Aggregating Individual-level Information in Large Clusters*

Myungkou Shin[†]

November 26, 2024

Click here for the latest version.

Abstract

When given an endogenous cluster-level explanatory variable in a clustered dataset, we cannot model the cluster-level heterogeneity to be fully flexible. To control for the cluster-level heterogeneity, I assume a finite-dimensional cluster-level latent factor that is one-to-one with the cluster-level distribution of individual-level characteristics. The cluster-level distribution aggregates the available information at the individual level in a flexible way. Thanks to the low-dimensionality of the latent factor, a variety of existing moment restriction models can be used to identify a parameter of interest in relation to the cluster-level distributions.

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

[†]School of Social Sciences, University of Surrey. Email: m.shin@surrey.ac.uk

1 Introduction

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

This paper builds up on this motivation and provides a generalized econometric framework for clustered datasets, with an additional source of the cluster-level heterogeneity: the cluster-level aggregation of individual-level information. The idea of using a cluster-level aggregation of the individual-level information to model the cluster-level heterogeneity goes back a long way in the econometrics literature: e.g., Mundlak (1978); Chamberlain (1982) and more. This idea can be motivated from two different perspectives. Firstly, it gives us an alternative method in controlling for the cluster-level heterogeneity when given a cluster-level explanatory variable of interest; a fully flexible method such as cluster fixed-effects is infeasible since it subsumes the variation from the cluster-level variable of interest. Secondly, the aggregation of the individual-level characteristics can be an explanatory variable of interest on its own when a researcher is interested in how the cluster-specific 'context' or 'equilibrium,' which is formulated from the within-cluster collection of the individual-level characteristics, affects the individual-level outcome.

As an illustrative example, consider the following linear regression model:

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

Y_{ij} is the individual-level outcome variable for individual i in cluster j, Z_j is the cluster-level explanatory variable for cluster j and X_{ij} is the individual-level control covariates for individual i in cluster j. The regression regression model (1) includes an element of cluster-level heterogeneity: the cluster fixed-effect α_j. Due to the multicollinearity between α_j and Z_j, the coefficient β is not identified in model (1).

In light of this problem, the researcher may want to impose cluster homogeneity and use the regression equation (2):

$$Y_{ij} = \alpha + \beta Z_j + X_{ij}{}^{\intercal}\theta + U_j, \tag{2}$$

When the true model is (1) and the cluster heterogeneity α_j correlates with Z_j, the OLS estimator for β from the regression (2) will be biased.[1]

Thus, I propose an alternative regression model (3), where we do not assume cluster homogeneity, but instead put restrictions on the cluster heterogeneity by modeling it with $\mathbf{F}_j$, the within-cluster distribution of X_{ij} for cluster j:

$$Y_{ij} = \alpha(\mathbf{F}_j) + \beta Z_j + X_{ij}{}^{\intercal}\theta + U_j. \tag{3}$$

[1]This problem closely relates to the treatment endogeneity problem; when Z_j is independent of the underlying heterogeneity α_j, the regression equation (2) identifies β.

This approach fully utilizes the rich information observed at the individual level. Given that Z_j varies across clusters with the same distribution of X_{ij}, the coefficient β is identified. In addition, the model (3) distinguishes the effect of the aggregate-level regressors and that of the individual-level regressors: (α, β) v. θ. $\alpha(\mathbf{F}_j)$ tells us how the individual-level characteristics collectively affects the individual-level outcomes. In fact, we can even further relax the regression model.

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

In the model (4), $\beta(\mathbf{F}_j)$ and $\theta(\mathbf{F}_j)$ discuss how the the effect of the cluster-level characteristic Z_j and that of the individual-level characteristic X_{ij} interact with the 'context' provided by the collection of the individual-level characteristics.

In this paper, the *distribution* function is used as a choice of the aggregation method to be applied to the individual-level characteristics. The dimension reduction property of the distribution as an aggregation method is particularly helpful when the clusters are large. When the clusters are large, the unordered collection $\{X_{ij}\}_i$ becomes high-dimensional even when X_{ij} itself is low-dimensional. In this regard, for the large clusters case, we need aggregation methods with some dimension reduction property. The distribution function is a sensible choice since there often does not exist any ordering among individuals in a clustered dataset; individuals are exchangeable within each cluster. For example, in a census data, the identification number has little meaning on its own.

The formal econometric framework of this paper consists of two parts: a constructive latent factor model for the cluster-level distributions, and a moment restriction model for a parameter of interest. In the latent factor model, the cluster-level distribution function of the individual-level control covariates is modeled to be an one-to-one function of a finite-dimensional cluster-level latent factor $\lambda_j \in \mathbb{R}^{\rho}$: the second layer of dimension reduction. For the moment restriction model, I impose a restriction that the moment function is invariant to a rotation on the latent factor in the sense that for any rotation on the latent factor space, there is a rotation on the parameter space such that the moment restriction model remains unchanged. Thanks to this rotation invariance, it suffices for the latent factor model to consistently estimate some rotation of the true latent factor.

Assuming a latent factor model for the cluster-level distribution has several merits. Firstly, it further reduces the dimension of a distribution function in a model-based manner. The two latent factor estimators discussed in this paper are both motivated by interpretable models: finite types of individuals model for the functional principal component analysis (PCA) estimator and finite types of clusters model for the K-means clustering estimator. Especially, the motivating assumption for the functional PCA estimator stays true to the fact that the high-dimensional object at hand is a *distribution* and it is easily interpretable in the context of the clustering structure.

