Clustered Treatment in Multilevel Models[*]

Myungkou Shin[†]

December 12, 2023

Click here for the latest version.

Abstract

I develop a multilevel model for empirical contexts where each individual belongs to a cluster and a treatment is endogenously assigned at the cluster level. When an explanatory variable of interest is assigned at the cluster level, e.g. clustered treatment, its effect on cluster-level or individual-level outcome cannot be identified in a model with fully flexible cluster heterogeneity. To put restrictions on cluster heterogeneity, I assume that the cluster-level heterogeneity is a function of the cluster-level distribution of individual-level characteristics within each cluster. Since the distribution function is a high-dimensional object for large clusters, two functional analysis methods with dimension reduction properties are used: K-means clustering and functional PCA.

Keywords: hierarchical models, clustered treatment, 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 vast majority of datasets used in economics are multilevel; units of observations have a hierarchical structure (see (Raudenbush and Bryk, 2002) for general discussion). For example, in a dataset that collects demographic characteristics of a country's population, e.g., the Current Population Survey (CPS) of the United States, each surveyee's geographical location are also recorded, up to some regional level; in development economics, field experiments are often run at the village level and thus participants of the experiments are clustered at the village level (Voors et al., 2012; Giné and Yang, 2009; Banerjee et al., 2015).[1] Throughout this paper, I use *individual* and *cluster* to refer to the lower level and the higher level of the hierarchical structure, respectively. In light of the multilevel nature of the dataset, a researcher often considers an econometric framework that utilizes the multilevel structure. For example, when regressing individual-level outcomes on individual-level regressors with the CPS data, heterogeneity across states is often addressed with state fixed-effects or by including some state-level regressors such as population, average income, political party of the incumbent governor, etc.

The goal of this paper is to develop an econometric framework that exploits the multilevel structure, when an explanatory variable of interest, such as a treatment variable, is observed at the cluster level and an outcome variable of interest is observed at the individual level; every individual in the same cluster is under the same treatment regime. Many research topics 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

[1] The multilevel structure is not confined to datasets with a person as their unit of observation. In datasets that record market share of each product for demand estimation, products are often clustered to a product category or a market so that different brands are compared within a given product category or a market. (Besanko et al., 1998; Chintagunta et al., 2002) The Standard Industrial Classification System (SIC) and the North American Industry Classification System (NAICS) are another example of multilevel structures widely used in economics. The systems assign a specific industry code to each business establishment and they have a hierarchical system: each business establishment belongs to a finely defined industry category, which belongs to a more coarsely defined industry category, and so on. (MacKay and Phillips, 2005; Lee, 2009; De Loecker et al., 2020)

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. When a treatment variable is assigned at the cluster level, *within-cluster* variation that compares individuals from the same cluster cannot be used to identify treatment effect; every individual in a given cluster is exposed to the treatment variable in the same way. Thus, a researcher has to compare individuals from at least two different clusters, i.e. *between-cluster* variation. In order to use *between-cluster* variation instead of *within-cluster* variation, restrictions on cluster-level heterogeneity need to be made. In a model with fully flexible cluster-level heterogeneity, cluster heterogeneity and treatment effect cannot be separated; the researcher cannot know whether the difference between given two clusters comes from their cluster-level heterogeneity or the difference in the cluster-level variable of interest.[2] Thus, we need restrictions on cluster-level heterogeneity.

To impose restrictions on cluster-level heterogeneity, I assume that the observable information for each individual, aggregated at the cluster level, is sufficiently rich that the cluster-level heterogeneity can be controlled for using that information. In particular, I aggregate the individual-level information at the cluster level by looking at within-cluster *distribution* of the individual-level covariates. Then, conditioning on the cluster-level distribution of individual-level covariates, the clusters are assumed to be homogeneous. Thus, by comparing clusters with the same distribution of individual-level covariates, the effect of the cluster-level explanatory variable of interest is identified. The motivation for using the distribution function as a control variable comes from the selection-on-observable assumption in the program evaluation literature. The main purpose of the selection-on-observable assumption that treatment is random conditioning on some observable control covariates is

[2] The cluster-level heterogeneity problem discussed in this paper is a treatment endogeneity/selection bias problem in a sense. If the cluster-level explanatory variable of interest is independent of the cluster-level heterogeneity, its effect is identified without controlling for cluster-level heterogeneity since the cluster-level heterogeneity will be balanced across different levels of the cluster-level explanatory variable.

to control for treatment endogeneity. To implement the selection-on-observable approach in a multilevel model where the treatment is assigned at the cluster level, a researcher would want to gather all the available information for each cluster since clusters are the units of treatment assignment. When the clusters are large, i.e. there are many individuals in each cluster, the cluster-level collection of the individual-level information is high-dimensional even when the individual-level control covariate X_{ij} is low-dimensional; the model induced by the selection-on-observable is not parsimonious.

Thus, I impose additional restrictions on the individual-level observable information. Let X_{ij} denote the individual-level control covariates for individual i in cluster j and N_j denote the number of individuals for cluster j. We are interested in cases where N_j is large. Firstly, I assume exchangeability within a cluster: the distribution of individuals within a cluster is invariant up to permutation on labeling. By assuming exchangeability, the names of each individual in a given cluster do not have any additional information in terms of treatment assignment.[3] Thanks to this condition, I can substitute the potentially high-dimensional object $\{X_{ij}\}_{i=1}^{N_j}$, with an empirical distribution of X_{ij} for each cluster:

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

By shifting from $\{X_{ij}\}_{i=1}^{N_j}$ to $\hat{\mathbf{F}}_j$, the dimension of the control variable reduces down.[4] Secondly, to have further dimension reduction, I assume that the expectation of $\hat{\mathbf{F}}_j$ given some cluster-level latent factor λ_j contains all the relevant information for treatment assignment. Consider $\mathbf{F}_j$ such that for all $x \in \mathbb{R}^p$

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

[3] For a formal statement in terms of potential outcomes, see Appendix.

[4] To illustrate this in a simpler setting, consider an one-dimensional X_{ij}. Then, $\hat{\mathbf{F}}_j$ has a one-to-one mapping to the vector of ordered statistics. By shifting from $\{X_{ij}\}_{i=1}^{N_j}$ to the ordered statistics, the support for the control variable reduces down.

