Clustered treatment in multilevel models*

Myungkou Shin†

November 12, 2022

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 treatment assignment is clustered, the treatment effect cannot be identified in a model with fully flexible cluster heterogeneity. To put restrictions on cluster heterogeneity, I assume that potential outcomes are independent of the treatment, conditioning on two sets of variables: cluster-level characteristics, and the distribution of individual-level characteristics for each cluster. With this *selection-on-distribution* framework, I control for treatment endogeneity and show how to recover treatment effect heterogeneity in both individual-level and cluster-level variables. To implement this idea, I propose a two-step estimation procedure based on a K-means algorithm. In the first step, I group clusters in terms of their distributions of individual-level characteristics. In the second step, I use the grouping structure to estimate the treatment effect. To illustrate the method, I study the disemployment effect of a raise in the minimum wage level on teenagers.

Keywords: hierarchical models, clustered treatment, heterogeneous treatment effect, selection on observable, functional regression, group fixed-effect

JEL classification codes: C13, C14, C31, C55

*I am deeply grateful to my advisors 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, and the participants of the metrics advising group and the metrics student group at the University of Chicago for their constructive comments and input. Any and all errors are my own.

†Kenneth C. Griffin Department of Economics, University of Chicago. Email: myungkoushin@uchicago.edu

1 Introduction

A vast majority of datasets used in economics are multilevel; units of observations have a hierarchical structure. For example, in a dataset that collects demographic characteristics of the US population, such as the Current Population Survey (CPS) or the Panel Study of Income Dynamics (PSID), each surveyee's residing county and state is also recorded; 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 a treatment is endogenously assigned at the cluster level and an outcome 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 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 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 treatment 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 either treated or not treated. 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

[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)

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 two clusters comes from their cluster-level heterogeneity or treatment status. In a simple example of linear regression model, the infeasibility of fully flexible cluster heterogeneity becomes evident:

$$Y_{ij} = \alpha_j + \beta D_j + U_{ij}. \tag{1}$$

Y_{ij} is an outcome variable for individual i in cluster j. D_j is a binary treatment variable for cluster j. Cluster fixed-effect α_j flexibly controls for the cluster-level heterogeneity in level. In the linear model (1), the treatment effect β is not identified due to multicollinearity between α_j and D_j, unless treatment is exogeneous, i.e. $\mathbf{E}[\alpha_j|D_j] = \mathbf{E}[\alpha_j]$.[2] Thus, we need restrictions on cluster-level heterogeneity.

To impose restrictions on cluster-level heterogeneity, I focus on cases where a researcher observes both individual-level and cluster-level covariates that are relevant for treatment assignment. Taking advantage of the available information, the econometric framework of this paper takes cluster-level covariates as they are and summarizes individual-level covariates at the cluster level by looking at their distribution. Then, conditioning on cluster-level covariates and distribution of individual-level covariates, treatment is assumed to be as good as random; treatment effect is identified. I call this approach *selection-on-distribution*.

The first step in motivating the *selection-on-distribution* assumption is the selection-on-observable assumption. The selection-on-observable assumption that treatment is random conditioning on some observable control covariates is widely used in the program evaluation literature to control for treatment endogeneity. To implement the selection-on-observable approach in a clustered treatment setup, a researcher needs to gather all the available information for each cluster since clusters are the units of treatment assignment. Consider the following linear regression model:

$$Y_{ij} = \beta D_j + \begin{pmatrix} Z_j & \mathbb{X}_j \end{pmatrix}^{\intercal} \theta^{cl} + U_{ij}. \tag{2}$$

Z_j is a vector of cluster-level control covariates for cluster j and X_{ij} is a vector of control covariates

[2]The cluster-level heterogeneity problem discussed in this paper is closely connected to a treatment endogeneity/selection bias problem. If the treatment is truly random, average treatment effect is identifed without controlling for cluster-level heterogeneity; cluster-level heterogeneity can be left fully flexible. However, when the treatment is endogenous and a researcher believes that cluster-level heterogeneity affects treatment assignment, cluster-level heterogeneity needs to be controlled for.

for individual i in cluster j. $\mathbb{X}_j = \{X_{ij}\}_{i=1}^{N_j}$ is the cluster-level collection of X_{ij} across all individuals in cluster j; there are N_j individuals in cluster j. A comparison with (1) can be made here. Though both models contain an element of cluster-level heterogeneity, model (1) stays flexible in terms of the cluster-level heterogeneity by using cluster fixed-effect α_j while model (2) imposes some structure on the cluster-level heterogeneity by using cluster-level regressors Z_j and $\mathbb{X}_j$. Though free of the multicollinearity problem, this direct application of the selection-on-observable assumption to a clustered treatment setup also has a drawback. Note that the dimension of the model parameter $\left(\beta, \theta, \theta^{cl}\right)$ is proportional to the cluster size N_j. Thus, even when the individual-level control covariate X_{ij} is low-dimensional, their cluster-level collection can be high-dimensional; the model induced by the selection-on-observable is not parsimonious.

Thus, I impose additional restrictions on the observable information $\mathbb{X}_j$. 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 $\mathbb{X}_j$, 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 $\mathbb{X}_j$ 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 $\hat{\mathbf{F}}_j$ contains all the relevant information for treatment assignment. Consider $\mathbf{F}_j$ such that for all $x \in \mathbb{R}^p$

$$\mathbf{F}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \Pr\{X_{ij} \leq x\} = \mathbf{E}\left[\hat{\mathbf{F}}_j(x)\right].$$

By shifting from $\hat{\mathbf{F}}_j$ to $\mathbf{F}_j$, clusters of different sizes can also be matched whereas in general empirical distribution functions cannot be the same for two clusters of different sizes. With $\mathbf{F}_j$, I formally state the key assumption of this paper: potential outcomes are independent of the treatment conditioning on the distribution function $\mathbf{F}_j$ and the cluster-level covariates Z_j, i.e. *selection-on-distribution*.

[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 $\mathbb{X}_j$ to its ordered statistics, the support for the control variable reduces down.

The *selection-on-distribution* assumption suggests that a researcher use cluster-level distribution of individual-level control covariates in modeling cluster-level treatment assignment and individual-level outcomes. To implement this strategy I use a K-means clustering algorithm, an unsupervised learning method to group clusters, to regress outcome variables on the distribution functions. The result of the K-means algorithm is a finite grouping on the set of clusters such that clusters in each group are similar to each other in terms of their empirical distributions of individual-level control covariates. With the grouping structure from the K-means algorithm, I suggest two separate sets of treatment effect estimators. Firstly when the dataset is cross-sectional and there is no control covariate at the cluster level, I propose nonparametric estimators with inverse probability weighting. I construct estimators for the average treatment effect (ATE), the average treatment effect on treated clusters (ATT), and the conditional average treatment effect ($CATE$). Secondly, when the dataset is repeated cross-section/panel data, or there exist cluster-level control covariates, I propose a least-square estimator under parametric models; an example is a linear regression model with group-specific time fixed-effects.

My main theoretical results discuss the asymptotic properties of the K-means grouping structure and the treatment effect estimators. To discuss the asymptotic properties of the K-means grouping structure, I assume that the distribution $\mathbf{F}_j$ is a function of a cluster-level latent factor λ_j: heterogeneity in $\mathbf{F}_j$ comes from a cluster-level random variable. Then, I additionally assume that the latent factor λ_j has a finite support: heterogeneity in $\mathbf{F}_j$ is finitely discrete. Under this discrete heterogeneity assumption, the K-means algorithm successfully assigns the clusters with the same value of λ_j to the same group; the probability of the K-means algorithm perfectly recovering the grouping structure induced by the latent factor λ_j goes to one when the number of individuals per clusters increases at a polynomial rate of the number of clusters.

Building on this perfect grouping result, I show consistency and asymptotic normality of the nonparametric estimators for ATE, ATT, and $CATE(\lambda)$, and the parametric least-square estimator. In all of the asymptotic distributions, the asymptotic variance has a closed-form expression that can be consistently estimated under regular assumptions. As a relaxation of the finiteness assumption, I also discuss an alternative assumption that the latent factor λ_j is a continuous random variable and its support is a compact set in $\mathbb{R}^q$. Under the continuity assumption, I show that the nonparametric treatment effect estimators for ATE and ATT are 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. 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. With the finite support assumption on the latent factor λ_j for the cluster-level distribution $\mathbf{F}_j$, I derive theoretical results for the estimation of the finite grouping structure in 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. Only the individual-level control covariates are used to group clusters and therefore the grouping structure does not suffer from overfitting. 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. Also, when the latent factor is assumed to be continuous, Bester and Hansen (2016) is closely comparable. The difference between Bester and Hansen (2016) and this paper is that Bester and Hansen (2016) mostly discusses the case where the grouping structure is readily observed to a researcher while in this paper the researcher has to construct one from the observable information.

Thirdly, this paper contributes to the distributional regression literature. To estimate propensity score and treatment effect with cluster-level distributions of individual-level covariates, the *selection-on-distribution* assumption calls for a functional regression method that regresses a one-dimensional variable onto a high-dimensional object such as distribution. By using the K-means grouping structure, this paper proposes a simple and easy-to-understand functional regression method, compared to the alternatives of kernel or functional principal component analysis: Póczos et al. (2013); Delicado (2011). The use of the K-means result as a functional regression can be understood as an extension of X-adaptive partition-based regression (Cattaneo et al., 2020); the K-means algorithm partitions clusters based on their distributions of individual-level control covariates, hence X-adaptive, and propensity score and treatment effect are estimated by projecting cluster-level treatment variable and individual-level outcome variable onto a step function that is constant within the partitions.

In addition, there are several literatures that my paper relates to. Firstly, the *selection-on-distribution* assumption is comparable to 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. While there is no ordering between the two assumptions in terms of flexibility, the difference is intuitive. 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. Section 5 extends the model in use. In Section 6, simulation results are presented and in Section 7, the empirical illustration of the econometric framework is provided.

2 Model

An econometrician observes $\left\{ \{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, D_j \right\}_{j=1}^{J}$ 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 $D_j \in \{0, 1\}$ is a cluster-level binary treatment variable for cluster j. Note that X_{ij} may include lagged outcomes if the econometrician observes panel data. There exist J clusters and each cluster contains N_j individuals: in total there are $N = \sum_{j=1}^{J} N_j$ individuals. To discuss treatment effect, I let the observed outcome Y_{ij} for individual i in cluster j be constructed from treated potential outcome $Y_{ij}(1)$ and untreated potential outcome $Y_{ij}(0)$:

$$Y_{ij} = D_j \cdot Y_{ij}(1) + (1 - D_j) \cdot Y_{ij}(0).$$

Potential outcomes are defined at the individual level but treatment is defined at the cluster level: the multilevel structure plays a key role in treatment assignment.

Now, I introduce three assumptions: iid-ness across clusters, the *selection-on-distribution*, and finiteness on the latent factor.

Assumption 1. *(independent and identically distributed clusters with a latent factor)*
There exists a cluster-level latent factor $\lambda_j \in \Lambda$. With λ_j,

$$\big(D_j, N_j, \lambda_j\big) \sim \textit{iid.}$$

Then, $H^{hyper}\big(\{D_j, N_j, \lambda_j\}_{j=1}^{J} \big)$, the conditional distribution of $\left\{ \{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j} \right\}_{j=1}^{J}$ given $\{D_j, N_j, \lambda_j\}_{j=1}^{J}$, is a product of $H(D_j, N_j, \lambda_j)$, the conditional distribution of $\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}$

given $\big(D_j, N_j, \lambda_j\big)$*:*

$$H^{hyper}\big(\{D_j, N_j, \lambda_j\}_{j=1}^{J}\big) = \prod_{j=1}^{J} H(D_j, N_j, \lambda_j).$$

In Assumption 1, I assume cluster-level iid-ness. Following Bugni et al. (2022), the iid-ness discussed in Assumption 1 comes from a two-step data generating process: firstly, cluster-level variables $\big(D_j, N_j, \lambda_j\big)$ are independently drawn from a distribution. Then, conditioning on the cluster-level variables $\big(D_j, N_j, \lambda_j\big)$, individual-level variables $\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}$ are drawn from a distribution denoted with H, independently of individual-level variables and cluster-level variables from all the other clusters; independence. The distribution function H is not cluster-specific; identicalness. Dependence structure within a cluster is unrestricted.

The cluster-level latent factor λ_j can be thought of as the latent heterogeneity across clusters in terms of the distribution of individual-level potential outcomes and individual-level control covariates. So far, no further restrictions are made on λ_j. Thus, by letting λ_j be cluster-specific conditional distribution function of $\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}$ given $\big(D_j, N_j\big)$, the latent factor λ_j becomes a placeholder and Assumption 1 can be rewritten with $H(D_j, N_j)$ and $H^{hyper}(\{D_j, N_j\}_{i=1}^{J})$.

Assumption 2 introduces more context on the latent factor λ_j and assumes conditional independence of the treatment.

Assumption 2. *(selection-on-distribution)*
Let $B(\mathbb{R}^p)$ *denote the space of distribution functions on* $\mathbb{R}^p$*, with a metric* $\|\cdot\|_{w,2}$ *defined with a weighting function* w *as follows:*

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

Then, there exists an injective function $G : \Lambda \to B(\mathbb{R}^p)$ *such that for every* $x \in \mathbb{R}^p$*,*

$$\mathbf{F}_j(x) := \frac{1}{N_j} \sum_{i=1}^{N_j} \Pr\big\{X_{ij} \leq x \big| D_j, N_j, \lambda_j\big\} = \big(G(\lambda_j)\big)(x).$$

w *satisfies that* $\Pr\Big\{\|G(\lambda_j)\|_{w,2} < \infty\Big\} = 1$*. Also,*

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

Assumption 2 has two parts. Firstly, Assumption 2 assumes that the latent factor λ_j is the cluster-

level heterogeneity in terms of the distribution of X_{ij}. The connection between the latent factor λ_j and the distribution of X_{ij} is through the injective function G. To define injectivity, a metric $\|\cdot\|_{w,2}$ is defined on the space of distribution functions. Secondly, Assumption 2 assumes that the individual-level potential outcomes and the individual-level control covariates are independent of the cluster-level treatment status, after conditioning on the cluster-level variables N_j and λ_j: $H(D_j, N_j, \lambda_j) = H(N_j, \lambda_j)$. Thanks to the injectivity of G, the individual-level potential outcomes are independent of the treatment conditioning on N_j and $\mathbf{F}_j$:

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

This means that the distribution of X_{ij} for each cluster contains sufficient information in treatment assignment process, along with N_j, so that the treatment is as good as random after conditioning on N_j and $\mathbf{F}_j$. I call this the *selection-on-distribution* assumption.

Remark 1. When $p > 1$, an additional assumption can be made on λ_j for model simplicity. Let X_{ijl} denote the l-th random variable of the p-dimensional random vector X_{ij}:

$$X_{ij} = \big(X_{ij1}, \cdots, X_{ijp}\big).$$

Assume the second part of Assumption 2 as is. In addition, assume that λ_j is a p-tuple of latent factors, i.e.

$$\lambda_j = \big(\lambda_{j1}, \cdots, \lambda_{jp}\big),$$

and repeat the first part of Assumption 2 with each of λ_{jl} and the marginal distribution of X_{ijl}: for every $x_l \in \mathbb{R}$,

$$\mathbf{F}_{jl}(x_l) := \frac{1}{N_j} \sum_{i=1}^{N_j} \Pr\big\{X_{ijl} \le x_l \big| D_j, N_j, \lambda_j\big\} = \big(G(\lambda_{jl})\big)(x).$$

This modification to Assumption 2 assumes that each of the marginal distributions of X_{ij} conveys information on one component of λ_j. Thus, we do not lose any information for the latent factor λ_j, by shifting the conditioning object from the joint distribution $\mathbf{F}_j$, to a collection of the p marginal distributions $\mathbf{F}_{j1}, \cdots, \mathbf{F}_{jp}$.

Assumption 3 assumes that the latent factor has a finite support.

Assumption 3. *(finite support) The latent factor λ_j has a finite support: with a fixed K,*

$$\Lambda = \left\{\lambda^1, \cdots, \lambda^K\right\}.$$

To reduce the dimension of $\mathbf{F}_j$, I assume that the support of the latent factor is finite. $\mathbf{F}_j$, without any restriction, is an infinite-dimensional object; under Assumption 3, $\mathbf{F}_j$ can only take K values. Thus the idea of *selection-on-distribution* from Assumption 2 is facilitated under Assumption 3; there are finite types of clusters in terms of their distribution of the individual control covariate X_{ij} and the question of treatment effect estimation becomes that of recovering the finite type for each cluster. I discuss the case where Assumption 3 is relaxed and Λ is assumed to be a compact subset of $\mathbb{R}^q$, in Section 5.

Remark 2. The parameter K is often unknown to an econometrician. An estimator of K with the information criterion will be discussed in Section 3.

2.1 Treatment effect

In this subsection, I define various treatment effect parameters that are identified under Assumptions 1-3. The multilevel nature of the model is evident in the definitions of the treatment effect parameters as well. I construct two sets of aggregate treatment effect parameters, depending on whether I put equal weights across clusters or across individuals. Also, for conditional treatment effect parameters, I consider both cluster-level variables and individual-level variables as a conditioning variable.

2.1.1 Aggregate treatment effect

Firstly, let us construct cluster-level aggregate treatment effect parameters:

$$ATE^{cl} = \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right], \tag{3}$$

$$ATT^{cl} = \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)|D_j = 1\right]. \tag{4}$$

I used the superscript *cl* to indicate that the treatment effect parameters are defined with cluster means, putting equal weights across clusters. Expanding this, we can construct individual-level

aggregate treatment effect parameters:

$$ATE = \mathbf{E}\left[\frac{N_j}{\mathbf{E}\left[N_j\right]}\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)\right], \tag{5}$$

$$ATT = \mathbf{E}\left[\frac{N_j}{\mathbf{E}\left[N_J|D_j = 1\right]}\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)\Big|D_j = 1\right] \tag{6}$$

When the cluster size does not vary, individual-level aggregate treatment effect parameters are equal to their cluster-level counterparts.

2.1.2 Conditional treatment effect

Now, let us discuss conditional treatment effect parameters. At the cluster level, I use the cluster-level latent factor λ_j as a conditioning variable:

$$CATE^{cl}(\lambda) = \mathbf{E}[\bar{Y}_j(1) - \bar{Y}_j(0)|\lambda_j = \lambda]. \tag{7}$$

At the individual level, we add an additional conditioning variable at the individual level to define conditional treatment effect parameters:

$$CATE(x, \lambda) = \mathbf{E}\left[Y_{ij}(1) - Y_{ij}(0)|X_{ij} = x, \lambda_j = \lambda\right]. \tag{8}$$

Note that conditioning on λ_j, $CATE$ and $CATT$ are the same.

