Clustered treatment in multilevel models[*]

Myungkou Shin[†]

November 8, 2022

Click here for the latest version.

Abstract

I develop a multilevel model for empirical contexts where treatment is possibly endogeneous and uniformly applied to individuals within a cluster. When treatment assignment is clustered, treatment effect is not identified in a model without any restriction on cluster-level heterogeneity. As a sensible restriction on cluster heterogeneity, I introduce *selection-on-distribution* assumption that a cluster-level latent factor behind the cluster-level distribution of individual-level control covariates sufficiently controls for cluster-level heterogeneity in treatment assignment. In doing so, I let the model fully incorporate the multilevel nature of the data; I characterize treatment effect parameters with aggregate heterogeneity in terms of the cluster-level distribution and individual heterogeneity in terms of the individual-level control. To implement the idea of *selection-on-distribution*, I propose a two-step estimation procedure based on a K-means algorithm. I derive two sets of asymptotic results for the estimator under different assumptions: consistency and asymptotic normality when the latent factor has a finite support; consistency when the latent factor is continuous. An empirical illustration of the estimators is provided as 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 or state is also recorded so that the observations can be clustered to each county or state. 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] In light of the multilevel nature of datasets, a researcher may want to consider an econometric framework that fully 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 by including state fixed-effects or by including some state-level regressors such as population, average income, political party of the incumbent governor, etc. Throughout this paper, I use *individual* and *cluster* to refer to the lower level and the higher level of the hierarchical structure, respectively.

Suppose a researcher is interested in treatment effect estimation in a multilevel setup where treatment is assigned at the cluster level and an outcome variable of interest is observed at the individual level. A lot of 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 a 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 units 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 units from at least two different clusters: *between-cluster* variation. In order to use *between-cluster* variation instead of *within-cluster* variation, restrictions on cluster-level heterogeneity need to be made. In a model with fully flexible cluster-level heterogeneity, cluster heterogeneity and treatment effect cannot be decomposed; 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:

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

[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)

Y_{ij} is an outcome variable for individual i in cluster j and D_{ij} is a binary treatment variable for individual i in cluster j. Cluster fixed-effect α_j flexibly controls for the cluster-level heterogeneity in level. Whenever $D_{ij} = D_j$, i.e. treatment is assigned at the cluster level, treatment effect β is not identified due to multicollinearity between α_j and D_j. Thus, we need some restrictions on cluster-level heterogeneity.

In particular, this paper focuses on cases where a researcher is given individual-level information that is relevant for cluster-level heterogeneity in treatment assignment. Taking advantage of the available information, this paper develops an econometric framework that utilizes the information to estimate treatment effect; I introduce *selection-on-distribution* assumption that treatment is random conditioning on cluster-level distribution of the available individual-level control covariates. In other words, I assume that there is no cluster-level heterogeneity in terms of treatment assignment, among clusters with the same distribution of individual-level control covariates; direct comparison between any two of those clusters identifies treatment effect.

The *selection-on-distribution* assumption has three components. The first component is the *selection-on-observable* assumption. The *selection-on-observable* assumption that treatment is random conditioning on some control covariates observed at the level of treatment assignment is widely used in the program evaluation literature to control for treatment endogeneity. For a running example, suppose a researcher wants to estimate the effect of a job training program on employment status and suspects endogenous selection into the program; a simple mean comparison between participants and non-participants is comparing apples to oranges. The *selection-on-observable* assumption in this example assumes that there exist key determinants in the selection process, e.g. one's high school GPA and course selections, and that the selection into the program is as good as random among potential participants with the same high school GPA and course selections. Or, if the researcher is given panel data and suspects dependence over time, the determinant can be the lagged outcome; employment status from the last period. A simple way to implement the *selection-on-observable* approach is to use the determinants in the selection process as control covariates in regression:

$$Y_{ij} = \beta D_{ij} + X_{ij}^\intercal \theta + U_{ij}.$$

$X_{ij} \in \mathbb{R}^p$ is the vector of control covariates for individual i in cluster j that contains information on individual i's education. With some linear additivity assumption, the *selection-on-observable* assumption implies that U_{ij}, the error term remaining after controlling for X_{ij}, is independent of treatment status D_{ij}, identifying treatment effect β.

The application of the *selection-on-observable* approach to a multilevel model is not trivial. A naive extension of the *selection-on-observable* assumption to a multilevel model with clustered treatment would be that treatment is random conditioning on all the relevant information at the cluster level, since the cluster is the level of treatment assignment. In this case, even when the control covariates for each individual within a cluster are low-dimensional, the cluster-level collection of all the individual-level control covariates is high-dimensional if the cluster size is large. Let us revisit the running example and suppose that a high school decides whether or not to mandate a job training program on its students; treatment is assigned at the school level and the determinants in the school-level decision are GPA and course selections of all students at the school. Note that the conditioning object is two-dimensional in the case of individual-level selection decision while the dimension of the conditioning object is two times the number of students in the case of school-level selection decision. Thus, the multilevel model induced by the *selection-on-observable* assumption is not parsimonious. Consider a linear model with cluster-level treatment:

$$Y_{ij} = \beta D_j + X_{ij}{}^{\intercal}\theta + \mathbb{X}_j{}^{\intercal}\theta^{cl} + U_{ij}.^{2} \tag{2}$$

$\mathbb{X}_j = \{X_{ij}\}_{i=1}^{N_j}$ is the cluster-level collection of individual-level control covariates X_{ij}, which is the conditioning object for conditional independence of treatment. In the model, the dimension of the model parameter $\left(\beta, \theta, \theta^{cl}\right)$ is proportional to the cluster size N_j. Note that the high-dimensionality problem illustrated here is not confined to linear models.

