Clustered Treatment in Multilevel Models[*]

Myungkou Shin[†]

June 21st, 2024

Click here for the latest version.

Abstract

I develop a multilevel model for empirical contexts where each individual belongs to a cluster and clusters are large. When given a possibly endogeneous cluster-level explanatory variable, we cannot model the cluster-level heterogeneity to be fully flexible. To put restrictions, I model the cluster-level heterogeneity with finite-dimensional cluster-level latent factors, which has a one-to-one relationship with cluster-level distributions of individual-level characteristics. The proposed estimation methods for the latent factors are motivated from an interpretable distribution model. Thanks to the parsimonity of the latent factor model, a variety of existing moment restriction models can be used to define a parameter of interest in relation to the cluster-level distributions and to explore aggregate-level heterogeneity in data.

Keywords: hierarchical models, multilevel models, functional analysis

JEL classification codes: C13

[*] I am deeply grateful to Stéphane Bonhomme, Christian Hansen and Azeem Shaikh, who have provided me invaluable support and insight. I would also like to thank Michael Dinerstein, Max Tabord-Meehan, Alex Torgovitsky, Martin Weidner and the participants of the metrics advising group and the metrics student group at the University of Chicago for their constructive comments and input. I acknowledge the support from the European Research Council grant ERC-2018-CoG-819086-PANEDA. Any and all errors are my own.

[†] Department of Economics, University of Oxford. Email: myungkou.shin@economics.ox.ac.uk

1 Introduction

A significant volume of datasets used in economics are multilevel; units of observations have a hierarchical structure (see (Raudenbush and Bryk, 2002) for general discussion). For example, a dataset that collects demographic characteristics of a country's population, e.g., the Current Population Survey (CPS) of the United States, often documents each surveyee's geographical location up to some regional level. Throughout this paper, I use *individual* and *cluster* to refer to the lower level and the higher level of this hierarchical structure, respectively: e.g., in CPS, individuals refer to surveyees and clusters refer to states. In light of the multilevel nature of the dataset, a researcher may want to consider a research design that utilizes the multilevel structure. For example, when regressing individual-level outcomes on individual-level regressors with the CPS data, it is customary to include some state-level regressors such as state population or state average income, to control for the cluster-level heterogeneity.

This paper builds up on this motivation and provides a generalized econometric framework where we stay true to the hierarchical structure of the dataset and control for the cluster-level heterogeneity by aggregating the within-cluster individual-level information as a *distribution* function. The idea that some cluster-level aggregation of the individual-level observable information is rich enough to model the cluster-level heterogeneity goes back a long way in the econometrics literature: e.g., Mundlak (1978); Chamberlain (1982) and more. This paper contributes to the literature by providing an interpretable model where the within-cluster distribution of the individual-level characteristics is justified as an appropriate measure of aggregation.

In developing models for multilevel datasets, the restriction that the cluster-level heterogeneity can be controlled with some within-cluster aggregation of the individual-level information is appealing for the following two reasons. Firstly, it allows us to control for the cluster-level heterogeneity while not subsuming the variation from a cluster-level explanatory variable. Consider a linear regression setup and suppose that an explanatory variable of interest only varies at the cluster level: e.g., every individual in the same cluster is subject to the same treatment regime. Y_{ij} is the individual-level outcome variable for individual i in cluster j, Z_j is the cluster-level explanatory variable for cluster j and X_{ij} is the individual-level control covariates for individual i in cluster j. $\mathbf{F}_j$ denotes the distribution of X_{ij} for cluster j.

$$Y_{ij} = \alpha_j + \beta Z_j + {X_{ij}}^\intercal \theta + U_j, \tag{1}$$

$$Y_{ij} = \alpha(\mathbf{F}_j) + \beta Z_j + {X_{ij}}^\intercal \theta + U_j \tag{2}$$

The coefficient β in the fixed-effect regression model (1) is not identified due to the multicollinearity between α_j and Z_j; cluster fixed-effect subsumes the variation from Z_j.[1] On the other hand, with some overlap condition on Z_j given $\mathbf{F}_j$, β is identified in the regression model (2). The identification comes at the cost of imposing restrictions on α_j such that α_j is a function of $\mathbf{F}_j$.

Secondly, in some empirical contexts, the aggregation of the individual-level characteristics can be an explanatory variable of interest on its own: e.g., the researcher is interested in a cluster-level equilibrium effect from the distribution of X_{ij} for cluster j.

$$Y_{ij} = \alpha_j + {X_{ij}}^\intercal \theta + U_j, \tag{3}$$

$$Y_{ij} = \alpha(\mathbf{F}_j) + {X_{ij}}^\intercal \theta + U_j \tag{4}$$

When the cluster-level heterogeneity comes only from the cluster-level aggregation of the individual-level characteristics, the cluster fixed-effect α_j in (3) successfully captures the heterogeneity. However, the fixed-effect model (3) cannot provide a prediction for a hypothetical out-of-sample cluster if it does not look similar to with any of the in-sample clusters. Thus, even without any regressor at the cluster-level, explicitly modeling the cluster-level

[1] In this sense, the issue of controlling for the cluster-level heterogeneity here is closely connected to the treatment endogeneity problem; when α_j is uncorrelated with Z_j, a regression with a constant intercept α identifies β.

heterogeneity to be a function of the cluster-level aggregation of the individual-level information as in (4) can be helpful when we need to do extrapolation for an out-of-sample prediction.

In this paper, the distribution function is used a choice of the aggregation method since I focus on multilevel datasets with large clusters. Suppose that the cluster sizes are small. Then, we can use more flexible methods of aggregation to model the cluster-level heterogeneity and treat the multilevel dataset as a cross-sectional dataset; the simple unordered within-cluster aggregation $\{X_{ij}\}_i$ is low-dimensional. On the contrary, when the clusters are large, the unordered collection $\{X_{ij}\}_i$ becomes high-dimensional even when X_{ij} itself is low-dimensional. In this regard, for the large clusters cases, we need aggregation methods with some dimension reduction property. The distribution function is a sensible choice since there often does not exist any ordering among individuals in the multilevel datasets; the individuals are exchangeable within each cluster. For example, in a census data, the identification number has little meaning on its own. Thus, there is no information loss from using a distribution function instead of an unordered collection, while there is gain of dimension reduction.[2]

The formal econometric framework of this paper consists of two parts: a constructive latent factor model for the cluster-level distributions, which only concerns $\{X_{ij}\}_{i,j}$, and a moment restriction model for a parameter of interest, which may use the entirety of the observed dataset. In the latent factor model, the cluster-level distribution function of the individual-level control covariates is modeled to be a function of a finite-dimensional cluster-level latent factor $\lambda_j \in \mathbb{R}^\rho$: the second layer of dimension reduction. By assuming a bijection between the latent factor and the distribution function, a large class of moment restriction models that take finite-dimensional vectors as inputs can be used to develop a model where a parameter of interest is identified in relation to the cluster-level distributions. The esti-

[2] All of the theoretical results from this paper does not assume any dimension restriction on X_{ij}. Thus, the distribution function $\mathbf{F}_j$ can contain all the information about individual-level heterogeneity, e.g., race, gender, age and etc, as long as it is observed in the dataset.

mation is done in a plug-in manner; the latent factors are estimated from the cluster-level distributions and the estimates are used as true latent factors in the moment restriction model, to estimate the parameter of interest.

The key assumption that makes the plug-in procedure valid is that the moment restriction model is invariant to a rotation on the latent factor while the latent factor model provides an estimator that estimates the latent factor up to some rotation. The 'invariance' is in the sense that we can always find a (interpretable) rotated parameter of interest in the parameter space to establish model equivalence between the original moment restriction model and a new moment restriction model where a rotation is applied to the latent factor. Thanks to the restriction, we do not need to estimate the face values of the latent factors; it suffices to estimate some rotation of the latent factors. The rotation invariance is helpful since it allows us to use machine learning methods with dimension reduction property even though their outputs are not readily interpretable on their own; as long as they are a rotation of an interpretable factor, we can motivate a latent factor model based on the machine learning method. Using this feature, two latent factor models for the cluster-level distributions are proposed in this paper. The first of the two models relates to the K-means clustering algorithm and the second relates to the functional principal component analysis (PCA). Under their respective latent factor models and given appropriate relative growth rate of the cluster size to the number of clusters, both estimators have fast enough estimation error so that the plug-in estimator using the estimates is consistent and asymptotically normal.

As an empirical illustration of the econometric framework of this paper, I revisit the disemployment effect estimation in the context of the US minimum wage and teen employment. By controlling for the state-level distribution of individual-level employment status history and the state-level distribution of individual-level wage income, I show that the aggregate heterogeneity in state-level labor market fundamentals matters in estimating the disemployment effect. In addition, I explore how the individual heterogeneity—age and race—interact with the aggregate heterogeneity in terms of the disemployment effect. I find differential disemployment effect in terms of both of the individual-level control variables and show that the differential also depends on labor market fundamentals.

This paper contributes to several literatures in econometrics. Firstly, this paper contributes to the literature of multilevel/hierarchical/clustered models. Similar to this paper, Yang and Schmidt (2021) identifies the effect of a possibly endogeneous—meaning that it is correlated with the cluster-level heterogeneity—cluster-level variable in a linear regression setup; instead of modeling the cluster-level heterogeneity, they use instruments for the cluster-level explanatory variable. Arkhangelsky and Imbens (2023) also considers a multilevel setup and uses aggregation of individual-level information to control for the cluster-level heterogeneity; however, their goal differs from mine in that they focus on small clusters with individual-level explanatory variable while I focus on large clusters.[3] In a slightly different path, Hansen et al. (2014) focuses on inference and discusses a randomization test for a null that a binary cluster-level treatment has no treatment effect; by shifting the magnitude of the cluster-level heterogeneity in treatment assignment, Hansen et al. (2014) conducts a sensitivity analysis on the power and the size of the test.

Secondly, this paper contributes to the literature of correlated random coefficient models. The simple linear regression examples (1) and (3) above can be thought of as a random coefficient model where the coefficient α_j is possibly correlated with Z_j and/or X_{ij}. When given a cluster-level variable Z_j, a fixed-effect type approach (e.g., Wooldridge (2005); Graham and Powell (2012); Arellano and Bonhomme (2012)) is not applicable. Thus, I impose distributional assumptions on the random effect; if we rewrite this paper's premise that the coefficients/partial effects are identified in relation to the cluster-level distribution in the language of the random coefficient model, it becomes that the random coefficients are un-

[3] Inherently, the problem they focus on only exists in small clusters setup; the within-cluster comparison will identify the effect of the individual-level explanatory variable when the clusters are large. On the other hand, the two motivations I give in this paper applies to both small and large cluster setups; however, the solution of this paper is only valid in large clusters setup since the latent factor model applies to population distribution functions. Thus, one can consider the approach in Arkhangelsky and Imbens (2023) for the motivations discussed in this paper, when the cluster-level distribution functions are not well-approximated due to clusters being small.

correlated with the explanatory variable after conditioning on the cluster-level distributions. In this sense, this paper is closer to Altonji and Matzkin (2005); Bester and Hansen (2009) that also impose some restrictions on the joint distribution of the latent heterogeneity and the observable information.

Lastly, this paper contributes to the literature of the factor model approach in causal inference/program evaluation. The factor model approach in the program evaluation literature assumes that the error term consists of a systemic part, modeled with a factor model, and an idiosyncratic error; the treatment endogeneity happens only through the factors. By assuming a latent factor model for the cluster-level distributions[4], which is sufficiently informative for the cluster-level heterogeneity, this paper also follows the same approach in solving the endogeneity problem of the cluster-level explanatory variable. On the contrary to the canonical synthetic control methods that aim to cancel out the latent factor (Abadie et al., 2010, 2015; Gunsilius, 2023) using pretreatment outcomes, this paper directly estimates the factors; in this sense, Xu (2017) is closer to this paper.

The rest of the paper is organized as follows. In Section 2, I formally discuss the two parts of the econometric framework of this paper: the latent factor model for the cluster-level distributions and the moment restriction model for the parameter of interest. In Section 3, I discuss two latent factor models, which motivate the use of the K-means clustering algorithm and the functional PCA in estimating the cluster-level latent factor. In Section 4, I discuss an extension of the two-level model. In Section 5, an empirical illustration of the econometric framework is provided.

[4] In a factor model for interactive fixed-effects in panel data, the two dimensions of the factor model are unit and time. In this paper, the two dimensions of the (latent) factor model is clusters and individuals. The difference is that the individuals are not ordered in a multilevel data as times are in a panel data; I need to find pairs of individuals from different clusters who share the common factor loadings, as the two observations Y_{it} and Y_{jt} do in panel data. For that, I order individuals with X_{ij}; the variation in the relative position of an individual with $X = x$ across two clusters only comes from the variation in the factors, since the factor loadings are fixed.

