Aggregating Individual-level Information in Large Clusters*

Myungkou Shin†

October 23, 2024

Click here for the latest version.

Abstract

When given a possibly endogenous cluster-level explanatory variable in a clustered dataset, we cannot model the cluster-level heterogeneity to be fully flexible due to multicollinearity. To model 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, fully utilizing the available information at the individual level. Thanks to the low-dimensionality of the latent factor model, 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 three regression equation models:

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

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

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

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. $\mathbf{F}_j$ denotes the distribution of X_{ij} for cluster j. The first regression equation (1) models the cluster-level heterogeneity in a fully flexible manner, with cluster fixed-effect α_j; the coefficient β is not identified in model (1) due to multicollinearity. In light of this problem, the researcher may want to use the second regression equation (2) assuming that there is no unobserved heterogeneity across clusters. However, when the cluster heterogeneity correlates with Z_j, the OLS estimator for β will be biased.[1] Thus, I propose an alternative regression equation (3), where we do not assume cluster homogeneity, but put restrictions on the cluster heterogeneity by modeling it with the cluster-level distribution $\mathbf{F}_j$. This approach fully utilizes the rich information observed at the individual level. 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

[1] This problem closely relates to the treatment endogeneity problem; when α_j is uncorrelated with Z_j, the second regression equation (2) also identifies β.

how the individual-level characteristics collectively affects the individual-level outcomes.

In this paper, the *distribution* function is used a choice of the aggregation method. The dimension reduction property of using the distribution is particularly helpful when the clusters are large. Suppose that the clusters are small. Then, we can use more flexible methods of aggregation than distribution to model the cluster-level heterogeneity, since the simple unordered cluster-level aggregation $\{X_{ij}\}_i$ is low-dimensional. On the contrary, 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. Thus, there is no information loss from using a distribution function instead of an unordered collection, while there is gain of dimension reduction.

The formal econometric framework of this paper consists of two parts: two constructive latent factor models for the cluster-level distributions, and a moment restriction model for a parameter of interest. In the latent factor models, the cluster-level distribution function of the individual-level control covariates is modeled to be a function of a finite-dimensional cluster-level latent factor $\lambda_j \in \mathbb{R}^\rho$: the second layer of dimension reduction. By assuming a bijection between the latent factor and the distribution function, a large class of moment restriction models that take finite-dimensional vectors as inputs can be used to develop a model where a parameter of interest is identified in relation to the cluster-level distributions.

The estimation is done in a plug-in manner; the latent factors are estimated from the cluster-level distributions and the estimates are used in the moment restriction model to estimate the parameter of interest. The key assumption that makes the plug-in procedure valid is that the moment restriction model is invariant to a rotation on the latent factor while the latent factor model provides an estimator that estimates the latent factor up to some rotation. The rotation invariance is helpful since it allows us to use machine learning methods that have desirable properties such as dimension reduction but do not provide interpretable outputs. As long as the outputs are a rotation of an interpretable factor, we can motivate a latent factor model based on the machine learning method. Using this feature, two latent factor models for the cluster-level distributions are proposed in this paper. The first of the two models relates to the K-means clustering algorithm and the second relates to the functional principal component analysis (PCA). Under their respective latent factor models and given appropriate relative growth rate of the cluster size to the number of clusters, both estimators have fast enough estimation error so that the plug-in estimator using the estimates is consistent and asymptotically normal.

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 (Z_j, X_{ij}) after conditioning on the cluster-level distributions. The parameter of interest in my model, $\alpha(\mathbf{F}_j)$ 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

[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.

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; in this sense, Xu (2017) is closer to this paper.

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 the parameter of interest. In Section 3, I discuss two latent factor models, which motivate the use of the K-means clustering algorithm and the functional PCA in estimating the cluster-level latent factor. In Sections 4-5, simulation results and an empirical illustration of the econometric framework that discuss 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

[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

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 $j = 1, \cdots, J$,

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

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

$$\big(Y_{ij}, X_{ij}\big) \;\big|\; \{Z_k, N_k, \lambda_k\}_{k=1}^{J} \overset{iid}{\sim} H(Z_j, N_j, \lambda_j; \xi) \tag{4}$$

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 (4) 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

[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

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 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,

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.

[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.

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 $\frac{\partial}{\partial x}\left(\mathbf{E}[Y_{ij}|(X_{ij}, \mathbf{F}_j) = (x, \mathbf{F})] - \mathbf{E}[Y_{ij}|(X_{ij}, \mathbf{F}_j) = (x, \mathbf{F}')]\right)$.

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 we want to construct a regression model where a binary outcome Y_{ij} depends on a distribution function $\mathbf{F}_j$, we can directly use the known results on the logistic model with finite-dimensional control covariates, by substituting λ_j for $\mathbf{F}_j$.

Given the clustered data model (4) 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,

$$\mathbf{E}\left[m(W_j^*;\theta)\right] = 0. \tag{5}$$

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 (5) 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:

$$\left(Y_{ij}(1), Y_{ij}(0), X_{ij}\right) \perp\!\!\!\perp Z_j \;\middle|\; (N_j, \lambda_j)\,. \tag{6}$$

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 (6) 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 (6) 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:

$$\begin{aligned}Y_{ij} &= {X_{ij}}^\intercal\theta_1+{Z_j}^\intercal\theta_2+{\lambda_j}^\intercal\theta_3+U_{ij}, \qquad (7)\\ 0 &= \mathbf{E}\left[U_{ij}|X_{ij},Z_j,N_j,\lambda_j\right].\end{aligned}$$

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 (5) 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 (5) 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 (5). 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 ${G_A}^{-1}$,

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

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

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

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

[7]The second equality holds since

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

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.

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 K-means clustering and the functional PCA 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 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 $\mathbf{F}_j$. However, we do not directly observe $\mathbf{F}_j$. Thus, as an estimator for $\mathbf{F}_j$, I use the empirical distribution function $\hat{\mathbf{F}}_j$: for all $x \in \mathbb{R}^p$,

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

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 $\hat{\mathbf{F}}_j$:

$$\left(\hat{\lambda}_1, \cdots, \hat{\lambda}_J, \hat{G}(1), \cdots, \hat{G}(\rho)\right) = \arg\min_{\lambda, G} \sum_{j=1}^{J} \left\|\hat{\mathbf{F}}_j - G(\lambda_j)\right\|_{w,2}^2. \tag{8}$$

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 (8), 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 $\mathbb{R}^\rho$,

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

[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.

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 (8) 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 the rate discussed in Theorems 1-2. Assumption 4 discusses a set of conditions for that.

Assumption 4. (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 4.a ensures that we observe positive measure of clusters for each value of the latent factor as J goes to infinity. Under Assumption 4.b, clusters with different values of the latent factor will be distinct from each other in terms of their distributions of X_{ij}. Thus, the K-means algorithm that uses $\hat{\mathbf{F}}_j$ is able to tell apart clusters with different values of λ_j, when $\hat{\mathbf{F}}_j$ is a consistent estimator for $\mathbf{F}_j$. Assumption 4.c assumes that the size of clusters goes to infinity as the number of clusters goes to infinity. This assumption limits our

attention to cases where clusters are large. It should be noted that Assumption 4.c excludes cases where the size of cluster increases only for some clusters and is fixed for some other clusters; the estimation of $\hat{\mathbf{F}}_j$ jointly improves as J increases.

The key element of the econometric model described in Assumption 4 is that there are finite types of clusters, in terms of their distribution of individual-level control covariates X_{ij}. Thus, using Assumption 4 to model the cluster-level heterogeneity would make the most sense when we expect that the heterogeneity across clusters are discrete and finite.

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

Proposition 1. *Suppose that 4 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,

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

Proposition 1 shows that the misclassification probability of the K-means algorithm grouping clusters with different values of λ_j together goes to zero when $J/{N_{\min}}^{\nu^*}$ goes to zero for some $\nu^* > 0$. When the misclassification probability converges to zero, the estimation error $\left(\sum_{j=1}^{J}\|\hat{\lambda}_j - A\lambda_j\|_2{}^2\right)^{\frac{1}{2}}$ is $o_p(a_n)$ for any sequence $\{a_n\}_{n=1}^{\infty}$ since for any $\varepsilon > 0$ the probability $\Pr\{a_n(\sum_{j=1}^{J}\|\hat{\lambda}_j - A\Lambda_j\|_2^2)^{\frac{1}{2}} > \varepsilon\}$ is bounded by the misclassification probability. The rate on

the misclassification probability is the same rate found in the literature: e.g., Bonhomme and Manresa (2015).

Under Assumption 4, we can apply the K-means clustering estimator for the latent factor to a variety of models with a grouping structure. For example, Example 1 in the previous section would be a clustered treatment model with latent group-specific propensity score. Example 2 in the previous section would be a group fixed-effect regression model.

Though the K-means clustering estimator has desirable qualities such as being concise and having a fast estimation rate, the finite support assumption can be too restrictive, depending on contexts. Thus, in the next subsection, I propose an alternative framework where the cluster-level heterogeneity λ_j is assumed to be continuous, using the functional PCA.

3.2 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. Again, 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 $\hat{q}_r = \left(\hat{q}_{r1}, \cdots, \hat{q}_{rJ}\right)^\intercal$ 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 5. (Finite Types of Individuals) *With some $C > 0$,*

a. *(finite mixture model for distribution)*

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

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 5.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 this sense, we can make a following comparison between the two latent factor models proposed in this paper: Assumption 4 for the K-means clustering algorithm assumes that there are finite types of *clusters*, while Assumption 5 for the functional PCA assumes that there are finite types of *individuals* across clusters.

In addition to motivating the finite mixture model for density, Assumption 5.a assumes that the underlying density functions $g_1, \cdots, g_\rho$ are smooth and bounded, up to their third derivatives. Under Assumption 5.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 5.b assumes that the underlying density functions $g_1, \cdots, g_\rho$ have sufficient variation, when measured with $\langle \cdot, \cdot \rangle_w$.

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

Proposition 2. *Suppose that Assumption 5 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 2 bounds the estimation error rate with the square root of the relative growth rate. Recall that $N_{\min} \propto J$ is a sufficiently fast growth rate in the case of the K-means clustering for the hypertheorems. However, for the functional PCA, $N_{\min} \propto J$ is not fast enough for the hypertheorems. Allowing for continuous cluster-level heterogeneity comes at the cost of requiring a faster growth rate on the cluster size.

4 Simulation

To discuss finite-sample performance of the K-means estimator and the functional PCA 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

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

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 clusters, satisfying Assumption 4. I apply the K-means estimator to the first DGP. The second DGP admits two types of individuals, satisfying Assumption 5. I apply the functional PCA 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 K-means estimator and the functional PCA estimator to the third DGP. The specifics of the DGPs are discussed in 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 K-means estimated factors, I simply take the sample mean: $\hat{\pi}(\lambda) = \frac{\sum_{j=1}^{J} D_j \mathbf{1}\{\hat{\lambda}_j = \lambda\}}{\sum_{j=1}^{J} \mathbf{1}\{\hat{\lambda}_j = \lambda\}}$. For the functional PCA estimated factors, I run OLS: $\hat{\pi}(\lambda) = \hat{\lambda}^{\intercal}\hat{\pi}$. 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. From the third and the fourth columns, we can see that the classification improves and thus the bias decreases as N increases. Table 2 contains the simulation results for the second 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. 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.

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 con-

tains 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.

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

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 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. The right panel of Table 4 suggests that the two-way fixed-effects sufficiently control for the state-level labor market heterogeneity while the left panel shows that distributional control matters when there is no time dimension to be used.

Table 5 extends the regression specification (9) 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. 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 K-means algorithm are the functional PCA 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 discuss 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.