With $CATE$ defined as above, multilevel nature of heterogeneity in treatment effect can be explored. A first use of the model is to allow for heterogeneity in treatment effect that comes from individual-level characteristics. Fix λ and let $CATE(x, \lambda)$ be a function of x: $IH_\lambda(x; \lambda) = CATE(x, \lambda)$. Then, $IH_\lambda(x)$ captures the individual-level heterogeneity in treatment effect. Secondly, the model finds heterogeneity in treatment effect in terms of cluster-level aggregation of the individual-level characteristics. Fix x and let $CATE(x, \lambda)$ be a function of λ: $AH_x(\lambda; x) = CATE(x, \lambda)$. $AH_x(\lambda)$ captures the aggregate heterogeneity. Individual heterogeneity discusses how the treatment affects individuals with different characteristics differently while aggregate heterogeneity discusses how the treatment affects the same individual differently depending on which cluster they belong to.

These heterogeneity parameters in treatment effects are often of interest in applications. In a typical regression specification to estimate treatment effect, interaction terms between some

control covariates and the binary treatment variable are often included. If the control covariate is an individual-level variable, the interaction term essentially captures individual heterogeneity in treatment effect and if the control covariate is a cluster-level variable, the interaction term captures aggregate heterogeneity. In addition to using cluster-level variables given from the dataset, the econometric framework of this paper provides another window to discuss heterogeneous treatment effect: the distribution of the individual-level control covariates. By looking at the distribution, I suggest a sensible assumption to aggregate the individual-level information available at each cluster and allow my model to capture heterogeneity in treatment effect in terms of the composition of individuals for each cluster.

2.2 Examples

There are a plenty of economic models where a distribution of individual-level control covariates is a key determinat in cluster-level treatment assignment, and treatment effect shows both individual heterogeneity and aggregate heterogeneity. In this subsection, I list three examples.

2.2.1 Minimum wage and unemployment

Let us construct a dynamic model where state legislators decide whether or not to increase their state's minimum wage level. At each time period, the state legislators observe the distribution of individual-level socioeconomic and demographic characteristics: with $X_{ijt} \in \mathbb{R}^p$ being the socioeconomic and demographic characteristics of individual i in state j at time t, the state legislators observe

$$\mathbf{F}_{jt}(x) = \Pr\{X_{ijt} \leq x\} \qquad \forall x \in \mathbb{R}^p,$$
$$\mathbf{F}_{jt} = \mathbf{F}(\lambda_{jt}, MinWage_{jt}/P_t).$$

The distribution $\mathbf{F}_{jt}$ has two determinants: underlying labor market fundamental λ_{jt} and the minimum wage level $MinWage_{jt}$. Note that the nominal minimum wage level is divided with a price level $P_t = (1+p)^t$. It is assumed that the price level increases in a deterministic way, at the rate of p, and the state of the labor market, λ_{jt}, follows a Markov process. Let us further assume that the state space Λ of λ_{jt} is finite: $\Lambda = \{\lambda^1, \cdots, \lambda^q\}$. Then, the transition probability is denoted with a $q \times q$ matrix: $\mathbb{P}$. The nominal minimum wage level, $MinWage_{jt}$, is determined by the state legislators, in the process described below.

At each time period, after observing the distribution $\mathbf{F}_{jt}$, the state legislators decide the minimum wage level for the next period. The decision to raise the minimum wage level comes at a cost c_{jt}. In deciding the minimum wage level for the next period, the state legislators maximize an infinite sum of a period-specific social welfare function:

$$\begin{aligned} SW_{jt} &= g(\mathbf{F}_{jt}) - c_{jt}\mathbf{1}\{MinWage_{jt+1} > MinWage_{j,t}\} \\ &= g(\lambda_{jt}, MinWage_{jt}/P_t) - c\mathbf{1}\{MinWage_{jt+1} > MinWage_{j,t}\}. \end{aligned}$$

g is labor market welfare function that takes the distribution $\mathbf{F}_{jt}$ as its input and evaluates the social welfare generated in the labor market. Suppose X_{ijt} includes two variables: Emp_{ijt}, the employment status of individual i, and $WageInc_{ijt}$, the wage income of individual i. If the state legislators only care about the unemployment rate, we would have $g(\mathbf{F}_{jt}) = g(\Pr\{Emp_{ijt} = 0\})$. If the state legislators care about the proportion of their constituents making below the federal poverty line, we would have $g(\mathbf{F}_{jt}) = g(\Pr\{WageInc_{ijt} \leq FederalPovertyLine\})$. In general, the function g would be more complex. c_{jt} is the menu cost of raising the nominal minimum wage level. I assume that the menu cost process has no autocorrelation and is independent of the labor market state: $c_{jt} \sim$ iid and $\{c_{jt}\}_t \perp\!\!\!\perp \{\lambda_{jt}\}_t$. The total period-specific social welfare is the labor market welfare minus the cost of changing the minimum wage level.

Based on the setup discussed above, let us construct a Bellman equation for the dynamic optimization problem:

$$V(\lambda, m, c) = \max_{m' \geq m} \left\{ g(\lambda, m) - c\mathbf{1}\{m' > m\} + \delta \mathbf{E}\left[V\left(\lambda', \frac{m'}{1+p}, c'\right) | \lambda \right] \right\}.$$

λ is the labor market state, m is the real minimum wage level and c is the menu cost of raising the minimum wage level: (λ, m, c) is the state of the dynamic programming and m' is the action. Given (λ, m, c), V is the value function that evaluates the discounted sum of social welfare. The expectation notation in the Bellman equation is a conditional expectation on λ since the labor market state has Markov property. Specifically,

$$\mathbf{E}\left[V\left(\lambda', \frac{m'}{1+p}, c'\right) | \lambda \right] = \begin{pmatrix} \mathbf{1}\{\lambda = \lambda^1\} & \cdots & \mathbf{1}\{\lambda = \lambda^q\} \end{pmatrix} \cdot \mathbb{P} \cdot \begin{pmatrix} \int V\left(\lambda^1, \frac{m'}{1+p}, c'\right) f(c')dc' \\ \vdots \\ \int V\left(\lambda^q, \frac{m'}{1+p}, c'\right) f(c')dc' \end{pmatrix}.$$

f is the density function of c_{jt}. The state legislators solve this dynamic optimization problem and set the minimum wage level: the optimal policy function $m^*(\lambda, m)$ sets the minimum wage level for the next period. It is evident in this model that the distribution $\mathbf{F}_{jt}$ is the key determinant in 'treatment' assignment process: *selection-on-distribution*.

In Section 7, I analyze the effect of a raise in the minimum wage level on employment status of teenagers. Relying on this framework, I control for the state-level heterogeneity in the minimum wage decision process, using the cluster-level distribution of individual-level control covariates and solve the selection bias problem.

2.2.2 Team-level performance pay

Suppose a company introduces a team-level performance pay scheme under which workers are rewarded $r > 0$ when the total output of their team is above some predetermined level y^*. The company does not introduce the performance pay scheme to all teams at once. Instead, the company considers each team's worker composition and decides whether or not to apply the performance pay scheme: $D_j = 1$ indicates that team j is under the performance pay scheme.

To discuss treatment effect heterogeneity in this example, let us consider a simple linear outcome model with latent effort level, which will be the main source of heterogeneity in treatment effect. Each worker's output level Y_{ij} is determined from their productivity level $X_{ij} \in [0, 1]$, latent binary effort level $E_{ij} \in \{0, 1\}$, and some idiosyncratic error U_{ij}:

$$Y_{ij} = \beta_1 X_{ij} + \beta_2 E_{ij} + U_{ij}.$$

The productivity level X_{ij} is observed to a researcher and comes from a distribution whose parameter is λ_j.

The act of putting in 'efforts' is not free; worker's utility decreases by $c(X_{ij})$ when $E_{ij} = 1$. With monotone decreasing c,

$$utility_{ij} = \begin{cases} r \cdot \mathbf{1}\{\sum_i Y_{ij} \geq y^*\} - c(X_{ij}) \cdot E_{ij}, & \text{if } D_j = 1 \\ -c(X_{ij}) \cdot E_{ij}, & \text{if } D_j = 0 \end{cases}$$

Without any reward on putting in efforts, effort level E_{ij} is always 0. With the performance pay scheme, a worker decides if they should put in efforts by looking at their team composition. Given some belief on the effort levels of his teammates, the optimal strategy of an worker who maximizes

expected payoff is to put in 'efforts' if and only if

$$\Pr_{X_{-j}}\left\{\beta_1\sum_i X_{ij}+\beta_2\sum_{i'\neq i}E_{ij}+\sum_i U_{ij}\geq y^*-\beta_2\right\}$$
$$-\Pr_{X_{-j}}\left\{\beta_1\sum_i X_{ij}+\beta_2\sum_{i'\neq i}E_{ij}+\sum_i U_{ij}\geq y^*\right\}\geq\frac{c(X_{ij})}{r}.$$

Note that the probability is an expectation over worker i's belief on the productivity level and the effort level of his teammates: the expectation is conditional on λ_j. As an equilibrium outcome of this game that workers play within a team, the optimal effort level $E^*_{ij}=e(X_{ij},\lambda_j)$ would be a function of one's own productivity level and the productivity distribution λ_j.

From the discussion above, it directly follows that the treatment effect on worker i is a function of both their own productivity level X_{ij} and their team's productivity distribution λ_j:

$$Y_{ij}(1)-Y_{ij}(0)=\beta_2E^*_{ij}(1)=\beta_2e\big(X_{ij},\lambda_j\big).$$

Firstly, we see that the treatment affects workers differently within a given team; for example, when $c(x)$ decreases in x, workers with higher productivity are more reactive to the treatment, thus having positive treatment effect, while workers with lower productivity may not react and have a zero treatment effect: individual heterogeneity. Secondly, the performance pay scheme affects workers with the same productivity level differently, when their team compositions vary. For example, the performance pay scheme may increase output from a worker of a certain productivity level when they are assigned to a high-productivity team, but not when they are assigned to a low-productivity team: aggregate heterogeneity. The construction of conditional treatment effect parameters as in $CATE(x,\lambda)$ above allows us to explore this heterogeneity in treatment effect.

2.2.3 School-level teaching strategy with peer effect

My third and last example is based on a network formation model. Suppose a school district experiments with a new teaching strategy across schools. In this example, I assume a latent network structure among students and resulting peer effect. Let Y_{ij}, test score of student i in school j, be determined from their own ability X_{ij} and their peers' ability:

$$Y_{ij}=(\theta_1+D_j\beta_1)\cdot X_{ij}+(\theta_2+D_j\beta_2)\cdot e_i{}^{\intercal}G_j\mathbb{X}_j+U_{ij}.$$

Note that the slope coefficients depend on D_j, the teaching strategy of school j. To allow for peer effect, a $N_J \times N_J$ (reweighted) network matrix G_j is used. G_j is constructed in a way that its i-th row j-th column element $\left(G_j\right)_{hi}$ is

$$\frac{W_{hij}}{\sum_{i'} W_{hi'j}}$$

where $W_{hij} \in \{0,1\}$ is a binary linkage variable indicating whether student i and student h in school j are friends. For example, $\left(G_j\right)_{hi} = 1/4$ means that student h has four friends and student i is one of them. $\mathbb{X}_j$ is a stacked vector of X_{ij}s for cluster j:

$$\mathbb{X}_j = \begin{pmatrix} X_{1j} & \cdots & X_{N_j j} \end{pmatrix}^\intercal.$$

Then, $G_j\mathbb{X}_j$ is a column vector of mean ability of peers, for students in school j. e_i is the standard unit vector whose i-th element is one and the rest are zeros; ${e_i}^\intercal G_j\mathbb{X}_j$ retrieves the mean ability of student i's peers.

The latent friendship network structure G_j is constructed from the following network formation model:

$$W_{hij} = \begin{cases} \mathbf{1}\{|\tilde{X}_{hj} - \tilde{X}_{ij}|^\intercal \eta + \varepsilon_{hij} \geq 0\}, & \text{if } h \neq i \\ 0, & \text{if } h = i \end{cases}$$

with some observable student characteristic $\tilde{X}_{ij}$: e.g. sex, race, address, etc. With $\eta < 0$, students with similar characteristic are more likely to be friends.

Let $\mathbf{F}(\lambda)$ denote the distribution of $\left(X_{ij}, \tilde{X}_{ij}\right)$ for a certain school and $(x, \tilde{x})$ denote an ability level and observable characteristics of a certain student at the school. Conditioning on $(x, \tilde{x}, \lambda)$,

$$\begin{aligned} CATE(x, \mathbf{F}) &= \beta_1 \cdot x + \beta_2 \cdot \sum_{i \neq 1} \mathbf{E}\left[\frac{W_{1ij}}{\sum_{i'} W_{1i'j}} \middle| \mathbf{F}(\lambda)\right] x_{ij} \\ &=: \beta_1 \cdot x + \beta_2 \cdot g(\tilde{x}, \lambda). \end{aligned}$$

It is easy to see that a change in $(x, \tilde{x})$ shifts both $\beta_1 \cdot x$, the direct treatment effect, and $\beta_2 \cdot g(\tilde{x}, \mathbf{F})$, the indirect peer effect, while a change in $\mathbf{F}$ only shifts the latter. Based on this observation, I make following connection to the network effect/peer effect literature: individual heterogeneity defined as in this paper refers to a shift in the total treatment effect, which is a sum of the direct treatment effect and the indirect peer effect, while aggregate heterogeneity refers to a shift only in the indirectly peer effect.

3 Estimation

In this section, I propose a two-step estimation procedure. The first step is to find a finite grouping structure on clusters by solving a K-means minimization problem with cluster-level distributions of individual-level control covariates. The second step is to use the finite grouping structure on clusters in treatment effect estimation. In the second step, I propose two sets of estimators; nonparametric estimators with inverse probability weightings and a least-squre estimator with parametric model.

3.1 First step: K-means grouping

In the first step, I construct a finite grouping structure by aggregating the individual-level information at the cluster-level to a distribution. In practice, the cluster-level distributions are not directly observed. Thus, as an estimator for the cluster-specific distribution of the individual-level control covariate, $\mathbf{F}_j$, I use the empirical distribution function $\hat{\mathbf{F}}_j$: for all $x \in \mathbb{R}^p$,

$$\hat{\mathbf{F}}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\{X_{ij} \leq x\}. \tag{9}$$

A key observation which directly follows Assumption 2 is that $\mathbf{E}\left[\hat{\mathbf{F}}_j(x)|D_j, N_j, \lambda_j\right] = \big(G(\lambda_j)\big)(x)$ for every $x \in \mathbb{R}^p$: $\hat{\mathbf{F}}_j$, the estimator I use for $\mathbf{F}_j$, is pointwise unbiased. In Section 4, I discuss conditions under which $\hat{\mathbf{F}}_j$ is a good estimator for $\mathbf{F}_j$ more rigorously.

To construct a finite grouping structure on clusters, I start with some predetermined $K \leq J$. With the predetermined K, I assign each cluster to one of K groups, based on $\|\cdot\|_{w,2}$ from Assumption 2, so that clusters within a group are similar to each other in terms of $\hat{\mathbf{F}}_j$:

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

The K-means minimization problem in (10) finds a grouping on J clusters, while minimizing the within-group variation of clusters measured in terms of $\|\cdot\|_{w,2}$. In the minimization problem, there are two arguments to minimize the objective over: k_j and $G(k)$. k_j is the group to which cluster j is assigned to: $k_j \in \{1, \cdots, K\}$. $G(k)$ is the distribution of X_{ij} for group k. For each cluster j, $\hat{k}_j$ will be the group which cluster j is closest to, measured in terms of $\left\|\hat{\mathbf{F}}_j - G(k)\right\|_{w,2}$. The solution to (10) maps $\hat{\mathbf{F}}_j$ to $\hat{k}_j$, a discrete variable with finite support: dimension reduction.

K, the dimension parameter of the finite grouping structure is often unknown. When K is unknown, an information criterion can be used to estimate K.[5] Assume in addition to Assumption 3 that we are given a fixed constant $K_{\max} < J$ such that $K \leq K_{\max}$ and let

$$Q_J(K) = \min_{k_j \in \{1,\cdots,K\}, G(1),\cdots,G(K)} \sum_{j=1}^{J} \left\| \hat{\mathbf{F}}_j - G(k_j) \right\|_{w,2}^2.$$

Then, for example, an estimator based on the Bayesian Information Criterion (BIC) is

$$\hat{K} = \arg \min_{K \leq K_{\max}} \big(Q_J(K) + K \log J \big)$$

and an estimator based on the Akaike Information Criterion (AIC) is

$$\hat{K} = \arg \min_{K \leq K_{\max}} \big(Q_J(K) + K \big).$$

Given estimated $\hat{K}$ or known K, I use an iterative algorithm, called the (naive) K-means clustering algorithm or Lloyd's algorithm, to solve the minimization problem (10). Find that at the optimum

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

The estimated $\hat{G}$ for group k will be the subsample mean of $\hat{F}_j$ where the subsample is the set of clusters that are assigned to group k under $\big(\hat{k}_1, \cdots, \hat{k}_J\big)$. Motivated by this observation, the iterative K-means algorithm finds the minimum as follows: given an initial grouping $\big(k_1^{(0)}, \cdots, k_N^{(0)}\big)$,

1. **(update G)** Given the grouping from the s-th iteration, update $G^{(s)}(k)$ to be the subsample mean of $\hat{\mathbf{F}}_j$ where the subsample is the set of clusters that are assigned to group k under $\big(k_1^{(s)}, \cdots, k_J^{(s)}\big)$:

$$\Big(G^{(s)}(k) \Big)(x) = \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{k_j^{(s)} = k\}} \sum_{j=1}^{J} \hat{\mathbf{F}}_j(x) \mathbf{1}\{k_j^{(s)} = k\}.$$

[5] Using an information criterion, Bai and Ng (2002) estimates the dimension of the latent factor in a factor model for panel data. More closely to the setup of this paper, Ke et al. (2016); Wang and Su (2021) also use an information criterion to estimate the dimension of a finite grouping structure in a panel data model with group fixed-effects. In the canonical models of Ke et al. (2016); Wang and Su (2021), slope coefficient estimates in a linear model are used to group units; in this paper, distribution functions in a multilevel model are used to group clusters.

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

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

3. Repeat 1-2 until $\left(k_1^{(s)}, \cdots, k_J^{(s)}\right)$ 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 $G^{(s)}(k)$ 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 $O(J^{dK+1})$. On the other hand, the iterative algorithm is computed in time $O(JKd)$. Thus, the iterative algorithm gives us computational gain, at the cost of not being able to guarantee the global minimum.[6] Thus, I suggest using multiple initial groupings and comparing the results of the K-means algorithm across initial groupings. For more discussion on how to choose the initial grouping, see Appendix.

3.2 Second step: treatment effect estimation

In the second step, I use the finite grouping structure from the first step to estimate treatment effect parameters. Specifically, I propose two sets of estimators for two different data contexts. Firstly, suppose that a researcher is given a cross-sectional dataset without any cluster-level control covariates relevant for treatment assignment; the hierarchical nature of the model only exists in terms of the clustering structure on individuals. Then, no additional function form assumptions other than the basic model described in Assumptions 1-3 are needed to model the data. Thus, in this case, I propose nonparametric estimators directly motivated from Assumptions 1-3, using the finite grouping structure from the first step and the inverse probability weighting principle. Secondly, suppose that a researcher is given a panel data or cluster-level control covariates relevant

[6] 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.

for treatment assignment. In this case, the researcher would want to impose more restrictions on the model to control for time heterogeneity, or the cluster-level control covariates. To that end, I propose a least-square estimator in a parametric model where the cluster-level latent factor is treated as a categorical variable.

