Distributional Treatment Effect with Finite Mixture

Myungkou Shin*

March 9, 2024

Abstract

Treatment effect heterogeneity is of a great concern when evaluating the treatment. However, even with a simple case of a binary treatment, the distribution of treatment effect is difficult to identify due to the fundamental limitation that we cannot observe both treated potential outcome and untreated potential outcome for a given individual. This paper assumes a finite mixture model on the potential outcomes and a vector of control covariates to address treatment endogeneity and imposes a Markov condition on the potential outcomes and covariates within each type to identify the treatment effect distribution. The mixture weights of the finite mixture model are consistently estimated with a nonnegative matrix factorization algorithm, thus allowing us to consistently estimate the component distribution parameters, including ones for the treatment effect distribution.

Keywords: finite mixture model, quantile treatment effect, distributional treatment effect, nonnegative matrix factorization

JEL classification codes: C13

*Department of Economics, University of Oxford. email: myungkou.shin@economics.ox.ac.uk

1 Introduction

The estimation of the treatment effect distribution is an important but very difficult task even in a binary treatment setup due to the fact that in general we cannot simultaneously observe the two potential outcomes—treated potential outcome and untreated potential outcome—for a given individual, even when the treatment is random. Thus, instead of estimating the whole distribution of treatment effect, researcher often try to estimate some summary measures of the treatment effect, such as the average treatment effect (ATE) or the quantile treatment effect (QTE). These summary measures of the treatment effect distribution often provide insights into the treatment effect distribution and thus help researchers with policy recommendations. However, there still remain a lot of research questions that can only be answered from the treatment effect distributions; e.g., is the treatment Pareto improving? what is the share of people who are worse off under the treatment regime? This paper aims to answer these questions, by estimating the distributional treatment effect, under a finite mixture model with conditional Markov property.

In this paper, I assume that the distribution of the two potential outcomes and some control covariates follow a finite mixture model, with a fixed number of types K. The latent type variable models the treatment endogeneity; within each type, the potential outcome distribution does not depend on the treatment status. Given the finite mixture model, I use an instrument that shifts the mixture weights; using the instrument, I construct a grouping structure on individuals in a way that different groups differ in terms of their type shares/mixture weight and treatment status: let J_0 denote the number of untreated groups and J_1 the number of treated groups.

The existence of the control covariates is crucial to the identification of the finite mixture. Since the treated potential outcome and the untreated potential outcomes are fundamentally different, we cannot use the outcome variable to identify the common finite mixture model for treated groups and untreated groups. Instead, I use the control covariates to identify the finite mixture model. Note that we now have J_0 untreated groups and J_1 treated groups thanks to the instrument, which differ in terms of their type share/mixture weights. This difference in terms of the mixture weights will be reflected in their distribution of the control covariates. Under a full rank assumption on the component distribution for the control covariates, any two groups with the same distribution of the control covariates have the same mixture weights across K types; the distribution of the control covariates is informative about the mixture weights. If we additionally assume $J_0 + J_1 \geq K$ and the mixture weights also satisfy full rank condition, the group-specific distribution of the control

covariates partially identify the finite mixture model: for more discussion, see Henry et al. (2014).

Given some mixture weights and component distribution functions in the identified set, a counterfactual outcome distribution can be easily constructed by applying the mixture weights for a untreated group to the component distribution function for treated potential outcome or vice versa. A key observation of this paper is that the mixture weights and the component distribution functions do not need to be point identified to recover a counterfactual outcome distribution. Any pair of the mixture weights and the component distribution functions in the identified set constructs the same counterfactual outcome distribution.

Based on this observation, I construct a modified nonnegative matrix factorization minimization problem to estimate a pair of the mixture weights and the component distribution function. Though the solution to the minimization problem is not unique, induced weights for a counterfactual outcome distribution consistently estimates the counterfactual outcome distribution and thus the quantile treatment effects.

Recall that the goal of this paper is to estimate the distributional treatment effect. Though the nonnegative matrix factorization problem finds a pair of the mixture weights and the component distribution functions, it only finds the component distribution function for the treated potential outcome and that for the untreated potential outcome separately. Thus, to estimate the distribution of treatment effect, I impose additional assumption; the conditional distribution of the potential outcomes only depends on a subvector of the control covariates and that subvector of the control covariates is not independent of the rest of the control covariates. Given this 'Markov' property, I can use the two separate distributions—the joint component distribution of the control covariates and the treated potential outcome and the joint component distribution of the control covariates and the untreated potential outcome—to recover the conditional component distribution of the treated potential outcome given the untreated potential outcome.

When this Markov condition is assumed within the control covariates themselves, the mixture weights is point identified and estimated by adding some additional quadratic constraints to the nonnegative matrix factorization problem. Given the consistent estimates of the mixture weights, the treatment effect distribution is estimated through a maximum likelihood estimation.

The assumptions of this paper are most suitable in a panel data setup where the lagged outcome variables are used as control covariates, making a connection to the diff-in-diff literature. Consider a short panel where a subset of individuals is treated only for the last time period. Then, the approach of this paper to identify the finite mixture model by looking at the distribution of the control

covariates is essentially to use the lagged outcome distribution to learn about the heterogeneity across groups defined with the instrument. Though the untreated potential outcomes for a treated group is not directly observed, we are assuming that the heterogeneity, i.e., the mixture weights, that we learn from the lagged outcome is persistent to the last period as well; the same principle as in the diff-in-diff. In this setup, the Markov condition assumes that the conditional distribution of the last period potential outcome only depends on a subset of the most recent lagged outcomes and the latent type. Thus, the lagged outcomes that are further away only affect the last period potential outcome through the subset of the most recent lagged outcomes.

This paper relates to the recent development of the proxy variable approach: Deaner (2023); Hu and Schennach (2008); Miao et al. (2018). This paper has two identification result: the identification of the counterfactual outcome distribution and the identification of the treatment effect distribution. The first identification result is directly implied by the previous works such as Deaner (2023); Miao et al. (2018). Kedagni (2023) has a similar identification result in the context of binary instrument and use the subpopulation of compliers, always-takers, etc, as the types. To my best knowledge, the second identification argument, the identification of the treatment effect distribution, is new to the literature.

In addition, this paper contributes to the program evaluation literature, especially on the estimation of the quantile treatment effect and the distributional treatment effect. For quantile treatment effect, Callaway and Li (2019) also discusses a short panel and assumes copular stability and independence between first-differenced potential outcome and treatment. For distributional treatment effect, Frandsen and Lefgren (2021) build on Fan and Park (2010) and assumes stochastic monotonicity to put bounds on the distributions of treatment effect; the distributional treatment effect is partially identified. Kaji and Cao (2023) develop two new summarizing measure of the treatment effect distribution and develop partial identification results on the two measures. Noh (2023) assumes conditional independence of treatment effect and untreated potential outcome and uses deconvolution to point identify the treatment effect distribution. This paper also discusses point identification of the treatment effect distribution, under an arguably more flexible setup, relying on a finite mixture model and restrictions applied only to the type-specific distributions.

The rest of the paper is organized as follows. Section 2 formally develops the finite mixture model. Section 3 discusses the two identification results the counterfactual outcome distribution and the treatment effect distribution. Section 4 explains the estimation procedure based on a nonnegative matrix factorization problem and proves the consistency of the mixture weights. Sec-

tion 5 applies the estimation procedure to the National Longitudinal Survey of Youth dataset and estimates quantile treatment effects of the disemployment.

2 Setup

An econometrican observes a dataset $\{Y_i, X_i, D_i, Z_i\}_{i=1}^n$ where $Y_i \in \mathcal{Y} \subset \mathbb{R}$, $X_i \in \mathcal{X} \subset \mathbb{R}^p$, $D_i \in \{0,1\}$ and $Z_i \in \mathcal{Z}$. Y_i is an outcome variable, X_i is a vector of control covariates, D_i is a binary treatment variable and Z_i is an instrument. The outcome Y_i is constructed with two potential outcomes.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0).$$

The treatment variable D_i is potentially endogenous. For notational brevity, let $W_i = \big(Y_i, X_i\big)$ and $W_i(d) = \big(Y_i(d), X_i\big)$. $W_i = W_i(d)$ when $D_i = d$.

To put restrictions on the treatment endogeneity, I assume that there is some latent type variable k_i such that the treatment is as good as random after conditioning on the latent type variable k_i and the instrument Z_i.

Assumption 1. (Unconfoundedness with a Latent Type)
There exists a latent type variable $k_i \in \{1, \cdots, K\}$ such that

$$\Big(Y_i(1), Y_i(0), X_i\Big) \perp\!\!\!\perp D_i \mid \big(k_i, Z_i\big).$$

Also, $\big(Y_i(1), Y_i(0), X_i, D_i, k_i, Z_i\big) \sim iid.$

Under Assumption 1, the conditional distribution of W_i given $\big(k_i, D_i = d, Z_i\big)$ is equal to the conditional distribution of $W_i(d)$ given $\big(k_i, Z_i\big)$. In addition, I assume the usual exclusion restriction condition for the instrument.

Assumption 2. (Exclusion Restriction)

$$\Big(Y_i(1), Y_i(0), X_i\Big) \perp\!\!\!\perp Z_i \mid k_i.$$

Assumption 2 assumes that the instrument Z_i does not affect the distribution of $\big(Y_i(1), Y_i(0), X_i\big)$ after conditioning on the latent type k_i. Lastly, I assume that there are K finite types.

Assumption 3. (FINITE SUPPORT) $k_i \in \{1, \cdots, K\}$.

Under Assumptions 1-3, the conditional distribution of $W_i(d)$ given $\big(D_i = d', Z_i = z\big)$ admits a finite mixture representation. Let $\mathbf{F}_d(\cdot|d', z)$ denote the conditional distribution of $W_i(d)$ given $\big(D_i = d', Z_i = z\big)$. Then,

$$\begin{aligned}
\mathbf{F}_d(w|d', z) &= \Pr\big\{W_i(d) \le w | D_i = d', Z_i = z\big\} \\
&= \sum_{k=1}^{K} \Pr\big\{W_i(d) \le w | k_i = k, D_i = d', Z_i = z\big\} \cdot \Pr\big\{k_i = k | D_i = d', Z_i = z\big\} \quad \cdots \text{ LIE} \\
&= \sum_{k=1}^{K} \Pr\{W_i(d) \le w | k_i = k, Z_i = z\} \cdot \Pr\big\{k_i = k | D_i = d', Z_i = z\big\} \quad \cdots \text{ Assumption 1} \\
&= \sum_{k=1}^{K} \Pr\{W_i(d) \le w | k_i = k\} \cdot \Pr\big\{k_i = k | D_i = d', Z_i = z\big\} \quad \cdots \text{ Assumption 2}
\end{aligned}$$

Note that $\mathbf{F}_d(\cdot|d', z)$ is identified when $d = d'$ and $\Pr\{D_i = d, Z_i = z\} > 0$. Let $\mathbf{G}_{dk}$ denote the conditional distribution of $W_i(d)$ given $k_i = k$. Note that $\mathbf{G}_{0k}$ and $\mathbf{G}_{1k}$ should be coherent in the sense that the marginal distribution of X_i should be the same; let $\mathbf{G}_k$ denote the conditional distribution of $(Y_i(1), Y_i(0), X_i)$ given $k_i = k$. $\mathbf{G}_{0k}$ and $\mathbf{G}_{1k}$ are induced from $\mathbf{G}_k$. Also, let $\lambda(k|d, z)$ denote the conditional probability of $k_i = k$ given $(D_i = d, Z_i = z)$. Then,

$$\mathbf{F}_d(w|d', z) = \sum_{k=1}^{K} \lambda(k|d', z) \cdot \mathbf{G}_{dk}(w).$$

Under Assumptions 1-3, the conditional distribution of $W_i(d)$ given $\big(D_i = d', Z_i = z\big)$ is decomposed into K mixture weights $\{\lambda(k|d', z)\}_{k=1}^{K}$ and K component distributions $\{\mathbf{G}_{dk}\}_k$.

The key assumption that will be maintained throughout the rest of the paper is that there is sufficient variation in the mixture weights and the mixture component distributions. This directly connects to the relevance of the instruments, meaning that the mixture weights $\lambda(\cdot|d, z)$ is not a trivial function of z. Below I present two empirical contexts where there is a natural choice for such an instrument and discuss what the relevance condition and Assumptions 1-2 mean in each context.