Thus, as the second component, I introduce *exchangebility* to partially solve the high-dimensionality problem. The *exchangeability* assumption assumes that the distribution of individuals within a cluster is invariant up to permutation on labeling: for any permutation σ on $\{1, \cdots, N_j\}$,

$$\left(Y_{1j}, X_{1j}, \cdots, Y_{N_j j}, X_{N_j j}, D_j\right) \overset{d}{\equiv} \left(Y_{\sigma(1)j}, X_{\sigma(1)j}, \cdots, Y_{\sigma(N_j)j}, X_{\sigma(N_j)j}, D_j\right).$$

Under the *exchangeability* assumption in addition to the *selection-on-observable* assumption, we can reduce the dimension of the conditioning object by replacing $\mathbb{X}_j$ with some function g of $\mathbb{X}_j$, as long as g is injective symmetric, i.e. $g(\mathbb{X}) = g(\mathbb{Z})$ if and only if $\mathbb{Z}$ is a permutation of $\mathbb{X}$.[3] When X_{ij} is one-dimensional, ordered statistics can be g:

$$g(\mathbb{X}_j) = \left(X_{(1)j}, \cdots, X_{(N_j)g}\right).$$

g maps $\mathbb{R}^{N_j}$ onto $\{x = (x_1, \cdots, x_{N_j}) \in \mathbb{R}^{N_j} : x_1 \leq \cdots \leq x_{N_j}\}$. Note that $g(\mathbb{X}) = g(\mathbb{Z})$ if and only if there exists a permutation σ such that $X_{ij} = Z_{\sigma(i)j}$ for every i and $g(\mathbb{X}_j)$ has strictly smaller support than $\mathbb{X}_j$. To allow for multivariate X_{ij}, I use empirical distribution function as g in this paper: for all $x \in \mathbb{R}^p$,

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

g maps $\mathbb{R}^{p \cdot N_j}$ onto a space of distribution functions on $\mathbb{R}^p$. The *exchangeability* assumption implies that the names of each individual in a given cluster does not have any additional information in terms of treatment assignment. In the case of the school-level selection decision of the running example, the *exchangeability*

[2] A comparison with (1) can be made here. Both of the models contain an element of cluster-level heterogeneity. However, in model (1), it is through cluster fixed-effect α_j while in model (1), it is through the cluster-level regressor $\mathbb{X}_j$.

[3] See Appendix for a formal statement in terms of potential outcomes.

assumption assumes that identity of each student does not matter for the decision of the school; school principals cannot make the decision solely based on their own child enrolled at their school.

The last component of the *selection-on-distribution* assumption is the decomposition of g into signal and noise in terms of treatment assignment. Specifically, I assume that only the expectation of g is the relevant information in terms of treatment assignment. Let us denote the expectation with $\mathbf{F}_j$: for all $x \in \mathbb{R}^p$

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

Then, I assume that treatment is random conditioning on the distribution function $\mathbf{F}_j$: *selection-on-distribution*. Note that through this decomposition, clusters with different cluster size N_j can also be matched in terms of the signal $\mathbf{F}_j$ whereas in general empirical distribution functions cannot be same for two clusters of different sizes.[4]

The *selection-on-distribution* assumption suggests that we 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 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 dataset is a cross-section and there is no control covariate at the cluster level, I propose nonparametric estimators with inverse probability weighting. Secondly, when dataset is a 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.

1.1 Related literature

This paper contributes to several literatures in econometrics. Firstly, this paper contributes to the treatment effect and program evaluation literature. The *selection-on-distribution* assumption of this paper requires that cluster-level distribution of individual-level control covariates be used in modeling cluster-level treatment assignment and individual-level outcome. By using both cluster-level distribution and individual-level control covariates for each individual, treatment effect is modelled with two types of heterogeneity: aggregate heterogeneity from the cluster-level distribution and individual heterogeneity from the individual-level control covariates. By estimating treatment effect while allowing for aggregate heterogeneity and individual heterogeneity, 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

[4] In the canonical model of this paper, I partially relax this assumption and model treatment assignment to depend on the cluster size as well, while the information from the individual-level control covariates X_{ij} only enters the treatment assignment process through its distribution function $\mathbf{F}_j$.

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?”, “what neighborhood characteristic of a destination location makes migration successful?”, “does individual-level demographic characteristic interact with neighborhood characteristic of the destination?”, by looking at individual heterogeneity, aggregate heterogeneity, and interactive hetergoeneity in treatment effect, respectively. Moreover, aggregate heterogeneity in treatment effect can also be thought of as a general equilibrium result of network/neighborhood/social interaction effect when there exist multiple, well-separated pools of potential network formation: Manski (1993); Bramoullé et al. (2009).

Secondly, this paper makes contribution to the literature of regression with heterogeneous slopes, and particularly to the group fixed-effect literature that assumes a finite grouping structure on unit-specific fixed effects or unit-specific slope coefficients in a panel data model. Whereas the group fixed-effect literature mostly focuses on a panel data, I apply the idea of assuming a finite grouping structure for tractability to multilevel models, which is possibly a cross-section. A key difference of the grouping approach in this paper from the 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) are comparable to this paper. Both papers use the information from the RHS of the equation to recover the slope heterogeneity. Also, when the grouping structure is thought of as an approximation of underlying continuous latent factor for cluster-level distribution $\mathbf{F}_j$, as will be discussed in Section 5, 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 makes contribution to the distributional regression literature. To estimate propensity score and treatment effect with cluster-level distribution of individual-level controls, 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 a X-adaptive partition-based regression (Cattaneo et al., 2020); a 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.

