Clustered treatment in multilevel models*

Myungkou Shin†

October 27, 2022

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, fully flexible cluster heterogeneity immediately fails identification of treatment effect. Thus, I use *selection-on-distribution* assumption that a cluster-level latent factor behind the cluster-level distribution of individual control covariate 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 macro heterogeneity in terms of the cluster-level distribution and micro 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 the 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 on the village level so that participants of the experiments can be clustered on the village level. (Voors et al., 2012; Giné and Yang, 2009; Banerjee et al., 2015)[1] In the 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 on the cluster level while an outcome variable of interest exists on the individual level. One can find such setup in many empirical contexts. 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 on the cluster level, it is not possible to model the cluster-level heterogeneity to be completely flexible in the outcome model, dropping *between* variation altogether and only using *within* variation. Any valid comparison to identify treatment effect has to be between at least two clusters and fully flexible cluster-level heterogeneity implies that a researcher is agnostic about which pair of clusters to compare. More observations per each cluster does not help us since the observations are all under the same regime; either treated or untreated. The infeasibilty of flexible cluster heterogeneity can be translated into the language of a linear regression as follows: cluster fixed-effects are infeasible with a cluster-level treatment variable due to multicollinearity.

The goal of this paper is to provide an econometric framework that fully incorporates the multilevel nature of such datasets and controls for cluster-level heterogeneity in treatment assignment and treatment effect,

[1]The multilevel structure is not confined to datasets where the unit of observation is a person. 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 also good examples. 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)

using the observable information. The key assumption used in this paper to achieve the goal is *selection-on-distribution* assumption: I assume that individual-level outcomes are independent of cluster-level treatment after conditioning on cluster-level control covariates and distribution of individual-level control covariates. With the *selection-on-distribution* assumption, I impose restriction on cluster heterogeneity: a pair of clusters with the same distribution of individuals, but with different treatment status, are deemed homogeneous and jointly identify treatment effect. The *selection-on-distribution* assumption naturally leads to the use of cluster-level distribution of individual control covariates in modeling cluster-level treatment assignment and individual-level outcomes. To implement that, I suggest the use of the K-means clustering algorithm, an unsupervised learning method to group clusters so 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 firstly estimate treatment effect with nonparametric estimators such as inverse probability weighting estimator, and secondly with parametric models such as linear regression with group-specific time fixed-effects. I apply both of the estimation strategies to the CPS data to reexamine the effect of minimum wage on teen employment and find heterogeneity both on the individual level and on the state level.

There are four main contributions of this paper. The first contribution of this paper is to the selection bias and treatment endogeneity literature. In the literature, *selection-on-observable* assumption that treatment is random after conditioning on observable information is often used to solve the selection bias problem. This *selection-on-observable* assumption is mostly used in contexts where treatment is assigned on the individual level; the information available for each individual is relatively low-dimensional, facilitating propensity score estimation. However, when treatment is assigned on the cluster level, the available information for a unit of treatment assignment is high-dimensional since it contains observable information for every individual in a given cluster, especially when the cluster is large. I solve this problem by adopting the *selection-on-distribution* assumption. By moving from unordered stack of individual-level control covariates to a distribution, the dimensionality of the conditioning object, which is vital to solve the selection bias problem with *selection-on-observable* approach, reduces down. Note that the distribution is invariant to any permutation on the individual labels within a cluster; the *selection-on-distribution* assumption essentially imposes that the label of individuals within a cluster should not matter. In this sense, *selection-on-distribution* can be thought of as a sum of two assumptions: *selection-on-observable* and exchangeability among individuals.

Secondly, I contribute to the literature of heterogeneous treatment effect. (Callaway and Sant'Anna, 2021; Borusyak et al., 2021; Sun and Abraham, 2021) The *selection-on-distribution* assumption of this paper requires that cluster-level distribution of individuals be used in modelling cluster-level treatment assignment and individual-level outcome. By using cluster-level distribution of individuals and individual-level control covariates together, treatment effect is modelled with two types of heterogeneity: macro heterogeneity from cluster-level information and micro heterogeneity from individual-level information. The notion of two layers of heterogeneity itself is not new; cluster-level control covariates are often used in regression of individual-

level outcomes. That being said, the framework from this paper provides a novel lens to document macro heterogeneity, the cluster-level distribution of individual control covariates. By documenting treatment effect heterogeneity in terms of the cluster-level distribution of individuals, the econometric framework of this paper answers a variety of intriguing research questions. For example, suppose a researcher is interested in how neighborhood of residence or migration affects individual outcomes, as in Derenoncourt (2022); Chetty et al. (2016). Natural research questions in such a setup are "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?", which would be micro heterogeneity, macro heterogeneity, and interaction between the two, in the terminology of this paper. The use of cluster-level distribution of individuals allows for these questions to be answered. Moreover, macro 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)

Thirdly, this paper makes contribution to the distributional regression literature. To estimate propensity score and treatment effect with cluster-level distribution of individuals, we need to use a functional regression method that regresses a one-dimensional variable onto a high-dimensional object such as distribution. By using the K-means algorithm, this paper proposes a simple and easy-to-understand way of such a functional regression, 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 special case of partition-based regression (Cattaneo et al., 2020); the K-means algorithm partitions clusters based on their distributions of individual-level control covariate and propensity score and treatment effect is projected onto a step function that is constant within a partition.

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 macro heterogeneity in state-level labor market fundamentals, by controlling for the distribution of individual employment status history. Then, I explore two channels of micro 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 to the discussion so far, there are several literatures that my paper relates to. Firstly, the *selection-on-distribution* assumption is comparable to the factor model assumption used in Pesaran (2006) and the synthetic control literature (Abadie et al., 2010, 2015). With a factor model, a linearity is imposed on a potentially high-dimensional time-series of observable control covariate as exchangeability is imposed on a stacked matrix of individuals in this paper. The difference is intuitive. In the case of panel data, the time dimension, the label of observations within each unit, conveys significant information. 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 to control for heterogeneity in the outcome model, similarly to this paper that matches clusters with the K-means grouping. Thirdly, in using the K-means algorithm as my functional regression method of choice, I assume a latent factor that governs the cluster-level distribution of individuals and and I assume the latent factor to be discrete with a finite support for the most part of the paper. In this sense, this paper relates to the group fixed-effect literature where a latent finite grouping structure is assumed. (Bonhomme and Manresa, 2015; Su et al., 2016; Ke et al., 2016) However, this paper differs from the literature in the sense that the grouping structure is not recovered from the outcome model, but from the distribution of control variables itself. Thus, the grouping result in this paper does not suffer from the overfitting problem. Lastly, when the latent factor is assumed to be continuous, as will be discussed in Section 5, this paper relates to Bester and Hansen (2016) in that I use the finite grouping structure as an approximation of more flexible latent factor structure.