2 Distribution as control variable

2.1 Setup

An econometrician observes $\left\{ \{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j \right\}_{j=1}^{J}$ where $Y_{ij} \in \mathbb{R}$ is an individual-level outcome variable for individual i in cluster j, $X_{ij} \in \mathbb{R}^p$ is a p-dimensional vector of individual-level control covariates for individual i in cluster j, and $Z_j \in \mathbb{R}^{p_{cl}}$ is a p_{cl}-dimensional vector of cluster-level control covariates for cluster j. There exist J clusters and each cluster contains N_j individuals: in total there are $N = \sum_{j=1}^{J} N_j$ individuals. Clusters are large; the asymptotic regime of this paper lets both J and $\min_j N_j$ go to infinity. In addition to the observable covariates X_{ij} and Z_j, there exists cluster-level latent factor $\lambda_j \in \Lambda$, which models the cluster-level heterogeneity. The individuals are assumed to be independent and identically distributed within clusters and the clusters are assumed to be independent and identically distributed. Since the cluster sizes are allowed to be uneven, the distributional identity is established through a conditional distribution function H:

$$\left(Z_j, N_j, \lambda_j\right) \sim \text{iid}$$

and for each $j = 1, \cdots, J$,

$$\left(Y_{ij}, X_{ij}\right) \;\middle|\; \{Z_k, N_k, \lambda_k\}_{k=1}^{J} \overset{iid}{\sim} H(Z_j, N_j, \lambda_j; \xi) \tag{5}$$

independently of $\left\{ \{Y_{ik}, X_{ik}\}_{i=1}^{N_k} \right\}_{k \neq j}$. The conditional distribution function H is a known function of the model parameter ξ. Note that $(Y_{1j}, X_{1j}), \cdots, \left(Y_{N_j j}, X_{N_j j}\right)$ are iid, conditioning on $\{Z_k, N_k, \lambda_k\}_{k=1}^{J}$: individual-level iidness within cluster. Also, the distribution of (Y_{ij}, X_{ij}) only depends on (Z_j, N_j, λ_j), independent of other clusters: cluster-level independence. Lastly, (Z_j, N_j, λ_j) are iid and the function H is not subscripted with j: cluster-level distributional identity.

In this model, the cluster-level latent factor λ_j models the cluster-level heterogeneity and I assume that there is an one-to-one relationship between the latent factor λ_j and the cluster-level distribution of X_{ij}. Let $\mathbf{F}_j$ denote the conditional distribution of X_{ij} given (Z_j, N_j, λ_j): for $x \in \mathbb{R}^p$,

$$\mathbf{F}_j(x) = \Pr\left\{X_{ij} \leq x | Z_j, N_j, \lambda_j\right\}.$$

$\mathbf{F}_j$ is a random function.

Assumption 1. $\Lambda \subset \mathbb{R}^\rho$. *There exists an injective function* $G : \Lambda \to [0,1]^{\mathbb{R}^p}$ *such that*

$$\mathbf{F}_j = G(\lambda_j) = G(\lambda_j; \xi).$$

The injectivity of G*: there exist a weighting function* $w : \mathbb{R}^p \to \mathbb{R}_+$ *and an induced* l_2 *norm* $\| \cdot \|_{w,2}$ *such that*

$$\|\mathbf{F}\|_{w,2} = \left(\int_{\mathbb{R}^p} \mathbf{F}(x)^2 w(x) dx\right)^{\frac{1}{2}}.$$

$\lambda \neq \lambda' \Rightarrow \|G(\lambda) - G(\lambda')\|_{w,2} > 0$ *and* $\Pr\left\{\|G(\lambda_j)\|_{w,2} < \infty\right\} = 1$.

Assumption 1 combined with the clustered data model (5) assumes that the cluster-level distribution of individual-level control covariates $\mathbf{F}_j$ sufficiently controls for the cluster-level heterogeneity; H is a function of $(N_j, Z_j, G^{-1}(\mathbf{F}_j))$.

Using the cluster-level distribution $\mathbf{F}_j$ as a control covariate in a model for the clustered data can be motivated in two different ways. Firstly, suppose that the econometrician is interested in identifying the effect of some cluster-level observable characteristic Z_j on individual-level outcome Y_{ij},[5] while suspecting cluster-level latent heterogeneity. The cluster-level heterogeneity cannot be modelled to be fully flexible, e.g., with a cluster fixed-effect, due

[5] Many research questions in economics fit this description. For example, economists study the effect of a raise in the minimum wage level, a state-level variable, on employment status, an individual-level variable (Allegretto et al., 2011, 2017; Neumark et al., 2014; Cengiz et al., 2019; Neumark and Shirley, 2022); the effect of a team-level performance pay scheme on worker-level output (Hamilton et al., 2003; Bartel et al., 2017; Bandiera et al., 2007); the effect of a local media advertisement on individual consumer choice (Shapiro, 2018); the effect of a class/school-level teaching method on student-level outcomes (Algan et al., 2013; Choi et al., 2021), etc.

to the limitation that there is no within-cluster variation in Z_j. An alternative to using cluster fixed-effects is to aggregate X_{ij} for each cluster, assuming that aggregating individual-level information for each cluster sufficiently controls for cluster-level heterogeneity. This approach is appealing when the cluster sizes are relatively small. However, when the clusters are large, a simple collection of the individual-level information $\{X_{ij}\}_{i=1}^{N_j}$ will be high dimensional. Note that in many empirical contexts, the order of individuals in a cluster does not provide additional information regarding the cluster-level heterogeneity; the order of individuals is often simply a random order of data collection.[6] In these empirical contexts, the cluster-level distribution $\mathbf{F}_j$ reduces the dimension of the simple collection $\{X_{ij}\}_{i=1}^{N_j}$, while preserving the relevant information; for more discussion, see Appendix. Thus, using $\mathbf{F}_j$ to model the cluster-level heterogeneity is a sensible dimension reduction approach when the econometrician wants to use the collection of X_{ij} to control for the cluster-level heterogenetiy.

Secondly, there are empirical contexts where the econometrician is directly interested in identifying the effect of the cluster-level distribution $\mathbf{F}_j$ on Y_{ij}. In these cases, the cluster-level distribution of individual-level control covariates X_{ij} is a 'regressor' of interest on its own: e.g., the effect of state-level wage income distribution on an individual's disemployment probability; the effect of a school's racial composition on student's academic performance, etc. In these examples, the effect of the distribution function $\mathbf{F}_j$ is often referred to as "contextual effect" or "equilibrium effect." For these research questions, having a model G for the distribution function as in Assumption 1 can be particularly helpful if we want to discuss out-of-sample prediction of the outcome Y_{ij} given an arbitrary distribution of individual-level covariates, in a reduced-form setup; cluster fixed-effects will successfully control for the cluster-level heterogeneity, but will not be able to give us a prediction for $\mathbf{E}\left[Y_{ij}|\mathbf{F}_j = \mathbf{F}\right]$ when $\mathbf{F}$ is different from $\{\mathbf{F}_1, \cdots, \mathbf{F}_J\}$.

In addition to suggesting that the cluster-level distribution $\mathbf{F}_j$ be used as a control covariate in controlling for the cluster-level heterogeneity, Assumption 1 also assumes that the

[6] Exceptions would be when the individuals are ordered in a specific way; e.g., siblings being ordered in their birth order within a family, workers being ordered in terms of seniority within a firm, etc.

latent factor λ_j is defined on a finite-dimensional space: $\Lambda \subset \mathbb{R}^{\rho}$. Thus, under Assumption 1, an infinite-dimensional object $\mathbf{F}_j$ is reduced to a finite-dimensional factor λ_j, through G. This adds a (additional) layer of dimension reduction, giving us practical benefits. For example, by modelling $\mathbf{F}_j$ to be a function of λ_j, the task of estimating an infinite-dimensional object $\mathbf{F}_j$ becomes an easier task of estimating a finite-dimensional factor λ_j. Also, a variety of econometric frameworks that use finite-dimensional control covariates become readily applicable by substituting λ_j for $\mathbf{F}_j$; e.g. when we want to construct a regression model where a binary outcome Y_{ij} depends on a distribution function $\mathbf{F}_j$, we can directly use the known results on the logistic model with finite-dimensional control covariates, by substituting λ_j for $\mathbf{F}_j$.

Given the clustered data model (5) and the (possibly infinite-dimensional) model parameter ξ, I assume that a finite-dimensional parameter of interest $\theta = \theta(\xi)$ is identified with a moment restriction model: at true value of θ,

$$\mathbf{E}\left[m(W_j^*; \theta)\right] = 0. \tag{6}$$

Let l denote the dimension of m and k denote the dimension of θ: $l \geq k$. W_j^* is a function of cluster-level random objects $\left(\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j, N_j, \lambda_j\right)$. Note that λ_j is latent; the superscript * is used to denote that W_j^* is not directly observed.

Example 1 *(clustered treatment)* Consider a binary treatment assigned at the cluster level: $Z_j \in \{0, 1\}$,

$$Y_{ij} = Y_{ij}(1) \cdot Z_j + Y_{ij}(0) \cdot (1 - Z_j) \tag{7}$$

and assume unconfoundedness with the cluster-level latent factor λ_j:

$$\left(Y_{ij}(1), Y_{ij}(0), X_{ij}\right) \perp\!\!\!\perp Z_j \;\middle|\; \left(N_j, \lambda_j\right). \tag{8}$$

The average treatment effect (ATE) is identified with moment restrictions using the inverse probability weighting: with some known function π such that $\mathbf{E}\left[Z_j | N_j, \lambda_j\right] = \pi(\lambda_j; \theta_\pi)$,

$$\begin{aligned}
\theta &= \left(\mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right], {\theta_\pi}^\intercal\right)^\intercal, \\
W_j^* &= \left(\bar{Y}_j, Z_j, \lambda_j\right)^\intercal, \\
m(W_j^*; \theta) &= \begin{pmatrix} \left(\frac{Z_j}{\pi(\lambda_j;\theta_\pi)} - \frac{1-Z_j}{1-\pi(\lambda_j;\theta_\pi)}\right)\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right] \\ \lambda_j\left(Z_j - \pi(\lambda_j;\theta_\pi)\right) \end{pmatrix}.
\end{aligned}$$

In this example, it is assumed that the binary treatment variable is clustered and nonrandom. The unconfoundedness assumption (8) assumes that the treatment is independent of the potential outcomes conditioning on the latent factor λ_j, i.e., the cluster-level distribution of X_{ij}; the treatment is as good as random between two clusters with the same distribution of individual-level characteristics. Suppose for example that the econometrician is interested in the effect of a state-wide policy on individual-level outcomes in the United States. The unconfoundedness assumption (8) would be to assume that the adoption of the policy is independent of the potential outcomes for a given state, conditioning on the state-level distribution of individuals; we compare two states with the same distribution of individual characteristics to estimate the effect of the policy adoption.

Example 2 *(linear regression)* Consider a regression model where X_{ij}, Z_j and λ_j enter the model linearly:

$$\begin{aligned}
Y_{ij} &= {X_{ij}}^\intercal\theta_1 + {Z_j}^\intercal\theta_2 + {\lambda_j}^\intercal\theta_3 + U_{ij}, \qquad (9) \\
0 &= \mathbf{E}\left[U_{ij} | X_{ij}, Z_j, N_j, \lambda_j\right].
\end{aligned}$$

Then, the slope coefficients are identified from $\mathbf{E}\left[U_{ij}|X_{ij}, N_j, Z_j, \lambda_j\right] = 0$:

$$
\begin{aligned}
\theta &= \left(\theta_1{}^\intercal, \theta_2{}^\intercal, \theta_3{}^\intercal\right)^\intercal, \\
W_j^* &= \left(\frac{1}{N_j}\sum_{i=1}^{N_j} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix} Y_{ij}, \frac{1}{N_j}\sum_{i=1}^{N_j} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix}^\intercal \right), \\
m\left(W_j^*; \theta\right) &= \frac{1}{N_j}\sum_{i=1}^{N_j} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix} Y_{ij} - \frac{1}{N_j}\sum_{i=1}^{N_j} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix} \begin{pmatrix} X_{ij} \\ Z_j \\ \lambda_j \end{pmatrix}^\intercal \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}.
\end{aligned}
$$

The linear regression model assumes that the individual-level characteristics X_{ij}, the cluster-level characteristics Z_j and the cluster-level distribution $\mathbf{F}_j$ enter the regression linearly. Specifically, the model assumes that $\mathbf{F}_j$ enters the model linearly in the sense that the function G^{-1} maps $\mathbf{F}_j$ to a finite-dimensional factor in which the model is linear: $\theta_3{}^\intercal G^{-1}(\mathbf{F}_j) = \lambda_j{}^\intercal \theta_3$. Given the linear regression model, the comparative statistics in terms of the cluster-level distribution $\mathbf{F}_j$ can be constructed with the inverse function G^{-1}:

$$
\begin{aligned}
\mathbf{E}\left[Y_{ij}|X_{ij} = x, Z_j = z, N_j = n, \mathbf{F}_j = \mathbf{F}'\right] - \mathbf{E}\left[Y_{ij}|X_{ij} = x, Z_j = z, N_j = n, \mathbf{F}_j = \mathbf{F}\right] \\
= \theta_3{}^\intercal \left(G^{-1}(\mathbf{F}') - G^{-1}(\mathbf{F})\right)
\end{aligned}
$$

when $\mathbf{F}, \mathbf{F}' \in G(\Lambda)$.