In addition, having a finite-dimensional latent factor in place of the cluster-level distribution allows us to use a wide variety of existing moment restriction models as they are, when defining a parameter of interest in relation to the cluster-level distribution. This is in contrast to semiparametric estimators based on machine learning techniques that often require an orthogonalization specific to the model in use. At the cost of assuming a finite-dimensional latent factor in the model, the implementation procedure for the "distribution-as-control" approach becomes a simple two-step plug-in estimation; a researcher first estimates (rotated) latent factors from the cluster-level distributions and then use the estimated latent factors in a moment restriction model of their choice without any further modification.

This paper contributes to several literatures in econometrics. Firstly, this paper contributes to the literature of multilevel/hierarchical/clustered models. Similar to this paper, Yang and Schmidt (2021) identifies the effect of a possibly endogenous—meaning that it is correlated with the cluster-level heterogeneity—cluster-level variable in a linear regression setup; instead of modeling the cluster-level heterogeneity, they use instruments for the cluster-level explanatory variable. Arkhangelsky and Imbens (2023) also considers a multilevel setup and uses aggregation of individual-level information to control for the cluster-level heterogeneity; however, their goal differs from mine since they focus on small clusters with individual-level explanatory variable while I focus on large clusters.[2]

Secondly, this paper contributes to the literature of correlated random coefficient models. The simple linear regression example (1) above can be thought of as a random coefficient model where the coefficient α_j is possibly correlated with Z_j and/or X_{ij}. When given a cluster-level variable Z_j, a fixed-effect type approach (e.g., Wooldridge (2005); Graham and Powell (2012); Arellano and Bonhomme (2012)) is not applicable. Thus, I impose distributional assumptions on the random effect that the random coefficients are uncorrelated with

[2]Inherently, the problem they focus on only exists in small clusters setup; the within-cluster comparison will identify the effect of the individual-level explanatory variable when the clusters are large. On the other hand, the two motivations I give in this paper applies to both small and large cluster setups. Another difference is that their solution works for both small and large clusters by imposing additional parametric structure on the model while the solution of this paper is only valid for the large clusters case. Thus, one can consider the approach in Arkhangelsky and Imbens (2023) as an alternative to the latent factor model of this paper when clusters are small, at the cost of imposing additional parametric structure on the model.

(Z_j, X_{ij}) after conditioning on the cluster-level distributions. The parameter of interest in my model, for example $\alpha(\mathbf{F}_j)$ in (3), can be thought of as a conditional expectation of the random effect given the cluster-level distribution. In this sense, this paper is closer to Altonji and Matzkin (2005) and Bester and Hansen (2009) which also impose some restrictions on the joint distribution of the latent heterogeneity and the observable information.

Lastly, this paper contributes to the literature of the factor model approach in causal inference/program evaluation. The factor model approach in the program evaluation literature assumes that the error term consists of a systemic part, modeled with a factor model, and an idiosyncratic error and that the treatment endogeneity happens only through the factors. By assuming a latent factor model for the cluster-level distributions,[3] this paper also follows the same approach in solving the endogeneity problem of the cluster-level explanatory variable. On the contrary to the canonical synthetic control methods that aim to cancel out the latent factor (Abadie et al., 2010, 2015; Gunsilius, 2023) using pretreatment outcomes, this paper directly estimates the factors as did Xu (2017). As for using cluster-level distributions to control for the cluster-level heterogeneity, this paper shares the same spirit with Gunsilius (2023).

The rest of the paper is organized as follows. In Section 2, I formally discuss the two parts of the econometric framework of this paper: the latent factor model for the cluster-level distributions and the moment restriction model for

[3] In a factor model for interactive fixed-effects in panel data, the two dimensions of the factor model are unit and time. In this paper, the two dimensions of the factor model is clusters and individuals. The difference is that the individuals are not ordered in a clustered data as times are in a panel data. Thus, instead of directly modeling the variables X_{ij} with a factor model, I use the factor model to model the cluster-level distribution function of X_{ij}. Then, the factor loadings are connected to a relative position of an individual with $X_{ij} = x$ within a cluster, instead of being tied to a specific individual index i as it is to a time index t in a panel data

the parameter of interest. In Section 3, I discuss two latent factor models, which motivate the use of the functional PCA and the K-means clustering algorithm in estimating the cluster-level latent factor. In Sections 4-5, simulation results and an empirical illustration of the "distribution-as-control" approach that discusses the disemployment effect of the state-wise minimum wage in the United States are provided. All of the proofs for the theoretical results are given in the Supplementary Appendix.

2 Distribution as a control variable

2.1 Model

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

$1, \cdots, J$,

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

and for $i = 1, \cdots, N_j$ with a given j,

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

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

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

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

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

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

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

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

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

$\lambda \neq \lambda' \Rightarrow \|G(\lambda) - G(\lambda')\|_{w,2} > 0$ *and* $\|G(\lambda_j)\|_{w,2}$ *is bounded.*

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

The idea of using the cluster-level distribution $\mathbf{F}_j$ as a control covariate for the cluster-level heterogeneity is naturally appealing in the following two contexts. Firstly, suppose that the econometrician is interested in identifying the effect of a cluster-level observable characteristic Z_j on individual-level outcome Y_{ij} while allowing for some cluster-level latent heterogeneity.[4] The cluster-level heterogeneity cannot be modeled to be fully flexible with cluster fixed-effects, due to the limitation that there is no within-cluster variation in Z_j. An alternative to using cluster fixed-effects is to aggregate X_{ij} for each cluster, assuming that aggregating individual-level information for each cluster sufficiently controls for cluster-level heterogeneity. This approach is easy-to-implement when the cluster sizes are relatively small. However, when the clusters are large, a simple collection of the individual-level information $\{X_{ij}\}_{i=1}^{N_j}$ will be high dimensional. Thus, to impose some dimension reduction on the naive collection of control covariates $\{X_{ij}\}_{i=1}^{N_j}$, the econometrician