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 nonparametric treatment effect estimators. In Section 4, I discuss asymptotic properties of the estimator, 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 covariate for individual i in cluster j, and $D_j \in \{0, 1\}$ is a cluster-level binary treatment variable for cluster j. Note that each individual belongs to one cluster and one cluster only: multiway clustering is excluded. There exist J clusters and N_j individuals for each cluster: in total there are $N = \sum_{j=1}^{J} N_j$ individuals. To discuss treatment effect, I let the observed outcome Y_{ij} 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 on the individual level but treatment is defined on the cluster level.

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

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

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

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

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

I assume cluster-level iid-ness with **Assumption 1**, with some cluster-level latent factor λ_j. The iid-ness discussed in **A1** comes from a two-step data generating process: firstly, cluster-level variables $\big(D_j, N_j, \lambda_j\big)$ are drawn from an iid distribution. Then, conditioning on the cluster-level variables $\big(D_j, N_j, Z_j\big)$, individual-level variables $\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}$ are drawn from distribution function H, which applies to every cluster; hence identical-ness, and independently of other clusters; hence independence. Dependence structure within a cluster is yet to be specified.

Assumption 2. *(selection-on-distribution) The latent factor λ_j contains all the information for the distribution of X_{ij}: there exists a function $G : \Lambda \to B(\mathbb{R}^p)$ such that for every $x \in \mathbb{R}^p$,*

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

Also,

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

From **Assumption 2**, the cluster-level latent factor λ_j contains all the relevant information on the distribution of X_{ij}; (average) conditional distribution of X_{ij} given $\big(D_j, N_j, \lambda_j\big)$ only depends on λ_j. In addition, **A2** assumes that the individual-level potential outcomes and the control covariates are independent of the cluster-level treatment status, after conditioning on the cluster-level variables: $H(D_j, N_j, \lambda_j) = H(N_j, \lambda_j)$. In other words, given the (average) conditional distribution of X_{ij} and the cluster size, the individual-level potential outcomes are independent of the treatment status. This can be thought of as *selection-on-observable*, but in the population sense; population quantity of $\mathbf{F}_j$ contains all the relevant information in treatment assignment so that the variation in D_j after conditioning on $\mathbf{F}_j$ is independent of the potential outcomes; thus I call **A2** *selection-on-distribution*.

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

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

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 **A2** is facilitated under **A3**; 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 learning this finite type for each cluster. I discuss the case where **A3** is relaxed and Λ is assumed to be a compact subset of $\mathbb{R}^q$, in Section 5.

2.1 Treatment effect

2.1.1 Aggregate treatment effect

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

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

$$ATT^{cl} = \mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)|D_j = 1\right]. \tag{2}$$

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:

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

$$ATT = \mathbf{E}\left[\frac{N_j}{\mathbf{E}\left[N_J|D_j = 1\right]}\left(\bar{Y}_j(1) - \bar{Y}_j(0)\right)\Big|D_j = 1\right] \tag{4}$$

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:

$$CATE^{cl}(\lambda) = \mathbf{E}[\bar{Y}_j(1) - \bar{Y}_j(0)|\lambda_j = \lambda]. \tag{5}$$

On the individual level, we again use **Assumption 2** and use an additional conditioning variable on the individual level to define conditional treatment effect parameters:

$$CATE(x, \lambda) = \mathbf{E}\left[Y_{ij}(1) - Y_{ij}(0)|X_{ij} = x, \lambda_j = \lambda\right]. \tag{6}$$

Note that after 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. The usual dimension to study treatment effect heterogeneity is on the individual-level conditioning variable:

$$\Delta_x CATE(x, \lambda).$$

I call this micro heterogeneity. In addition, we can also document the treatment effect heterogeneity in terms of cluster-level conditioning variables:

$$\Delta_\lambda CATE(x, \lambda).$$

I call this macro heterogeneity. Micro heterogeneity discusses how the treatment affects individuals with different characteristics differently while macro heterogeneity discusses how the treatment affects the same individual differently depending on which cluster they belong to.

The idea of the distinction between micro heterogeneity and macro heterogeneity is not unconventional. 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 micro heterogeneity and if the control covariate is a cluster-level variable, the interaction term captures macro 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.

There are a plenty of economic models and examples where we see both micro heterogeneity and macro heterogeneity in treatment effect. In the next subsection, I list three examples where the distribution of individual-level variables matters for treatment effect.

2.2 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 wage income distribution of their state: with X_{ijt} being the wage income of individual i in state j at time t, the state legislators observe

$$\mathbf{F}_{jt}(x) = \Pr\left\{X_{ijt}/P_t \leq x\right\}.$$

Inside the probability, the nominal wage income X_{ijt} is divided with a price level $P_t = (1+p)^t$: $\mathbf{F}_{jt}$ is the distribution of real wage income. It is assumed that the price level increases in a deterministic way, at the rate of p. The distribution $\mathbf{F}_{jt}$ has two determinants: underlying labor market fundamental and the minimum wage level. Let us denote the two with λ_{jt} and $MinimumWage_{jt}$, respectively:

$$\mathbf{F}_{jt} = \mathbf{F}(\lambda_{jt}, MinimumWage_{jt}/P_t).$$

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, $MinimumWage_{jt}$, is determined by the state legislators, in the process described below.

At each time period, the state legislators observe the realized wage income distribution. After observing the wage income distribution, 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:

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

g is labor market welfare function that takes the real wage distribution $\mathbf{F}_{jt}$ as its input and evaluates the social welfare generated in the labor market. If the state legislators only care about the unemployment rate, we would have $g(\mathbf{F}_{jt}) = g(\mathbf{F}_{jt}(0))$. If the state legislators care about the proportion of their constituents making below the federal poverty line, we would have $g(\mathbf{F}_{jt}) = g(\mathbf{F}_{jt}(\text{federal poverty line}))$. If the state legislators take many different aspects of the wage income distribution into their account, 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 real wage distribution is the key determinant in 'treatment' assignment process. Due to the Markov property, the potential outcomes, $X_{ijt'}$ in the next period, are independent of the treatment after conditioning on the distribution.

