \right. \\
&\qquad \left. \leq 2\left\|\hat{G}(\lambda_j) - G(\lambda_j)\right\|_{w,2}^2 + 3\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\} \\
&\leq \Pr\left\{\exists\, r \neq \lambda_j \text{ s.t. } \frac{1}{4}\|G(r) - G(\lambda_j)\|_{w,2}^2 \right. \\
&\qquad \left. \leq \frac{5}{2}\sum_{r'=1}^{\rho} \left\|\hat{G}(r') - G(r')\right\|_{w,2}^2 + 3\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\} \\
&\leq \Pr\left\{\frac{c^*}{4} \leq \frac{5}{2}\sum_{r=1}^{\rho} \left\|\hat{G}(r) - G(r)\right\|_{w,2}^2 + 3\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\} \\
&\leq \Pr\left\{\frac{c^*}{12} - \frac{5}{6}\varepsilon \leq \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2\right\}
\end{aligned}$$

The last inequality is from the construction of the event A_ε. In the last inequality A_ε can be dropped since the probability does not require σ being bijective to be well-defined. Set

$\varepsilon^* = \frac{c^*}{20}$ so that $\frac{c^*}{12} - \frac{5}{6}\varepsilon^* = \frac{c^*}{24} > 0$.

By repeating the expansion for every j,

$$\begin{aligned}\Pr\left\{\exists\ j \text{ s.t. } \hat{\lambda}_j \neq \lambda_j\right\} &\leq \Pr\left\{\exists\ j \text{ s.t. } \hat{\lambda}_j \neq \lambda_j, A_{\varepsilon^*}\right\} + \Pr\left\{{A_{\varepsilon^*}}^c\right\} \\ &\leq \sum_{j=1}^{J} \Pr\left\{\frac{c^*}{24} \leq \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2}\right\} + \Pr\left\{{A_{\varepsilon^*}}^c\right\}.\end{aligned}$$

We already know $\Pr\left\{{A_{\varepsilon^*}}^c\right\} = o(1)$ as $J \to \infty$. It remains to show that the first quantity in the RHS of the inequality is $o(J/N_{\min}^{\nu})$ for any $\nu > 0$. Let ε^{**} denote $\frac{c^*}{24}$. Choose an arbitrary $\nu > 0$. From the within-cluster iidness,

$$\begin{aligned}\Pr\left\{\varepsilon^{**} \leq \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^{\ 2}\right\} &\leq \mathbf{E}\left[\Pr\left\{\varepsilon^{**} \leq \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{\infty}^{\ 2} \middle| N_j, Z_j, \lambda_j\right\}\right] \\ &\leq \mathbf{E}\left[C^*(N_j+1)\exp\left(-2N_j\varepsilon^{**}\right)\right]\end{aligned}$$

with some constant $C^* > 0$, by taking the least favorable case over $\lambda_j = 1, \cdots, \rho$ and applying the Dvoretzky–Kiefer–Wolfowitz inequality. Thus, for any $\nu > 0$,

$$\begin{aligned}\frac{{N_{\min}}^{\nu}}{J}\sum_{j=1}^{J}\Pr\left\{\varepsilon^{**} \leq \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2}\right\} &= {N_{\min}}^{\nu}\mathbf{E}\left[C^*(N_j+1)\exp\left(-2N_j\varepsilon^{**}\right)\right] \\ &\leq \frac{C^*{N_{\min}}^{\nu}(N_{\min}+1)}{\exp\left(2N_{\min}\varepsilon^{**}\right)} = o(1)\end{aligned}$$

as $J \to \infty$. The inequality holds for large n; $n \mapsto (n+1)\exp(-2n\varepsilon^{**})$ is decreasing in n for large n.