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

may want to use the fact that the order of individuals in a cluster often does not provide additional information in a clustered dataset.[5] In these empirical contexts, the cluster-level distribution $\mathbf{F}_j$ reduces the dimension of the simple collection $\{X_{ij}\}_{i=1}^{N_j}$, while preserving the relevant information regarding the cluster-level heterogeneity.[6]

Secondly, there are empirical contexts where the econometrician is directly interested in identifying the effect of the cluster-level distribution $\mathbf{F}_j$ on Y_{ij}, i.e. the aggregate-level equilibrium/contextual effect. In these cases, the cluster-level distribution of individual-level control covariates is a 'regressor' of interest on its own: e.g., the effect of state-level wage income distribution on an individual's disemployment probability; the effect of a school's racial composition on student's academic performance, etc. For these research questions, having a model G for the distribution function as in Assumption 1 can be particularly helpful if we want to discuss out-of-sample prediction for the outcome Y_{ij} given an unprecedented aggregate-level policy intervention for the cluster-level distribution: e.g., what happens when policymakers exogenously shifts the racial composition of a school? Cluster fixed-effects will successfully control for the cluster-level heterogeneity, but will not be able to give us a prediction for $\mathbf{E}[Y_{ij}|\mathbf{F}_j = \mathbf{F}]$ when $\mathbf{F}$ is different from $\{\mathbf{F}_1, \cdots, \mathbf{F}_J\}$. Moreover, by explicitly modeling the 'context' with $\mathbf{F}_j$, we can discuss how a counterfactual policy at the aggregate level differentially affects individuals, by looking at $\frac{\partial}{\partial x}\left(\mathbf{E}[Y_{ij}|(X_{ij}, \mathbf{F}_j) = (x, \mathbf{F})] - \mathbf{E}[Y_{ij}|(X_{ij}, \mathbf{F}_j) = (x, \mathbf{F}')]\right)$.

[5] In this sense, a panel data cannot be thought of as an example of a clustered dataset discussed in this paper. Also, in some cases, clustered datasets order individuals in a specific way as well; e.g., siblings being ordered in their birth order within a family, workers being ordered in terms of seniority within a firm, etc. In these cases, the order of the individual may contain information and therefore should not be ignored.

[6] See Section A of the Supplementary Appendix on how within-cluster exchangeability motivates the use of the cluster-level distribution as a control.

In addition to suggesting that the cluster-level distribution $\mathbf{F}_j$ be used as a control covariate in controlling for the cluster-level heterogeneity, Assumption 1 also assumes that the latent factor λ_j is finite-dimensional: $\mathcal{S}_\lambda \subset \mathbb{R}^\rho$. Thus, under Assumption 1, an infinite-dimensional object $\mathbf{F}_j$ is reduced to a finite-dimensional factor λ_j, through G. This adds an additional layer of dimension reduction, on top of aggregating the individual-level information to a distribution. By modeling $\mathbf{F}_j$ to be a function of λ_j, the task of estimating an infinite-dimensional object $\mathbf{F}_j$ becomes an easier task of estimating a finite-dimensional factor λ_j. Also, a variety of econometric frameworks that use finite-dimensional control covariates become readily applicable by substituting λ_j for $\mathbf{F}_j$. For example, when a researcher wants to construct a model where a binary outcome Y_{ij} depends on a distribution function $\mathbf{F}_j$, they can directly use a pre-existing logistic regression command in a statistical software, by substituting λ_j for $\mathbf{F}_j$. The low-dimensionality of the latent factor λ_j makes the "distribution-as-control" approach easily implementable.

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

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

Let l denote the dimension of m and k denote the dimension of θ: $l \geq k$. W_j^* is a function of cluster-level random objects $\left(\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j, N_j, \lambda_j\right)$. Note that λ_j is latent; the superscript * is used to denote that W_j^* is not directly observed. In addition, W_j^* is set to be a function of the latent factor λ_j, instead of the cluster-level distribution $\mathbf{F}_j$. In this sense, the model (6) can be understood as a derived model that we get from applying Assumption

1 to an original model written in terms of the cluster-level distribution $\mathbf{F}_j$. Assumption 1 will not be explicitly invoked for the rest of the paper.

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

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

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

$$\big(Y_{ij}(1), Y_{ij}(0), X_{ij}\big) \perp\!\!\!\perp Z_j \;\big|\; (N_j, \lambda_j)\,. \tag{7}$$

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

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

In this example, the binary treatment variable varies at the cluster level. The unconfoundedness assumption (7) assumes that the treatment is independent of the potential outcomes conditioning on the latent factor λ_j, i.e., the cluster-level distribution of X_{ij}. The treatment is as good as random between two clusters with the same distribution of individual-level characteristics. Suppose for example that the econometrician is interested in the effect of a state-wide policy on individual-level outcomes in the United States. The unconfoundedness assumption (7) would be to assume that the adoption of the policy is independent of the potential outcomes of a given state, conditioning on the state-level distribution of individuals. Then, we compare two states with the same distribution of individual characteristics to estimate the effect of the policy adoption.

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

$$\begin{aligned} Y_{ij} &= X_{ij}^{\intercal}\theta_1 + Z_j^{\intercal}\theta_2 + \lambda_j^{\intercal}\theta_3 + U_{ij}, \\ 0 &= \mathbf{E}\left[U_{ij} | X_{ij}, Z_j, N_j, \lambda_j\right]. \end{aligned} \tag{8}$$

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

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

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

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

when $\mathbf{F}, \mathbf{F}' \in G(\mathcal{S}_\lambda)$.