Example 1. *(clustered treatment)* Suppose that there exist an observed clustering structure in the dataset and the treatment is assigned at the cluster level. Then, I suggest using the clustering

structure as an instrument. Let

$$Z_i \in \{1, \cdots, J\} \quad \text{and} \quad D_i = D(Z_i).$$

There are J clusters in the dataset and the unit-level treatment status D_i equals the cluster-level treatment status $D(Z_i)$. Assumption 1 is trivially satisfied since D_i is a function of Z_i. Assumption 2 assumes that the cluster membership Z_i does not have any information on the unit-level heterogeneity, after conditioning on the unit-level latent type k_i. Lastly, the relevance condition assumes that the clusters are heterogeneous in terms of their unit-level type distributions.

Example 2. *(event-study)* Suppose that the dataset is now a panel data with multiple treatment timings: $\left\{\{Y_{it}\}_{t=1}^{T}, E_i\right\}_{i=1}^{n}$. $E_i \in \{2, \cdots, T, \infty\}$ is the treatment timing for unit i. In this canonical staggered-adoption event-study setup, we can use the treatment timing as an instrument. To that end, we first focus on a subpopulation of units that were not treated immediately; fix some time period $t^* > 1$ and drop units such that $E_i < t^*$. Let

$$Y_i = Y_{it^*}, \quad X_i = \{Y_{it}\}_{t=1}^{t^*-1}, \quad D_i = \mathbf{1}\{E_i = t^*\} \quad \text{and} \quad Z_i = E_i.$$

Units that are not treated until $t = t^*$ are coded as 'untreated' and units that are treated at $t = t^*$ are coded as 'treated.' Again, Assumption 1 trivially holds since D_i is a function of $E_i = Z_i$. Assumption 2 assumes that the treatment timing is as good as random after conditioning on the unit-level latent type k_i. The relevance condition assumes that the earlier-treated units and the later-treated units are heterogeneous in terms of their unit-level type distributions.

Though the two empirical examples above both use a discrete instrument Z_i, none of the theoretical results of this paper requires Z_i to be discrete. As long as the assumptions used in the following sections hold for a discretized version of Z_i, a continuous Z_i can be used. Thus, for notational brevity, I only consider discrete Z_i; whenever the original instrument variable in the dataset is continuous, an appropriate discretization is implicitly imposed. Specifically, I consider discretization where $Z_i \in \{1, \cdots, J\}$ with some $J = J_0 + J_1 \geq K$ and $D_i = 0 \Leftrightarrow Z_i \leq J_0$. Thus, the instrument Z_i forms a grouping structure over units so that there are J_0 untreated groups and J_1 treated groups; each group has nonzero measure. For the rest of the paper, I drop D_i in the

conditioning set in modelling the finite mixture: for $d \in \{0,1\}$ and $j \in \{1, \cdots, J\}$,

$$\mathbf{F}_d(w|j) = \sum_{k=1}^{K} \lambda(k|j) \cdot \mathbf{G}_{dk}(w) \tag{1}$$

$\mathbf{F}_d(\cdot|j)$ is directly observed from the data when $d = 0, j \leq J_0$ or $d = 1, j > J_0$. Also, I call subpopulation $\{i : Z_i = j\}$ the group j.

3 Identification

In this section, I discuss identification of two objects of interest: distribution of counterfactual outcome and distribution of treatment effect. Depending on the context, the counterfactual outcome refers to treated potential outcome for untreated units, or untreated potential outcome for treated units. For example, $\mathbf{F}_0(\cdot|j)$ for some $j > J_0$ is a distribution of counterfactual outcome where the counterfactual is the treated group j being untreated.

The first identification result, which is for distribution of counterfactual outcome, is tantamount to those of Miao et al. (2018); Deaner (2023); under some relevance condition on the instrument and sufficient variation condition on the component distributions, the marginal distribution is point identified. The second identification result, which requires identifying the joint distribution of potential outcomes, is new to the literature to my best knowledge. For the second identification result, I further assume that a subvector of X works as an instrument for the rest of X. Then, the joint distribution of untreated and treated potential outcome is identified from the conditional distributions of the potential outcomes.

3.1 Identification of counterfactual outcome distribution

In this subsection, I discuss identification of counterfactual outcome distribution. For example, suppose we are interested in $\mathbf{F}_0(\cdot|j)$ for some $j > J_0$: untreated potential outcome distribution for the treated group j. Once we identify $\mathbf{F}_0(\cdot|j)$, we can identify various treatment effect parameters: e.g.,

$$\begin{aligned} CATT(j) &= \mathbf{E}\left[Y_i(1) - Y_i(0) | D_i = 1, Z_i = j\right], \\ CQTT(j,\tau) &= \mathbf{F}_1^{-1}(\tau|j) - \mathbf{F}_0^{-1}(\tau|j). \end{aligned}$$

$CATT(j)$ is the conditional average treatment effect on the treated group j and $CQTT(j,\tau)$ is the conditional quantile treatment effect on the treated group j, for the τ-th quantile.

The following is a modification of Miao et al. (2018) so that X_i is allowed to be discrete, continuous or mixed random vector. The identification relies on two invertible matrices. Firstly, Assumption 4 assumes sufficient variation across types, in the domain of X_i. Let $\mathbf{G}_{x1}, \cdots, \mathbf{G}_{xK}$ denote the marginal distributions of X_i, constructed from $\mathbf{G}_1, \cdots, \mathbf{G}_K$, respectively.

Assumption 4. (RANK) *There exist some $x_1, \cdots, x_K \in \mathcal{X}$ such that the $K \times K$ matrix*

$$\Gamma_x = \begin{pmatrix} \mathbf{G}_{x1}(x_1) & \cdots & \mathbf{G}_{xK}(x_1) \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{x1}(x_K) & \cdots & \mathbf{G}_{xK}(x_K) \end{pmatrix}$$

is invertible.

Assumption 4 assumes that there is sufficient variation in the marginal distributions of X_i given k_i so that it preserves the variation in the mixture weights when we look at the marginal distributions of X_i given Z_i. Assumption 4 is plausible in the context of the event-study setup in Example 2, if we believe that the variation in the outcome distribution would be preserved in the lagged outcome distribution as well.

Secondly, Assumption 5 assumes that instrument Z_i sufficiently shifts the mixture weights $\lambda(k|j) = \Pr\{k_i = k | Z_i = j\}$.

Assumption 5. (RELEVANCE)

(i) *The $K \times J_0$ matrix*

$$\Lambda_0 = \begin{pmatrix} \lambda(1|1) & \cdots & \lambda(1|J_0) \\ \vdots & \ddots & \vdots \\ \lambda(K|1) & \cdots & \lambda(K|J_0) \end{pmatrix}$$

has rank K.

(ii) *The $K \times J_1$ matrix*

$$\Lambda_1 = \begin{pmatrix} \lambda(1|J_0+1) & \cdots & \lambda(1|J) \\ \vdots & \ddots & \vdots \\ \lambda(K|J_0+1) & \cdots & \lambda(K|J) \end{pmatrix}$$

has rank K.

Note that Assumption 5-(i) only assumes relevance of Z_i among untreated units while Assumption 5-(ii) only assumes relevance of Z_i among treated units. To identify $\mathbf{F}_0(\cdot|j)$ for some $j > J_0$, I only use Assumption 5-(i); it suffices to have full rank on the mixture weights matrix for untreated units to identify a counterfactual distribution of untreated potential outcome.

The following lemma establishes that the counterfactual outcome distribution could be identified through the identified set of the mixture weights and the component distributions.

Lemma 1. *Consider an identified set of* $\left(\{\mathbf{G}_{0k}, \mathbf{G}_{1k}\}_{k=1}^K, \Lambda_0, \Lambda_1\right)$ *that satisfies the following, as in Henry et al. (2014):*

1. $\{\mathbf{G}_{0k}, \mathbf{G}_{1k}\}_{k=1}^K$ *are component distribution functions on* $\mathcal{Y} \times \mathcal{X}$*:*
   $\{\mathbf{G}_{0k}, \mathbf{G}_{1k}\}_{k=1}^K$ *are monotone increasing, right-continuous and have left limit of zero and right limit of one. Also, for any* $x \in \mathcal{X}$*,*

$$\int_{\mathcal{Y}} \mathbf{G}_{0k}(y, x)dy = \int_{\mathcal{Y}} \mathbf{G}_{1k}(y, x)dy \quad \forall k = 1, \cdots, K.$$

2. Λ_0 *is a* $K \times J_0$ *mixture weight matrix and* Λ_1 *is a* $K \times J_1$ *mixture weight matrix:*
   Λ_0 *and* Λ_1 *are nonnegative and each column of* Λ_0 *and* Λ_1 *sum to one.*

3. *For any* $w \in \mathcal{Y} \times \mathcal{X}$*,*

$$\sum_{k=1}^K \lambda(k|j) \cdot \mathbf{G}_{0k}(w) = \Pr\{W_i \leq w | Z_i = j\} \quad \forall j = 1, \cdots, J_0,$$
$$\sum_{k=1}^K \lambda(k|j) \cdot \mathbf{G}_{1k}(w) = \Pr\{W_i \leq w | Z_i = j\} \quad \forall j = J_0 + 1, \cdots, J.$$

Assumptions 1-4 hold. For any two tetrads $\left(\{\mathbf{G}'_{0k}, \mathbf{G}'_{1k}\}_{k=1}^K, \Lambda'_0, \Lambda'_1\right), \left(\{\mathbf{G}''_{0k}, \mathbf{G}''_{1k}\}_{k=1}^K, \Lambda''_0, \Lambda''_1\right)$ *in the identified set, Assumption 5-(i) implies*

$$\sum_{k=1}^K \lambda(k|j)' \cdot \mathbf{G}'_{0k} = \sum_{k=1}^K \lambda(k|j)'' \cdot \mathbf{G}''_{0k} \quad \forall j = J_0 + 1, \cdots, J$$

and Assumption 5-(ii) implies

$$\sum_{k=1}^K \lambda(k|j)' \cdot \mathbf{G}'_{1k} = \sum_{k=1}^K \lambda(k|j)'' \cdot \mathbf{G}''_{1k} \quad \forall j = 1, \cdots, J_0.$$

Proof. Direct from identification argument as in Miao et al. (2018). See Appendix. □

In general, the identified set is not a singleton. However, building on the identification argument as seen in Miao et al. (2018); Deaner (2023), Lemma 1 states that the extrapolation for $\mathbf{F}_0(\cdot|j)$ for $j > J_0$ using the identified set is unique when Assumption 5-(i) holds and therefore $\mathbf{F}_0(\cdot|j)$ is identified through the identified set of $\left(\{\mathbf{G}_{0k}\}_{k=1}^K, \Lambda_0, \Lambda_1\right)$ as well.

Remark 1. Given $\{x_k\}_{k=1}^K$ satisfying Assumption 4 and Assumption 5-(i), we can construct a linear coefficient ψ such that the distribution of untreated potential outcome for the treated group j is a linear combination of the (observed) distributions of the untreated groups $\{\mathbf{F}_0(\cdot|j)\}_{j=1}^{J_0}$. Let $\mathbf{F}_x(\cdot|j)$ denote the conditional distribution of X_i given $Z_i = j$ and let

$$\psi = \begin{pmatrix} \mathbf{F}_x(x_1|1) & \cdots & \mathbf{F}_x(x_K|1) \\ \vdots & \ddots & \vdots \\ \mathbf{F}_x(x_1|J_0) & \cdots & \mathbf{F}_x(x_K|J_0) \end{pmatrix} \left(\begin{pmatrix} \mathbf{F}_x(x_1|1) & \cdots & \mathbf{F}_x(x_1|J_0) \\ \vdots & \ddots & \vdots \\ \mathbf{F}_x(x_K|1) & \cdots & \mathbf{F}_x(x_K|J_0) \end{pmatrix} \begin{pmatrix} \mathbf{F}_x(x_1|1) & \cdots & \mathbf{F}_x(x_K|1) \\ \vdots & \ddots & \vdots \\ \mathbf{F}_x(x_1|J_0) & \cdots & \mathbf{F}_x(x_K|J_0) \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{F}_x(x_1|j) \\ \vdots \\ \mathbf{F}_x(x_K|j) \end{pmatrix}$$

for some $j > J_0$. Then, for any $w \in \mathcal{Y} \times \mathcal{X}$,

$$\mathbf{F}_0(w|j) = \psi^\intercal \begin{pmatrix} \mathbf{F}_0(w|1) \\ \vdots \\ \mathbf{F}_0(w|J_0) \end{pmatrix}.$$

The linear coefficient ψ always sum to one but is not always nonnegative.