Lastly, I apply the econometric framework proposed in this paper to revisit the question whether a raise in the minimum wage level has disemployment effect on teenager population of US. In doing so, I control for aggregate heterogeneity in state-level labor market fundamentals, by controlling for the distribution of individual employment status history. Then, I explore two channels of individual heterogeneity: age and race. I find differential disemployment effect in terms of both individual-level control variables and find that the differential effect also depends on labor market fundamentals.

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 at 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 *selection-on-distribution* assumption. In Section 3, I explain the K-means algorithm and 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 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} can be multi-dimensional and X_{ij} can include lagged outcomes if the econometrician observes panel data, as discussed in the previous section. 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 $Y_{ij}(0)$:

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

Note that potential outcomes are defined at the individual level but treatment is defined at the cluster level: multilevel structure is relevant to treatment assignment.

Now, I introduce three assumptions to develop a multilevel model with finite cluster-level latent factor.

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

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

Also, $H^{hyper}\big(\{D_j, N_j, \lambda_j\}_{j=1}^{J}\big)$, the conditional distribution of $\Big\{\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}\Big\}_{j=1}^{J}$ 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. 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 a distribution function $H(D_j, N_j, \lambda_j)$, independently of any other individual-level variables or cluster-level variables from different 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 restriction is 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 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 $x \in \mathbb{R}^p$,

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

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 finds the cluster-level latent factor λ_j an interpretation in the data generating process; the latent factor λ_j captures 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} in cluster j 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 status:

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

Assumption 2 is the *selection-on-distribution* assumption.

Remark 1. Assumption 2 does not impose any restriction on p. The cluster-level distribution function $\mathbf{F}_j$ can be multivariate distribution.

Remark 2. 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}:

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

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

$$\lambda_j = \left(\lambda_{1j}, \cdots, \lambda_{pj}\right),$$

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

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

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 of $X_{ij} = \left(X_{ij1}, \cdots, X_{ijp}\right)$, to the p marginal distributions of $X_{ij1}, \cdots, X_{ijp}$.

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 3. 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

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. If the latent factor λ_j is observed, **Assumption 2** identifies all of the treatment effect parameters, with some uniform overlap condition on D_j across $\left(N_j, \lambda_j\right)$.

2.1.2 Conditional treatment effect

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

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

On the individual level, we again use **Assumption 2** and use 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. Again, if the latent factor λ_j is observed and an overlap condition is given, **Assumption 2** identifies both $CATE^{cl}$ and $CATE$.

With CATE defined as above, multilevel nature of heterogeneity in treatment effect can be explored. A first use of the model is to estimate heterogeneous treatment effect where the heterogeneity comes from individual 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 heterogeneity. Secondly, the model estimates heterogeneous treatment effect in terms of cluster-level conditioning variables. 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 a binary treatment variable are often included. If the control covariate is an individual-level variable, the interaction term essentially captures individual heterogeneity and if the control covariate is a cluster-level variable, the interaction term captures aggregate heterogeneity. When analyzing results from such a regression specification, a researcher understands such distinction and notes where the treatment effect heterogeneity comes from. The distinction that I make here is of the same nature: the difference is that instead of using a given cluster-level variable, I construct a new cluster-level object from individual-level data.

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 to increase their state's minimum wage level or not. 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 on 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. In the terminology of dynamic programming, (λ, m, c) is the state 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 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

$$\begin{aligned}\mathrm{Pr}_{X_{-j}} &\left\{\beta_1 \sum_i X_{ij} + \beta_2 \sum_{i' \neq i} E_{ij} + \sum_i U_{ij} \geq y^* - \beta_2\right\} \\ &- \mathrm{Pr}_{X_{-j}} \left\{\beta_1 \sum_i X_{ij} + \beta_2 \sum_{i' \neq i} E_{ij} + \sum_i U_{ij} \geq y^*\right\} \geq \frac{c(X_{ij})}{r}.\end{aligned}$$

Note that the probability is over every other worker in team j who is not worker i. 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:

$$Y_{ij}(1) - Y_{ij}(0) = \beta_2 E_{ij}^*(1) = \beta_2 e\big(X_{ij}, \lambda_j\big).$$

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 connects to the 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 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 $\big(G_j\big)_{hi}$ is

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

where $W_{hij} \in \{0,1\}$ is a binary friendship 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. 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 $(x, \tilde{x}, \lambda)$,

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

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 in the indirectly peer effect only.

3 Estimation

In this section, I propose a two-step estimation procedure of estimating $ATE^{cl}, ATT^{cl}, ATE, ATT, CATE^{cl}$ and $CATE$. The first step is to find a finite grouping structure on cluster heterogeneity from cluster-level distributions of individual-level control covariates. The second step is to use the finite grouping structure on clusters in estimations. In the second step, I propose two sets of estimators; nonparametric estimators with inverse probability weightings and least-squre estimators with parametric model.

3.1 First step: the 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 $\mathbf{E}\left[\hat{\mathbf{F}}_j | D_j, N_j, \lambda_j\right] = G(\lambda_j)$: $\hat{\mathbf{F}}_j$, the estimator I use for $\mathbf{F}_j$, is 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

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

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

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

[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.

Monte Carlo simulation results support the use of an estimator for K based on an information criterion.

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

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

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 $\left(\hat{k}_1, \cdots, \hat{k}_J\right)$. Motivated by this observation, the iterative K-means algorithm finds the minimum as follows: given an initial grouping $\left(k_1^{(0)}, \cdots, k_N^{(0)}\right)$,

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 $\left(k_1^{(s)}, \cdots, k_J^{(s)}\right)$:

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

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, a most naive choice is to stop the algorithm after a fixed number of iterations.

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 optimized objective function $Q_J(K)$ across initial groupings. For more discussion on how to choose those initial values, see Appendix.