2.2 Plug-in estimation with the latent factor estimates

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

Consider an invertible $\rho \times \rho$ matrix A and the rotated latent factor $\tilde{\lambda} = A\lambda \in A\mathcal{S}_\lambda$. Then, by letting $G_A(\tilde{\lambda}) = G(A^{-1}\tilde{\lambda})$, Assumption 1 holds with G_A. Likewise, by modifying the construction of W_j so that it is a function of $\left(\{Y_{ij}, X_{ij}\}_{i=1}^{N_j}, Z_j, N_j, A^{-1}\tilde{\lambda}_j\right)$, the moment restriction model (6) holds as well. Thus, the rotated moment restriction model can also be thought of as an implication of Assumption 1 and an original moment restriction model defined with $\mathbf{F}_j$.

To discuss the moment function m, both in the context of the true latent factor and the rotated latent factor, construct a function W which takes cluster-level observable variables and the latent factor λ and computes the observation relevant for the moment restriction model (6). Then,

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

A slight abuse of notation is applied here since the dimension of the input changes with N_j. Also, let

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

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

Assumption 2.

- ***a.*** Θ, *the parameter space for* θ, *is a compact subset of* $\mathbb{R}^k$.
  The true value of θ, *denoted with* θ^0, *lies in the interior of* Θ.

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

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

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

- ***d.*** *There are (random) invertible* $\rho \times \rho$ *matrices* A *and* $\tilde{A}$ *such that for each*

$\theta \in \Theta$, $W_j = W_j(A\lambda_j)$ satisfies

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

almost surely.

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

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

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

Assumption 2.a-c are the standard sufficient conditions for consistency of an extremum estimator. Assumption 2.d assumes that the model is invariant to a rotation on the latent factor. Assumption 2.e assumes that the first derivative of the moment function with regard to the latent factor is bounded in expectation when evaluated within a small neighborhood around the true latent factor. Assumption 2.f assumes that the rotation does not change the scale of the latent factor and the parameter θ.

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

enters the moment function m as a single index of ${\lambda_j}^\intercal \theta_\lambda$ when $\theta = ({\theta_\lambda}^\intercal, {\theta_{-\lambda}}^\intercal)^\intercal$.

The appeal of this assumption in an empirical researcher's perspective hinges on whether the rotated parameter of interest $\tilde{A}\theta$ still has an interpretable implication as the original parameter of interest θ. In Example 1, assuming the rotation invariance on the propensity score model does not affect the ATE parameter $\mathbf{E}\left[\bar{Y}_j(1) - \bar{Y}_j(0)\right]$; the ATE parameter does not lose its casual interpretation. In Example 2, it is straightforward to see that the linear regression model satisfies the rotation invariance and the rotated parameter of interest is $\tilde{A}\theta = \left({\theta_1}^\intercal, {\theta_2}^\intercal, \left({A^\intercal}^{-1}\theta_3\right)^\intercal\right)^\intercal$. The slope coefficients on X_{ij} and Z_j remain unchanged. Moreover, the comparative statistics in terms of $\mathbf{F}_j$ can still be constructed using $\tilde{A}\theta$: given ${\theta_3}^\intercal A^{-1}$ and ${G_A}^{-1}$,

$$
\begin{aligned}
&\mathbf{E}\left[Y_{ij} | (X_{ij}, Z_j, N_j, \mathbf{F}_j) = (x, z, n, \mathbf{F}')\right] - \mathbf{E}\left[Y_{ij} | (X_{ij}, Z_j, N_j, \mathbf{F}_j) = (x, z, n, \mathbf{F})\right] \\
&\quad = {\theta_3}^\intercal \left(G^{-1}\left(\mathbf{F}'\right) - G^{-1}\left(\mathbf{F}\right)\right) \qquad \cdots \textit{what we want} \\
&\quad = {\theta_3}^\intercal A^{-1} \left({G_A}^{-1}\left(\mathbf{F}'\right) - {G_A}^{-1}\left(\mathbf{F}\right)\right) \qquad \cdots \textit{what we construct from data}
\end{aligned}
$$

when $\mathbf{F}, \mathbf{F}' \in G(\mathcal{S}_\lambda)$.[7]

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

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

[7]The second equality holds since

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

Theorem 1. *Suppose that Assumption 2 holds and there exists an consistent estimator* $\{\hat{\lambda}_j\}_{j=1}^J$ *for* $\{\lambda_j\}_{j=1}^J$ *such that*

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

Then,

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

as $J \to \infty$.

Theorem 1 assumes that the researcher is given some $\sqrt{J}$-consistent estimator for the rotated latent factor $A\lambda_j$: $\sum_{j=1}^J \left\| \hat{\lambda}_j - A\lambda_j \right\|_2^2 = o_p(1)$.

In addition to Assumption 2, Assumption 3 assumes additional conditions on the differentiability of m with regard to θ. Theorem 2 establishes the asymptotic normality.

Assumption 3.

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

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

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

$$\sup_{\theta' \in \Theta} \left\| \frac{1}{J}\sum_{j=1}^J \frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta'} - \mathbf{E}\left[\frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta'}\right] \right\|_F \xrightarrow{p} 0$$

as $J \to \infty$.

Theorem 2. *Suppose that $\hat{\theta}$ satisfies*

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

in addition to the conditions in Theorem 1. Then,

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

as $J \to \infty$, with some consistently estimable covariance matrix Σ.

The approximation error rate condition in Theorem 1 is sufficiently fast that the plug-in estimator $\hat{\theta}$ constructed with the estimated factors $\{\hat{\lambda}_j\}_{j=1}^J$ admits the same asymptotic distribution as the infeasible GMM estimator constructed with the rotated true factors $\{A\lambda_j\}_{j=1}^J$ and thus no additional correction is needed for inference on $\tilde{A}\theta^0$.