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

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

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 may be more reactive to the treatment, thus having positive treatment effect, while workers with lower productivity may not react and have a zero treatment effect. 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. 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

Suppose a school district experiments with a new teaching strategy across schools. Again, the decision of assigning teaching strategies to schools is not random and determined after considering student body of each school.

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, $\big(G_j\big)_{hi} = 1/4$ 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 $(x, \tilde{x}, \lambda)$,

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

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: micro 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 macro heterogeneity refers to a shift in the indirectly peer effect only.

3 Estimation

In this section, I propose a two-step approach of estimating $ATE^{cl}, ATT^{cl}, ATE, ATT, CATE^{cl}$ and $CATE$ in a nonparametric way. To estimate the treatment effect parameters with observables, there needs to be an estimator for the cluster-specific distribution of the individual-level control covariate, $\mathbf{F}_j$. I use the empirical distribution function: for all $x \in \mathbb{R}^p$,

$$\hat{\mathbf{F}}_j(x) = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\{X_{ij} \leq x\}. \tag{7}$$

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.

3.1 First step: propensity score

In the first step, I estimate the propensity score

$$\pi(\lambda) = \mathbf{E}[D_j | \lambda_j = \lambda] \tag{8}$$

as a function of the empirical distribution function $\hat{\mathbf{F}}_j$. I first propose an estimator based on K-means algorithm, which partitions J clusters into K groups. The asymptotic results and the empirical results of this paper are obtained using the K-means estimators. Then, I discuss various alternative estimators that one can consider instead of the K-means estimator. Though there is no estimator that clearly dominates

another, in terms of desirable statistical properties, I make a case on the K-means estimator for several expository qualities: the K-means estimator clearly defines control units to compare treated units to; the K-means estimator motivates a regression specification that naturally expands to cluster fixed-effects.

3.1.1 K-means estimator

When N_j is large, the conditioning covariate $\hat{\mathbf{F}}_j$ is a high-dimensional object. The high-dimensionality of $\hat{\mathbf{F}}_j$ makes functional regression methods with dimension reduction property more desirable. Though there are a lot of function regression methods that attain the dimension reduction property, I choose the K-means algorithm (Bonhomme et al., 2022). The K-means algorithm partitions observations into K groups, while minimizing the within-group variation of the observations. In our case, the unit of observations for the K-means algorithm is the clusters: $\hat{\mathbf{F}}_j$ is a cluster-level object.

To apply the K-means algorithm to propensity score estimation, we start with some predetermined $K \leq J$. With the predetermined K, the K-means algorithm assigns each cluster to one of K groups, based on $\|\cdot\|_{w,2}$, by solving the following minimization problem:

$$\left(\hat{k}_1, \cdots, \hat{k}_J, \hat{G}(1), \cdots, \hat{G}(K)\right) = \arg\min_{k,G} \sum_{j=1}^{J} \left\|\hat{\mathbf{F}}_j - G(k_j)\right\|_{w,2}^2. \tag{9}$$

The K-means minimization problem in (9) 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 (9) maps $\hat{\mathbf{F}}_j$ to $\hat{k}_j$, a discrete variable with finite support: dimension reduction.

The minimization problem does not have an analytical solution. Thus, an iterative algorithm is often used to solve the minimization problem. 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.

Once the grouping is complete, the propensity score estimates are

$$\begin{aligned} \hat{\pi}(k) &= \frac{1}{\sum_{j=1}^{J} \mathbf{1}\{\hat{k}_j = k\}} \sum_{j=1}^{J} D_j \mathbf{1}\{\hat{k}_j = k\}, \\ \hat{\pi}_j &= \hat{\pi}(\hat{k}_j). \end{aligned} \tag{10}$$

In practice, to avoid no overlap, we can either use a trimmed version of propensity score estimates or drop groups without overlap. This will be discussed more in Section 4.

3.1.2 Alternative estimators

There are multiple ways to do a functional regression other than the K-means algorithm. Firstly, there is a kernel estimator with l^2 norm (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 $\mathbf{F}$. Note that the kernel estimator does not have the dimension reduction property. Secondly, functional principal component analysis (functional PCA) can be an alternative to the K-means algorithm. (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

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

Then, by choosing the first K largest components of M, with some predetermined $K \leq J$, functional PCA maps $\mathbf{F}_j$ to a $K \times 1$ vector: dimension reduction. 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 $\left(\mu_1(\hat{\mathbf{F}}_j), \cdots \mu_K(\hat{\mathbf{F}}_J)\right)$ to a $\tilde{K} \times 1$ vector

$\big(\mu_{k_1}(\hat{\mathbf{F}}_j), \cdots, \mu_{k_{\tilde{K}}}(\hat{\mathbf{F}}_J)\big)$. In addition to the three examples I listed here, there are much more methods to do a functional regression.

That being said, I believe there are some qualitative benefits to the K-means algorithm. 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. Recall that Proposition 1 shows us that clusters with the same distribution of X_{ij} are comparable; a simple sample mean comparison across treated clusters and untreated clusters with the same distribution retrieves treatment effect. In practice, when N_J is large, it may not be straightforward from only the distribution functions $\hat{\mathbf{F}}_j$ to see which clusters are comparable to which. 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. Also, the grouping presents a natural guideline to summarize cluster heterogeneity. When cluster heterogeneity is continuously modeled, a researcher has to make a conscious choice in summarizing the heterogeneity. In the example of LASSO, by including an interaction term between treatment status variable and some moment $\mu_k(\hat{\mathbf{F}}_j)$, a researcher is implicitly deciding to impose linearity in treatment effect heterogeneity. With the K-means grouping, the researcher flexibly chooses which subsets of cluster to summarize the heterogeneity over.