One can think of the distinction between the empirical distribution function $\hat{\mathbf{F}}_j$ and the true distribution function $\mathbf{F}_j$ as observed value with noise and signal. Since the clusters are assumed to be large in the paper, the noise will disappear as the number of individuals per clusters grows. Lastly, I assume that λ_j is a low-dimensional factor and propose two different models for G, based on the two dimension reduction methods I apply to the distribution functions: the K-means clustering and the functional PCA. With these three layers of dimension reduction, the potentially high-dimensional object is reduced down to a finite-dimensional latent factor λ_j. In implementation, I apply the K-means clustering to the empirical distribution function and the functional PCA to the kernel density estimation and use the outcome of the two algorithms as my estimate for the latent factor.

To discuss my main theoretical results, I characterize a class of moment condition models where the model parameter and the latent factor for the cluster-level distribution $\mathbf{F}_j$ can rotate simultaneously. For the class of models, we do not need to estimate the latent factor perfectly; we only need to estimate some linear rotation of the latent factors. For both of the estimation strategies, the K-means clustering and the functional PCA, an interpretable distributional model can be constructed. For the K-means clustering estimator, I assume that the cluster-level heterogeneity is finitely discrete; there are only finite types of clusters. In addition, I assume that the finite types are well separated in terms of the distribution function $\mathbf{F}_j$. For the functional PCA, I assume that each cluster is made up of a finite types of individuals. Also, I assume that there is sufficient variation across the types of individuals in terms of their underlying density functions. Under these conditions, both estimation strategy estimate the latent factors fast enough that the plug-in estimator using the estimates is consistent.

As an empirical illustration, I apply the econometric framework proposed in this paper to revisit the disemployment effect of the minimum wage on teenagers. Using the econometric framework of this paper, I address aggregate heterogeneity in state-level labor market fundamentals by controlling for the distribution of individual employment status history and the distribution of wage income. Also, I explore how the two channels of individual heterogeneity—age and race—interact with the aggregate heterogeneity. 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.

1.1 Related literature

This paper contributes to several literatures in econometrics. Firstly, this paper contributes to the treatment effect and program evaluation literature. This paper is the first to use a selection-on-observable type assumption in solving the treatment endogeneity problem of a clustered treatment. Arkhangelsky and Imbens (2022); Hansen et al. (2014) use similar selection-on-observable type assumptions at the cluster level but Arkhangelsky and Imbens (2022) focus on individual-level treatment and Hansen et al. (2014) take pairs of comparable clusters as given. Also, by using both cluster-level distribution and individual-level control covariates, this paper models treatment effect to have two types of heterogeneity: aggregate heterogeneity from the cluster-level distribution and individual heterogeneity from the individual-level control covariates. With these two types of heterogeneity in treatment effect, the econometric framework of this paper answers a variety of novel research questions. For example, suppose a researcher is interested in how neighborhood of residence or migration affects individual outcomes, as in Derenoncourt (2022); Chetty et al. (2016). In the framework of this paper, a researcher can answer questions such as "what demographic characteristic of an individual makes migration successful?", "does the demographic composition of a destination neighborhood matter?", and "does individual-level demographic characteristic interact with the demographic composition of the destination?" by looking at individual heterogeneity, aggregate heterogeneity, and interactive hetergoeneity in treatment effect, respectively.

Secondly, this paper contributes to the literature of regression with heterogeneous slopes, and particularly to the group fixed-effect literature. Whereas the group fixed-effect literature mostly focuses on panel data and assumes a finite grouping structure on unit-specific fixed effects, I apply the idea of a finite grouping structure to a cross-sectional multilevel model. A key difference of the grouping approach in this paper from most of the group fixed-effect literature is that the grouping structure is not recovered from the LHS of the outcome model (Bonhomme and Manresa, 2015; Su et al., 2016; Ke et al., 2016; Wang and Su, 2021), but from the RHS of the outcome model, the individual-level control covarites X_{ij}. In this sense, Pesaran (2006) is comparable to this paper. Both papers use the information from the RHS of the equation to recover the slope heterogeneity.

In addition, there are several literatures that my paper relates to. Firstly, both latent factor models used in this paper are essentially a variant of the factor model: Abadie et al. (2010, 2015); Bai (2009). With a factor model, a linearity is imposed on a potentially high-dimensional time-series of observable control covariates whereas in this paper exchangeability is imposed on individuals within a cluster. In the case of panel data, the time dimension, the label of observations within each unit, conveys significant information; thus, exchangeability is not desirable. However, in the case of multilevel data, the individual identity, the label of observations within each cluster, has little information. Secondly, Auerbach (2022); Zeleneev (2020) discuss a dataset with network structure and suggest matching units based on the observable information, such as network links, to control for heterogeneity in the outcome model. The idea of using the particular structure of dataset in hand and using the observable information to control for latent heterogeneity is present in both this paper and their works.

The rest of the paper is organized as follows. In Section 2, I formally discuss the model with the *selection-on-distribution* assumption. In Section 3, I explain the K-means algorithm and the treatment effect estimators. In Section 4, I discuss asymptotic properties of the estimators, under the finiteness assumption. In Section 5, the empirical illustration of the econometric framework is provided.

2 Distribution as control variable

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. The econometrician is interested in estimating the effect of Z_j on Y_{ij}.

In this paper, I assume that the individuals are independent and identically distributed within clusters and the clusters are independent and identically distributed. To characterize the cluster-level heterogeneity, I consider an additional random object that is not observed in the dataset: the cluster-level distribution of X_{ij}.

Assumption 1. *(iid-ness within and across clusters)* $\mathbf{F}_j$ *is a cluster-level* p*-dimensional random field. Then,*

a. $\left(Z_j, N_j, \mathbf{F}_j\right) \sim$ *iid.*

b. $\Pr\{\mathbf{F}_j$ *is a well-defined distribution function*$\} = 1$.

c. *For each* j,

$$\left(Y_{ij}, X_{ij}\right) \mid \left(Z_j, N_j, \mathbf{F}_j\right) \overset{iid}{\sim} H(Z_j, N_j, \mathbf{F}_j),$$
$$X_{ij} \mid \left(Z_j, N_j, \mathbf{F}_j\right) \overset{iid}{\sim} \mathbf{F}_j.$$

The model that I consider in Assumption 1 imposes restriction on cluster-level heterogeneity in the sense that the cluster-level observable information (Z_j, N_j) and the distribution of individual-level observable information X_{ij} sufficiently control the cluster-level heterogeneity in terms of the joint distribution of (Y_{ij}, X_{ij}): H is not subscripted with j. In addition, from iid-ness within cluster, it is assumed that there is no spillover across individuals within each cluster after conditioning on the random distribution function $\mathbf{F}_j$.