2.2 Plug-in estimation with an estimator for the latent factor λ_j

In the previous subsection, the moment function m in (6) was constructed with the true cluster-level factor λ_j, which is unobservable. In practice, even when G is known, λ_j is not directly observed since $\mathbf{F}_j$ is not directly observed; λ_j has to be estimated. To have a broad applicability, I impose a relatively relaxed condition on the latent factor estimation that there is a consistent estimator for some linear transformation of λ_j.

Consider an invertible $\rho \times \rho$ matrix A and the transformed latent factor $\tilde{\lambda} = A\lambda \in A\Lambda$. Then, by letting $G_A(\tilde{\lambda}) = G(A^{-1}\tilde{\lambda})$, Assumption 1 holds with G_A. Likewise, by modifying the construction of W_j so that it is a function of $\left(\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j, N_j, A^{-1}\tilde{\lambda}_j\right)$, the moment restriction model (6) holds as well. Based on this observation, I assume that there exists an estimator for some linear transformation of the latent factor; $\hat{\lambda}_j$ is an estimator for $A\lambda_j$, for $j = 1, \cdots, J$. Specific examples of the model for the cluster-level distribution function G and the estimators for the latent factor λ_j are discussed in the next section.

Consider a function W which takes cluster-level observable variables and the latent factor λ_j and computes the observation relevant for the moment restriction model (6). Let

$$W_j(\lambda) = W\left(\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j, N_j, \lambda\right).[7]$$

$W_j(\lambda)$ takes a latent factor λ and computes W using observable information from cluster j; $W_j^* = W_j(\lambda_j)$ is the infeasible true observation for cluster j, used in the moment restriction model in the previous subsection, and $\widehat{W}_j = W_j(\hat{\lambda}_j)$ is the feasible observation for cluster j, used in the estimation. Recall that l denotes the dimension of m and k denotes the dimension of θ.

Assumption 2. *There are (random) invertible $\rho \times \rho$ matrices A and $\tilde{A}$. Assume*

- ***a.*** *Θ, the parameter space for θ, is a compact subset of $\mathbb{R}^k$.*

  The true value of θ, denoted with θ^0, lies in the interior of Θ.

- ***b.*** $\mathbf{E}[m(W_j^*; \theta^0)] = 0$ *and for any $\varepsilon > 0$,*

$$\inf_{\|\theta - \theta^0\|_2 \geq \varepsilon} \left\|\mathbf{E}\left[m(W_j^*; \theta)\right]\right\|_2 > 0.$$

- ***c.*** $\sup_{\theta \in \Theta} \left\|\frac{1}{J}\sum_{j=1}^{J} m(W_j^*; \theta) - \mathbf{E}\left[m(W_j^*; \theta)\right]\right\|_2 \xrightarrow{p} 0$ *as $J \to \infty$.*

[7] There is a slight abuse of notation here since the dimension of the input depends on N_j.

d. *For each $\theta \in \Theta$, $W_j = W_j(A\lambda_j)$ satisfies*

$$m(W_j^*; \theta) = m\big(W_j; \tilde{A}\theta\big)$$

almost surely.

e. *For each $\theta \in \Theta$, the map $\lambda \mapsto m(W_j(\lambda); \theta)$ is almost surely continuously differentiable. Also, there are some $\eta, M > 0$ such that*

$$\mathbf{E}\left[\sup_{\|\lambda' - \lambda_j\|_2 \le \eta} \sup_{\theta \in \Theta} \left\| \frac{\partial}{\partial \lambda} m\Big(W_j(\lambda); \theta\Big)\Big|_{\lambda = \lambda'} \right\|_F^2 \right] \le M.$$

f. *There is some $\tilde{M} > 0$ such that* $\Pr\left\{\|A^{-1}\|_F \le \tilde{M}\right\} \to 1$ *as $J \to \infty$.*

Assumption 2.a-c are the usual sufficient conditions for consistency of an extremum estimator. Assumption 2.d assumes that the model is invariant to a rotation on the latent factor. Assumption 2.e assumes that the first derivative of the moment function with regard to the latent factor is bounded in expectation when evaluated within a small neighborhood around the true latent factor.

Assumption 2.d adds an extra restriction to the moment restriction model; the same moment restriction can still be used with the rotated factor $W_j = W_j(A\lambda_j)$, as long as the parameter of interest θ is adjusted accordingly. This restriction is particularly helpful since it allows us to estimate the (rotated) parameter of interest while not knowing the rotation A; we cannot retrieve $W_j^* = W(\lambda_j)$ from $A\lambda_j$ when A is unknown. A sufficient condition for Assumption 2.d is to assume a single index restriction that the latent factor λ_j enters the moment function m as a single index of ${\lambda_j}^\intercal \theta_\lambda$ when $\theta = ({\theta_\lambda}^\intercal, {\theta_{-\lambda}}^\intercal)^\intercal$.

The key prerequisite for this assumption in an empirical researcher's perspective is that the rotated parameter of interest $\tilde{A}\theta$ still has an interpretable implication as the original parameter of interest θ. In Example 1, when we assume Assumption 2.d for the propensity score model, the rotation in the parameter of interest θ only applies to a subvector of θ; the

ATE parameter $\mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right]$ remains unchanged. Thus, we can estimate the ATE given estimates of the rotated latent factor $A\lambda_j$. In Example 2, it is straightforward to see that the linear regression model satisfies Assumption 2.d and $\tilde{A}\theta = \left({\theta_1}^\intercal, {\theta_2}^\intercal, \left({A^\intercal}^{-1}\theta_3\right)^\intercal\right)^\intercal$. The slope coefficients on X_{ij} and Z_j remain unchanged. Moreover, the comparative statistics in terms of $\mathbf{F}_j$ can still be constructed using $\tilde{A}\theta$: given ${\theta_3}^\intercal A^{-1}$ and ${G_A}^{-1}$,

$$\begin{aligned}
&\mathbf{E}\left[Y_{ij}|X_{ij} = x, Z_j = z, N_j = n, \mathbf{F}_j = \mathbf{F}'\right] - \mathbf{E}\left[Y_{ij}|X_{ij} = x, Z_j = z, N_j = n, \mathbf{F}_j = \mathbf{F}\right] \\
&\quad = {\theta_3}^\intercal A^{-1}\left(AG^{-1}\left(\mathbf{F}'\right) - AG^{-1}\left(\mathbf{F}\right)\right) \\
&\quad = {\theta_3}^\intercal A^{-1}\left({G_A}^{-1}\left(\mathbf{F}'\right) - {G_A}^{-1}\left(\mathbf{F}\right)\right)
\end{aligned}$$

when $\mathbf{F}, \mathbf{F}' \in G(\Lambda)$. The second equality holds since

$$\begin{aligned}
{G_A}^{-1}(\mathbf{F}) &= {G_A}^{-1}\left(G\left(G^{-1}(\mathbf{F})\right)\right) = {G_A}^{-1}\left(G\left(A^{-1}AG^{-1}(\mathbf{F})\right)\right) \\
&= {G_A}^{-1}\left(G_A\left(AG^{-1}(\mathbf{F})\right)\right) = AG^{-1}(\mathbf{F}). \qquad \left(\because G(A^{-1}\lambda) = G_A(\lambda)\right)
\end{aligned}$$

Theorem 1 establishes the consistency of the GMM estimator for the rotated parameter of interest.

Theorem 1. *Assumptions 1-2 hold. There is an consistent estimator* $\left\{\hat{\lambda}_j\right\}_{j=1}^J$ *for* $\left\{\lambda_j\right\}_{j=1}^J$ *such that*

$$\left\|\begin{pmatrix}\hat{\lambda}_1 & \cdots & \hat{\lambda}_J\end{pmatrix} - A\begin{pmatrix}\lambda_1 & \cdots & \lambda_J\end{pmatrix}\right\|_F = o_p\left(1\right).$$

Let $\widehat{W}_j = W_j\left(\hat{\lambda}_j\right)$ *be the estimated observation for cluster* j. $\hat{\theta}$ *solves*

$$\min_{\theta \in \tilde{A}\Theta}\left\|\frac{1}{J}\sum_{j=1}^J m\left(\widehat{W}_j; \theta\right)\right\|_2.$$

Then,

$$\hat{\theta} \xrightarrow{p} \tilde{A}\theta^0$$

as $J \to \infty$.

Proof. See Appendix. □

Theorem 1 assumes that the researcher is given some $\sqrt{J}$-consistent estimator for the rotated latent factor $A\lambda_j$:

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

Assumption 3 introduces additional regularity conditions used to derive asymptotic normality of the GMM estimator. Theorem 2 establishes the asymptotic normality.

Assumption 3. *Assume*

- ***a.*** *Let $\tilde{m}$ denote a component of the moment function m. The map $\theta \mapsto \tilde{m}(W_j;\theta)$ is almost surely twice continuously differentiable and there is some $\eta, M > 0$ such that*

$$\mathbf{E}\left[\sup_{\|\theta' - \tilde{A}\theta^0\|_2 \leq \eta} \left\| \frac{\partial^2}{\partial\theta\partial\theta^\intercal} \tilde{m}\left(W_j;\theta\right) \Big|_{\theta=\theta'} \right\|_2 \right] \leq M$$

- ***b.*** $\mathbf{E}\left[\frac{\partial}{\partial\theta} m\left(W_j;\theta\right)\Big|_{\theta=\tilde{A}\theta^0}\right]$ *has full rank.*

Theorem 2. *Assumptions 1-3 and conditions in Theorem 1 hold. $\hat{\theta}$ satisfies*

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

and the estimator for the latent factor satisfies

$$\left\| \begin{pmatrix} \hat{\lambda}_1 & \cdots & \hat{\lambda}_J \end{pmatrix} - A \begin{pmatrix} \lambda_1 & \cdots & \lambda_J \end{pmatrix} \right\|_F = o_p\left(\frac{1}{\sqrt{J}}\right).$$

Then,

$$\sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right)$$

as $J \to \infty$, where

$$\begin{aligned}\Sigma = &\left(\mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)^\intercal\right]\mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)\right]\right)^{-1}\\ &\cdot \mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)^\intercal\right]\mathbf{E}\left[m\left(W_j;\tilde{A}\theta^0\right)m\left(W_j;\tilde{A}\theta^0\right)^\intercal\right]\mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)\right]\\ &\cdot\left(\mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)^\intercal\right]\mathbf{E}\left[m_\theta\left(W_j;\tilde{A}\theta^0\right)\right]\right)^{-1}.\end{aligned}$$

Proof. See Appendix. □

Theorem 2 assumes a faster rate on the latent factor estimation. Now the rotated latent factor $A\lambda_j$ is assumed to have a J-consistent estimator:

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

3 Latent factor models for distribution

A notable feature of the hypertheorems in Section 2 is that the latent factor λ_j and the model parameter θ are both discussed in terms of some rotation A and a corresponding shift $\theta \mapsto \tilde{A}\theta$; the hypertheorems are confined to moment restriction models that are invariant to some rotation of the latent factor. Thanks to the rotation invariance, we can use the machine learning algorithms that summarize patterns of high-dimensional objects, such as distributions, and construct low-dimensional outputs, even though they are often not readily interpretable in the context of an econometric model. In this section, I take the K-means clustering and the functional PCA as examples of such an algorithm and develop two different econometric models for the cluster-level distribution of individual-level characteristics G and construct rotated estimators for the latent factor λ_j.

3.1 K-means clustering

The K-means clustering algorithm is an algorithm that solves a minimization problem called the K-means minimization problem. The K-means minimization problem takes J data points and a predetermined number of groups ρ and finds a grouping structure on the J data points such that the sum of the distance between data points and their closest group centeroid is minimized. In this paper, a data point is a cluster-level distribution of the individual-level control covariate $\mathbf{F}_j$. However, we do not directly observe $\mathbf{F}_j$. Thus, as an estimator for $\mathbf{F}_j$, I use the empirical distribution function $\hat{\mathbf{F}}_j$: for all $x \in \mathbb{R}^p$,

$$\hat{\mathbf{F}}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\{X_{ij} \leq x\}.$$

A key observation which directly follows Assumption 1 is that

$$\mathbf{E}\left[\hat{\mathbf{F}}_j(x) | Z_j, N_j, \lambda_j\right] = \big(G(\lambda_j)\big)(x)$$

for every $x \in \mathbb{R}^p$: $\hat{\mathbf{F}}_j$, the estimator I use for $\mathbf{F}_j$, is pointwise unbiased.

Now that we have estimates for the cluster-level distributions, a feasible version of the K-means minimization problem can be defined for some $\rho \leq J$. With the predetermined ρ, the minimization problem assigns each cluster to one of ρ groups so that clusters within a group are similar to each other in terms of the l_2 norm $\|\cdot\|_{w,2}$ on $\hat{\mathbf{F}}_j$:

$$\left(\hat{\lambda}_1, \cdots, \hat{\lambda}_J, \hat{G}(1), \cdots, \hat{G}(\rho)\right) = \arg\min_{\lambda, G} \sum_{j=1}^{J} \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2. \tag{10}$$

In the minimization problem, there are two arguments to minimize the objective over: λ_j and $G(\lambda)$. λ_j is the group to which cluster j is assigned to: $\lambda_j \in \{1, \cdots, \rho\}$. $G(\lambda)$ is the distribution of X_{ij} for group λ. For each cluster j, $\hat{\lambda}_j$ will be the group which cluster j is closest to, measured in terms of $\left\|\hat{\mathbf{F}}_j - \hat{G}(\lambda)\right\|_{w,2}$. Note that the algorithm maps $\hat{\mathbf{F}}_j$ to $\hat{\lambda}_j$, a discrete variable with finite support: dimension reduction.