C.4 Proposition 2

Step 1. Firstly, let us discuss the rotation on the latent factor. For notational simplicity, let

$$V = \begin{pmatrix} \int_{\mathbb{R}} g_1(x)^2 w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x) g_1(x) w(x)dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}} g_1(x) g_\rho(x) w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)^2 w(x)dx \end{pmatrix},$$
$$\Lambda = \begin{pmatrix} \lambda_1 & \cdots & \lambda_J \end{pmatrix}.$$

Suppose $rank(M) = rank(\Lambda^\intercal V \Lambda) = \rho$ and consider an eigen-decomposition for M with orthonormal eigenvectors, using the ρ positive eigenvalues: $V_1, \cdots, V_\rho$. Let Q be a $J \times \rho$ matrix with the orthonormal eigenvectors as columns and let $\tilde{\Lambda} = \sqrt{J} Q^\intercal$. Then, $\frac{1}{J}\tilde{\Lambda}\tilde{\Lambda}^\intercal = Q^\intercal Q = I_\rho$ and

$$\Lambda^\intercal V \Lambda = M = Q \text{diag}\left(V_1, \cdots, V_\rho\right) Q^\intercal = \tilde{\Lambda}^\intercal \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right) \tilde{\Lambda}.$$

Let

$$A^\intercal = V \left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right) \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1},$$

we have

$$\begin{aligned} \Lambda^\intercal A^\intercal &= \Lambda^\intercal V \left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right) \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1} \\ &= \tilde{\Lambda}^\intercal \text{diag}\left(\frac{\nu_1}{J}, \cdots, \frac{\nu_\rho}{J}\right) \frac{1}{J}\tilde{\Lambda}\tilde{\Lambda}^\intercal \text{diag}\left(\frac{\nu_1}{J}, \cdots, \frac{\nu_\rho}{J}\right)^{-1} = \tilde{\Lambda}^\intercal. \end{aligned}$$

We have a rotation between the matrix of the true latent factor Λ and the matrix of (rescaled) eigenvectors $\tilde{\Lambda}$.

The rotation matrix A in Proposition 2 satisfies Assumption 2.f:

$$\left\|A^{-1}\right\|_F = \left\| diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1} V^{-1}\right\|_F .$$

Find that

$$\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal \cdot diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right) \cdot \frac{1}{J}\tilde{\Lambda}\Lambda^\intercal = \frac{1}{J}\Lambda\Lambda^\intercal \cdot V \cdot \frac{1}{J}\Lambda\Lambda^\intercal$$

$$\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1} = \left(\frac{1}{J}\Lambda\Lambda^\intercal \cdot V \cdot \frac{1}{J}\Lambda\Lambda^\intercal\right)^{-1} \cdot \frac{1}{J}\Lambda\tilde{\Lambda}^\intercal \cdot diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)$$

and since the Frobenius norm is invariant to a unitary operation

$$\left\|\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right\|_F \leq \frac{1}{\sqrt{J}}\left\|\Lambda\right\|_F = \left(\frac{1}{J}\sum_{j=1}^{J}\left\|\lambda_j\right\|_2{}^2\right)^{\frac{1}{2}} = O_p(1).$$

$\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1}$ is also $O_p(1)$, satisfying Assumtion 2.f.

Step 2. Now, we show the estimate $\widehat{M}$ is close to the true matrix M. The following convergence rate on $\left\|\widehat{M} - M\right\|_F$ is a multivariate extension of Proposition 1 and Theorem 1 of Kneip and Utikal (2001).

$$\left\|\widehat{M} - M\right\|_F = O_p\left(\frac{J}{\sqrt{\min_j N_j}}\right).$$

To avoid notational complexity, I will use subscript λ to indicate that the expectation is conditioning on (N_j, Z_j, λ_j). Find that

$$\mathbf{E}_\lambda\left[\left(\widehat{M}_{jk} - M_{jk}\right)^2\right] = \mathrm{Var}_\lambda\left(\widehat{M}_{jk}\right) + \left(\mathbf{E}_\lambda\left[\widehat{M}_{jk}\right] - M_{jk}\right)^2$$

From the kernel estimation,

$$\begin{aligned}
&\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)\right] \\
&= \int_{\mathbb{R}^p}\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-x')\right)\mathbf{f}_j(x')dx' \\
&= \int_{\mathbb{R}^p}K(t)\mathbf{f}_j\big(x-H^{\frac{1}{2}}t\big)dt \qquad \text{by letting } x'=x-H^{\frac{1}{2}}t \\
&= \int_{\mathbb{R}^p}K(t)\left(\mathbf{f}_j(x)-\mathbf{f}_j^{(1)}(x)^\intercal H^{\frac{1}{2}}t+t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}t\right)dt \\
&= \mathbf{f}_j(x)+\int_{\mathbb{R}^p}K(t)\cdot t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}tdt
\end{aligned}$$

for some $\tilde{x}$ depending on x and $x-H^{\frac{1}{2}}t$. The second equality holds from the differentiability in Assumption 5.a and the last equality holds from the conditions i. and ii. given in Proposition 2. Lastly, from the condition iii. in Proposition 2 and the boundedness from Assumption 5.a,

$$\left|\int_{\mathbb{R}^p}K(t)\cdot t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}tdt\right| \le \frac{p^2C}{2}\cdot\max_x\left\|H^{\frac{1}{2}\intercal}\mathbf{f}_j^{(2)}(x)H^{\frac{1}{2}}\right\|_F \le \frac{p^3C^2}{2}\cdot\left\|H^{\frac{1}{2}}\right\|_F^{\,2}.$$