3 Latent factor models for distribution

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

3.1 Functional principal component analysis

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

$$\hat{M}_{jk} = \begin{cases} \frac{\sum_{i=1}^{N_j}\sum_{i'=1}^{N_k}}{N_j N_k} \int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)}{det(H)^{\frac{1}{2}}} \cdot \frac{K\left(H^{-\frac{1}{2}}(x-X_{i'k})\right)}{det(H)^{\frac{1}{2}}} w(x)dx, & \text{if } j \neq k \\ \frac{\sum_{i=1}^{N_j}\sum_{i'\neq i}}{N_j(N_j-1)} \int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)}{det(H)^{\frac{1}{2}}} \cdot \frac{K\left(H^{-\frac{1}{2}}(x-X_{i'j})\right)}{det(H)^{\frac{1}{2}}} w(x)dx, & \text{if } j = k, \end{cases}$$

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

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

where $\mathbf{f}_j$ is the cluster-level density function of the individual-level control covariates X_{ij} for cluster j. Note that the density function is not directly estimated; only the $J(J+1)/2$ moments are estimated.

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

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

where $\hat{q}_r = \big(\hat{q}_{r1}, \cdots, \hat{q}_{rJ}\big)^{\intercal}$ is the eigenvector associated with the r-th eigenvalue. The rescaling with $\sqrt{J}$ is introduced so that the estimated latent factor $\hat{\lambda}_j$ does not converge to zero as J grows: $\hat{q}_r^{\intercal}\hat{q}_r = 1$ is imposed in the eigenvalue decomposition. Again, the estimated latent factor $\hat{\lambda}_j$ is not unique. In an eigenvalue decomposition, the eigenvectors are uniquely determined only up to a sign even when the eigenvalues are all distinct.

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

Assumption 4. (Finite Types of Individuals) *With some $C > 0$,*

a. *(finite mixture model for distribution)*

$$\mathcal{S}_\lambda = \left\{ \lambda = \big(\lambda_1, \cdots, \lambda_\rho\big)^{\intercal} \in \mathbb{R}^\rho : \lambda_r \geq 0 \ \forall r \ \text{and} \ \sum_{r=1}^{\rho} \lambda_r = 1 \right\}$$

and there exist thrice continuously differentiable distribution functions $G_1, \cdots, G_\rho$ such that for any $x \in \mathbb{R}^p$,

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

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

$r = 1, \cdots, \rho,$

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

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

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

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

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

Assumption 4.a assumes that the cluster-level distribution function $\mathbf{F}_j$ is a mixture of ρ underlying distributions $G_1, \cdots, G_\rho$, with the latent factor λ_j as the mixture weights. In addition to motivating the finite mixture model for density, Assumption 4.a assumes that the underlying density functions $g_1, \cdots, g_\rho$ are smooth and bounded, up to their third derivatives. Under Assumption 4.a, the product matrix M can be rewritten as follows:

$$M_{jk} = \int_{\mathbb{R}^p} \mathbf{f}_j(x)\mathbf{f}_k(x)w(x)dx = \sum_{r,r'} \lambda_{jr}\lambda_{kr'} \int_{\mathbb{R}^p} g_r(x)g_{r'}(x)w(x)dx$$

$$M = \begin{pmatrix} \lambda_1^\intercal \\ \vdots \\ \lambda_J^\intercal \end{pmatrix} \underbrace{\begin{pmatrix} \int_{\mathbb{R}^p} g_1(x)^2 w(x)dx & \cdots & \int_{\mathbb{R}^p} g_\rho(x)g_1(x)w(x)dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}^p} g_1(x)g_\rho(x)w(x)dx & \cdots & \int_{\mathbb{R}^p} g_\rho(x)^2 w(x)dx \end{pmatrix}}_{=:V} \begin{pmatrix} \lambda_1^\intercal \\ \vdots \\ \lambda_J^\intercal \end{pmatrix}^\intercal.$$

Assumption 4.b assumes that the underlying density functions $g_1, \cdots, g_\rho$ have sufficient variation, when measured with $\langle \cdot, \cdot \rangle_w$. Particularly, the asymptotic separation of eigenvalues can be thought of as a strong factor separation condition in the canonical factor model for panel data. Assumption 4.c assumes that the cluster sizes grow jointly.

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

Proposition 1. *Suppose that Assumption 4 holds and that the kernel K used in the estimation procedure satisfies*

i. *K is bounded, nonnegative and symmetric around zero in the sense that $\int_{\mathbb{R}^\rho} t_r K(t) dt = 0$ for any r-th component t_r of t.*

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

iii. *$\int_{\mathbb{R}^\rho} |t_r t_{r'}| K(t) dt \leq C$ for any two components t_r and $t_{r'}$ of t.*

and the positive definite weighting matrix H satisfies

iv. *$N_{\min} \cdot det(H)^{\frac{1}{2}}$ goes to infinity as $J \to \infty$.*

v. *$\left\| H^{\frac{1}{2}} \right\|_F \propto N_{\min}^{-\nu}$ for some $\nu \in [0.25, 1)$.*

Then,

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

with some $\rho \times \rho$ matrix A, which is invertible and bounded with probability going to one.

Proposition 1 bounds the estimation error rate with the square root of the relative growth rate. When $J/N_{\min} \to 0$ as $J \to \infty$, the approximation error condition for Theorem 1 is satisfied. In the next subsection, I take the K-means clustering algorithm and provide an alternative latent factor model that allows for a slower growth rate on $N_{\min}$, at the cost of restricting across-cluster heterogeneity.