Appendix

A Tables

Table 1—Finite Types of Clusters

		mean diff.		K-means		
J	N	bias	MSE	Pr {perfect class.}	bias	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 $\sigma(\lambda) = 1$.

Table 2—Finite Types of Individuals

		mean diff.		fPCA		
J	N	bias	MSE	$\mathbf{E}[R^2]$	bias	MSE
	10	0.121	0.099	0.559	0.024	0.094
30	50	0.126	0.092	0.891	0.003	0.067
	100	0.118	0.092	0.946	-0.011	0.056
	10	0.145	0.065	0.583	-0.028	2.761
50	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 $\sigma(\lambda) = 1$. $\mathbf{E}[R^2]$ denotes the average R^2 of regressing λ_j on $\hat{\lambda}_j$.

TABLE 3—NONLINEARITY IN CLUSTER-LEVEL DISTRIBUTIONS

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

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

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)	(0.079)	(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.

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^{**}	-0.037^{**}
			(0.018)	(0.018)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.047^{**}	-0.047^{**}
			(0.023)	(0.023)
$\{Age = 19\}$	-0.004	-0.004		
	(0.019)	(0.019)		
$\times\{\lambda_{EmpHistory} = e_1\}$			-0.001	-0.001
			(0.020)	(0.020)
$\times\{\lambda_{EmpHistory} = e_2\}$			-0.002	-0.002
			(0.018)	(0.019)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.023	-0.023
			(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.

Supplementary Appendix

Myungkou Shin*

October 23, 2024

A Exchangeability

Assumption 1 assumes that the cluster-level distribution contains sufficient information on the cluster heterogeneity λ_j. To motivate this assumption, let us consider a simple binary treatment model $Z_j \in \{0, 1\}$. When we consider a population distribution with a fixed number of individual per cluster and random sampling, Assumption 1 is a direct result of selection-on-observable and exchangeability. Let N_j^* denote the population number of individuals per cluster. N_j out of N_j^* individuals are randomly sampled. The observed dataset is

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

where $Y_{ij} = Y_{ij}(1) \cdot Z_j + Y_{ij}(0) \cdot (1 - Z_j)$ and the underlying population is

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

Clusters are independent of each other. Assume the following three assumptions:

(*random sampling*) *There is a random injective function* $\sigma_j : \{1, \cdots, N_j\} \to \{1, \cdots, N_j^*\}$

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

such that

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

σ_j is independent of $\left(\left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N_j*}, Z_j\right)$. Also, for any distinct $\left(i_1, \cdots, i_{N_j}\right)$

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

(*unconfoundedness*)

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

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

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

Note that the *exchangeability* assumption restricts dependence structure within a given cluster in a way that the labelling of individuals should not matter. However, it still allows individual-level outcomes within a cluster to be arbitrarily correlated after conditioning on control covariates: for example, when X_{ij} includes a location variable, individuals close to each other is allowed to be more correlated than individuals further away from each other. Proposition A.1 follows immediately.

Proposition A.1. *Under random sampling, unconfoundedness and exchangeability,*

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

where $\mathbf{F}_j(x) = \frac{1}{N_j^*}\sum_{i=1}^{N_j^*} \mathbf{1}\{X^*_{ij} \leq x\}$.

Proof. Firstly, find that $\mathbf{E}[Z_j|\mathbf{F}_j]$ is an weighted average of $\mathbf{E}[Z_j|X^*_{\sigma^*(1)j}, \cdots, X^*_{\sigma^*(N_J)j}]$ across

all possible permutations σ^*. Thus, under the *exchangeability*,

$$\mathbf{E}[Z_j|\mathbf{F}_j] = \mathbf{E}[Z_j|X^*_{1j}, \cdots, X^*_{N_j j}] = \mathbf{E}[Z_j|X^*_{\sigma^*(1)j}, \cdots, X^*_{\sigma^*(N_j)j}]$$

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

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

The first equality holds since $\mathbf{F}_j$ is a function of $\{X^*_{ij}\}_{i=1}^{N^*_j}$ and $\{Y_{ij}(1), Y_{ij}(0)\}_{i=1}^{N_j}$ is a function of $\{Y_{ij}(1)^*, Y_{ij}(0)^*\}_{i=1}^{N^*_j}$ and σ_j. The second equality holds since *random sampling* implies that Z_j is independent of σ_j given $\left\{Y_{ij}(1)^*, Y_{ij}(0)^*, X^*_{ij}\right\}_{i=1}^{N^*_j}$. The third equality is from *unconfoundedness*. □

Proposition A.1 suggests propensity score matching based on $\mathbf{F}_j$, the population distribution function of X_{ij} for cluster j. In this example, the population distribution is assumed to be discrete to explicitly invoke the exchangeability condition. Assumption 1 extends on this idea and assumes that the population distribution is possibly continuous and can be written as a function of a latent low-dimensional factor λ_j, which controls for the cluster-level heterogeneity, as does the propensity score $\pi(\mathbf{F}_j)$ in this example.

B Additional discussion on empirical illustration

B.1 Background

There exists a unique opportunity in research design when studying the question of whether an increase in minimum wage level leads to higher unemployment rate in the United States: the state-level variation in minimum wage. In the United States, each state has their own minimum wage level in addition to the federal minimum wage level and thus we see states with different minimum wage levels for the same time period. The state-level policy variation is helpful since it allows us to control for time heterogeneity in a flexible way, by comparing contemporaneous outcomes across states.

However, there could still be spatial heterogeneity that affects both minimum wage level and employment at the state level, which complicates the causal interpretation of a minimum wage regression. The literature has suggested several remedies for this spatial heterogeneity problem. For example, difference-in-differences (DiD) compares over-the-time difference in employment rate across states, assuming that spatial heterogeneity only exists as state heterogeneity and the state heterogeneity is cancelled out by taking the over-the-time difference (Card and Krueger, 1994). Some researchers limit their scope of analysis to counties that are located near the state border to account for spatial heterogeneity (Dube et al., 2010). Some use a more relaxed functional form assumption on state heterogeneity than DiD, such as state-specific linear trends (Allegretto et al., 2011, 2017). Some have the data construct a synthetic state that is comparable to an observed state (Neumark et al., 2014).

The clustered data setup in the paper fits the empirical context of the US minimum wage application well. Firstly, employment status, the outcome of interest, is observed at the individual level while the minimum wage level, the regressor of interest, is observed at the state level, i.e. the dataset is hierarchical. Secondly, an assumption that is shared in the minimum wage literature as a common denominator is that there is no dependence across states. In other words, it is believed that the decision of whether and how much the state

minimum wage level changes is only determined by what happens within the state. This corresponds to the clusters being independent.

Building on this observation, I apply the results of Sections 2 and 3 in the main text to control for the spatial heterogeneity in estimating the disemployment effect of the minimum wage. The key assumption in doing so is that the state-level distribution of individual-level demographic and socioeconomic characteristics sufficiently controls for the spatial heterogeneity. If the information that state legislators look at when deciding their state's minimum wage level is completely incorporated in the state-level distribution, the assumption would naturally hold. This 'distribution-as-control' approach is complementary to assuming that there exists some unrestricted and time-invariant state-level heterogeneity as in the two-way fixed-effect specification in the DiD literature. In the 'distribution-as-control' approach, the state-level heterogeneity is allowed to vary over time, but restricted in the sense that it is a function of the (near-)observable state-level distribution of individual-level characteristics.

B.2 Estimation

Following Allegretto et al. (2011); Neumark et al. (2014); Allegretto et al. (2017), I focus on the teen employment since it is likely that teenagers work at jobs that pay near the minimum wage level compared to adults, thus being more responsive to a change in the minimum wage level. I constructed a dataset by pooling the Current Population Survey (CPS) data from 2000 to 2021, collecting the same demographic control covariates on teenagers as Allegretto et al. (2011), and additional control covariates on all individuals. The additional variables were collected for every individual to construct state-level distributions, since information only from teenagers may not accurately reflect the state-level labor market status. Let $\mathcal{I}_{jt}$ denote the set of teens in state j at time t and $\tilde{\mathcal{I}}_{jt}$ denote the set of all individuals in state j at time t, from the CPS: $\mathcal{I}_{jt} \subset \tilde{\mathcal{I}}_{jt}$. Since the CPS is collected every month, the dataset contains $264 = 12 \cdot 22$ time periods in total.

The main regression specification I use is motivated from Allegretto et al. (2011). As

one of the two main regression specifications, Allegretto et al. (2011) estimates the following linear model: for teen i in state j at time t,

$$Y_{ijt} = \alpha_j + \delta_{cd(j)t} + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{1}$$

There are two noteworthy observations to be made here. Firstly, the regressor of interest $MinWage_{jt}$ varies on the state-by-time level, making state-by-time fixed-effects infeasible. This is exactly the same type of multicollinearity problem discussed in Section 2 of the main text. When treatment is assigned at the cluster level, treatment effects cannot be identified under a model with fully flexibly cluster heterogeneity. Thus, Allegretto et al. (2011) uses census-division-by-time fixed-effects by grouping 50 states and Washington D.C. into 9 census divisions: $\delta_{cd(j)t}$. Secondly, Equation (1) already implements the idea of aggregating individual-level information: the state-by-time employment rate $EmpRate_{jt}$ computed from Y_{ijt}. In using $EmpRate_{jt}$, a conscious choice was made by the researcher to use the mean to summarize the individual-level information for each state.

In this paper, I build upon the two observations above and develop a more flexible regression model:

$$Y_{ijt} = \alpha_j + \lambda_{jt}^{\intercal}\delta_t + \beta \log MinWage_{jt} + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \tag{2}$$

In the regression model, λ_{jt} is a time-varying state-level latent factor that I assume to be one-to-one with state-level distributions of individual-level characteristics. Specifically, I use the following two variables: $EmpHistory_{ijt}$ and $WageInc_{ijt}$. By using λ_{jt} as a control, I implement the 'distribution-as-control' approach. The latent factor λ_{jt} allows us to control for the spatial heterogeneity while not subsuming the variation in $MinWage_{jt}$. In doing so, λ_{jt} summarizes the available information at the individual level, in a more flexible way than the simple mean as in $EmpRate_{jt}$. In the next two paragraphs, I provide more detail on how I estimate the latent factor λ_{jt}, from the two state-level distributions.

Firstly, I apply the K-means clustering algorithm to the distribution of $EmpHistory_{ijt}$, an individual-level employment history variable:

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

$EmpStatus_{ijt}$ is an employment status variable for individual i in state j at time t. It is a categorical variable with three possible values: being employed, being unemployed, and not being in the labor force. $EmpHistory_{ijt}$ concatenates $EmpStatus_{ij\tau}$ for $\tau = t-4, \cdots, t-1$; $EmpHistory_{ijt}$ is a four-month-long history of employment status. Since $EmpStatus_{ijt}$ is a categorical variable with a finite support of three elements, $EmpHistory_{ijt}$ has a finite support of 81 elements. Note that $Y_{ijt} = 1 \Leftrightarrow EmpStatus_{ijt} = Emp$ and thus $EmpHistory_{ijt}$ can be understood as a vector of lagged outcome variables, but defined for both teenagers and adults. To aggregate the information from $EmpHistory_{ijt}$ to learn about the labor market fundamental of a given state, I compute the empirical distribution function: for $x \in \mathcal{X}$,

$$\hat{\mathbf{F}}_{jt}(x) = \frac{1}{\sum_{i \in \tilde{\mathcal{I}}_{jt}} \tilde{\omega}_i} \sum_{i \in \tilde{\mathcal{I}}_{jt}} \mathbf{1}\{EmpHistory_{ijt} = x\}\tilde{\omega}_i$$

$\{\tilde{\omega}_i\}_i$ are the longitudinal weights provided by the IPUMS-CPS to construct a four-month-long panel using the CPS sample. Note that $\tilde{\mathcal{I}}_{jt}$ is used instead of $\mathcal{I}_{jt}$; information from adults' employment history is included. When evaluating the distance between states measured in terms of $\hat{\mathbf{F}}_{jt}$, I use the uniform weighting function since $\mathcal{X}$ is a finite set. By applying the K-means algorithm to $\left\{\hat{\mathbf{F}}_{jt}\right\}_{j=1}^{J}$, I get $\{\hat{\lambda}_{jt,EmpHistory}\}_{j=1}^{J}$.