Secondly, the use of fixed-effects based on the grouping in the regression specification, which will be discussed in Section 5, can be thought of as a natural adjustment of cluster fixed-effect. Cluster fixed-effect is one of the most common estimation strategies to control for cluster heterogeneity in a regression specification, when the dataset has a clustering structure. However, when a regressor of interest has no variation across individuals in a cluster, as the binary treatment variable described in the model of this paper in Section 2, cluster fixed-effects are infeasible due to the multicollinearity problem. Group fixed-effect suggested in Section 5 introduces fixed-effects on the group-level, instead of the cluster-level, which solves the multicollinearity problem as long as the regressor of interest has variation in each group. In the case of the treatment status, the restriction translates to that there needs to be overlap in treatment status in each group. Group fixed-effect approach imposes a restriction that clusters with similar observable characteristic share the fixed-effects and allow us to enjoy the flexibility of fixed-effect approach.

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. In contract, sparsity assumption with LASSO with a lot of moments does not have a clear interpretation as the K-means.

3.2 Second step: treatment effect

Given the propensity score estimators from (10), the cluster-level aggregate treatment effect are estimated as follows. Using the inverse probability weighting principle,

$$\widehat{ATE}^{cl} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j}\right), \tag{11}$$

$$\widehat{ATT}^{cl} = \frac{1}{\sum_{j=1}^{J} D_j}\sum_{j=1}^{J}\left(D_j\bar{Y}_j - \frac{(1-D_j)\hat{\pi}_j\bar{Y}_j}{1-\hat{\pi}_j}\right). \tag{12}$$

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

$$\widehat{ATE} = \frac{1}{N}\sum_{j=1}^{J} N_j\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j} - \frac{(1-D_j)\bar{Y}_j}{1-\hat{\pi}_j}\right), \tag{13}$$

$$\widehat{ATT} = \frac{1}{\sum_{j=1}^{J} D_j N_j}\sum_{j=1}^{J} N_j\left(D_j\bar{Y}_j - \frac{(1-D_j)\hat{\pi}_j\bar{Y}_j}{1-\hat{\pi}_j}\right). \tag{14}$$

Lastly, for each group estimated in the first step, we construct conditional treatment effect estimators. Here it is assumed that the support of X_{ij} is finite. $CATE$ with continuous X_{ij} is discussed in Section 5 under parametric model assumption.

$$\widehat{CATE}^{cl}(k) = \frac{\sum_{j=1}^{J}\bar{Y}_j D_j \mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J} D_j\mathbf{1}\{\hat{k}_j = k\}} - \frac{\sum_{j=1}^{J}\bar{Y}_j (1-D_j)\mathbf{1}\{\hat{k}_j = k\}}{\sum_{j=1}^{J}(1-D_j)\mathbf{1}\{\hat{k}_j = k\}}, \tag{15}$$

$$\widehat{CATE}(x,k) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_J} Y_{ij} D_j \mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}{\sum_{j=1}^{J}\sum_{i=1}^{N_J} D_j\mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_J} Y_{ij} (1-D_j)\mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}{\sum_{j=1}^{J}\sum_{i=1}^{N_J}(1-D_j)\mathbf{1}\{X_{ij} = x, \hat{k}_j = k\}}. \tag{16}$$

Note that $CATE^{cl}$ and $CATE$ are estimated with the categorical group membership variable $\hat{k}_j$, instead of an estimator on the latent factor λ_j, with which $CATE^{cl}$ and $CATE$ are defined with. 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 macro heterogeneity. It suffices to know which clusters have the same value of λ_j.

4 Asymptotic results

In this section, I discuss asymptotic properties of treatment effect estimators from Section 3. To ensure overlap, I trim the propensity score estimator to be on $[h, 1-h]$:

$$\hat{\pi}_j = \hat{\pi}(\hat{k}_j) = \min\left\{1-h, \max\left\{h, \frac{\sum_{l=1}^{J} D_l\mathbf{1}\{\hat{k}_l = \hat{k}_j\}}{\sum_{l=1}^{J}\mathbf{1}\{\hat{k}_l = \hat{k}_j\}}\right\}\right\} \tag{17}$$

with some $h \in (0, 0.5)$.

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

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

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

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

Also,

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

for large J.

Assumption 4.a) ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. **Assumption 4.b)** assumes that we have (uniform) overlap across treated clusters and control clusters, for each value of the latent factor. Under **Assumption 4.c)**, clusters with different values of the latent factor will be distinct from each other in terms of their distributions of X_{ij}. Thus, the K-means algorithm that uses $\hat{\mathbf{F}}_j$ is able to tell apart clusters with different values of λ_j, when $\hat{\mathbf{F}}_j$ is a good estimator for $\mathbf{F}_j$. **Assumption 4.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.

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. The condition that the size of cluster goes to infinity, at least for some clusters, is in some sense necessary for the high dimensionality problem of doing *selection-on-observable* in clustered treatment context to exist. When the size of cluster is uniformly bounded by a fixed constant, the stacked vector of the conditioning covariate $\{X_{ij}\}_{i=1}^{N_j}$ is a finite-dimensional object. That being said, it should be noted that **A4.d)** excludes cases where the size of cluster increases only for some clusters and is fixed for some other clusters, thus the estimator $\hat{\mathbf{F}}_j$ not improving as J grows for the second set of clusters. In these cases, the most straightforward strategy would be to consider the two sets of clusters separately.

The following theorem 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.

To get asymptotic results on the treatment effect estimators, let us adopt additional assumption on the individual-level variables. Firstly, find that for any (d, k), the expectation of $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\bar{Y}_j$ is equal to the expectation of $\mathbf{1}\{D_j = d, \lambda_j = \lambda^k\}\mathbf{E}\left[\bar{Y}_j(d)|N_j, \lambda_j = \lambda^k\right]$:

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

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

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

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

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

c) *Let*

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

Then,

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

as $J \to \infty$*, with*

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

d) *Let*

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

Then,

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

as $J \to \infty$*, with*

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

Assumption 5.a) puts a bound on 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

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

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)****,*

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

as $J \to \infty$*, where*

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

Proof. See Appendix. □

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.

