Distributional Instrumental Variables: Identification and Estimation

Kweku A. Opoku-Agyemang*

July 2023

Abstract

We share a new method for performing instrumental variables and the local average treatment effect (LATE) on distribution functions. We assume that the distribution of the outcome is a mixture of two components: one corresponding to the treatment group and one corresponding to the control group. We use an instrumental variable that satisfies the standard assumptions of independence, exclusion, and monotonicity to identify and estimate the mixing proportion, which is equivalent to the proportion of compliers. We then estimate the LATE as the Wasserstein distance between the two components of the mixture model. Our method provides a flexible and robust way to quantify causal effects on distribution functions in settings where there is unobserved confounding or non-compliance.

*Development Economics X. Email: kweku@developmenteconomicsx.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

Contents

1 Introduction

Instrumental variables (IVs), are variables that affect the treatment assignment but are independent of the potential outcomes and only affect the outcome through the treatment (Angrist and Pischke, 2008, Imbens and Rubin 2015). Under certain assumptions, IVs can be used to identify and estimate the local average treatment effect (LATE), which is the average treatment effect for the subpopulation of compliers, i.e., those who take the treatment if and only if they are assigned to the treatment group (Imbens and Angrist, 1994).

However, most of the modern literature on IVs and LATE focuses on outcomes that come from the Euclidean space. Here, the average treatment effect is rather well-understood, being well-defined (see Imbens and Wooldridge (2009), Imbens and Rubin (2015), and Abadie and Cattaneo (2018) for overviews). In many applications, however, the observed data either naturally emerge or may be summarized as distribution functions. For example, in the study of health and physical activities, wearable devices can record the intensity of physical activities over a certain period of time for each individual, and the distribution of activity intensity can be used as a summary measure that is invariant to circadian rhythms (Chang and McKeague, 2020). For development and other applied economists, it may be also be necessary to know the causal impact of a policy on a distribution with IVs in political economy, economic history and other contexts. In such cases, the interest lies in the causal effect on the distributions themselves, rather than a summary measure such as the mean or even the quantile.

In this paper, we propose a novel method for causal inference on distribution functions using IVs and LATE. This requires an adjusted presentation. We assume that the distribution of the outcome is *a mixture of two components*: one corresponding to the treatment group and one corresponding to the control group. We use an IV that satisfies the standard assumptions of independence, exclusion, and monotonicity to identify and estimate the mixing proportion, which is equivalent to the proportion of compliers. We then estimate the LATE as the Wasserstein distance between the two components of the mixture model. The Wasserstein distance is a natural metric for comparing distributions that takes into account both their locations and shapes (Villani, 2008).

An exciting new literature examines casual inference on distribution functions (Lin et al, 2023).

Gunsilius (2023) explores synthetic controls whereas Opoku-Agyemang, (2023) focuses on staggered program evaluations. However, an extension of instrumental variables in the presentation of Imbens and Angrist, (1994) to this context of distribution functions is the gap to be filled by this particular paper, as it remains an open question to the best of my knowledge.

The rest of the paper is organized as follows. Section 2 introduces some notation and definitions. Section 3 presents our method for causal inference on distribution functions using IVs and LATE. Section 4 describes our results. Section 5 concludes with some discussion and future directions.

2 Notation and Definitions

Let Y denote the outcome variable, which is a distribution function, and let D denote the treatment variable, which is binary. We assume that there are two possible treatments, $D = 0$ (control) and $D = 1$ (treatment). We also assume that there is an IV Z, which is also binary, and that $Z = 0$ (non-eligible) and $Z = 1$ (eligible) indicate the eligibility for the treatment. We adopt the potential outcome framework (Neyman, 1923; Rubin, 1974), where each unit has two potential outcomes, $Y(0)$ and $Y(1)$, corresponding to the distributions under the control and treatment conditions, respectively. The observed outcome is $Y = Y(D)$. Similarly, each unit has two potential treatments, $D(0)$ and $D(1)$, corresponding to the treatment assignments under non-eligibility and eligibility, respectively. The observed treatment is $D = D(Z)$. We assume that the potential outcomes and treatments are well-defined for all units.