Secondly, I apply the functional PCA to the distribution of $WageInc_{ijt}$. $WageInc_{ijt}$ is a wage income variable for individual i in state j at time t. Since the current and past unemployment rates are already controlled with $EmpRate_{jt}$ and the distribution of $EmpHistory_{ijt}$, I consider a truncated distribution of $WageInc_{ijt}$ by focusing on individuals

whose wage income is strictly positive. The wage income variable comes from the March Annual Social and Economic Supplement (ASEC). The ASEC sample is collected only once a year in March and is different from the basic monthly CPS sample. Let $\breve{\mathcal{I}}_{jt}$ denote the set of all individuals with positive wage income in state j, from the most recent ASEC sample at time t. Then, $\breve{\mathcal{I}}_{jt} = \breve{\mathcal{I}}_{jt+1}$ except when t corresponds to a month of March and $\breve{\mathcal{I}}_{jt} \neq \tilde{\mathcal{I}}_{jt}$ in general. To aggregate the information from $WageInc_{ijt}$, I compute the product of the state-level conditional densities of $\log WageInc_{ijt}$. The j-th row and k-th column component of the estimated product matrix $\hat{M}_t$ is

$$\hat{M}_{jkt} = \begin{cases} \frac{\sum_{i\in\breve{\mathcal{I}}_{jt}, i'\in\breve{\mathcal{I}}_{kt}} \breve{\omega}_i\breve{\omega}_{i'}}{\sum_{i\in\breve{\mathcal{I}}_{jt}, i'\in\breve{\mathcal{I}}_{kt}} \breve{\omega}_i\breve{\omega}_{i'}} \int_{\mathbb{R}} \frac{\breve{\omega}_i\breve{\omega}_{i'}}{h^2} K\left(\frac{x-\log WageInc_{ijt}}{h}\right) \cdot K\left(\frac{x-\log WageInc_{i'kt}}{h}\right) w(x)dx, & \text{if } j \neq k \\ \frac{\sum_{i,i'\in\breve{\mathcal{I}}_{jt}, i\neq i'} \breve{\omega}_i\breve{\omega}_{i'}}{\sum_{i,i'\in\breve{\mathcal{I}}_{jt}, i\neq i'} \breve{\omega}_i\breve{\omega}_{i'}} \int_{\mathbb{R}} \frac{\breve{\omega}_i\breve{\omega}_{i'}}{h^2} K\left(\frac{x-\log WageInc_{ijt}}{h}\right) \cdot K\left(\frac{x-\log WageInc_{i'jt}}{h}\right) w(x)dx, & \text{if } j = k \end{cases}.$$

$\{\breve{\omega}_i\}_i$ are the cross-sectional weights provided by the IPUMS-CPS to construct a cross-section with the ASEC sample. For the weighting function w, I use the uniform weighting on [0,15]: $w(x) = \frac{1}{1001}\mathbf{1}\{x \in \{0, 15/1000, \cdots, 15\}\}$. By applying the eigenvalue decomposition to $\hat{M}_t$, I get $\{\hat{\lambda}_{jt,WageInc}\}_{j=1}^J$. An estimate for the entire latent factor λ_{jt} is obtained from concatenating $\hat{\lambda}_{jt,EmpHistory}$ and $\hat{\lambda}_{jt,WageInc}$.

B.2.1 Cross-validation on the dimension of the latent factor

Both of the latent factor estimation methodologies introduced in the paper involve an unknown parameter: ρ, the dimension of the latent factor. To decide on ρ, I conduct a 5-fold cross-validation exercise for a given time t.

1. Fix ρ and randomly split the individual indices for a given state into five subsets: $\tilde{\mathcal{I}}_{jt} = \cup_{k=1}^5 \tilde{\mathcal{I}}_{jt,k}$ and $\breve{\mathcal{I}}_{jt} = \cup_{k=1}^5 \breve{\mathcal{I}}_{jt,k}$, respectively for $EmpHistory_{ijt}$ and $WageInc_{ijt}$. For each k, define the train sets $\left\{\tilde{\mathcal{I}}_{jt,-k} = \tilde{\mathcal{I}}_{jt} \setminus \tilde{\mathcal{I}}_{jt,k}\right\}_{j=1}^J$ and $\left\{\breve{\mathcal{I}}_{jt,-k} = \breve{\mathcal{I}}_{jt} \setminus \breve{\mathcal{I}}_{jt,k}\right\}_{j=1}^J$.

2. For each k, construct $\left\{\hat{\mathbf{F}}_{jt,-k}(x)\right\}_{j=1}^J$ and $\hat{M}_{t,-k}$ from their respective train sets and estimate $\lambda_{jt,EmpHistory}$ and $\lambda_{jt,WageInc}$ with the predetermined value of ρ.

3. Evaluate the out-of-sample performance of the estimated models from Step 2, using the test sets. For each k, construct $\left\{\hat{\mathbf{F}}_{jt,k}\right\}_{j=1}^{J}$ and $\hat{M}_{t,k}$ with their respective test sets $\left\{\tilde{\mathcal{I}}_{jt,k}\right\}_{j=1}^{J}$ and $\left\{\breve{\mathcal{I}}_{jt,k}\right\}_{j=1}^{J}$ and let

$$\text{SSFE}_{t,EmpHistory}(\rho) = \frac{1}{5}\sum_{k=1}^{5}\sum_{j=1}^{J}\sum_{x\in\mathcal{X}}\left(\hat{\mathbf{F}}_{jt,k}(x) - \hat{G}_{-k,EmpHistory}\left(\hat{\lambda}_{jt,-k,EmpHistory}\right)\right)^2,$$

$$\text{SSFE}_{t,WageInc}(\rho) = \frac{1}{5}\sum_{k=1}^{5}\left\|\hat{M}_{t,k} - \tilde{M}_{t,-k}\right\|_F^2.$$

$\hat{G}_{-k,EmpHistory}\left(\hat{\lambda}_{jt,-k,EmpHistory}\right)$ is the fitted value of the empirical distribution function $\mathbf{F}_{jt}$ from applying the K-means algorithm with ρ groups to $\left\{\hat{\mathbf{F}}_{jt,-k}\right\}_{j=1}^{J}$ from the train set. $\tilde{M}_{t,-k}$ is the fitted value of the product matrix M from applying the eigenvalue decomposition to the estimated product matrix $\hat{M}_{t,-k}$ from the train set and suppressing the $J-\rho$ smallest eigenvalues to zero.

The random splitting is a valid strategy in constructing a test set and a train set since the individuals are assumed to be iid within a cluster. To evaluate the performance of a latent factor model with the dimension ρ, I use the same criteria used in estimating the latent factor model. To see if the cross-validation result is stable across t, I consider the first and the last months of the timeframe—January 1990 and December 2021—and a month in the middle—January 2007—, which is used for a cross-sectional regression in the main text.

	ρ			
t	2	3	5	7
January 1990	0.4742	0.4862	0.4611	0.4567
January 2007	0.6043	0.6173	0.6281	0.6225
December 2021	0.7553	0.7142	0.7399	0.7398
average	0.6113	0.6059	0.6097	0.6063

Table 1: $\text{SSFE}_{t,EmpHistory}(\rho)$

t	ρ				
	1	2	3	5	7
March 1989	0.5856	0.5681	0.5685	0.5689	0.5690
March 2006	0.9985	0.9809	0.9816	0.9826	0.9826
March 2021	0.8984	0.8246	0.8254	0.8261	0.8264
average	0.8275	0.7912	0.7918	0.7925	0.7927

Table 2: $\text{SSFE}_{t,WageInc}(\rho)$

The distribution of $WageInc_{ijt}$ is only observed once a year, in March. Thus, the distributions of $WageInc_{ijt}$ at the time periods above are used as control for the three months I consider: January 1990, January 2007 and December 2021.
The triangular kernel is used and the tuning parameter h is selected by the *density* function in R.

Table 1 contains the cross-validation results for the K-means algorithm on the distribution of $EmpHistory_{ijt}$ and Table 2 contains the cross-validation results for the functional PCA on the distribution of $WageInc_{ijt}$. Table 1 shows that the cross-validation result is not stable across t for the K-means algorithm on the distribution of $EmpHistory_{ijt}$. The cross-validation result from January 1990 suggests using a latent factor model with larger dimension while the cross-validation result from January 2007 suggests using a latent factor model with $\rho_{Kmeans} = 2$. I take the average of the three cross-validation results and let $\rho_{Kmeans} = 3$. As a robustness check, I also present the estimation results from $\rho_{Kmeans} = 5$ below: Section B.3.2. On the other hand, the cross-validation result is stable across t for the functional PCA on the distribution of $WageInc_{ijt}$. I let $\rho_{fPCA} = 2$.

When the sole purpose of estimating the cluster-level latent factors λ_{jt} is to use the factors as controls for the spatial/state-level heterogeneity, one could repeat the cross-validation exercise for every t and let ρ vary across t. However, when the state-level heterogeneity is an object of interest on this own, letting ρ time-invariant can be helpful since then we could connect the support of the latent factor $\mathcal{S}_\lambda$ across different time periods and obtain a pooled estimate on the equilibrium/contextual effect that the state-level distribution $\mathbf{F}_{jt}$ has on

individual-level outcomes. This point on the aggregate-level heterogeneity will be reiterated in Section B.3.2.

B.3 Empirical results

B.3.1 Latent factor estimation for January 2007

Before discussing the estimation results from the regression model (2), here I illustrate how the two latent factor estimation methods are implemented on an actual dataset, by looking at a snapshot of the dataset. As for the timing of the snapshot, I choose January 2007 as I did in the main text, since January 2007 was when the most states raised their minimum wage levels without a federal minimum wage raise.