When we are interested in estimating the effect of X_{ij} on Y_{ij}, we may not need such restrictions on the cluster-level heterogeneity; given sufficient variation in X_{ij} within a cluster, cluster-level distribution of (Y_{ij}, X_{ij}) may identify the effect of X_{ij} on Y_{ij}. However, in this paper, I focus on cases where a researcher is interested in the effect of a cluster-level variable Z_j on individual-level outcome variable Y_{ij}. Thus, abstracting away from the cluster-level heterogeneity is infeasible. Let us consider a very simple example of a regression model:

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

$$Y_{ij} = \tilde{\alpha}_j + X_{ij}{}^{\intercal}\theta + U_{ij}. \tag{2}$$

Suppose that the true model is Equation (1). The cluster fixed-effect α_j models unrestricted cluster-level heterogeneity in the level of Y_{ij}. Due to the multicollinearity problem with α_j, β is not identified while θ is still identified by using a reduced Equation (2) and letting $Z_j{}^{\intercal}\beta$ be subsumed in the cluster fixed-effect $\tilde{\alpha}_j$. The sample problem exists in more complicated multilevel models as well. Thus, I impose the restriction on the cluster-level heterogeneity as in Assumption 1-c and use the distribution function as a control variable.

The use of the cluster-level distribution $\mathbf{F}_j$ as a control variable to model the cluster-level heterogeneity is sensible in many empirical contexts. When given a clustering structure where the clusters are large, a simple collection of the individual-level information will be very high dimensional: $\{X_{ij}\}_{i=1}^{N_j}$. Also, there is often no natural ordering of the individuals within a cluster: e.g. students in a school, workers at a firm, individuals in a neighborhood. In most cases, it is only the distribution that matters at the cluster level. Let us consider the main empirical example of the minimum wage in the United States. We would think that the decision makers of the minimum wage look at the distribution of wage income rather than selected specific individuals, when they decide on the minimum wage level. This does not mean that they do not care about heterogeneity across individuals; when X_{ij} contains socioeconomic characteristics such as wage income as well as demographic characteristics such as race and age, the distribution of X_{ij} also has information such as racial gap in wage income distribution. Assumption 1-c formalizes this argument and assumes that it is only the distribution of individual-level characteristics that matters for the cluster-level heterogeneity.

To model the effect of Z_j on Y_{ij} in a general way, I consider a finite-dimensional treatment effect parameter β and assume that β and a nuisance parameter θ are identified with a momenct function m: at true values of β and θ,

$$\mathbf{E}\left[m(W_j;\beta^0,\theta^0)\right]=0. \tag{3}$$

W_j is a function of cluster-level random objects $\left(\{Y_{ij},X_{ij},\}_{i=1}^{N_j},Z_j,\mathbf{F}_j\right)$. The leading example considered throughout the paper is a binary treatment assigned at the cluster level where the treatment is random after conditioning on the cluster-level distribution of individual-level control covariates. Let $Z_j\in\{0,1\}$ and

$$Y_{ij}=Y_{ij}(1)\cdot Z_j+Y_{ij}(0)\cdot(1-Z_j). \tag{4}$$

In addition, assume conditional independence of Z_j given N_j and $\mathbf{F}_j$:

$$\{Y_{ij}(1),Y_{ij}(0),X_{ij}\}_{i=1}^{N_j}\;\Big|\;(Z_j,N_j,\mathbf{F}_j)\overset{iid}{\sim}H^*(N_j,\mathbf{F}_j). \tag{5}$$

H^* contains additional information compared to H: H^* tells us how $Y_{ij}(1)$ and $Y_{ij}(0)$ depend on each other. Suppose that $\mathbf{F}_j$ is directly observed for now; in practice, $\mathbf{F}_j$ is only indirectly observed through $\{X_{ij}\}_{i=1}^{N_j}$. Then, the average treatment effect (ATE) is identified as follows:

$$\beta=\mathbf{E}\left[\bar{Y}_j(1)-\bar{Y}_j(0)\right]=\mathbf{E}\left[\mathbf{E}\left[\bar{Y}_j|Z_j=1,N_j,\mathbf{F}_j\right]-\mathbf{E}\left[\bar{Y}_j|Z_j=0,N_j,\mathbf{F}_j\right]\right].$$

We can rewrite the identification result in the context of Equation (3): with some known function π,

$$\begin{aligned}
W_j &= \left(\bar{Y}_j, Z_j, N_j, \mathbf{F}_j\right)^\intercal, \\
\mathbf{E}[Z_j|N_j, \mathbf{F}_j] &= \pi(N_j, \mathbf{F}_j; \theta), \\
m_1(W_j; \beta, \theta) &= \left(\frac{Z_j}{\pi(N_j, \mathbf{F}_j; \theta)} - \frac{1 - Z_j}{1 - \pi(N_j, \mathbf{F}_j; \theta)}\right)\bar{Y}_j - \beta.
\end{aligned}$$

Only one component of m regarding β is given above. Likewise, the conditional average treatment effect (CATE) can be identified as well:

$$\begin{aligned}
\mathbf{E}[Y_{ij}(1) - Y_{ij}(0)|X_{ij}, N_j, \mathbf{F}_j] &= \mathbf{E}[Y_{ij}(1)|X_{ij}, Z_j = 1, N_j, \mathbf{F}_j] - \mathbf{E}[Y_{ij}(0)|X_{ij}, Z_j = 0, N_j, \mathbf{F}_j] \\
&= \mathbf{E}[Y_{ij}|X_{ij}, Z_j = 1, N_j, \mathbf{F}_j] - \mathbf{E}[Y_{ij}|X_{ij}, Z_j = 0, N_j, \mathbf{F}_j].
\end{aligned}$$