3.2 Identification of treatment effect distribution

The two marginal distributions of potential outcomes provide much information about treatment effect heterogeneity; we can look at the quantile treatment effect and see how the treatment disproportionately affects the distribution of outcome. However, there still remain a plenty of questions which cannot be answered by the two marginal distributions: is the treatment Pareto improving?; what is the proportion of individuals that are better off under the treatment? does the treatment benefit individuals who would have been at the bottom of the counterfactual distribution?

To answer these question, I impose an additional assumption on X_i.

Assumption 6. (Markov Property)

(i) $X_i = (X_{i1}^\intercal, X_{i2}^\intercal)^\intercal$ *such that* $X_{i1} \in \mathcal{X}_1$, $X_{i2} \in \mathcal{X}_2$ *and*

$$\Big(Y_i(1), Y_i(0)\Big) \perp\!\!\!\perp X_{i1} \;\Big|\; (k_i, X_{i2})\,.$$

(ii) $X_i = (X_{i0}^\intercal, X_{i1}^\intercal, X_{i2}^\intercal)^\intercal$ *such that* $X_{i0} \in \mathcal{X}_0$, $X_{i1} \in \mathcal{X}_1$, $X_{i2} \in \mathcal{X}_2$ *and*

$$\begin{aligned}\Big(Y_i(1), Y_i(0)\Big) \;\Big|\; (k_i, X_i) &\overset{d}{\equiv} \Big(Y_i(1), Y_i(0)\Big) \;\Big|\; (k_i, X_{i2})\,,\\ X_{i2} \;\Big|\; (k_i, X_{i0}, X_{i1}) &\overset{d}{\equiv} X_{i2} \;\Big|\; (k_i, X_{i1})\,,\end{aligned}$$

Assumption 6-(i) divides X_i into two parts, X_{i1} and X_{i2}, and assumes that the conditional distribution of $\big(Y_i(1), Y_i(0)\big)$ given $\big(k_i, X_i\big)$ only depends on $\big(k_i, X_{i2}\big)$. Assumption 6-(ii) assumes that X_i can be decomposed into three parts and additionally assumes that the conditional distribution of X_{i2} given $\big(k_i, X_{i0}, X_{i1}\big)$ only depends on $\big(k_i, X_{i1}\big)$. In the context of the event-study setup in Example 2, Assumption 6 is equivalent with assuming that the outcome process satisfy the Markov property, conditioning on the type k_i. Under some additional conditions, Assumption 6-(i) gives us a partial identification result on the distribution of treatment effect and Assumption 6-(ii) gives us a point identification result.

An important observation to be made here is that all of the conditional independence statements in Assumption 6 have the latent type variable k_i as a conditioning variable; I am not assuming that the control covariates X_i satisfy some Markov property marginally.

Let us discuss the partial identification result first. Suppose Assumptions 1-5 hold; for more discussion, see Henry et al. (2014). Then, the component distribution functions $\{\mathbf{G}_{0k}, \mathbf{G}_{1k}\}_{k=1}^K$ in the mixture model (1) are partially identified. Fix an arbitrary $\{\mathbf{G}_{0k}, \mathbf{G}_{1k}\}_{k=1}^K$ in the identified set and construct

$$\begin{aligned}G_{1|x_1,k}(y|x) &= \Pr\{Y_i(1) \le y | k_i = k, X_{i1} = x\}\,,\\ \Phi_k(u|y, x_2) &= \Pr\{Y_i(1) - Y_i(0) \le u | Y_i(0) = y, k_i = k, X_{i2} = x_2\}\,,\\ G_{0x_2|x_1,k}(y, x_2|x_1) &= \Pr\{Y_i(0) \le y, X_{i2} \le x_2 | k_i = k, X_{i1} = x_1\}\end{aligned}$$

based on the component distributions. Let $g_{1|x_1,k}, \phi_k, g_{0x_2|x_1,k}$ be the corresponding densities. Then,

$$\begin{aligned}
G_{1|x_1,k}(y_1|x_1) &= \mathbf{E}\left[\Pr\left\{Y_i(1) \le y_1 | Y_i(0), k_i = k, X_{i1} = x_1, X_{i2}\right\} | k_i = k, X_{i1} = x_1\right] \\
&= \mathbf{E}\left[\Pr\left\{Y_i(1) - Y_i(0) \le y_1 - Y_i(0) | Y_i(0), k_i = k, X_{i2}\right\} | k_i = k, X_{i1} = x_1\right] \\
&= \mathbf{E}\left[\Phi_k(y_1 - Y_i(0) | Y_i(0), X_{i2}) | k_i = k, X_{i1} = x_1\right] \\
&= \int_{\mathcal{X}_2} \int_{\mathcal{Y}} \Phi_k(y_1 - y_0 | y_0, x_2) g_{0x_2|x_1,k}(y_0, x_2 | x_1) dy_0 dx_2.
\end{aligned}$$

The second equality holds when we assume Assumption 6-(i). Then,

$$g_{1|x_1,k}(y_1|x_1) = \int_{\mathcal{Y}\times\mathcal{X}_2} \phi_k(y_1 - y_0 | y_0, x_2) g_{0x_2|x_1,k}(y_0, x_2 | x_1) d\left(y_0, x_2\right). \tag{2}$$

After fixing y_1, $g_{1|x_1,k}$ can be written as an integral transform of ϕ_k with $g_{0x_2|x_1,k}$ as the integral kernel. By 'inverting' the integral operator, we retrieve ϕ_k, the conditional distribution of $Y_i(1) - Y_i(0)$ given $\left(Y_i(0), k_i = k, X_{i2}\right)$. Consider a simple case where Y_i, X_i are discrete random variables and $|\mathcal{X}_1| \ge |\mathcal{Y} \times \mathcal{X}_2|$. When

$$\begin{pmatrix}
g_{0x_2|x_1,k}(w_1|x_1) & \cdots & g_{0x_2|x_1,k}(w_1|x_{|\mathcal{X}_1|}) \\
\vdots & \ddots & \vdots \\
g_{0x_2|x_1,k}(w_{|\mathcal{Y}\times\mathcal{X}_2|}|x_1) & \cdots & g_{0x_2|x_1,k}(w_{|\mathcal{Y}\times\mathcal{X}_2|}|x_{|\mathcal{X}_1|})
\end{pmatrix}$$

has full rank, ϕ_k is identified from multiplying its pseudo-inverse to Equation (2). Note that the outcome depends on the input distributions, which come from the partially identified component distributions $\mathbf{G}_{dk}$. The following assumptions extend the full rank condition to continuous X_i and assumes completeness on $G_{0x_2|x_1,k}$.

Assumption 7. (Bounded Density) *There exist some weighting function* $\xi_1 : \mathcal{X}_1 \to (0, \infty)$ *and* $\xi_2 : \mathcal{Y} \times \mathcal{X}_2 \to (0, \infty)$ *such that*

$$\begin{aligned}
\int_{\mathcal{X}_1} g_{1|x_1,k}(y|x)\xi_1(x)dx < \infty \quad \forall k = 1, \cdots, K \text{ and } y \in \mathcal{Y}, \\
\int_{\mathcal{Y}\times\mathcal{X}_2} \phi_k(y_1 - y_0 | y_0, x)\xi_2(y_0, x) d\left(y_0, x\right) < \infty \quad \forall k = 1, \cdots, K \text{ and } y_1 \in \mathcal{Y},
\end{aligned}$$

Assumption 8. (Completeness) *For each* $k = 1, \cdots, K$,

$$\int_{\mathcal{Y}\times\mathcal{X}_2} |\phi(y,x)|\xi_2(y,x)d\,(y,x) < \infty,$$
$$\int_{\mathcal{Y}\times\mathcal{X}_2} \phi(y,x_2)g_{0x_2|x_1,k}(y,x_2|x_1)d\,(y,x_2) = 0 \quad \forall x_1 \in \mathcal{X}_1$$

implies $\phi(y,x_2) = 0$ *for all* $(y,x_2) \in \mathcal{Y}\times\mathcal{X}_2$.

Then, $\{\phi_k\}_{k=1}^K$ and therefore the conditional distribution of $Y_i(1) - Y_i(0)$ given k_i are partially identified when Assumptions 1-5, 6-(i) and 7-8 hold, thanks to the partial identification result in Henry et al. (2014).

For point identification, I impose additional restriction on the shape of $\mathbf{G}_k$; $\mathbf{G}_{xk}$, the observed part of $\mathbf{G}_k$, satisfy the Markov property: Assumption 6-(ii). Let $g_{x_2|x_1,k}$ denote the conditional density of X_{i2} given $\left(k_i = k, X_{i1}\right)$ and let $g_{x_0|x_1,k}$ denote the conditional density of X_{i0} given $\left(k_i = k, X_{i1}\right)$.

Assumption 9. (Sufficient Variation in X_i) *There exist some set* $\tilde{\mathcal{X}}_1 \subset \mathcal{X}_1$ *such that*

(i) *For any* $k \neq k'$, *there exist some* $(x_1,x_2) \in \tilde{\mathcal{X}}_1 \times \mathcal{X}_2$ *such that* $g_{x_2|x_1,k}(x_2|x_1) \neq g_{x_2|x_1,k'}(x_2|x_1)$.

(ii) *For any* $x_1 \in \tilde{\mathcal{X}}_1$,

- $g_{x_0|x_1,1}(\cdot|x_1), \cdots, g_{x_0|x_1,K}(\cdot|x_1)$ *are linearly independent;*
- $g_{x_0|x_1,1}(\cdot|x_1), \cdots, g_{x_0|x_1,K}(\cdot|x_1)$ *are continuously differentiable;*
- *there is some* $x \in \mathcal{X}_0$ *such that* $\lim_{x_0\to x} g_{x_0|x_1,k}(x_0|x_1) = 0$ *for* $k = 1, \cdots, K$.

Assumption 9 assumes that there is some set $\tilde{\mathcal{X}}_1$ in $\mathcal{X}_1$ such that the conditional distribution of X_{i2} given X_{i1} and the conditional distribution of X_{i0} given X_{i1} show sufficient variations. Assumption 9 ensures that the identified set with this additional restriction is a singleton.

Theorem 1. *Assumptions 1-9 hold. Then, the joint component distribution* $\{\mathbf{G}_k\}_{k=1}^K$ *and the mixture weight matrices* Λ_0 *and* Λ_1 *are point identified.*

Proof. See Appendix.

□

4 Implementation

In this section, I discuss how to estimate the distribution of counterfactual outcome and the distribution of treatment effect, based on the identification result. In estimating the distribution

of counterfactual outcome, I rely on Assumption 3 and use the nonnegative matrix factorization (NMF), to directly estimate the component distribution functions of the mixture model. Though the solution to the minimization problem is not unique, I show that the induced linear coefficients $\hat{\psi}$ converges to the true ψ as defined in Remark 1 and thus the estimator for the counterfactual outcome distribution is consistent. In estimating the distribution of treatment effect, I further impose the Markov property in the first step of the NMF. Then, given the estimated component distribution functions, I use MLE to estimate the treatment effect distribution.

4.1 Counterfactual outcome distribution

When $|\mathcal{X}| = K$, using the sample analogue of the weights ψ defined in Remark 1 is a straightforward way of estimating the distribution of counterfactual potential outcomes. However, when $|\mathcal{X}| > K$, it is often implausible to assume that there is only one set of $(x_1, \cdots, x_K)$ which Assumption 4 holds and the researcher knows that specific set. Thus, relying on Lemma 1, I suggest estimating the weights ψ through the estimation of the mixture weights Λ and the component distributions $\mathbf{G}_{dk}$.