The outcome of the K-means latent factor estimation is a grouping structure on states. Since $EmpHistory_{ijt}$ captures the latest four month history of individual employment status, the latent factor estimation for January 2007 assigns 50 states and Washington D.C. into one of the $\rho_{Kmeans} = 3$ groups based on the state-level distribution of employment history from September 2006 to December 2006. Figure 1 contains the grouping result and below is the list of states in each group:

Group 1: **Arizona***, Arkansas, **California***, DC, Louisiana, Michigan, Mississippi, New Mexico, **New York***, Oklahoma, **Oregon***, South Carolina, Tennessee, West Virginia

Group 2: Alabama, **Connecticut***, **Delaware***, **Florida***, Georgia, **Hawaii***, Idaho, Illinois, Indiana, Kentucky, Maine, Maryland, **Massachusetts***, **Missouri***, Nevada, New Jersey, **North Carolina***, **Ohio***, **Pennsylvania***, **Rhode Island***, Texas, Utah, Virginia

Group 3: Alaska, **Colorado***, Iowa, Kansas, Minnesota, **Montana***, Nebraska, New Hampshire, North Dakota, South Dakota, **Vermont***, **Washington***, Wisconsin, Wyoming

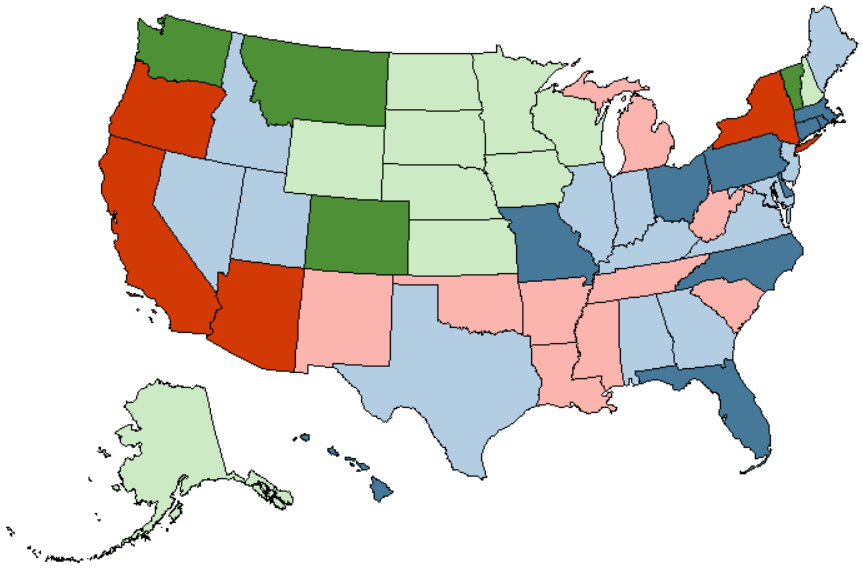

Figure 1: Grouping of states from the distribution of $EmpHistory_{ijt}$, January 2007

50 states and Washing D.C. are grouped into three groups based on the state-level distribution of individual-level employment history from September 2006 to December 2006, which tracks employment, unemployment, and labor force participation. Colors—red, blue, green—denote different groups and darker shades denote an increase in the minimum wage level in January 2007.

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. We can estimate a 'treatment effect,' by interpreting the increase in the minimum wage level as a binary treatment. The within-group comparison is free of the potential treatment endogeneity problem when the distribution of $EmpHistory_{ijt}$ gives us unconfoundedness.

Table 3 shows how the groups estimated using the distribution of $EmpHistory_{ijt}$ differ from one another. Table 3 takes three subsets of $\mathcal{X}$ and computes the proportion of each subset across groups, putting equal weights over states. The three subsets are:

- Always-employed: $\{Emp\}^4$
- Ever-unemployed: $\left\{(EmpStatus_{-1}, \cdots) : EmpStatus_\tau = Unemp \text{ for some } \tau\right\}$
- Never-in-the-labor-force: $\{NotInLaborForce\}^4$

group	1	2	3
Always-employed	0.520	0.588	0.645
Ever-unemployed	0.076	0.060	0.060
Never-in-the-labor-force	0.337	0.281	0.227

Table 3: Heterogeneity across states, January 2007

The table reports proportions of three types of employment history, across 50 states and Washington D.C. The proportions of each employment history are firstly computed within states and then the group mean is computed by putting equal weights on states.

'Always-employed' is the proportion of individuals who have been continuously employed from September 2006 to December 2006, 'Ever-unemployed' is the proportion of individuals who was unemployed for at least one month, and 'Never-in-the-labor-force' is the proportion of individuals who have never been in the labor force from September 2006 to December 2006. Group 1 states have the lowest employment rate and Group 3 states have the highest.

Secondly, to illustrate how the functional PCA is applied to a real dataset, I look at March 2006 ASEC sample; this sample is used for the latent factor estimation on the distribution of $WageInc_{ijt}$ for January 2006, due to the ASEC sample being observed only once a year. After applying the eigenvalue decomposition to the product matrix computed from the conditional densities of $\log WageInc_{ijt}$ given $WageInc_{ijt} > 0$ across 50 states and Washington D.C., the second to the fourteenth largest eigenvalues are plotted in Figure 2. The biggest eigenvalue is much bigger than the rest of the eigenvalues, with the associated eigenvectors being mostly constant across states, and is therefore omitted. We can see that the second biggest eigenvalue is much bigger than the third to the fourteenth eigenvalues. This means that the additional gain in explaining the variation across the state-level conditional densities of $\log WageInc_{ijt}$ is much bigger when we increase ρ_{fPCA} from one to two, than when we further increase ρ_{fPCA} from two. This observation is coherent with the cross-validation results from the previous subsection that choose $\rho_{fPCA} = 2$.

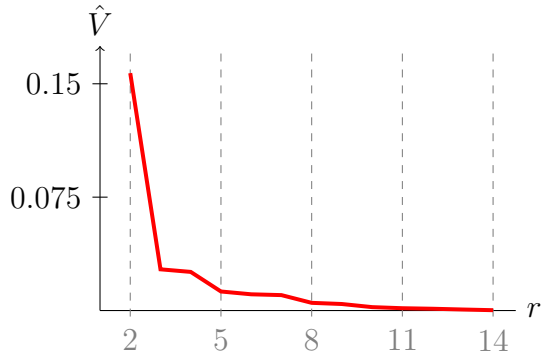

Figure 2: The scree plot of eigenvalues from the distribution of $WageInc_{ijt}$, March 2006

March 2006 ASEC sample is used in constructing the wage income densities. $WageInc_{ijt}$ is truncated at $WageInc_{ijt} > 0$ and logged. The biggest eigenvalue is not included in the plot. Its value is 15.37.

In addition, I plotted the second component of the estimated latent factor in Figure 3. Several northeastern states and Alaska have higher values of the second component of $\hat{\lambda}_{jt,WageInc}$ while some southern states such as Arkansas and Mississippi and mountain states such as Montana have lower values. We do not have an interpretation for the second

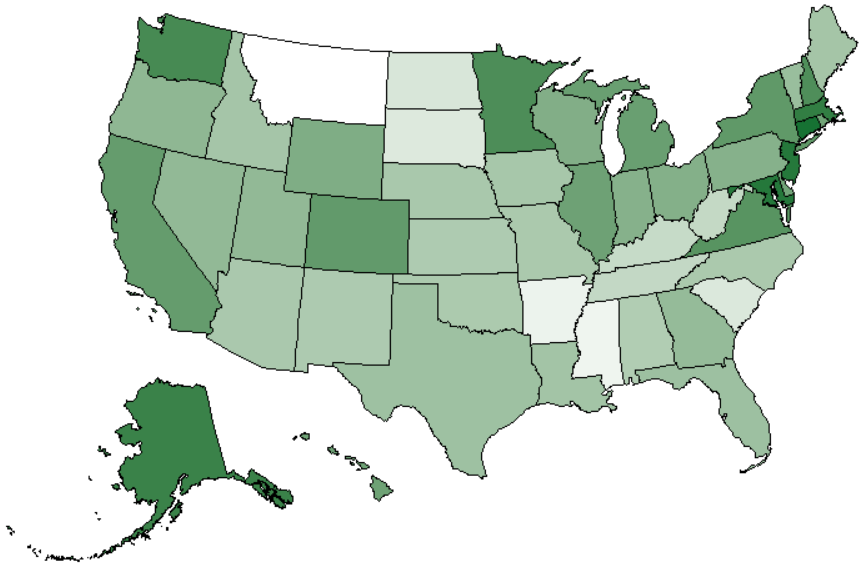

Figure 3: The second component of $\hat{\lambda}_{jt,WageInc}$ across states, March 2006

March 2006 ASEC sample is used in constructing the wage income densities. $WageInc_{ijt}$ is truncated at $WageInc_{ijt} > 0$ and logged. The darker shade corresponds to a higher value of $\hat{\lambda}_{2jt,WageInc}$ and the lighter shade correspond to a lower value.

component of $\hat{\lambda}_{jt,WageInc}$; Figure 3 only tells us which states are similar in that regard. Not being able to interpret the value of $\hat{\lambda}_{jt,WageInc}$ is due to the rotation on the latent factor and is a definite caveat of the latent factor models suggested in the paper. However, as discussed in the main text, not being able to interpret the estimates does not stop us from having an interpretable model and we can still conduct comparative statistics in terms of the distribution of $WageInc_{ijt}$.

B.3.2 Pooled regression on disemployment effect

Now, I discuss the regression results from (2). Table 4 expands Table 4 of the main text and includes estimation results when $\rho_{Kmeans} = 5$. As in the main text, I use time-specific coefficients for $\lambda_{jt,EmpHistory}$ and time-invariant coefficients for $\lambda_{jt,WageInc}$. Columns (2) and (5) contain the estimation results when $\rho_{Kmeans} = 3$ and columns (3) and (6) contain the estimation results when $\rho_{Kmeans} = 5$. The estimation results are stable across the choice of ρ_{Kmeans}.

$\hat{\beta}$	(1)	(2)	(3)	(4)	(5)	(6)
	-0.109*	-0.052	-0.061	-0.029*	-0.030*	-0.033**
	(0.061)	(0.079)	(0.086)	(0.017)	(0.017)	(0.016)
time FE	X	X	X	O	O	O
$EmpHistory$	X	O ($\rho = 3$)	O ($\rho = 5$)	X	O ($\rho = 3$)	O ($\rho = 5$)
$WageInc$	X	O	O	X	O	O
T	1 (January 2007)			264 (2000-2021)		

Table 4: Disemployment effect estimates across specifications

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.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

In the regression model (2), the state minimum wage level $MinWage_{jt}$ enters after taking logarithm, following the convention in the literature. Thus, by dividing the slope coefficient on $\log MinWage_{jt}$ with the average teen employment rate from the dataset, which is 0.328, we get the elasticity interpretation. Based on columns (4)-(6) of Table 4, an one percentage point increase in the minimum wage level reduces teen employment by 0.087-0.099 percentage point. Neumark and Shirley (2022) provides a meta-analysis of studies on teen employment and minimum wage and find that the mean of the estimates across studies is -0.170 and the median is -0.122. By controlling for the state-level heterogeneity in a more rigorous manner using the state-level distributions, I find that the existing literature slightly overestimates the wage elasticity of teen employment.

$\hat{\beta}$	(1)	(2)	(3)
$\{\lambda_{EmpHistory} = e_1\}$	-0.027	-0.027	-0.027
	(0.017)	(0.016)	(0.017)
$\{\lambda_{EmpHistory} = e_2\}$	-0.029*	-0.028	-0.028
	(0.017)	(0.017)	(0.017)
$\{\lambda_{EmpHistory} = e_3\}$	-0.030*	-0.042*	-0.042*
	(0.017)	(0.023)	(0.023)
time FE	O	O	O
$EmpHistory$	X	O	O
$WageInc$	X	X	O
T	264 (2000-2021)		

Table 5: Aggregate-level heterogeneity in disemployment effect

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.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

Table 5 discuss the aggregate heterogeneity in disemployment effect:

$$Y_{ijt} = \log MinWage_{jt} \cdot \left(\sum_r \beta_r \mathbf{1}\{\lambda_{jt,EmpHistory} = e_r\} \right) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (3)$$

The coefficient on the minimum wage is a function of the distribution of $EmpHistory_{ijt}$. To connect the 'labels' of the grouping structure across different time periods, I reordered $\lambda_{jt,EmpHistory}$ across t so that Group 1 (i.e. $\lambda_{jt,EmpHistory} = e_1$) is always the group of states with lower employment rate and lower labor force participation rate and Group 3 (i.e. $\lambda_{jt,EmpHistory} = e_3$) is always the group of states with higher employment rate and higher labor force participation rate. Columns (3)-(4) show us that teens in Group 3 states are more affected by the minimum wage than teens in Group 1 states. This may happen for a variety of reasons; e.g., Group 3 states may have thicker labor supply on lower end of the wage distribution and thus low-skilled teenagers get replaced more easily.

