Instrumental variable estimation with observed and unobserved heterogeneity of the treatment and instrument effect: A latent class approach

Pablo Rodriguez[1] and Mauricio Sarrias[2]

[1]Facultad de Economía y Negocios, Universidad de Talca, Talca, Chile. pablo.rodriguez@utalca.cl

[2]Facultad de Economía y Negocios, Universidad de Talca, Talca, Chile. mauricio.sarrias@utalca.cl

September 28, 2023

Abstract

This article presents a latent class approach for estimating the impact of a continuous treatment on an outcome variable while accounting for unobserved heterogeneity in both the treatment and the instrument effect, without requiring the assumption of monotonicity. Our approach, which is fully parametric and estimated using maximum likelihood estimator, allows for the parameters to vary across different classes (groups) of individuals. However, as the membership of each individual to a given class is unknown, we estimate it simultaneously with the parameters of each group, using a discrete distribution. We perform a Monte Carlo experiment to evaluate the performance of our estimator under assumptions similar to those of the traditional instrumental variable (IV) model. Our results indicate that the proposed estimator can accurately estimate the true degree of unobserved heterogeneity across classes and the population average treatment effect. Finally, we apply the proposed method to an empirical example to illustrate its practical implementations. Our proposed method provides a flexible and robust approach to estimating treatment effects in the presence of unobserved heterogeneity, and we believe it can be a valuable tool for researchers in a variety of fields.

1 Introduction

The instrumental variables (IV) estimator has become a cornerstone in estimating causal effects, particularly when traditional regression models fail due to endogeneity. Since the work of Imbens and Angrist (1994), it is well-known that the IV estimator provides a consistent estimate of the treatment effect for those who comply with the instrument's assignment (the compliers), under the assumption that the instrument satisfies the monotonicity condition.[1] This estimated causal effect, known as the local average treatment effect (LATE), provides valuable insights when monotonicity holds. However, when the treatment and instrument effects exhibit heterogeneity within the population, the IV estimator may no longer provide informative estimates for either the average treatment effect (ATE) or LATE.

Heterogeneity in treatment parameters and its implications have been thoroughly explored in the literature (e.g., Angrist, 2004; Wooldridge, 2005; Heckman et al., 2006). In general, when the treatment effects are heterogeneous the IV estimate still can be interpreted as the LATE parameter. However, when heterogeneity extends to instrument effects, a critical assumption must be met: all individuals affected by the instrument must experience effects in the same direction. Without this condition, the IV estimator may fail to represent any specific group (Heckman et al., 2006; Klein, 2010).[2]

Recent methodological advancements have relaxed the assumption of homogeneity. Wooldridge (1997, 2003) and Heckman and Vytlacil (1998) introduced a correlated random coefficient (CRC) model, which allows coefficients to correlate with the regressors, capturing unobserved heterogeneity in the endogenous variable-outcome relationship. Heckman and Vytlacil (1999, 2007) proposed to estimate several treatment effects parameters using local IV methods (LIV). The LIV estimator relies on the local linear regression approach, which consists of fitting a linear regression model using a kernel-weighted average of the observations in a neighborhood othe point where the regression is being evaluated. Benini and Sperlich (2022) also propose a semiparametric approach to deal with both observed and unobserved heterogeneity. The control function (CF) approach has also been suggested to estimate IV models with heterogeneous parameters (Wooldridge, 2005; Florens et al., 2008; Newey and Stouli, 2022). However, the identification in these approaches still hinges on the monotonicity assumption.

Some studies have tackled the issue of heterogeneity in first-stage parameters. Recently, Abadie et al. (2023) propose an estimator that assumes that the first-stage

[1] Monotonicity condition should also be satisfied with more than one instrument. For a more recent discussion see Mogstad et al. (2021).

[2] Klein (2010) provides approximations of the bias term (that depends on estimable quantities) when monotonicity does not hold.

correlation between the instrument and the endogenous variable is heterogeneous across groups of individuals. Groups are assumed to be known and based on observables. However, they also assume that the instrument affects the endogenous regressor in the same direction across all groups.

Abrevaya and Xu (2021) extend the classical IV model and allows the heteroskedasticity of the error term to depends upon the treatment variable so that treatment generates both mean and variance effects on the outcome. Their approach does not restrict causal interpretation to compliers.