For that, I firstly partition $\mathcal{X}$ into M_x sets: $\{A_m\}_{m=1}^{M_x}$. For each set in the partition $\{A_m\}_{m=1}^{M_x}$, I construct a partition on $\mathcal{Y}$: $\{B_{mm'}\}_{m'=1}^{L_m}$. Let $M = \sum_{m=1}^{M_x} L_m$. $\{C_m\}_{m=1}^{M} := \{A_m \times B_{mm'}\}_{m,m'}$ forms a well-defined partition on $\mathcal{Y} \times \mathcal{X}$. $\{C_m\}_{m=1}^{M}$ is ordered lexicographically; i.e., $C_1 = A_1 \times B_1^1, C_2 = A_1 \times B_2^1, \cdots$. Then, I construct a $M \times J_0$ matrix and a $M \times J_1$ matrix whose elements are conditional probability masses for each set of the partition given Z_i: let

$$\mathbb{H}_0 = \begin{pmatrix} \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_1, Z_i = 1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = 1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_1, Z_i = J_0\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J_0\}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_M, Z_i = 1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = 1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_M, Z_i = J_0\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J_0\}} \end{pmatrix},$$

$$\mathbb{H}_1 = \begin{pmatrix} \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_1, Z_i = J_0+1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J_0+1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_1, Z_i = J\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J\}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_M, Z_i = J_0+1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J_0+1\}} & \cdots & \frac{\sum_{i=1}^n \mathbf{1}\{W_i \in C_M, Z_i = J\}}{\sum_{i=1}^n \mathbf{1}\{Z_i = J\}} \end{pmatrix},$$

Each column of $\mathbb{H}_0$ is a (discretized) conditional distribution function of W_i given $\big(D_i = 0, Z_i\big)$ and each column of $\mathbb{H}_1$ is a (discretized) conditional distribution function of W_i given $\big(D_i = 1, Z_i\big)$. The

NMF solves the following minimization problem: with some $\gamma > 0$,

$$\min_{\Lambda_0,\Lambda_1,\Gamma_0,\Gamma_1} \|\mathbb{H}_0 - \Gamma_0\Lambda_0\|_F{}^2 + \|\mathbb{H}_1 - \Gamma_1\Lambda_1\|_F{}^2 \tag{3}$$

subject to

$$\begin{gathered}
\Lambda_0 \in \mathbb{R}_+{}^{K\times J_0}, \quad \Lambda_1 \in \mathbb{R}_+{}^{K\times J_1}, \quad \Gamma_0 \in \mathbb{R}_+{}^{M\times K}, \quad \Gamma_1 \in \mathbb{R}_+{}^{M\times K}, \\
\mathbf{1}_K{}^\intercal\Lambda_0 = \mathbf{1}_{J_0}{}^\intercal, \quad \mathbf{1}_K{}^\intercal\Lambda_1 = \mathbf{1}_{J_1}{}^\intercal, \quad \mathbf{1}_M{}^\intercal\Gamma_0 = \mathbf{1}_K{}^\intercal, \quad \mathbf{1}_M{}^\intercal\Gamma_1 = \mathbf{1}_K{}^\intercal, \\
P\Gamma_0 = P\Gamma_1
\end{gathered}$$

where P is a $M_x \times M$ matrix such that

$$\mathbb{I} = \begin{pmatrix} \mathbf{1}_{L_1}{}^\intercal & \mathbf{0}_{L_2}{}^\intercal & \cdots & \mathbf{0}_{L_{M_x}}{}^\intercal \\ \mathbf{0}_{L_1}{}^\intercal & \mathbf{1}_{L_2}{}^\intercal & \cdots & \mathbf{0}_{L_{M_x}}{}^\intercal \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{L_1}{}^\intercal & \mathbf{0}_{L_2}{}^\intercal & \cdots & \mathbf{1}_{L_{M_x}}{}^\intercal \end{pmatrix}.$$

Let $\widehat{\Lambda}_0$, $\widehat{\Lambda}_1$, $\widehat{\Gamma}_0$ and $\widehat{\Gamma}_1$ denote the solution to the minimization problem.

Note that the minimization problem (3) is a quadratic program when either (Λ_0, Λ_1) or (Γ_0, Γ_1) is fixed. Thus, I suggest an iterative algorithm to solve the minimization problem.

1. Initialize $\Gamma_0^{(0)}, \Gamma_1^{(0)}$.

2. *(Update Λ)* Given $\Gamma_0^{(0)}, \Gamma_1^{(0)}$, solve the following quadratic program:

$$\left(\Lambda_0^{(s)}, \Lambda_1^{(s)}\right) = \arg\min_{\Lambda_0,\Lambda_1} \left\|\mathbb{H}_0 - \Gamma_0^{(s)}\Lambda_0\right\|_F{}^2 + \left\|\mathbb{H}_1 - \Gamma_1^{(s)}\Lambda_1\right\|_F{}^2$$

subject to $\Lambda_0 \in \mathbb{R}_+{}^{K\times J_0}, \Lambda_1 \in \mathbb{R}_+{}^{K\times J_1}, \mathbf{1}_K{}^\intercal\Lambda_0 = \mathbf{1}_{J_0}{}^\intercal$ and $\mathbf{1}_K{}^\intercal\Lambda_1 = \mathbf{1}_{J_1}{}^\intercal$.

3. *(Update Γ)* Given $\left(\Lambda_0^{(s)}, \Lambda_1^{(s)}\right)$, solve the following quadratic program:

$$\left(\Gamma_0^{(s+1)}, \Gamma_1^{(s+1)}\right) = \arg\min_{\Gamma_0,\Gamma_1} \left\|\mathbb{H}_0 - \Gamma_0\Lambda_0^{(s)}\right\|_F{}^2 + \left\|\mathbb{H}_1 - \Gamma_1\Lambda_1^{(s)}\right\|_F{}^2$$

subject to $\Gamma_0 \in \mathbb{R}_+{}^{M\times K}, \Gamma_1 \in \mathbb{R}_+{}^{M\times K}, \mathbf{1}_M{}^\intercal\Gamma_0 = \mathbf{1}_K{}^\intercal, \mathbf{1}_M{}^\intercal\Gamma_1 = \mathbf{1}_K{}^\intercal$ and $P\Gamma_0 = P\Gamma_1$.

4. Repeat **2-3** until convergence.

As discussed in Lemma 1, a distribution of counterfactual outcomes is identified when either Assumption 6-(i) or Assumption 6-(ii) holds. For example, when Assumption 6-(i) holds, there are at least $J_0 \geq K$ untreated groups whose observed distributions contain nonzero shares of each of the K types. Thus, to initialize $\Gamma_0^{(0)}, \Gamma_1^{(1)}$, we can consider every K-combination of columns from $\mathbb{H}_0$ for small enough J_0 and choose the combination that minimizes the objective: $\binom{J_0}{K}$ initial values. Likewise, we can consider randomly drawn K set of weights that sum to one and use the weighted sums of columns of $\mathbb{H}_0$ as an initial value. Alternatively, we can select the eigenvectors associated with the first K largest eigenvalues of ${\mathbb{H}_0}^\intercal \mathbb{H}_0$ as an initial value. In the empirical example discussed in Section 5, the factorization result was stable across the choice of the initial value. Likewise, when $J_1 \leq K$ and we assume Assumption 6-(ii), we can use $\mathbb{H}_1$ to initialize $\Gamma_0^{(0)}, \Gamma_1^{(0)}$.

Note that

$$\begin{aligned}\begin{pmatrix}\mathbf{F}_0(\cdot|1) & \cdots & \mathbf{F}_0(\cdot|J_0)\end{pmatrix} &= \begin{pmatrix}\mathbf{G}_{01} & \cdots & \mathbf{G}_{0K}\end{pmatrix}\Lambda_0, \\ \begin{pmatrix}\mathbf{F}_0(\cdot|J_0+1) & \cdots & \mathbf{F}_0(\cdot|J)\end{pmatrix} &= \begin{pmatrix}\mathbf{G}_{01} & \cdots & \mathbf{G}_{0K}\end{pmatrix}\Lambda_1.\end{aligned}$$

When $(\Lambda_0 {\Lambda_0}^\intercal)^{-1}$ exists, we get

$$\begin{pmatrix}\mathbf{F}_0(\cdot|J_0+1) & \cdots & \mathbf{F}_0(\cdot|J)\end{pmatrix} = \begin{pmatrix}\mathbf{F}_0(\cdot|1) & \cdots & \mathbf{F}_0(\cdot|J_0)\end{pmatrix}{\Lambda_0}^\intercal \left(\Lambda_0 {\Lambda_0}^\intercal\right)^{-1}\Lambda_1.$$

Building on this observation, I estimate $\mathbf{F}_0(\cdot|j)$ for treated group $j > J_0$, as a linear combination of empirical distribution functions for untreated groups $\{\hat{\mathbf{F}}_0(\cdot|j)\}_{j=1}^{J_0}$, using $\widehat{\Lambda}_0\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\widehat{\Lambda}_1$ as weights. Likewise, when $(\Lambda_1 {\Lambda_1}^\intercal)^{-1}$ exists, I can estimate $\mathbf{F}_1(\cdot|j)$ for untreated group $j \leq J_0$.

The following Assumption replaces Assumption 5, in the context of the estimation.

Assumption 10.

(i) *Let*

$$\Gamma_0 = \begin{pmatrix}\Pr\{W_i(0) \in C_1 | k_i = 1\} & \cdots & \Pr\{W_i(0) \in C_1 | k_i = K\} \\ \vdots & \ddots & \vdots \\ \Pr\{W_i(0) \in C_M | k_i = 1\} & \cdots & \Pr\{W_i(0) \in C_M | k_i = K\}\end{pmatrix}.$$

Γ_0*,* $P\Gamma_0$ *and* Λ_0 *have rank* K*.*

(ii) *Let*

$$\Gamma_1 = \begin{pmatrix} \Pr\{W_i(1) \in C_1 | k_i = 1\} & \cdots & \Pr\{W_i(1) \in C_1 | k_i = K\} \\ \vdots & \ddots & \vdots \\ \Pr\{W_i(0) \in C_M | k_i = 1\} & \cdots & \Pr\{W_i(0) \in C_M | k_i = K\} \end{pmatrix}.$$

Γ_1, $P\Gamma_1$ *and* Λ_1 *have rank* K.

Theorem 2. *Assumptions 1-3 hold. When Assumption 10-(i) holds,*

$$\widehat{\Lambda}_0^\intercal \left(\widehat{\Lambda}_0 \widehat{\Lambda}_0^\intercal\right)^{-1} \widehat{\Lambda}_1 \xrightarrow{p} \Lambda_0^\intercal \left(\Lambda_0 \Lambda_0^\intercal\right)^{-1} \Lambda_1$$

as $n \to \infty$. *When Assumption 10-(ii) holds,*

$$\widehat{\Lambda}_1^\intercal \left(\widehat{\Lambda}_1 \widehat{\Lambda}_1^\intercal\right)^{-1} \widehat{\Lambda}_0 \xrightarrow{p} \Lambda_1^\intercal \left(\Lambda_1 \Lambda_1^\intercal\right)^{-1} \Lambda_0$$

as $n \to \infty$.

Proof. See Appendix. □

Let $\mathbf{F}_{y(0)}(\cdot|j)$ denote the conditional distribution of $Y_i(0)$ given $Z_i = j$. For $j \leq J_0$, $\mathbf{F}_{y(0)}(\cdot|j)$ is directly observed from the dataset. Corollary 1 shows that there is a consistent estimator of $\mathbf{F}_{y(0)}(\cdot|j)$ for the treated group $j > J_0$, when Assumption 10-(i) holds.

Corollary 1. *Assumptions 1-3 and 10-(i) hold. Let*

$$\mathbb{F}_{y(0)}(y|j) = \frac{1}{\sum_{i=1}^n \mathbf{1}\{D_i = 0, Z_i = j\}} \sum_{i=1}^n \mathbf{1}\{Y_i \leq y, D_i = 0, Z_i = j\} \quad \forall y \in \mathcal{Y}$$

for each $j = 1, \cdots, J_0$ *and let*

$$\left(\mathbb{F}_{y(0)}(y|J_0+1) \quad \cdots \quad \mathbb{F}_{y(0)}(y|J)\right) = \left(\mathbb{F}_{y(0)}(y|1) \quad \cdots \quad \mathbb{F}_{y(0)}(y|J_0)\right) \widehat{\Lambda}_0^\intercal \left(\widehat{\Lambda}_0 \widehat{\Lambda}_0^\intercal\right) \widehat{\Lambda}_1.$$

Then, for each $j = J_0 + 1, \cdots, J$,

$$\sup_{y \in \mathbb{R}} \left|\mathbb{F}_{y(0)}(y|j) - \mathbf{F}_{y(0)}(y|j)\right| \xrightarrow{p} 0$$

as $n \to \infty$.