The first equality holds since the joint distribution of $(Y_{ij}(1), Y_{ij}(0), X_{ij})$ is independent of Z_j conditioning on $(N_j, \mathbf{F}_j)$. A connection to Equation (3) can be made here as well. Fix some $(x, n, \mathbf{F})$. Then, the identification of the CATE parameter

$$\beta(x, n, \mathbf{F}) = \mathbf{E}[Y_{ij}(1) - Y_{ij}(0)|X_{ij} = x, N_j = n, \mathbf{F}_j = \mathbf{F}] \tag{6}$$

can be rewritten with the following moment function:

$$\begin{aligned}
W_j &= \left(\bar{Y}_j(x), Z_j, N_j, \mathbf{F}_j\right)^\intercal, \\
\bar{Y}_j(x) &= \frac{\sum_{i=1}^{N_j} Y_{ij}\mathbf{1}\{X_{ij} = x\}}{\mathbf{f}_j(x)}, \\
\theta = (\theta_1, \theta_2) &= \Big(\mathbf{E}[Z_j|N_j = n, \mathbf{F}_j = \mathbf{F}], \Pr\{N_j = n, \mathbf{F}_j = \mathbf{F}\}\Big) \\
m_1(W_j; \beta, \theta) &= \left(\frac{Z_j}{\theta_1} - \frac{1 - Z_j}{1 - \theta_1}\right) \cdot \frac{\mathbf{1}\{N_j = n, \mathbf{F}_j = \mathbf{F}\}}{\theta_2} \cdot \bar{Y}_j(x) - \beta.
\end{aligned}$$

For simplicity, the individual-level control covariate X_{ij} and the cluster-level distribution function $\mathbf{F}_j$ are treated as discrete random variables and only the component of m that is relevant for β is given above.

The CATE parameter as defined in (6) is particularly useful in discussing treatment effect heterogeneity. The CATE parameter has both individual-level information and the cluter-level information in the conditioning set: X_{ij} and $\mathbf{F}_j$. Thus, it captures treatment effect heterogeneity at both levels. With $\beta(x, n, \mathbf{F}) - \beta(x', n, \mathbf{F})$ for some $x \neq x'$, we capture the individual-level treatment effect heterogeneity; with $\beta(x, n, \mathbf{F}) - \beta(x, n, \mathbf{F}')$ for some $\mathbf{F} \neq \mathbf{F}'$, we capture the aggregate-level treatment effect heterogeneity. Moreover, by taking double differences, we look at how the individual-level treatment effect heterogeneity interacts with the aggregate-level treatment effect heterogeneity. In this sense, the construction of the CATE parameter in (6) is true to the multilevel nature of the datasets.

Many empirical contexts benefit from this multilevel construct of the CATE parameter. Let us go back to the empirical example of minimum wage in the United States. The disemployment effect of minimum wage may depend on both individual-level characteristics such as education level or age and aggregate-level characteristics such as labor market status of the state. More importantly, it may depend on both; the minimum wage may affect the same low-skilled worker differently depending on the wage income levels of the state that they live in, whereas it does not affect high-skilled workers at all, regardless of their location. The construction of $\beta(x, n, \mathbf{F})$ allows for this discussion and is consistent with the empirical practice that often includes interaction terms between some control covariates and the treatment variable in a regression specification.

There are two hardships in applying the GMM method directly, when a distribution function $\mathbf{F}_j$ is used as a control variable: $\mathbf{F}_j$ is not directly observed and infinite-dimensional. For that I assume that $\mathbf{F}_j$ is a function of a finite-dimensional cluster-level factor.

Assumption 2. *(cluster-level factor) There exists some cluster-level latent factor* $\lambda_j \in \Lambda \subset$

$\mathbb{R}^\rho$ and an injective function $G : \Lambda \to [0,1]^{\mathbb{R}^p}$ such that

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

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* $\Pr\{\|G(\lambda_j)\|_{w,2} < \infty\} = 1$

From the injectivity of G, we can construct an inverse G^{-1} such that

$$\Pr\left\{\lambda_j = G^{-1}(\mathbf{F}_j)\ \forall j\right\} = 1.$$

Assumption 1-a,c hold by replacing $\mathbf{F}_j$ with λ_j. By assuming that the cluster-level distribution of the individual-level control covariates is a function of a finite-dimensional cluster-level factor, I reduced the dimension of the distribution function. Note that the injectivity of G allows us to repeat Assumption 2 for any monomorphism on Λ. For example, consider an invertible $\rho \times \rho$ matrix A and the transformed latent factor $\tilde{\lambda} \in A\Lambda$. By letting $G_A(\tilde{\lambda}) = G(A^{-1}\tilde{\lambda})$, we have Assumption 2 hold for G_A as well.

Now, I present the hypertheorem which can be used in the distributional control model described with Assumptions 1-2. Consider a function W which takes cluster-level observable variables and the latent factor λ_j and computes the observation relevant for the moment condition model m. Let

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

There is a slight abuse of notation here since the dimension of the argument depends on N_j, which is a random variable. Since the cluster-level factor λ_j is latent, $W_j(\lambda)$ takes the factor as an input and compute W. Let $W_j^* = W_j(\lambda_j)$ be the infeasible true observation for cluster j. The moment condition model (3) holds for W_j^*. For notational simplicity, I use θ to denote the vector of both the treatment effect parameter and the nuisance parameter. Let l denote the dimension of m and k denote the dimension of θ: $l \geq k$.

Assumption 3. *There is an (random) invertible $\rho \times \rho$ matrix A. Assume*

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

The true value of θ 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 $J \to \infty$.*

d. *There is a function that maps A to an invertible matrix $\tilde{A}$ such that $W_j = W_j(A\lambda_j)$ satisfies*

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

almost surely.

e. *The map $\lambda \mapsto m(W_j(\lambda);\tilde{A}\theta)$ is almost surely continuously differentiable and there is some $\eta, M > 0$ such that*

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

Theorem 1. *Assumptions 1-3 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, with some invertible matrix $\tilde{A}$,

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

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-a,b,c are the usual sufficient conditions for consistency of an extremum estimator. Assumption 3-d discusses the rotation invariance of the model. From Assumption 3-e, the first derivative of the moment function with regard to the latent factor, evaluated at the estimated latent factor, is bounded in expectation when the estimation error is small. Theorem 1 has the consistency result.

Assumption 4. *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*

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

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 1-4 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, with some invertible matrix $\tilde{A}$,

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

as $J \to \infty$, where

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

Proof. See Appendix. □

3 Latent factor models for distribution

In the previous section, I have not discussed what the latent factor λ_j means and how to construct a consistent estimator for the latent factor. In this section, I discuss two latent factor models for $G : \Lambda \to [0,1]^{\mathbb{R}^p}$, which give us some interpretation on λ_j, and construct estimators for λ_j. Note that the two models discussed here are not the only models with estimators satisfying Assumptions 3-4.