This paper contributes to the literature of heterogeneous effects by proposing a parametric approach to estimate treatment and instrument effect parameters with observed and unobserved heterogeneity in the parameters. Building upon the latent class framework, we assume that the parameters in the structural and first-stage equation vary across classes or groups of individuals following an unknown discrete distribution.[3] Our approach does not require imposing any restriction on the domain of the instrument effect parameters, allowing the identification of the average treatment effect without the assumption of monotonicity, as long as the instrument satisfies the assumption of exogeneity and relevance. We use the maximum likelihood estimator to jointly estimate both the individuals' class-membership probabilities and the parameters of each class under the assumption of joint normality. We demonstrate the effectiveness of our method through a Monte Carlo study, and we illustrate its practical implementation in an empirical example (To be defined)

2 IV approach with latent classes

2.1 Intuition and causal parameters

Before introducing our model, it is essential to explore the implications of unobserved heterogeneity (both on treatment and instrument effect) and violation of the monotonicity assumption on the IV estimator.[4]

For simplicity consider a simple model with one endogenous continuous variable and one instrument. We focus in this simple case as it clarifies they key ideas. Assume that there exist $q = 1, \ldots, Q$ classes of individuals with different treatment and instrument effects in each class. Also, each individual $i = 1, \ldots, N$ belongs to one and only one class. The structural and first-stage equations are

$$\begin{aligned} y_{1i} &= \beta_0 + \gamma_q y_{2i} + \epsilon_i, \\ y_{2i} &= \delta_0 + \delta_{1q} z_i + \upsilon_i, \end{aligned} \tag{1}$$

[3] Our approach is similar to Sarrias (2021). However, he focuses on binary outcome variables and does not provide the consequences of unobserved and observed heterogeneity on the IV estimator.

[4] For simplicity in the exposition we focused on the IV estimator. However, our main conclusion also hold for the 2SLS estimator under the assumption outlined in Mogstad et al. (2021).

where $y_{1i} \in \mathbb{R}$ is the dependent variable for individual i; $y_{2i} \in \mathbb{R}$ is the endogenous treatment variable, i.e, $\mathbb{E}(\epsilon_i \upsilon_i) \neq 0$; $z_i \in \mathbb{R}$ is the exogenous instrument such that $\mathbb{E}(z_i \upsilon_i) = \mathbb{E}(z_i \epsilon_i) = 0$. The parameter γ_q is the class-specific impact of the treatment, whereas δ_{1q} is the class-specific effect of the instrument on y_{2i}. However, we do not know how these parameters vary across classes. All we know is that they vary in the population according to an unknown non-degenerate discrete distribution $(\gamma, \delta_1) \sim g(\gamma_q, \delta_{1q})$, where

$$g(\gamma_q, \delta_{1q}) = \begin{cases} (\gamma_1, \delta_{11}) & \text{with probability } \pi_1, \\ (\gamma_2, \delta_{12}) & \text{with probability } \pi_2, \\ \vdots & \vdots \\ (\gamma_Q, \delta_{1Q}) & \text{with probability } \pi_Q, \end{cases} \tag{2}$$

and it is independent of z_i. Since the probabilities are constant across individuals, π_q represents the share of individual in each class $q = 1, \ldots, Q$.

In general, our interest is to estimate the average treatment effect (ATE). Under our model (1) and the discrete distribution (2), the ATE in the population is simply

$$\text{ATE} = \mathbb{E}(\gamma) = \sum_{q=1}^{Q} \pi_q \gamma_q, \tag{3}$$

which is the weighted average of the treatment effect across classes.

Using Equations (1) and (2), and assuming that $\mathbb{E}(\delta_1) \neq 0$, it can be shown that the IV estimator will converge to (see Appendix B)

$$\widehat{\gamma}_{IV} \xrightarrow{p} \frac{\mathbb{C}(y_{1i}, z_i)}{\mathbb{C}(y_{2i}, z_i)} = \frac{\mathbb{E}(\gamma \delta_1)}{\mathbb{E}(\delta_1)} = \sum_{q=1}^{Q} \gamma_q \pi_q \left(\frac{\delta_{1q}}{\sum_{q=1}^{Q} \delta_{1q} \pi_q} \right). \tag{4}$$

It is well known that the IV estimator will deliver different causal effects based on different assumptions regarding monotonicity and the degree of heterogeneity of the instrument effect (Imbens and Angrist, 1994; Angrist, 2004). If we add the additional assumption of monotonicity (i.e, either $\delta_{1q} \geq 0$ or $\delta_{1q} \leq 0$ but not both, $\forall q$) then $\widehat{\gamma}_{IV}$ will consistently estimate the weighted local average treatment effect (WLATE), where the weights are given by $\left(\frac{\delta_{1q} \pi_q}{\sum_{q=1}^{Q} \delta_{1q} \pi_q} \right)$.