Proof. The proof is direct from the Glivenko-Cantelli Theorem and Theorem 2. □

Similarly, when Assumption 10-(ii) holds, we have a consistent estimator for $\mathbf{F}_{y(1)}(\cdot|j)$ for the untreated group j.

4.2 Treatment effect distribution

For estimation of the treatment effect distribution, the estimation procedure is three steps. Firstly, by solving some NMF problem, we get $\left(\widehat{\Lambda}_0, \widehat{\Lambda}_1\right)$. Then, the treatment effect distribution is estimated with MLE applied to (2). Suppose that the conditional distribution of $(Y_i(0), X_{i2})$ given (k_i, X_{i1}) is parametrized with some finite-dimensional parameter ξ and the conditional distribution of $Y_i(1) - Y_i(0)$ given $(Y_i(0), k_i, X_{i2})$ is parametrized with some finite-dimensional parameter θ. WLOG suppose $\theta = (\theta_1, \cdots, \theta_K)$ and $\xi = (\xi_1, \cdots, \xi_K)$ where θ_k and ξ_k denote the distributional parameters for type k in the finite mixture. ξ is a nuisance parameter. The second step is to estimate the nuisance parameter, using the untreated groups:

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \sum_{i: Z_i \leq J_0} \log \left(\sum_{k=1}^{K} \hat{\lambda}(k|Z_j) \cdot g_{0x_2|x_1}(Y_i, X_{i2}|X_{i1}; \xi_k) \right).$$

Given $\left(\widehat{\Lambda}_0, \widehat{\Lambda}_1, \hat{\xi}\right)$, the last step is to estimate the parameter of interest, using the treated groups:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i: Z_i > J_0} \log \left(\sum_{k=1}^{K} \hat{\lambda}(k|Z_j) \cdot \int_{\mathcal{Y} \times \mathcal{X}_2} \phi(Y_i - y_0|y_0, x_2; \theta_k) g_{0x_2|x_1}(y_0, x_2|X_{i1}; \hat{\xi}_k) d\left(y_0, x_2\right) \right).$$

For partial identification of (Λ_0, Λ_1), I use the same NMF algorithm from the previous subsection. For point identification, I slightly modify the NMF problem. The partition on $\mathcal{X}$ is now constructed in a way that Assumption 6-(ii) can be checked. Let $\{D_m\}_{m=1}^{M_{x_1}}$ be a partition of $\mathcal{X}_1$. For each A_m, construct a partition on $\mathcal{X}_0$ and a partition on $\mathcal{X}_2$: $\{E_{ml}\}_{l=1}^{M_{m,x_0}}$ and $\{F_{ml}\}_{l=1}^{M_{m,x_2}}$. Let $\{A_m\}_{m=1}^{M_x} = \{D_m \times E_{ml} \times F_{ml'}\}_{m,l,l'}$. Then, $M_x = \sum_{m=1}^{M_{x_1}} M_{m,x_0} \cdot M_{m,x_2}$. Construct $\{C_m\}_{m=1}^{M}$ in the same way as in the previous subsection. Assumption 6-(ii) implies

$$\begin{aligned}
&\Pr\left\{X_{i0} \in E_{ml}, X_{i1} \in D_m, X_{i2} \in F_{ml'} | k_i = k\right\} \\
&= \int_{D_m} \Pr\left\{X_{i0} \in E_{ml}, X_{i2} \in F_{ml'} | k_i = k, X_{i1} = x\right\} g_{x_1,k}(x_1) dx_1 \\
&= \int_{D_m} \Pr\left\{X_{i0} \in E_{ml} | k_i = k, X_{i1} = x\right\} \cdot \Pr\left\{X_{i2} \in F_{ml'} | k_i = k, X_{i1} = x\right\} g_{x_1,k}(x_1) dx_1.
\end{aligned}$$

Suppose $|\Pr\{X_{i0} \in E_{ml} | k_i = k, X_{i1} = x\} - \Pr\{X_{i0} \in E_{ml} | k_i = k, X_{i1} \in D_m\}| \leq \eta$ for all $x \in D_m$. Then,

$$\begin{aligned}
&|\Pr\{X_{i0} \in E_{ml}, X_{i1} \in D_m, X_{i2} \in F_{ml'} | k_i = k\} \cdot \Pr\{X_{i1} \in D_m | k_i = k\} \\
&\quad - \Pr\{X_{i0} \in E_{ml}, X_{i1} \in D_m | k_i = k\} \cdot \Pr\{X_{i1} \in D_m, X_{i2} \in F_{ml'} | k_i = k\}| \\
&\leq \eta \Pr\{X_{i1} \in D_m | k_i = k\}^2 . \qquad (4)
\end{aligned}$$

The inequality imposes a quadratic constraints on (Γ_0, Γ_1) since all of the four probabilities that appear in Equation (4) are linear in $P\Gamma_0 = P\Gamma_1$. Note that the quadratic constraints on (Γ_0, Γ_1) are not positive definite and thus the minimization problem (3) may not be convex even when (Λ_0, Λ_1) is fixed. In this modified version of the NMF problem, there is clear tradeoff in increasing M; with a finer partition of $\mathcal{X}$, we can check Assumption 6-(ii) more tightly, giving us a smaller identified set at the cost of having bigger noise in $\mathbb{H}_0, \mathbb{H}_1$. Thus, I let M_x grow with n; we use a finer partition on $\mathcal{X}$ when there are more observations. The dependency on n is omitted for notational brevity.

The following assumption replaces Assumptions 6 and 9 in the context of estimation and assumes smoothness on the conditional expectation so that we can use the inequality in (4).

Assumption 11. *There exists some $\tilde{x} \in \mathcal{X}_1$ such that*

(i) $X_{i0} \perp\!\!\!\perp X_{i2} \big| X_{i1} = \tilde{x}$.

(ii) *There exist some partitions $\{\tilde{E}_l\}_{l=1}^{\tilde{M}_{x_0}}, \{\tilde{F}_{ml}\}_{l=1}^{\tilde{M}_{x_2}}$, respectively on $\mathcal{X}_0, \mathcal{X}_2$, such that*

$$\begin{aligned}
p_k &= \left(\Pr\left\{X_{i0} \in \tilde{E}_1 | k_i = k, X_{i1} = \tilde{x}\right\} \quad \cdots \quad \Pr\left\{X_{i0} \in \tilde{E}_{\tilde{M}_{x_0}} | k_i = k, X_{i1} = \tilde{x}\right\}\right)^\intercal , \\
q_k &= \left(\Pr\left\{X_{i2} \in \tilde{F}_1 | k_i = k, X_{i1} = \tilde{x}\right\} \quad \cdots \quad \Pr\left\{X_{i2} \in \tilde{F}_{\tilde{M}_{x_2}} | k_i = k, X_{i1} = \tilde{x}\right\}\right)^\intercal
\end{aligned}$$

satisfy the following: for any $k \neq k'$, $q_k \neq q_{k'}$ and $p_k, p_{k'}$ are linearly independent; $\{\tilde{D}_n\}_{n=1}^\infty \subset \mathcal{X}_1$ contains $\tilde{x}$ and

$$\left\| \left(\Pr\left\{X_{i0} \in \tilde{E}_l, X_{i2} \in \tilde{F}_l | k_i = k, X_{i1} \in \tilde{D}_n\right\}\right)_{l,l'} - p_k q_k^\intercal \right\|_F \leq \eta_n$$

with some $\{\eta_n\}_{n=1}^\infty$ converging to zero.

(iii) *The partition used in the NMF algorithm includes $\{\tilde{D}_n \times \tilde{E}_l \times \tilde{F}_{l'}\}_{l,l'}$ from Assumption (ii)*

for each n. Also, for (Γ_0, Γ_1) defined as in Assumption 10, which now changes with n, there exists some $K \times M$ matrix $\tilde{P}$ whose elements are either one or zero, in a way that product of any two rows is zero and $\|\tilde{P}\|_F = M$. $\tilde{P}\Gamma_1$ converge to some invertible matrices as $M \to \infty$. Lastly, $M^3/n \to 0$ as $n \to \infty$.

Assumption 11-(i) replaces Assumption 6 and Assumption 11-(ii) replaces Assumption 9.

Theorem 3. *Assumptions 1-3, 10-11 hold. Up to some permutation on $\{1, \cdots, K\}$,*

$$\widehat{\Lambda}_0 \xrightarrow{p} \Lambda_0 \quad \text{and} \quad \widehat{\Lambda}_1 \xrightarrow{p} \Lambda_1$$

as $n \to \infty$.

Proof. See Appendix. □

The following corollary follows directly.

Corollary 2. *Assumptions 1-3, 6-8, 10-11 and some regularity conditions on the density functions $g_{0x_2|x_1}, \phi$ hold. Up to some permutation on $\{1, \cdots, K\}$,*

$$\hat{\xi} \xrightarrow{p} \xi \quad \text{and} \quad \hat{\theta} \xrightarrow{p} \theta$$

as $n \to \infty$.

Proof. Direct from the convergence of $\widehat{\Lambda}_0$. □

5 Empirical illustration

In this section, I revisit the question of measuring disemployment effect on earnings and apply the nonnegative matrix factorization to five distribution functions retrieved from the National Longitudinal Survey of Youth (NLSY). I focus on the annual earnings distribution in 1987 and use the annual earning in 1985 and the Armes Forces Qualification Test (AFQT) as control variables. For instrument, I follow the same spirit as in Example 2 (diff-in-diff) and use the disemployment

timing as an instrument.

$$X_i : \text{Armed Forces Qualification Test, Annual earnings in 1985,}$$

$$Y_i : \text{Annual earnings in 1987,}$$

$$Z_i : \text{categorical variable for disemployment timing.}$$

Z_i finds five groups: disemployed in 1987, disemployed in 1989, disemployed in 1991, dismploeyed in 1993, never disemployed until 1993. Since I am using the annual earnings in 1987 as the outcome variable, Z_i finds one treated group and four untreated groups.

The main model of the paper (1) assumes that there are finite types of individuals which follow different distributions of the AFQT score and annual earnings. Assumption 4 assumes that we have sufficient variation across these type-specific component distribution in terms of their marginal distribution of the AFQT score and the past earning, which we observe for every individual. Assumption 5 assumes that each group defined with the disemployment timing show sufficient variation in terms of their type shares; earlier-disemployed individuals are systemically different from later-disemployed individuals, in terms of their types. Lastly, Assumption 6 assumes that the conditional distribution of 1987 earnings does not depend on the AFQT score, when conditioned on the 1985 earnings and the type; the (observed) 1985 earnings and the unobserved type contain sufficient information about individual's potential earning distribution so that the AFQT score does not affect the conditional distribution of the 1987 earnings.

5.1 Preliminary result: quantile treatment effect

I present some preliminary empirical results on the quantile treatment effect in this subsection. For the NMF minimization problem, I used $M_x = 4$ and $L_m = 2$; the support of X_i is partitioned into four sets and the support of Y_i is partitioned into two sets, for each of the four sets in the partition for X_i. For the number of types K, I used $K = 2$ and $K = 3$.

When $K = 2$, the extrapolation weight $\widehat{\Lambda}_0^\intercal \left(\widehat{\Lambda}_0 \widehat{\Lambda}_0^\intercal\right)^{-1} \widehat{\Lambda}_1$ is $(0.381, 0.286, 0.258, 0.075)$, on the four untreated groups: disemployed in 1989, disemployed in 1991, disempolyed in 1993, never disemployed until 1993. When $K = 3$, the extrapolation weight is $(0.312, 0.379, 0.228, 0.082)$. The never-disemployed group gets less weights than the other untreated groups, which makes sense intuitively; the type distributions among individuals in the disemployed groups are similar to each other while the type distribution among individuals in the never-disemployed group is distinctvely

different.