3.2 K-means clustering

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

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

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

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

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

To solve (9), I use the (naive) K-means clustering algorithm. Find that at the optimum

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

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

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

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

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

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

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

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

Once the K-means minimization problem is solved, I use the estimated group $\hat{\lambda}_j$ as the estimated latent factor, by transforming it to a categorical variable: with $e_1, \cdots, e_\rho$ being the elementary vectors of $\mathbb{R}^\rho$,

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

Note that the estimated latent factor $\hat{\lambda}_j$ is not unique. Given the grouping structure $\hat{\lambda}_j$ and the centeroids $\hat{G}(\lambda)$, we can find a relabeling on $\hat{\lambda}_j$ and $\hat{G}(\lambda)$ such that the minimum for (9) is still attained. Thus, we cannot take the face value of $\hat{\lambda}_j$ and interpret it to be an estimator for the true latent factor λ_j.

Now, it remains to develop an econometric model where the estimator for the latent factor using the K-means clustering algorithm is actually a consistent estimator for the true latent factor with sensible interpretation, at

[8] For simplicity of the discussion, let the weighting function w in $\|\cdot\|_{w,2}$ be discrete and finite: with some $x^1, \cdots, x^d \in \mathbb{R}^p$,

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

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

the rate discussed in Theorems 1-2. Assumption 5 discusses a set of conditions for that.

Assumption 5. (Finite Types of Clusters)

a. *(no measure zero type)* $\mathcal{S}_\lambda = \{e_1, \cdots, e_\rho\}$ *and* $\mu(r) := \Pr\{\lambda_j = e_r\} > 0$ $\forall r = 1, \cdots, \rho$.

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

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

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

Assumption 5.a ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. Under Assumption 5.b, clusters with different values of the latent factor will be distinct from each other in terms of their distributions of X_{ij}, analogous to the eigenvalue separation condition in Assumption 4. Thus, the K-means algorithm that uses $\hat{\mathbf{F}}_j$ is able to tell apart clusters with different values of λ_j when N_j is large and thus $\hat{\mathbf{F}}_j$ is a good estimate for $\mathbf{F}_j$. Assumption 5.c assumes that the cluster sizes grows jointly.

The key element of the econometric model described in Assumption 5 is that there are finite types of clusters, in terms of their distribution of individual-level control covariates X_{ij}. Thus, using Assumption 5 to model the cluster-level heterogeneity would make sense when we expect that the heterogeneity across clusters are discrete and finite. Also, we can make a following comparison between the two latent factor models proposed in this paper: Assumption 5 for the K-means clustering algorithm assumes that there are finite types of *clusters*, while Assumption 4 for the functional PCA assumes that there are finite types of *individuals* across clusters.

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

Proposition 2. *Suppose that 5 holds. Then, there is a rotation matrix A such that*

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

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

$$\left\|\begin{pmatrix}\hat{\lambda}_1 & \cdots & \hat{\lambda}_J\end{pmatrix} - A\begin{pmatrix}\lambda_1 & \cdots & \lambda_J\end{pmatrix}\right\|_F = \left(2\sum_{j=1}^{J}\mathbf{1}\left\{\hat{\lambda}_j \neq A\lambda_j\right\}\right)^{\frac{1}{2}} = o_p\left(1\right).$$

Proposition 2 shows that the misclassification probability of the K-means algorithm grouping clusters with different values of λ_j together goes to zero when $J/{N_{\min}}^{\nu^*}$ goes to zero for some $\nu^* > 0$. The rate on the misclassification probability is the same rate found in the literature: e.g., Bonhomme and Manresa (2015). Note that $N_{\min} \propto J$ is a sufficiently fast growth rate in the case of the K-means clustering for the hypertheorems as opposed to the case of the functional PCA; further restricting cluster-level heterogeneity allows for a slower growth rate on the cluster size.

4 Simulation

To discuss finite-sample performance of the functional PCA estimator and the K-means estimator, I simulated 500 cluster-level random samples for three different data generating processes. The baseline model specification is as follows: for $j = 1, \cdots, J$ and $i = 1, \cdots, N$,

$$Y_{ij} = 2D_j + \mu(\lambda_j) + \varepsilon_{ij}$$

where $D_j \perp\!\!\!\perp \{X_{ij}, \varepsilon_{ij}\}_{i=1}^{N} \mid \lambda_j$ and

$$\begin{aligned} D_j \mid \lambda_j &\sim \text{Bernoulli}\big(\pi(\lambda_j)\big) \\ \begin{pmatrix} X_{ij} \\ \varepsilon_{ij} \end{pmatrix} \Bigg| \lambda_j &\overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu(\lambda_j) \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma(\lambda_j)^2 & 0 \\ 0 & N/4 \end{pmatrix} \right). \end{aligned}$$

D_j is a cluster-level binary treatment variable that is potentially correlated with $\mu(\lambda_j)$. The variance of ε_{ij} is set to be proportional to N so that the only gain from larger N is improvement in the latent factor estimation and the variance of $\bar{Y}_j$ stays the same. The functions π, μ and σ and the distribution of λ_j vary across the three DGPs.

The first DGP admits two types of individuals, satisfying Assumption 4. I apply the functional PCA estimator to the first DGP. The second DGP admits two types of clusters, satisfying Assumption 5. I apply the K-means estimator to the second DGP. Lastly, to discuss misspecification, the third DGP is constructed in a way that it does not reduce down to any of the two latent factor models discussed in Section 3. I apply both the functional PCA estimator and the K-means estimator to the third DGP. The specifics of the

DGPs are provided in the notes for Tables 1-3.