5 Extension

5.1 Parametric model

In this section, I extend Theorem 1 and apply the result to a parametric multilevel model with a cluster-level variable $Z_j \in \mathbb{R}^{p^{cl}}$: with some function $g : \mathbb{R}^p \times \{0,1\} \times \mathbb{R}^{p^{cl}} \times \Lambda \to \mathbb{R}$,

$$Y_{ij} = g(X_{ij}, D_j, Z_j, \lambda_j; \theta_0) + U_{ij}, \tag{18}$$

$$0 = \mathbf{E}\left[U_{ij}|X_{ij}, D_j, Z_j, \lambda_j\right]. \tag{19}$$

To include a 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, $H^{hyper}\big(\{D_j, Z_j, \lambda_j\}_{j=1}^J\big)$, the conditional distribution of $\Big\{\{Y_{ij}(1), Y_{ij}(0), X_{ij}\}_{i=1}^{N_j}\Big\}_{j=1}^J$ given $\{D_j, Z_j, \lambda_j\}_{j=1}^J$, 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 (D_j, Z_j, λ_j):

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

The parametric model in (18) specifies H.

With the modified version of **Assumption 1** and the parametric model g, I adopt the following assumption for GMM estimation.

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) *$\theta = \big(\theta^1, \cdots, \theta^K\big)$ 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 **A6.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.

A GMM estimator of θ is

$$\hat{\theta} = \left(\hat{\theta}^1, \cdots, \hat{\theta}^K\right) = \arg\min_{\theta \in \Theta} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(Y_{ij} - \tilde{g}(X_{ij}, D_j, Z_j; \theta^{\hat{k}_j})\right)^2. \tag{20}$$

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. □

A direct application of Corollary 2 is a linear regression specification, such as group fixed-effects. Let

$$\begin{aligned} Y_{ij} &= g(X_{ij}, D_j, Z_j, \lambda_j; \theta_0) + U_{ij} \\ &= \delta_{\lambda_j} + \beta_{\lambda_j} D_j + {Z_j}^\intercal \eta^{cl} + {X_{ij}}^\intercal \eta + U_{ij}, \\ 0 &= \mathbf{E}\left[U_{ij} | X_{ij}, D_j, Z_j, \lambda_j\right]. \end{aligned} \tag{21}$$

The parameter of the model is $\theta = \left(\delta_1, \cdots, \delta_K, \beta_1, \cdots, \beta_K, \eta^{cl}, \eta\right)$. Note that the model satisfies **A6.c)**: $\theta^k = \left(\delta_k, \beta_k, \eta^{cl}, \eta\right)$. The exact value of λ_j does not matter.

5.2 Continuous λ

So far, 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) *(finite dimensional heterogeneity)* Λ *is a compact subset of* $\mathbb{R}^q$.

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

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

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

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

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

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

e) *$N_{\min,J} \to \infty$ as $J \to \infty$.*

f) *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.c-d) 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.f)** 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** and **7**,*

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

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.

5.3 Multilevel models with 3+ levels

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

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

which is high-dimensional, is summarized to

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

Given these cluster-level group membership $\hat{k}_{jl}$, construct the hypercluster-level distribution with cluster-level observables: for every $z \in \mathbb{R}^{p^{cl}}$ and $k \in \{k, 1 \cdots, K\}$,

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

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

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

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 Simulation

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 $j = 1, \cdots \mathrm{J}$. The data generating process I consider is

$$\begin{aligned}
\lambda_j &\in [-2, 2], \\
D_j &\overset{\text{iid}}{\sim} \text{Bernoulli}\left(\pi(\lambda)\right), \\
Y_{ij} &= \beta(\lambda_j) D_j + U_{ij}, \\
\left(U_{ij}, X_{ij}\right) \mid \left(D_j, \lambda_j\right) &\overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \text{diag}(1/4, 1)\right)
\end{aligned}$$

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

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

Figure 1 shows how the propensity score π and the treatment effect β changes with the latent factor λ.

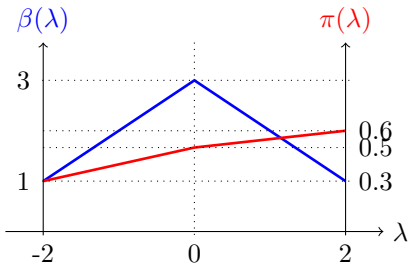

Figure 1: π and β

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.

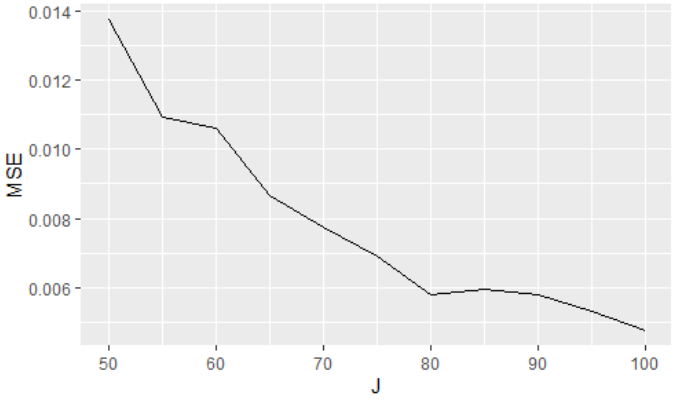

Figure 2: MSE as J and N_J increase: $J = 50, \cdots, 100$ and $N_j = 100, \cdots, 300$.

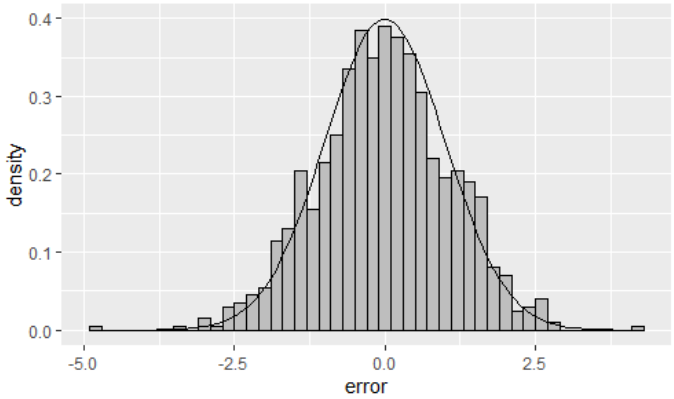

Figure 3: Asymptotic normality: $J = 100$ and $N_J = 300$.

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.