Lastly, I study the interaction between the aggregate-level heterogeneity and the individual-level heterogeneity in terms of race. The left panel of Table 6 estimates

$$Y_{ijt} = \log MinWage_{jt} \cdot \Big(\beta_1 \mathbf{1}\{White_{ij} = 1\} + \beta_0 \mathbf{1}\{White_{ij} = 0\} \Big) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (4)$$

Disemployment effect is modeled to be heterogeneous at the individual level in terms of race. β_1 is the disemployment effect coefficient on white teenagers and β_0 is the disemployment effect coefficient on non-white teenagers. The right panel of Table 6 estimates

$$Y_{ijt} = \log MinWage_{jt} \cdot \Big(\sum_{w=0,1} \sum_r \beta_{w,r} \mathbf{1}\{White_{ij} = w, \lambda_{jt,EmpHistory} = e_r\} \Big) + \alpha_j + \lambda_{jt}^{\intercal}\delta_t + X_{ijt}^{\intercal}\theta_1 + \theta_2 EmpRate_{jt} + U_{ijt}. \quad (5)$$

The aggregate-level heterogeneity in terms of the distribution of $EmpHistory_{ijt}$ is introduced, in addition to the individual-level heterogeneity in terms of race. $\beta_{1,r}$ is the disemployment effect coefficient on white teenagers in Group r states while $\beta_{0,r}$ is the disemployment effect coefficient on non-white teenagers in Group r states.

From the left panel of Table 6, 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. The racial disparity interacts with the labor market fundamentals. The right panel of Table 6 shows us that the racial disparity persists across groups and interact with the aggregate heterogeneity in a way that the employment effect for non-white teenagers is mitigated in Group 3. Figure 4 contains confidence intervals for interactive disemployment effect coefficients from Column (4) of Table 5 in the main text and Column (4) of Table 6.

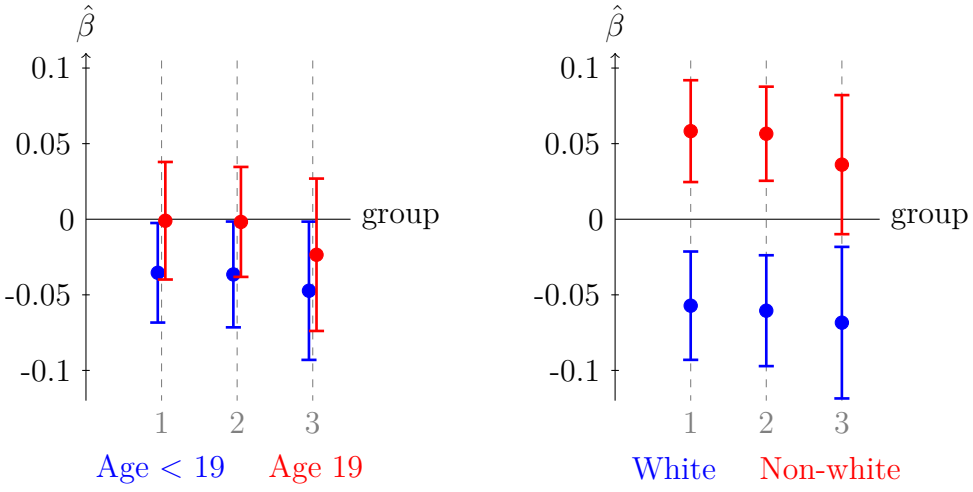

Figure 4: 95% confidence intervals on disemployment effect coefficient

The x-axis denotes the group. The color denotes the individual-level control covariate.
The y-axis is estimates and confidence interval.
Comparison across colors at each point of the x-axis relates to individual heterogeneity and comparison across x-axis for the same color relates to aggregate heterogeneity.

$\hat{\beta}$	(1)	(2)	(3)	(4)
$\{White = 1\}$	-0.061***	-0.061***		
	(0.018)	(0.018)		
$\times\{\lambda_{EmpHistory} = e_1\}$			-0.057***	-0.057***
			(0.018)	(0.018)
$\times\{\lambda_{EmpHistory} = e_2\}$			-0.060***	-0.060***
			(0.019)	(0.019)
$\times\{\lambda_{EmpHistory} = e_3\}$			-0.068**	-0.068**
			(0.026)	(0.026)
$\{White = 0\}$	0.054***	0.054***		
	(0.016)	(0.016)		
$\times\{\lambda_{EmpHistory} = e_1\}$			0.058***	0.058***
			(0.017)	(0.017)
$\times\{\lambda_{EmpHistory} = e_2\}$			0.057***	0.057***
			(0.016)	(0.016)
$\times\{\lambda_{EmpHistory} = e_3\}$			0.036	0.036
			(0.024)	(0.023)
$EmpHistory$	O	O	O	O
$WageInc$	X	O	X	O
T	264 (2000-2021)			

Table 6: Individual-level and interactive heterogeneity in disemployment effect

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.

*, **, ** denote significance level 0.1, 0.05, 0.001, respectively.

C Proofs

C.1 Theorem 1

Firstly, we want to show that the objective function constructed with the estimated latent factors is close to the infeasible objective function with the rotated true latent factors: for any $\theta \in \tilde{A}\Theta$,

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

By taking the first-order Taylor's expansion of m with regard to $\hat{\lambda}_j$ around $A\lambda_j$,

$$\begin{aligned} \left| \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) \right\|_2 - \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \right| &\leq \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \theta\right) - \frac{1}{J} \sum_{j=1}^{J} m\left(\widehat{W}_j; \theta\right) \right\|_2 \\ &= \left\| \frac{1}{J} \sum_{j=1}^{J} R_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right) \right\|_2. \end{aligned}$$

We can apply the Taylor's expansion since the mapping $\lambda \mapsto m\big(W_j(\lambda); \theta\big)$ is continuously differentiable on $A\mathcal{S}_\lambda$ for each $\theta \in \tilde{A}\Theta$: for any $\theta \in \tilde{A}\Theta$ and any λ' in the interior of $A\mathcal{S}_\lambda$,

$$\begin{aligned} m(W_j(\lambda'); \theta) &= m\big(W_j(A^{-1}\lambda'); \tilde{A}^{-1}\theta\big) \\ \frac{\partial}{\partial \lambda} m(W_j(\lambda); \theta)\Big|_{\lambda=\lambda'} &= \frac{\partial}{\partial \lambda} m\big(W_j(A^{-1}\lambda); \tilde{A}^{-1}\theta\big)\Big|_{\lambda=\lambda'} \\ &= \frac{\partial}{\partial \lambda} m\big(W_j(\lambda); \tilde{A}^{-1}\theta\big)\Big|_{\lambda=A^{-1}\lambda'} \cdot A^{-1} \end{aligned}$$

The first two equalities hold from Assumption 2.d. The last equality holds from the chain rule and the differentiability of the mapping $\lambda \mapsto m\big(W_j(\lambda); \theta\big)$ at $A^{-1}\lambda' \in \mathcal{S}_\lambda$ for $\tilde{A}^{-1}\theta \in \Theta$ from Assumption 2.e.

Note that $R_j(\cdot, \cdot)$ in the remainder term is a $l \times \rho$ matrix; if λ_j is a scalar, R_j would be a first-order derivative of m with regard to λ_j, evaluated at some point between $A\lambda_j$ and $\hat{\lambda}_j$.

Let $\tilde{R}_j$ denote an arbitrary row of R_j. By applying the Cauchy-Schwarz inequality to the j-th cluster in the summation,

$$\left|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right| \leq \left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2 \left\|\hat{\lambda}_j - A\lambda_j\right\|_2.$$

By applying the Cauchy-Schwarz inequality again,

$$\begin{aligned}\left|\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right| &\leq \frac{1}{J}\sum_{j=1}^{J}\left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2 \left\|\left(\hat{\lambda}_j - A\lambda_j\right)\right\|_2 \\ &\leq \left(\frac{1}{J}\sum_{j=1}^{J}\left\|\tilde{R}_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_2^{\,2}\right)^{\frac{1}{2}}\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}\right)^{\frac{1}{2}}.\end{aligned}$$

Then, by summing over the rows of R_j, we get

$$\left\|\frac{1}{J}\sum_{j=1}^{J}R_j\left(\hat{\lambda}_j, A\lambda_j\right)\left(\hat{\lambda}_j - A\lambda_j\right)\right\|_2^{\,2} \leq \left(\frac{1}{J}\sum_{j=1}^{J}\left\|R_j\left(\hat{\lambda}_j, A\lambda_j\right)\right\|_F^{\,2}\right)\left(\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}\right).$$

$\frac{1}{J}\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^{\,2}$ is $\frac{1}{J}\cdot o_p(1)$ from the conditions of Theorem 1.

It remains to show that $\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}$ is $O_p(1)$. From the Taylor's theorem, the matrix R_j can be written as an integral as follows:

$$\begin{aligned}R_j &= \int_0^1 \frac{\partial}{\partial\lambda} m\left(W_j(\lambda);\theta\right)\Big|_{\lambda = A\lambda_j + t\left(\hat{\lambda}_j - A\lambda_j\right)} dt \\ &= \int_0^1 \frac{\partial}{\partial\lambda} m\left(W_j(\lambda);\tilde{A}^{-1}\theta\right)\Big|_{\lambda = \lambda_j + t\left(A^{-1}\hat{\lambda}_j - \lambda_j\right)} \cdot A^{-1} dt.\end{aligned}$$

Find that

$$\|R_j\|_F^{\,2} \leq l\rho\left(\rho \sup_{t\in[0,1]}\left\|m\big(W_j(\lambda);\tilde{A}^{-1}\theta\big)\Big|_{\lambda=\lambda_j+t\left(A^{-1}\hat{\lambda}_j-\lambda_j\right)}\right\|_\infty \cdot \left\|A^{-1}\right\|_\infty\right)^2$$

by finding the components of $\frac{\partial}{\partial\lambda}m$ and A^{-1} with the biggest absolute values. Then,

$$\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2} \leq \frac{l\rho^3}{J}\sum_{j=1}^{J}\sup_{t\in[0,1]}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\tilde{A}^{-1}\theta\right)\Big|_{\lambda=\lambda_j+t\left(A^{-1}\hat{\lambda}_j-\lambda_j\right)}\right\|_F^{\,2}\cdot\left\|A^{-1}\right\|_F^{\,2}.$$

Lastly, from Assumption 2.f, $\left\|A^{-1}\right\|_F$ is bounded with probability going to one. Thus, $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2 \leq \left\|A^{-1}\right\|_F\cdot\max_j\left\|\hat{\lambda}_j-A\lambda_j\right\|_2 \leq \eta$ holds with probability going to one, from the condition of Theorem 1. In addition, conditioning on the event that $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\eta$ and $\left\|A^{-1}\right\|_F\leq\tilde{M}$, we have

$$\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2} \leq \frac{\tilde{M}^2l\rho^3}{J}\sum_{j=1}^{J}\sup_{\|\lambda'-\lambda_j\|_2\leq\eta}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\theta\right)\Big|_{\lambda=\lambda'}\right\|_F^{\,2}.$$

From Assumption 2.e, the RHS of the inequality above is bounded in expectation by $M\tilde{M}^2l\rho^3$.