Figure 1 contains the estimation result when $K = 2$. The left panel is the observed distribution of the 1985 earnings and the fitted distribution of the 1985 earnings, for the treated group. Note that the distributions are fitted quite well even though only a small number of moments ($M_x = 4$) were used. The right panel is the observed distribution of 1987 earnings, thus the treated potential outcomes, and the estimated counterfactual distribution of the 1987 earnings, thus the untreated potential outcomes, for the treated group. The difference between the two distributions denote the treatment effect. Figure 2 contains the same estimation result for $K = 3$.

Table 1 contains the estimates for τ-th quantile; $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. We see significant disemployment effect on the earning distributions, except at the right tail. The disemployment decreases the annual earnings by approximately 500 dollars at the 0.9-th quantile while the disemployment effects are mostly larger than 1500 dollars at the rest of the four quantiles.

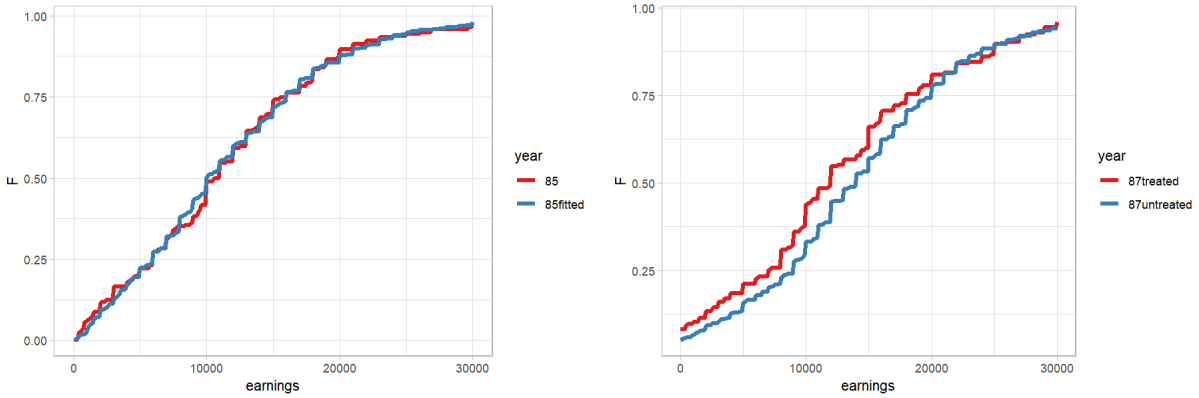

Figure 1: The annual earnings distribution, $K = 2$

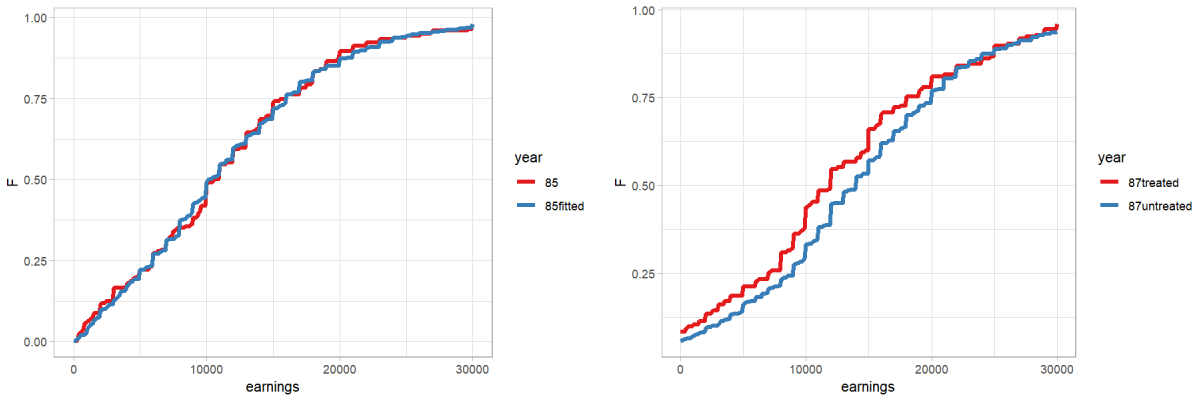

Figure 2: The annual earnings distribution, $K = 2$

τ	0.1	0.25	0.5	0.75	0.9
$K = 2$	-2024.8	-2060.6	-1989.2	-2068 8	-32.8
$K = 3$	-1445.4	-2063.5	-1988.2	-2044.7	-536.1

Table 1: Quantile treatment effects

6 Conclusion

This paper proposes a new estimation strategy using an NMF algorithm for the quantile treatment effects and the distributional treatment effects, based on a finite mixture model. The quantile treatment effects are identified and estimated under a fairly relaxed assumptions such as a full rank assumption on the mixture weight matrix. A key observation is that the mixture weights and the component distribution functions need not be point identified when the object of interest is the counterfactual outcome distribution. For the identification and the estimation of the distributional treatment effects, I impose additional assumption that the observed control covariates satisfy a Markov property, conditioning on the type. The Markov property condition is particularly suitable when we have a panel data and we can use lagged outcomes as controls; the outcome process satisfies the Markov property, conditioning on the latent type. The NMF algorithm with some additional constraints that reflect the Markov property consistently estimate the mixture weights, allowing us the recover the component distribution functions.

References

Callaway, Brantly and Tong Li, "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

Deaner, Ben, "Proxy Controls and Panel Data," 2023.

Fan, Yanqin and Sang Soo Park, "Sharp bounds on the distribution of treatment effects and their statistical inference," *Econometric Theory*, 2010, *26* (3), 931–951.

Frandsen, Brigham R and Lars J Lefgren, "Partial identification of the distribution of treatment effects with an application to the Knowledge is Power Program (KIPP)," *Quantitative Economics*, 2021, *12* (1), 143–171.

Henry, Marc, Yuichi Kitamura, and Bernard Salanié, "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 2014, *5* (1), 123–144.

Hu, Yingyao and Susanne M Schennach, "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 2008, *76* (1), 195–216.

Kaji, Tetsuya and Jianfei Cao, "Assessing Heterogeneity of Treatment Effects," 2023.

Kedagni, Desire, "Identifying treatment effects in the presence of confounded types," *Journal of Econometrics*, 2023, *234* (2), 479–511.

Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, 2018, *105* (4), 987–993.

Noh, Sungho, "Nonparametric identification and estimation of heterogeneous causal effects under conditional independence," *Econometric Reviews*, 2023, *42* (3), 307–341.

APPENDIX

A Proofs

A.1 Proof for Lemma 1

Suppose that Assumption 5-(i) holds. For any $w \in \mathcal{Y} \times \mathcal{X}$,

$$\sum_{k=1}^{K} \lambda'(k|j) \cdot \mathbf{G}'_{0k}(w) = \begin{pmatrix} \mathbf{G}'_{01}(w) & \cdots & \mathbf{G}'_{0K}(w) \end{pmatrix} \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{F}_0(w|0,1) & \cdots & \mathbf{F}_0(w|0,J_0) \end{pmatrix} \Lambda_0'^{\intercal} \left(\Lambda'_0 \Lambda_0'^{\intercal}\right)^{-1} \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix}.$$

$\Lambda'_0 \Lambda_0'^{\intercal}$ is invertible since

$$\begin{pmatrix} \mathbf{G}'_{x1}(x_1) & \cdots & \mathbf{G}'_{xK}(x_1) \\ \vdots & \ddots & \vdots \\ \mathbf{G}'_{x1}(x_K) & \cdots & \mathbf{G}'_{xK}(x_K) \end{pmatrix} \Lambda'_0 = \begin{pmatrix} \mathbf{F}_x(x_1|1) & \cdots & \mathbf{F}_x(x_1|J_0) \\ \vdots & \ddots & \vdots \\ \mathbf{F}_x(x_K|1) & \cdots & \mathbf{F}_x(x_K|J_0) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{G}_{x1}(x_1) & \cdots & \mathbf{G}_{xK}(x_1) \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{x1}(x_K) & \cdots & \mathbf{G}_{xK}(x_K) \end{pmatrix} \Lambda_0$$

is full rank from Assumption 4 and Assumption 5-(i) and therefore Λ'_0 has full rank. $\mathbf{F}_x(\cdot|j)$ denotes the conditional distribution of X_i given $Z_i = j$. It remains to show that

$$\Lambda_0'^{\intercal} \left(\Lambda'_0 \Lambda_0'^{\intercal}\right)^{-1} \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix} = \Lambda_0''^{\intercal} \left(\Lambda''_0 \Lambda_0''^{\intercal}\right)^{-1} \begin{pmatrix} \lambda''(1|j) \\ \vdots \\ \lambda''(K|j) \end{pmatrix}.$$

Let Γ'_x and Γ''_x be defined as in Assumption 4. Find that both Γ'_x and Γ''_x are invertible, $\Gamma'_x\Lambda'_0 = \Gamma''_x\Lambda''_0$, and

$$\Gamma'_x \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix} = \begin{pmatrix} \mathbf{F}_x(x_1|j) \\ \vdots \\ \mathbf{F}_x(x_K|j) \end{pmatrix} = \Gamma''_X \begin{pmatrix} \lambda''(1|j) \\ \vdots \\ \lambda''(K|j) \end{pmatrix}$$

since both of the two tetrads we consider are in the identified set. Thus,

$$\begin{aligned} \Lambda_0'^{\intercal} \left(\Lambda'_0\Lambda_0'^{\intercal}\right)^{-1} \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix} &= \Lambda_0'^{\intercal}\Gamma_x'^{\intercal} \left(\Gamma_x'^{\intercal}\right)^{-1} \left(\Lambda'_0\Lambda_0'^{\intercal}\right)^{-1} \left(\Gamma'_x\right)^{-1} \Gamma'_x \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix} \\ &= \left(\Gamma'_x\Lambda'_0\right)^{\intercal} \left(\Gamma'_x\Lambda'_0 \left(\Gamma'_x\Lambda'_0\right)^{\intercal}\right)^{-1} \Gamma'_x \begin{pmatrix} \lambda'(1|j) \\ \vdots \\ \lambda'(K|j) \end{pmatrix}. \end{aligned}$$

We can repeat the same argument for the second half of the lemma as well.

A.2 Proof for Theorem 1

Let $g_{x,k}$ denote the conditional density of X_i given $k_i = k$ and $g_{x_0x_1,k}$ denote the conditional density of $\left(X_{i0}, X_{i1}\right)$ given $k_i = k$. Assumption 6-(ii) assumes that

$$g_{x,k}(x_0, x_1, x_2) = g_{x_2|x_1,k}(x_2|x_1)g_{x_0x_1,k}(x_0, x_1).$$

Suppose a possibly misspecified component density function $g = \sum_{k=1}^K \alpha_k g_{x,k}$ such that $\sum_k \alpha_k = 1$. Then,

$$\begin{aligned} g(x_0, x_1, x_2) &= \sum_{k=1}^K \alpha_k g_{x_2|x_1,k}(x_2|x_1)g_{x_0x_1,k}(x_0, x_1) \\ g_{x_0x_1}(x_0, x_1) &= \sum_{k=1}^K \alpha_k g_{x_0x_1,k}(x_0, x_1) \\ \frac{g(x_0, x_1, x_2)}{g_{x_0x_1}(x_0, x_1)} &= \frac{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x_0|x_1)g_{x_1x_2,k}(x_1, x_2)}{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x_0|x_1)g_{x_1,k}(x_1)} \end{aligned} \tag{5}$$

When g is correctly specified, i.e. $g = g_k$ for some k and $\alpha_{k'} = 0$ for all $k' \neq k$, the quantity on the RHS of Equation (5) is equal to $g_{x_2|x_1,k}(x_2|x_1)$ and is a trivial function of x_0.