[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.

3.2 Second step: treatment effect estimation

In the second step, I use the finite grouping structure from the step to estimate treatment effect parameters. Specifically, I propose two types of estimators for two different data contexts. Firstly, suppose that a researcher is given a cross-section 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 context. 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 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:

$$\begin{aligned}\hat{\pi}(k) &= \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{\hat{k}_j = k\}} \sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\}, \qquad (14)\\ \hat{\pi}_j &= \hat{\pi}(\hat{k}_j).\end{aligned}$$

The propensity score estimates are computed as a sample mean of treatment status variable D_j for each group. Note that $\hat{\pi}(k)$ is allowed to be zeros or ones. There are multiple remedies to this problem of finite-sample no overlap. For example, we may drop the group without overlap altogether. 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 the distance on $\hat{\mathbf{F}}_j$ and vice versa. 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 (14), 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 cluster-level control covariates. To include the cluster-level variable Z_j, let us modify **Assumption 1**, by replacing N_j with an arbitrary cluster-level random vector Z_j that includes N_j.

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

Also, $H^{hyper}\left(\{D_j, Z_j, \lambda_j\}_{j=1}^{J}\right)$*, 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* $\left(D_j, Z_j, \lambda_j\right)$*:*

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

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 $\theta = \left(\theta^1, \cdots, \theta^K\right)$ 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 4. 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 5. 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,

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

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 definite benefits. First of all, the grouping from the K-means algorithm 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, there exist theoretical results on asymptotic behavior of the K-means estimators. A vast literature has studied the asymptotic behaviors of various 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 from Assumption 3 helps us discuss theoretical properties of the induced estimators while being in touch with the economic insight.

Thirdly, the parametric model under the finite grouping structure can motivate a linear regression model 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 are three distinctive groups among J clusters.

4 Asymptotic results

In this section, I discuss asymptotic properties of treatment effect estimators from Section 3. Firstly, I introduce Assupmtion 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) $\mu(k) := \Pr\left\{\lambda_j = \lambda^k\right\} > 0\ \forall k$.*
- ***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) $N_{\min,J} = \max_n\{\Pr\left\{\min_j N_j \geq n\right\} = 1\} \to \infty$ 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 $J/\min_j {N_j}^{\nu^*}$ goes to zero for some ν^*. Thus, when $\min_j {N_j}^{\nu^*}$ 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* $\mathbf{E}\left[\bar{Y}_j^2\middle|\left\{X_{ij}\right\}_{i=1}^{N_j}, D_j, N_j, \lambda_j\right] < M$ *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 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 A5.c-d) assume asymptotic normality on $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\bar{Y}_j$, with relevant scaling 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 weights equal to $\frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\{\lambda_j = \lambda^k\}$. 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

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

Proof. See Appendix. □

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 reduce down to the population mean.

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

and

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

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. □

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 for a researcher to choose.

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* $\Pr\left\{\eta \leq \pi(\lambda_j) \leq 1-\eta\right\} = 1$.

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

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

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,*

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

as $J, K \to \infty$*.*

Proof. See Appendix. □

Theorem 2 characterizes the convergence rate of $\widehat{ATE}^{cl} - ATE^{cl}$. 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 to 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.

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 state, students in schools in school district, workers in firms in industries, etc.

The researcher wants his model to allow for the county-level heterogeneity and the state-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

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

which is high-dimensional, is summarized to

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

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 $k \in \{k, 1 \cdots, K\}$,

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

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 simulate datasets to apply the K-means estimators to and reaffirm the asymptotic properties discussed in Section 4. For simplicity, I let each cluster to be of the same cluster size and let the cluster size to depend on the number of clusters I generate. To denote this, I use N_J: $N_J = N_{\min,J} = N_j$ for every $j = 1, \cdots$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 $i = 1, \cdots, N_J$ 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 how the propensity score π and the treatment effect β changes with the latent factor λ.

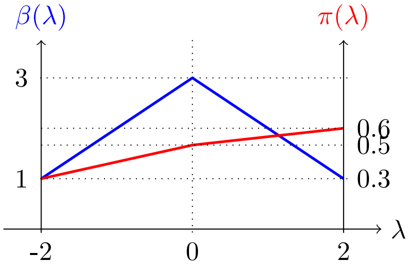

Figure 1: π and β

Firstly, I let λ_j be discrete:

$$\Pr\{\lambda_j = \lambda\} = \begin{cases} 0.1 & \text{for } \lambda = 2, 2 \\ 0.3 & \text{for } \lambda = 1, -1 \\ 0.2 & \text{for } \lambda = 0 \\ 0 & \text{otherwise} \end{cases}$$

I generate 1,000 datasets following the DGP and estimate the average treatment effect with $\widehat{ATE}^{cl}$ for each dataset. Figure 2 shows how the mean squared error (MSE) computed across the 1,000 datasets changes as I shift J, the number of clusters, and N_J, the cluster size. I increase both J and N_j at a rate that N_J/J also increases: this is to guarantee that there exists some positive constant ν such that $J/{N_J}^{\nu}$ goes to zero. The MSE decreases as J and N_J/J increases. The grouping based on $\hat{\mathbf{F}}_j$ improves as N_J/J increases and the variance in estimating ATE decreases as J increases. Also, for the specification where J and N_J are largest, I draw the distribution of the estimator $\widehat{ATE}^{cl}$, normalized with the infeasible asymptotic variance. Figure 3 shows the asymptotic normality and the 0.05 significance level test has the rejection rate of 0.068.