3.2.1 Nonparametric estimator

When the finite grouping structure $\{\hat{k}_1, \cdots, \hat{k}_J\} \in \{1, \cdots, K\}^J$ successfully recovers the latent factor $\{\lambda_j, \cdots, \lambda_J\} \in \{\lambda^1, \cdots, \lambda^1\}^J$, a direct mean comparison within a group is a natural estimator for $CATE^{cl}(\lambda)$, from the *selection-on-distribution* assumption. Thus, for each group estimated in the first step, I construct cluster-level conditional treatment effect estimators as follows: for $k = 1, \cdots, K$,

$$\widehat{CATE}^{cl}(k) = \frac{\sum_{j=1}^{J} \bar{Y}_j D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\}} - \frac{\sum_{j=1}^{J} \bar{Y}_j (1 - D_j) \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J} (1 - D_j) \mathbf{1}\{\hat{k}_j = k\}}. \tag{11}$$

Note that I cannot construct an estimator for specific λ^k. From the construction of the model, the realized value of λ_j cannot be identified, nor is it necessary to know the realized value of λ_j to discuss aggregate heterogeneity. Thus, while I construct K distinct estimators with $\widehat{CATE}^{cl}(k)$, I remain agnostic about how the estimators connect to $CATE^{cl}(\lambda^k)$. In addition, when X_{ij} is discrete, I estimate individual-level treatment effect parameter $CATE(x, \lambda)$ as follows:

$$\widehat{CATE}(x, k) = \frac{\sum_{j=1}^{J} \sum_{i=1}^{N_J} Y_{ij} D_j \mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}{\sum_{j=1}^{J} \sum_{i=1}^{N_J} D_j \mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}} - \frac{\sum_{j=1}^{J} \sum_{i=1}^{N_J} Y_{ij} (1 - D_j) \mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}{\sum_{j=1}^{J} \sum_{i=1}^{N_J} (1 - D_j) \mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}. \tag{12}$$

When X_{ij} is continuous, we can use kernel smoothing to construct a nonparametric estimator, or use a parametric model as will be discussed in the next subsection. By comparing $\widehat{CATE}^{cl}(k)$ across $k = 1, \cdots, K$, I estimate aggregate heterogeneity in treatment effect. Similarly, by fixing k and comparing $\widehat{CATE}(x, k)$ across x, I estimate individual heterogeneity in treatment effect.

To construct aggregate treatment effect estimators, I estimate propensity score

$$\pi(\lambda) = \mathbf{E}[D_j | \lambda_j = \lambda] \tag{13}$$

as follows:

$$\hat{\pi}(k) = \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{\hat{k}_j = k\}} \sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\}, \tag{14}$$
$$\hat{\pi}_j = \hat{\pi}(\hat{k}_j).$$

The propensity score estimates are computed as a sample mean of treatment status variable D_j for each group. Note that the propensity score estimator in (14) does not guarantee overlap. There are multiple remedies to this problem of no overlap. For example, we may drop the group without overlap altogether in estimating the aggregate treatment effect. Or, we may pair the clusters before the K-means algorithm so that each treated cluster is matched with the closest untreated cluster in terms of $\hat{\mathbf{F}}_j$ and have the K-means algorithm to group the pairs, instead of the clusters. In this paper, I choose the trimming strategy. I trim the propensity score estimator to be on $[h, 1-h]$: with some $h \in (0, 0.5)$,

$$\hat{\pi}_j = \hat{\pi}(\hat{k}_j) = \min\left\{1-h, \max\left\{h, \frac{\sum_{l=1}^{J} D_l \mathbf{1}\{\hat{k}_l = \hat{k}_j\}}{\sum_{l=1}^{J} \mathbf{1}\{\hat{k}_l = \hat{k}_j\}}\right\}\right\}. \tag{15}$$

Given the propensity score estimators from (15), the cluster-level aggregate treatment effect are estimated as follows. Using the inverse probability weighting principle,

$$\widehat{ATE}^{cl} = \frac{1}{J} \sum_{j=1}^{J} \left(\frac{D_j \bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j} \right), \tag{16}$$

$$\widehat{ATT}^{cl} = \frac{1}{\sum_{j=1}^{J} D_j} \sum_{j=1}^{J} \left(D_j \bar{Y}_j - \frac{(1-D_j)\hat{\pi}_j \bar{Y}_j}{1-\hat{\pi}_j} \right). \tag{17}$$

Likewise, the individual-level aggregate treatment effect estimators are:

$$\widehat{ATE} = \frac{1}{N} \sum_{j=1}^{J} N_j \left(\frac{D_j \bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j} \right), \tag{18}$$

$$\widehat{ATT} = \frac{1}{\sum_{j=1}^{J} D_j N_j} \sum_{j=1}^{J} N_j \left(D_j \bar{Y}_j - \frac{(1-D_j)\hat{\pi}_j \bar{Y}_j}{1-\hat{\pi}_j} \right). \tag{19}$$

3.2.2 Parametric estimator

In the baseline model discussed in Section 2, an econometrician only observes control covariates at the individual level. In this subsection, I extend the baseline model to include cluster-level control covariates, at the cost of parametrization. The econometrician now observes

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

where $Z_j \in \mathbb{R}^{p^{cl}}$ is a p^{cl}-dimensional vector of cluster-level control covariates. To include the cluster-level covariates Z_j, let us modify Assumption 1, by replacing N_j with an arbitrary cluster-level random vector Z_j that includes N_j.

$$\big(D_j, Z_j, \lambda_j\big) \sim \text{iid.}$$

Also, $H^{hyper}\big(\{D_j, Z_j, \lambda_j\}_{j=1}^{J}\big)$*, the conditional distribution of* $\left\{\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}\right\}_{j=1}^{J}$ *given* $\{D_j, Z_j, \lambda_j\}_{j=1}^{J}$*, is a product of* $H(D_j, Z_j, \lambda_j)$*, the conditional distribution of* $\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}$ *given* $\big(D_j, Z_j, \lambda_j\big)$*:*

$$H^{hyper}\big(\{D_j, Z_j, \lambda_j\}_{j=1}^{J}\big) = \prod_{j=1}^{J} H(D_j, Z_j, \lambda_j).$$

In addition, there exists some function $g : \mathbb{R}^p \times \{0,1\} \times \mathbb{R}^{p^{cl}} \times \Lambda \to \mathbb{R}$,

$$Y_{ij} = g(X_{ij}, D_j, Z_j, \lambda_j; \theta_0) + U_{ij}, \tag{20}$$

$$0 = \mathbf{E}\left[U_{ij} | X_{ij}, D_j, Z_j, \lambda_j\right], . \tag{21}$$

Recall that the finite grouping structure from the first step cannot retrieve the specific values of λ_j. Thus, additional assumption is made on θ and g. Let $\theta = \big(\theta^1, \cdots, \theta^K\big)$ and $\theta_j = \sum_{k=1}^{K} \theta^k \mathbf{1}\{\lambda_j = \lambda^k\}$. With some $\tilde{g} : \mathbb{R}^p \times \{0,1\} \times \mathbb{R}^{p^{cl}} \to \mathbb{R}$,

$$g(x, d, z, \lambda^k; \theta) = \tilde{g}(x, d, z; \theta_j). \tag{22}$$

The parametric model in (20)-(22) adds restrictions on H. From this model, I construct a least-square estimator as follows:

$$\hat{\theta} = \left(\hat{\theta}^1, \cdots, \hat{\theta}^K\right) = \arg\min_{\theta \in \Theta} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(Y_{ij} - \tilde{g}(X_{ij}, D_j, Z_j; \theta^{\hat{k}_j})\right)^2. \tag{23}$$

Again, each of the estimator $\hat{\theta}^k$ does not directly estimate θ^k; $\hat{\theta}$ as a whole estimates $\left(\theta^1, \cdots, \theta^K\right)$, up to a relabeling.

Remark 3. Though the conditional treatment effect parameters are not directly estimated here, a sufficiently flexible parametric model $\tilde{g}$ addresses aggregate heterogeneity and individual heterogeneity in treatment effect. $\theta \mapsto \tilde{g}(x, 1, z, \theta) - \tilde{g}(x, 0, z, \theta)$ captures aggregate heterogeneity and $x \mapsto \tilde{g}(x, 1, z, \theta) - \tilde{g}(x, 0, z, \theta)$ capture individual heterogeneity.

Remark 4. A direct connection to the group fixed-effect estimators can be made here. The parametric model in this paper can be understood as a group fixed-effects where a unit fixed-effect θ_j takes one of the K values: $\theta^1, \cdots, \theta^K$. In this sense, the least-square estimator in (23) is a group fixed-effect estimators.

Example 1. An example of the parametric model discussed here is a linear regression model with group-specific time fixed-effects and group-specific slope coefficients.

$$\begin{aligned} Y_{ij} &= g(X_{ij}, D_j, Z_j, \lambda_j; \theta_0) + U_{ij} \\ &= \delta_j + \beta_j D_j + {Z_j}^\intercal \eta^{cl} + {X_{ij}}^\intercal \eta + U_{ij}, \\ 0 &= \mathbf{E}\left[U_{ij} | X_{ij}, D_j, Z_j, \lambda_j\right]. \end{aligned}$$

where $\delta_j = \sum_{k=1}^{K} \delta^k \mathbf{1}\{\lambda_j = \lambda^k\}$ and $\beta_j = \sum_{k=1}^{K} \beta^k \mathbf{1}\{\lambda_j = \lambda^k\}$. The parameter of the model is $\theta = \left(\delta^1, \cdots, \delta^K, \beta^1, \cdots, \beta^K, \eta^{cl}, \eta\right)$ and $\theta^k = \left(\delta^k, \beta^k, \eta^{cl}, \eta\right)$. In Section 7, I extend the cross-sectional linear regression model to panel data linear regression model.

3.3 Alternative estimators

The K-means grouping structure is by no means the only way to implement the *selection-on-distribution* approach. There are other functional regression methods that we can use to run a regression on distribution. Firstly, there is a kernel estimator: Póczos et al. (2013). With some

tuning parameter h_F and kernel κ,

$$\hat{\pi}^{\kappa}(\mathbf{F}) = \frac{\sum_{j=1}^{J} D_j \kappa\big(\|\mathbf{F} - \hat{\mathbf{F}}_j\|_{w,2}/h_F\big)}{\sum_{j=1}^{J} \kappa\big(\|\mathbf{F} - \hat{\mathbf{F}}_j\|_{w,2}/h_F\big)}$$

estimates the propensity score of a cluster with given distribution $\mathbf{F}$. Then, the inverse probability weighting estimators can be constructed as before. Note that the kernel estimator does not have the dimension reduction property.

Secondly, there is functional principal component analysis (functional PCA): Delicado (2011); Hron et al. (2016); Kneip and Utikal (2001). Functional PCA constructs the following $J \times J$ matrix M whose j-th row l-th column element is

$$M_{jl} = \left\|\hat{\mathbf{F}}_j - \hat{\mathbf{F}}_l\right\|_{w,2} \quad \text{or} \quad \left\|\hat{\mathbf{f}}_j - \hat{\mathbf{f}}_l\right\|_{w,2}$$

where $\hat{\mathbf{f}}_j$ is the estimated density function of cluster j. Then, by choosing the first r largest singular values of M, with some predetermined $r \leq J$, functional PCA maps $\hat{\mathbf{F}}_j$ or $\hat{\mathbf{f}}_j$ to a r-dimensional factor: dimension reduction. Building on functional PCA, one can solve the K-means minimization problem in terms of the euclidean distance between the r-dimensional factors for each cluster; spectral clustering. By matching cluster with the estimated factor itself or the grouping variable from the spectral clustering, nonparametric estimation is possible. Also, since functional PCA has nice dimension reduction property, we may use the factor directly in a parametric model.

Thirdly, another alternative with the dimension reduction property is regularized regressions 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 $K \gg J$, regress

$$D_j = \beta_1 \mu_1(\hat{\mathbf{F}}_j) + \cdots + \beta_K \mu_K(\hat{\mathbf{F}}_j) + V_j$$

with LASSO. Suppose LASSO selects $\tilde{K}$ variables: $\{k_1, \cdots, k_{\tilde{K}}\} \subset \{1, \cdots, K\}$. Then, the variable selection property has reduced the dimension from the $K \times 1$ vector $\big(\mu_1(\hat{\mathbf{F}}_j), \cdots \mu_K(\hat{\mathbf{F}}_J)\big)$ to a $\tilde{K} \times 1$ vector $\big(\mu_{k_1}(\hat{\mathbf{F}}_j), \cdots, \mu_{k_{\tilde{K}}}(\hat{\mathbf{F}}_J)\big)$ and selected the moments of X that are relevant for treatment assignment. Again, we can use the selected moments to match clusters for nonparametric estimation, or use the selected moments in a parametric model.

Compared to these alternative estimation strategies, the estimation strategy based on the K-means algorithm has several definitive benefits. First of all, the grouping from the K-means al-

gorithm by itself is an interesting descriptive statistics. The grouping from the K-means gives us clearly defined "controls" in estimating treatment effect. In the case of the kernel estimator, for example, the 'control' would be some nonexistent hypothetical cluster that is constructed to be a weighted average of untreated clusters. Under the discrete structure of the K-means grouping, a researcher clearly sees which untreated clusters are used as a 'control' for a given treated cluster. This simple structure of finite grouping also gives us nice visual representations that help the audience understand the data structure, as will be shown in Section 7.

Secondly, I can derive theoretical results on asymptotic behavior of the K-means estimators, using the finite grouping structure assumption in Assumption 3. A vast literature has studied estimators motivated from a finite grouping structure and justification for the assumptions used to derive desirable asymptotic properties has been made with regard to models with economic interpretation: Hahn and Moon (2010). In this sense, the finite grouping structure assumption in Assumption 3 helps me with developing theoretical results for the induced estimators while being in touch with the economic insight.

Thirdly, the finite grouping structure from the K-means algorithm can motivate parametric models with group fixed-effects. As discussed in Section 1, a linear regression model with cluster fixed-effects is not identified in a clustered treatment context due to the multicollinearity problem. Given the preference for a parsimonious model among empirical researchers, the adaptation of the linear regression model with cluster fixed-effect to accommodate the restrictions imposed from clustered treatment assignment would be appealing. The finite grouping structure assumption from Assumption 3 and the K-means algorithm as estimation strategy directly motivate the use of group fixed-effects and allow empirical researchers to develop a parametric model that suits their data contexts while allowing for the aggregated individual-level information to enter the model in a parsimonious way.

Lastly, the dimension reduction assumption in the K-means algorithm has a straightforward interpretation; the number of groups K is the degree of discretization. For example, $K = 3$ means that a researcher believes that there exist three distinctive patterns of the distribution $\mathbf{F}_j$ among J clusters. On the other hand, for example, a modification of the sparsity assumption used in LASSO that only a small number of moments of the distribution are relevant for treatment assignment has little interpretable implications.

4 Asymptotic results

In this section, I discuss asymptotic properties of treatment effect estimators from Section 3. Firstly, I introduce Assumption 4 to discuss the asymptotic behavior of the first step grouping structure estimator.

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

a) *(no measure zero type) $\mu(\lambda^k) := \Pr\left\{\lambda_j = \lambda^k\right\} > 0 \ \forall k$.*

b) *(overlap) There exists some $\eta \in (h, 0.5)$ such that $\eta \leq \pi(\lambda^k) \leq 1 - \eta$ for every k.*

c) *(sufficient separation) For every $k \neq k'$,*

$$\left\|G(\lambda^k) - G(\lambda^{k'})\right\|_{w,2}^2 =: c(k, k') > 0.$$

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

e) *For any $\varepsilon > 0$,*

$$\Pr\left\{\varepsilon < \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2\right\} \leq C_1 \exp\left(-C_2 N_{\min,J}\varepsilon\right)$$

with some $C_1, C_2 > 0$ that do not depend on j.

Also,

$$\mathbf{E}\left[N_j\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2\right] \leq M.$$

for large J.

Assumption 4.a) ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. Assumption 4.b) assumes that we have (uniform) overlap across treated clusters and control clusters, for each value of the latent factor. Under Assumption 4.c), 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 4.d) assumes that the size of clusters goes to infinity as the number of clusters goes to infinity. This assumption limits our attention to cases where clusters are large. It should be noted that Assumption 4.d) excludes cases where the size of cluster increases only for some clusters and is fixed for some other clusters; the estimator

$\hat{\mathbf{F}}_j$ improves uniformly as J increases. Assumption 4.e) discusses the properties of the empirical distribution function $\hat{\mathbf{F}}_j$. The first part assumes that the tail probability of the distance between $\hat{\mathbf{F}}_j$ and $\mathbf{F}_j$ in terms of $\|\cdot\|_{w,2}$ goes to zero exponentially. The second part assumes that the distance is bounded in expectation when normalized with N_j.

Theorem 1 derives a rate on the probability of the first step grouping from the K-means algorithm retrieving the latent factor.

Theorem 1. *Under Assumptions 1-4, up to some relabeling on* Λ*,*

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

for any $\nu > 0$ *as* $J \to \infty$*.*

Proof. See Appendix. □

Theorem 1 shows that the probability of the first step grouping from the K-means algorithm making a mistake such that clusters with different values of λ_j are grouped together goes to zero when $J/{N_{\min,J}}^{\nu^*}$ goes to zero for some ν^*. Thus, when ${N_{\min,J}}^{\nu^*}$ increases faster than J for some $\nu^* > 0$, we can use the grouping from the first step as if the true values of λ_j are known to us.

Now, I prove asymptotic normality of the nonparametric treatment effect estimators under some regular assumptions. Before stating the formal assumptions, find that for any (d, k), the expectation of $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\bar{Y}_j$ is equal to the expectation of $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j = \lambda^k\right]$:

$$\begin{aligned}
&\mathbf{E}\left[\frac{\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}}{N_j}\sum_{i=1}^{N_j}\left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j = \lambda^k\right]\right)\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\frac{\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}}{N_j}\sum_{i=1}^{N_j}\left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j = k\right]\right)\Big|D_j, N_j, \lambda_j\right]\right] \\
&= \mathbf{E}\left[\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\left(\mathbf{E}\left[\bar{Y}_j|D_j, N_j, \lambda_j\right] - \mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j = \lambda^k\right]\right)\right] = 0
\end{aligned}$$

from Assumption 2, under some finite moments assumptions on $\mathbf{E}\left[\bar{Y}_j|D_j, N_j, \lambda_j\right]$. Assumption 5 formalizes the finite moments assumptions and assumes asymptotic normality on the difference.