Consequently, we have that $\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}$ is $O_p(1)$: for any $\varepsilon>0$, find large enough J^* such that the probability that $\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\frac{\eta}{\tilde{M}}$ and $\left\|A^{-1}\right\|_F\leq\tilde{M}$ holds is bigger than $1-\frac{\varepsilon}{3}$ and large enough M^* such that the probability of the RHS of the inequality above being bigger than M^* is smaller than $\frac{\varepsilon}{3}$. Then, for $J\geq J^*$,

$$\begin{aligned}
&\Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}\geq M^*\right\}\\
&\leq \Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\|R_j\|_F^{\,2}\geq M^*,\max_j\left\|A^{-1}\hat{\lambda}_j-\lambda_j\right\|_2\leq\eta,\left\|A^{-1}\right\|_F\leq\tilde{M}\right\}+\frac{\varepsilon}{3}\\
&\leq \Pr\left\{\frac{\tilde{M}^2l\rho^3}{J}\sum_{j=1}^{J}\sup_{\|\lambda'-\lambda_j\|_2\leq\eta}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\lambda}m\left(W_j(\lambda);\theta\right)\Big|_{\lambda=\lambda'}\right\|_F^{\,2}\geq M^*\right\}+\frac{\varepsilon}{3}\leq\frac{2\varepsilon}{3}.
\end{aligned}$$

Note that the stochastic boundedness is uniform across $\theta\in\tilde{A}\Theta$ since the quantity in the last probability involves a supremum over Θ.

Having shown that the feasible objective function is close to the infeasible objective function, I now show that the consistency of the infeasible GMM estimator leads to the

consistency of the feasible GMM estimator. Find that

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

The inequality is from the definition of the GMM estimator. The first equality holds for a random object $\hat{\theta}$ since the stochastic boundedness of $\frac{1}{J}\sum_{j=1}^{J} \|R_j\|_F{}^2$ does not depend on the choice of θ. The second to the last equality is from Assumption 2.c-d. Again, from Assumption 2.c-d, we get

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

The first equality holds from Assumption 2.d and the second equality holds from Assumption 2.c. Then,

$$\left\| \mathbf{E}\left[m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \right] \right\|_2 = \left\| \mathbf{E}\left[m\left(W_j; \hat{\theta}\right) \right] \right\|_2 = \left\| \frac{1}{J} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) \right\|_2 + o_p(1) = o_p(1).$$

We get the consistency of $\tilde{A}^{-1}\hat{\theta}$ to θ^0 from Assumption 2.b and thus the consistency of $\hat{\theta}$ to

$\tilde{A}\theta^0$ from Assumption 2.f: for any $\varepsilon > 0$,

$$\begin{aligned}\Pr\left\{\left\|\hat{\theta} - \tilde{A}\theta^0\right\|_2 \geq \varepsilon\right\} &\leq \Pr\left\{\left\|\tilde{A}\right\|_F \cdot \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \varepsilon\right\} \\ &\leq \Pr\left\{\left\|\tilde{A}\right\|_F \cdot \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \varepsilon, \left\|\tilde{A}\right\|_F \leq \tilde{M}\right\} + \Pr\left\{\left\|\tilde{A}\right\|_F > \tilde{M}\right\} \\ &\leq \Pr\left\{\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \geq \frac{\varepsilon}{\tilde{M}}\right\} + \Pr\left\{\left\|\tilde{A}\right\|_F > \tilde{M}\right\} = o(1).\end{aligned}$$

C.2 Theorem 2

Recall that

$$\begin{aligned}\left\|\frac{1}{J}\sum_{j=1}^{J} m\left(\widehat{W}_j;\theta\right) - \frac{1}{J}\sum_{j=1}^{J} m\left(W_j;\theta\right)\right\|_2 &= O_p(1)\cdot\frac{1}{\sqrt{J}}\cdot\left(\sum_{j=1}^{J}\left\|\hat{\lambda}_j - A\lambda_j\right\|_2^2\right)^{\frac{1}{2}} \\ &= O_p(1)o_p(1)\frac{1}{\sqrt{J}}\end{aligned}$$

from the proof of Theorem 1 and therefore

$$\begin{aligned}&\left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right)\right\|_2 \\ &\geq \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 - \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right) - \frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 \\ &= \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 + o_p(1).\end{aligned}$$

From the condition of Theorem 2, we get

$$\begin{aligned}o_p(1) = \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(\widehat{W}_j;\hat{\theta}\right)\right\|_2 &\geq \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2 + o_p(1) \\ o_p(1) &= \left\|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} m\left(W_j;\hat{\theta}\right)\right\|_2.\end{aligned}$$

Step 1.

For asymptotic normality result, we need a stronger consistency result for $\tilde{A}^{-1}\hat{\theta}$ than Theorem 1. For that, let us apply the first-order Taylor's expansion to the objective function with regard to the parameter of interest θ^0:

$$\begin{aligned} o_p(1) &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j; \hat{\theta}\right) = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j^*; \tilde{A}^{-1}\hat{\theta}\right) \\ &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} m\left(W_j^*; \theta^0\right) + \frac{1}{J} \sum_{j=1}^{J} R_{1j}\left(\tilde{A}^{-1}\hat{\theta}, \theta^0\right) \cdot \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right). \end{aligned}$$

We can apply the Taylor's expansion since Assumption 3.a assumes twice-differentiability of m.

R_{1j} is a $l \times k$ matrix for the first-order remainder term in the expansion. The remainder term coefficient matrix R_{1j} can be rewritten as

$$R_{1j}\left(\tilde{A}^{-1}\hat{\theta}, \theta^0\right) = \int_0^1 \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt.$$

Find that

$$\begin{aligned} &\left\| \frac{1}{J}\sum_{j=1}^{J} \int_0^1 \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt - \int_0^1 \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] dt \right\|_F \\ &= \left\| \int_0^1 \left(\frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] \right) dt \right\|_F \\ &\leq \sqrt{lk} \cdot \sup_{t\in[0,1]} \left\| \frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*; \theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}\right] \right\|_F \\ &= o_p(1). \end{aligned}$$

The first equality holds from Fubini's theorem since the integral and the summation are both defined with σ-finite measures on $\{1, \cdots, J\}$ and $[0, 1]$. The inequality holds from finding that any component of the $l \times k$ matrix $\frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \theta} m - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\right]$ for a given $t \in [0, 1]$ and therefore its integral over $[0, 1]$ are bounded by the supremum in the Frobenius norm. The

second equality holds from Assumption 3.b. Lastly, find that

$$\left\| \int_0^1 \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} \right] dt - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] \right\|_F$$
$$\leq \sqrt{lk} \cdot \sup_{t\in[0,1]} \left\| \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} \right] - \mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] \right\|_F = o_p(1).$$

The inquality holds since any component of the $l\times k$ matrix $\mathbf{E}\left[\frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta^0=t(\tilde{A}^{-1}\hat{\theta}-\theta^0)}\right] dt$ for a given $t\in[0,1]$ is bounded by the supremum over the Frobenius norm. $\theta \mapsto \frac{\partial}{\partial\theta} m(W_j^*;\theta)$ is continuously differentiable from Assumption 3.a. From the Leibniz's rule, its expectation is also differentiable and thus continuous; the equality holds. $\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right)$ converges to a full rank matrix $\mathbf{E}\left[\frac{\partial}{\partial\theta} m(W_j^*;\theta)\big|_{\theta=\theta^0}\right]$.

Lastly, since $\frac{1}{\sqrt{J}}\sum_{j=1}^J m\left(W_j^*;\theta^0\right)$ is $O_p(1)$ from the CLT,

$$\left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right) \cdot \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) = O_p(1).$$

Therefore $\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) = O_p(1)$ by finding a small neighborhood around $\mathbf{E}\left[\frac{\partial}{\partial\theta} m\left(W_j^*;\theta\right)\big|_{\theta=\theta^0}\right]$ such that $\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right)$ is full rank and the Frobenius norm of its left inverse is bounded: for any M^* and M_R,

$$\begin{aligned}
&\Pr\left\{ \left\| \sqrt{J}\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right) \right\|_2 \geq M^* \right\} \\
&\leq \Pr\left\{ \left\| \left(\frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right) \right)^{-1} \frac{1}{\sqrt{J}} \sum_{j=1}^J \left(m\left(W_j^*;\tilde{A}^{-1}\hat{\theta}\right) - m\left(W_j^*;\theta^0\right) \right) \right\|_2 \geq M^* \right\} \\
&\quad + \Pr\left\{ \frac{1}{J}\sum_{j=1}^J R_{1j}\left(\tilde{A}^{-1}\hat{\theta},\theta^0\right) \text{ is not full rank} \right\} \\
&= \Pr\left\{ \left\| \left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right)^{-1} \right\|_F \cdot \left\| O_p(1) \right\|_F \geq M^* \right\} + o(1) \\
&\leq \Pr\left\{ \left\| O_p(1) \right\|_F \geq \frac{M^*}{M_R} \right\} + \Pr\left\{ \left\| \left(\mathbf{E}\left[\frac{\partial}{\partial \theta} m\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0} \right] + o_p(1) \right)^{-1} \right\|_F > M_R \right\} + o(1).
\end{aligned}$$

The second probability in the last inequality goes to zero for large enough M_R from the

continuous mapping theorem since each component of the inverse matrix is a continuous function of the original matrix. Given some $\varepsilon > 0$, first choose large enough M_R so that the second probability in the last inequality is arbitrarily small for large J and then choose large enough M^* so that the first probability is arbitrarily small as well. Then, we can find some J^* such that $\Pr\left\{\left\|\sqrt{J}(\tilde{A}^{-1}\hat{\theta} - \theta)\right\|_F \leq M^*\right\} \leq \varepsilon$ for $J \geq J^*$.

Step 2.

Again, let $\tilde{m}$ denote an arbitrary component of m. From the component-wise second-order Taylor's expansion,

$$\begin{aligned} o_p(1) &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \tilde{m}\left(W_j; \hat{\theta}\right) \\ &= \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \tilde{m}\left(W_j; \tilde{A}\theta^0\right) + \frac{1}{J} \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \tilde{m}\left(W_j; \theta\right)\big|_{\theta = \tilde{A}\theta^0} \cdot \sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) \\ &\qquad + \left(\hat{\theta} - \tilde{A}\theta^0\right)^{\intercal} \cdot \frac{1}{J} \sum_{j=1}^{J} \tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right) \cdot \sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) \end{aligned}$$

where $\tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right)$ is a $k \times k$ matrix for the second-order remainder term. Find that

$$\tilde{R}_{2j}\left(\hat{\theta}, \tilde{A}\theta^0\right) = \left(\tilde{A}^{\intercal}\right)^{-1} \cdot \int_0^1 (1-t) \frac{\partial^2}{\partial \theta \partial \theta^{\intercal}} \tilde{m}\left(W_j^*; \theta\right)\Big|_{\theta = \theta^0 + t\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)} dt \cdot \tilde{A}^{-1}$$

from the Taylor's theorem and by applying the chain rule. For any $M^{**} > 0$,

$$\begin{aligned} &\Pr\left\{\left\|\frac{1}{J}\sum_{j=1}^{J}\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt\right\|_F \geq M^{**}\right\} \\ &\leq \Pr\left\{\frac{1}{J}\sum_{j=1}^{J}\left\|\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)} dt\right\|_F \geq M^{**}, \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \leq \eta\right\} \\ &\quad + \Pr\left\{\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 > \eta\right\} \\ &\leq \Pr\left\{\frac{k}{J}\sum_{j=1}^{J}\sup_{\|\theta' - \theta^0\|_2 \leq \eta}\left\|\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta'}\right\|_F \geq M^{**}\right\} + o(1). \end{aligned}$$

The last inequality holds from the fact that any component of $\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}$ for a given $t \in [0,1]$ and therefore any component of the integral $\int(1-t)\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}dt$ are bounded by the supremum over the Frobenius norm, when $\left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \leq \eta$. Given some $\varepsilon > 0$, we can find large enough M^{**} and J^{**} such that

$$\Pr\left\{\left\|\frac{1}{J}\sum_{j=1}^{J}\int_0^1 (1-t)\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\tilde{m}\left(W_j^*;\theta\right)\Big|_{\theta=\theta^0+t\left(\tilde{A}^{-1}\hat{\theta}-\theta^0\right)}dt\right\|_F \geq M^{**}\right\} \leq \varepsilon$$

for any $J \geq J^{**}$, from Assumption 3.a. $\tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}$ is $O_p(1)$.