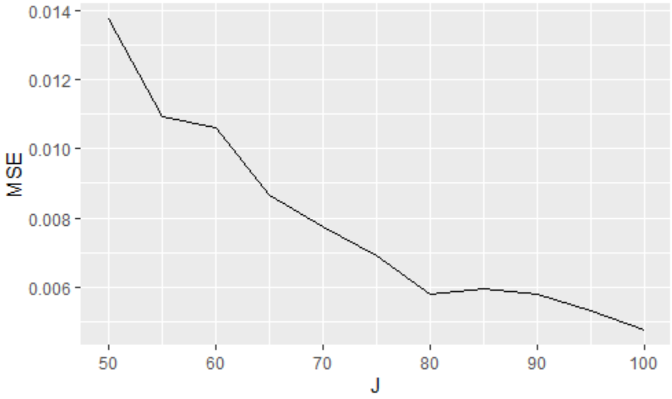

Figure 2: MSE as J and N_J increase: $J = 50, \cdots, 100$ and $N_j = 100, \cdots, 300$.

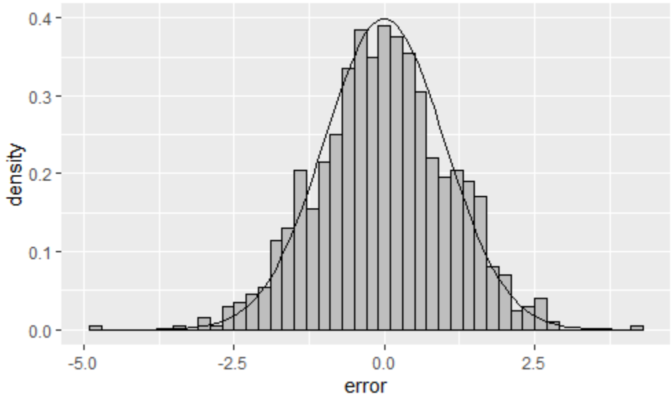

Figure 3: Asymptotic normality: $J = 100$ and $N_J = 300$.

Secondly, I let λ_j be continuous:

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

Again, I generate 1,000 datasets and estimate ATE^{cl}. Figure 4 shows that the MSE decreases as J increases. As shown in the convergence rate of Theorem 2, larger J reduces the variance in estimating the propensity score, hence decreasing the MSE. Figure 5 shows that the MSE is U-shaped in terms of K. Recall that the convergence rate had the tradeoff in terms of K. When K is smaller, the improvement in the approximation bias dominates the cost of having less sample to estimate the propensity score for each group and thus the MSE decreases with K. When K is bigger, the approximation of Λ to $\{1, \cdots, K\}$ is sufficiently improved and thus the MSE increases with K, due to the increased variance of the propensity score estimation.

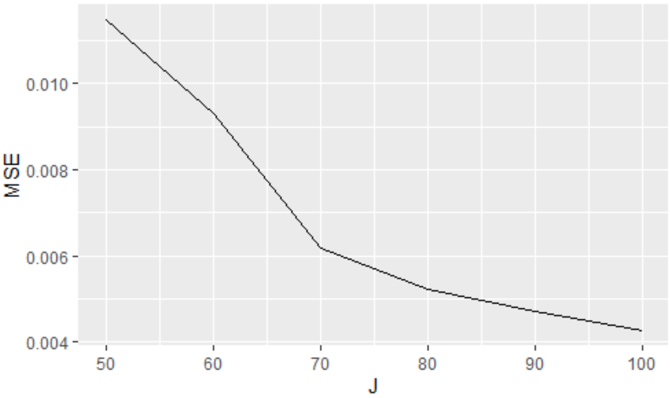

Figure 4: MSE as J increases: $J = 50, \cdots, 100$, $K = 5$ and $N_j = 200$

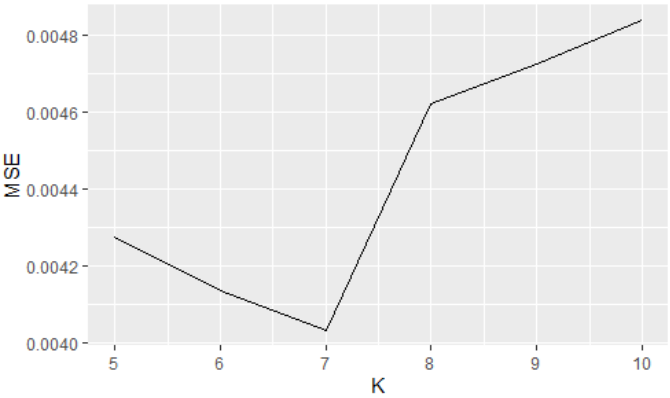

Figure 5: MSE as K increases: $J = 100$, $K = 5, \cdots, 10$ and $N_j = 200$

7 Application: effect of minimum wage on teen employment

I apply the K-means estimator suggested in this paper and revisit the question of whether an increase in minimum wage level leads to higher unemployment rate in US labor market. This quintessential question in labor economics has often been answered using a state-level policy variation; since each state has their own minimum wage level in addition to federal minimum wage level in the United States, 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 specific heterogeneity. Still, there could be a spatial heterogeneity with which treatment may be endogenous with labor market outcome. For that researchers have long been debating on how best to estimate the causal effect of minimum wage increase on employment while accounting for spatial heterogeneity. For example, difference-in-differences (DID) compares over-the-time difference in employment rate for each state, assuming that spatial heterogeneity only exists in the form of state heterogeneity and the state heterogeneity is controlled 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 to compare for a state with minimum wage level increase (Neumark et al., 2014). All of the preexisting literature rely on varying identifying assumptions and enjoy corresponding distinctive benefits, which explains why there is no clear winner.