To solve (10), I use the (naive) K-means clustering algorithm. Find that at the optimum

$$\left(\hat{G}(\lambda)\right)(x) = \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{\hat{\lambda}_j = \lambda\}} \sum_{j=1}^{J} \hat{\mathbf{F}}_j(x)\mathbf{1}\{\hat{\lambda}_j = \lambda\}.$$

The estimated $\hat{G}$ for group λ will be the subsample mean of $\hat{F}_j$ where the subsample is the set of clusters that are assigned to group λ under $\left(\hat{\lambda}_1, \cdots, \hat{\lambda}_J\right)$. Motivated by this observation, the iterative K-means algorithm finds the minimum as follows: given an initial grouping $\left(\lambda_1^{(0)}, \cdots, \lambda_J^{(0)}\right)$,

1. **(update G)** Given the grouping from the s-th iteration, update $G^{(s)}(\lambda)$ to be the subsample mean of $\hat{\mathbf{F}}_j$ where the subsample is the set of clusters that are assigned to group λ under $\left(\lambda_1^{(s)}, \cdots, \lambda_J^{(s)}\right)$:

$$\left(G^{(s)}(\lambda)\right)(x) = \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{\lambda_j^{(s)} = \lambda\}} \sum_{j=1}^{J} \hat{\mathbf{F}}_j(x)\mathbf{1}\{\lambda_j^{(s)} = \lambda\}.$$

2. **(update λ)** Given the subsample means from the s-th iteration, update $\lambda_j^{(s)}$ for each cluster by letting $\lambda_j^{(s+1)}$ be the solution to the following minimization problem: for $j = 1, \cdots, J$,

$$\min_{\lambda \in \{1,\cdots,\rho\}} \left\| \hat{\mathbf{F}}_j - G^{(s)}(\lambda) \right\|_{w,2}.$$

3. Repeat 1-2 until $\left(\lambda_1^{(s)}, \cdots, \lambda_J^{(s)}\right)$ is not updated, or some stopping criterion is met.

For stopping criterion, popular choices are to stop the algorithm after a fixed number of iterations or to stop the algorithm when updates in $G^{(s)}(\lambda)$ are sufficiently small. While the iterative algorithm is extremely fast, giving us computational gain, there is no guarantee that the algorithm gives us the global minimum.[8] Thus, I suggest using multiple initial groupings and comparing the results of the K-means algorithm across initial groupings.

Once the K-means minimization problem is solved, I use the estimated group $\hat{\lambda}_j$ as the

[8] For simplicity of the discussion, let the weighting function w in $\|\cdot\|_{w,2}$ be discrete and finite: with some

estimated latent factor, by transforming it to a categorical variable: with $e_1, \cdots, e_\rho$ being the elementary vectors of $\mathbb{R}^\rho$,

$$\hat{\lambda}_j \in \{e_1, \cdots, e_\rho\} =: \Lambda.$$

Note that the estimated latent factor $\hat{\lambda}_j$ is not unique. Given the grouping structure $\hat{\lambda}_j$ and the centeroids $\hat{G}(\lambda)$, we can find a relabeling on $\hat{\lambda}_j$ and $\hat{G}(\lambda)$ such that the minimum for (10) is still attained. Thus, we cannot take the face value of $\hat{\lambda}_j$ and interpret it to be an estimtaor for the true latent factor λ_j.

Now, it remains to develop an econometric model where the estimator for the latent factor using the K-means clustering algorithm is actually a consistent estimator for the true latent factor with sensible interpretation, at the rate discussed in Theorems 1-2. Assumption 4 discusses a set of conditions for that.

Assumption 4. *Assume with some constant $C > 0$,*

a. *(no measure zero type) $\mu(r) := \Pr\{\lambda_j = e_r\} > 0 \ \forall r = 1, \cdots, \rho$.*

b. *(sufficient separation) For every $r \neq r'$,*

$$\|G(e_r) - G(e_{r'})\|_{w,2}{}^2 =: c(r, r') > 0.$$

c. *(growing clusters) $N_{\min} = \max_n\{\Pr\{\min_j N_j \geq n\} = 1\} \to \infty$ as $J \to \infty$.*

Assumption 4.a ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. Under Assumption 4.b, clusters with different values of the

$x^1, \cdots, x^d \in \mathbb{R}^p$,

$$\|\mathbf{F}\|_{w,2} = \left(\sum_{\tilde{d}=1}^{d} \left(\mathbf{F}(x^{\tilde{d}}) \right)^2 w(x^{\tilde{d}}) \right)^{\frac{1}{2}}.$$

Then, Inaba et al. (1994) shows that the global minimum can be computed in time $O(J^{d\rho+1})$. On the other hand, the iterative algorithm is computed in time $O(J\rho d)$; the computation time becomes proportional to J by using the iterative algorithm. A number of alternative algorithms with computation time linear in J have been proposed and some of them, e.g. Kumar et al. (2004), have certain theoretical guarantees. However, most of the alternative algorithms are complex to implement.

latent factor will be distinct from each other in terms of their distributions of X_{ij}. Thus, the K-means algorithm that uses $\hat{\mathbf{F}}_j$ is able to tell apart clusters with different values of λ_j, when $\hat{\mathbf{F}}_j$ is a consistent estimator for $\mathbf{F}_j$. Assumption 4.c assumes that the size of clusters goes to infinity as the number of clusters goes to infinity. This assumption limits our attention to cases where clusters are large. It should be noted that Assumption 4.c excludes cases where the size of cluster increases only for some clusters and is fixed for some other clusters; the estimation of $\hat{\mathbf{F}}_j$ jointly improves as J increases.

The key element of the econometric model described in Assumption 5 is that there are finite types of clusters, in terms of their distribution of individual-level control covariates X_{ij}. Thus, using Assumption 4 to model the cluster-level heterogeneity would make the most sense when we expect that the heterogeneity across clusters are discrete and finite.

Proposition 1 derives a rate on the estimation error of the latent factor.

Proposition 1. *Assumptions 1-2, 4 hold. Then, there is a rotation matrix A such that*

$$\Pr\left\{\exists\ j\ \ s.t.\ \ \hat{\lambda}_j \neq A\lambda_j\right\} = o\left(\frac{J}{{N_{\min}}^{\nu}}\right) + o(1)$$

for any $\nu > 0$ as $J \to \infty$. Suppose there is some $\nu^* > 0$ such that ${N_{\min}}^{\mu^*}/J \to \infty$ as $J \to \infty$. Then,

$$\left\|\widehat{\Lambda} - A\Lambda\right\|_F = \left(2\sum_{j=1}^{J}\mathbf{1}\left\{\hat{\lambda}_j \neq A\lambda_j\right\}\right)^{\frac{1}{2}} = o_p\left(\frac{1}{\sqrt{J}}\right).$$

Proof. See Appendix. □

Proposition 1 shows that the misclassification probability of the K-means algorithm grouping clusters with different values of λ_j goes to zero when $J/{N_{\min}}^{\nu^*}$ goes to zero for some $\nu^* > 0$. When the misclassification probability converges to zero, the estimation error $\|\hat{\Lambda} - A\Lambda\|_F$ is $o_p(a_n)$ for any sequence $\{a_n\}_{n=1}^{\infty}$ since for any $\varepsilon > 0$ the probability $\Pr\left\{a_n\|\hat{\Lambda} - A\Lambda\|_F > \varepsilon\right\}$ is bounded by the misclassification probability.

Under Assumption 5, we can apply the K-means clustering estimator for the latent factor to a variety of models with a grouping structure. For example, Example 1 in the previous section will be a clustered treatment model with latent group-specific propensity score. Example 2 in the previous section will be a group fixed-effect regression model.

Though the K-means clustering estimator has desirable qualities such as being concise and having a fast estimation rate, the finite support assumption can be too restrictive, depending on contexts. Thus, in the next subsection, I propose an alternative framework where the cluster-level heterogeneity λ_j is assumed to be continuous, using the functional PCA.

3.2 Functional principal component analysis

The functional PCA is an extension of the matrix PCA technique to a functional dataset. Given J functions, the functional PCA computes their product matrix and apply the eigenvalue decomposition to the product matrix to extract a finite number of eigenvectors that explain the most of the variation across J functions. In this paper, cluster-level density function of the individual-level control covariates X_{ij} is used as the functions to which the functional PCA is applied. Again, the density functions are not directly observed. Thus, we compute the product matrix using kernel estimation. Given some kernel K and bandwidth h,

$$\hat{M}_{jk} = \begin{cases} \dfrac{1}{N_j N_k} \displaystyle\sum_{i=1}^{N_j} \sum_{i'=1}^{N_k} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - X_{ij}}{h}\right) \cdot \frac{1}{h} K\left(\frac{x - X_{i'k}}{h}\right) w(x) dx, & \text{if } j \neq k \\ \dfrac{1}{N_j \left(N_j - 1\right)} \displaystyle\sum_{i=1}^{N_j} \sum_{i' \neq i} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - X_{ij}}{h}\right) \cdot \frac{1}{h} K\left(\frac{x - X_{i'j}}{h}\right) w(x) dx, & \text{if } j = k, \end{cases}$$

$\hat{M}$ is an estimator for $J \times J$ matrix M such that

$$M_{jk} = \int_{\mathbb{R}} \mathbf{f}_j(x) \mathbf{f}_k(x) w(x) dx$$

where $\mathbf{f}_j$ is the cluster-level density function of the individual-level control covariates X_{ij} for cluster j. Note that the density function is not directly estimated; only the J^2 moments are estimated.

Given the estimate for the product matrix, I apply the eigenvalue decomposition to $\hat{M}$ and compute the eigenvectors: $\hat{p}_1, \cdots \hat{p}_J$. Each component of the r-th eigenvectors captures one dimension of heterogeneity across clusters and the value of the r-th eigenvalue denotes the magnitude of the corresponding dimension. Thus, with some predetermined $\rho \leq J$, taking eigenvectors associated with the first ρ largest eigenvalues finds a collection of ρ-dimensional vectors that explain the variation across clusters the most. Estimate λ_j by taking the j-th components of the eigenvectors:

$$\hat{\lambda}_j = \sqrt{J}\,(\hat{p}_{1j}, \cdots, \hat{p}_{\rho j})^\intercal$$

where

$$\hat{\rho}_r = \begin{pmatrix} \hat{p}_{r1} \\ \vdots \\ \hat{p}_{rJ} \end{pmatrix}$$

is the eigenvector associated with the r-th eigenvalue. The rescaling with $\sqrt{J}$ is introduced so that the estimated latent factor $\hat{\lambda}_j$ does not converge to zero as J grows: $\hat{p}_r^\intercal \hat{p}_r = 1$ is imposed in the eigenvalue decomposition. Again, the estimated latent factor $\hat{\lambda}_j$ is not unique. In an eigenvalue decomposition, the eigenvectors are uniquely determined only up to a sign even when the eigenvalues are all distinct.

The following set of assumptions motivate a finite mixture model for the cluster-level density of individual-level characteristics and discuss conditions under which an estimation error rate for the functional PCA estimators is derived.

Assumption 5. *Assume with some constant $C > 0$,*

a. *(finite mixture model for distribution) There are thrice continuously differenciable dis-*

tribution function $G_1, \cdots, G_\rho$ and the latent factor λ_j is nonnegative and sum to one: for any $x \in \mathbb{R}$,

$$\Big(G(\lambda)\Big)(x) = \sum_{r=1}^{\rho} \lambda_r G_r(x).$$

$g_1, \cdots, g_\rho$ are the corresponding density functions. For $a = 0, 1, 2$ and $r = 1, \cdots, \rho$,

$$\sup_{x \in \mathbb{R}} \left\| g_r^{(a)}(x) \right\|_2 \leq C.$$

b. *(sufficient variation in $\{g_r\}_{r=1}^{\rho}$ and $\{\lambda_j\}_{j=1}^{J}$) Let $(V_1, \cdots, V_\rho)$ denote the vector of the ordered eigenvalues of M. There exists some $\tilde{J}$ such that $\Pr\{V_1 > \cdots > V_\rho > 0\} = 1$ when $J \geq \tilde{J}$. Also,*

$$\frac{1}{J}(V_1, \cdots, V_\rho) \xrightarrow{p} (v_1^*, \cdots, v_\rho^*)$$

for some $\{v_r^*\}_{r=1}^{\rho}$ such that $v_1^* > \cdots > v_\rho^* > 0$.

c. *(growing clusters) $N_{\min} = \max_n \{\Pr\{\min_j N_j \geq n\} = 1\} \to \infty$ as $J \to \infty$.*

Assumption 5.a assumes that the cluster-level distribution function $\mathbf{F}_j$ is a mixture of ρ underlying distributions $G_1, \cdots, G_\rho$ with the latent factor λ_j as the mixture weights. In this sense, we can construct a following comparison between the two frameworks proposed in this paper: Assumption 4 for the K-means clustering algorithm assumes that there are finite types of *clusters*, while Assumption 5 for the functional PCA assumes that there are fintie types of *individuals* across clusters.