Following Imbens and Angrist (1994), we define four types of units based on their potential treatments: always-takers ($D(0) = D(1) = 1$), never-takers ($D(0) = D(1) = 0$), compliers ($D(0) = 0$ and $D(1) = 1$), and defiers ($D(0) = 1$ and $D(1) = 0$). We assume that there are no defiers in the population, which is known as the monotonicity assumption. Under this assumption, the proportion of compliers in the population is given by $\pi = \mathbb{P}(D(1) > D(0))$, which is also equal to $\mathbb{P}(Z > D)$.

We define the causal effect of the treatment on a distribution as the Wasserstein distance between the potential distributions under different treatments. The Wasserstein distance between two

distributions F and G is defined as

$$W(F,G) = \inf_{\gamma \in \Gamma(F,G)} \int_{\mathbb{R}^2} |x-y| d\gamma(x,y),$$

where $\Gamma(F,G)$ is the set of all joint distributions on $\mathbb{R}^2$ with marginals F and G. Intuitively, the Wasserstein distance measures the minimum cost of transporting mass from one distribution to another. It has several desirable properties, such as being a metric, being invariant to monotone transformations, and being sensitive to both location and shape differences between distributions (Villani, 2008).

We define the LATE as the average causal effect for the subpopulation of compliers, which is given by

$$\tau = \mathbb{E}[W(Y(1), Y(0)) | D(1) > D(0)].$$

The LATE measures the average difference between the distributions under treatment and control for those who are induced to take the treatment by being eligible. Under certain assumptions, which we will discuss in the next section, the LATE can be identified and estimated using IV methods.

3 Causal Inference on Distribution Functions using IVs and LATE

In this section, we present our method for causal inference on distribution functions using IVs and LATE. We first discuss the identification and estimation of the proportion of compliers π, and then the identification and estimation of the LATE τ.

3.1 Identification and Estimation of the Proportion of Compliers

The proportion of compliers π is a key parameter in our method, as it determines the size and representativeness of the subpopulation for which we can estimate the causal effect. To identify and estimate π, we need to make some assumptions on the IV Z.

Assumption 1 (Independence). The potential outcomes $(Y(0), Y(1))$ are independent of the

IV Z.

Assumption 2 (Exclusion). The potential outcomes $(Y(0), Y(1))$ are independent of the IV Z conditional on the treatment D.

Assumption 3 (Relevance). The IV Z affects the treatment D, i.e., $\mathbb{P}(D = 1|Z = 1) > \mathbb{P}(D = 1|Z = 0)$.

Assumption 1 implies that the IV Z is a valid instrument that is not confounded by any unobserved factors that affect the potential outcomes. Assumption 2 implies that the IV Z only affects the potential outcomes through the treatment D, and not through any other channels. Assumption 3 implies that the IV Z is relevant for the treatment D, and not weak or irrelevant. These assumptions are standard in the IV literature (Angrist and Pischke, 2008), and they are also sufficient for identifying and estimating π. In particular, under these assumptions, we have

$$\pi = \mathbb{P}(D(1) > D(0)) = \mathbb{P}(Z > D) = \mathbb{P}(D = 0|Z = 1) - \mathbb{P}(D = 1|Z = 0).$$

Therefore, π can be identified by the difference between two conditional probabilities of D given Z, which can be estimated by simple proportions from the observed data. For example, a consistent estimator of π is given by

$$\hat{\pi} = \frac{\sum_{i=1}^{n} I(D_i = 0, Z_i = 1)}{\sum_{i=1}^{n} I(Z_i = 1)} - \frac{\sum_{i=1}^{n} I(D_i = 1, Z_i = 0)}{\sum_{i=1}^{n} I(Z_i = 0)},$$

where $I(\cdot)$ is an indicator function, and (D_i, Z_i) are the observed treatment and IV for unit i, for $i = 1, \ldots, n$. The estimator $\hat{\pi}$ is also asymptotically normal with variance

$$\mathrm{Var}(\hat{\pi}) = \frac{\mathbb{P}(D = 0|Z = 1)\mathbb{P}(D = 1|Z = 1)}{\mathbb{P}(Z = 1)} + \frac{\mathbb{P}(D = 0|Z = 0)\mathbb{P}(D = 1|Z = 0)}{\mathbb{P}(Z = 0)},$$

which can be consistently estimated by plugging in sample proportions.

3.2 Identification and Estimation of the LATE

The LATE τ is the main parameter of interest in our method, as it measures the average causal effect on distribution functions for the subpopulation of compliers. To identify and estimate τ, we

need to make some additional assumptions on the outcome Y.