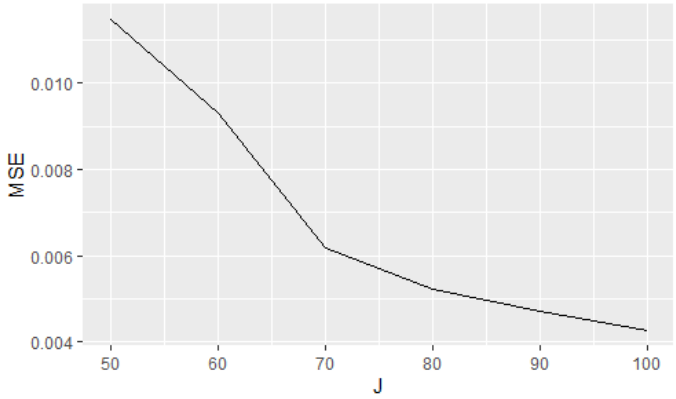

Figure 4: MSE as J increases: $J = 50, \cdots, 100$, $K = 5$ and $N_j = 200$

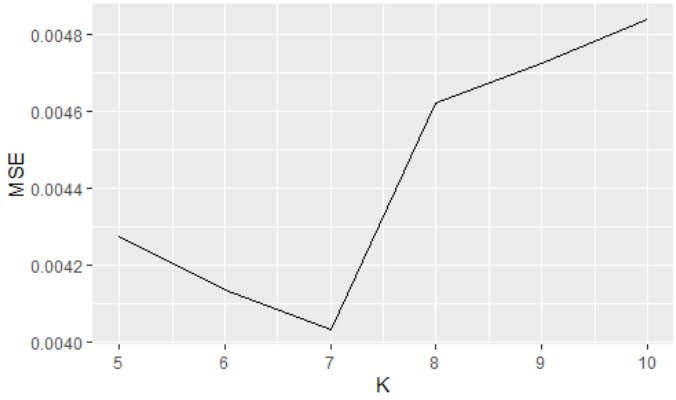

Figure 5: MSE as K increases: $J = 100$, $K = 5, \cdots, 10$ and $N_j = 200$

7 Application

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. If we would like to expand this, we can have a three-level model, for example, with counties as clusters or census divisions as hyperclusters. 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.

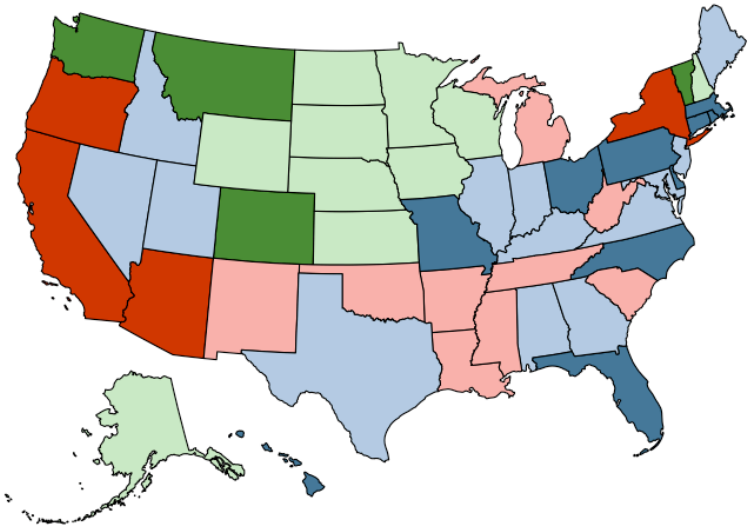

Figure 6: Grouping of states when $K = 3$

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.

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

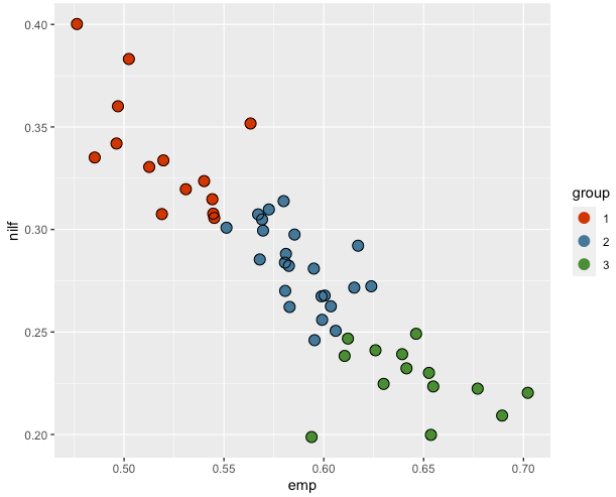

Figure 7: group heterogeneity

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 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,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{22}$$

$EmpRate_{jt}$ 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:

$$Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \beta \log MinWage_{jt} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \tag{23}$$

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 teen employment by 0.18 percentage point. Neumark and Shirley (2022) provides a meta-analysis of studies on teen employment and minimum wage and find that the mean of the estimates across studies is -0.170 and the median is -0.122. The estimate from the group fixed-effect specification is slightly above the mean.

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

Table 3: heterogeneity in treatment effect

The columns (4)-(6) of Table 3 discuss the macro 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 micro heterogeneity and macro heterogeneity. The left panel of Table 4 is from the linear specification

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \beta^1 \log MinWage_{jt} \mathbf{1}\{Age_{ijt} \leq 18\} \\ + \beta^2 \log MinWage_{jt} \mathbf{1}\{Age_{ijt} = 19\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{24}$$

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

$$\begin{aligned} Y_{ijt} = \alpha_j + \delta_{\hat{k}_{jt}t} + \sum_{k=1}^{3} \beta^1(k) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} \leq 18, \hat{k}_{jt} = k\} \\ + \sum_{k=1}^{3} \beta^2(k) \log MinWage_{jt} \mathbf{1}\{Age_{ijt} = 19, \hat{k}_{jt} = k\} + X_{ijt}{}^{\intercal}\eta + \eta^{cl} EmpRate_{jt} + U_{ijt}. \end{aligned} \tag{25}$$

Interaction between micro heterogeneity in terms of age and macro 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 micro heterogeneity in evident in all of the three specifications I considered, the interaction between the micro heterogeneity and the macro 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 micro 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 macro 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.