In addition to motivating the finite mixture model for density, Assumption 5.a assumes that the underlying density functions $g_1, \cdots, g_\rho$ are smooth and bounded, up to third derivative. Under Assumption 5.a, the product matrix M can be rewritten as follows:

$$\begin{aligned}
M_{jk} &= \int_{\mathbb{R}} \mathbf{f}_j(x)\mathbf{f}_k(x)w(x)dx \\
&= \sum_{r,r'} \lambda_{jr}\lambda_{kr'} \int_{\mathbb{R}^p} g_r(x)g_{r'}(x)w(x)dx \\
M &= \begin{pmatrix} \lambda_1{}^{\mathsf{T}} \\ \vdots \\ \lambda_J{}^{\mathsf{T}} \end{pmatrix} \underbrace{\begin{pmatrix} \int_{\mathbb{R}} g_1(x)^2 w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)g_1(x)w(x)dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}} g_1(x)g_\rho(x)w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)^2 w(x)dx \end{pmatrix}}_{=:V} \begin{pmatrix} \lambda_1 & \cdots & \lambda_J \end{pmatrix}.
\end{aligned}$$

Assumption 5.b assumes that the underlying density functions $g_1, \cdots, g_\rho$ have sufficient variation, when measured with $\langle \cdot, \cdot \rangle_w$, and the latent factor λ_j span the column space of the covariance matrix V constructed with $\{g_r\}_{r=1}^{\rho}$.

Proposition 2 derives a rate on the estimation error of the latent factor.

Proposition 2. *Assumptions 1-2, 5 hold. The kernel K used in the estimation procedure satisfy that*

i. *K is bounded, symmetric around zero, and nonnegative.*

ii. *$\int_{\mathbb{R}} K(t)dt = 1$.*

iii. *$\int_{\mathbb{R}} t^2 K(t)dt \leq C$.*

$h \propto N_{\min}{}^{-\nu}$ for some $\nu \in [0.25, 1)$. $\tilde{\Lambda}$ is a matrix whose rows are the eigenvectors of M associated with nonzero eigenvalues and $A^{\mathsf{T}} = V\left(\frac{1}{J}\Lambda\tilde{\Lambda}^{\mathsf{T}}\right)$ diag$\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1}$. Then,

$$\left\|\widehat{\Lambda} - A\Lambda\right\|_F = O_p\left(\frac{\sqrt{J}}{\sqrt{N_{\min}}}\right).$$

Proof. See Appendix. □

Proposition 2 bounds the estimation error rate with the square root of the relative growth rate. Recall that $J/N_{\min} \to 0$ is a sufficiently fast growth rate in the case of the K-means clustering for both consistency and asymptotic normality of the GMM estimator. However, for the functional PCA, $J/N_{\min} \to 0$ is not fast enough for us to have an asymptotic normality; the functional PCA requires larger clusters.

4 Extension

4.1 Generalized multilevel models

A possible direction of generalizing the model in hand is to allow for more than two levels. Suppose the econometrician observes

$$\left\{ \left\{ \{Y_{ijl}, X_{ijl}\}_{i=1}^{N_{jl}}, Z_{jl} \right\}_{j=1}^{J_l} W_l \right\}_{l=1}^{L},$$

where i denotes *individual*, j denotes *cluster*, and l denotes *hypercluster*. Each individual belong to a cluster and each cluster belong to a hyper-cluster. Thus, Y_{ijl} is an outcome variable for individual i in cluster j in hypercluster l. There are various data contexts that are relevant to this model: individuals in counties in states, students in schools in school districts, workers in firms in industries, etc.

The researcher wants their model to incorporate the cluster-level heterogeneity and the hypercluster-level heterogeneity, in terms of the observable information. To stay true to this multilevel structure, firstly construct the cluster-level distribution with individual-level control covariate as before: for every $x \in \mathbb{R}^p$,

$$\hat{\mathbf{F}}_{jl}(x) = \frac{1}{N_{jl}} \sum_{i=1}^{N_{jl}} \mathbf{1}\{X_{ijl} \le x\}.$$

Then, use either the K-means algorithm or the functional PCA to estimate the cluster-level latent factors: $\left\{ \left\{ \hat{\lambda}_{jl} \right\}_{j=1}^{J_l} \right\}_{l=1}^{L}$. Note that the latent factor estimation was done irrespective of each cluster's hypercluster membership: as long as $\hat{\mathbf{F}}_{jl}$ are the same, the subscript l does not matter. Then, the cluster-level observable information

$$\left(\{X_{ijl}\}_{i=1}^{N_{jl}}, Z_{jl} \right),$$

which is high-dimensional, is summarized to

$$\left(\hat{\lambda}_{jl}, Z_{jl} \right).$$

Given these cluster-level latent factor estimates $\hat{\lambda}_{jl}$, construct the hypercluster-level distribution with cluster-level objects: for every $z \in \mathbb{R}^{p_{cl}}$ and $\lambda \in \mathbb{R}^{\rho}$,

$$\hat{\mathbf{F}}_l(k, z) = \frac{1}{J_l} \sum_{j=1}^{J_l} \mathbf{1}\{\hat{\lambda}_{jl} \leq \lambda, Z_{jl} \leq z\}.$$

By applying the K-means or the functional PCA again to estimate the hypercluster-level latent factor $\left\{ \hat{\lambda}_l \right\}_{l=1}^{L}$, we reduce the dimension of the hypercluster-level objects

$$\left(\left\{ \{X_{ijl}\}_{i=1}^{N_{jl}}, Z_{jl} \right\}_{j=1}^{J_l}, W_l \right)$$

into

$$\left(\hat{\lambda}_l, W_l \right).$$

Note that the dimension reduction property is crucial in a multilevel models with more than two levels since we use $\hat{\lambda}_{jl}$, the dimension-reduced summary of the cluster-level distribution $\hat{\mathbf{F}}_{jl}$, to construct a hypercluster-level distribution $\hat{\mathbf{F}}_l$. If we were to use $\hat{\mathbf{F}}_{jl}$ as is, we need to construct a distribution of distributions, which there is yet to be a widely accepted solution to.

5 Empirical illustration: disemployment effect of minimum wage

5.1 Background

In this section, I revisit the question of whether an increase in minimum wage level leads to higher unemployment rate in the United States labor market, while using the state-level distribution of individual-level characteristics as a control variable. This quintessential question in labor economics has often been answered using a state-level policy variation; each state has their own minimum wage level in addition to federal minimum wage level in the United States and thus we see states with different minimum wage levels for the same time period. The state-level policy variation is helpful in that it allows us to control for time heterogeneity. However, there could still be spatial heterogeneity that possibly affects both minimum wage level and labor market outcome of a given state simultaneously, and researchers have long been debating how to estimate the causal effect of minimum wage on employment while controlling for spatial heterogeneity. 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 limited 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).

Note that the multilevel model in the paper fits the empirical context of the minimum wage application very 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 multilevel. 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 iid.

Building on this observation, I apply the results of Sections 2 and 3 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; variation in unemployment rate across states with the same distribution of individual-level characteristics only comes from observable state-level characteristics, including the minimum wage level, and not from the underlying state-level heterogeneity. This approach is complementary to assuming that there exists some unrestricted and time-invariant state-level heterogeneity as in the DID literature; the state-level heterogeneity is allowed to vary over time, but restricted in the sense that it can be fully controlled by the time-varying state-level distribution of individual-level characteristics: $\mathbf{F}_{jt}$.

5.2 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 teenangers 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. 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: $\mathcal{I}_{jt} \subset \tilde{\mathcal{I}}_{jt}$. 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,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}{}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{11}$$

$logMinWage_{jt}$ is the logged minimum wage level of state j at time t. Y_{ijt} is employment status of teen i in state j at time t. X_{ijt} is the control covariates of teen i: age, race, sex, marital status, education. $EmpRate_{jt}$ is the average of Y_{ijt} for every individual in state j while the regression runs only on teens: $EmpRate_{jt} = 1/|\tilde{\mathcal{I}}_{jt}| \sum_{i \in \tilde{\mathcal{I}}_{jt}} Y_{ijt}$. In addition to the observable regressors, cluster fixed-effects α_j and census division time fixed-effects $\delta_{cd(j)t}$ are included.

Let us make two connections between (11) and the discussion on a multilevel model from the previous sections. Firstly, the regressor of interest $MinWage_{jt}$ varies on the state-by-month level, making state-specific time fixed-effects infeasible. This is exactly the same type of multicollinearity problem discussed in Section 2; 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 time fixed-effects by grouping 50 states and Washington D.C. into 9 census divisions: $\delta_{cd(j)t}$. Secondly, (11) already implements the idea of aggregating some individual-level information and using the summary statistic in the regression: $EmpRate_{jt}$. In Allegretto et al. (2011), a conscious choice was made by a researcher to use the mean of Y_{ijt} for every individual in state j at time t, to control for the state-level heterogeneity with observable information.

Building on (11), I motivate a linear regression model with a time-varying state-level latent factor, which will be estimated using the time-specific state-level distribution $\mathbf{F}_{jt}$:

$$Y_{ijt} = \alpha_j + \lambda_{jt}{}^{\intercal}\delta_t + \beta \log MinWage_{jt} + X_{ijt}{}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{12}$$

As implied with the use of $EmpRate_{jt}$, the fundamentals of the state labor market should play a role in an individual's employment status and/or the state legislator's decision on the minimum wage level. To control for that, I firstly apply the K-means clustering estimator to group states at each month using their distributions of individual-level employment history. Thus, I assume that there are finite types of states in terms of their distributions of individual-level employment history. Specifically, I use

$$EmpHistory_{ijt} = \big(Emp_{ijt-1}, \cdots, Emp_{ijt-4}\big) \in \mathcal{X} := \{Emp, Unemp, NotInLaborForce\}^4.$$

Emp_{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. $EmpHistory_{ijt}$ collects $Emp_{ij\tau}$ for $\tau = t-4, \cdots, t-1$; $EmpHistory_{ijt}$ is a four-month-long history of employment status. Since Emp_{ijt} is a categorical variable with a finite support of three elements, $EmpHistory_{ijt}$ has a finite support of 81 elements. Note that $Y_{ijt} = 1 \Leftrightarrow Emp_{ijt} = Emp$ 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 collect $EmpHistory_{ijt}$ for every individual and compute the empirical distribution function: for $x \in \mathcal{X}$,

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

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.

Secondly, I apply the functional PCA estimator to states at each month using their distributions of individual-level wage income: $WageInc_{ijt}$. $WageInc_{ijt}$ is a wage income variable for individual i in state j at time t. The wage income variable comes from the

March Annual Social and Economic Supplement (ASEC); it is observed only once a year and the individuals in the ASEC sample differ from the individuals in the basic monthly CPS sample. Also, since the monthly employment rate is used as a control, using the CPS sample, I only collected individuals from the ASEC sample whose wage income is nonzero: $\breve{\mathcal{I}}_{jt}$. To aggregate the information from $WageInc_{ijt}$, I compute the covariance matrix across states:

$$\hat{M}_{jkt} = \begin{cases} \frac{\sum_{i \in \breve{\mathcal{I}}_{jt}, i' \in \breve{\mathcal{I}}_{kt}}}{\left|\breve{\mathcal{I}}_{jt}\right| \cdot \left|\breve{\mathcal{I}}_{kt}\right|} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - WageInc_{ijt}}{h}\right) \cdot \frac{1}{h} K\left(\frac{x - WageInc_{i'kt}}{h}\right) w(x) dx, & \text{if } j \neq k \\ \frac{\sum_{i, i' \in \breve{\mathcal{I}}_{jt}, i \neq i'}}{\left|\breve{\mathcal{I}}_{jt}\right| \left(\left|\breve{\mathcal{I}}_{jt}\right| - 1\right)} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - WageInc_{ijt}}{h}\right) \cdot \frac{1}{h} K\left(\frac{x - WageInc_{i'jt}}{h}\right) w(x) dx, & \text{if } j = k, \end{cases}$$

For the weighting function w, I use the uniform weighting across zero and the 90th quantile of $WageInc_{ijt}$, computed pooling 22 years.

Then, by combining the two estimates for the latent factors of the distributions—the state-by-month distribution of $EmpHistory_{ijt}$ and the state-by-year conditional distribution of $WageInc_{ijt}$ given $WageInc_{ijt} > 0$—, I construct $\hat{\lambda}_{jt}$. Then, to control for the time heterogeneiety, time-specific coefficient for the latent factor is used: ${\lambda_{jt}}^{\intercal}\delta_t$. By using the two distribution as control variables, I control for the state-level labor market heterogeneity.