Lastly, since $\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right) = O_p(1)$, the second-order remainder term in the second-order approximation is $o_p(1)$:

$$\begin{aligned}
&\left|\left(\hat{\theta} - \tilde{A}\theta^0\right)^\intercal \cdot \frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\cdot\sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right)\right| \\
&\left|\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)^\intercal \cdot \tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}\cdot\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)\right| \\
&\leq \left\|\tilde{A}^{-1}\hat{\theta} - \theta^0\right\|_2 \cdot \left\|\tilde{A}^\intercal\frac{1}{J}\sum_{j=1}^{J}\tilde{R}_{2j}\left(\hat{\theta},\tilde{A}\theta^0\right)\tilde{A}\right\|_F \cdot \left\|\sqrt{J}\left(\tilde{A}^{-1}\hat{\theta} - \theta^0\right)\right\|_2 = o_p(1).
\end{aligned}$$

Thus,

$$\sqrt{J}\left(\hat{\theta} - \tilde{A}\theta^0\right) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{\partial}{\partial\theta}m\left(W_j;\tilde{A}\theta^0\right)\right)^{-1}\frac{1}{\sqrt{J}}\sum_{j=1}^{J}m\left(W_j;\tilde{A}\theta^0\right) + o_p(1).$$

C.3 Proposition 1

For the convenience of notation, let $\lambda_j \in \{1,\cdots,\rho\}$ for true latent factor λ_j as well.

Step 1

From the within-cluster iidness,

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

The second equality holds from exchanging the order of integration and expectation. Thus,

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

Step 2

Let us connect $\hat{G}(1), \cdots, \hat{G}(\rho)$ to $G(1), \cdots, G(\rho)$. Define $\sigma(r)$ such that

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

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

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

The last inequality holds since $\sum_{j=1}^{J}\left\|\hat{G}(\hat{\lambda}_j)-\hat{\mathbf{F}}_j\right\|_{w,2}^{2} \leq \sum_{j=1}^{J}\left\|G(\lambda_j)-\hat{\mathbf{F}}_j\right\|_{w,2}^{2}$ from the definition of $\hat{G}$ and $\hat{\lambda}$. From Assumption 4.a, $\sum_{j=1}^{J}\mathbf{1}\{\lambda_j = r\}/J \xrightarrow{p} \mu(r) > 0$ as $J \to \infty$. Thus,

$$\left\|\hat{G}(\sigma(r))-G(r)\right\|_{w,2}^{2} \xrightarrow{p} 0$$

as $J \to \infty$ from Assumption 4.c and Step 1.

Note that for any $r' \neq r$,

$$\left\|\hat{G}(\sigma(r))-G(r')\right\|_{w,2}^{2} \geq \frac{1}{2}\|G(r)-G(r')\|_{w,2}^{2} - \left\|\hat{G}\left(\sigma\left(r\right)\right)-G(r)\right\|_{w,2}^{2} = \frac{1}{2}c(r,r') + o_p(1)$$

as $J \to \infty$ from the same argument from above and Assumption 4.c.

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

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

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

Step 3

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

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

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

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

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

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

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

The last inequality is from the construction of the event A_ε. In the last inequality A_ε can be dropped since the probability does not require σ being bijective to be well-defined. Set

$\varepsilon^* = \frac{c^*}{20}$ so that $\frac{c^*}{12} - \frac{5}{6}\varepsilon^* = \frac{c^*}{24} > 0$.

By repeating the expansion for every j,

$$\begin{aligned}\Pr\left\{\exists\ j \text{ s.t. } \hat{\lambda}_j \neq \lambda_j\right\} &\leq \Pr\left\{\exists\ j \text{ s.t. } \hat{\lambda}_j \neq \lambda_j, A_{\varepsilon^*}\right\} + \Pr\left\{{A_{\varepsilon^*}}^c\right\} \\ &\leq \sum_{j=1}^{J} \Pr\left\{\frac{c^*}{24} \leq \left\|G(\lambda_j) - \hat{\mathbf{F}}_j\right\|_{w,2}^{\ 2}\right\} + \Pr\left\{{A_{\varepsilon^*}}^c\right\}.\end{aligned}$$

We already know $\Pr\left\{{A_{\varepsilon^*}}^c\right\} = o(1)$ as $J \to \infty$. It remains to show that the first quantity in the RHS of the inequality is $o(J/N_{\min}^{\nu})$ for any $\nu > 0$. Let ε^{**} denote $\frac{c^*}{24}$. Choose an arbitrary $\nu > 0$. From the within-cluster iidness,

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

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

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

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

C.4 Proposition 2

Step 1. Firstly, let us discuss the rotation on the latent factor. For notational simplicity, let

$$V = \begin{pmatrix} \int_{\mathbb{R}} g_1(x)^2 w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x) g_1(x) w(x)dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}} g_1(x) g_\rho(x) w(x)dx & \cdots & \int_{\mathbb{R}} g_\rho(x)^2 w(x)dx \end{pmatrix},$$
$$\Lambda = \begin{pmatrix} \lambda_1 & \cdots & \lambda_J \end{pmatrix}.$$

Suppose $rank(M) = rank(\Lambda^\intercal V \Lambda) = \rho$ and consider an eigen-decomposition for M with orthonormal eigenvectors, using the ρ positive eigenvalues: $V_1, \cdots, V_\rho$. Let Q be a $J \times \rho$ matrix with the orthonormal eigenvectors as columns and let $\tilde{\Lambda} = \sqrt{J} Q^\intercal$. Then, $\frac{1}{J}\tilde{\Lambda}\tilde{\Lambda}^\intercal = Q^\intercal Q = I_\rho$ and

$$\Lambda^\intercal V \Lambda = M = Q \text{diag}\left(V_1, \cdots, V_\rho\right) Q^\intercal = \tilde{\Lambda}^\intercal \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right) \tilde{\Lambda}.$$

Let

$$A^\intercal = V \left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right) \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1},$$

we have

$$\begin{aligned} \Lambda^\intercal A^\intercal &= \Lambda^\intercal V \left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right) \text{diag}\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)^{-1} \\ &= \tilde{\Lambda}^\intercal \text{diag}\left(\frac{\nu_1}{J}, \cdots, \frac{\nu_\rho}{J}\right) \frac{1}{J}\tilde{\Lambda}\tilde{\Lambda}^\intercal \text{diag}\left(\frac{\nu_1}{J}, \cdots, \frac{\nu_\rho}{J}\right)^{-1} = \tilde{\Lambda}^\intercal. \end{aligned}$$

We have a rotation between the matrix of the true latent factor Λ and the matrix of (rescaled) eigenvectors $\tilde{\Lambda}$.

The rotation matrix A in Proposition 2 satisfies Assumption 2.f:

$$\left\|A^{-1}\right\|_F = \left\| diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1} V^{-1}\right\|_F .$$

Find that

$$\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal \cdot diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right) \cdot \frac{1}{J}\tilde{\Lambda}\Lambda^\intercal = \frac{1}{J}\Lambda\Lambda^\intercal \cdot V \cdot \frac{1}{J}\Lambda\Lambda^\intercal$$

$$\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1} = \left(\frac{1}{J}\Lambda\Lambda^\intercal \cdot V \cdot \frac{1}{J}\Lambda\Lambda^\intercal\right)^{-1} \cdot \frac{1}{J}\Lambda\tilde{\Lambda}^\intercal \cdot diag\left(\frac{V_1}{J}, \cdots, \frac{V_\rho}{J}\right)$$

and since the Frobenius norm is invariant to a unitary operation

$$\left\|\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right\|_F \leq \frac{1}{\sqrt{J}}\left\|\Lambda\right\|_F = \left(\frac{1}{J}\sum_{j=1}^{J}\left\|\lambda_j\right\|_2{}^2\right)^{\frac{1}{2}} = O_p(1).$$

$\left(\frac{1}{J}\Lambda\tilde{\Lambda}^\intercal\right)^{-1}$ is also $O_p(1)$, satisfying Assumtion 2.f.

Step 2. Now, we show the estimate $\widehat{M}$ is close to the true matrix M. The following convergence rate on $\left\|\widehat{M} - M\right\|_F$ is a multivariate extension of Proposition 1 and Theorem 1 of Kneip and Utikal (2001).

$$\left\|\widehat{M} - M\right\|_F = O_p\left(\frac{J}{\sqrt{\min_j N_j}}\right).$$

To avoid notational complexity, I will use subscript λ to indicate that the expectation is conditioning on (N_j, Z_j, λ_j). Find that

$$\mathbf{E}_\lambda\left[\left(\widehat{M}_{jk} - M_{jk}\right)^2\right] = \mathrm{Var}_\lambda\left(\widehat{M}_{jk}\right) + \left(\mathbf{E}_\lambda\left[\widehat{M}_{jk}\right] - M_{jk}\right)^2$$

From the kernel estimation,

$$\begin{aligned}
&\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)\right] \\
&= \int_{\mathbb{R}^p}\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-x')\right)\mathbf{f}_j(x')dx' \\
&= \int_{\mathbb{R}^p}K(t)\mathbf{f}_j\big(x-H^{\frac{1}{2}}t\big)dt \qquad \text{by letting } x'=x-H^{\frac{1}{2}}t \\
&= \int_{\mathbb{R}^p}K(t)\left(\mathbf{f}_j(x)-\mathbf{f}_j^{(1)}(x)^\intercal H^{\frac{1}{2}}t+t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}t\right)dt \\
&= \mathbf{f}_j(x)+\int_{\mathbb{R}^p}K(t)\cdot t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}tdt
\end{aligned}$$

for some $\tilde{x}$ depending on x and $x-H^{\frac{1}{2}}t$. The second equality holds from the differentiability in Assumption 5.a and the last equality holds from the conditions i. and ii. given in Proposition 2. Lastly, from the condition iii. in Proposition 2 and the boundedness from Assumption 5.a,

$$\left|\int_{\mathbb{R}^p}K(t)\cdot t^\intercal H^{\frac{1}{2}\intercal}\frac{\mathbf{f}_j^{(2)}(\tilde{x})}{2}H^{\frac{1}{2}}tdt\right| \le \frac{p^2C}{2}\cdot\max_x\left\|H^{\frac{1}{2}\intercal}\mathbf{f}_j^{(2)}(x)H^{\frac{1}{2}}\right\|_F \le \frac{p^3C^2}{2}\cdot\left\|H^{\frac{1}{2}}\right\|_F^{\,2}.$$

The first inequality is from the condition iii. and the second inequality is from Assumption 5.a. Then,

$$\begin{aligned}
&\left|\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ij})\right)\right]\mathbf{E}_\lambda\left[\frac{1}{det(H)^{\frac{1}{2}}}K\left(H^{-\frac{1}{2}}(x-X_{ik})\right)\right]-\mathbf{f}_j(x)\mathbf{f}_k(x)\right| \\
&\le C_1\left\|H^{\frac{1}{2}}\right\|_F^{\,2}
\end{aligned}$$

with some $C_1 > 0$ that does not depend on λ_j or H. By extending this,

$$\begin{aligned}
&\left|\mathbf{E}_\lambda\left[\widehat{M}_{jk} - M_{jk}\right]\right| \\
&\leq \int_{\mathbb{R}^p} \left|\mathbf{E}_\lambda\left[\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{1j}\right)\right)}{det(H)^{\frac{1}{2}}}\right]\mathbf{E}_\lambda\left[\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{2k}\right)\right)}{det(H)^{\frac{1}{2}}}\right] - \mathbf{f}_j(x)\mathbf{f}_k(x)\right| w(x)dx \\
&\leq C_1 \left\|H^{\frac{1}{2}}\right\|_F{}^2.
\end{aligned}$$