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

a) $\mathbf{E}\left[{Y_{ij}}^2|X_{ij}, D_j, N_j, \lambda_j\right] < M$ *and* $\mathbf{E}\left[\bar{Y}_j^2\middle|\{X_{ij}\}_{i=1}^{N_j}, D_j, N_j, \lambda_j\right] < M$ *uniformly.*

b) *$N/J - \mathbf{E}_J[N_j] = o_p(1)$ as $J \to \infty$. Also, $\mathbf{E}_J[N_j] \leq MN_{\min,J}$ for large J.*

c) *Let*

$$W_j^{cl} = \begin{pmatrix} \sqrt{\frac{\mathbf{E}[N_j]}{N_j}} \frac{D_j \mathbf{1}\{\lambda_j = \lambda^1\}}{\sqrt{N_j}} \sum_{i=1}^{N_j} \left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(1) | N_j, \lambda_j = \lambda^1\right]\right) \\ \vdots \\ \sqrt{\frac{\mathbf{E}[N_j]}{N_j}} \frac{(1-D_j)\mathbf{1}\{\lambda_j = \lambda^K\}}{\sqrt{N_j}} \sum_{i=1}^{N_j} \left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(0) | N_j, \lambda_j = \lambda^K\right]\right) \end{pmatrix}$$

Then,

$$\frac{1}{\sqrt{J}} \sum_{j=1}^{J} W_j^{cl} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \Sigma_{W^{cl}}\right)$$

as $J \to \infty$, with

$$\Sigma_{W^{cl}} = \lim_{J \to \infty} \mathrm{Var}_J\left(W_j^{cl}\right).$$

d) *Let*

$$W_j = \begin{pmatrix} \sqrt{\frac{N_j}{\mathbf{E}[N_j]}} \frac{D_j \mathbf{1}\{\lambda_j = \lambda^1\}}{\sqrt{N_j}} \sum_{i=1}^{N_j} \left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(1) | N_j, \lambda_j = \lambda^1\right]\right) \\ \vdots \\ \sqrt{\frac{N_j}{\mathbf{E}[N_j]}} \frac{(1-D_j)\mathbf{1}\{\lambda_j = \lambda^K\}}{\sqrt{N_j}} \sum_{i=1}^{N_j} \left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j(0) | N_j, \lambda_j = \lambda^K\right]\right) \end{pmatrix}$$

Then,

$$\frac{1}{\sqrt{J}} \sum_{j=1}^{J} W_j \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \Sigma_W\right)$$

as $J \to \infty$, with

$$\Sigma_W = \lim_{J \to \infty} \mathrm{Var}_J\left(W_j\right).$$

Assumption 5.a) puts a bound on the conditional first and second moments of Y_{ij} and $\bar{Y}_j$. Assumption 5.b) assumes that N/J is a consistent estimator of $\mathbf{E}[N_j]$ and the ratio of the average cluster size $\mathbf{E}[N_j]$ and the minimum cluster size $N_{\min,J}$ cannot diverge. Assumption 5.c-d) assume asymptotic normality on $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\bar{Y}_j$, with relevant rescaling with regard to the cluster size. Note that the expectation of N_j and the variance of W_j is subscripted with J to denote that they depend on J.

Corollary 1. *Suppose $J/N_{\min,J}^{\nu*} \to 0$ as $J \to \infty$ for some $\nu^* > 0$. Under Assumptions 1-4 and Assumption 5.a-c), up to some relabeling on Λ,*

$$\sqrt{N}\begin{pmatrix} \widehat{CATE}^{cl}(1) - \overline{CATE}^{cl}(\lambda^1) \\ \vdots \\ \widehat{CATE}^{cl}(K) - \overline{CATE}^{cl}(\lambda^K) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \Sigma^{cl}\right)$$

as $J \to \infty$, where

$$\overline{CATE}^{cl}(\lambda^k) = \frac{\sum_{j=1}^J \mathbf{E}\left[\bar{Y}_j(1)|N_j, \lambda_j = \lambda^k\right] D_j \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\lambda_j = \lambda^k\}} - \frac{\sum_{j=1}^J \mathbf{E}\left[\bar{Y}_j(0)|N_j, \lambda_j = \lambda^k\right] (1-D_j) \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\lambda_j = \lambda^k\}}.$$

It directly follows that

$$\sqrt{N}\left(\widehat{ATE}^{cl} - \overline{ATE}^{cl}\right) \xrightarrow{d} \mathcal{N}\left(0, {\sigma^{cl}}^2\right)$$

as $J \to \infty$, where $\overline{ATE}^{cl}$ is the weighted average of $\overline{CATE}^{cl}$ with $\frac{1}{J}\sum_{j=1}^J \mathbf{1}\{\lambda_j = \lambda^k\}$ as weights. Also, under Assumptions 1-4 and Assumption 5.a-b,d),

$$\sqrt{N}\left(\widehat{ATE} - \overline{ATE}\right) \xrightarrow{d} \left(0, \sigma^2\right)$$

as $J \to \infty$, where

$$\overline{ATE} = \sum_{k=1}^K \frac{\sum_{j=1}^J \mathbf{1}\{\lambda_j = \lambda^k\}}{N} \left(\frac{\sum_{j=1}^J \mathbf{E}[\bar{Y}_j(1)|N_j, \lambda_j = \lambda^k] D_j N_j \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\lambda_j = \lambda^k\}} - \frac{\sum_{j=1}^J \mathbf{E}[\bar{Y}_j(0)|N_j, \lambda_j = \lambda^k](1-D_j) N_j \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\lambda_j = \lambda^k\}} \right)$$

Proof. See Appendix. □

Remark 5. By repeating the same argument that connects the asymptotic normality of $\widehat{CATE}^{cl}(k)$ to that of $\widehat{CATE}^{cl}$, but with different weighting, I derive the asymptotic normality of $\widehat{ATT}^{cl}$ and similarly for $\widehat{ATT}$.

Remark 6. The closed-form expression of the asymptotic variances are as follows:

$$\Sigma^{cl} = \begin{pmatrix} \frac{1}{\pi(\lambda^1)\mu(\lambda^1)} & -\frac{1}{(1-\pi(\lambda^1))\mu(\lambda^1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{(1-\pi(\lambda^K))\mu(\lambda^K)} \end{pmatrix} \Sigma_{W^{cl}} \begin{pmatrix} \frac{1}{\pi(\lambda^1)\mu(\lambda^1)} & -\frac{1}{(1-\pi(\lambda^1))\mu(\lambda^1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{(1-\pi(\lambda^K))\mu(\lambda^K)} \end{pmatrix}^{\intercal},$$

$$\sigma^{cl\,2} = \left(\mu(\lambda^1), \cdots, \mu(\lambda^K)\right) \Sigma^{cl} \left(\mu(\lambda^1), \cdots, \mu(\lambda^K)\right)^{\intercal}$$

and similarly for σ^2. Given consistent estimators for $\Sigma_{W^{cl}}$ and Σ_W, consistent estimators for $\Sigma^{cl}, \sigma^{cl}, \sigma$ can be constructed.

With Corollary 1, we have the consistency and the asymptotic normality of the treatment effect estimators. Note that the target parameter in the asymptotic distribution is a weighted sum of *conditional* treatment effects. This is because the asymptotic distributions in Corollary 1 are at the rate of $\sqrt{N}$: the variation from the cluster-level variables such as N_j is approximated to the population mean at the rate of $\sqrt{J}$, not $\sqrt{N}$.

When the potential outcomes are conditionally mean independent of the cluster size, i.e.,

$$\mathbf{E}\left[\bar{Y}(d)|N_j, \lambda_j = \lambda^k\right] = \mathbf{E}\left[\bar{Y}(d)|\lambda_j = \lambda^k\right]$$

for every k, the target parameters in the asymptotic distributions reduce down to the treatment effect parameters defined in Section 2.

$$\begin{aligned} \overline{CATE}^{cl}(\lambda^k) &= \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)|\lambda_j = \lambda^k\right] = CATE^{cl}(\lambda^k), \\ \overline{ATE}^{cl} &= \sum_{k=1}^{K} \frac{\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = \lambda^k\}}{J} CATE^{cl}(\lambda^k), \\ \overline{ATE} &= \sum_{k=1}^{K} \frac{\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = \lambda^k\}}{J} \left(\frac{\sum_{j=1}^{J} D_j N_j \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{\lambda_j = \lambda^k\}} \Big/ \frac{N}{J} CATE^{cl}(\lambda^k) \right. \\ &\qquad \left. \frac{\sum_{j=1}^{J} (1-D_j) N_j \mathbf{1}\{\lambda_j = \lambda^k\}}{\sum_{j=1}^{J} (1-D_j) \mathbf{1}\{\lambda_j = \lambda^k\}} \Big/ \frac{N}{J} CATE^{cl}(\lambda^k) \right). \end{aligned}$$

It is straightforward to see that the weights on the target parameter $\overline{ATE}^{cl}$ are sensible: the weights

are sample analogues of $\mu(\lambda^k)$, the population weights for ATE^{cl}.

$$\begin{aligned} ATE^{cl} &= \mathbf{E}[\bar{Y}_j(1) - \bar{Y}_j(0)] \\ &= \sum_{k=1}^{K} \mu(\lambda^k) \cdot CATE^{cl}(\lambda^k). \end{aligned}$$

In the case of $\overline{ATE}$, the weights on $\mathbf{E}\left[\bar{Y}_j(1)|\lambda_j = \lambda^k\right]$ are sample analogues of

$$\mu(\lambda^k) \cdot \mathbf{E}\left[N_j | D_j = 1, \lambda_j = \lambda^k\right] / \mathbf{E}[N_j].$$

When the cluster size is conditionally mean independent of the treatment status, i.e.

$$\mathbf{E}\left[N_j | D_j, \lambda_j = \lambda^k\right] = \mathbf{E}[N_j | \lambda_j = \lambda^k],$$

for every k,

$$\begin{aligned} ATE &= \mathbf{E}\left[\frac{N_j}{\mathbf{E}[N_j]}\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\frac{N_j}{\mathbf{E}[N_j]}\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)|N_j, \lambda_j\right]\right] \\ &= \mathbf{E}\left[\frac{N_j}{\mathbf{E}[N_j]}\mathbf{E}\left[\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)|\lambda_j\right]\right] \\ &= \sum_{k=1}^{K} \mu(\lambda^k) \frac{\mathbf{E}[N_j|\lambda_j = \lambda^k]}{\mathbf{E}[N_j]} \cdot CATE^{cl}(\lambda^k). \end{aligned}$$

Both of the target parameters $\overline{ATE}^{cl}$ and $\overline{ATE}$ can be thought of as the population parameter ATE^{cl} and ATE whose weights on $CATE^{cl}(\lambda^k)$ are replaced with their sample analogues.

Lastly, I show that the least-square estimator from the parametric model (20)-(22) is asymptotically normal, under regular assumptions on a GMM estimator.

Assumption 6. *Assume with some $M > 0$,*

a) *Θ, the parameter space of θ, is a compact subset of $\mathbb{R}^{rK}$.*
Also, the true value θ_0 lies in the interior of Θ.

b) *$(X_{ij}, U_{ij}) \,|\, (D_j, Z_j, \lambda_j) \sim$ iid.*

c) *$\theta = \left(\theta^1, \cdots, \theta^K\right)$ and there exists $\tilde{g} : \mathbb{R}^p \times \{0,1\} \times \mathbb{R}^{p^{cl}} \to \mathbb{R}$ such that for every k,*

$$g(x, d, z, \lambda^k; \theta) = \tilde{g}(x, d, z; \theta^k).$$

d) *(identification) Let g_θ be the first derivative of g with regard to θ.*

$$\mathbf{E}\left[(Y_{ij} - g(X_{ij}, D_j, Z_j, \lambda_j; \theta)) \cdot g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta)\right] = 0$$

only if $\theta = \theta_0$.

e) *(continuity of g) $\theta \mapsto g(x, d, z, \lambda; \theta)$ is twice continuously differentiable at every (x, d, z, λ).*

f) $\mathbf{E}\left[\sup_{\theta\in\Theta} \|g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta)\|_{sup}\right] < M,$
$\mathbf{E}\left[\sup_{\theta\in\Theta} \|g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta) g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta)^\intercal\|_{sup}\right] < M,$
$\mathbf{E}\left[\sup_{\theta\in\Theta} \|g_{\theta\theta^\intercal}(X_{ij}, D_j, Z_j, \lambda_j; \theta)\|_{sup}\right] < M.$

g) $\mathbf{E}\left[-g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta_0) g_\theta(X_{ij}, D_j, Z_j, \lambda_j; \theta_0)^\intercal\right]$
$+\mathbf{E}\left[(Y_{ij} - g(X_{ij}, D_j, Z_j, \lambda_j; \theta)) g_{\theta\theta^\intercal}(X_{ij}, D_j, Z_j, \lambda_j; \theta_0)\right]$ *has full rank.*

Assumption 6.a) assumes that the parameter space of θ is compact. Assumption 6.b) assumes that the individual-level control covariate X_{ij} and the idiosyncratic error U_{ij} are independently and identically distributed, after conditioning on the cluster-level covariates (D_j, Z_j, λ_j). Assumption 6.c) assumes that the latent factor λ_j is treated as a categorical variable in the model. Thanks to Assumption 6.c), the group membership variable $\hat{k}_j$ estimated as in Section 3 can be used to substitue for λ_j. Assumption 6.d-g) are the regularity assumptions for the infeasible GMM estimator.

Corollary 2. *Suppose $J/N_{\min,J}^{\nu*} \to 0$ as $J \to \infty$ for some $\nu* > 0$. Under Assumption 1-4, 5.a) and 6, up to some relabeling on Λ,*

$$\sqrt{N}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \Sigma^{gmm}\right)$$

as $J \to \infty$.

Proof. See Appendix. □

Remark 7. The $\sqrt{N}$-asymptotic variance of the least-square estimator is equal to the asymptotic variance of the infeasible oracle estimator with known $\lambda_1, \cdots, \lambda_J$. Suppose that there exists a consistent estimator for Σ^{gmm} when true $\lambda_1, \cdots, \lambda_J$ are known and that the estimator does not depend on the values of $\lambda_1, \cdots, \lambda_J$, but only depends on the induced grouping structure: for *Example 1*, the White estimator satisfies the conditions. Then, the naive approach of using the infeasible variance estimator under the estimated grouping structure $\hat{k}_1, \cdots, \hat{k}_J$ consistently estimates Σ^{gmm}.

5 Extension

5.1 Continuous λ

Throughout Sections 2-4, the support of the latent factor λ_j is assumed to be a finite set $\Lambda = \{\lambda^1, \cdots, \lambda^K\}$. With the finiteness assumption, the grouping structure based on $\hat{\mathbf{F}}_j$ can be directly thought of as an estimate of the latent factor λ_j. However, in some contexts, the assumption that Λ is finite, i.e. there are only finite types of clusters in terms of their distribution of X_{ij}, is not sensible. Thus, in this section, I discuss the asymptotic properties of the K-means treatment effect estimator when Λ is not a finite set, but a compact subset of $\mathbb{R}^q$. With this assumption, K is not a population parameter anymore; it is a tuning parameter that a researcher chooses in estimation.

Assumption 7. *Assume with some $M > 0$,*

a) *Either $\mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j\right] = \mathbf{E}\left[\bar{Y}_j(d)|\lambda_j\right]$ or $\mathbf{E}\left[D_j|N_j, \lambda_j\right] = \mathbf{E}\left[D_j|\lambda_j\right]$ with probability one.*

b) *(dimension of heterogeneity) Λ is a compact subset of $\mathbb{R}^q$.*

c) *(overlap) There exists some $\eta \in (h, 0.5)$ such that $\Pr\{\eta \leq \pi(\lambda_j) \leq 1 - \eta\} = 1$.*

d) *For any $\lambda, \lambda' \in \Lambda$ and $\alpha \in (0, 1)$, there exists $\lambda^* \in \Lambda$ such that*

$$\|\alpha G(\lambda) + (1 - \alpha)G(\lambda) - G(\lambda^*)\|_{w,2} = 0$$

Also, G and its inverse function are τ-Lipshitz:

$$\left\|G(\lambda) - G(\lambda')\right\|_{w,2} \leq \tau \left\|\lambda - \lambda'\right\|_2, \qquad \left\|\lambda - \lambda'\right\|_2 \leq \tau \left\|G(\lambda) - G(\lambda')\right\|_{w,2}.$$

e) *π is twice differentiable. $\frac{\partial^2}{\partial\lambda\partial\lambda^\intercal}\pi$ is uniformly bounded.*

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

g) *For large* J,

$$\mathbf{E}\left[N_j \left\| \hat{\mathbf{F}}_j - G(\lambda_j) \right\|_{w,2}^2\right] \leq M.$$

Assumption 7.a) assumes that either cluster-level mean of outcome variable Y_{ij} or treatment status variable D_j is mean independent of the cluster size N_j given the latent factor λ_j. Assumption 7.a) can easily be relaxed when the support for N_j is finite, by estimating the propensity score as a function of both N_j and $\hat{k}_j$: $\hat{\pi}(n,k)$. Assumption 7.d-e) assume that the clusters that are close to each other in terms of their distance measured with $\mathbf{F}_j = G(\lambda_j)$ should have similar λ_j and the functions G and π are smooth. Assumption 7.g) assumes that the empirical distribution function $\hat{\mathbf{F}}_j$ is a good estimate of the true distribution function $G(\lambda_j)$, when the cluster size N_j is large. Combined together, these conditions allow us to use the grouping structure based on $\hat{\mathbf{F}}_j$ as a good approximation of a grouping structure based on λ_j.

Theorem 2. *Under Assumptions 1-2, 5.a) and 7,*

$$\widehat{ATE}^{cl} - ATE^{cl} = O_p\left(\sqrt{\frac{K}{N_{\min,J}} + \frac{1}{K^{\frac{2}{q}}} + \frac{K}{J}}\right),$$
$$\widehat{ATT}^{cl} - ATT^{cl} = O_p\left(\sqrt{\frac{K}{N_{\min,J}} + \frac{1}{K^{\frac{2}{q}}} + \frac{K}{J}}\right),$$

as $J, K \to \infty$.

Proof. See Appendix. □

Theorem 2 characterizes the convergence rate of $\widehat{ATE}^{cl}$ and $\widehat{ATT}^{cl}$. The rate has three terms: $K/N_{\min,J}$, $1/K^{\frac{2}{q}}$ and K/J. The first term $K/N_{\min,J}$ is the variance of the distribution function estimator $\hat{\mathbf{F}}_j$. The second term $1/K^{\frac{2}{q}}$ is from the approximation bias of projecting Λ onto a grouping structure with finite K. The third term K/J is the variance of the propensity score estimator $\hat{\pi}(k)$. It is straightforward to see the classical bias-variance tradeoff in the choice of the tuning parameter K. When K is large, a continuous variable of λ_j is better approximated with a group membership variable $\hat{k}_j$, hence smaller bias, while the estimation of the propensity score worsens, hence larger variance.

Remark 8. The number of groups K in the first step of the estimation procedure is not a parameter of the model anymore; K is a tuning parameter. More discussion on the choice of K is in Appendix.