The identification is complete by showing that there exist some $(x_0, x_1, x_2) \in \mathcal{X}$ such that

$$\frac{g(x_0, x_1, x_2)}{g_{x_0x_1}(x_0, x_1)} \cdot \frac{g_{x_1}(x_1)}{g_{x_1x_2}(x_1, x_2)} \neq 1.$$

whenever g is misspecified. Under misspecification, there are at least two k such that $\alpha_k \neq 0$. From Assumption 9-(i), we can find some (x_1, x_2) such that $\left(\alpha_1 g_{x_1x_2,1}(x_1, x_2), \cdots, \alpha_K g_{x_1x_2,K}(x_1, x_2)\right)$ and $\left(\alpha_1 g_{x_1,1}(x_1), \cdots, \alpha_K g_{x_1,K}(x_1)\right)$ are not linearly independent. Assume to the contrary that the LHS on Equation (5) is a trivial function of x_0 and let

$$\tilde{g}(x) = \frac{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x|x_1) g_{x_1x_2,k}(x_1, x_2)}{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x|x_1) g_{x_1,k}(x_1)}.$$

Then,

$$\begin{aligned}\tilde{g}'(x) = &\frac{1}{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x|x_1) g_{x_1,k}(x_1)} \begin{pmatrix} \alpha_1 g_{x_1x_2,1}(x_1, x_2) & \cdots & \alpha_K g_{x_1x_2,K}(x_1, x_2) \end{pmatrix} \begin{pmatrix} g'_{x_0|x_1,1}(x|x_1) \\ \vdots \\ g'_{x_0|x_1,K}(x|x_1) \end{pmatrix} \\ &- \frac{\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x|x_1) g_{x_1x_2,k}(x_1, x_2)}{\left(\sum_{k=1}^K \alpha_k g_{x_0|x_1,k}(x|x_1) g_{x_1,k}(x_1)\right)^2} \begin{pmatrix} \alpha_1 g_{x_1,1}(x_1) & \cdots & \alpha_K g_{x_1,K}(x_1) \end{pmatrix} \begin{pmatrix} g'_{x_0|x_1,1}(x|x_1) \\ \vdots \\ g'_{x_0|x_1,K}(x|x_1) \end{pmatrix}.\end{aligned}$$

Since $\left(\alpha_1 g_{x_1x_2,1}(x_1, x_2), \cdots, \alpha_K g_{x_1x_2,K}(x_1, x_2)\right)$ and $\left(\alpha_1 g_{x_1,1}(x_1), \cdots, \alpha_K g_{x_1,K}(x_1)\right)$ are not linearly independent, $\tilde{g}'(x) = 0$ implies that there is some nonzero vector $c \in \mathbb{R}^K$ such that

$$\begin{pmatrix} g'_{x_0|x_1,1}(x|x_1) & \cdots & g'_{x_0|x_1,K}(x|x_1) \end{pmatrix} c = 0 \quad \forall x \in \mathcal{X}_0$$

and therefore

$$\begin{pmatrix} g_{x_0|x_1,1}(x|x_1) & \cdots & g_{x_0|x_1,K}(x|x_1) \end{pmatrix} c = C \quad \forall x \in \mathcal{X}_0$$

with some constant C. From Assumption 9-(ii), $C = 0$ and therefore there cannot be a nonzero vector c satisfying the above. $\tilde{g}$ is not a trivial function of x_0.

A.3 Proof for Theorem 2

Let us prove the first half of the theorem. Let

$$\mathbf{H}_0 = \begin{pmatrix} \Pr\{W_i \in C_1 | Z_i = 1\} & \cdots & \Pr\{W_i \in C_1 | Z_i = J_0\} \\ \vdots & \ddots & \vdots \\ \Pr\{W_i \in C_M | Z_i = 1\} & \cdots & \Pr\{W_i \in C_M | Z_i = J_0\} \end{pmatrix} = \Gamma_0 \Lambda_0,$$

$$\mathbf{H}_1 = \begin{pmatrix} \Pr\{X_i \in C_1 | Z_i = J_0 + 1\} & \cdots & \Pr\{X_i \in C_1 | Z_i = J\} \\ \vdots & \ddots & \vdots \\ \Pr\{X_i \in C_M | Z_i = J_0 + 1\} & \cdots & \Pr\{X_i \in C_M | Z_i = J\} \end{pmatrix} = \Gamma_1 \Lambda_1.$$

From iid-ness of observations, we have

$$\|\mathbb{H}_0 - \mathbf{H}_0\|_F = O_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \|\mathbb{H}_1 - \mathbf{H}_1\|_F = O_p\left(\frac{1}{\sqrt{n}}\right).$$

From the definition of $\widehat{\Lambda}_0$ and $\widehat{\Lambda}_1$, we have

$$\begin{aligned} &\left\|\mathbb{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 + \left\|\mathbb{H}_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F^2 \\ &\leq \|\mathbb{H}_0 - \Gamma_0\Lambda_0\|_F^2 + \|\mathbb{H}_1 - \Gamma_1\Lambda_1\|_F^2 \\ &\leq (\|\mathbb{H}_0 - \mathbf{H}_0\|_F + \|\mathbf{H}_0 - \Gamma_0\Lambda_0\|_F)^2 + (\|\mathbb{H}_1 - \mathbf{H}_1\|_F + \|\mathbf{H}_1 - \Gamma_1\Lambda_1\|_F)^2 \\ &= \|\mathbb{H}_0 - \mathbf{H}_0\|_F^2 + \|\mathbb{H}_1 - \mathbf{H}_1\|_F^2 = O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Then,

$$\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 = \left\|\mathbf{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F^2 \leq \left(\|\mathbf{H}_0 - \mathbb{H}_0\|_F + \left\|\mathbb{H}_0 - \widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F\right)^2 = O_p\left(\frac{1}{\sqrt{n}}\right)$$

and likewise for $\left\|\Gamma_1\Lambda_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F = \left\|\mathbf{H}_1 - \widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F$. From the submultiplicavity of $\|\cdot\|_F$, we also get $\left\|P\Gamma_1\Lambda_1 - P\widehat{\Gamma}_1\widehat{\Lambda}_1\right\|_F = \left\|P\Gamma_0\Lambda_1 - P\widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F = O_p\left(\frac{1}{\sqrt{n}}\right)$.

Find that when $(\Gamma_0{}^\intercal\Gamma_0)^{-1}, (\Gamma_0{}^\intercal P^\intercal P\Gamma_0)^{-1}, \left(\widehat{\Gamma}_0^\intercal\Gamma_0\right)^{-1}$ and $\left(\Gamma_0{}^\intercal P^\intercal P\widehat{\Gamma}_0\right)^{-1}$ exist,

$$\begin{aligned} \Lambda_0{}^\intercal(\Lambda_0\Lambda_0{}^\intercal)^{-1}\Lambda_1 &= \Lambda_0{}^\intercal\Gamma_0{}^\intercal\Gamma_0(\Gamma_0{}^\intercal P^\intercal P\Gamma_0\Lambda_0\Lambda_0{}^\intercal\Gamma_0{}^\intercal\Gamma_0)^{-1}\Gamma_0{}^\intercal P^\intercal P\Gamma_0\Lambda_1, \\ \widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\widehat{\Lambda}_1 &= \widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\Gamma_0\left(\Gamma_0{}^\intercal P^\intercal P\widehat{\Gamma}_0\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\Gamma_0\right)^{-1}\Gamma_0{}^\intercal P^\intercal P\widehat{\Gamma}_0\widehat{\Lambda}_1 \end{aligned}$$

and therefore

$$\left\|\Lambda_0^\intercal \left(\Lambda_0\Lambda_0^\intercal\right)^{-1}\Lambda_1 - \widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\widehat{\Lambda}_1\right\|_F = o_p\left(1\right)$$

from $\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F = o_p(1)$ and $\left\|P\Gamma_0\Lambda_1 - P\widehat{\Gamma}_0\widehat{\Lambda}_1\right\|_F = o_p(1)$. Assumption 10-(i) implies that Γ_0 and $P\Gamma_0$ have full rank. $\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ exists with probability going to one since:

$$\left\|\Gamma_0^\intercal\Gamma_0\Lambda_0 - \Gamma_0^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F \leq \|\Gamma_0\|_F \cdot o_p(1) = o_p(1).$$

Thus, the determinant of $\Gamma_0^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0$ converges to a nonzero constant in probability and thus $\widehat{\Lambda}_0$ has a full rank with probability converging to one. It remains to show $\widehat{\Gamma}_0^\intercal\Gamma_0$ and $\Gamma_0^\intercal P^\intercal P\widehat{\Gamma}_0$ have full rank, with probability converging to one.

Suppose $\widehat{\Gamma}_0^\intercal\Gamma_0$ is not invertible. Then, there is some nonzero vector $v \in \mathbb{R}^K$ such that $\widehat{\Gamma}_0^\intercal\Gamma_0 v = \mathbf{0}_K$. $\Gamma_0 v$ lies in the left null space of $\widehat{\Gamma}_0$. Find some nonzero vector $u \in \mathbb{R}^{J_0}$ such that $v = \Lambda_0 u$. Such u always exist since Λ_0 has full rank. Then, for any $\widehat{\Lambda}_0 u$,

$$\Gamma_0\Lambda_0 u - \widehat{\Gamma}_0\widehat{\Lambda}_0 u = \Gamma_0 v - \widehat{\Gamma}_0\widehat{\Lambda}_0 u \neq 0.$$

WLOG we can find v such that $\Gamma_0 v$ is orthogonal to the columns of $\widehat{\Gamma}_0$ and $\|\Gamma_0 v\|_2 = 1$. Also, we put a bound on $\|u\|_2$ by letting $u = \Lambda_0^\intercal\left(\Lambda_0\Lambda_0^\intercal\right)^{-1}v$; $\|v\|_2 \leq 1$ from the observation that each element of Γ_0 lies between 0 and 1. Then, when $\widehat{\Gamma}_0^\intercal\Gamma_0$ is not invertible,

$$1 = \|\Gamma_0 v\|_2 \leq \left\|\Gamma_0 v - \widehat{\Gamma}_0\widehat{\Lambda}_0 u\right\|_2 = \left\|\Gamma_0\Lambda_0 u - \widehat{\Gamma}_0\widehat{\Lambda}_0 u\right\|_2 \leq \left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F \|u\|_2$$

holds, giving us a contradiction. The first inequality holds since $(a+b)^\intercal(a+b) = a^\intercal a + b^\intercal b$ when $a^\intercal b = 0$. Therefore

$$\Pr\left\{\widehat{\Gamma}_0^\intercal\Gamma_0 \text{ is not invertible}\right\} \leq \Pr\left\{\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F \geq \frac{1}{\|u\|_2}\right\} = o(1).$$

We can repeat the same argument for $\Gamma_0^\intercal P^\intercal P\widehat{\Gamma}_0$.

A.4 Proof for Theorem 3

Let us first prove the consistency of $\widehat{\Lambda}_0$.

Step 1. $\left\|\Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F = O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$.

Note that $\mathbb{H}_0, \mathbb{H}_1, \mathbf{H}_0$ and $\mathbf{H}_1$ now have growing number of rows, depending on the (sequence of) partition we use. Find that

$$\begin{aligned}
&\Pr\left\{\frac{n}{M}\sum_{j=1}^{J_0}\sum_{m=1}^{M}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{W_i\in C_m, Z_i=j\}-\Pr\{W_i\in C_m, Z_i=j\}\right)^2\geq\varepsilon\right\}\\
&\leq\sum_{j=1}^{J_0}\sum_{m=1}^{M}\Pr\left\{n\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{W_i\in C_m, Z_i=j\}-\Pr\{W_i\in C_m, Z_i=j\}\right)^2\geq\frac{\varepsilon}{J_0}\right\}\\
&\leq\frac{J_0}{\varepsilon}\sum_{j=1}^{J_0}\sum_{m=1}^{M}\mathrm{Var}\left(\mathbf{1}\{W_i\in C_m, Z_i=j\}\right)\leq\frac{J_0}{\varepsilon}
\end{aligned}$$

The second inequality holds from Markov's inequality. The last inequality holds since

$$\sum_{j=1}^{J_0}\sum_{m=1}^{M}\Pr\{W_i\in C_m, Z_i=j\}\leq 1$$

and thus the summation of variances has an upper bound of $\sum_{j,m}\frac{J_0M-1}{{J_0}^2M^2}\leq 1$. Thus,

$$\left\|\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}\{W_i\in C_m, Z_i=j\}\right)_{m,j\leq J_0}-\left(\Pr\{W_i\in C_m, Z_i=j\}\right)_{m,j\leq J_0}\right\|_F=O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right).$$

We can repeat the same for $j>J_0$. Thus, $\|\mathbb{H}_0-\mathbf{H}_0\|_F=O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$ and $\|\mathbb{H}_1-\mathbf{H}_1\|_F=O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$. From the same argument in the proof of Theorem 2, $\left\|\Gamma_0\Lambda_0-\widehat{\Gamma}_0\widehat{\Lambda}_0\right\|_F=O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$.