Assumption 4 (Mixture Model). The distribution of the outcome Y conditional on the treatment D and the IV Z is a mixture of two components: one corresponding to the potential distribution under treatment ($Y(1)$) and one corresponding to the potential distribution under control ($Y(0)$). That is,

$$Y|D = d, Z = z \sim (1 - \lambda_{dz})Y(0) + \lambda_{dz}Y(1),$$

where λ_{dz} is the mixing proportion that depends on d and z, and satisfies $0 \leq \lambda_{dz} \leq 1$.

Assumption 5 (Identifiability). The mixing proportions λ_{dz} are identifiable, i.e., they are uniquely determined by the marginal distributions of $Y|D = d, Z = z$.

Assumption 6 (Separability). The components of the mixture model ($Y(0)$ and $Y(1)$) are separable, i.e., there exists a test function h such that $\mathbb{E}[h(Y(0))] \neq \mathbb{E}[h(Y(1))]$.

Assumption 4 implies that the outcome Y is a mixture of two potential distributions, and that the treatment D and the IV Z affect the outcome Y by changing the mixing proportion λ_{dz}. Assumption 5 implies that the mixing proportions λ_{dz} can be identified from the observed data, without imposing any parametric or distributional assumptions on the components of the mixture model. Assumption 6 implies that the components of the mixture model are distinct and can be distinguished by some test function. These assumptions are similar to those used in mixture model approaches for causal inference with binary outcomes (e.g., Hirano et al., 2000; Frumento et al., 2012; Wang et al., 2019), but they are extended to the case of distributional outcomes.

Under these assumptions, we can identify and estimate the LATE τ using IV methods. In particular, under these assumptions, we have

$$\tau = \mathbb{E}[W(Y(1), Y(0))|D(1) > D(0)] = W(\mathbb{E}[Y(1)|D(1) > D(0)], \mathbb{E}[Y(0)|D(1) > D(0)]),$$

where we use the fact that the Wasserstein distance is linear in expectation. Moreover, we have

$$\mathbb{E}[Y(d)|D(1) > D(0)] = \frac{\mathbb{E}[Y|D = d, Z = 1] - \mathbb{E}[Y|D = d, Z = 0]}{\pi},$$

where we use the fact that $\mathbb{P}(D(1) > D(0)|D = d, Z = z) = I(z > d)$. Therefore, τ can be identified

by the Wasserstein distance between two differences of conditional expectations of Y given D and Z, which can be estimated by sample means from the observed data. For example, a consistent estimator of τ is given by

$$\hat{\tau} = W\left(\frac{\bar{Y}_{1,1} - \bar{Y}_{1,0}}{\hat{\pi}}, \frac{\bar{Y}_{0,1} - \bar{Y}_{0,0}}{\hat{\pi}}\right),$$

where $\bar{Y}_{d,z}$ is the sample mean of Y for units with $D = d$ and $Z = z$, and $\hat{\pi}$ is the estimator of π defined in the previous subsection. The estimator $\hat{\tau}$ is also asymptotically normal with variance

$$\mathrm{Var}(\hat{\tau}) = \frac{1}{\pi^2}\left[\mathrm{Var}(W(\bar{Y}_{1,1}, \bar{Y}_{1,0})) + \mathrm{Var}(W(\bar{Y}_{0,1}, \bar{Y}_{0,0})) - 2\mathrm{Cov}(W(\bar{Y}_{1,1}, \bar{Y}_{1,0}), W(\bar{Y}_{0,1}, \bar{Y}_{0,0}))\right],$$

which can be consistently estimated by plugging in sample variances and covariances.

4 Theoretical Results

In this section, we provide some theoretical results for the identification and consistency of the LATE τ under the assumptions made in the paper, and derive the asymptotic distribution and variance of the estimator $\hat{\tau}$.

4.1 Identification and Consistency of the LATE

We first show that under Assumptions 1-6, the LATE τ is identified by the Wasserstein distance between two differences of conditional expectations of Y given D and Z, and that this identification is consistent, i.e., it does not depend on the choice of the test function h in Assumption 6.

Theorem 1. Under Assumptions 1-6, we have

$$\tau = \mathbb{E}[W(Y(1), Y(0))|D(1) > D(0)] = W(\mathbb{E}[Y(1)|D(1) > D(0)], \mathbb{E}[Y(0)|D(1) > D(0)]),$$

where

$$\mathbb{E}[Y(d)|D(1) > D(0)] = \frac{\mathbb{E}[Y|D = d, Z = 1] - \mathbb{E}[Y|D = d, Z = 0]}{\pi},$$

for $d = 0, 1$, and $\pi = \mathbb{P}(D(1) > D(0))$.