The first inequality is from the condition iii. and the second inequality is from Assumption 5.a. Then,

$$\begin{aligned}
&\left|\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)\right]\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ik})\right)\right]-\mathbf{f}_j(x)\mathbf{f}_k(x)\right| \\
&\le C_1\left\|H^{\frac{1}{2}}\right\|_F^{\,2}
\end{aligned}$$

with some $C_1 > 0$ that does not depend on λ_j or H. By extending this,

$$\begin{aligned}
&\left|\mathbf{E}_\lambda\left[\widehat{M}_{jk} - M_{jk}\right]\right| \\
&\leq \int_{\mathbb{R}^p} \left|\mathbf{E}_\lambda\left[\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{1j}\right)\right)}{det(H)^{\frac{1}{2}}}\right]\mathbf{E}_\lambda\left[\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{2k}\right)\right)}{det(H)^{\frac{1}{2}}}\right] - \mathbf{f}_j(x)\mathbf{f}_k(x)\right| w(x)dx \\
&\leq C_1 \left\|H^{\frac{1}{2}}\right\|_F{}^2.
\end{aligned}$$

$\mathbf{E}_\lambda$ and $\int_{\mathbb{R}}$ are interchangeable from Fubini's theorem. For $\mathrm{Var}_\lambda(\widehat{M}_{jk})$, find that

$$\begin{aligned}
\mathrm{Var}_\lambda\left(\widehat{M}_{jk}\right) = &\frac{\sum_{i=1}^{N_j}\sum_{i'=1}^{N_k}}{{N_j}^2 {N_k}^2}\left(\mathrm{Var}_\lambda\left(A_{ii'}\right) + \sum_{l\neq i}\mathrm{Cov}_\lambda\left(A_{ii'}, A_{li'}\right) + \sum_{l\neq i'}\mathrm{Cov}_\lambda\left(A_{ii'}, A_{il}\right)\right)\mathbf{1}\{j \neq k\} \\
&+ \frac{\sum_{i=1}^{N_j}\sum_{i'=i}}{{N_j}^2\left(N_j - 1\right)^2}\left(\mathrm{Var}_\lambda\left(A_{ii'}\right) + \sum_{l\neq i,i'}\mathrm{Cov}_\lambda\left(A_{ii'}, A_{li'}\right) + \sum_{l\neq i,i'}\mathrm{Cov}_\lambda\left(A_{ii'}, A_{il}\right)\right)\mathbf{1}\{j = k\}
\end{aligned}$$

where

$$A_{ii'} = \int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}\left(x - X_{ij}\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{i'k}\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx.$$

We have that for some $l \neq i'$,

$$\begin{aligned}
&\mathbf{E}_\lambda\left[{A_{ii'}}^2\right] \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}\left(x - x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x - x''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right)^2 \mathbf{f}_j(x')\mathbf{f}_k(x'')dx'dx'' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} K\left(t\right)\frac{K\big(t + H^{-\frac{1}{2}}(x' - x'')\big)}{det(H)^{\frac{1}{2}}}w\big(x' + H^{\frac{1}{2}}t\big)dt\right)^2 \mathbf{f}_j(x')\mathbf{f}_k(x'')dx'dx'' \\
&= \frac{1}{det(H)^{\frac{1}{2}}}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} K\left(t\right)K\left(t + s\right)w\big(x'' + H^{\frac{1}{2}}(t + s)\big)dt\right)^2 \mathbf{f}_j\big(x'' + H^{\frac{1}{2}}s\big)\mathbf{f}_k(x'')dsdx''
\end{aligned}$$