For example, assume that there are two classes of individuals in the population, $Q = 2$. The first class represents 70% of the population ($\pi_1 = 0.7$), while the remaining 30% belongs to the second class, ($\pi_2 = 0.3$). The treatment effects in each class are $\gamma_1 = 2$ and $\gamma_2 = 1$, respectively, so that the ATE in the population equals $\text{ATE} = (0.7)(2) + (0.3)(1) = 1.7$ (see Equation (3)).

Consider the implications of the compliers. If both classes are compliers such that $\delta_{11} = 1$ and $\delta_{12} = 0.5$, then $\gamma_{IV} = (2)(0.7)(1/0.85) + (1)(0.3)(0.5/0.85) = 1.82$ which is the weighted average of local average treatment effects. If the second class were composed of noncompliers (e.g., 'always-takers' and 'never-takers'), such that $\delta_{12} = 0$, then $\gamma_{IV} = 2$ which is just the treatment effect of the first class. Thus, WLATE provides no information about about the causal response of the second class because the instrument does not affect its treatment status. Note that the weights depend on the probability of compliance. A low probability of compliance in a group reduces the weight of its treatment parameter in γ_{IV}.

Now consider the implications of defiers. Assume the same previous set up, but now $\delta_{11} = 1$ and $\delta_{12} = -1$, i.e., the second class is now defier. Then, the IV estimator in Equation (4) will converge to $(0.7)(2)(1/0.4) + (0.3)(1)(-1/0.4) = 2.75$, which is not representative of any class (or group of individuals). This illustrate that the WLATE interpretation of the IV estimator requires some limitation of the degree of heterogeneity. Although some of the individuals (or groups of individuals in our example) may not respond to the instrument, the rest of the individuals should respond in the same direction to changes in the instrument.

Finally, if the instrument effect is homogeneous, ($\delta_{1q} = \delta_1 \neq 0, \forall q$), then the weights equal one and WATE = ATE = $\mathbb{E}(\gamma)$ (Heckman and Vytlacil, 1998; Wooldridge, 1997, 2003). Although this can be easily observed from Equation (4), this results can also be explained by the concept of essential heterogeneity (Heckman et al., 2006). Put simply, essential heterogeneity implies that the treatment and instrument effects are correlated, $\mathbb{C}(\gamma, \delta_1) \neq 0$. This arises for example if people make rational choices about whether to participate in any offered treatment (Heckman et al., 2006; Ebenstein, 2009; Ravallion, 2015). In other words, the assignment to treatment is correlated with receiving the treatment.[5] If the instrument effect is homogeneous, then is the case that $\mathbb{C}(\gamma, \delta_1) = 0$.[6]

Specifically, note that Equation (4) can also be written as:

$$\widehat{\gamma}_{IV} \xrightarrow{p} \frac{\mathbb{E}(\gamma)\mathbb{E}(\delta_1) + \mathbb{C}(\gamma, \delta_1)}{\mathbb{E}(\delta_1)}. \tag{5}$$

If $\mathbb{C}(\gamma, \delta_1) = 0$, then $\widehat{\gamma}_{IV} \xrightarrow{p}$ ATE. In other words, the higher the covariance between the treatment and instrument effect, the higher the difference between ATE and WLATE (see for example Masten and Torgovitsky, 2016; Klein, 2010).

[5] This is also known as selective take-up in randomized experiments. For more examples, see Breen and Ermisch (2021), Ravallion (2015) and Huntington-Klein (2020).

[6] Note that $\mathbb{C}(\gamma, \delta_1) = \mathbb{E}(\gamma, \delta_1) - \mathbb{E}(\gamma)\mathbb{E}(\delta_1)$. Since δ_1 is constant across classes, it is also independent of γ. Then $\mathbb{E}(\gamma, \delta_1) = \mathbb{E}(\gamma)\mathbb{E}(\delta_1)$. Now, consider our example above and assume that $\delta_{11} = \delta_{12} = 2$. Then $\mathbb{E}(\gamma) = 1.7$, $\mathbb{E}(\delta_1) = 2$, and $\mathbb{E}(\gamma, \delta_1) = (2)(0.7)(2) + (1)(0.3)(2) = 3.4 = \mathbb{E}(\gamma)\mathbb{E}(\delta_1)$.