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 rotations A and $\tilde{A}$; the hypertheorems are confined to models that are invariant to some rotation of the latent factor and the model parameter. The value of λ_j in and of itself does not matter. Recall that the purpose of assuming Assumption 2 is to reduce the dimension of the distribution function $\mathbf{F}_j$. In the field of unsupervised learning in machine learning, many algorithms that summarize patterns of high-dimensional data such as distributions have been proposed. In most cases, the low-dimensional output of such algorithms by itself is not readily interpretable. Therefore, it is difficult to directly use the output as an estimator for λ_j and develop an econometric model where the estimator is consistent for the true latent factor. To bypass this, the rotation invariance is imposed.

In this section, I discuss two examples of such algorithms and their associated data generating process assumptions: K-means clustering and functional principal component analysis (functional PCA).

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 finds a fixed number of centeroids such that the sum of the distance between a data point and its closest centeroid is minimized. In this paper, a data point that the K-means minimization problem takes 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 Assumptions 1-2 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 $\hat{\mathbf{F}}_j$:

$$\left(\hat{\lambda}_1, \cdots, \hat{\lambda}_J, \hat{G}(1), \cdots, \hat{G}(\rho)\right) = \arg\min_{\lambda, G} \sum_{j=1}^{J} \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^{2}. \tag{7}$$

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 $\left\|\hat{\mathbf{F}}_j - G(\lambda)\right\|_{w,2}$. The solution to (7) maps $\hat{\mathbf{F}}_j$ to $\hat{\lambda}_j$, a discrete variable with finite support: dimension reduction.

To solve (7), I use the (naive) K-means clustering algorithm or Lloyd's 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 $\left(\lambda_1^{(0)}, \cdots, \lambda_N^{(0)}\right)$,

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.

There is no guarantee that the result of the iterative algorithm is indeed the global minimum. For simplicity of the discussion, let the weighting function w in $\|\cdot\|_{w,2}$ be discrete and finite: with some $x^1, \cdots, x^d \in \mathbb{R}^p$,

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

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 $O(J\rho d)$. Thus, the iterative algorithm gives us computational gain, at the cost of not being able to guarantee the global minimum.[5] 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 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 (7)

[5] 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.

is still attained.

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 Theorem 1. Assumption 5 discusses a set of conditions for that.

Assumption 5. *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 5-a ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. Under Assumption 5-b, clusters with different values of the 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 good estimator for $\mathbf{F}_j$. Assumption 5-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 5-c excludes cases where the size of cluster increases only for some clusters and is fixed for some other clusters; the estimation of $\hat{\mathbf{F}}_j$ improves uniformly as J increases.

One big restriction that the econometric model described in Assumption 5 imposes on the cluster-level heterogeneity is that there is a discrete grouping structure on clusters, in terms of their distribution of individual-level control covariates X_{ij}. The latent factor $\lambda_j \in \Lambda$ is a categorical variable indicating which group each of the clusters belongs to; the face value of λ_j does not have any information on cluster j, other than that it tells us which of the remaining clusters belongs to the same group with cluster j. Within each group, the clusters with the same cluster-level control covariates Z_j are homogeneous in terms of the joint distribution of (Y_{ij}, X_{ij}). Another big restrictions that Assumption 5 makes are that the number of groups is fixed even when the number of clusters grows and that the groups are well-separated even at the margin: Assumption 5-b. Thus, using Assumption 5 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, 5 hold. Then, there is a transition 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 together 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. Here are two examples.

Example 1 *(ATE with group-specific treatment propensity)* Let $Z_j \in \{0, 1\}$ and assume (4) and (5). A moment condition model can be established for ATE:

$$
\begin{aligned}
W_j &= \left(\bar{Y}_j, Z_j, \mathbf{F}_j\right)^\intercal, \\
\mathbf{E}[Z_j|N_j, \lambda_j] &= {\lambda_j}^\intercal \theta, \\
m(W_j; \beta, \theta) &= \begin{pmatrix} \left(\frac{Z_j}{{\lambda_j}^\intercal\theta} - \frac{1-Z_j}{1-{\lambda_j}^\intercal\theta}\right)\bar{Y}_j - \beta \\ Z_j \mathbf{1}\left\{\lambda_j = (1,0,\cdots,0)^\intercal\right\} - \theta_1 \mathbf{1}\left\{\lambda_j = (1,0,\cdots,0)^\intercal\right\} \\ \vdots \\ Z_j \mathbf{1}\left\{\lambda_j = (0,\cdots,0,1)^\intercal\right\} - \theta_\rho \mathbf{1}\left\{\lambda_j = (0,\cdots,0,1)^\intercal\right\} \end{pmatrix}.
\end{aligned}
$$

The nuisance parameter $\theta = (\theta_1, \cdots, \theta_\rho)^\intercal$ is the group-specific propensity to treatment. With some overlap condition, i.e. $\Theta \subset [\varepsilon, 1-\varepsilon]^\rho$, Assumption 3 is satisfied.

Example 2 *(Group fixed-effects regression)* Consider a regression model:

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

${\lambda_j}^\intercal\theta_1$ is the group fixed-effect that controls for the cluster-level heterogeneity.

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

Note that both θ_2 and β do not change for any rotation A applied to λ_j. The fact that we cannot get the order of the groups from the estimated latent factor $\hat{\lambda}_j$ correctly does not stop us from estimating the parameter of interest β.

3.2 Functional principal component analysis

The functional PCA is an extension of the PCA technique, often applied to large matrices, to functional dataset. Given J functions, the functional PCA computes their covariance matrix and apply the eigenvalue decomposition to the covariance 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 is used. Again, the density functions are not directly observed. Thus, we compute the covariance matrix using kernel estimation. Given some kernel K and bandwidth h,

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

$\hat{M}$ 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.

Then, apply the eigenvalue decomposition to 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 cluster and the value of the r-th eigenvalue denotes the magnitude of the corresponding dimension. Thus, with some predetermined $\rho \leq J$, taking the first ρ largest eigenvectors 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:

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

The rescaling is introduced so that the estimated latent factor $\hat{\lambda}_j$ does not converge to zero as J grows. 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.