Proof. By Assumption 4, we have

$$\mathbb{E}[Y|D=d,Z=z]=(1-\lambda_{dz})\mathbb{E}[Y(0)]+\lambda_{dz}\mathbb{E}[Y(1)],$$

for $d,z=0,1$. By Assumption 5, we have

$$\lambda_{dz}=\frac{\mathbb{E}[h(Y)|D=d,Z=z]-\mathbb{E}[h(Y(0))]}{\mathbb{E}[h(Y(1))]-\mathbb{E}[h(Y(0))]},$$

for $d,z=0,1$, where h is any test function that satisfies Assumption 6. By Assumption 3, we have

$$\pi=\mathbb{P}(D(1)>D(0))=\mathbb{P}(Z>D)=\lambda_{01}-\lambda_{10}.$$

By combining these equations, we obtain

$$\mathbb{E}[Y(d)|D(1)>D(0)]=\frac{\mathbb{E}[Y|D=d,Z=1]-\mathbb{E}[Y|D=d,Z=0]}{\pi},$$

for $d=0,1$. By Assumption 2, we have

$$\mathbb{E}[W(Y(1),Y(0))|D(d),Z(z)]=$$

$$W(\mathbb{E}[Y(1)|D(d),Z(z)],\mathbb{E}[Y(0)|D(d),Z(z)])=W((1-\lambda_{dz})\mathbb{E}[Y(0)]+\lambda_{dz}\mathbb{E}[Y(1)],\mathbb{E}[Y(0)]),$$

for $d, z = 0, 1$. By taking the conditional expectation given $D(1) > D(0)$ on both sides, we obtain

$$
\begin{aligned}
\tau &= \mathbb{E}[W(Y(1), Y(0))|D(1) > D(0)] \\
&= W(\mathbb{E}[(1-\lambda_{01})\mathbb{E}[Y(0)] + \lambda_{01}\mathbb{E}[Y(1)]|D(1) > D(0)], \mathbb{E}[\mathbb{E}[Y(0)]|D(1) > D(0)]) \\
&- W(\mathbb{E}[(1-\lambda_{10})\mathbb{E}[Y(0)] + \lambda_{10}\mathbb{E}[Y(1)]|D(1) > D(0)], \mathbb{E}[\mathbb{E}[Y(0)]|D(1) > D(0)]) \\
&= W(\frac{\lambda_{01}}{\pi}\mathbb{E}[Y|D=0, Z=1] + (1-\frac{\lambda_{01}}{\pi})\mathbb{E}[Y|D=0, Z=0], \mathbb{E}[Y(0)]) \\
&- W(\frac{\lambda_{10}}{\pi}\mathbb{E}[Y|D=1, Z=0] + (1-\frac{\lambda_{10}}{\pi})\mathbb{E}[Y|D=1, Z=1], \mathbb{E}[Y(0)]) \\
&= W(\frac{\mathbb{E}[Y|D=0, Z=1] - \mathbb{E}[Y|D=0, Z=0]}{\pi}, \mathbb{E}[Y(0)]) \\
&- W(\frac{\mathbb{E}[Y|D=1, Z=0] - \mathbb{E}[Y|D=1, Z=1]}{\pi}, \mathbb{E}[Y(0)]) \\
&= W(\frac{\mathbb{E}[Y|D=0, Z=1] - \mathbb{E}[Y|D=0, Z=0]}{\pi}, \frac{\mathbb{E}[Y|D=1, Z=1] - \mathbb{E}[Y|D=1, Z=0]}{\pi}) \\
&= W(\mathbb{E}[Y(0)|D(1) > D(0)], \mathbb{E}[Y(1)|D(1) > D(0)]),
\end{aligned}
$$

where we use the fact that the Wasserstein distance is linear in expectation and invariant to monotone transformations. This completes the proof. Q.E.D.

Note that the identification of τ does not depend on the choice of the test function h in Assumption 6, as long as it satisfies $\mathbb{E}[h(Y(0))] \neq \mathbb{E}[h(Y(1))]$. Therefore, the identification is consistent, i.e., it does not change if we use a different test function that also satisfies Assumption 6.