2.2 Model and assumptions

Instead of assuming that the parameters vary continuously across individuals (c.f., Heckman and Vytlacil, 1998; Wooldridge, 1997), our approach accounts for heterogeneity by assuming that it arises from discrete groups of individuals. In this framework, unobserved heterogeneity is accommodated through a finite number, denoted as Q, of distinct and unobserved classes or segments of individuals, each characterized by unique values for both structural and first-stage parameters.

This formulation gives rise to the Latent Class Instrumental Variables (LCIV) model, structured as follows:

$$y_{1iq} = \mathbf{x}_{1iq}^\top \boldsymbol{\beta}_{1q} + \gamma_q y_{2iq} + \epsilon_{iq}, \tag{6}$$

$$= \mathbf{x}_{iq}^\top \boldsymbol{\beta}_q + \epsilon_{iq}, \tag{7}$$

$$y_{2iq} = \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q + \upsilon_{iq}, \tag{8}$$

where y_{1i} is the continuous dependent variable for individual $i = 1, \ldots, N$ which belongs to class $q = 1, \ldots, Q$; $\mathbf{x}_{iq} = (\mathbf{x}_{1iq}^\top, y_{2iq})^\top$ is a $K \times 1$ column vector of explanatory variables where y_{2iq} represents the continuous and endogenous treatment variable, and $\mathbf{x}_{1iq}$ is a set of predetermined (exogenous) variables; and $\mathbf{z}_{iq} = (\mathbf{x}_{1iq}^\top, \mathbf{x}_{2iq}^\top)^\top$ is a $P \times 1$ vector of predetermined variables where $\mathbf{x}_{2iq}$ is the vector of instruments (additional predetermined variables) for y_{2iq}.

Our goal is to estimate $\boldsymbol{\beta}_q$ and $\boldsymbol{\delta}_q$ for each $q = 1, \ldots, Q$, along with the ATE parameter. If we knew in advance which class each individual belongs to, and willing to assume monotonicity, then the parameters could be estimated using Abadie et al. (2023)'s approach.[7] However, this can lead to biased and inefficient estimates if heterogeneity is unobserved. We propose an approach that involves the simultaneous estimation of class-specific parameters and the assignment probability of each individual, assuming a discrete probability distribution

Assumption 1 (Distribution of random parameters)**.** *Assume that the number of classes Q is fixed and known. Let $\boldsymbol{\psi}_q = (\boldsymbol{\beta}_q^\top, \boldsymbol{\delta}_q^\top)^\top$ be a $(K+P)$-dimensional vector that collects the random parameters for Equations (7) and (8) for each class $q = 1, ..., Q$ and independent of both $\mathbf{z}_{iq}$ and $(\epsilon_{iq}, \upsilon_{iq}), \forall q$. The parameters $\boldsymbol{\psi}_q$ are assumed to vary across individuals following a discrete distribution:*

$$g(\boldsymbol{\psi}_q) = \begin{cases} \boldsymbol{\psi}_1 & \textit{with probability } \pi_{i1} \\ \boldsymbol{\psi}_2 & \textit{with probability } \pi_{i2} \\ \vdots & \vdots \\ \boldsymbol{\psi}_Q & \textit{with probability } \pi_{iQ} \end{cases}, \tag{9}$$

[7] See Assumption 1.2 in Abadie et al. (2023).

where individual i belongs to class q with probability π_{iq}, such that $\sum_{q=1}^{Q} \pi_{iq} = 1$, $\forall i$, and $\pi_{iq} > 0, \forall i, q$. Let $\boldsymbol{\Omega}_{\psi}$ be the $(K+P)\times(K+P)$ variance-covariance matrix of the random parameters. This matrix is assumed to be symmetric and positive definite.

Assumption 1 lays out the framework for how the random parameters in the LCIV model are distributed across individuals and classes. It assumes the number of latent classes, denoted by Q, is fixed and known in advance by the researcher. This is a common characteristic of latent class models but may require careful consideration when choosing an appropriate value for Q (For guidance on selecting Q, see McLachlan and Peel (2004)). It also assumes that the distribution of the random parameters, represented by $g(\boldsymbol{\psi}_q)$, is also independent of both $\mathbf{z}_{iq}$ and the error terms $(\epsilon_{iq}, \upsilon_{iq}), \forall q$. Thus, the ATE for γ in the population can be computed as a weighted sum of the treatment effects within each latent class, with weights given by the class assignment probabilities π_{iq}:

$$\text{ATE}_{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\sum_{q=1}^{Q}\pi_{iq}\gamma_q. \tag{10}$$