Assumption 6. *Assume with some constant $C > 0$,*

a. *(finite mixture model for distribution) There are thrice continuously differenciable distribution 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 6-a assumes that the cluster-level distribution function $\mathbf{F}_j$ a mixture of ρ underlying distributions $G_1, \cdots, G_\rho$ and the latent factor λ_j is the mixture weights across the distributions. In addition, it assumes that the underlying density functions $g_1, \cdots, g_\rho$ are smooth and bounded, up to third derivative. Under Assumption 6-a, the covariance matrix M can be rewritten as follows:

$$\begin{aligned} M_{jk} &= \int_{\mathbb{R}} \mathbf{f}_j(x)\mathbf{f}_k(x)w(x)dx \\ &= \sum_{r,r'} \lambda_{jr}\lambda_{kr'} \int_{\mathbb{R}^p} g_r(x)g_{r'}(x)w(x)dx \\ M &= \begin{pmatrix} \lambda_1{}^\intercal \\ \vdots \\ \lambda_J{}^\intercal \end{pmatrix} \underbrace{\begin{pmatrix} \int_{\mathbb{R}} g_1(x)^2 w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)g_1(x)w(x)dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}} g_1(x)g_\rho(x)w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)^2 w(x)dx \end{pmatrix}}_{=:V} \begin{pmatrix} \lambda_1 & \cdots & \lambda_J \end{pmatrix}. \end{aligned}$$

Assumption 6-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 constructed with $\{g_r\}_{r=1}^{\rho}$.

When combined with a moment condition model as described in Assumption 3, the econometric model described in Assumption 6 imposes some linearity condition across the model for the cluster-level distribution and the moment condition model. The latent factor λ_j that characterizes the mixture weights of the underlying distribution that make up $\mathbf{F}_j$ also enters the the moment condition model (3) linearly. This is in contrast to the grouping structure model that was described in Assumption 5; the distribution function $\mathbf{F}_j$ and the moment condition model can be connected in an arbitrary way as long as the grouping structure was consistent aross the two models. To see this linearity more clearly, consider an arbitrary function on $\mathbb{R}^p$, ϕ. Find that

$$\int_{\mathbb{R}} \phi(x)\mathbf{f}_j(x)w(x)dx = \sum_{r=1}^{\rho} \lambda_{jr} \underbrace{\int_{\mathbb{R}} \phi(x)g_r(x)w(x)dx}_{=:\theta_r} = \lambda_j{}^\intercal\theta.$$

Being linear in the latent factor λ_j is equivalent with being linear in the density function; the moment condition model can only admit linear functions of the density $\mathbf{f}_j$. Thus, using Assumption 6 to model the cluster-level heterogeneity is more suitable when the heterogeneity across clusters is continuous and the cost of being linear in density is thought to be small.

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

Proposition 2. *Assumptions 1-2, 6 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 where the eigenvectors of M with nonzero eigenvalues are rows and $A^\intercal = V\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right) diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1}$. Then,

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

Proof. See Appendix. □

Proposition 2 shows that the estimation error goes to zero at the same rate with the root ratio of the number of clusters to the smallest cluster size.

Example 1 discussed in the previous subsection still applies to the functional PCA estimated latent factor. However, in the case of the functional PCA, the propensity score will not be group-specific, but linear in the density function $\mathbf{f}_j$. A sufficient overlap condition is

$$\Pr\left\{\varepsilon \leq {\lambda_j}^\intercal\theta \leq 1 - \varepsilon\right\} = 1$$

with some $\varepsilon > 0$.

Example 3 *(Linear-in-density regression)* Consider a regression model:

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

$\lambda_j{}^\intercal\theta_1$ is linear term in the density $\mathbf{f}_j$ that controls for the cluster-level heterogeneity.

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

Again, note that both θ_2 and β do not change for any rotation A applied to λ_j. Though the two models from Example 2 and 3 look exactly the same, they differ significantly in terms of modelling the cluster-level heterogeneity. The group fixed-effect regression model in Example 2 assumes that the cluster-level heterogeneity is discrete and some clusters are perfectly homogeneous. The linear-in-density regression model in Example 3 allows the cluster-level heterogeneity to be continuous and allows every pair of clusters to be heterogeneous, to some extent. However, the flexibility in the linear-in-density model comes at the cost of assuming that the cluster-level heterogeneity is linear in the cluster-level density $\mathbf{f}_j$.

Surely, the K-means clustering and the functional PCA are not the only two options in reducing the dimension of a distribution function. A dimension reduction method widely studied in Econometrics is regularized regression with variable selection property: e.g. LASSO (Tibshirani, 1996). Set $p = 1$ for brevity and let $\mu_k(\mathbf{F})$ be the k-th moment of some random vector X such that $X \sim \mathbf{F}$. With some large $\rho_{\max} \gg J$, we can consider regression specifications such as

$$
Y_{ij} = \begin{pmatrix} \mu_1(\hat{\mathbf{F}}_j) & \cdots & \mu_{\rho_{\max}}(\hat{\mathbf{F}}_j) \end{pmatrix}\theta_1 + X_{ij}{}^\intercal\theta_2 + Z_j{}^\intercal\beta + U_{ij}
$$

with l_1 penalty on $\mu_r(\hat{\mathbf{F}}_j)$. Suppose LASSO selects ρ variables:

$$\mu_{r_1}(\hat{\mathbf{F}}_j), \cdots, \mu_{r_\rho}(\hat{\mathbf{F}}_j)$$

Then, the variable selection property has reduced the dimension from the $\rho_{\max} \times 1$ vector to a $\rho \times 1$ vector and selected the moments of X_{ij} that are relevant in explaining the variation of Y_{ij}. However, the regularized regression approach is fundamentally different from the two approached discussed here since it also uses information from the outcome variable Y_{ij} while the two approaches discussed here only uses information from X_{ij}; the K-means clustering and the functional PCA are a 'summary' of the distribution function $\mathbf{F}_j$ in a truer sense.

4 Empirical illustration: disemployment effect of minimum wage

4.1 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).

In addition to the existing approaches, I would like to use the state-level distribution of individual-level information and allow for more flexible patterns of heterogeneity across states. The multilevel model with clustered treatment described 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 in that state. This corresponds to Assumption 1. Thus, I believe the latent factor models for the cluster-level distribution and the corresponding estimation strategy suggested in this paper are a naturally appealing approach when studying the effect of the minimum wage.

4.2 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,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}{}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{8}$$

$logMinWage_{jt}$ 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: $EmpRate_{jt} = 1/|\tilde{\mathcal{I}}_{jt}| \sum_{i \in \tilde{\mathcal{I}}_{j}t} Y_{ijt}$. 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 (8) 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, (8) 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 (8), 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 $\mathbf{F}_{jt}$:

$$Y_{ijt} = \alpha_j + \lambda_{jt}^\intercal \delta_t + \beta \log MinWage_{jt} + X_{ijt}^\intercal \theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{9}$$

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.