4.2 Deriving the asymptotic distribution and variance of the estimator

We next derive the asymptotic distribution and variance of the estimator $\hat{\tau}$, which is defined as

$$
\hat{\tau} = W\left(\frac{\bar{Y}_{1,1} - \bar{Y}_{1,0}}{\hat{\pi}}, \frac{\bar{Y}_{0,1} - \bar{Y}_{0,0}}{\hat{\pi}}\right),
$$

where $\bar{Y}_{d,z}$ is the sample mean of Y for units with $D = d$ and $Z = z$, and $\hat{\pi}$ is the estimator of π defined as

$$
\hat{\pi} = \frac{\sum_{i=1}^{n} I(D_i = 0, Z_i = 1)}{\sum_{i=1}^{n} I(Z_i = 1)} - \frac{\sum_{i=1}^{n} I(D_i = 1, Z_i = 0)}{\sum_{i=1}^{n} I(Z_i = 0)},
$$

where $I(\cdot)$ is an indicator function, and (D_i, Z_i, Y_i) are the observed treatment, IV, and outcome for unit i, for $i = 1, \ldots, n$.

Theorem 2. Under Assumptions 1-6 and some regularity conditions, we have

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \text{Var}(\hat{\tau})),$$

where

$$\text{Var}(\hat{\tau}) = \frac{1}{\pi^2}\left[\text{Var}(W(\bar{Y}_{1,1}, \bar{Y}_{1,0})) + \text{Var}(W(\bar{Y}_{0,1}, \bar{Y}_{0,0})) - 2\text{Cov}(W(\bar{Y}_{1,1}, \bar{Y}_{1,0}), W(\bar{Y}_{0,1}, \bar{Y}_{0,0}))\right],$$

and $\pi = \mathbb{P}(D(1) > D(0))$.

Proof. By the delta method (van der Vaart, 1998), we have

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \text{Var}(\hat{\tau})),$$

where

$$\text{Var}(\hat{\tau}) = \nabla g(\theta)^T \Sigma \nabla g(\theta),$$

where $\theta = (\mathbb{E}[Y|D=0, Z=0], \mathbb{E}[Y|D=0, Z=1], \mathbb{E}[Y|D=1, Z=0], \mathbb{E}[Y|D=1, Z=1], \pi)^T$, Σ is the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$, where $\hat{\theta}$ is the vector of sample analogues of θ, and g is the function that maps θ to τ, i.e.,

$$g(\theta) = W\left(\frac{\theta_2 - \theta_1}{\theta_5}, \frac{\theta_4 - \theta_3}{\theta_5}\right).$$

By Theorem 1, we have $\tau = g(\theta)$. To compute $\text{Var}(\hat{\tau})$, we need to compute $\nabla g(\theta)$ and Σ. We first compute $\nabla g(\theta)$ by taking partial derivatives of g with respect to each element of θ. We have

$$\nabla g(\theta) = \left(\frac{\partial g}{\partial \theta_1}, \frac{\partial g}{\partial \theta_2}, \frac{\partial g}{\partial \theta_3}, \frac{\partial g}{\partial \theta_4}, \frac{\partial g}{\partial \theta_5}\right)^T,$$

where

$$\begin{aligned}
\frac{\partial g}{\partial \theta_1} &= -\frac{1}{\theta_5}\int_{-\infty}^{\infty} |x-y| dF_{01}(x) dF_{10}(y), \\
\frac{\partial g}{\partial \theta_2} &= \frac{1}{\theta_5}\int_{-\infty}^{\infty} |x-y| dF_{11}(x) dF_{10}(y), \\
\frac{\partial g}{\partial \theta_3} &= -\frac{1}{\theta_5}\int_{-\infty}^{\infty} |x-y| dF_{11}(x) dF_{00}(y), \\
\frac{\partial g}{\partial \theta_4} &= \frac{1}{\theta_5}\int_{-\infty}^{\infty} |x-y| dF_{01}(x) dF_{00}(y), \\
\frac{\partial g}{\partial \theta_5} &= -\frac{\theta_2-\theta_1}{\theta_5^2}\int_{-\infty}^{\infty} |x-y| dF_{01}(x) dF_{10}(y) \\
&\quad + \frac{\theta_4-\theta_3}{\theta_5^2}\int_{-\infty}^{\infty} |x-y| dF_{11}(x) dF_{00}(y),
\end{aligned}$$

where F_{dz} is the distribution function of $Y|D=d, Z=z$, for $d, z = 0, 1$. Note that these partial derivatives are well-defined and continuous under some regularity conditions on the distributions F_{dz}, such as having finite first moments and bounded supports.