5.3 Results

5.3.1 Latent factor estimation

Before providing the estimation results under the main regression specification, I illustrate how the two latent factor estimation methods are implemented on an actual dataset, by looking at a snapshot of the dataset. Firstly, to illustrate how the K-means clustering algorithm is applied to a real dataset, I chose January 2007 since eighteen states raised their minimum wage levels then. It is the timing where the most states raised their minimum wage levels without a federal minimum wage raise. Since $EmpHistory_{ijt}$ captures the latest four month history of individual employment status, the K-means grouping step that uses

$\tilde{X}_{ij,Jan07}$ and assigns 50 states and Washington D.C. into one of the K groups is based on the distribution of employment status history from September 2006 to December 2006. Figure 1 contains the grouping result when there are three groups. Each group is shaded with different color: red, blue and green. 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

Treated states, 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. Find that we have overlap for each group.

Table 1 contains empirical evidence that the groups estimated using the distribution of $\tilde{X}_{ij,Jan07}$ are heterogeneous. Table 1 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: $\{Emp, Unemp\}^4 \setminus \big(Emp, Emp, Emp, Emp\big)$

- Never-in-the-labor-force: $\{NotInLaborForce\}^4$

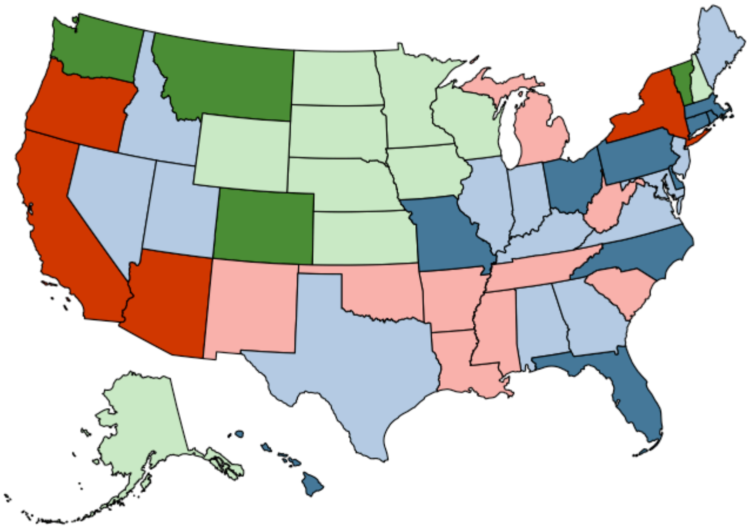

Figure 1: Grouping of states when $\rho_{Kmeans} = 3$, January 2007

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.

'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 have been continuously in the labor force, but 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.

Secondly, to illustrate how the functional PCA is applied to a real dataset, I choose March 2007 ASEC sample, to be compatible with the K-means clustering timeframe. The $WageInc_{ijt}$ captures the annual wage income distribution of individuals across states and Washington D.C., conditioning on the wage income being nonzero, for year 2006. The second to the 14-th largest eigenvalues are plotted in Figure 2; the biggest eigenvalue is much bigger than the rest of the eigenvalues and the associated first component of the estimated latent factor is mostly constant across states and therefore omitted. For the regression, $\rho_{fPCA} = 3$ chosen.

group	1	2	3
Always-employed	0.532	0.586	0.642
Ever-unemployed	0.034	0.031	0.030
Never-in-the-labor-force	0.325	0.282	0.229

Table 1: Heterogeneity across states, Janury 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, using the longitudinal weights provided by the IPUMS-CPS to connect individuals across different months. Then, the group mean is computed by putting equal weights on states.

Hotelling's multivariate t-test rejects the null of same mean for any pair of two groups at significance level 0.001.

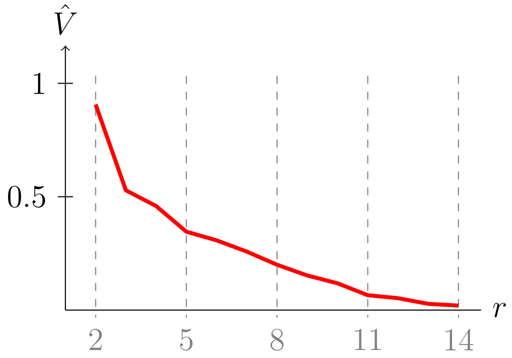

Figure 2: The scree plot of eigenvalues from the wage income distribution, March 2007

March 2007 ASEC sample is used in constructing the wage income distributions.
Bandwidth $h = 10, 100, 500$ are used in the functional PCA and the plot given above uses results from $h = 10$. The eigenvalues are rescaled by multiplying 10^6.
The biggest eigenvalue is not included in the plot: its value was 133.63.

Since the first component of the estimated latent factor is mostly constant across states, I plotted the second component of the estimated latent factor in Figure 3. Several northeastern states have the highest value of the second component $\hat{\lambda}_{jt}$ while some southern states such as Arkansas have the lowest value. Since we do not have an interpretation for the value of $\hat{\lambda}_{jt}$ itself, Figure 3 only provides qualitative results telling us which states are similar.

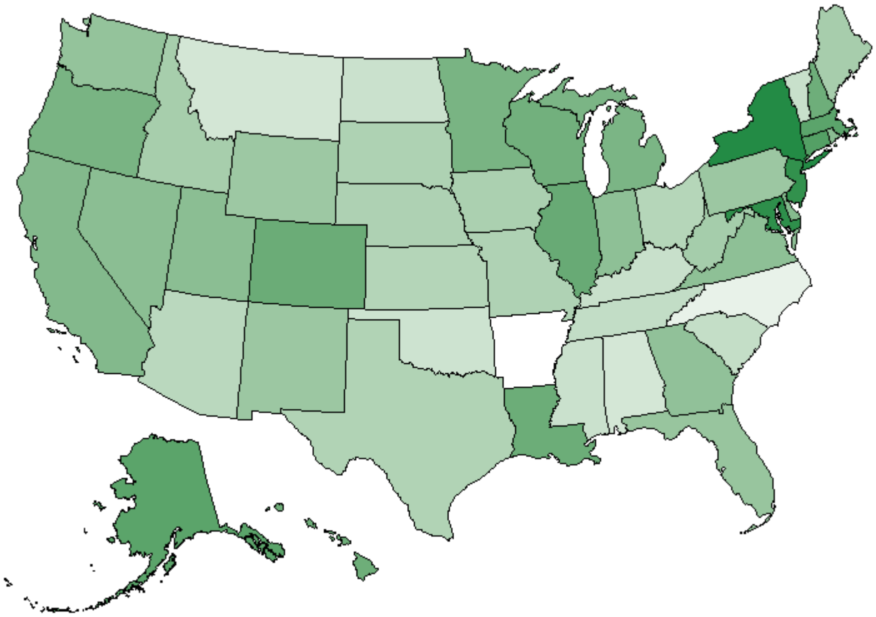

Figure 3: $\hat{\lambda}_{j2}$ across states for March 2007

March 2007 ASEC sample is used in constructing the wage income distributions. Bandwidth $h = 10, 100, 500$ are used in the functional PCA and the plot given above uses results from $h = 10$.

5.3.2 Disemployment effect regression

Now, I discuss the regression results from (12). For the pooled estimation, I repeated the K-means clustering with three groups, i.e. $\rho_{Kmeans} = 3$ for every month and the functionl PCA with three-dimensional factors, i.e. $\rho_{fPCA} = 3$ for every year. Then, combining the estimated latent factors as given, I ran the linear regression of (12). Table 2 contains the estimation result, along with the estimation results for several alternative specifications as benchmarks. In the regression model, the state minimum wage level $MinWage_{jt}$ enters after taking logarithm, following the convention in the literature. Thus, by diving the slope coefficient on $\log MinWage_{jt}$ with the average teen employment rate from the dataset, which is 0.326, we get the elasticity interpretation. Based on columns (3)-(5), the elasticity of teen employment lies between -0.054 and -0.074, meaning that an one percentage point increase in the minimum wage level reduces teen employment by 0.05-0.07 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 distribution, I find that the existing literature overestimates the wage elasticity of teen employment.

β	(1)	(2)	(3)	(4)	(5)
pooled	-0.024	-0.035**	-0.024	-0.023	-0.018
	(0.017)	(0.015)	(0.016)	(0.014)	(0.015)
${\lambda_{jt}}^\intercal\delta_t$	TWFE	Census Div.	K-means	fPCA	K-means and fPCA

Table 2: Impact of minimum wage on teen employment, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment.
When divided by 0.326, the estimates have the elasticity interpretation.
The standard errors are clustered at the state level.
*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

Table 3 discuss the aggregate heterogeneity in treatment effect:

$$Y_{ijt} = \alpha_j + {\lambda_{jt}}^\intercal\delta_t + \beta(\lambda_{jt}) \log MinWage_{jt} + {X_{ijt}}^\intercal\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{13}$$

Note that the slope coefficient for $\log MinWage_{jt}$ is a function of the latent factor λ_{jt}. To make the model parsimonious in the latent factor, it is assumed that

$$\beta(\lambda_{jt}) = {\lambda_{jt,Kmeans}}^\intercal\beta$$

when $\lambda_{jt} = ({\lambda_{jt,Kmeans}}^\intercal, {\lambda_{jt,fPCA}}^\intercal)^\intercal$. The slope is a function of the grouping from the employment history distribution only. Also, to connect the 'labels' of the grouping structure across different time periods, I reordered $\lambda_{jt,Kmeans}$ across t so that Group 1 (i.e. $\lambda_{jt,Kmeans} = e_1$) is always the group of states with lower employment rate and lower labor force participation rate and Group 3 (i.e. $\lambda_{jt,Kmeans} = e_3$)is always the group of states with higher employment rate and higher labor force participation rate.

β	(1)	(2)	(3)	(4)
Group 1	-0.022	-0.034**	-0.019	-0.018
	(0.017)	(0.015)	(0.017)	(0.014)
Group 2	-0.024	-0.035**	-0.023	-0.016
	(0.017)	(0.015)	(0.016)	(0.015)
Group 3	-0.026	-0.038**	-0.037	-0.028
	(0.017)	(0.015)	(0.024)	(0.024)
$\lambda_{jt}{}^{\intercal}\delta_t$	TWFE	Census Div.	K-means	K-means and fPCA

Table 3: Impact of minimum wage on teen employment, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment.
When divided by 0.326, the estimates have the elasticity interpretation.
The standard errors are clustered at the state level.
*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

Columns (3)-(4) show us that teens in Group 1 states where the proportion of 'Always-employed' is lower and the proportion of 'Never-in-the-labor-force' is higher are less affected by the minimum wage and their counter parts in Group 3. However, none of the estimates is significantly away from zero at the significance level 0.1.

In addition to aggregate heterogeneity, I further extend (12)-(13) to discuss individual heterogeneity and aggregate heterogeneity simultaneously. The left panel of Table 4 estimates

$$\begin{aligned} Y_{ijt} = \alpha_j + \lambda_{jt}{}^{\intercal}\delta_t + \beta_{yt} \log MinWage_{jt}\mathbf{1}\{Age_{ijt} \leq 18\}. \\ + \beta_{ot} \log MinWage_{jt}\mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}{}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{14}$$

Treatment effect is heterogeneous in terms of age, at the individual level: β_{yt} is the treatment effect on younger teens and β_{ot} is the treatment effect on older teens. The right panel of

Table 4 estimates

$$
\begin{aligned}
Y_{ijt} = \alpha_j + \lambda_{jt}^\intercal \delta_t + \beta_{yt}(\lambda_{jt}) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} \leq 18\}. \\
+ \beta_{ot}(\lambda_{jt}) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}^\intercal \theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (15)
\end{aligned}
$$

Again, $\beta_{yi}(\lambda_{jt})$ and $\beta_{ot}(\lambda_{jt})$ are assumed to be linear functions of the latent factor estimated with the employment history distribution only; interaction between individual heterogeneity in terms of age and aggregate heterogeneity in terms of employment history is introduced.

Table 4 shows that younger teens, who are under the age of nineteen, are more affected by a raise in the minimum wage level than older teens of the age nineteen in general. In Columns (3)-(4), we see how this individual-level heterogeneity in disemployment effect interacts with aggregate-level heterogeneity. Younger teens tend to be more affected by a raise in the minimum wage level and that tendency is stronger for group 3 states where the employment rate and the labor force participation rate are higher.

Table 5 repeats the same regression specification, but in terms of race; Table 5 documents individual heterogeneity in terms of white teens against non-white teens. From the left panel of Table 5, 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.[9] Again, the racial disparity interacts with the labor market fundamentals. From the right panel of Table 5, it is shown that the racial disparity persists across groups and interact with the aggregate heterogeneity in a way that the disemployment effect is bigger for Group 3 states where the employment rate and the labor force participation rate are high; the employment effect for non-white teenagers is mitigated in Group 3. Figure 4 contains confidence intervals of treatment effect estimates from Column (4) of Table 4 and Column (4) of Table 5.

[9] Suppose that teens with more financial burdens actually increase their labor supply when the minimum wage goes up. Since the regression specification does not control for household financial variables, the racial gap in disemployment effect may be attributed to the racial gap in household finances.