5.2 Generalized multilevel models

Another nontrivial direction of generalizing the model in hand is to allow for more than two levels. Suppose an econometrician observes

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

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

The researcher wants his model to incorporate the cluster-level heterogeneity and the hypercluster-level heterogeneity, in terms of the observables. To implement this multilevel property with the K-means algorithm, firstly construct the cluster-level distribution with individual-level control covariate as before: for every $x \in \mathbb{R}^p$,

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

Then, use the K-means algorithm to group clusters into K groups: $\hat{k}_{jl} \in \{1, \cdots, K\}$. Note that the grouping was done irrespective of each cluster's hypercluster membership: as long as $\hat{\mathbf{F}}_{jl}$ are the same, the subscript l does not matter. Then, the cluster-level observable information

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

which is high-dimensional, is summarized to

$$\Big(\hat{k}_{jl}, Z_{jl} \Big).$$

Given these cluster-level group membership $\hat{k}_{jl}$, construct the hypercluster-level distribution with

cluster-level observables: for every $z \in \mathbb{R}^{p^{cl}}$ and $k \in \{1 \cdots, K\}$,

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

By applying the K-means again to group the hyperclusters with K^{hyper}, which may not be equal to K, we reduce the dimension of the hypercluster-level observable

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

into

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

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

6 Monte Carlo simulations

In this section, I present two sets of Monte Carlo results where I apply the K-means estimators to simulated datasets and confirm the theoretical results from Section 4-5. In simulated datasets, I let each cluster to be of the same size: $N_1 = N_j$ for $j = 1, \cdots, J$. The data generating process I consider is as follows: given λ_j,

$$\begin{aligned} D_j \mid \lambda_j &\sim \text{Bernoulli}\left(\pi(\lambda_j)\right), \\ \left(U_{ij}, X_{ij}\right) \mid \left(D_j, \lambda_j\right) &\overset{\text{iid}}{\sim} \mathcal{N}\left((0, \lambda_j)^\intercal, I_2\right), \\ Y_{ij} &= \beta(\lambda_j) D_j + U_{ij} \end{aligned}$$

for $i = 1, \cdots, N_1$ and $j = 1, \cdots, J$ where

$$\pi(\lambda) = \frac{\lambda}{10} - \frac{\lambda}{20}\mathbf{1}\{\lambda \geq 0\} + \frac{1}{2}, \qquad \beta(\lambda) = \lambda - 2\lambda \mathbf{1}\{\lambda \geq 0\} + 3.$$

Figure 1 shows the propensity score π and the treatment effect β as functions of λ.

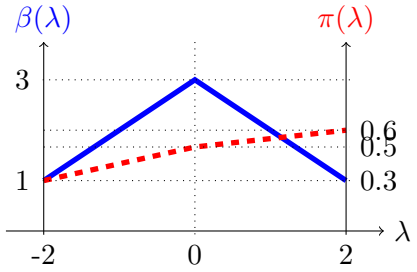

Figure 1: π and β

The red dashed line is the propensity score $\pi(\lambda)$; the blue solid lines is the treatment effect $\beta(\lambda)$. Both $\pi(\lambda)$ and $\beta(\lambda)$ have a kink at $\lambda = 0$.

Firstly, I let λ_j be discrete:

$$\Pr\{\lambda_j = \lambda\} = \frac{1}{4} \text{ for } \lambda = -1.5, -0.5, 0.5, 1.5.$$

I generate 2,000 datasets following the DGP and estimate the average treatment effect with $\widehat{ATE}^{cl}$ for each dataset. Table 1 contains simulation results across different values of the cluster size N_j while the number of clusters J is fixed to be 50 and then 200. The fourth column, which contains

J	N_j	Bias	rMSE	Coverage	prob. of grouping mistake
50	50	-0.005	0.091	0.939	0.126
	70	-0.004	0.088	0.939	0.036
	90	-0.009	0.090	0.950	0.018
	110	-0.007	0.088	0.952	0.009
	130	-0.005	0.084	0.954	0.013
	150	-0.007	0.086	0.952	0.011
200	50	0.001	0.041	0.950	0.317
	70	0.001	0.040	0.954	0.068
	90	0.000	0.039	0.968	0.016
	110	0.000	0.038	0.961	0.002
	130	0.001	0.037	0.975	0.002
	150	0.002	0.037	0.973	0.001

Table 1: Simulation results for $\widehat{ATE}^{cl}$, under discrete λ_j

J clusters are simulated with varying cluster size $N_j = 50, \cdots, 150$.
In the K-means grouping, $K = 4$ and 50 randomly drawn initial groupings were used.
The bias and rMSE are computed around the true value of $CATE^{cl} = 2$.

the simulated probability of the estimated grouping structure being different from the true grouping structure, directly relates to Theorem 1. As the cluster size N_j increase, the probability of making a mistake in the K-means grouping step goes to zero, for fixed J. Also, the 95% confidence interval constructed with the true asymptotic variance discussed in *Remark 6* contains the true $ATE^{cl} = 2$ with the probability of 0.95.

Secondly, I let λ_j be continuous:

$$\lambda_j \overset{\text{iid}}{\sim} \text{unif}[-2, 2].$$

Again, I generate 2,000 datasets and estimate ATE^{cl}. Table 2 contains simulation results across different values of the number of clusters J and the tuning parameter K, while N_j is fixed to be 150. As shown in the convergence rate of Theorem 2, larger J reduces the variance in estimating the propensity score, hence decreasing the rMSE. Also, Table 2 shows that bias and rMSE are U-shaped in terms of K: for $K \leq 5$, both bias and rMSE improve and for $K > 5$, both bias and rMSE worsen.

J	K	Bias	rMSE
30	5	-0.052	0.193
40	5	-0.024	0.142
50	5	-0.009	0.114
60	5	0.000	0.093
70	5	0.000	0.082
80	5	0.003	0.077
90	5	0.006	0.073
100	5	0.003	0.065
110	5	0.006	0.064
120	5	0.004	0.060

J	K	Bias	rMSE
75	3	0.011	0.084
75	4	0.008	0.080
75	5	0.002	0.078
75	6	-0.004	0.081
75	7	-0.005	0.086
75	8	-0.010	0.090
75	9	-0.014	0.090
75	10	-0.019	0.099
75	11	-0.026	0.104
75	12	-0.022	0.099

Table 2: Simulation results for $\widehat{ATE}^{cl}$, under continuous λ_j

For each cluster, 150 individuals were drawn randomly: $N_j = 150$.
In the K-means grouping, 50 randomly drawn initial groupings were used.
The bias and rMSE are computed around the true value of $CATE^{cl} = 2$.

7 Empirical illustration: effect of minimum wage on employment

7.1 Background

I apply the K-means two-step estimation strategy to revisit the question of whether an increase in minimum wage level leads to higher unemployment rate in the US labor market. 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 *selection-on-distribution* approach suggested in this paper to study the effect of minimum wage on employment, especially focusing on the heterogeneity in treatment effect. 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 *selection-on-distribution* assumption and the K-means estimation strategy suggested in this paper are a naturally appealing approach when studying the effect of the minimum wage.

7.2 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 susceptible 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 that will be use in the *selection-on-distribution*. 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,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{24}$$

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

Let us make two connections between (24) 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 1; 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, (24) 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 (24), I motivate a linear regression model with group fixed-effects, where each state is assigned to one of the K groups at each time t:

$$Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{25}$$

As implied with the use of $EmpRate_{jt}$ and from the dynamic programming example from Section 2, 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 apply the first step of the two-step estimation procedure of this paper and group states at each month using their distributions of individual-level employment history. Specifically, I use

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

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. $\tilde{X}_{ijt}$ collects $Emp_{ij\tau}$ for $\tau = t-4, \cdots, t-1$; $\tilde{X}_{ijt}$ is a four-month-long history of employment status. Since Emp_{ijt} is a categorical variable with a finite support of three elements, X_{ij} has a finite support of 81 elements. Note that $Y_{ijt} = 1 \Leftrightarrow Emp_{ijt} = Emp$ and thus $\tilde{X}_{ijt}$ can be understood as a vector of lagged outcome variables. To aggregate the information from $\tilde{X}_{ijt}$ to learn about the labor market fundamental of a given state, I collect $\tilde{X}_{ijt}$ for every individual and compute the empirical distribution function: for $x \in \mathcal{X}$,

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

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.[7] While (25) is the main regression specification of this paper, I also consider modifications of (25) to discuss treatment effect heterogeneity.

[7] I also consider a continuous control covariate in Appendix.

7.3 Results

7.3.1 Motivational snapshot

Before providing the pooled estimation results under the main regression specification, I illustrate how the K-means grouping step is implemented on an actual dataset, by looking at a snapshot of the pooled data. Out of the available 264 time periods, 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. By taking out a month of the pooled data and treating it as cross-section, I create a binary treatment D_j, where $D_j = 1$ means that state j increased their minimum wage starting January 1st, 2007:

$$D_j = \mathbf{1}\{MinWage_{j,Jan07} - MinWage_{j,Dec06}\}.$$

Since $\tilde{X}_{ijt}$ 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 2 contains the grouping result when $K = 3$. 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.

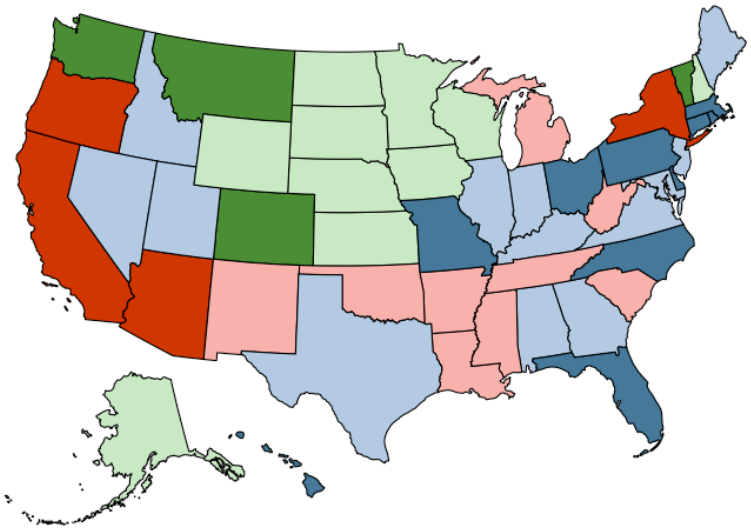

Figure 2: Grouping of states when $K = 3$, January 2007

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.

Table 3 and Figure 3 contain empirical evidence that the groups estimated using the distribution of $\tilde{X}_{ij,Jan07}$ are heterogeneous. Table 3 takes three subsets of $\mathcal{X}$ and computes the proportion of each subset across groups, putting equal weights over states. The three subsets are:

- Always-employed: $\{Emp\}^4$
- Ever-unemployed: $\{Emp, Unemp\}^4 \setminus \big(Emp, Emp, Emp, Emp\big)$
- 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.

In addition, Figure 3 takes the first and the last types of employment history, 'Always-employed' and 'Never-in-the-labor-force', and plots the states in terms of their state-level proportions. It is clear that there is negative correlation between the two types: the bigger the proportion of always-employed individuals is, the lower the proportion of never-in-the-labor-force individuals

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 3: 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.

is. Specifically, Group 1 states such as California and New York have lower proportion of always-employed and higher proportion of never-in-the-labor-force while Group 3 states such as Washington and Wisconsin have higher proportion of always-employed and lower proportion of never-in-the-labor-force.

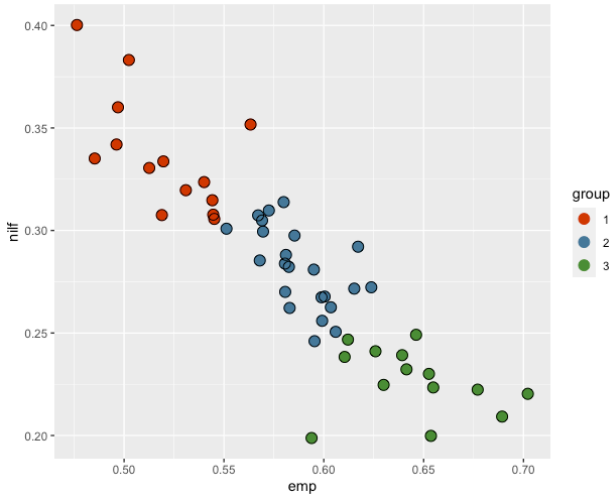

Figure 3: Heterogeneity across states, Janury 2007

This figure scatter plots 50 states and Washington D.C. The x-axis is the proportion of always-employed individuals in each state and the y-axis is the proportion of never-in-the-labor-force individuals in each state. Again, colors — red, blue, green — denote the estimated group.

Based on the K-means grouping, I estimate ATT^{cl} and $CATE^{cl}$ for each group. In estimation, instead of using $Y_{ij,Jan07}$ in level, I used the over-the-year difference outcome variables, to further control for state heterogeneity and seasonality: $Y_{ij}^{post} = Y_{ij,Jan07}$ is a binary employment status variable from January 2007 and $Y_{ij}^{pre} = Y_{ij,Jan06}$ is a binary employment status variable from

January 2006. The treatment effect estimator for each state is

$$\bar{Y}_j^{post} - \bar{Y}_j^{pre} - \left(\bar{Y}_{control}^{post} - \bar{Y}_{control}^{pre}\right).$$

$\bar{Y}_j$ is the sample mean of Y_{ij} for teens in state j and $\bar{Y}_{control}$ is the average of those sample means in the 'control' group, which is to be all of the untreated states for the DID estimator, and the untreated states from the same group for the K-means estimator. Note that by averaging the estimates within each group, I get the nonparmetric estimator $\widehat{CATE}^{cl}(k)$ from (12).

Table 4 contains the estimates. Overall, one percentage point raise in the minimum wage level leads to 0.291 percentage point decrease in the teen employment rate. Also, there seems to be a huge heterogeneity across states in terms of the employment history distribution. In Group 1 state, where the proportion of always-employed was low and the proportion of never-in-the-labor-force was high, the raise in the minimum wage level reduced the teen employment while in Group 3 states, the direction was the opposite. However, these findings are not statistically significant with t-test with the grouping structure as given, due to the small size of the dataset, except for $CATE^{cl}$ for Group 2.

	DID	K-means
ATT	-0.275	-0.291
	(0.189)	(0.191)
Group 1		-0.433
		(0.312)
Group 2		-0.396*
		(0.211)
Group 3		0.982
		(0.630)

Table 4: Impact of minimum wage on teen employment, January 2007

The table reports the effect of a raise in the minimum wage level on teen employment, by comparing state which raised their minimum wage levels in January 2007 with states which did not. The estimates are DID estimates where the over-the-time difference was made between state teen employment rate for January 2006 and that for January 2006, to control for seasonality.

The DID estimates are reweighted with the size of the minimum wage level increase, so that each estimate is interpreted to be an elasticity:

$$\frac{\bar{Y}_j^{post} - \bar{Y}_j^{pre} - \left(\bar{Y}_{control}^{post} - \bar{Y}_{control}^{pre}\right)}{\log MinWage_j^{Jan07} - \log MinWage_j^{Jan06}} \cdot \frac{1}{\bar{Y}^{pre}}.$$

The standard errors are computed at the state level. * denotes significance level 0.1.

7.3.2 Pooled least-square estimation

Now, I discuss the pooled estimation results from (25). For the pooled estimation, I repeated the K-means grouping step I did for January 2007 for every month from 2000 to 2021. Then, taking the estimated grouping structure as given, I ran the linear regression of (25). Table 5 contains the estimation result of the group fixed-effect specification, along with the estimation results for the TWFE specification and the census division fixed-effects specification as benchmarks. In the pooled estimation, the state minimum wage level $MinWage_{jt}$ is not converted into a binary treatment variable; the logarithm of $MinWage_{jt}$ is used as a regressor. Thus, by diving the slope coefficient on $\log MinWage_{jt}$ with the average teen employment rate from the pooled dataset, which is 0.326, we get the elasticity interpretation. Based on column (3), the preferred specification for pooled estimate, the elasticity of teen employment is -0.181, meaning that an one percentage point increase in the minimum wage level reduces teen employment by 0.18 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. The estimate from the group fixed-effect specification is slightly above the mean.

β	(1)	(2)	(3)	(4)	(5)	(6)
pooled	-0.024	-0.035**	-0.059***			
	(0.017)	(0.015)	(0.017)			
Group 1				-0.022	-0.034**	-0.066***
				(0.017)	(0.015)	(0.017)
Group 2				-0.024	-0.035**	-0.037**
				(0.017)	(0.015)	(0.019)
Group 3				-0.026	-0.038**	0.010
				(0.017)	(0.015)	(0.026)
δ_{jt}	TWFE	Census Div.	GFE	TWFE	Census Div.	GFE

Table 5: 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. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (3) and (6) contain the results from the preferred specification.

Columns (4), (5) and (6) report the group-specific minimum wage effect. Group 1 is the group of states with lower employment rate and lower labor force participation rate while Group 3 is the group of states with higher employment rate and higher labor force participation rate.

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.

The columns (4)-(6) of Table 5 discuss the aggregate heterogeneity in treatment effect:

$$Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \beta(\hat{k}_{jt}) \log MinWage_{jt} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{26}$$

Note that the slope coefficient for $\log MinWage_{jt}$ is a function of the estimated group membership variable $\hat{k}_{jt}$. In the main regression specification, the labels of groups across different time periods did not matter; the group membership variable $\hat{k}_{jt}$ only entered the regression through time fixed-effects. However, in (26), states with the same 'label' of group across different time periods are pooled together to estimate $\beta_1, \beta_2, \beta_3$. Thus, I relabeled the grouping structure from each time period and connect groups across months based on their relative position so that Group 1 is always the group of states with lower employment rate and lower labor force participation rate and Group 3 is always the group of states with higher employment rate and higher labor force participation rate.

Column (6) shows 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 more affected by the minimum wage and their counter parts in Group 3. We see that the labor market fundamental measured with the employment history distribution affects the treatment effect in a way that lower employment rate and lower labor force participation rate leads to bigger disemployment effect of the minimum wage increase among teens.

In addition to aggregate heterogeneity, I further extend (25)-(26) to discuss individual heterogeneity and aggregate heterogeneity simultaneously. The left panel of Table 6 estimates

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} &+ \beta_{yt} \log MinWage_{jt} \mathbf{1}\{Age_{ijt} \leq 18\} \\ &+ \beta_{ot} \log MinWage_{jt} \mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{27}$$

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 6 estimates

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} &+ \sum_{k=1}^{3} \beta_{yt}(k) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} \leq 18, \hat{k}_{jt} = k\} \\ &+ \sum_{k=1}^{3} \beta_{ot}(k) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} = 19, \hat{k}_{jt} = k\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{28}$$

Interaction between individual heterogeneity in terms of age and aggregate heterogeneity in terms of employment history is introduced.

β	(1)	(2)	(3)	(4)	(5)	(6)
$Age_{ijt} \leq 18$	-0.032*	-0.043***	-0.067***			
	(0.017)	(0.016)	(0.017)			
× Group 1				-0.030*	-0.042**	-0.074***
				(0.017)	(0.016)	(0.017)
× Group 2				-0.032*	-0.044***	-0.045**
				(0.017)	(0.016)	(0.019)
× Group 3				-0.032*	-0.044***	-0.015
				(0.017)	(0.016)	(0.027)
$Age_{ijt} = 19$	0.002	-0.009	-0.034			
	(0.020)	(0.016)	(0.021)			
× Group 1				0.005	-0.007	-0.039**
				(0.020)	(0.017)	(0.019)
× Group 2				0.003	-0.009	-0.010
				(0.019)	(0.016)	(0.021)
× Group 3				-0.008	-0.020	0.008
				(0.018)	(0.016)	(0.026)
δ_{jt}	TWFE	Census Div.	GFE	TWFE	Census Div.	GFE

Table 6: 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. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (3) and (6) contain the results from the preferred specification.