As shown in previous section, another feature of the latent class structure in Assumption 1 is that it allows for correlation between coefficients as long as the two coefficients in question take on more than one value across the Q classes (Hess, 2014).[8] This is an important feature of the model, as it accommodates situations where coefficients may vary together across classes (Heckman and Vytlacil, 1998; Wooldridge, 1997, 2003). Furthermore, since $\boldsymbol{\Omega}_{\psi}$ is assumed to positive definite the distribution of the random parameters is identified (see Gao and Pesaran, 2023).

Importantly, Assumption 1 does not impose any restriction on the sign and magnitude of $\boldsymbol{\delta}_q$. Consequently, it does not require the monotonicity assumption for the identification of either $\boldsymbol{\psi}_q$ or the ATE estimate.

The class-assignment probability π_{iq} is unknown. However, the following assumption proposes a semi-parametric multinomial logit formulation which links these probabilities to characteristics of individuals.[9]

Assumption 2 (Class-assignment probability)**.** *Let $\mathbf{h}_i$ be a $L \times 1$ i.i.d vector of exogenous variables having finite and nonsingular matrix $\mathbb{E}(\mathbf{h}_i\mathbf{h}_i^{\top})$. Assume that there exists a latent continuous variable F_{iq}^{*} that determines the class-assignment of individual i in class $q = 1, ..., Q$ and is given by a linear function of $\mathbf{h}_i$ such that $F_{iq}^{*} = \mathbf{h}_i^{\top}\boldsymbol{\lambda}_q + \xi_{iq}$, where $\boldsymbol{\lambda}_q$ is a vector of parameters to be estimated. Assuming that ξ_{iq} are i.i.d Extreme Value Type I, the probability for individual i to belong to a*

[8] In other words, it must be the case that $\boldsymbol{\psi}_1 < \boldsymbol{\psi}_2 < ... < \boldsymbol{\psi}_Q$.

[9] This is the most common specification rather than an absolute requirement (Hess, 2014).

particular class q is given by:

$$\pi_{iq}(\boldsymbol{\lambda}_q) = \frac{\exp\left(\mathbf{h}_i^\top \boldsymbol{\lambda}_q\right)}{\sum_{c=1}^{Q} \exp\left(\mathbf{h}_i^\top \boldsymbol{\lambda}_c\right)}, \quad q = 1, ..., Q. \tag{11}$$

The parameters for some class are normalized to zero for identification of the probabilities. We set $\boldsymbol{\lambda}_1 = \mathbf{0}$.

Equation (11) assigns each individual to one and only one class and ensures that $\sum_{q=1}^{Q} \pi_{iq} = 1$, $\forall i$, and $\pi_{iq} > 0, \forall i, q$. Note also that $\mathbf{h}_i$ in Equation (11) is a vector of exogenous variables that determines the assignment of each individual in a given class. Thus, our formulation accounts for both observed and unobserved sources of heterogeneity.

Additionally, the assumption that $\mathbb{E}(\mathbf{h}_i \mathbf{h}_i^\top)$ is finite and nonsingular matrix is crucial for the identification of the probabilities (McFadden, 1974).

Under Assumption 1 the probability π_{iq} varies across individuals based on their characteristics,$\mathbf{h}_i$. However, an alternative assumption can be made where the class allocation probabilities are considered constant across individuals, represented as $\pi_{iq} = \pi_q$ for all individuals, $i = 1, \ldots, N$. In this case, π_q reflects the population share of individuals in each latent class. The formulation for class probabilities in Equation (11) simplifies to:

$$\pi_q = \frac{\exp\left(\lambda_q\right)}{\sum_{c=1}^{Q} \exp\left(\lambda_c\right)}, q = 1, ..., Q, \tag{12}$$

such that $\lambda_1 = 0$ for identification.