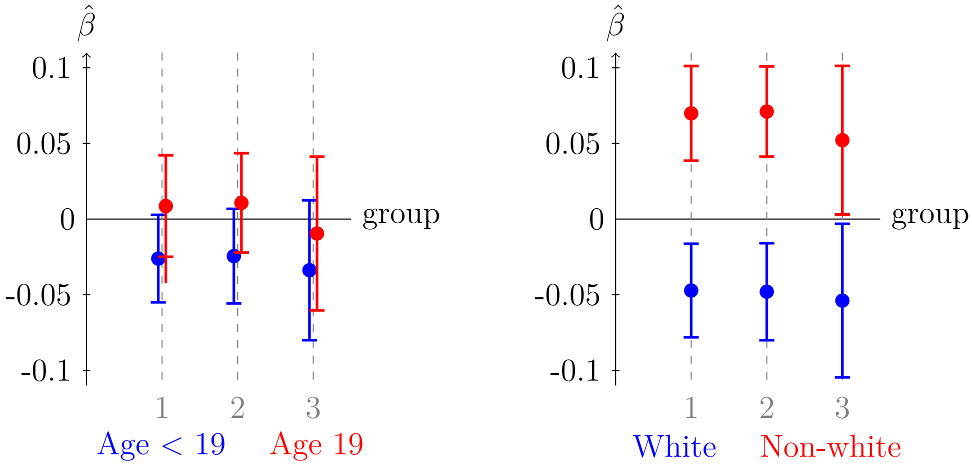

Figure 4: Interaction between individual and aggregate heterogeneity

The figure reports 95% confidence interval of the minimum wage effect estimators, under the group fixed-effects specification where the minimum wage effect is allowed to interact with both an indivdual-level covariate—age or race—and the state-level group membership.

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.

β	(1)	(2)	(3)	(4)
$Age_{ijt} \leq 18$	-0.032*	-0.027*		
	(0.017)	(0.015)		
× Group 1			-0.027	-0.026*
			(0.017)	(0.015)
× Group 2			-0.031*	-0.024
			(0.017)	(0.016)
× Group 3			-0.043*	-0.034
			(0.024)	(0.024)
$Age_{ijt} = 19$	0.002	0.008		
	(0.020)	(0.017)		
× Group 1			0.007	0.009
			(0.021)	(0.017)
× Group 2			0.004	0.011
			(0.018)	(0.017)
× Group 3			-0.019	-0.010
			(0.027)	(0.026)
$EmpHistory$	O	O	O	O
$WageInc$	X	O	X	O

Table 4: Impact of minimum wage on teen employment in terms of age, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment. The regression pools teenagers between the age of 16 and 19 and allows the minimum wage effect to differ across teens younger than 19 and teens of age 19.

When divided by 0.326, the estimates have the elasticity interpretation.

The standard errors are clustered at the state level.

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

β	(1)	(2)	(3)	(4)
$White_{ij} = 1$	-0.055***	-0.049***		
	(0.018)	(0.016)		
× Group 1			-0.049**	-0.047***
			(0.019)	(0.016)
× Group 2			-0.054***	-0.048***
			(0.018)	(0.016)
× Group 3			-0.064**	-0.054**
			(0.027)	(0.026)
$White_{ij} = 0$	0.060***	0.068***		
	(0.016)	(0.01)		
× Group 1			0.067***	0.070***
			(0.018)	(0.016)
× Group 2			0.063***	0.071***
			(0.016)	(0.015)
× Group 3			0.040	0.052**
			(0.025)	(0.025)
$EmpHistory$	O	O	O	O
$WageInc$	X	O	X	O

Table 5: Impact of minimum wage on teen employment in terms of age, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment. The regression pools teenagers between the age of 16 and 19 and allows the minimum wage effect to differ across teens younger than 19 and teens of age 19.

When divided by 0.326, the estimates have the elasticity interpretation.

The standard errors are clustered at the state level.

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

6 Conclusion

This paper motivates the use of the cluster-level distribution of individual-level control covariates to control for the cluster-level heterogeneity in a multilevel model. This framework is most relevant when the clusters are large, so that the cluster-level distributions are well-estimated. By explicitly controlling for the distribution of individuals, two different dimensions of heterogeneity in data are modeled, being true to the multilevel nature of the dataset: individual heterogeneity and aggregate heterogeneity. I apply the estimation method of this paper to revisit the question whether a raise in the minimum wage level has disemployment effect on teens in the United States. I find the disemployment effect to be heterogeneous both at the individual level and the state level, and the two dimensions of heterogeneity interact.

In implementing the idea of distributional control, the K-means algorithm are the functional PCA are used in this paper. The two approaches complement each other; one tolerates slower growth rate of cluster size at the cost of discrete cluster-level heterogeneity and the other allows for continuous cluster-level heterogeneity, requiring larger clusters. However, based on empirical contexts, yet another dimension reduction method on distributions may be more suitable, calling for follow-up research that discuss different functional analysis methods. Also, this paper mostly focuses on cross-section data and non-dynamic panel data. Though the empirical section discusses panel data, the cluster-level latent factor are assumed to be strictly exogenous. An exciting direction for future research is to extend this and study a dynamic multilevel model where the distribution of individuals for each cluster is modelled to be a dynamic process.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 2010, *105* (490), 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Comparative politics and the synthetic control method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

Algan, Yann, Pierre Cahuc, and Andrei Shleifer, "Teaching practices and social capital," *American Economic Journal: Applied Economics*, 2013, *5* (3), 189–210.

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.

Altonji, Joseph G and Rosa L Matzkin, "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 2005, *73* (4), 1053–1102.

Arellano, Manuel and Stéphane Bonhomme, "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 2012, *79* (3), 987–1020.

Arkhangelsky, Dmitry and Guido W Imbens, "Fixed Effects and the Generalized Mundlak Estimator," *Review of Economic Studies*, 2023, p. rdad089.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Incentives for managers and inequality among workers: Evidence from a firm-level experiment," *The Quarterly Journal of Economics*, 2007, *122* (2), 729–773.

Bartel, Ann P, Brianna Cardiff-Hicks, and Kathryn Shaw, "Incentives for Lawyers: Moving Away from "Eat What You Kill"," *ILR Review*, 2017, *70* (2), 336–358.

Bester, C Alan and Christian Hansen, "Identification of marginal effects in a nonparametric correlated random effects model," *Journal of Business & Economic Statistics*, 2009, *27* (2), 235–250.

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.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The effect of minimum wages on low-wage jobs," *The Quarterly Journal of Economics*, 2019, *134* (3), 1405–1454.

Chamberlain, Gary, "Multivariate regression models for panel data," *Journal of econometrics*, 1982, *18* (1), 5–46.

Choi, Syngjoo, Booyuel Kim, Minseon Park, and Yoonsoo Park, "Do Teaching Practices Matter for Cooperation?," *Journal of Behavioral and Experimental Economics*, 2021, *93*, 101703.

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.

Graham, Bryan S and James L Powell, "Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models," *Econometrica*, 2012, *80* (5), 2105–2152.

Gunsilius, Florian F, "Distributional synthetic controls," *Econometrica*, 2023, *91* (3), 1105–1117.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan, "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation," *Journal of political Economy*, 2003, *111* (3), 465–497.

Hansen, Ben B, Paul R Rosenbaum, and Dylan S Small, "Clustered treatment assignments and sensitivity to unmeasured biases in observational studies," *Journal of the American Statistical Association*, 2014, *109* (505), 133–144.

Inaba, Mary, Naoki Katoh, and Hiroshi Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in "Proceedings of the tenth annual symposium on Computational geometry" 1994, pp. 332–339.

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.

Kumar, Amit, Yogish Sabharwal, and Sandeep Sen, "A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions," in "45th Annual IEEE Symposium on Foundations of Computer Science" IEEE 2004, pp. 454–462.

Mundlak, Yair, "On the pooling of time series and cross section data," *Econometrica: journal of the Econometric Society*, 1978, pp. 69–85.

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.

Raudenbush, Stephen W and Anthony S Bryk, *Hierarchical linear models: Applications and data analysis methods*, Vol. 1, sage, 2002.

Shapiro, Bradley T, “Positive spillovers and free riding in advertising of prescription pharmaceuticals: The case of antidepressants,” *Journal of political economy*, 2018, *126* (1), 381–437.

Wooldridge, Jeffrey M, “Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models,” *Review of Economics and Statistics*, 2005, *87* (2), 385–390.

Xu, Yiqing, “Generalized synthetic control method: Causal inference with interactive fixed effects models,” *Political Analysis*, 2017, *25* (1), 57–76.

Yang, Yimin and Peter Schmidt, “An econometric approach to the estimation of multi-level models,” *Journal of Econometrics*, 2021, *220* (2), 532–543.

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* $\sigma_j : \{1, \cdots, N_j\} \to \{1, \cdots, N_j^*\}$ *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 $\left(i_1, \cdots, i_{N_j}\right)$

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

(*unconfoundedness*)

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

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

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

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 3 follows immediately.

Proposition 3. *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*,

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

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

$$\begin{aligned}
&\Pr\Big\{Z_j = 1\big|\mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\Big\} \\
&\quad = \mathbf{E}\Big[\mathbf{E}\Big[Z_j\big|\, \{Y_{ij}(1)^*, Y_{ij}(0)^*, X_{ij}^*\}_{i=1}^{N_j^*}, \sigma_j\Big] \big|\mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\Big] \\
&\quad = \mathbf{E}\Big[\mathbf{E}\Big[Z_j\big|\, \{Y_{ij}(1)^*, Y_{ij}(0)^*, X_{ij}^*\}_{i=1}^{N_j^*}\Big] \big|\mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\Big] \\
&\quad = \mathbf{E}\Big[\mathbf{E}\Big[Z_j\big|\, \{X_{ij}^*\}_{i=1}^{N_j^*}\Big] \big|\mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\Big] \\
&\quad = \mathbf{E}\Big[\pi(\mathbf{F}_j)\big|\mathbf{F}_j, \{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}\Big] = \pi(\mathbf{F}_j) = \Pr\big\{Z_j = 1\big|\mathbf{F}_j\big\}.
\end{aligned}$$

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*. $\square$

Proposition 3 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 Proofs

B.1 Theorem 1

We want to show that for any $\theta \in \tilde{A}\Theta$,

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

From the first-order Taylor's expansion of m around $A\lambda_j$,

$$\begin{aligned} \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) \right\|_2 &- \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \\ &\leq \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) - \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \\ &= \left\| \frac{1}{J} \sum_{j=1}^{J} \frac{\partial}{\partial \lambda} m\left(W_j(\lambda); \theta\right)^{\intercal} \Big|_{\lambda \in [A\lambda_j, \hat{\lambda}_j]} \left(\hat{\lambda}_j - A\lambda_j\right) \right\|_2 .^{10} \end{aligned}$$

Let $\tilde{m}$ an arbitrary component of the moment function m. Since the dimension of m is fixed, it suffices to show that

$$\frac{1}{J} \sum_{j=1}^{J} \frac{\partial}{\partial \lambda} \tilde{m}\left(W_j(\lambda); \theta\right)^{\intercal} \Big|_{\lambda \in [A\lambda_j, \hat{\lambda}_j]} \left(\hat{\lambda}_j - A\lambda_j\right) = o_p(1).$$

By applying the Cauchy-Schwarz inequality for the j-th cluster in the summation,

$$\left| \frac{\partial}{\partial \lambda} \tilde{m}\left(W_j(\lambda); \theta\right)^{\intercal} \Big|_{\lambda \in [A\lambda_j, \hat{\lambda}_j]} \left(\hat{\lambda}_j - A\lambda_j\right) \right| \leq \left\| \frac{\partial}{\partial \lambda} \tilde{m}\left(W_j(\lambda); \theta\right)^{\intercal} \Big|_{\lambda \in [A\lambda_j, \hat{\lambda}_j]} \right\|_2 \left\| \hat{\lambda}_j - A\lambda_j \right\|_2 .$$

[10] An abuse of notation is used here; $\lambda \in [A\lambda_j, \hat{\lambda}_j]$ indicates that λ lies on the line between $A\lambda_j$ and $\hat{\lambda}_j$.