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 translates to this minimum wage application very well. Firstly, the outcome of interest that most papers in the literature focus on is by construction a multilevel quantity: employment rate is defined as a labor market participant's employment status, averaged on the state level. In the simplest two-level model, state would be the higher level and labor market participant would be the lower level. 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 should be increased is only determined by what happens in that state. This corresponds to **Assumption 1**. Thus, I believe the K-means estimator suggested in the paper is a naturally appealing candidate if we are interested in an estimator that is motivated from the *selection-on-distribution* approach.

In estimation, I use the Current Population Survey (CPS) data, which is publicly available. Though there are numerous timings where multiple states raised their minimum wage levels, I chose January 2007, where eighteen states raised their minimum wage levels. It is the timing where the most states raised their minimum wage levels without a federal minimum wage raise. Since the estimation method suggested in this paper is a matching estimator, more states with treatment helps with overlap. Also, instead of looking at broader labor market outcomes, I focus on the teen employment, as in Neumark et al. (2014) and Allegretto et al. (2017) since it is more likely that teens work at jobs that pay near the minimum wage level compared to adults, thus being more susceptible to treatment. That being said, since the teen population may not contain all the information about state heterogeneity, I use distribution of demographic characteristics for entire population in the first step of the estimation when grouping states.

Formal connections between the terminology that I have been using in the paper and the context of the minimum wage application are as follows. I have $J = 51$ clusters: 50 states and the District of Columbia. For each state, I observe one cluster-level treatment status D_j, where $D_j = 1$ means that state j increased their minimum wage starting January 1st, 2007:

$$D_j = \mathbf{1}\{MinWage_j^{Jan07} - MinWage_j^{Dec06}\}.$$

Also, for each state, I observe two sets of individual-level variables $\{Y_{ij}\}_{i\in\mathcal{I}_j^{teen}}$ and $\{X_{ij}\}_{i\in\mathcal{I}_j}$. Here I observe two different sets of individual-level variables since the universe of the outcome variable that I am interested in is the teen population while that of the control covariate covers the entire labor force. Thus, $\{Y_{ij}\}_{i\in\mathcal{I}_j^{teen}}$ are employment status variable for teens of state j and $\{X_{ij}\}_{i\in\mathcal{I}_j}$ is some control covariate for residents of state j. Specifically, I let

$$\begin{aligned} X_{ij} &= EmpHistory_{ij} \\ &= \left(Emp_{ij}^{Sep06}, \cdots, Emp_{ij}^{Dec06}\right) \in \mathcal{X} := \{Emp, Unemp, NotInLaborForce\}^4, \end{aligned}$$

which is a four-month-long history of employment status, from September 2006 to December 2006, for individual i in state j, categorized into three categories: if the individual is employed, unemployed, or not in the labor force. $\mathbf{X}$, the support of X_{ij}, is a finite set of 81 elements.

7.1 K-means groups and propensity score

Firstly, I compute the empirical distribution of X_{ij} for each state: for each $x \in \mathcal{X}$,

$$\hat{\mathbf{F}}_j(x) = \frac{1}{\mathcal{I}_j|} \sum_{i\in\mathcal{I}_j} \mathbf{1}\{EmpHistory_{ij} = x\}.$$

When evaluating the distance between states measured in terms of $\hat{\mathbf{F}}_j$, I use the uniform weighting function since the support of X_{ij} is a finite set. Then, I apply the K-means algorithm to group states, in terms of $\|\cdot\|_2$. Figure 6 contains the grouping result when $K = 3$. Each group is shaded with different color: red, blue, green and purple. Here 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.

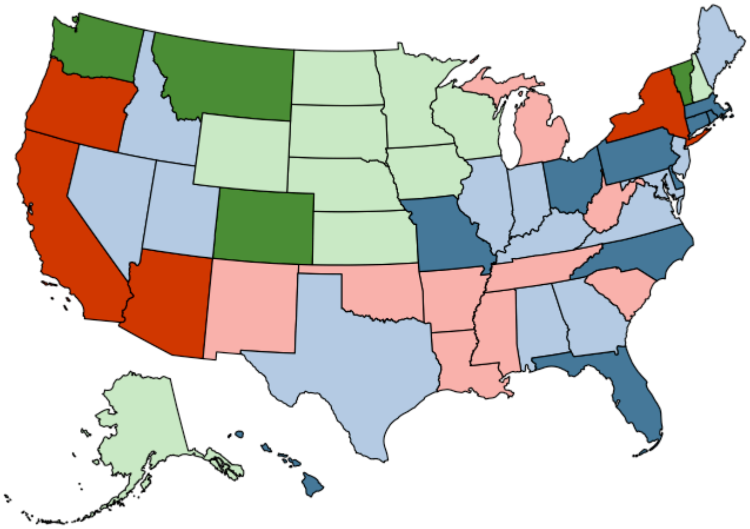

Figure 6: Grouping of states when $K = 3$

Table 1 and Figure 7 contain empirical evidence that the groups estimated using the distribution of $EmpHistory_{ij}$ are heterogeneous. Table 1 takes a subvector of $EmpHistory_{ij}$ and copmutes their sample means for each group, putting equal weights over states. The selected subvector contains three types of employment history: having been employed continuously for the four months, having been in the labor force continuously and unemployed at least for one month, and having been out of the labor force continuously. A simple t-test that takes the grouping as given and tests a null hypothesis that certain two groups have the same group mean, e.g.

$$H_0 : \mathbf{E}[\bar{X}_j | j \in \text{group } 1] = \mathbf{E}[\bar{X}_j | j \in \text{group } 2],$$

is rejected at significance level 0.001 for any pair of two groups: groups are heterogeneous.