by letting $x = x' + H^{\frac{1}{2}}t$ and $x' = x'' + H^{\frac{1}{2}}s$ and

$$
\begin{aligned}
&\mathbf{E}_\lambda\left[A_{ii'}A_{il}\right] \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x-x''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx''' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)\frac{K\left(t+H^{-\frac{1}{2}}(x'-x'')\right)}{det(H)^{\frac{1}{2}}}w\big(x'+H^{\frac{1}{2}}t\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)\frac{K\left(t+H^{-\frac{1}{2}}(x'-x''')\right)}{det(H)^{\frac{1}{2}}}w\big(x'+H^{\frac{1}{2}}t\big)dt\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx''' \\
&= \frac{1}{det(H)^{\frac{1}{2}}}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s\right)w\big(x''+H^{\frac{1}{2}}(t+s)\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s+H^{-\frac{1}{2}}(x''-x''')\right)w\big(x''+H^{\frac{1}{2}}(t+s)\big)dt\right) \\
&\quad\cdot\mathbf{f}_j\big(x''+H^{\frac{1}{2}}s\big)\mathbf{f}_k(x'')\mathbf{f}_k(x''')dsdx''dx''' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s\right)w\big(x'''+H^{\frac{1}{2}}(t+s+u)\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s+u\right)w\big(x'''+H^{\frac{1}{2}}(t+s+u)\big)dt\right) \\
&\quad\cdot\mathbf{f}_j\big(x''+H^{\frac{1}{2}}s\big)\mathbf{f}_k\big(x'''+H^{\frac{1}{2}}u\big)\mathbf{f}_k(x''')dsdudx'''
\end{aligned}
$$

by letting $x = x' + H^{\frac{1}{2}}t$, $x' = x'' + H^{\frac{1}{2}}s$ and $x'' = x''' + H^{\frac{1}{2}}u$. Thus, with some constant $C_2 > 0$ that does not depend on λ_j or λ_k, $\mathrm{Var}_\lambda(A_{ii'}) \leq C_2/det(H)^{\frac{1}{2}}$ and $|\mathrm{Cov}_\lambda(A_{ii'}, A_{il})| \leq C_2$ and

$$
\mathrm{Var}_\lambda\left(\hat{M}_{jk}\right) \leq \begin{cases} C_2\left(\dfrac{1}{N_jN_kdet(H)^{\frac{1}{2}}} + \dfrac{1}{N_j} + \dfrac{1}{N_k}\right), & \text{if } j \neq k \\ C_2\left(\dfrac{1}{N_j(N_j-1)det(H)^{\frac{1}{2}}} + \dfrac{2}{N_j-1}\right), & \text{if } j = k \end{cases}
$$

Since $\min_j N_j det(H)^{\frac{1}{2}} \to \infty$ and $\min_j N_j \left\|H^{\frac{1}{2}}\right\|_F^4 = O(1)$ as $J \to \infty$, we have

$$\sum_{j=1}^{J}\sum_{k=1}^{J} \mathbf{E}_\lambda \left[\left(\widehat{M}_{jk} - M_{jk}\right)^2\right] = O\left(\frac{J^2}{\min_j N_j}\right)$$

$$\left\|\widehat{M} - M\right\|_F = \left(\sum_{j=1}^{J}\sum_{k=1}^{J}\left(\widehat{M}_{jk} - M_{jk}\right)^2\right)^{\frac{1}{2}} = O_p\left(\frac{J}{\sqrt{\min_j N_j}}\right)$$

Step 3. Lastly, given the rate on $\left\|\widehat{M} - M\right\|_F$, the convergence rate on $\left\|\tilde{\Lambda} - \widehat{\Lambda}\right\|_F$ is obtained by applying Lemma A.1.b of Kneip and Utikal (2001), as in Theorem 1.b of Kneip and Utikal (2001).

Firstly, let $\hat{V}_r$ denote the r-the largest eigenvalue of $\widehat{M}$; $\hat{V}_r$ is an estimate of V_r, as defined in Assumption 5. Note that $V_r = 0$ for $\rho < r \leq J$. Also, let $\hat{q}_r$ denote the (orthonormal) eigenvector of $\widehat{M}$ associated with the r-th eigenvalue and similarly for q_r. Recall that

$$\widehat{\Lambda} = \sqrt{J}\hat{Q}^\intercal = \sqrt{J}\begin{pmatrix}\hat{q}_1 & \cdots & \hat{q}_\rho\end{pmatrix}^\intercal$$

$$\tilde{\Lambda} = \sqrt{J}Q^\intercal = \sqrt{J}\begin{pmatrix}q_1 & \cdots & q_\rho\end{pmatrix}^\intercal$$

$$I_J = \begin{pmatrix}q_1 & \cdots & q_J\end{pmatrix}\begin{pmatrix}q_1{}^\intercal \\ \vdots \\ q_J{}^\intercal\end{pmatrix} = \sum_{r=1}^{J} q_r q_r{}^\intercal$$