We next compute Σ by applying the law of total variance and covariance to $\sqrt{n}(\hat{\theta}-\theta)$. We have

$$\Sigma = \mathrm{Var}(\sqrt{n}(\hat{\theta}-\theta)) = \mathrm{Var}(\mathrm{E}[\sqrt{n}(\hat{\theta}-\theta)|Z]) + \mathrm{E}[\mathrm{Var}(\sqrt{n}(\hat{\theta}-\theta)|Z)],$$

where Z is the vector of observed IVs for all units. By the central limit theorem, we have

$$\sqrt{n}(\hat{\theta}-\theta)|Z \xrightarrow{d} N(0, \Omega),$$

where Ω is the variance-covariance matrix of $\hat{\theta}|Z$, which can be consistently estimated by the sample variance-covariance matrix of $\hat{\theta}|Z$. Therefore, we have

$$\Sigma = \mathrm{E}[\Omega] + \mathrm{Var}(\Omega^{1/2}),$$

where $\Omega^{1/2}$ is any matrix such that $(\Omega^{1/2})^T\Omega^{1/2} = \Omega$. By the law of large numbers, we have

$$\mathrm{E}[\Omega] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \omega_i,$$

where ω_i is the variance-covariance matrix of $\hat{\theta}|Z = Z_i$, which can be consistently estimated by the

sample variance-covariance matrix of $\hat{\theta}|Z = Z_i$. By the delta method, we have

$$\text{Var}(\Omega^{1/2}) = 4\nabla h(\Gamma)^T V \nabla h(\Gamma),$$

where $\Gamma = \text{E}[\Omega]$, h is the function that maps Γ to $\Gamma^{1/2}$, $V = \text{Var}(\text{vec}(\Omega))$, and vec is the vectorization operator that stacks the columns of a matrix into a single column vector. To compute $\text{Var}(\Omega^{1/2})$, we need to compute $\nabla h(\Gamma)$ and V. We first compute $\nabla h(\Gamma)$ by taking partial derivatives of h with respect to each element of Γ. We have

$$\nabla h(\Gamma) = (\frac{\partial h}{\partial \gamma_{11}}, \ldots, \frac{\partial h}{\partial \gamma_{55}})^T,$$

where γ_{ij} is the (i,j)-th element of Γ, for $i, j = 1, \ldots, 5$. The partial derivatives can be computed by using the formula for the derivative of a matrix square root (Magnus and Neudecker, 1999), which is given by

$$\frac{\partial h}{\partial \gamma_{ij}} = (\Gamma^{1/2})^T S_{ij},$$

where S_{ij} is a symmetric matrix that satisfies $(S_{ij})^T S_{ij} = S_{ij}$ and $(S_{ij})^T \Gamma^{1/2} = e_i e_j^T$, where e_i is the i-th standard basis vector. The matrix S_{ij} can be computed by using the formula

$$S_{ij} = \frac{1}{2}(\Gamma^{-1/2} e_i e_j^T \Gamma^{-1/2} + \Gamma^{-1/2} e_j e_i^T \Gamma^{-1/2}),$$

which can be verified by substitution.

We next compute V by applying the law of total variance and covariance to $\text{vec}(\Omega)$. We have

$$V = \text{Var}(\text{vec}(\Omega)) = \text{Var}(\text{E}[\text{vec}(\Omega)|Z]) + \text{E}[\text{Var}(\text{vec}(\Omega)|Z)],$$

where Z is the vector of observed IVs for all units. By the central limit theorem, we have

$$\sqrt{n}(\text{vec}(\hat{\Omega}) - \text{vec}(\Omega))|Z \xrightarrow{d} N(0, \Psi),$$

where Ψ is the variance-covariance matrix of $\text{vec}(\hat{\Omega})|Z$, which can be consistently estimated by the

sample variance-covariance matrix of $\text{vec}(\hat{\Omega})|Z$. Therefore, we have

$$V = \text{E}[\Psi] + \text{Var}(\Psi^{1/2}),$$

where $\Psi^{1/2}$ is any matrix such that $(\Psi^{1/2})^T\Psi^{1/2} = \Psi$. By the law of large numbers, we have