For a random sample from each of the three DGPs, I firstly apply the corresponding latent factor estimation method (both for the misspecification exercise) to estimate the latent factor. Then, I estimate the propensity to be treated, using the estimated latent factor. For the functional PCA estimated factors, I run OLS: $\hat{\pi}(\lambda) = \hat{\lambda}^{\intercal}\hat{\pi}$. For the K-means estimated factors, I simply take the sample mean: $\hat{\pi}(\lambda) = \frac{\sum_{j=1}^{J} D_j \mathbf{1}\{\hat{\lambda}_j = \lambda\}}{\sum_{j=1}^{J} \mathbf{1}\{\hat{\lambda}_j = \lambda\}}$. Using the estimated propensity score, β is estimated using the inverse probability weighting characterization: $\hat{\beta} = \sum_{j=1}^{J} \left(\frac{D_j}{\hat{\pi}\left(\hat{\lambda}_j\right)} - \frac{1-D_j}{1-\hat{\pi}\left(\hat{\lambda}_j\right)} \right) \bar{Y}_j$. As a benchmark, I also computed a simple mean difference estimator.

Table 1 contains the simulation result for the first DGP and the third column shows that the average R^2 of regressing λ_j on $\hat{\lambda}_j$ improves as N increases. Though the functional PCA estimators explain on average 90% of variation in the true latent factors when $N \geq 50$, the estimator suffers from large variance when N is small. Table 2 contains the simulation results for the second DGP. From the third and the fourth columns, we can see that the classification im-

Table 1—Finite Types of Individuals

		mean diff.		fPCA		
J	N	bias	MSE	$\mathbf{E}\left[R^2\right]$	bias	MSE
30	10	0.121	0.099	0.559	0.024	0.094
	50	0.126	0.092	0.891	0.003	0.067
	100	0.118	0.092	0.946	-0.011	0.056
50	10	0.145	0.065	0.583	-0.028	2.761
	50	0.137	0.062	0.891	0.011	0.029
	100	0.138	0.068	0.937	-0.017	0.039

Notes: $\lambda_j \sim \text{unif}[0,1]$, $\pi(\lambda) = 0.4 + 0.2 \cdot \lambda$, $\mu(\lambda) = -1 + 2 \cdot \lambda$ and $\sigma(\lambda) = 1$. $\mathbf{E}\left[R^2\right]$ denotes the average R^2 of regressing λ_j on $\hat{\lambda}_j$.

Table 2—Finite Types of Clusters

J	N	mean diff. bias	mean diff. MSE	K-means Pr {perfect class.}	K-means bias	K-means MSE
	10	0.206	0.112	0.126	0.056	0.057
30	50	0.210	0.103	0.990	0.013	0.035
	100	0.188	0.100	1.000	-0.003	0.039
	10	0.204	0.082	0.032	0.048	0.030
50	50	0.201	0.079	0.986	-0.003	0.023
	100	0.213	0.083	1.000	0.013	0.021

Notes: $\Pr\{\lambda_j = 1\} = \Pr\{\lambda_j = 2\} = \frac{1}{2}$, $\pi(\lambda) = 0.2 + 0.2 \cdot \lambda$, $\mu(\lambda) = -1.5 + \lambda$ and $\sigma(\lambda) = 1$.

proves and thus the bias decreases as N increases. Lastly, Table 3 contains the simulation result for the third DGP. The functional PCA estimator seems to outperform the K-means estimator thanks to its flexibility when both J and N are larger. However, the functional PCA estimator suffers from high variability when J and N are smaller.

Table 3—Nonlinearity in Cluster-level Distributions

J	N	mean diff. bias	mean diff. MSE	functional PCA bias	functional PCA MSE	K-means bias	K-means MSE
	10	0.134	0.100	0.069	0.088	0.074	0.072
30	50	0.156	0.096	0.039	1.148	0.055	0.045
	100	0.114	0.096	-0.001	0.301	0.016	0.056
	10	0.113	0.057	0.043	0.041	0.064	0.038
50	50	0.135	0.063	0.029	0.039	0.048	0.031
	100	0.131	0.063	-0.008	0.031	0.035	0.030

Notes: $\lambda_j \sim \text{unif}[0,1]$, $\pi(\lambda) = 0.4 + 0.2 \cdot \lambda$, $\mu(\lambda) = -1 + 2 \cdot \lambda$ and $\sigma(\lambda) = 1 + \lambda$.

5 Empirical Illustration

As an empirical illustration of the "distribution-as-control" approach, I revisit the question of whether an increase in the state-level minimum wage leads to a decrease in teen employment rate in the United States, using the Current Population Survey (CPS) from 2000-2021. To control for the state-level heterogeneity with regard to state-level labor market equilibria, I use two individual-level variables: $EmpHistory_{ijt}$ and $WageInc_{ijt}$. $EmpHistory_{ijt}$ is a discrete, monthly variable that concatenates the last four months' employment status variables with three categories: employed, unemployed, and not-in-the-labor-force. The K-means estimator is applied to the distribution of the $EmpHistory_{ijt}$, with $\rho_{Kmeans} = 3$. $WageInc_{ijt}$ is a continuous wage income variable recorded annually: March Annual Social and Economic Supplement (ASEC). $WageInc_{ijt}$ is truncated at zero, since $EmpHistory_{ijt}$ already contains information on unemployment, and then logged. The functional PCA is applied to the truncated $\log WageInc_{ijt}$, with $\rho_{fPCA} = 2$.

To estimate the disemployment effect of minimum wage, I use a regression model developed in accordance with practices in the minimum wage literature: see Allegretto et al. (2011); Neumark et al. (2014); Allegretto et al. (2017) for more.