Columns (4), (5) and (6) report the group-specific minimum wage effect, while interacting with race. Group 1 is the group of states with lower employment rate and lower labor force participation rate while Group 3 is the group of states with higher employment rate and higher labor force participation rate.

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 6 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. Though the individual heterogeneity in evident in all of the three specifications I considered, the interaction between the individual heterogeneity and the aggregate heterogeneity is most evident in the group fixed-effects

specification of Column (6). Younger teens tend to be more affected by a raise in the minimum wage level and that tendency is stronger for group 1 states where the employment rate and the labor force participation rate are lower whereas in group 3 states a raise in the minimum wage level does not really affect either of younger and older teens.

β	(1)	(2)	(3)	(4)	(5)	(6)
$White_{ij} = 1$	-0.055***	-0.070***	-0.091***			
	(0.019)	(0.017)	(0.019)			
× Group 1				-0.052***	-0.067***	-0.098***
				(0.019)	(0.018)	(0.019)
× Group 2				-0.055***	-0.069***	-0.069***
				(0.019)	(0.017)	(0.020)
× Group 3				-0.054*	-0.069***	-0.037
				(0.019)	(0.018)	(0.028)
$White_{ij} = 0$	0.060***	0.048***	0.023			
	(0.017)	(0.017)	(0.018)			
× Group 1				0.063***	0.051***	0.016
				(0.017)	(0.018)	(0.016)
× Group 2				0.062***	0.051***	0.048**
				(0.017)	(0.017)	(0.018)
× Group 3				0.050***	0.038**	0.067**
				(0.016)	(0.017)	(0.025)
δ_{jt}	TWFE	Census Div.	GFE	TWFE	Census Div.	GFE

Table 7: Impact of minimum wage on teen employment in terms of race, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment. The regression allows the minimum wage effect to differ across white teens and non-white teens. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (3) and (6) contain the results from the preferred specification.

Columns (4), (5) and (6) report the group-specific minimum wage effect, while interacting with age. Group 1 is the group of states with lower employment rate and lower labor force participation rate while Group 3 is the group of states with higher employment rate and higher labor force participation rate.

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 7 repeats the same regression specification, but in terms of race; Table 7 documents individual heterogeneity in terms of white teens against non-white teens. From the left panel of Table 7, 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. This finding is reasonable in the sense that a financial standing of a family should affect a teenager's labor market choices; non-white teens may have more financial burdens and thus the effect of increased labor supply from non-white teens can dominate the effect of decreased labor demand. Again, the racial disparity interacts with the labor market fundamentals. From Column (6) of Table 7, it is shown that the racial disparity persists across groups and interact with the aggregate heterogeneity in a way that a raise in the minimum wage level has insignificant disemployment effect on non-white teens of states with lower employment rate and lower labor force participation rate, but has statistically significant employment effect on non-whote teens of states with higher employment rate and higher labor force participation rate. For white teens, a raise in the minimum wage level has statistically significant disemployment effect in states with lower employment rate and lower labor force participation rate, but has insignificant disemployment effect in states with higher employment rate and higher labor force participation rate.

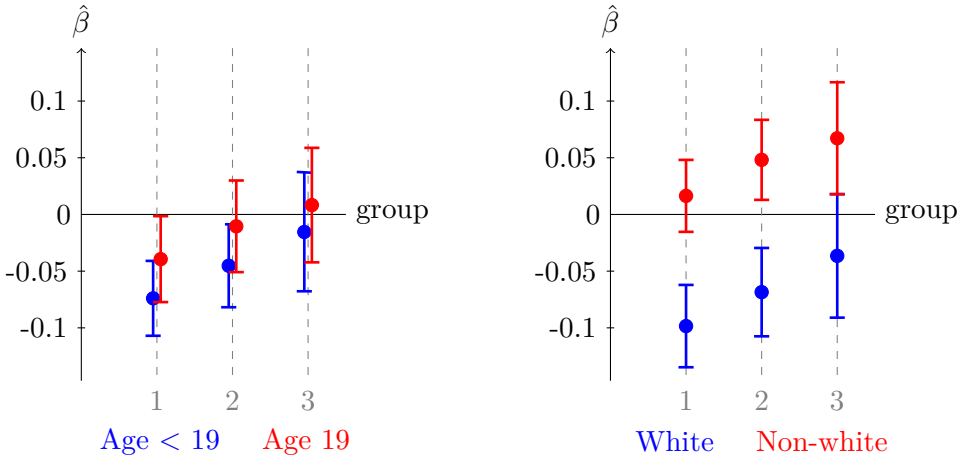

Figure 4: Interaction between individual and aggregate heterogeneity in minimum wage effect

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.

Figure 4 contains confidence intervals of treatment effect estimates from Column (6) of Table 6 and Column (6) of Table 7 and summarizes the interaction between individual heterogeneity in terms of age and race and aggregate heterogeneity in terms of employment history distribution. The individual heterogeneity in treatment effect is more evident in terms of race than age.

8 Conclusion

This paper extends the idea of the selection-on-observable assumption and motivates the *selection-on-distribution* assumption that individual-level potential outcomes are independent of cluster-level treatment after conditioning on the distribution of individual-level control covariate. Under the *selection-on-distribution* assumption, treatment effects are identified by comparing clusters with different treatment status but with the same distribution of individuals. 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 functional regression on distributions, the K-means algorithm is used in this paper. Though the K-means algorithm as a functional regression has several definitive benefits, application of an alternative functional regression method to the *selection-on-distribution* assumption would complement this paper by allowing for different sets of DGP assumptions on the cluster-level distribution. Also, this paper mostly focuses on cross-section data and non-dynamic panel data. 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. Lastly, there exist illustrative benefits to the K-means estimator even when the distribution of individuals is not thought to be discrete. This paper advocates the use of the K-means estimator in such contexts, though to a limited extent, with Theorem 2 where the K-means estimator is proven to be consistent when the latent factor for the distribution of individuals is continuous. Further discussion on asymptotic properties of the K-means estimator with a continuous latent factor would be an interesting direction for future research.

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.

Arthur, David and Sergei Vassilvitskii, "k-means++: The Advantages of Careful Seeding," Technical Report 2006-13, Stanford InfoLab June 2006.

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.

Bai, Jushan and Serena Ng, "Determining the number of factors in approximate factor models," *Econometrica*, 2002, *70* (1), 191–221.

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.

Bester, C Alan and Christian B Hansen, "Grouped effects estimators in fixed effects models," *Journal of Econometrics*, 2016, *190* (1), 197–208.

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

Bugni, Federico, Ivan Canay, Azeem Shaikh, and Max Tabord-Meehan, "Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes," *arXiv preprint arXiv:2204.08356*, 2022.

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.

Cattaneo, Matias D, Max H Farrell, and Yingjie Feng, "Large sample properties of partitioning-based series estimators," *The Annals of Statistics*, 2020, *48* (3), 1718–1741.

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.

Delicado, Pedro, "Dimensionality reduction when data are density functions," *Computational Statistics & Data Analysis*, 2011, *55* (1), 401–420.

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.

Graf, Siegfried and Harald Luschgy, "Rates of convergence for the empirical quantization error," *The Annals of Probability*, 2002, *30* (2), 874–897.

Hahn, Jinyong and Hyungsik Roger Moon, "Panel data models with finite number of multiple equilibria," *Econometric Theory*, 2010, *26* (3), 863–881.

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.

Hron, Karel, Alessandra Menafoglio, Matthias Templ, K Hrůzová, and Peter Filzmoser, "Simplicial principal component analysis for density functions in Bayes spaces," *Computational Statistics & Data Analysis*, 2016, *94*, 330–350.

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.

Newey, Whitney K and Daniel McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 1994, *4*, 2111–2245.

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

Póczos, Barnabás, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman, "Distribution-free distribution regression," in "Artificial Intelligence and Statistics" PMLR 2013, pp. 507–515.

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

Assume the following two assumptions:

(selection-on-observable)

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

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

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

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 1** follows immediately.

Proposition 1. *Under selection-on-observable and exchangeability,*

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

where

$$\hat{\mathbf{F}}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\{X_{ij} \leq x\}. \tag{29}$$

Proof. Firstly, find that $\mathbf{E}[D_j|\hat{\mathbf{F}}_j]$ is an weighted average of $\mathbf{E}[D_j|X_{\sigma(1)j}, \cdots, X_{\sigma(N_J)j}]$ across all possible permutations σ_J. Thus, under the *exchangeability*,

$$\mathbf{E}[D_j|\hat{\mathbf{F}}_j] = \mathbf{E}[D_j|X_{1j}, \cdots, X_{N_j j}] = \mathbf{E}[D_j|X_{\sigma(1)j}, \cdots, X_{\sigma(N_j)j}]$$

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

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

The third equality is from the selection-on-observable. □

Proposition 1 suggests propensity score matching based on $\hat{\mathbf{F}}_j$, the empirical distribution function of X_{ij} for cluster j. We can repeat the same argument with any mapping on $\{X_{1j}, \cdots, X_{N_j j}\}$ that is isomorphic to $\hat{\mathbf{F}}_j$: e.g. order statistics when $p = 1$.

B Addtional discussion on estimation strategy

B.1 Choice of initial values in the K-means grouping

Arthur and Vassilvitskii (2006) proposes an intuitive way of drawing an initial grouping for the naive K-means algorithm: K-means++

1. Randomly draw a cluster from $\{1, \cdots, J\}$ with uniform probability. Let j_1 denote the drawn cluster.

2. Given $\{j_1, \cdots, j_k\}$ from the k-th iteration, let

$$d^k(j) = \min_{1 \le k' \le k} \left\| \hat{\mathbf{F}}_j - \hat{\mathbf{F}}_{j_{k'}} \right\|_{w,2}^2.$$

3. Given d^k from Step 2, randomly draw a cluster from $\{1, \cdots, J\} \backslash \{j_1, \cdots, j_k\}$, with probability

$$\frac{d^k(j)}{\sum_{j'=1}^{J} d^k(j')}.$$

   Let j_{k+1} denote the drawn cluster.

4. Repeat Step 2-3 until $k = K$ and use $\hat{\mathbf{F}}_{j_1}, \cdots, \hat{\mathbf{F}}_{j_K}$ as the initial values $G^{(1)}(1), \cdots, G^{(1)}(K)$ for the naive algorithm.

The motivation for this approach is that a desirable initial grouping structure should already separate the clusters well. To that end, the K-means++ approach draws a cluster with probability proportional to the minimum distance to clusters that have already been chosen as $G^{(1)}(k)$.

B.2 Choice of K as tuning parameter

As discussed in Section 5, the number of groups used in the K-means grouping step is a tuning parameter under Assumption 7. I suggest the following cross-validation approach to choose K. Suppose we consider $K = K_{\min}, \cdots, K_{\max}$ as potential choices of the tuning parameter, with some $K_{\min}$ and $K_{\max}$.

1. Randomly split J clusters into L subsamples of an equal size: for each $l = 1, \cdots, L$, the l-th subsample $\mathcal{J}_l$ contains approximately J/L clusters. Depending on $\{N_j\}_{j=1}^J$, the number of individuals for each subsample may vary.

2. Given K and l, take the l-th subsample $\mathcal{J}_l$ as a 'test set' and take the rest $\{1, \cdots, J\} \setminus \mathcal{J}_l$ as a 'training set.' Apply the K-means algorithm to the training set and construct a grouping with K groups, denoted with $\hat{k}_j^{-l}$ and $\hat{G}(k)^{-l}$. Assign clusters from the test set to the groups constructed from the training set, based on $\|\cdot\|_{w,2}$: for each $j \in \mathcal{J}_l$

$$\hat{k}_j^{-l} = \min_{1,\cdots,K} \left\| \hat{\mathbf{F}}_j - \hat{G}^{-l}(k) \right\|_{w,2}.$$

   Evaluate the grouping structure with the test set:

$$\sum_{j \in \mathcal{J}_l} \left(\bar{Y}_j - \frac{D_j \sum_{j' \notin \mathcal{J}_l} \bar{Y}_{j'} D_{j'} \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}}{\sum_{j' \notin \mathcal{J}_l} D_{j'} \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}} - \frac{(1 - D_j) \sum_{j' \notin \mathcal{J}_l} \bar{Y}_{j'} (1 - D_{j'}) \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}}{\sum_{j' \notin \mathcal{J}_l} (1 - D_{j'}) \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}} \right)^2. \tag{30}$$

3. Repeat 2 for every $l = 1, \cdots, L$ and evaluate K by summing (30) over $l = 1, \cdots, L$:

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{J}_l} \left(\bar{Y}_j - \frac{D_j \sum_{j' \notin \mathcal{J}_l} \bar{Y}_{j'} D_{j'} \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}}{\sum_{j' \notin \mathcal{J}_l} D_{j'} \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}} - \frac{(1 - D_j) \sum_{j' \notin \mathcal{J}_l} \bar{Y}_{j'} (1 - D_{j'}) \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}}{\sum_{j' \notin \mathcal{J}_l} (1 - D_{j'}) \mathbf{1}\{\hat{k}_{j'}^{-l} = \hat{k}_j^{-l}\}} \right)^2. \tag{31}$$

4. Repeat 2-3 for every $K = K_{\min}, \cdots, K_{\max}$ and choose K that minimizes (31).

C Proofs

C.1 Theorem 1

For the convenience of notation, let us construct a new cluster-level variable k_j: $k_j = k \Leftrightarrow \lambda_j = \lambda^k$.

Step 1

WTS

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

From **A5.e)**,

$$\mathbf{E}\left[\frac{1}{J}\sum_{l=1}^{J} N_{\min,J}\left\|\hat{\mathbf{F}}_l - G(\lambda_l)\right\|_{w,2}^2\right] \leq \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}\left[N_j\left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2\right] \leq M$$

for large J.

Step 2

Let us connect $\hat{G}(1),\cdots,\hat{G}(K)$ to $G(\lambda^1),\cdots,G(\lambda^K)$. Define $\sigma(k)$ such that

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

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

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

The last inequality holds since $\sum_{j=1}^{J}\left\|\hat{G}(\hat{k}_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^2 \leq \sum_{j=1}^{J}\left\|G(\lambda^{k_j}) - \hat{\mathbf{F}}_j\right\|_{w,2}^2$ from the defini-

tion of $\hat{G}$ and $\hat{k}$. From **A5.a)**, $\sum_{j=1}^{J} \mathbf{1}\{k_j = k\}/J \to \mu(k) > 0$ as $J \to \infty$. Thus,

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

as $J \to \infty$ from **A5.d)** and Step 1.

For $k' \neq k$,

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

as $J \to \infty$ from **A5.c-d)** and Step 1.

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

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

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

Step 3

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

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

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

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

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

Then,

$$\begin{aligned}
\Pr\left\{\hat{k}_j \neq k_j, A_\varepsilon\right\} &\leq \Pr\left\{\exists\, k \neq k_j \text{ s.t. } \left\|\hat{G}(k) - \hat{\mathbf{F}}_j\right\|_{w,2} \leq \left\|\hat{G}(k_j) - \hat{\mathbf{F}}_j\right\|_{w,2}, A_\varepsilon\right\} \\
&\leq \Pr\left\{\exists\, k \neq k_j \text{ s.t. } \frac{1}{2}\left\|\hat{G}(k) - G(\lambda^{k_j})\right\|_{w,2}^2 - \left\|\hat{\mathbf{F}}_j - G(\lambda^{k_j})\right\|_{w,2}^2 \right. \\
&\qquad\qquad \left. \leq 2\left\|\hat{G}(k_j) - G(\lambda^{k_j})\right\|_{w,2}^2 + 2\left\|G(\lambda^{k_j}) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\} \\
&\leq \Pr\left\{\exists\, k \neq k_j \text{ s.t. } \frac{1}{4}\left\|G(\lambda^{\sigma^{-1}(k)=k}) - G(\lambda^{k_j})\right\|_{w,2}^2 - \frac{1}{2}\left\|\hat{G}(k) - G(\lambda^k)\right\|_{w,2}^2 \right. \\
&\qquad\qquad \left. \leq 2\left\|\hat{G}(k_j) - G(\lambda^{k_j})\right\|_{w,2}^2 + 3\left\|G(\lambda^{k_j}) - \hat{\mathbf{F}}_j\right\|_{w,2}^2, A_\varepsilon\right\}
\end{aligned}$$

The bijective-ness of σ is used in the third inequality to link $\left\|\hat{G}(k) - G(\lambda^{k_j})\right\|_{w,2}$ to $\left\|G(\lambda^k) - G(\lambda^{k_j})\right\|_{w,2}$: for every k, we can connect $\hat{G}(k)$ to $G(k)$. Then,

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

The second inequality is from **A5.c)**. The third 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. $(K-1)$ comes from repeating the argument for every $k \neq k_j$.

Set $\varepsilon^{**} = \frac{1}{20}\min_{k\neq k'} c(k,k')$ so that

$$\frac{1}{12}\min_{k\neq k'} c(k,k') - \frac{5}{6}\varepsilon^{**} = \frac{1}{24}\min_{k\neq k'} c(k,k') > 0.$$

By repeating the expansion for every j,

$$\begin{aligned}\Pr\left\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\right\} &\leq \Pr\left\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j, A_{\varepsilon^{**}}\right\} + \Pr\left\{{A_{\varepsilon^{**}}}^c\right\} \\ &\leq (K-1)\sum_{j=1}^{J}\Pr\left\{\frac{1}{24}\min_{h\neq h'} c(h,h') \leq \left\|G(\lambda^{k_j}) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2}\right\} + \Pr\left\{{A_{\varepsilon^{**}}}^c\right\}.\end{aligned}$$

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 $o(J/\min_j {N_j}^\nu)$ for any $\nu > 0$. Let ε^{***} denote $\frac{1}{24}\min_{k\neq k'} c(k,k')$. Choose an arbitrary $\nu > 0$. From **A5.e)**,

$$\begin{aligned}(K-1)\sum_{j=1}^{J}\Pr\left\{\varepsilon^{***} \leq \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^{\ 2}\right\} &\leq J(K-1)C_1\exp\left(-C_2 N_{\min,J}\varepsilon^{***}\right) \\ &= (K-1)C_1\cdot\left(\frac{J}{{N_{\min,J}}^\nu}\right)\cdot\frac{{N_{\min,J}}^\nu}{\exp\left(C_2 N_{\min,J}\varepsilon^{***}\right)}.\end{aligned}$$

Thus, for any $\nu > 0$, ${N_{\min,J}}^\nu/J \cdot \Pr\left\{\exists\ \hat{k}_j \neq k_j\right\} \to 0$ as $J \to \infty$.