4.3 Results

4.3.1 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:

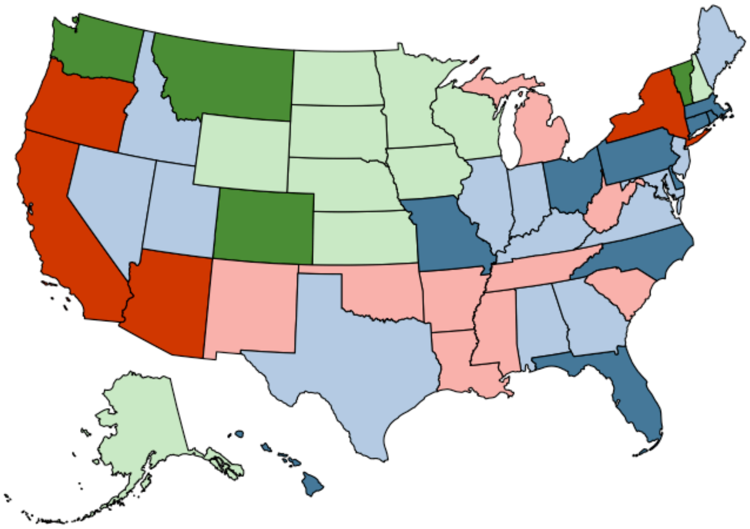

Figure 1: Grouping of states when $\rho_{Kmeans} = 3$, January 2007

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: $\{Emp\}^4$
- Ever-unemployed: $\{Emp, Unemp\}^4 \setminus \big(Emp, Emp, Emp, Emp\big)$
- Never-in-the-labor-force: $\{NotInLaborForce\}^4$

'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

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.

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.

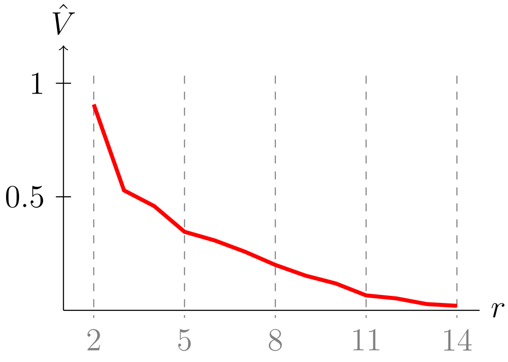

Figure 2: The scree plot of eigenvalues from the wage income distribution, March 2007

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.

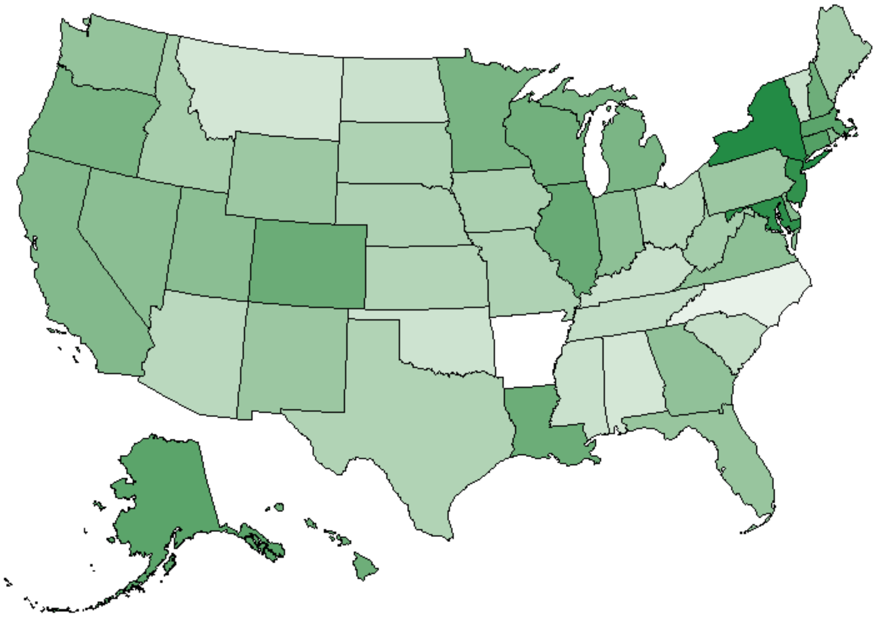

Figure 3: $\hat{\lambda}_{j2}$ across states for March 2007

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 $h = 10$.

4.3.2 Disemployment effect regression

Now, I discuss the regression results from (9). 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 (9). 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:

$$Y_{ijt} = \alpha_j + {\lambda_{jt}}^\intercal\delta_t + \beta(\lambda_{jt})\log MinWage_{jt} + {X_{ijt}}^\intercal\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \qquad (10)$$

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 (9)-(10) to discuss individual heterogeneity and aggregate heterogeneity simultaneously. The left panel of Table 4 estimates

$$\begin{aligned} Y_{ijt} = \alpha_j + {\lambda_{jt}}^\intercal\delta_t + \beta_{yt}\log MinWage_{jt}\mathbf{1}\{Age_{ijt} \leq 18\}. \\ + \beta_{ot}\log MinWage_{jt}\mathbf{1}\{Age_{ijt} = 19\} + {X_{ijt}}^\intercal\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \qquad (11) \end{aligned}$$

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

$$
\begin{aligned}
Y_{ijt} = \alpha_j + \lambda_{jt}{}^{\intercal}\delta_t + \beta_{yt}(\lambda_{jt}) \log MinWage_{jt}\mathbf{1}\{Age_{ijt} \leq 18\}. \\
+ \beta_{ot}(\lambda_{jt}) \log MinWage_{jt}\mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}{}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (12)
\end{aligned}
$$

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.[6] 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.

[6] 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.

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.