For some $r \leq \rho$,

$$\hat{q}_r = \left(q_r q_r{}^\intercal + \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal\right)\hat{q}_r = \left(q_r{}^\intercal \hat{q}_r\right) q_r + \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r.$$

Since $\hat{q}_r^\intercal \hat{q}_r = {q_r}^\intercal q_r = 1$, we have $1 = ({q_r}^\intercal \hat{q}_r)^2 + \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$. Thus,

$$
\begin{aligned}
{q_r}^\intercal \hat{q}_r &= \pm \left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r\right)^{\frac{1}{2}}, \\
\hat{q}_r - q_r &= \left(\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r\right)^{\frac{1}{2}} - 1\right) q_r + \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r.
\end{aligned}
$$

The second equality holds by changing signs of $\hat{q}_r$ and q_r so that ${q_r}^\intercal \hat{q}_r > 0$. Note that RHS will be zero when $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r = 0$ and $\sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$ is a zero vector.

Firstly, let us find a bound on $\sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$. Note that

$$
\begin{aligned}
\left(M - V_r I_J\right) \hat{q}_r &= \left(\widehat{M} - \left(\widehat{M} - M\right) - V_r I_J\right) \hat{q}_r \\
&= \left(\hat{V}_r - V_r\right) \hat{q}_r - \left(\widehat{M} - M\right) \hat{q}_r
\end{aligned}
$$

since $\hat{V}_r$ is the corresponding eigenvalue of $\widehat{M}$ for $\hat{q}_r$. Let $S_r = \sum_{r' \neq r} \frac{1}{V_{r'} - V_r} q_{r'} {q_{r'}}^\intercal$. S_r is well-defined from Assumption 5.b. By multiplying S_r to the equality above, we get

$$
\begin{aligned}
S_r \left(\left(\hat{V}_r - V_r\right) \hat{q}_r - \left(\widehat{M} - M\right) \hat{q}_r\right) &= S_r \left(M - V_r I_j\right) \hat{q}_r \\
&= S_r \left(\sum_{r'=1}^{\rho} V_{r'} q_{r'} {q_{r'}}^\intercal - V_r I_j\right) \hat{q}_r \\
&= \left(\sum_{r' \neq r} \frac{V_{r'}}{V_{r'} - V_r} q_{r'} {q'_r}^\intercal - \sum_{r' \neq r} \frac{V_r}{V_{r'} - V_r} q_{r'} {q_{r'}}^\intercal\right) \hat{q}_r \\
&= \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r.
\end{aligned}
$$

Find that $\left|\hat{V}_r - V_r\right| \leq \left\|\widehat{M} - M\right\|_{Ind,2} \leq \|\widehat{M} - M\|_F$ (see Chapter 8 Theorem 9 of Bellman

(1997)). $\|\cdot\|_{Ind,2}$ denotes the matrix norm induced by the vector norm $\|\cdot\|_2$. Also,

$$\begin{aligned}
\|S_r\|_{Ind,2} &= \Big\| \sum_{r' \neq r} \frac{1}{V_{r'} - V_r} q_{r'} q_{r'}{}^\intercal \Big\|_{Ind,2} \\
&= \sup_v \Big\| \sum_{r' \neq r} \frac{1}{V_{r'} - V_r} q_{r'} q_{r'}{}^\intercal v \Big\|_2 \qquad \text{s.t. } v = \sum_{r'=1}^{J} c_{r'} q_{r'} \text{ and } |v^\intercal v| = \Big| \sum_{r'} {c_{r'}}^2 \Big| \leq 1 \\
&= \sup_{c_1, \cdots, c_J} \Big\| \sum_{r' \neq r} \frac{c_{r'}}{V_{r'} - V_r} q_{r'} \Big\|_2 \qquad \text{s.t. } \Big| \sum_{r'} {c_{r'}}^2 \Big| \leq 1 \\
&= \sup_{c_1, \cdots, c_J} \left(\sum_{r' \neq r} \left(\frac{c_{r'}}{V_{r'} - V_r} \right)^2 \right)^{\frac{1}{2}} \qquad \text{s.t. } \Big| \sum_{r'} {c_{r'}}^2 \Big| \leq 1 \\
&\leq \frac{1}{\min_{r' \neq r} |V_{r'} - V_r|}.
\end{aligned}$$