In addition, Figure 7 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 mean. 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 is. Specifically, group 1 states such as California and

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

Table 1: group heterogeneity

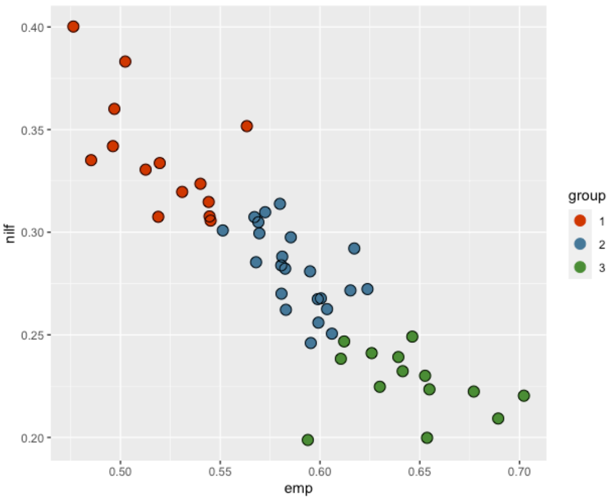

Figure 7: group heterogeneity

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.

7.2 Treatment effects

Based on the K-means grouping, I estimate $ATTcl$ and $CATE^{cl}$ for each group. In estimation, instead of using Y_{ij} in level, I used the over-the-year difference outcome variables, to further control for state heterogeneity and seasonality: $Y_{ij}^{post} = Emp_{ij}^{Jan07}$ is a binary employment status variable from January 2007 and $Y_{ij}^{pre} = Emp_{ij}^{Jan06}$ 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. Table 2 contains the estimates, reweighted with the size of the minimum wage level increase, so that each estimate is interpreted to be an elasticity:

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

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 2: Treatment effect estimates as elasticities

7.3 Regression specification

The estimates in Table 2 can be thought of as a snapshot estimation that pm;y focuses on immediate effects, leading to a larger estimates than usually reported in the literature. For comparability and bigger power, I use Corollary 2 and consider various linear regression specifications that are more consistent with the literature, and use the entirety of the dataset spanning from 2000 to 2021: since the CPS dataset is collected every month, the time subscript t ranges from $1, \cdots, 264 = 12 \cdot 22$. With the pooled dataset, I commit to one of the two main regression specification from Allegretto et al. (2011): for teen i in state j at time t,

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

$EmpRate_{jt}$ is the average of Y_{ijt} for every individual in state j while the regression runs only on teens. Note that the variable of interest $MinWage_{jt}$ varies on the state level and the month level, making state-specific time fixed-effects infeasible. Thus, census division fixed-effects are used instead by grouping 51 states into 9 census divisions: $\delta_{cd(j)t}$ and $cd(j) \in \{1, \cdots, 9\}$.

Building on this, I use a linear regression specification with group fixed-effects, where each states are partitioned into K groups, based on their distribution of four-month-long individual-level employment history, at each time t:

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

Instead of the census division fixed-effect $\delta_{cd(j)t}$, the group fixed-effect $\delta_{\hat{k}_{jt}t}$ is used. Since I use a long time-series, I group states each month, and connect group memberships across months based on their relative position: at each month, the group of states with the highest proportion of always-employed is labeled group 1 and the group of states with the lower proportion of always-employed is labeled group 3.

Note that the idea of aggregating the state-level information and using the summary statistic in the regression is not new: $EmpRate_{jt}$ is used in the original specification. In the original specification, 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. By using the group fixed-effects, I allow for the state-level heterogeneity to be a more flexible function of Y_{ijt}s; I look at the history of employment status and I look at their distribution. If a lagged employment rate were to be used, $EmpRate_{jt-1}$ would indeed be a summary statistic of $\hat{\mathbf{F}}_{jt}$, the employment history distribution used for grouping.

Moreover, the group fixed-effects are comparable to the census division fixed-effects in the sense that they are also an adaptation of the state-specific time fixed-effects. Suppose a researcher believes that the state heterogeneity only exists as a level shift. Then he would want to use state-specific time fixed-effects to control for the state heterogeneity. However, since we have a variable of interest that does not vary within a state at a given time, the state-specific time fixed-effects are infeasible due to multicollinearity. Thus, an adaptation is made to circumvent the multicollinearity. For example, in a two-way fixed-effect (TWFE) specification, time fixed-effects are assumed to be constant across every state and the state heterogeneity only exists in the state fixed-effect: $\delta_{jt} = \delta_t$. In Allegretto et al. (2011), the census division fixed-effects are used to allow for more heterogeneity across states by letting the time fixed-effects to vary across census divisions at each time t, but still assume that the census division structure does not change over time: $\delta_{jt} = \delta_{cd(j)t}$. In contrast, the group fixed-effects still impose that the state heterogeneity be constant across states within a group, which comes from the state-level observable information, but allow the pattern of heterogeneity to vary over time: $\delta_{jt} = \delta_{\hat{k}_{jt}t}$.

Table 3 contains the estimation result of the TWFE specification, the census division fixed-effects specification, and the group fixed-effect specification. Again, by dividing the estimate with the average teen employment rate, we get the elasticity interpretation: the average teen employment rate from the pooled dataset is 0.326. 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

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 3: heterogeneity in treatment effect

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.

The columns (4)-(6) of Table 3 discuss the aggregate heterogeneity in treatment effect. 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.