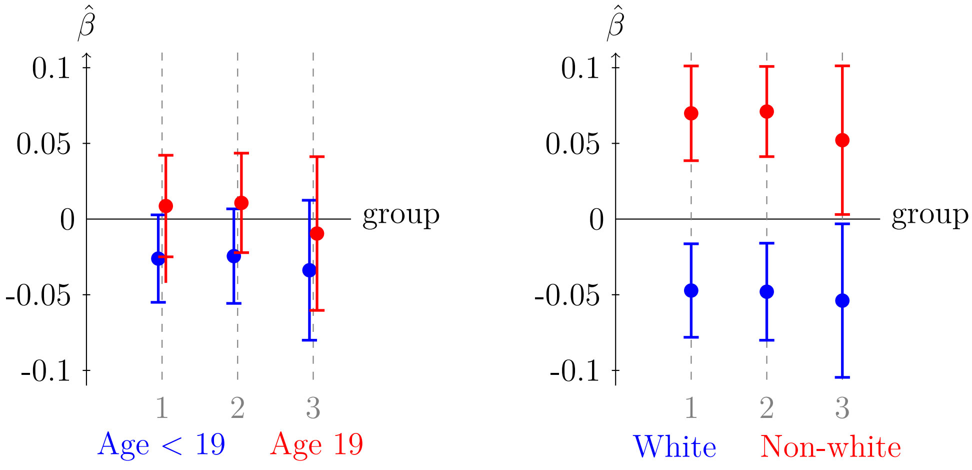

Figure 4: Interaction between individual and aggregate heterogeneity

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.

5 Conclusion

This paper extends the idea of the selection-on-observable assumption and motivates the use of the cluster-level distribution of individual-level control covariates to control for the cluster-level heterogeneity using the observable information. This framework is most relevant when the clusters are large, so that the cluster-level distributions are well-estimated, and the individuals within clusters are independent and identically distributed given the cluster-level heterogeneity. By explicitly controlling for the distribution of individuals, two different dimensions of heterogeneity in treatment effect are modelled, 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 cluster level, and the two dimensions of heterogeneity interact.

This paper serves as a first step in developing multilevel models where the distribution of individuals is used as a cluster-level object. For the choice of the dimension reduction method on distributions, the K-means algorithm are the functional PCA are used in this paper. The two approaches complement each other; one allows for flexible connection between the outcome model and the distribution model at the cost of discrete cluster-level heterogeneity and the other allows for continuous cluster-level heterogeneity at the cost of linearity. However, based on empirical contexts, a new dimension reduction method on distributions may be more suitable, calling for follow-up research that discuss different distributional 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.

Arkhangelsky, Dmitry and Guido Imbens, “The Role of the Propensity Score in Fixed Effect Models,” *arXiv e-prints*, 2022, pp. arXiv–1807.

Auerbach, Eric, “Identification and estimation of a partially linear regression model using network data,” *Econometrica*, 2022, *90* (1), 347–365.

Bai, Jushan, “Panel data models with interactive fixed effects,” *Econometrica*, 2009, *77* (4), 1229–1279.

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.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan, “The miracle of microfinance? Evidence from a randomized evaluation,” *American economic journal: Applied economics*, 2015, *7* (1), 22–53.

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.

Besanko, David, Sachin Gupta, and Dipak Jain, “Logit demand estimation under competitive pricing behavior: An equilibrium framework,” *Management Science*, 1998, *44* (11-part-1), 1533–1547.

Bonhomme, Stéphane and Elena Manresa, “Grouped patterns of heterogeneity in panel data,” *Econometrica*, 2015, *83* (3), 1147–1184.

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.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz, “The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment,” *American Economic Review*, 2016, *106* (4), 855–902.

Chintagunta, Pradeep K, Andre Bonfrer, and Inseong Song, “Investigating the effects of store-brand introduction on retailer demand and pricing behavior,” *Management Science*, 2002, *48* (10), 1242–1267.

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

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

Derenoncourt, Ellora, "Can you move to opportunity? Evidence from the Great Migration," *American Economic Review*, 2022, *112* (2), 369–408.

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.

Giné, Xavier and Dean Yang, "Insurance, credit, and technology adoption: Field experimental evidencefrom Malawi," *Journal of development Economics*, 2009, *89* (1), 1–11.

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.

Ke, Yuan, Jialiang Li, and Wenyang Zhang, "Structure identification in panel data analysis," *The Annals of Statistics*, 2016, *44* (3), 1193–1233.

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.

Lee, Jim, "Does size matter in firm performance? Evidence from US public firms," *international Journal of the economics of Business*, 2009, *16* (2), 189–203.

MacKay, Peter and Gordon M Phillips, "How does industry affect firm financial structure?," *The review of financial studies*, 2005, *18* (4), 1433–1466.

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.

Pesaran, M Hashem, "Estimation and inference in large heterogeneous panels with a multifactor error structure," *Econometrica*, 2006, *74* (4), 967–1012.

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.

Su, Liangjun, Zhentao Shi, and Peter CB Phillips, "Identifying latent structures in panel data," *Econometrica*, 2016, *84* (6), 2215–2264.

Tibshirani, Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, *58* (1), 267–288.

Voors, Maarten J, Eleonora EM Nillesen, Philip Verwimp, Erwin H Bulte, Robert Lensink, and Daan P Van Soest, "Violent conflict and behavior: a field experiment in Burundi," *American Economic Review*, 2012, *102* (2), 941–64.

Wang, Wuyi and Liangjun Su, "Identifying latent group structures in nonlinear panels," *Journal of Econometrics*, 2021, *220* (2), 272–295.

Zeleneev, Andrei, "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," *working paper*, 2020.

A Exchangeability

Assumption 1 assumes that the cluster-level distribution contains sufficient information on the cluster heterogeneity. 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* $\sigma_j : \{1, \cdots, N_j\} \to \{1, \cdots, N_j^*\}$,

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

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 $\left(i_1, \cdots, i_{N_j}\right)$

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

(*selection-on-observable*)

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

(*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 selection-on-observable 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,

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

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 *selection-on-observable.* □

Proposition 3 suggests propensity score matching based on $\mathbf{F}_j$, the population distribution function of X_{ij} for cluster j.

B Proofs

B.1 Theorem 1

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

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

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

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

Then,

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

Then, $\hat{\theta} - \tilde{A}\theta^0$ 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,

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

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,

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

The last inequality holds since $\sum_{j=1}^{J}\left\|\hat{G}(\hat{\lambda}_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{2} \le \sum_{j=1}^{J}\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{2}$ 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 $\Pr\left\{A_\varepsilon\right\} \to 1$ as $J \to \infty$ for any $\varepsilon > 0$.

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

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

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,

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

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

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

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

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,

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

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

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

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

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

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, $\mathrm{Var}_\lambda(A_{ii'}) \le C_2/h$ and $|\mathrm{Cov}_\lambda(A_{ii'}, A_{il})| \le C_2$ and

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

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 $\rho < r \le J$. 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 $r \le \rho$,

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

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 $\left|\hat{V}_r - V_r\right| \le \|\widehat{M} - M\|_2 \le \|\widehat{M} - M\|_F$ and

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

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

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

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

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

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

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