By applying the Cauchy-Schwarz inequality once more,

$$\begin{aligned}
&\left|\frac{1}{J}\sum_{j=1}^{J}\frac{\partial}{\partial\lambda}\tilde{m}\left(W_j(\lambda);\theta\right)^{\intercal}\Big|_{\lambda\in[A\lambda_j,\hat{\lambda}_j]}\left(\hat{\lambda}_j-A\lambda_j\right)\right|\\
&\le\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\frac{\partial}{\partial\lambda}\tilde{m}\left(W_j(\lambda);\theta\right)^{\intercal}\Big|_{\lambda\in[A\lambda_j,\hat{\lambda}_j]}\right\|_2^2\right)^{\frac{1}{2}}\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j-A\lambda_j\right\|_2^2\right)^{\frac{1}{2}}
\end{aligned}$$

Since $\sum_{j=1}^{J}\left\|\hat{\lambda}_j-A\lambda_j\right\|_2^2=o_p(1)$, $\Pr\left\{\left\|\hat{\lambda}_j-A\lambda_j\right\|_2\le\varepsilon\ \forall j\right\}\to 1$ as $J\to\infty$, for any $\varepsilon>0$. From Assumption 2.e, the Frobenius norm of the gradient matrix is bounded in expectation when $\hat{\lambda}_j$ is close to $A\lambda_j$. From Assumption 2.f, $\left\|A^{-1}\right\|_F$ is bounded with probability going to one. Thus, $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\le\eta$ holds with probability going to one. From Assumption 2.d, we have

$$\begin{aligned}
\left\|\frac{\partial}{\partial\lambda}\tilde{m}\left(W_j(\lambda);\theta\right)^{\intercal}\Big|_{\lambda\in[A\lambda_j,\hat{\lambda}_j]}\right\|_2^2&=\left\|\frac{\partial}{\partial\lambda}\tilde{m}\left(W_j(\lambda);\tilde{A}^{-1}\theta\right)^{\intercal}\Big|_{\lambda\in[\lambda_j,A^{-1}\hat{\lambda}_j]}\right\|_2^2\\
&\le\sup_{\|\lambda'-\lambda_j\|_2\le\eta}\ \sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\lambda}m\Big(W_j(\lambda);\theta\Big)\Big|_{\lambda=\lambda'}\right\|_F^2
\end{aligned}$$

when $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\le\eta$ holds and therefore

$$\left|\frac{1}{J}\sum_{j=1}^{J}\frac{\partial}{\partial\lambda}\tilde{m}\left(W_j(\lambda);\theta\right)^{\intercal}\Big|_{\lambda\in[A\lambda_j,\hat{\lambda}_j]}\left(\hat{\lambda}_j-A\lambda_j\right)\right|=O_p(1)\cdot o_p\left(\frac{1}{\sqrt{J}}\right)=o_p(1).$$

Lastly,

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

The inequality is from the definition of the GMM estimator. The second to the last equality is from Assumption 2.c-d. From Assumption 2.d, we get

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

Then,

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

$\hat{\theta} - \tilde{A}\theta^0$ as $J \to \infty$.

B.2 Theorem 2

From the proof of Theorem 1,

$$o_p\left(\frac{1}{\sqrt{J}}\right) = \left\|\frac{1}{J}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right)\right\|_2^2 = \left\|\frac{1}{J}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2^2.$$

We can repeat the argument below for every component of m,

$$\begin{aligned}
o_p(1) &= \frac{1}{\sqrt{J}}\sum_{j=1}^{J} \tilde{m}\left(W_j;\tilde{A}\theta^0\right) + \frac{1}{\sqrt{J}}\sum_{j=1}^{J} \tilde{m}_\theta\left(W_j;\tilde{\theta}\right)^\intercal\left(\hat{\theta}-\tilde{A}\theta^0\right) \\
o_p(1) &= \frac{1}{\sqrt{J}}\sum_{j=1}^{J} \tilde{m}\left(W_j;\tilde{A}\theta^0\right) + \frac{1}{\sqrt{J}}\sum_{j=1}^{J} \tilde{m}_\theta\left(W_j;\tilde{A}\theta\right)^\intercal\left(\hat{\theta}-\tilde{A}\theta^0\right) \\
&\quad + \sqrt{J}\left(\hat{\theta}-\tilde{A}\theta^0\right)^\intercal \frac{1}{J}\sum_{j=1}^{J} \tilde{m}_{\theta\theta^\intercal}\left(W_j;\tilde{\theta}\right)^\intercal\left(\hat{\theta}-\tilde{A}\theta^0\right)
\end{aligned}$$

From the usual asymptotic argument, we have the asymptotic normality.

B.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 Assumptions 1-2,

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

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' group that cluster j would have been assigned to, when

$\mathbf{F}_j$ is observed and $\hat{G}(1), \cdots, \hat{G}(\rho)$ are given. Then,

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

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 5-a, $\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = r\}/J \xrightarrow{p} \mu(r) > 0$ as $J \to \infty$. Thus,

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

as $J \to \infty$ from Assumption 5-d and Step 1.

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

$$\begin{aligned}
&\left\|\hat{G}(\sigma(r)) - G(r')\right\|_{w,2}^{2} \\
&= \frac{J}{\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = r\}} \cdot \frac{1}{J} \sum_{j=1}^{J} \left\|\hat{G}(\sigma(r)) - G(\lambda_j) + G(\lambda_j) - G(r')\right\|_{w,2}^{2} \mathbf{1}\{\lambda_j = r\} \\
&\geq \frac{1}{2} \|G(r) - G(r')\|_{w,2}^{2} - \frac{J}{\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = r\}} \cdot \frac{1}{J} \sum_{j=1}^{J} \left\|\hat{G}(\sigma(r)) - G(\lambda_j)\right\|_{w,2}^{2} \mathbf{1}\{\lambda_j = r\} \\
&\to \frac{1}{2} c(r, r') > 0.
\end{aligned}$$

as $J \to \infty$ from the same argument from above and Assumption 5-b.

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 k. 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 $\hat{G}(r)$ is thought of as 'matched' with $G(r)$. 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 $c^* = \min_{r \neq r'} c(r, r') > 0$,

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

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. 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 Assumptions 1-2,

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

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 $\nu > 0$,

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

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

B.4 Proposition 2

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 P be a $J \times \rho$ matrix with the orthonormal eigenvectors as columns and let $\tilde{\Lambda} = \sqrt{J} P^\intercal$. Then, $\frac{1}{J}\tilde{\Lambda}\tilde{\Lambda}^\intercal = P^\intercal P = I_\rho$ and

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

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 $\tilde{\Lambda}$.

Given the rotation, let us estimate M and the eigenvectors $\tilde{\Lambda}$. For that firstly 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 from Proposition 1 and Theorem 1 of Kneip and Utikal (2001).

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

We aim to show $\hat{M}_{jk} = M_{jk} + O_p\left(\frac{1}{\sqrt{J}}\right)$. To avoid notational complexity, I will use subscript λ to note that the expectation is conditioning on 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,

$$\begin{aligned}\mathbf{E}_\lambda\left[\frac{1}{h}K\left(\frac{x - X_{ij}}{h}\right)\right] &= \int_{\mathbb{R}} \frac{1}{h}K\left(\frac{x' - x}{h}\right)\mathbf{f}_j(x')dx' = \int K(t)\mathbf{f}_j(x+th)dt \\ &= \int_{\mathbb{R}} K(t)\left(\mathbf{f}_j(x) + \mathbf{f}_j^{(1)}(x)th + \frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}t^2h^2\right)dt \\ &= \mathbf{f}_j(x) + h^2\int_{\mathbb{R}} \frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}t^2K(t)dt\end{aligned}$$

for some $\tilde{x}$ depending on x and $x + th$, from Assumption 6-a. Thus,

$$\left|\mathbf{E}_\lambda\left[\frac{1}{h}K\left(\frac{x - X_{ij}}{h}\right)\right]\mathbf{E}_\lambda\left[\frac{1}{h}K\left(\frac{x - X_{ik}}{h}\right)\right] - \mathbf{f}_j(x)\mathbf{f}_k(x)\right| \leq Ch^2$$

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

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

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

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

where $A_{ii'}=\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-X_{ij}}{h}\right)\frac{1}{h}K\left(\frac{x-X_{i'k}}{h}\right)w(x)dx$. We have that for some $l\neq i'$,

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

by letting $t=(x-x')/h$ and $s=(x'-x'')/h$.

$$\begin{aligned}\mathbf{E}_\lambda\left[A_{ii'}A_{il}\right] &= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-x'}{h}\right)\frac{1}{h}K\left(\frac{x-x''}{h}\right)w(x)dx\right)\\ &\qquad\cdot\left(\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-x'}{h}\right)\frac{1}{h}K\left(\frac{x-x'''}{h}\right)w(x)dx\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx'''\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}K\left(t\right)\frac{1}{h}K\left(t+\frac{x'-x''}{h}\right)w(x'+th)dt\right)\\ &\qquad\cdot\left(\int_{\mathbb{R}}K\left(t\right)\frac{1}{h}K\left(t+\frac{x'-x'''}{h}\right)x(x'+th)dt\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx'''\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}K\left(t\right)\frac{1}{h}K\left(t+s+\frac{x''-x'''}{h}\right)w(x''+(t+s)h)dt\right)\\ &\qquad\cdot\left(\int_{\mathbb{R}}K\left(t\right)K\left(t+s\right)w(x''+(t+s)h)dt\right)\mathbf{f}_j(x''+sh)\mathbf{f}_k(x'')\mathbf{f}_k(x''')dsdx''dx'''\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}K\left(t\right)K\left(t+s\right)w(x'''+(t+s+u)h)dt\right)\\ &\qquad\cdot\left(\int_{\mathbb{R}}K\left(t\right)K\left(t+s+u\right)w(x'''+(t+s+u)h)dt\right)\\ &\qquad\cdot\mathbf{f}_j(x''+sh)\mathbf{f}_k(x'''+uh)\mathbf{f}_k(x''')dsdudx'''\end{aligned}$$

by letting $w = (x - x')/h$, $s = (x' - x'')/h$ and $u = (x'' - x''')/h$. Thus, with some constant $C_2 > 0$ that does not depend on λ_j or λ_k, $\mathrm{Var}_\lambda(A_{ii'}) \leq C_2/h$ and $|\mathrm{Cov}_\lambda(A_{ii'}, A_{il})| \leq C_2$ and

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

Since $\min_j N_j h \to \infty$ and $\min_j N_j h^4 = O(1)$ as $J \to \infty$, we have

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

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

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 6. Note that $V_r = 0$ for $\rho < r \leq J$. Also, let $\hat{p}_r$ denote the (orthonormal) eigenvector of $\widehat{M}$ associated with the r-th eigenvalue and similarly for p_r. Recall that

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

$$\tilde{\Lambda} = \sqrt{J}P^\intercal = \sqrt{J}\begin{pmatrix}p_1 & \cdots & p_\rho\end{pmatrix}^\intercal$$

$$I_J = \begin{pmatrix}p_1 & \cdots & p_J\end{pmatrix}\begin{pmatrix}{p_1}^\intercal \\ \vdots \\ {p_J}^\intercal\end{pmatrix} = \sum_{r=1}^{J} p_r {p_r}^\intercal$$

For some $r \leq \rho$,

$$\hat{p}_r = \left(p_r{p_r}^\intercal + \sum_{r' \neq r} p_{r'}{p_{r'}}^\intercal\right)\hat{p}_r = \left({p_r}^\intercal\hat{p}_r\right)p_r + \sum_{r'\neq r} p_{r'}{p_{r'}}^\intercal\hat{p}_r.$$

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

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

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

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

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

Let $S_r = \sum_{r' \neq r} \frac{1}{V_{r'} - V_r} p_{r'} p_{r'}{}^\intercal$. S_r is well-defined from Assumption 6-b. By multiplying S_r to the equality above, we get

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

We know that $\left|\hat{V}_r - V_r\right| \leq \|\widehat{M} - M\|_2 \leq \|\widehat{M} - M\|_F$ and

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

Since $\|\hat{p}_r\|_2 = \|\hat{p}_r\|_F = (\hat{p}_r^\intercal \hat{p}_r)^{\frac{1}{2}} = 1$,

$$\begin{aligned}
\Big\| \sum_{r' \neq r} p_{r'} {p_r}^\intercal \hat{p}_r \Big\|_2 &\leq \left| \hat{V}_r - V_r \right| \|S_r \hat{p}_r\|_2 + \left\| S_r \left(\widehat{M} - M \right) \hat{p}_r \right\|_2 \\
&\leq 2 \|S_r\|_2 \left\| \widehat{M} - M \right\|_F = \frac{2\|\widehat{M} - M\|_F}{\min_{r' \neq r} |\nu_{r'} - \nu_r|} \\
&= O_p\left(\frac{1}{\sqrt{\min_j N_j}} \right).
\end{aligned}$$

The last equality holds from Assumption 6-b.

Secondly, using the same result again,

$$\begin{aligned}
\hat{p}_r^\intercal \sum_{r' \neq r} p_{r'} {p_{r'}}^\intercal \hat{p}_r &= \left(\sum_{r' \neq r} p_{r'} {p_{r'}}^\intercal \hat{p}_r \right)^\intercal \sum_{r' \neq r} p_{r'} {p_{r'}}^\intercal \hat{p}_r \\
&= \Big\| \sum_{r' \neq r} p_{r'} {p_{r'}}^\intercal \hat{p}_r \Big\|_F^2 = \Big\| \sum_{r' \neq r} p_{r'} {p_{r'}}^\intercal \hat{p}_r \Big\|_2^2 = O_p\left(\frac{1}{\min_j N_j} \right).
\end{aligned}$$

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

$$\begin{aligned}\left\| \left(\left(1 - \hat{p}_r^\intercal \sum_{r' \neq r} p_{r'} p_{r'}{}^\intercal \hat{p}_r \right)^{\frac{1}{2}} - 1 \right) p_r \right\|_2 &\leq \left| \left(1 - \hat{p}_r^\intercal \sum_{r' \neq r} p_{r'} p_{r'}{}^\intercal \hat{p}_r \right)^{\frac{1}{2}} - 1 \right| \\ &\leq \hat{p}_r^\intercal \sum_{r' \neq r} p_{r'} p_{r'}{}^\intercal \hat{p}_r = O_p \left(\frac{1}{\min_j N_j} \right)\end{aligned}$$

By combining the two bounds, we have

$$\|\hat{p}_r - p_r\|_F = O_p \left(\frac{1}{\sqrt{\min_j N_j}} \right).$$

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

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