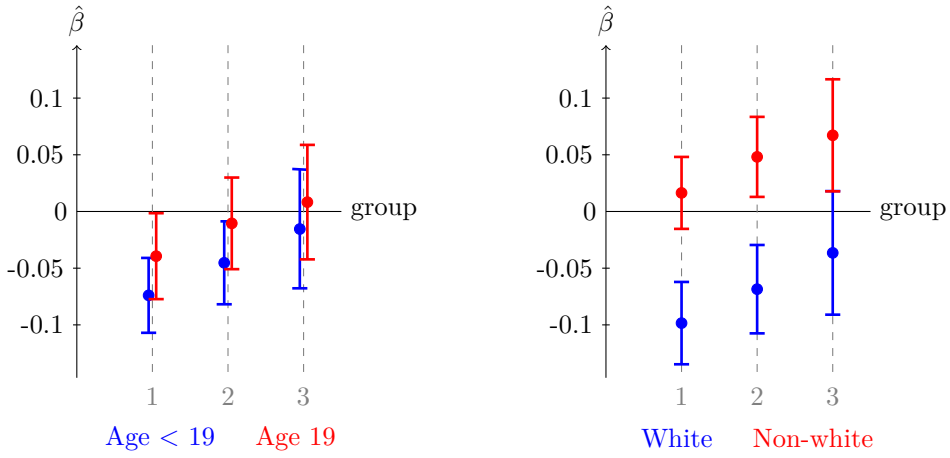

Figure 8: Interaction between micro and macro heterogeneity

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: micro heterogeneity and macro 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.

To my best knowledge, 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.

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

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.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, "Discretizing unobserved heterogeneity," *Econometrica*, 2022, *90* (2), 625–643.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2021.

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

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

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 evidence-from 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.

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.

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.

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.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

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.

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

A Proofs

A.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,

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

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

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

The second inequality is from **A5.c)**. The third inequality is from the construction of the event A_ε. In the last inequality A_ε can be dropped since the probability does not require σ being bijective. $(K-1)$ comes from repeating the argument for every $k \neq k_j$.

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

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

By repeating the expansion for every j,

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

We already know $\Pr\left\{{A_{\varepsilon^{**}}}^c\right\} = o(1)$ as $J \to \infty$. It remains to show that the first quantity in the RHS of the

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

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

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

A.2 Corollary 1

Let

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

with some abuse of notation. I let

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

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

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

as $J \to \infty$.

Claim 1

Firstly, find that

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

and

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

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

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

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

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

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

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

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

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

where

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

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,

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

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

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

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

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

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

Averaging: $\widehat{ATE}$

Again, with some abuse of notations with zero denominators,

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

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

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

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

A.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

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

A.4 Theorem 2

Let π_j, γ_{1j} and γ_{0j} denote $\pi(\lambda_j)$, $\gamma^{cl}(1,\lambda_j)$ and $\gamma^{cl}(0,\lambda_j)$ respectively. Define

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

Find that

$$\begin{aligned}
\widehat{ATE}^{cl}_{oracle} - \mathbf{E}\left[Y_{ij}(1) - Y_{ij}(0)\right] &= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\gamma_{1j}}{\pi_j} - \frac{(1-D_j)\gamma_{0j}}{1-\pi_j} - \mathbf{E}\left[Y_{ij}(1) - Y_{ij}(0)\right]\right) + \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j}{\pi_j} - \frac{1-D_j}{1-\pi_j}\right)\bar{U}_j \\
&= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\gamma_{1j}}{\pi_j} - \frac{(1-D_j)\gamma_{0j}}{1-\pi_j} - \mathbf{E}\left[Y_{ij}(1) - Y_{ij}(0)\right]\right) \\
&\qquad + \left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j}{\pi_j} - \frac{1-D_j}{1-\pi_j}\right)^2\right)^{\frac{1}{2}}\left(\frac{1}{N_J}\frac{1}{J}\sum_{j=1}^{J}\left(\sqrt{N_J}\bar{U}_j\right)^2\right)^{\frac{1}{2}} \\
&= o_p(1)
\end{aligned}$$

as $J \to \infty$. Next,

$$\begin{aligned}
\widehat{ATE}^{cl}_{ipw} - \widehat{ATE}^{cl}_{oracle} &= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right) D_j \bar{Y}_j - \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{1-\hat{\pi}_j} - \frac{1}{\pi_j}\right)(1-D_j)\bar{Y}_j \\
&= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right) D_j \gamma_{1j} - \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{1-\hat{\pi}_j} - \frac{1}{\pi_j}\right)(1-D_j)\gamma_{0j} + o_p(1), \\
\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j}\right) D_j\gamma_{1j} &= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\bar{\pi}_j} + \frac{1}{\bar{\pi}_j} - \frac{1}{\pi_j}\right) D_j\gamma_{1j} \\
&= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j} - \frac{1}{\bar{\pi}_j}\right) D_j\gamma_{1j} + \frac{1}{J}\sum_{j=1}^{J}\frac{1}{\bar{\pi}_j\pi_j}\left(\pi_j - \bar{\pi}_j\right) D_j\gamma_{1j} \\
&\leq \frac{1}{J}\sum_{j=1}^{J}\frac{D_j\gamma_{1j}}{\hat{\pi}_j\bar{\pi}_j}\left(\bar{\pi}_j - \hat{\pi}_j\right) + \frac{1}{J}\min_{\tilde{k}}\sum_{j=1}^{J}\left\|\lambda_j - \bar{\lambda}(\tilde{k}_j)\right\|_2^2 \\
&= -\frac{1}{J}\sum_{j=1}^{J}\frac{D_j\gamma_{1j}}{\hat{\pi}_j\bar{\pi}_j}\left(\bar{V}_j + (h-\bar{\pi}_j)\mathbf{1}\{\bar{V}_j \leq h - \bar{\pi}_j\} + (1-h-\bar{\pi}_j)\mathbf{1}\{\bar{V}_j \geq 1-h-\bar{\pi}_j\}\right) \\
&\quad + O_p\left(\frac{K}{N_J} + \frac{1}{K^{\frac{2}{q}}}\right) \\
&= O_p\left(\frac{K}{N_J} + \frac{1}{K^{\frac{2}{q}}} + \frac{K}{J}\right).
\end{aligned}$$

where

$$\begin{aligned}
\bar{\pi}_j = \bar{\pi}(\hat{k}_j) &= \frac{\sum_{l=1}^{J}\pi_l \mathbf{1}\{\hat{k}_l = \hat{k}_j\}}{\sum_{l=1}^{J}\mathbf{1}\{\hat{k}_l = \hat{k}_j\}}, \\
\bar{V}_j = \bar{V}(\hat{k}_j) &= \frac{\sum_{l=1}^{J}V_l \mathbf{1}\{\hat{k}_l = \hat{k}_j\}}{\sum_{l=1}^{J}\mathbf{1}\{\hat{k}_l = \hat{k}_j\}}.
\end{aligned}$$

For

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

I used

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

Step 1.