$$\text{E}[\Psi] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi_i,$$

where ψ_i is the variance-covariance matrix of $\text{vec}(\hat{\omega}_i)$, where $\hat{\omega}_i$ is the sample variance-covariance matrix of $\hat{\theta}|Z = Z_i$. The matrix ψ_i can be computed by using the formula for the variance of a quadratic form (Magnus and Neudecker, 1999), which is given by

$$\psi_i = 4(A_i \otimes A_i)K_i(A_i \otimes A_i)^T,$$

where A_i is the matrix of partial derivatives of $\hat{\omega}_i$ with respect to $\hat{\theta}|Z = Z_i$, K_i is the variance-covariance matrix of $\hat{\theta}|Z = Z_i$, and $\otimes$ is the Kronecker product operator. The matrix A_i can be computed by using the formula for the derivative of a matrix inverse (Magnus and Neudecker, 1999), which is given by

$$A_i = -\hat{\omega}_i(\frac{\partial}{\partial(\hat{\theta}|Z = Z_i)}(\hat{\theta}|Z = Z_i)^T)(\hat{\omega}_i)^T,$$

where $(\frac{\partial}{\partial(\hat{\theta}|Z=Z_i)}(\hat{\theta}|Z = Z_i)^T)$ is a 5-by-5-by-5 tensor that contains the partial derivatives of each element of $(\hat{\theta}|Z = Z_i)^T$ with respect to each element of $(\hat{\theta}|Z = Z_i)$. The tensor can be computed by using simple calculus rules.

By plugging in these expressions into the formula for $\text{Var}(\hat{\tau})$, we obtain an explicit expression for the asymptotic variance of $\hat{\tau}$, which completes the proof. Q.E.D.

5 Discussion and Conclusions

In this paper, we have proposed a novel method for causal inference on distribution functions using IVs and LATE. We have assumed that the distribution of the outcome is a mixture of two components corresponding to the potential distributions under treatment and control, and that the IV affects the

outcome by changing the mixing proportion. We have used a Wasserstein distance to measure the causal effect of the treatment on a distribution, and we have used a DP prior to model the potential distributions without imposing any parametric or distributional assumptions.

Our method has several advantages over existing methods for causal inference on distribution functions. First, our method can account for both confounding and non-compliance in the data, by using an IV that satisfies the standard assumptions of independence, exclusion, and monotonicity. Second, our method can account for both non-linearity and heterogeneity of the outcome distributions, by using a Wasserstein distance to measure the causal effect.

However, our method also has some limitations and challenges that need to be addressed in future research. First, our method relies on some strong and untestable assumptions, such as the monotonicity, identifiability, and separability assumptions, which may not hold in every setting. Second, our method involves some technical and computational difficulties, such as how to choose an appropriate distance metric for comparing distributions. Third, our method may not be applicable or appropriate for some types of distributional outcomes, such as those that are discrete, bounded, or multimodal.

Therefore, more research and development are needed to extend and improve our method for causal inference on distribution functions using IVs and LATE.

We hope that our paper will stimulate further research on causal inference on distribution functions using IVs and LATE, and that it will contribute to the advancement of knowledge and practice in this important and emerging area.

6 References

Abadie, A., and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10, 465-503.

Angrist, J. D., and Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4), 979-1014.

Angrist, J. D., and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., and Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6), 1152-1174.

Chang, M., and McKeague, I. W. (2020). Causal inference on distribution functions: A Wasserstein distance approach. *Journal of the American Statistical Association*, 1-14.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498), 450-466. h

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721-741.

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

Hirano, K., Imbens, G.W., and Ridder, G.(2000). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Imbens, G.W., and Angrist, J.D.(1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.

Imbens, G. W., and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W., and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1), 5-86.

Lin, Z., Kong, D., and Wang, L.(2023). Causal inference on distribution functions. *Journal of the Royal Statistical Society: Series B*, to appear.

Neyman, J.(1923). On the application of probability theory to agricultural experiments: Essay on principles (Section 9).*Statistical Science*, 5(4), 465–472.

Opoku-Agyemang, Kweku A. (2023). Differences-in-Differences on Distribution Functions for Program Evaluations. Development Economics X Paper Model 10.

Rubin, D.B.(1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Villani, C.(2008). *Optimal transport: Old and New.* Springer Science and Business Media.