C.2 Corollary 1

Let

$$\widetilde{CATE}^{cl}(k) = \frac{\sum_{j=1}^{J}\bar{Y}_j D_j \mathbf{1}\{k_j = k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{k_j = k\}} - \frac{\sum_{j=1}^{J}\bar{Y}_j (1-D_j)\mathbf{1}\{k_j = k\}}{\sum_{j=1}^{J}(1-D_j)\mathbf{1}\{k_j = k\}}$$

with some abuse of notation. I let

$$\widetilde{CATE}^{cl}(k) = \begin{cases} -\frac{\sum_{j=1}^{J}\bar{Y}_j(1-D_j)\mathbf{1}\{k_j=k\}}{(1-h)\sum_{j=1}^{J}\mathbf{1}\{k_j=k\}}, & \text{if } \sum_{j=1}^{J}\mathbf{1}\{k_j=k\} > 0 \text{ and } \sum_{j=1}^{J} D_j\mathbf{1}\{k_j=k\} = 0, \\ \frac{\sum_{j=1}^{J}\bar{Y}_j D_j\mathbf{1}\{k_j=k\}}{h\sum_{j=1}^{J}\mathbf{1}\{k_j=k\}}, & \text{if } \sum_{j=1}^{J}\mathbf{1}\{k_j=k\} > 0 \text{ and } \sum_{j=1}^{J}(1-D_j)\mathbf{1}\{k_j=k\} = 0, \\ 0, & \text{if } \sum_{j=1}^{J}\mathbf{1}\{k_j=k\} = 0 \end{cases}$$

This adaptation is made so that $\widetilde{CATE}^{cl}(k)$ is well-defined and identical to $\widehat{CATE}^{cl}(k)$, with respect to $\widehat{ATE}^{cl}$, under the same grouping structure. With $\widetilde{CATE}^{cl}(k)$, I make two claims:

$$\widetilde{CATE}^{cl}(k) - \overline{CATE}^{cl}(\lambda^k) = O_p\left(\frac{1}{\sqrt{N}}\right),$$
$$\widehat{CATE}^{cl}(k) - \widetilde{CATE}^{cl}(k) = o_p\left(\frac{1}{\sqrt{N}}\right).$$

as $J \to \infty$.

Claim 1

Firstly, find that

$$\widetilde{CATE}^{cl}(k) - \overline{CATE}^{cl}(\lambda^k)$$
$$= \frac{\sum_{j=1}^{J}\left(\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(1)|N_j, k_j = k\right]\right) D_j \mathbf{1}\{k_j = k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{k_j = k\}} - \frac{\sum_{j=1}^{J}\left(\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(0)|N_j, k_j = k\right]\right)(1 - D_j)\mathbf{1}\{k_j = k\}}{\sum_{j=1}^{J}(1 - D_j)\mathbf{1}\{k_j = k\}}$$

and

$$\sqrt{N}\left(\frac{\sum_{j=1}^{J}\left(\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(1)|N_j, k_j = k\right]\right) D_j \mathbf{1}\{k_j = k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{k_j = k\}}\right)$$
$$= \sqrt{\frac{N}{J\mathbf{E}\left[N_j\right]}} \cdot \frac{\frac{1}{\sqrt{J}} \cdot \sqrt{\frac{\mathbf{E}[N_j]}{N_j}} \cdot \frac{D_j \mathbf{1}\{k_j = k\}}{\sqrt{N_j}} \sum_{i=1}^{N_j}\left(Y_{ij} - \mathbf{E}\left[\bar{Y}_j|D_j = 1, N_j, k_j = k\right]\right)}{\frac{1}{J}\sum_{j=1}^{J} D_j \mathbf{1}\{k_j = k\}}$$

and similarly for the second quantity in $\widetilde{CATE}^{cl}(k) - \overline{CATE}^{cl}(\lambda^k)$. From **A6.b)**,

$$\frac{N}{J\mathbf{E}[N_j]} - 1 = o_p\left(\frac{1}{\mathbf{E}[N_j]}\right).$$

Thus, $\sqrt{\frac{N}{J\mathbf{E}[N_j]}} \xrightarrow{p} 1$ as $J \to \infty$. From **A1** and **A5.a-b)**,

$$\frac{1}{J}\sum_{j=1}^{J} D_j \mathbf{1}\{k_j = k\} \xrightarrow{p} \mathbf{E}\left[D_j \mathbf{1}\{k_j = k\}\right] = \pi(k)\mu(k) > 0$$

as $J \to \infty$. Thus, from **A6.c)**,

$$\widetilde{CATE}^{cl}(k) - \overline{CATE}^{cl}(\lambda^k) \xrightarrow{d} \mathcal{N}\left(0, e_k{}^{\intercal}\Sigma_{W^{cl}} e_k\right)$$

where e_k is a $(2K) \times 1$ column vectors whose components except for the $(2k-1)$-th and and $2k$-th components are zeroes. The $(2k-1)$-th component is $1/\pi(k)\mu(k)$ and the $2k$-th component is $1/(1-\pi(k))\mu(k)$. By repeating this for every k, we obtain

$$\begin{pmatrix} \widetilde{CATE}^{cl}(1) - \overline{CATE}^{cl}(\lambda^1) \\ \vdots \\ \widetilde{CATE}^{cl}(K) - \overline{CATE}^{cl}(\lambda^K) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma^{cl}\right)$$

where

$$\Sigma^{cl} = \begin{pmatrix} \frac{1}{\pi(1)\mu(1)} & -\frac{1}{(1-\pi(1))\mu(1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{(1-\pi(K))\mu(K)} \end{pmatrix} \Sigma_W \begin{pmatrix} \frac{1}{\pi(1)\mu(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{(1-\pi(K))\mu(K)} \end{pmatrix}.$$

The first claim has been proven.

Claim 2

It suffices to show the second claim to finish the proof. Find that $\widehat{CATE}^{cl}(k) = \widetilde{CATE}^{cl}(k)$ for every k if $\hat{k}_j = k_j$ for every j.

$$\begin{aligned} &\left|\widehat{CATE}^{cl}(k) - \widetilde{CATE}^{cl}(k)\right| \\ &= \left|\widehat{CATE}^{cl}(k) - \widetilde{CATE}^{cl}(k)\right| \mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\} \\ &\leq \left(\left|\widehat{CATE}^{cl}(k)\right| + \left|\widetilde{CATE}^{cl}(k)\right|\right) \mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\}. \end{aligned}$$

Firstly, find that the indicator function converge to zero in probability at a rate faster than $1/\sqrt{N}$. Fix $\varepsilon > 0$:

$$\begin{aligned} \Pr\left\{\sqrt{N}\mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\} > \varepsilon\right\} &\leq \Pr\left\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\right\} \cdot \frac{\sqrt{N}}{\varepsilon} \\ &= \Pr\left\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\right\} \sqrt{JN_{\min,J}} \sqrt{\frac{\mathbf{E}[N_j]}{N_{\min,J}}} \sqrt{\frac{N}{J\mathbf{E}[N_j]}} \frac{1}{\varepsilon}. \end{aligned}$$

From Theorem 1, with any $\nu > 0$,

$$\Pr\left\{\sqrt{N}\mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\} > \varepsilon\right\} \leq \Pr\left\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\right\} \frac{{N_{\min,J}}^{\nu}}{J} \cdot J^{\frac{3}{2}} {N_{\min,J}}^{\frac{1}{2}-\nu} \sqrt{\frac{\mathbf{E}[N_j]}{N_{\min,J}}} \sqrt{\frac{N}{J\mathbf{E}[N_j]}} \frac{1}{\varepsilon}$$
$$= J^{\frac{3}{2}} {N_{\min,J}}^{\frac{1}{2}-\nu} o(1) M(1 + o_p(1)) \frac{1}{\varepsilon}$$

for large J. By letting $\nu > \frac{3\nu^*+1}{2} > 0$,

$$\frac{J^{\frac{3}{2}}}{{N_{\min,J}}^{\nu-\frac{1}{2}}} \leq \frac{J^{\frac{3}{2}}}{{N_{\min,J}}^{\frac{3\nu^*}{2}}} = \left(\frac{J}{{N_{\min,J}}^{\nu^*}}\right)^{\frac{3}{2}} \to 0$$

as $J \to \infty$. Thus, $\sqrt{N}\mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\} = o_p(1)$.

It remains to show that $\left|\widehat{CATE}^{cl}(k)\right|$ and $\left|\widetilde{CATE}^{cl}(k)\right|$ are bounded in expectation:

$$\begin{aligned}
\mathbf{E}\left[\left|\widehat{CATE}^{cl}(k)\right|\right] &= \mathbf{E}\left[\left|\frac{\sum_{j=1}^J \bar{Y}_j D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\hat{k}_j = k\}} - \frac{\sum_{j=1}^J \bar{Y}_j (1-D_j) \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\hat{k}_j = k\}}\right|\right] \\
&\leq \mathbf{E}\left[\frac{\sum_{j=1}^J \left|\bar{Y}_j\right| D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\hat{k}_j = k\}} + \frac{\sum_{j=1}^J \left|\bar{Y}_j\right| (1-D_j) \mathbf{1}\{\hat{k}_j = k\}\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\hat{k}_j = k\}}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\frac{\sum_{j=1}^J \left|\bar{Y}_j\right| D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\hat{k}_j = k\}} + \frac{\sum_{j=1}^J \left|\bar{Y}_j\right| (1-D_j) \mathbf{1}\{\hat{k}_j = k\}\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\hat{k}_j = k\}} \middle| \left\{\{X_{ij}\}_{i=1}^{N_j}, D_j, N_j, k_j\right\}_{j=1}^J\right]\right] \\
&= \mathbf{E}\left[\frac{\sum_{j=1}^J \mathbf{E}\left[\left|\bar{Y}_j\right| \middle| \{X_{ij}\}_{i=1}^{N_j}, D_j, N_j, k_j\right] D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^J D_j \mathbf{1}\{\hat{k}_j = k\}}\right] \\
&\quad + \mathbf{E}\left[\frac{\sum_{j=1}^J \mathbf{E}\left[\left|\bar{Y}_j\right| \middle| \{X_{ij}\}_{i=1}^{N_j}, D_j, N_j, k_j\right] (1-D_j) \mathbf{1}\{\hat{k}_j = k\}\}}{\sum_{j=1}^J (1-D_j) \mathbf{1}\{\hat{k}_j = k\}}\right] \\
&\leq M.
\end{aligned}$$

The third equality is from **A1** and $\left\{\hat{k}_j\right\}_j$ being a function of $\left\{\{X_{ij}\}_{i=1}^{N_j}\right\}_{j=1}^J$. The last equality is from **A6.a)**. By repeating the same argument, $\mathbf{E}\left[\widetilde{CATE}^{cl}(k)\right]$ is bounded in expectation as well. Then,

$$\sqrt{N}\left|\widehat{CATE}^{cl}(k) - \widetilde{CATE}^{cl}(k)\right| = O_p(1) \cdot o_p(1)$$

as $J \to \infty$. By repeating this for every K,

$$\sqrt{N}\begin{pmatrix} \left|\widehat{CATE}^{cl}(1) - \widetilde{CATE}^{cl}(1)\right| \\ \vdots \\ \left|\widehat{CATE}^{cl}(K) - \widetilde{CATE}^{cl}(K)\right| \end{pmatrix} = O_p(1) \cdot o_p(1)$$

By combining the two claims in the beginning,

$$\sqrt{N}\begin{pmatrix} \widehat{CATE}^{cl}(1) - \overline{CATE}^{cl}(\lambda^1) \\ \vdots \\ \widehat{CATE}^{cl}(K) - \overline{CATE}^{cl}(\lambda^K) \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).$$

Averaging: $\widehat{ATE}^{cl}$

Find that, with some abuse of notations with zero denominators,

$$\begin{aligned}
\widehat{ATE}^{cl} &= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j}\right) \\
&= \sum_{k=1}^{K}\frac{1}{J}\left(\sum_{j=1}^{J}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}(k)} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}(k)}\right)\mathbf{1}\{\hat{k}_j = k\}\right) \\
&= \sum_{k=1}^{K}\frac{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}}{J}\left(\frac{\sum_{j=1}^{J}\bar{Y}_j D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\}} - \frac{\sum_{j=1}^{J}\bar{Y}_j(1-D_j)\mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J}(1-D_j)\mathbf{1}\{\hat{k}_j = k\}}\right) \\
&= \sum_{k=1}^{K}\frac{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}}{J} \cdot \widehat{CATE}^{cl}(k)
\end{aligned}$$

since $\hat{\pi}(k) = \sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\} / \sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}$. The asymptotic normality of $\widehat{ATE}^{cl}$ directly follows from repeating the two claims, with $\widehat{ATE}^{cl}$ and

$$\widetilde{ATE}^{cl} = \sum_{k=1}^{K}\frac{\sum_{j=1}^{J}\mathbf{1}\{k_j = k\}}{J} \cdot \widetilde{CATE}^{cl}(k).$$

Averaging: $\widehat{ATE}$

Again, with some abuse of notations with zero denominators,

$$\begin{aligned}
\widehat{ATE} &= \frac{1}{N}\sum_{j=1}^{J} N_j \left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j}\right) \\
&= \frac{\sqrt{\mathbf{E}[N_j]}}{N}\cdot\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{N_j}{\mathbf{E}[N_j]}}\cdot\sqrt{N_j}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j}\right) \\
&= \frac{\sqrt{\mathbf{E}[N_j]}}{N}\cdot\sum_{k=1}^{K}\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{N_j}{\mathbf{E}[N_j]}}\cdot\sqrt{N_j}\left(\frac{\bar{Y}_j D_j\mathbf{1}\{\hat{k}_j=k\}}{\hat{\pi}(k)} - \frac{\bar{Y}_j(1-D_j)\mathbf{1}\{\hat{k}_j=k\}}{1-\hat{\pi}(k)}\right) \\
&= \frac{\sqrt{\mathbf{E}[N_j]}}{N}\cdot\sum_{k=1}^{K}\frac{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j=k\}}{J}\sum_{j=1}^{J}\sqrt{\frac{N_j}{\mathbf{E}[N_j]}}\cdot\sqrt{N_j}\left(\frac{\bar{Y}_j D_j\mathbf{1}\{\hat{k}_j=k\}}{\sum_{j=1}^{J}D_j\mathbf{1}\{\hat{k}_j=k\}} - \frac{\bar{Y}_j(1-D_j)\mathbf{1}\{\hat{k}_j=k\}}{\sum_{j=1}^{J}(1-D_j)\mathbf{1}\{\hat{k}_j=k\}}\right).
\end{aligned}$$

By repeating the same argument for $\sqrt{N}\left(\widehat{ATE} - ATE\right)$, with

$$\widetilde{ATE} = \frac{1}{N}\sum_{j=1}^{J} N_j\left(D_j\bar{Y}_j\frac{\sum_{l=1}^{J}\mathbf{1}\{k_l=k\}}{\sum_{l=1}^{J}D_j\mathbf{1}\{k_l=k\}} - (1-D_j)\bar{Y}_j\frac{\sum_{l=1}^{J}\mathbf{1}\{k_l=k\}}{\sum_{l=1}^{J}(1-D_j)\mathbf{1}\{k_l=k\}}\right)$$

as an intermediary, we have the asymptotic normality of $\widehat{ATE}$.

C.3 Corollary 2

Consider an infeasible GMM estimator $\tilde{\theta}$:

$$\tilde{\theta} = \arg\min_{\theta\in\Theta}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\left(Y_{ij} - \tilde{g}(X_{ij}, D_j, Z_j; \theta^{k_j})\right)^2.$$

From Theorem 2.6. and 3.4. of Newey and McFadden (1994), we have the asymptotic normality for $\sqrt{N}\left(\tilde{\theta} - \theta_0\right)$. As in Corollary 1, find that

$$\sqrt{N}\left|\hat{\theta} - \tilde{\theta}\right| \leq M\sqrt{N}\mathbf{1}\{\exists\ j \text{ s.t. } \hat{k}_j \neq k_j\} = o_p(1).$$

C.4 Theorem 2

Let $\pi_j = \pi(\lambda_j)$ and

$$\widetilde{ATE}^{cl} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j}{\pi_j} - \frac{1-D_j}{1-\pi_j}\right)\bar{Y}_j.$$

Find that

$$\begin{aligned}\widetilde{ATE}^{cl} - \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right] &= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j}{\pi_j}\left(\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(1)\right]\right) - \frac{1-D_j}{1-\pi_j}\left(\bar{Y}_j - \mathbf{E}\left[\bar{Y}_j(0)\right]\right)\right) \\ &\quad + \left(\frac{1}{J}\sum_{j=1}^{J}\frac{D_j}{\pi_j} - 1\right)\mathbf{E}\left[\bar{Y}_j(1)\right] - \left(\frac{1}{J}\sum_{j=1}^{J}\frac{1-D_j}{1-\pi_j} - 1\right)\mathbf{E}\left[\bar{Y}_j(0)\right] \\ &= o_p(1)\end{aligned}$$

as $J \to \infty$ since

$$\begin{aligned}\mathbf{E}\left[\frac{D_j}{\pi_j}\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{D_j}{\pi_j}\middle|\lambda_j\right]\right] = 1, \\ \mathbf{E}\left[\frac{D_j\bar{Y}_j}{\pi_j}\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{D_j\bar{Y}_j}{\pi_j}|N_j,\lambda_j\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\frac{D_j\bar{Y}_j(1)}{\pi_j}\middle|N_j,\lambda_j\right]\right] \\ &= \mathbf{E}\left[\frac{1}{\pi_j}\mathbf{E}\left[D_j|N_j,\lambda_j\right]\cdot\mathbf{E}\left[\bar{Y}_j(1)|N_j,\lambda_j\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\frac{1}{\pi_j}\mathbf{E}\left[D_j|N_j,\lambda_j\right]\cdot\mathbf{E}\left[\bar{Y}_j(1)|N_j,\lambda_j\right]\middle|\lambda_j\right]\right] = \mathbf{E}\left[\bar{Y}_j(1)\right]\end{aligned}$$

and similarly for $(1-D_j)/(1-\pi_j)$ and $(1-D_j)\bar{Y}_j/(1-\pi_j)$. The fourth equality is from **A2** and the last equality is from **A7.a)**. The consistency is from **A1** and **A5.a)**.

Next, let $\gamma_{1j} = \mathbf{E}\left[\bar{Y}_j(1)|\lambda_j\right]$ and $\gamma_{0j} = \mathbf{E}\left[\bar{Y}_j(0)|\lambda_j\right]$. Then, it remains to show

$$\widehat{ATE}^{cl} - \widetilde{ATE}^{cl} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right)D_j\bar{Y}_j - \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{1-\hat{\pi}_j} - \frac{1}{1-\pi_j}\right)(1-D_j)\bar{Y}_j = o_p(1).$$

Step 1.