As a next step, I further extend the linear specification in use to discuss individual heterogeneity and aggregate heterogeneity. The left panel of Table 4 is from the linear specification

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \beta^1 \log MinWage_{jt}\mathbf{1}\{Age_{ijt} \leq 18\} \\ + \beta^2 \log MinWage_{jt}\mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{26}$$

Individual heterogeneity in treatment effect is introduced in terms of age. The right panel of Table 4 is from the linear specification

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \sum_{k=1}^{3} \beta^1(k) \log MinWage_{jt}\mathbf{1}\{Age_{ijt} \leq 18, \hat{k}_{jt} = k\} \\ + \sum_{k=1}^{3} \beta^2(k) \log MinWage_{jt}\mathbf{1}\{Age_{ijt} = 19, \hat{k}_{jt} = k\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{27}$$

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)
$Age_{ij} \leq 18$	-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)
$Age_{ij} = 19$	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 4: heterogeneity in treatment effect, in terms of age

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 5: heterogeneity in treatment effect, in terms of race

Table 4 shows that younger teens, who are under the age of nineteen, are more affected by a raise in the minimum wage level than older teens of the age nineteen. 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.

Table 5 repeats the same regression specification, but in terms of white; Table 5 documents individual heterogeneity in terms of white teens against non-white teens. From the left panel of Table 5, we see that a raise in the minimum wage level decreases the employment rate of white teens and increases the employment rate of non-white teens. 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 5, 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 their counterparts 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 8 contains the confidence intervals of the interaction estimates from Column (6) of Table 4 and Column (6) of Table 5.

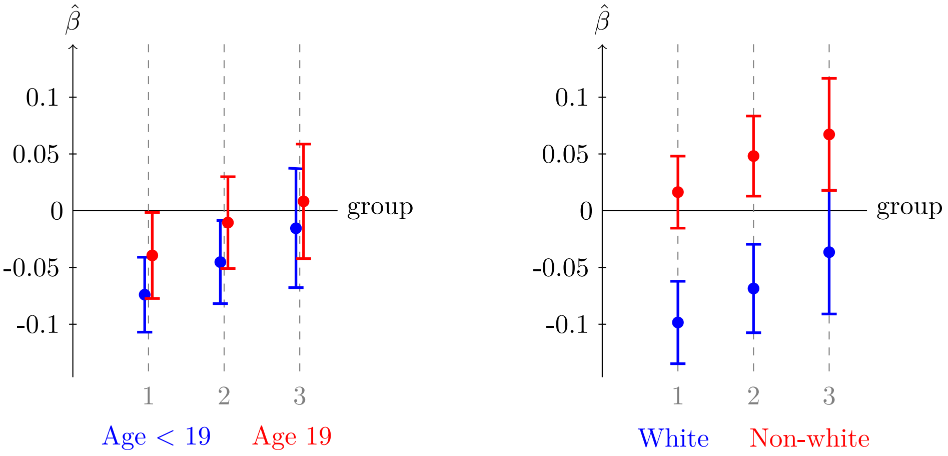

Figure 8: Interaction between individual and aggregate heterogeneity

8 Conclusion

This paper extends the idea of *selection-on-observable* 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 allowed, being true to the multilevel nature of the model: 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 US. I find the disemployment effect to be heterogeneous both on the individual level and the cluster level, and the two dimensions of heterogeneity interacts.

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 desirable qualities in terms of exposition, an application of an alternative functional regression method to the *selection-on-distribution* assumption would complement this paper by allowing for different sets of assumptions on the cluster-level distribution. Also, this paper mostly focuses on a cross-section and a non-dynamic panel data. An exciting direction for future research is to expand 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.

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.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, “Identification of peer effects through social networks,” *Journal of econometrics*, 2009, *150* (1), 41–55.

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.

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.

Manski, Charles F, "Identification of endogenous social effects: The reflection problem," *The review of economic studies*, 1993, *60* (3), 531–542.

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

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

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* $\{1, \cdots, N_j\}$,

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

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

$$\hat{\mathbf{F}}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\{X_{ij} \leq x\}. \tag{28}$$

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,

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

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 $\left\{X_{1j}, \cdots, X_{N_j j}\right\}$ that is isomorphic to $\hat{\mathbf{F}}_j$: e.g. order statistics when $p = 1$.

B Choice of initial values

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
$$d^k(j) = \min_{1 \leq k' \leq k} \left\| \hat{\mathbf{F}}_j - \hat{\mathbf{F}}_{j_{k'}} \right\|_{w,2}^2 .$$
3. Given d^k from Step 2, randomly draw a cluster from $\{1, \cdots, J\} \setminus \{j_1, \cdots, j_k\}$, 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 $G^{(1)}(k)$.

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 definition 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 $\varepsilon^{*} = \min_{k \neq k'} \frac{1}{8}c(k,k')$,

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

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 $\|G(\lambda^{k}) - G(\lambda^{k_j})\|_{w,2}$: for every k, we can connect $\hat{G}(k)$ to $G(k)$. Then,

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

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:

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

as $J \to \infty$.

Claim 1

Firstly, find that

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

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

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

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 $\nu > 0$,

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

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,

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

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

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

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 $\widehat{ATE}^{cl} - \widetilde{ATE}^{cl}$:

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

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 $(a+b)^2 \geq 0$,

$$\frac{M}{2J}\sum_{j=1}^{J}\left(\hat{\pi}_j - \pi_j\right)^2 \leq \frac{M}{J}\sum_{j=1}^{J}\left[\left(\hat{\pi}_j - \pi\big(\bar{\lambda}(\hat{k}_j)\big)\right)^2 + \left(\pi\big(\bar{\lambda}(\hat{k}_j)\big) - \pi_j\right)^2\right] \tag{29}$$

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

$$G\big(\bar{\lambda}(k)\big) = \frac{\sum_{j=1}^{J} G(\lambda_j)\mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j = k\}} \tag{30}$$

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 (29).

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

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

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

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,

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

with some constant $C>0$.

Step 3.

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

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

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.

With a slight abuse of notation,

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

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

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

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

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

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,

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