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

$$\left|\frac{1}{J}\sum_{j=1}^{J}\left(\frac{D_j\bar{Y}_j}{\hat{\pi}_j}-\frac{D_j\bar{Y}_j}{\pi\left(N_j,Z_j,\lambda_j\right)}\right)\right| \leq \left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j}-\frac{1}{\pi\left(N_j,Z_j,\lambda_j\right)}\right)^2\right)^{\frac{1}{2}}\cdot\left(\frac{1}{J}\sum_{j=1}^{J}\bar{Y}_j^2D_j\right)^{\frac{1}{2}}$$
$$= \left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j}-\frac{1}{\pi\left(N_j,Z_j,\lambda_j\right)}\right)^2\right)^{\frac{1}{2}}O_p(1)$$

from **A2.a**. Then, from Taylor's expansion, **A2.b** and **A3.a**,

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

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

$$\frac{M}{2J}\sum_{j=1}^{J}\left(\hat{\pi}_j-\pi\left(N_j,Z_j,\lambda_j\right)\right)^2 \leq \frac{M}{J}\sum_{j=1}^{J}\left[\left(\hat{\pi}_j-\pi\big(N_j,Z_j,\bar{\lambda}(\hat{k}_j)\big)\right)^2+\left(\pi\big(N_j,Z_j,\bar{\lambda}(\hat{k}_j)\big)-\pi\left(N_j,Z_j,\lambda_j\right)\right)^2\right] \tag{26}$$

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

$$G\big(\bar{\lambda}(k)\big)=\frac{\sum_{j=1}^{J}G(\lambda_j)\mathbf{1}\{\hat{k}_j=k\}}{\sum_{j=1}^{J}\mathbf{1}\{\hat{k}_j=k\}} \tag{27}$$

for $k=1,\cdots,K$. The existence of such $\bar{\lambda}$ and its uniqueness is guaranteed from **A2.d**.

Step 2.

Let us focus on the second quantity from (26).

$$\frac{1}{J}\sum_{j=1}^{J}\left(\pi\big(N_j,Z_j,\bar{\lambda}(\hat{k}_j)\big)-\pi\big(N_j,Z_j,\lambda_j\big)\right)^2 \leq \frac{M}{J}\sum_{j=1}^{J}\left\|\bar{\lambda}(\hat{k}_j)-\lambda_j\right\|_1^2$$
$$\leq \frac{M}{J}\sum_{j=1}^{J}q\left\|\bar{\lambda}(\hat{k}_j)-\lambda_j\right\|_2^2$$

with some constant $M>0$. The first inequality is from Taylor's expansion and **A2.c** and the second inequality is from Cauchy-Schwarz inequality.

From **A1.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 (9). Lastly, from **A3.b**, we have

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

and similarly

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

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,

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

with some constant $C>0$.

Step 3.

Let us apply Donsker's theorem to put a bound on $\frac{K}{J}\sum_{j=1}^{J}\left\|G(\lambda_j)-\hat{\mathbf{F}}_j\right\|_{\infty}^{2}$. From **A2.e**,

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

Thus,

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

(**A2.e** assumes X_{ij} being iid only after conditioning on cluster-level variables... does this matter?)

Step 4.

From **A2.g** let us construct a partition on $\{1,\cdots,J\}$ so that each set contains clusters where $\left(N_j,Z_j,\hat{k}_j\right)$ is identical. Suppose that the partition contains L sets. Let $\breve{k}_j\in\{1,\cdots,L\}$ be the membership indicator

where $\left(N_j, Z_j, \hat{k}_j\right) = (n_l, z_l, k_l)$ for every cluster j such that $\breve{k}_j = l$. From **A3.c**,

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}\left(\hat{\pi}_j - \pi\big(N_j, Z_j, \bar{\lambda}(\hat{k}_j)\big)\right)^2 \\
&= \frac{1}{J}\sum_{l=1}^{L}\breve{\#}(l)\left(\frac{\sum_{j=1}^{J}\pi\big(n_l, z_l, \lambda_j\big)\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)} + \frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)} - \pi\big(n_l, z_l, \bar{\lambda}(k_l)\big)\right)^2 \\
&\leq \frac{2}{J}\sum_{l=1}^{L}\breve{\#}(l)\left[\left(\frac{\sum_{j=1}^{J}\Big(\pi\big(n_l, z_l, \lambda_j\big) - \pi\big(n_l, z_l, \bar{\lambda}(k_l)\big)\Big)\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)}\right)^2 + \left(\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)}\right)^2\right] \\
&\leq \frac{2}{J}\sum_{l=1}^{L}\breve{\#}(l)\left[\frac{\sum_{j=1}^{J}\Big(\pi\big(n_l, z_l, \lambda_j\big) - \pi\big(n_l, z_l, \bar{\lambda}(k_l)\big)\Big)^2\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)} + \left(\frac{\sum_{j=1}^{J}V_j\mathbf{1}\{\breve{k}_j = l\}}{\breve{\#}(l)}\right)^2\right].
\end{aligned}$$

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

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

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

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

Now, it remains to put a bound on

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

From **A1.b** and **A2.f**,

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

Thus,

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

From **A2.g**, $L\propto K$. Thus,

$$\frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{\hat{\pi}_j}-\frac{1}{\pi\big(N_j,Z_j,\lambda_j\big)}\right)^2\leq O_p\left(\frac{K}{N_J^{\min}}+K^{-\frac{2}{q}}+\frac{K}{J}\right).$$

In the case of $\widehat{ATT}^{ipw}$, once we apply Taylor's expansion to get

$$\frac{1}{J}\sum_{j=1}^{J}\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}-\frac{\pi\big(N_j,Z_j,\lambda_j\big)}{1-\pi\big(N_j,Z_j,\lambda_j\big)}\right)^2\leq\frac{M}{2J}\sum_{j=1}^{J}\big(\hat{\pi}_j-\pi\big(N_j,Z_j,\lambda_j\big)\big)^2,$$

the rest follows.