Using the two inequalities, we get

$$\begin{aligned}
\Big\| \sum_{r' \neq r} q_{r'} q_r{}^\intercal \hat{q}_r \Big\|_2 &\leq \left| \hat{V}_r - V_r \right| \|S_r \hat{q}_r\|_2 + \left\| S_r \left(\widehat{M} - M \right) \hat{q}_r \right\|_2 \\
&\leq \left\| \widehat{M} - M \right\|_F \|S_r\|_{Ind,2} \|\hat{q}_r\|_2 + \|S_r\|_{Ind,2} \left\| \widehat{M} - M \right\|_{Ind,2} \|\hat{q}_r\|_2 \\
&\leq \frac{2 \|\widehat{M} - M\|_F}{\min_{r' \neq r} |V_{r'} - V_r|} \\
&= \frac{1}{J} O_p \left(\frac{J}{\sqrt{\min_j N_j}} \right) = O_p \left(\frac{1}{\sqrt{\min_j N_j}} \right).
\end{aligned}$$

The last equality holds from Assumption 5.b: $\frac{\min_{r' \neq r} |V_{r'} - V_r|}{J}$ converges to a nonzero constant in probability. $\sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ converges to a zero vector, when $\min_j N_j$ goes to infinity.

Secondly, let us put a bound on $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ to show that $\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r \right)^{\frac{1}{2}}$ converges to one. The convergence of $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ to zero directly follows from the

convergence above:

$$\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r = \left(\sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r$$
$$= \| \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \|_2^2 = O_p \left(\frac{1}{\min_j N_j} \right).$$

Note that for $x \in [0,1]$, $|(1-x)^{\frac{1}{2}} - 1| = 1 - (1-x)^{\frac{1}{2}} \leq x$. Thus,

$$\left\| \left(\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^{\frac{1}{2}} - 1 \right) q_r \right\|_2 \leq \left| \left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^{\frac{1}{2}} - 1 \right|$$
$$\leq \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r = O_p \left(\frac{1}{\min_j N_j} \right)$$

By combining the two bounds, we have

$$\|\hat{q}_r - q_r\|_2 = O_p \left(\frac{1}{\sqrt{\min_j N_j}} \right)$$

for $r \leq \rho$, by some sign change on $\hat{q}_r$. Accordingly,

$$\left\| \hat{\Lambda} - \tilde{\Lambda} \right\|_F = \left(\sum_{r=1}^{\rho} J \|\hat{q}_r - q_r\|_F^2 \right)^{\frac{1}{2}} = O_p \left(\frac{\sqrt{J}}{\sqrt{\min_j N_j}} \right).$$

References

Allegretto, Sylvia A, Arindrajit Dube, and Michael Reich, "Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data," *Industrial Relations: A Journal of Economy and Society*, 2011, *50* (2), 205–240.

Allegretto, Sylvia, Arindrajit Dube, Michael Reich, and Ben Zipperer, "Credible research designs for minimum wage studies: A response to Neumark, Salas, and Wascher," *ILR Review*, 2017, *70* (3), 559–592.

Bellman, Richard, *Introduction to matrix analysis*, SIAM, 1997.

Card, David and Alan B Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 1994, *84* (4), 772–793.

Dube, Arindrajit, T William Lester, and Michael Reich, "Minimum wage effects across state borders: Estimates using contiguous counties," *The review of economics and statistics*, 2010, *92* (4), 945–964.

Kneip, Alois and Klaus J Utikal, "Inference for density families using functional principal component analysis," *Journal of the American Statistical Association*, 2001, *96* (454), 519–542.

Neumark, David and Peter Shirley, "Myth or measurement: What does the new minimum wage research say about minimum wages and job loss in the United States?," *Industrial Relations: A Journal of Economy and Society*, 2022, *61* (4), 384–417.

Neumark, David, JM Ian Salas, and William Wascher, "Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater?," *Ilr Review*, 2014, *67* (3_suppl), 608–648.