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

Y_{ijt} is the binary employment status variable for teenager i in state j at month t. X_{ijt} is the socioeconomic covariates of age, race, sex, marital status and education. $EmpRate_{jt}$ is the state-level employment rate. The state-level heterogeneity is controlled in two different ways; firstly with state fixed-effect α_j and

TABLE 4—DISEMPLOYMENT EFFECT ESTIMATES ACROSS SPECIFICATIONS

$\hat{\beta}$	(1)	(2)	(3)	(4)	(5)	(6)
	-0.109*	-0.069	-0.052	-0.029*	-0.030*	-0.030*
	(0.061)	(0.060)	(0.078)	(0.017)	(0.017)	(0.017)
time FE	X	X	X	O	O	O
$EmpHistory$	X	O	O	X	O	O
$WageInc$	X	X	O	X	X	O
T	1 (January 2007)			264 (2000-2021)		

Notes: The categorical latent factors from the distribution of $EmpHistory_{ijt}$ are given time-varying loadings while the continuous latent factors from the distribution of $WageInc_{ijt}$ are given time-invariant loadings:

$$\lambda_{jt}{}^{\intercal}\delta_t = \lambda_{jt,EmpHistory}{}^{\intercal}\delta_{t,EmpHistory} + \lambda_{jt,WageInc}{}^{\intercal}\delta_{WageInc}.$$

The standard errors are clustered at the state level.

* denotes p-value<0.1; ** denotes p-value<0.05; *** denotes p-value<0.01.

secondly with the state-level latent factor λ_{jt} that I estimate from the distributions of $EmpHistory_{ijt}$ and $WageInc_{ijt}$: $\lambda_{jt} = \left(\lambda_{jt,EmpHistory}{}^{\intercal}, \lambda_{jt,WageInc}{}^{\intercal}\right)^{\intercal}$.

Table 4 compares the estimate on β across different specifications; particularly, I compare a cross-section regression that only uses observations from January 2007, when the most number of states raised their minimum wage level, with a pooled regression that uses all of the 22 years. Table 4 suggests that the rich information at the individual-level controls for the cluster-level heterogeneity in a cross-section, to a similar magnitude with the flexible two-way fixed-effect specification relying on the time dimension.

Table 5 extends the regression specification (10) and explores aggregate-level and individual-level heterogeneity in the disemployment effect of minimum wage, by letting β depend on Age_{ij} and $\lambda_{jt,EmpHistory}$. The left panel shows how the minimum wage affects older teenagers and younger teenagers differently; the disemployment effect is mostly coming from younger teenagers.

TABLE 5—INTERACTION BETWEEN AGE AND STATE LABOR MARKET

$\hat{\beta}$	(1)	(2)	(3)	(4)
$\{Age \leq 18\}$	-0.038^{**}	-0.038^{**}		
	(0.017)	(0.017)		
$\times\{\lambda_{EmpHistory} = e_1\}$			-0.035^{**}	-0.035^{**}
			(0.017)	(0.017)
$\times\{\lambda_{EmpHistory} = e_2\}$			-0.036^{**}	-0.036^{**}
			(0.018)	(0.018)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.047^{**}	-0.048^{**}
			(0.024)	(0.023)
$\{Age = 19\}$	-0.004	-0.004		
	(0.019)	(0.019)		
$\times\{\lambda_{EmpHistory} = e_1\}$			-0.001	-0.001
			(0.019)	(0.020)
$\times\{\lambda_{EmpHistory} = e_2\}$			-0.002	-0.002
			(0.018)	(0.019)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.024	-0.024
			(0.026)	(0.026)
$EmpHistory$	O	O	O	O
$WageInc$	X	O	X	O

Notes: The categorical latent factors from the distribution of $EmpHistory_{ijt}$ are given time-varying loadings while the continuous latent factors from the distribution of $WageInc_{ijt}$ are given time-invariant loadings:

$$\lambda_{jt}{}^{\intercal}\delta_t = \lambda_{jt,EmpHistory}{}^{\intercal}\delta_{t,EmpHistory} + \lambda_{jt,WageInc}{}^{\intercal}\delta_{WageInc}.$$

Across different months, $\lambda_{jt,EmpHistory}$ are ordered in a way that $\lambda_{jt,EmpHistory} = e_1$ indicates states with higher employment and higher labor force participation while $\lambda_{jt,EmpHistory} = e_3$ indicates states with lower employment and lower labor force participation.

The standard errors are clustered at the state level.

* denotes p-value<0.1; ** denotes p-value<0.05; *** denotes p-value<0.01.

The right panel interacts Age_{ijt} with the categorical latent factor $\lambda_{jt,EmpHistory}$; it discusses how the aggregate-level heterogeneity from the state-level labor market interacts with the individual-level heterogeneity from age. Across the two age groups, the disemployment effect is stronger in Group 3 states, where both the employment rate and the labor force participation rate are higher.

6 Conclusion

This paper motivates the use of the cluster-level distribution of individual-level control covariates as a control for the cluster-level heterogeneity in a clustered data. This framework is most relevant when the clusters are large, so that the estimation errors on the cluster-level distributions are negligible. By explicitly controlling for the distribution of individuals, two different dimensions of heterogeneity in data are modeled, being true to the hierarchical nature of the dataset: individual heterogeneity and aggregate heterogeneity.

To implement the idea of "distribution-as-control," the functional PCA and the K-means algorithm are used in this paper. The two approaches complement each other; one attains consistency under slower growth rate of cluster size while the other allows for continuous cluster-level heterogeneity. Based on empirical contexts, a yet another dimension reduction method on distributions may be more suitable, calling for follow-up research that discusses alternative functional analysis methods. Also, this paper mostly focuses on cross-section data. In Section 5, the cluster-level latent factor are assumed to be strictly exogenous. A natural direction for future research is to extend this and study a panel data model with clustering structure where the time-varying distribution of individuals for each cluster is modeled to be a dynamic process.

References

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

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

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

Allegretto, Sylvia A, Arindrajit Dube, and Michael Reich, "Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data," *Industrial Relations: A Journal of Economy and Society*, 2011, *50* (2), 205–240.

Allegretto, Sylvia, Arindrajit Dube, Michael Reich, and Ben Zipperer, "Credible research designs for minimum wage studies: A response to Neumark, Salas, and Wascher," *ILR Review*, 2017, *70* (3), 559–592.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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