Let us focus on the one side of $\widehat{ATE}^{cl} - \widetilde{ATE}^{cl}$:

$$\begin{aligned}\left|\frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{D_j\bar{Y}_j}{\pi_j}\right)\right| &\leq \left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right)^2\right)^{\frac{1}{2}}\cdot\left(\frac{1}{J}\sum_{j=1}^{J}\bar{Y}_j^2D_j\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right)^2\right)^{\frac{1}{2}}O_p(1)\end{aligned}$$

from Cauchy-Schwarz inequality and **A5.a)**. Then, from Taylor's expansion and **A7.e)**,

$$\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j}-\frac{1}{\pi_j}\right)^2 \leq \frac{M}{2J}\sum_{j=1}^{J}\left(\hat{\pi}_j-\pi_j\right)^2$$

with some constant $M>0$. Lastly, since $(a+b)^2 \geq 0$,

$$\frac{M}{2J}\sum_{j=1}^{J}\left(\hat{\pi}_j-\pi_j\right)^2 \leq \frac{M}{J}\sum_{j=1}^{J}\left[\left(\hat{\pi}_j-\pi\big(\bar{\lambda}(\hat{k}_j)\big)\right)^2+\left(\pi\big(\bar{\lambda}(\hat{k}_j)\big)-\pi_j\right)^2\right] \tag{32}$$

with $\bar{\lambda}(k)$ defined as

$$G\big(\bar{\lambda}(k)\big)=\frac{\sum_{j=1}^{J}G(\lambda_j)\mathbf{1}\{\hat{k}_j=k\}}{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j=k\}} \tag{33}$$

for $k=1,\cdots,K$. The existence of such $\bar{\lambda}$ and its uniqueness is guaranteed from **A7.d)**.

Step 2.

Let us focus on the second quantity from (32).

$$\begin{aligned}\frac{1}{J}\sum_{j=1}^{J}\left(\pi\big(\bar{\lambda}(\hat{k}_j)\big)-\pi_j\right)^2 &\leq \frac{M}{J}\sum_{j=1}^{J}\left\|\bar{\lambda}(\hat{k}_j)-\lambda_j\right\|_1^2 \\ &\leq \frac{M}{J}\sum_{j=1}^{J}q\left\|\bar{\lambda}(\hat{k}_j)-\lambda_j\right\|_2^2\end{aligned}$$

with some constant $M>0$. The first inequality is from Taylor's expansion and **A7.e)** and the second inequality is from Cauchy-Schwarz inequality.

From **A7.d** and $\left\|\vec{a}+\vec{b}\right\|_2^{\,2} \le 2\|\vec{a}\|_2^{\,2} + 2\left\|\vec{b}\right\|_2^{\,2}$, we have

$$
\begin{aligned}
&\sum_{j=1}^{J} \left\|\bar{\lambda}(\hat{k}_j) - \lambda_j\right\|_2^{\,2} \\
&\quad\le \sum_{j=1}^{J} \left[\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - G(\lambda_j)\right\|_{w,2}^{\,2}\right] \\
&\quad\le \sum_{j=1}^{J} \left[2\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\,2} + 2\tau^2 \left\|\hat{G}(\hat{k}_j) - G(\lambda_j)\right\|_{w,2}^{\,2}\right] \\
&\quad\le \sum_{j=1}^{J} \left[2\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\,2} + 4\tau^2 \left\|\hat{G}(\hat{k}_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\,2} + 4\tau^2 \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^{\,2}\right] \\
&\quad\le \sum_{j=1}^{J} \left[2\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\,2} + 4\tau^2 \left\|G\big(\tilde{\lambda}(\tilde{k}_j)\big) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\,2} + 4\tau^2 \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^{\,2}\right] \\
&\quad\le \sum_{j=1}^{J} \left[2\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\,2} + 8\tau^2 \left\|G\big(\tilde{\lambda}(\tilde{k}_j)\big) - G(\lambda_j)\right\|_{w,2}^{\,2} + 12\tau^2 \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\,2}\right] \\
&\quad\le \sum_{j=1}^{J} \left[2\tau^2 \left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\,2} + 8\tau^4 \left\|\tilde{\lambda}(\tilde{k}_j) - \lambda_j\right\|_2^{\,2} + 12\tau^2 \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\,2}\right]
\end{aligned}
$$

where $\tilde{\lambda}(k)$ and $\tilde{k}_j$ are defined as

$$\left(\tilde{k}_1, \cdots, \tilde{k}_J, \tilde{\lambda}(1), \cdots, \tilde{\lambda}(K)\right) = \arg\min \sum_{j=1}^{J} \left\|\lambda_j - \tilde{\lambda}(\tilde{k}_j)\right\|_2^{\,2}.$$

The fourth inequality is from the fact that $\hat{G}(k)$ and $\hat{k}_j$ solve the minimization problem (10). Lastly, from the observation that at the optimal solution of (10), $\hat{G}(k)$ is the average of $\hat{\mathbf{F}}_j$s such

that $\hat{k}_j = k$, we have

$$\begin{aligned}
\frac{1}{J}\sum_{j=1}^{J}\left\|G\big(\bar{\lambda}(\hat{k}_j)\big) - \hat{G}(\hat{k}_j)\right\|_{w,2}^{\ 2} &= \frac{1}{J}\sum_{k=1}^{K}\#(k)\cdot\left\|\frac{\sum_{j=1}^{J}\Big(G(\lambda_j) - \hat{\mathbf{F}}_j\Big)\mathbf{1}\{\hat{k}_j = k\}}{\#(k)}\right\|_{w,2}^{\ 2} \\
&= \frac{1}{J}\sum_{k=1}^{K}\frac{1}{\#(k)}\int\Big(\sum_{j=1}^{J}\Big(G(\lambda_j) - \hat{\mathbf{F}}_j\Big)\mathbf{1}\{\hat{k}_j = k\}\Big)^2(x)w(x)dx \\
&\leq \frac{1}{J}\sum_{k=1}^{K}\frac{1}{\#(k)}\int\Big(\sum_{j=1}^{J}\Big(G(\lambda_j) - \hat{\mathbf{F}}_j\Big)^2(x)\Big)\cdot\Big(\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}\Big)w(x)dx \\
&= \frac{K}{J}\int\sum_{j=1}^{J}\Big(G(\lambda_j) - \hat{\mathbf{F}}_j\Big)^2(x)w(x)dx \\
&\leq \frac{K}{J}\sum_{j=1}^{J}\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2}
\end{aligned}$$

and similarly

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

where $\#(k) = \sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}$. The first inequality is from Cauchy-Schwarz inequality. Note that

$$\frac{1}{J}\sum_{j=1}^{J}\left\|\lambda_j - \tilde{\lambda}(\tilde{k}_j)\right\|_2^{\ 2} = O_p\left(K^{-\frac{2}{q}}\right)$$

as $J, K \to \infty$ (Graf and Luschgy, 2002). Thus,

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right)^2 \\
&\qquad \leq \frac{M}{J}\sum_{j=1}^{J}\Big(\hat{\pi}_j - \pi\big(\bar{\lambda}(\hat{k}_j)\big)\Big)^2 + C\left[\frac{K}{J}\sum_{j=1}^{J}\left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2} + O_p\left(K^{-\frac{2}{q}}\right)\right]
\end{aligned}$$

with some constant $C > 0$.

Step 3.

From **A7.f-g)**,

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

Thus,

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

Step 4.

Let $V_j=D_j-\pi_j$. With a slight abuse of notation,

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}\left(\hat{\pi}_j-\pi\big(\bar{\lambda}(\hat{k}_j)\big)\right)^2\\
&=\frac{1}{J}\sum_{k=1}^{K}\#(k)\left(\frac{\sum_{j=1}^{J}\pi_j\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}+\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}-\pi\big(\bar{\lambda}(k)\big)\right)^2\\
&\leq\frac{2}{J}\sum_{k=1}^{K}\#(k)\left[\left(\frac{\sum_{j=1}^{J}\big(\pi_j-\pi\big(\bar{\lambda}(k)\big)\big)\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}\right)^2+\left(\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}\right)^2\right]\\
&\leq\frac{2}{J}\sum_{k=1}^{K}\#(k)\left[\frac{\sum_{j=1}^{J}\big(\pi_j-\pi\big(\bar{\lambda}(k)\big)\big)^2\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}+\left(\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}\right)^2\right].
\end{aligned}$$

The last inequality is from Cauchy-Schwarz inequality. The first quantity rearranges to

$$\frac{2}{J}\sum_{j=1}^{J}\left(\pi_j-\pi\big(\bar{\lambda}(\hat{k}_j)\big)\right)^2.$$

By repeating the argument from **Step 2-3**,

$$\frac{2}{J}\sum_{j=1}^{J}\left(\pi_j-\pi\big(\bar{\lambda}(\hat{k}_j)\big)\right)^2=O_p\left(\frac{K}{N_{\min,J}}+K^{-\frac{2}{q}}\right).$$

Now, it remains to put a bound on

$$\frac{2}{J}\sum_{k=1}^{K}\#(k)\left(\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\hat{k}_j=k\}}{\#(k)}\right)^2=\frac{2}{J}\sum_{k=1}^{K}\frac{1}{\#(k)}\left(\sum_{j=1}^{J}V_j\mathbf{1}\{\hat{k}_j=k\}\right)^2.$$

Note that

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{\#(k)}\left(\sum_{j=1}^{J} V_j \mathbf{1}\{\hat{k}_j = k\}\right)^2\right] &= \sum_{j=1}^{J} \mathbf{E}\Big[\frac{1}{\#(k)} \sum_{j'=1}^{J} V_j V_{j'} \mathbf{1}\{\hat{k}_j = \hat{k}_{j'} = k\}\Big] \\
&= \sum_{j=1}^{J} \mathbf{E}\Big[\frac{1}{\#(k)} \mathbf{E}\left[{V_j}^2 | N_j, \lambda_j, \{X_{ij}\}_{i,j}\right] \mathbf{1}\{\hat{k}_j = k\}\Big] \\
&\quad + \sum_{j=1}^{J} \mathbf{E}\Big[\frac{1}{\#(k)} \sum_{j' \neq j} \mathbf{E}\left[V_j V_{j'} | N_j, N_{j'}, \lambda_j, \lambda_{j'}, \{X_{ij}\}_{i,j}\right] \mathbf{1}\{\hat{k}_j = \hat{k}_{j'} = k\}\Big] \\
&= \sum_{j=1}^{J} \mathbf{E}\Big[\frac{1}{\#(k)} \mathbf{E}\left[{V_j}^2 | \lambda_j, \{X_{ij}\}_{i,j}\right] \mathbf{1}\{\hat{k}_j = k\}\Big] \leq 1.
\end{aligned}$$

The second equality holds since $\hat{k}_j$s are constructed only with $\hat{\mathbf{F}}_j$s and the third equality holds from **A1** and **A2**:

$$\mathbf{E}\left[V_j V_{j'} | N_j, N_{j'}, \lambda_j, \lambda_{j'}, \{X_{ij}\}_{i,j}\right] = \mathbf{E}\left[V_j V_{j'} | N_j, N_{j'}, \lambda_j, \lambda_{j'}\right] = \mathbf{E}\left[V_j | N_j, \lambda_j\right] \cdot \mathbf{E}\left[V_j | N_{j'}, \lambda_{j'}\right] = 0.$$

Thus,

$$\mathbf{E}\left[\frac{1}{J}\sum_{k=1}^{K} \frac{1}{\#(k)}\left(\sum_{j=1}^{J} V_j \mathbf{1}\{\hat{k}_j = k\}\right)^2\right] \leq \frac{K}{J}.$$

Thus,

$$\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right)^2 = O_p\left(\frac{K}{N_{\min,J}} + K^{-\frac{2}{q}} + \frac{K}{J}\right).$$

We may repeat the same argument for ATT^{cl}.

D Additional empirical results

In Section 7, I use the distribution of the individual-level employment history to capture the cluster-level heterogeneity in labor market fundamentals. In this section, I provide empirical results with an alternative individual-level control variable: wage income. The basic monthly CPS data does not contain information on income. Thus, I used the March Annual Social and Economic Supplement (ASEC) of the CPS to find information on the wage income: for each month t, $y(t)$ is

the calendar year that month t belongs to.

$$\tilde{X}_{ijt} = WageIncome_{ijy)}.$$

Thus, the K-means grouping step is not repeated for every month, but for every year: 22 grouping structures were estimated. Also, while $EmpHistroy_{ijt}$ has a finite support, $WageIncome_{ijy}$ is a continuous variable. Thus, I used the weighting function w that puts equal weights on the 200-quantiles that are estimated from the pooled dataset within each year. Figure 5 contains the empirical distribution functions of the three groups from the 2007 ASEC supplement.

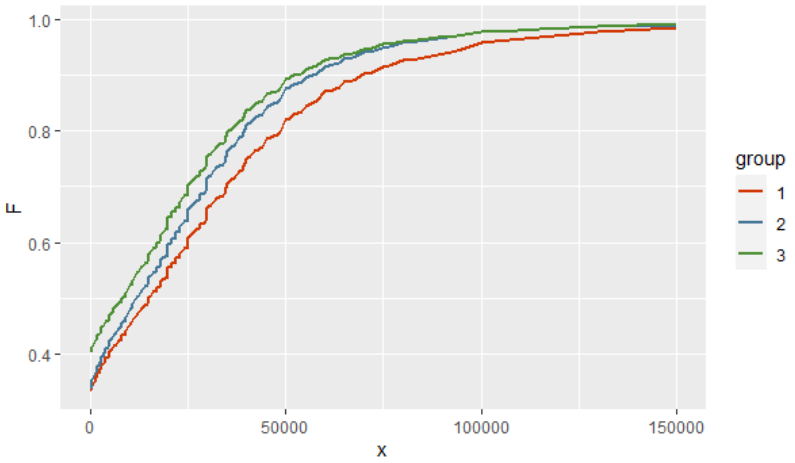

Figure 5: Heterogeneity across states, March 2007

This figure plots average of the empirical distribution functions of wage income for each group.

Group 1 contains states whose wage income distributions are more skewed to the right and Group 3 contains state whose wage income distributions are more skewed to the left.

β	(1)	(2)	(3)	(4)
pooled	-0.065***			
	(0.015)			
Group 1		-0.025	-0.036**	-0.075***
		(0.017)	(0.015)	(0.014)
Group 2		-0.025	-0.036**	-0.028**
		(0.017)	(0.015)	(0.014)
Group 3		-0.027	-0.039**	0.020
		(0.018)	(0.015)	(0.021)
δ_{jt}	GFE	TWFE	Census Div.	GFE

Table 8: 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. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (1) and (4) contain the results from the preferred specification.

Columns (2), (3) and (4) report the group-specific minimum wage effect. Group 1 is the group of states whose wage income distributions are skewed to the right while Group 3 is the group of states whose wage income distributions are skewed to the left.

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)
$Age_{ijt} \leq 18$	-0.072***			
	(0.015)			
× Group 1		-0.032*	-0.043***	-0.083***
		(0.018)	(0.016)	(0.014)
× Group 2		-0.034*	-0.045***	-0.038***
		(0.018)	(0.016)	(0.013)
× Group 3		-0.038*	-0.050***	0.009
		(0.019)	(0.016)	(0.021)
$Age_{ijt} = 19$	-0.039**			
	(0.019)			
× Group 1		-0.001	-0.012	-0.053***
		(0.019)	(0.017)	(0.018)
× Group 2		0.005	-0.006	0.002
		(0.019)	(0.016)	(0.017)
× Group 3		0.010	-0.002	0.057**
		(0.019)	(0.017)	(0.022)
δ_{jt}	GFE	TWFE	Census Div.	GFE

Table 9: 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. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (3) and (6) contain the results from the preferred specification.

Columns (4), (5) and (6) report the group-specific minimum wage effect, while interacting with race. Group 1 is the group of states whose wage income distributions are skewed to the right while Group 3 is the group of states whose wage income distributions are skewed to the left.

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)
$White_{ij} = 1$	-0.094***			
	(0.016)			
× Group 1		-0.053**	-0.066***	-0.101***
		(0.020)	(0.017)	(0.016)
× Group 2		-0.056***	-0.070***	-0.059***
		(0.019)	(0.017)	(0.015)
× Group 3		-0.057***	-0.073***	-0.011
		(0.020)	(0.018)	(0.022)
$White_{ij} = 0$	0.024			
	(0.017)			
× Group 1		0.053***	0.043**	0.008
		(0.017)	(0.018)	(0.016)
× Group 2		0.061***	0.050***	0.062***
		(0.017)	(0.017)	(0.014)
× Group 3		0.059***	0.047**	0.108***
		(0.019)	(0.019)	(0.020)
δ_{jt}	GFE	TWFE	Census Div.	GFE

Table 10: Impact of minimum wage on teen employment in terms of race, 2000-2021

The table reports the effect of a raise in the minimum wage level on teen employment. The regression allows the minimum wage effect to differ across white teens and non-white teens. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (3) and (6) contain the results from the preferred specification.

Columns (4), (5) and (6) report the group-specific minimum wage effect, while interacting with age.Group 1 is the group of states whose wage income distributions are skewed to the right while Group 3 is the group of states whose wage income distributions are skewed to the left.

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.

Lastly, I consider a specification where I use both $EmpHistroy_{ijt}$ and $WageIncome_{ijy}$ in the grouping structure. As discussed in *Remark 1*, I separately construct two grouping structures based on $EmpHistory_{ijt}$ and $WageIncome_{ijy}$: $\hat{k}_{jt}$ is the monthly grouping structure estimated with $EmpHistory_{ijt}$ and $\hat{l}_{jy}$ is the yearly grouping structure estimated with $WageIncome_{ijy}$.

$$Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}\hat{l}_{jy(t)}t} + \beta(\hat{k}_{jt}, \hat{l}_{jy(t)}) \log MinWage_{jt} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{34}$$

In total, I have $13464 = 51 \cdot 22 \cdot 12$ of state-by-month pairs to apply the two grouping structures. Table 11 contains the proportions of the state-by-month pairs in each category defined with the two grouping structures.

$WageIncome$ \ $EmpHistory$	Group 1	Group 2	Group 3
Group 1	0.011	0.118	0.179
Group 2	0.071	0.236	0.102
Group 3	0.170	0.107	0.008

Table 11: Grouping structures based on $EmpHistory$ and $WageIncome$

The rows denote the grouping structure with $WageIncome$ and the columns denote the grouping structure with $EmpHistory$. For example, out of 13464 state-by-month pairs, approximately 11% are assigned to Group 1 under the $WageIncome$ grouping and Group 1 under the $EmpHistory$ grouping.

β		(1)	(2)	(3)	(4)
pooled		-0.065***			
		(0.015)			
WageIncome	*EmpHistory*				
Group 1	Group 1		-0.020	-0.032**	-0.075**
			(0.017)	(0.015)	(0.032)
	Group 2		-0.025	-0.037**	-0.042*
			(0.017)	(0.015)	(0.023)
	Group 3		-0.027	-0.038**	-0.029
			(0.017)	(0.015)	(0.030)
Group 2	Group 1		-0.023	-0.035**	-0.013
			(0.017)	(0.015)	(0.025)
	Group 2		-0.026	-0.037**	-0.021
			(0.017)	(0.015)	(0.020)
	Group 3		-0.029	-0.040**	-0.061**
			(0.017)	(0.015)	(0.026)
Group 3	Group 1		-0.026	-0.038**	0.046*
			(0.018)	(0.015)	(0.025)
	Group 2		-0.027	-0.039**	-0.019
			(0.018)	(0.015)	(0.023)
	Group 3		-0.035	-0.050***	0.123*
			(0.018)	(0.016)	(0.071)
δ_{jt}		GFE	TWFE	Census Div.	GFE

Table 12: 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. For each specification, in addition to the fixed effects, individual-level control covariates — age, race, sex, marital status, education — and state-level employment rate are included as regressors.

Columns (1) and (4) contain the results from the preferred specification.

Columns (2), (3) and (4) report the group-specific minimum wage effect.

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.