Step 2. There is an envelope for the identified set of Γ_0 that shrinks to an arbitrarily small set as $n\to\infty$.

Define a $\tilde{M}_{x_0}\times\tilde{M}_{x_2}$ matrix $\tilde{\Gamma}_k$ such that

$$\tilde{\Gamma}_k=\left(\Pr\left\{X_{i0}\in\tilde{E}_l, X_{i2}\in\tilde{F}_{l'}|k_i=k, X_{i1}\in\tilde{D}_n\right\}\right)_{l,l'}.$$

$\tilde{\Gamma}_k$ depends on n. Let $\mathcal{A}_n$ be the set of linear coefficients on the mixture component distributions that satisfy some of the constraints in the NMF minimization problem, in the context of the

partition used at n:

$$\mathcal{A}_n = \Big\{ (\alpha_1, \cdots, \alpha_K) \in \mathbb{R}^{K\times K} : \ \mathbf{1}_K{}^\intercal \alpha_k = 1, \min_{p,q} \Big\| \sum_{l=1}^{K} \alpha_{kl} \tilde{\Gamma}_l - pq^\intercal \Big\|_F \leq \eta_n + \tilde{\eta} \ \forall k, \\ \exists \Lambda \in [0,1]^{K\times K} \text{ s.t. } \big\| (\alpha_1, \cdots, \alpha_K)\Lambda - \Lambda_0 \big\|_\infty \leq \tilde{\eta} \Big\}. \qquad (6)$$

$\tilde{\eta}$ is an arbitrarily small number such that $\tilde{\eta} \leq \frac{\min_{j,k} \lambda(k|j)}{2}$. α_k is a vector of linear coefficients to construct a type k component distribution, under the rotation $(\alpha_1, \cdots, \alpha_K)$; the rotated component distributions still need to satisfy constraints used in the NMF, including Equation (4).

WTS $\mathcal{A}_n \subset \bar{\mathcal{A}}_n := \big\{ (\alpha_1, \cdots, \alpha_K) \in \mathbb{R}^{K\times K} : \|\alpha_k - e_{\pi(k)}\|_\infty \leq \tilde{\varepsilon}_n \ \forall k \text{ with some permutation } \pi \big\}$ with some $\{\tilde{\varepsilon}_n\}_{n=1}^{\infty}$ converging to zero, where e_k is a $K \times 1$ vector whose k-th element is one and the rest are zeros. Consider some $a \in \mathbb{R}^K$ satisfying the conditions given in (6). Then,

$$\begin{aligned} \Big\| \sum_k a_k \tilde{\Gamma}_k - pq^\intercal \Big\|_F &= \Big\| \sum_k a_k \left(\tilde{\Gamma}_k - p_k q_k{}^\intercal \right) + \sum_k a_k p_k q_k{}^\intercal - pq^\intercal \Big\|_F \\ &\geq \Big\| \sum_k a_k p_k q_k{}^\intercal - pq^\intercal \Big\|_F - \Big\| \sum_k a_k \left(\tilde{\Gamma}_k - p_k q_k{}^\intercal \right) \Big\|_F \\ &\geq \Big\| \sum_k a_k p_k q_k{}^\intercal - pq^\intercal \Big\|_F - \eta_n. \end{aligned}$$

The last inequality is from Assumption 11-(ii) and $\mathbf{1}_K{}^\intercal a = 1$. We get $\Big\| \sum_k a_k p_k q_k{}^\intercal - pq^\intercal \Big\|_F \leq 2\eta_n + \tilde{\eta}$. Let

$$\rho(\varepsilon) = \min \left\{ \text{the second largest singular value of } \sum_k a_k p_k q_k{}^\intercal : \min_k \|a - e_k\|_\infty \geq \varepsilon, \mathbf{1}_K{}^\intercal a = 1 \right\}.$$

$\rho(\varepsilon)$ is decreasing in ε and bounded from above by ε; consider $(1+\varepsilon)p_1 q_1{}^\intercal - \varepsilon p_2 q_2{}^\intercal$. Also, from Assumption 11-(ii), $\rho(\varepsilon) > 0$ whenever $\varepsilon > 0$. By letting $\tilde{\varepsilon}_n$ such that $\rho(\tilde{\varepsilon}_n) = 3\eta_n + \tilde{\eta}$,

$$\Big\| \sum_k a_k p_k q_k{}^\intercal - pq^\intercal \Big\|_F \leq 2\eta_n + \tilde{\eta} \ \ \Rightarrow \ \ \min_k \|a - e_k\|_\infty < \tilde{\varepsilon}_n.$$

For any $a \in \mathcal{A}_n$, there exists some e_k that a is close to; the inequality restriction in (4) ensures that each of the rotated component distributions is close to one of the true component distributions.

It remains to show that the rotation retains all of the K component distributions. Suppose there exists some k such that $\|\alpha_l - e_k\|_\infty > \tilde{\varepsilon}_n$ for al l. Then, the k-th row of the $K \times K$ matrix $(\alpha_1, \cdots, \alpha_K)$ lies in $[0, \tilde{\varepsilon}_n]^K$. Thus, for any $\Lambda \in [0,1]^{K\times K}$, the k-th row of $(\alpha_1, \cdots, \alpha_K)\Lambda$ lies in

$[0, \tilde{\varepsilon}_n]^K$, leading to a contradiction. Thus, for small enough $\tilde{\varepsilon}_n$, the rotation matrix $(\alpha_1, \cdots, \alpha_K)$ is close to an identity matrix.

Step 3. There exist some A such that $\left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F = O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$ and $\Pr\{A \in \mathcal{A}_n\} \to 1$. Recall that $\left\|\widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0\Lambda_0\right\|_F = O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$. Let $E = \widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0\Lambda_0$. Then,

$$\widehat{\Gamma}_0 = \Gamma_0\Lambda_0\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1} + E\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$$

$\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ exists with probability converging to one, from a similar argument as in the proof of Theorem 2: with $\tilde{P}$ from Assumption 11-(iii),

$$\left\|\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0 - \tilde{P}\Gamma_0\Lambda_0\right\|_F \leq M \cdot O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right).$$

The determinant of $\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0$ converges to a nonzero constant as $M^3/n \to 0$. $\tilde{P}\widehat{\Gamma}_0$ is invertible and $\widehat{\Lambda}_0$ has full rank with probability converging to one. When $\left(\tilde{P}\widehat{\Gamma}_0\right)^{-1}$ and $\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ exist,

$$\left\|\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F \leq \left\|\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right\|_F \left\|\left(\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F \left\|\tilde{P}\widehat{\Gamma}_0\right\|_F.$$

$\left\|\tilde{P}\widehat{\Gamma}_0\right\|_F$ is bounded by K^2 and $\left\|\left(\tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\widehat{\Gamma}_0^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F$ converges to $\left\|\left(\tilde{P}\Gamma_0\Lambda_0\Lambda_0^\intercal\Gamma_0^\intercal\tilde{P}^\intercal\right)^{-1}\right\|_F$. $\left\|\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F$ is bounded. $\left\|E\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}\right\|_F = O_p\left(\frac{\sqrt{M}}{\sqrt{n}}\right)$.

It remains to show that $A := \Lambda_0\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1}$ belongs in $\mathcal{A}_n$. Firstly, find that

$$\begin{aligned}
\mathbf{1}_M{}^\intercal\widehat{\Gamma}_0 &= \mathbf{1}_M{}^\intercal\Gamma_0 A + \left(\mathbf{1}_M{}^\intercal\widehat{\Gamma}_0\widehat{\Lambda}_0 - \mathbf{1}_M{}^\intercal\Gamma_0\Lambda_0\right)\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1} \\
\mathbf{1}_K{}^\intercal &= \mathbf{1}_K{}^\intercal A + \left(\mathbf{1}_K{}^\intercal - \mathbf{1}_K{}^\intercal\right)\widehat{\Lambda}_0^\intercal\left(\widehat{\Lambda}_0\widehat{\Lambda}_0^\intercal\right)^{-1} \\
&= \mathbf{1}_K{}^\intercal A.
\end{aligned}$$

The first condition is satisfied. Secondly, let $\widehat{\tilde{\Gamma}}_k$ denote the $\widehat{\Gamma}_0$-equivalent of $\tilde{\Gamma}_k$. Find that

$$\begin{aligned}
min_{p,q}\left\|\sum_{l=1}^K \alpha_{kl}\tilde{\Gamma}_l - pq^\intercal\right\|_F &\leq \left\|\sum_{l=1}^K \alpha_{kl}\tilde{\Gamma}_l - \widehat{\tilde{\Gamma}}_k\right\|_F + min_{p,q}\left\|\widehat{\tilde{\Gamma}}_k - pq^\intercal\right\|_F \\
&\leq \left\|\sum_{l=1}^K \alpha_{kl}\tilde{\Gamma}_l - \widehat{\tilde{\Gamma}}_k\right\|_F + \eta_n \leq \left\|\widehat{\Gamma}_0 - \Gamma_0 A\right\|_F + \eta_n.
\end{aligned}$$

The second inequality is from the construction of $\widehat{\Gamma}_0$. The probabiltiy of the second condition being satisfied goes to one as $n \to \infty$. Lastly, find that

$$\begin{aligned}
\left\| A\widehat{\Lambda}_0 - \Lambda_0 \right\|_F &\leq \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P} \right\|_F \left\| \Gamma_0 \left(\Lambda_0 - A\widehat{\Lambda}_0 \right) \right\|_F \\
&\leq \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P} \right\|_F \left(\left\| \Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0 \right\|_F + \left\| \widehat{\Gamma}_0\widehat{\Lambda}_0 - \Gamma_0 A\widehat{\Lambda}_0 \right\|_F \right) \\
&= O_p(M) \cdot O_p\left(\frac{\sqrt{M}}{\sqrt{n}} \right).
\end{aligned}$$

The probability of the third condition being satisifies goes to one as $n \to \infty$.

Step 4. For an arbitrary $\varepsilon > 0$, $\Pr\left\{ \left\| \widehat{\Lambda}_0 - \Lambda_0 \right\|_F > \varepsilon \right\} \to 0$ as $n \to \infty$.

Find that

$$\begin{aligned}
\left\| \Lambda_0 - \widehat{\Lambda}_0 \right\|_F &\leq \left\| \Lambda_0 - \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0 \right\|_F + \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P}\widehat{\Gamma}_0\widehat{\Lambda}_0 - \widehat{\Lambda}_0 \right\|_F \\
&\leq \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P} \right\|_F \cdot \left\| \Gamma_0\Lambda_0 - \widehat{\Gamma}_0\widehat{\Lambda}_0 \right\|_F + \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P}\widehat{\Gamma}_0 - I_K \right\|_F \cdot \left\| \widehat{\Lambda}_0 \right\|_F \\
&\leq O_p\left(\frac{M\sqrt{M}}{\sqrt{n}} \right) + \left\| \widehat{\Lambda}_0 \right\|_F \cdot \left(\left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P} \left(\widehat{\Gamma}_0 - \Gamma_0 A \right) \right\|_F + \left\| \left(\tilde{P}\Gamma_0 \right)^{-1} \tilde{P}\Gamma_0 \left(A - I_K \right) \right\|_F \right) \\
&= O_p\left(\frac{M\sqrt{M}}{\sqrt{n}} \right) + \left\| \widehat{\Lambda}_0 \right\|_F \cdot \left\| A - I_K \right\|_F.
\end{aligned}$$

For some fixed $\tilde{\eta}$, $\Pr\left\{ A \in \bar{\mathcal{A}}_n \right\} \to 1$ as $n \to \infty$. Thus, we can relabel the types so that A is close to I_K. Then, for some $\tilde{\varepsilon} = \tilde{\varepsilon}(\tilde{\eta}) > 0$, $\Pr\left\{ \|A - I_K\|_F \leq K^2\tilde{\varepsilon}(\tilde{\eta}) \right\} \to 1$. By choosing $\tilde{\eta}$ appropriately for a given ε, we have

$$\Pr\left\{ \left\| \widehat{\Lambda}_0 - \Lambda_0 \right\|_F > \varepsilon \right\} \to 0$$

as $n \to \infty$.