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

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

where

$$A_{ii'} = \int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}\left(x - X_{ij}\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x - X_{i'k}\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx.$$

We have that for some $l \neq i'$,

$$\begin{aligned}
&\mathbf{E}_\lambda\left[{A_{ii'}}^2\right] \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} \frac{K\left(H^{-\frac{1}{2}}\left(x - x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x - x''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right)^2 \mathbf{f}_j(x')\mathbf{f}_k(x'')dx'dx'' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} K\left(t\right)\frac{K\big(t + H^{-\frac{1}{2}}(x' - x'')\big)}{det(H)^{\frac{1}{2}}}w\big(x' + H^{\frac{1}{2}}t\big)dt\right)^2 \mathbf{f}_j(x')\mathbf{f}_k(x'')dx'dx'' \\
&= \frac{1}{det(H)^{\frac{1}{2}}}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p} K\left(t\right)K\left(t + s\right)w\big(x'' + H^{\frac{1}{2}}(t + s)\big)dt\right)^2 \mathbf{f}_j\big(x'' + H^{\frac{1}{2}}s\big)\mathbf{f}_k(x'')dsdx''
\end{aligned}$$

by letting $x = x' + H^{\frac{1}{2}}t$ and $x' = x'' + H^{\frac{1}{2}}s$ and

$$
\begin{aligned}
&\mathbf{E}_\lambda\left[A_{ii'}A_{il}\right] \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x-x''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'\right)\right)}{det(H)^{\frac{1}{2}}}\frac{K\left(H^{-\frac{1}{2}}\left(x-x'''\right)\right)}{det(H)^{\frac{1}{2}}}w(x)dx\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx''' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)\frac{K\left(t+H^{-\frac{1}{2}}(x'-x'')\right)}{det(H)^{\frac{1}{2}}}w\big(x'+H^{\frac{1}{2}}t\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)\frac{K\left(t+H^{-\frac{1}{2}}(x'-x''')\right)}{det(H)^{\frac{1}{2}}}w\big(x'+H^{\frac{1}{2}}t\big)dt\right)\mathbf{f}_j(x')\mathbf{f}_k(x'')\mathbf{f}_k(x''')dx'dx''dx''' \\
&= \frac{1}{det(H)^{\frac{1}{2}}}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s\right)w\big(x''+H^{\frac{1}{2}}(t+s)\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s+H^{-\frac{1}{2}}(x''-x''')\right)w\big(x''+H^{\frac{1}{2}}(t+s)\big)dt\right) \\
&\quad\cdot\mathbf{f}_j\big(x''+H^{\frac{1}{2}}s\big)\mathbf{f}_k(x'')\mathbf{f}_k(x''')dsdx''dx''' \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s\right)w\big(x'''+H^{\frac{1}{2}}(t+s+u)\big)dt\right) \\
&\quad\cdot\left(\int_{\mathbb{R}^p}K\left(t\right)K\left(t+s+u\right)w\big(x'''+H^{\frac{1}{2}}(t+s+u)\big)dt\right) \\
&\quad\cdot\mathbf{f}_j\big(x''+H^{\frac{1}{2}}s\big)\mathbf{f}_k\big(x'''+H^{\frac{1}{2}}u\big)\mathbf{f}_k(x''')dsdudx'''
\end{aligned}
$$

by letting $x = x' + H^{\frac{1}{2}}t$, $x' = x'' + H^{\frac{1}{2}}s$ and $x'' = x''' + H^{\frac{1}{2}}u$. Thus, with some constant $C_2 > 0$ that does not depend on λ_j or λ_k, $\mathrm{Var}_\lambda(A_{ii'}) \leq C_2/det(H)^{\frac{1}{2}}$ and $|\mathrm{Cov}_\lambda(A_{ii'}, A_{il})| \leq C_2$ and

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

Since $\min_j N_j det(H)^{\frac{1}{2}} \to \infty$ and $\min_j N_j \left\|H^{\frac{1}{2}}\right\|_F^4 = O(1)$ as $J \to \infty$, we have

$$\sum_{j=1}^{J}\sum_{k=1}^{J} \mathbf{E}_\lambda \left[\left(\widehat{M}_{jk} - M_{jk}\right)^2\right] = O\left(\frac{J^2}{\min_j N_j}\right)$$

$$\left\|\widehat{M} - M\right\|_F = \left(\sum_{j=1}^{J}\sum_{k=1}^{J}\left(\widehat{M}_{jk} - M_{jk}\right)^2\right)^{\frac{1}{2}} = O_p\left(\frac{J}{\sqrt{\min_j N_j}}\right)$$

Step 3. Lastly, given the rate on $\left\|\widehat{M} - M\right\|_F$, the convergence rate on $\left\|\tilde{\Lambda} - \widehat{\Lambda}\right\|_F$ is obtained by applying Lemma A.1.b of Kneip and Utikal (2001), as in Theorem 1.b of Kneip and Utikal (2001).

Firstly, let $\hat{V}_r$ denote the r-the largest eigenvalue of $\widehat{M}$; $\hat{V}_r$ is an estimate of V_r, as defined in Assumption 5. Note that $V_r = 0$ for $\rho < r \leq J$. Also, let $\hat{q}_r$ denote the (orthonormal) eigenvector of $\widehat{M}$ associated with the r-th eigenvalue and similarly for q_r. Recall that

$$\widehat{\Lambda} = \sqrt{J}\hat{Q}^\intercal = \sqrt{J}\begin{pmatrix}\hat{q}_1 & \cdots & \hat{q}_\rho\end{pmatrix}^\intercal$$

$$\tilde{\Lambda} = \sqrt{J}Q^\intercal = \sqrt{J}\begin{pmatrix}q_1 & \cdots & q_\rho\end{pmatrix}^\intercal$$

$$I_J = \begin{pmatrix}q_1 & \cdots & q_J\end{pmatrix}\begin{pmatrix}q_1{}^\intercal \\ \vdots \\ q_J{}^\intercal\end{pmatrix} = \sum_{r=1}^{J} q_r q_r{}^\intercal$$

For some $r \leq \rho$,

$$\hat{q}_r = \left(q_r q_r{}^\intercal + \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal\right)\hat{q}_r = \left(q_r{}^\intercal \hat{q}_r\right) q_r + \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r.$$

Since $\hat{q}_r^\intercal \hat{q}_r = {q_r}^\intercal q_r = 1$, we have $1 = ({q_r}^\intercal \hat{q}_r)^2 + \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$. Thus,

$$
\begin{aligned}
{q_r}^\intercal \hat{q}_r &= \pm \left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r\right)^{\frac{1}{2}}, \\
\hat{q}_r - q_r &= \left(\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r\right)^{\frac{1}{2}} - 1\right) q_r + \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r.
\end{aligned}
$$

The second equality holds by changing signs of $\hat{q}_r$ and q_r so that ${q_r}^\intercal \hat{q}_r > 0$. Note that RHS will be zero when $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r = 0$ and $\sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$ is a zero vector.

Firstly, let us find a bound on $\sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r$. Note that

$$
\begin{aligned}
\left(M - V_r I_J\right) \hat{q}_r &= \left(\widehat{M} - \left(\widehat{M} - M\right) - V_r I_J\right) \hat{q}_r \\
&= \left(\hat{V}_r - V_r\right) \hat{q}_r - \left(\widehat{M} - M\right) \hat{q}_r
\end{aligned}
$$

since $\hat{V}_r$ is the corresponding eigenvalue of $\widehat{M}$ for $\hat{q}_r$. Let $S_r = \sum_{r' \neq r} \frac{1}{V_{r'} - V_r} q_{r'} {q_{r'}}^\intercal$. S_r is well-defined from Assumption 5.b. By multiplying S_r to the equality above, we get

$$
\begin{aligned}
S_r \left(\left(\hat{V}_r - V_r\right) \hat{q}_r - \left(\widehat{M} - M\right) \hat{q}_r\right) &= S_r \left(M - V_r I_j\right) \hat{q}_r \\
&= S_r \left(\sum_{r'=1}^{\rho} V_{r'} q_{r'} {q_{r'}}^\intercal - V_r I_j\right) \hat{q}_r \\
&= \left(\sum_{r' \neq r} \frac{V_{r'}}{V_{r'} - V_r} q_{r'} {q'_r}^\intercal - \sum_{r' \neq r} \frac{V_r}{V_{r'} - V_r} q_{r'} {q_{r'}}^\intercal\right) \hat{q}_r \\
&= \sum_{r' \neq r} q_{r'} {q_{r'}}^\intercal \hat{q}_r.
\end{aligned}
$$

Find that $\left|\hat{V}_r - V_r\right| \leq \left\|\widehat{M} - M\right\|_{Ind,2} \leq \|\widehat{M} - M\|_F$ (see Chapter 8 Theorem 9 of Bellman

(1997)). $\|\cdot\|_{Ind,2}$ denotes the matrix norm induced by the vector norm $\|\cdot\|_2$. Also,

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

Using the two inequalities, we get

$$\begin{aligned}
\Big\| \sum_{r' \neq r} q_{r'} q_r{}^\intercal \hat{q}_r \Big\|_2 &\leq \left| \hat{V}_r - V_r \right| \|S_r \hat{q}_r\|_2 + \left\| S_r \left(\widehat{M} - M \right) \hat{q}_r \right\|_2 \\
&\leq \left\| \widehat{M} - M \right\|_F \|S_r\|_{Ind,2} \|\hat{q}_r\|_2 + \|S_r\|_{Ind,2} \left\| \widehat{M} - M \right\|_{Ind,2} \|\hat{q}_r\|_2 \\
&\leq \frac{2 \|\widehat{M} - M\|_F}{\min_{r' \neq r} |V_{r'} - V_r|} \\
&= \frac{1}{J} O_p \left(\frac{J}{\sqrt{\min_j N_j}} \right) = O_p \left(\frac{1}{\sqrt{\min_j N_j}} \right).
\end{aligned}$$

The last equality holds from Assumption 5.b: $\frac{\min_{r' \neq r} |V_{r'} - V_r|}{J}$ converges to a nonzero constant in probability. $\sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ converges to a zero vector, when $\min_j N_j$ goes to infinity.

Secondly, let us put a bound on $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ to show that $\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r \right)^{\frac{1}{2}}$ converges to one. The convergence of $\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}{}^\intercal \hat{q}_r$ to zero directly follows from the

convergence above:

$$\hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r = \left(\sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r$$
$$= \| \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \|_2^2 = O_p \left(\frac{1}{\min_j N_j} \right).$$

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

$$\left\| \left(\left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^{\frac{1}{2}} - 1 \right) q_r \right\|_2 \leq \left| \left(1 - \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r \right)^{\frac{1}{2}} - 1 \right|$$
$$\leq \hat{q}_r^\intercal \sum_{r' \neq r} q_{r'} q_{r'}^\intercal \hat{q}_r = O_p \left(\frac{1}{\min_j N_j} \right)$$

By combining the two bounds, we have

$$\|\hat{q}_r - q_r\|_2 = O_p \left(\frac{1}{\sqrt{\min_j N_j}} \right)$$

for $r \leq \rho$, by some sign change on $\hat{q}_r$. Accordingly,

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

References

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.

Bellman, Richard, *Introduction to matrix analysis*, SIAM, 1997.

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.

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.

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.

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.