Assumption 3 (Error terms)**.** *Within each class $q = 1, ..., Q$, assume that $(\epsilon_{iq}, \upsilon_{iq})$ is i.i.d (independent of $\mathbf{z}_{iq}$) distributed as a bivariate normal distribution as follows:*

$$(\epsilon_{iq}, \upsilon_{iq})|\mathbf{z}_{iq} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{q,\epsilon} & \rho_q \sigma_{q,\epsilon} \sigma_{q,\upsilon} \\ \rho_q \sigma_{q,\epsilon} \sigma_{q,\upsilon} & \sigma^2_{q,\upsilon} \end{pmatrix}\right].$$

Assumption 3 imposes homoskedasticity within each class, but heteroskedasticity across classes. Moreover, each class it is allowed to have different degree of endogeneity given by $\mathrm{corr}(\epsilon_{iq}, \upsilon_{iq}|\mathbf{z}_{iq}) = \rho_q$. If $\rho_q = 0, \forall q$, then the treatment variable is exogenous within each class and the parameters can be estimated using a LC model. If in addition $\boldsymbol{\beta}_q = \boldsymbol{\beta}, \forall q$, then we obtain the traditional linear regression model and the parameters can be estimated by OLS.

Assumption 3 implicitly states that that within each class the $P \times 1$ vector $\mathbf{z}_{iq}$ is exogenous: $\mathbb{E}(\epsilon_{iq}|\mathbf{z}_{iq}) = \mathbb{E}(\upsilon_{iq}|\mathbf{z}_{iq}) = 0, \forall q$.

We also assume the following regularity conditions needed for consistency and asymptotic normality of MLE presented in the next Section.

Assumption 4 (Data). *The sequence of observed data* $\{y_{1iq}, y_{2iq}, \mathbf{x}_{iq}, \mathbf{z}_{iq}\}_{i=1}^{N}$ *is i.i.d. for all* $q = 1, ..., Q$

Assumption 5 (Identification conditions). *(i) The* $P \times K$ *matrix* $\mathbb{E}(\mathbf{z}_{iq}\mathbf{x}_{iq}^{\top})$ *if full column rank, i.e., its rank equals* K *for all* $q = 1, ..., Q$. *(ii)* $\mathbb{E}(\mathbf{z}_{iq}\mathbf{z}_{iq}^{\top})$ *is positive definite and finite for all* $q = 1, ..., Q$. *(iii)* $\mathbb{E}(\mathbf{x}_{iq}\mathbf{x}_{iq}^{\top})$ *exists and is nonsingular for all* $q = 1, ..., Q$; *(iv)* $\mathbb{E}\|(z_{ipq}x_{ikq})^2\|$ *exists and is finite for all* $k = 1, ..., K$, $p = 1, ..., P$ *and* $q = 1, ..., Q$.

This assumption outlines specific identification conditions and properties of the instrumental variables and covariates within the LCIV model. Assumption 5(i) requires that $P \geq K$ for all $q = 1, ..., Q$. This means that $\mathbf{z}_{iq}$ is sufficiently linearly related to $\mathbf{x}_{iq}$ in each class (also known as the relevance assumption). This assumption is critical for the identification of $\boldsymbol{\gamma}_q$ if $\rho_q \neq 0$ for a given class.

Assumptions 5(ii)-(iii) are needed for identification and consistency. In particular, under 5(ii) $\boldsymbol{\delta}_q$ is identified for all $q = 1, \ldots, Q$.

Assumption 5(iv) implies that the fourth moment exists. This is needed so that the expected value of the hessian exists.

Assumption 6 (Compact parameter space). *The vector of parameters of the model* $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_Q, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_Q, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_Q, \rho_1, ..., \rho_Q, \sigma_{1,\epsilon}...\sigma_{Q,\epsilon}, \sigma_{1,\upsilon}...\sigma_{Q,\upsilon})$ *belongs to* $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\boldsymbol{\beta}} \times \boldsymbol{\Theta}_{\boldsymbol{\delta}} \times \boldsymbol{\Theta}_{\boldsymbol{\lambda}} \times \boldsymbol{\Theta}_{\boldsymbol{\rho}} \times \boldsymbol{\Theta}_{\boldsymbol{\sigma}_\epsilon} \times \boldsymbol{\Theta}_{\boldsymbol{\sigma}_\upsilon} \subset \mathbb{R}^{KQ} \times \mathbb{R}^{PQ} \times \mathbb{R}^{L(Q-1)} \times \mathbb{R}^{Q}_{(-1,1)} \times \mathbb{R}^{Q}_{++} \times \mathbb{R}^{Q}_{++}$, *a subset of the* $(K+P+3)Q+L(Q-1)$ *dimensional Euclidean space,* $\mathbb{R}^{(K+P+3)Q+L(Q-1)}$. *In addition,* $\boldsymbol{\Theta}$ *is a compact set and includes the true value of* $\boldsymbol{\theta}$, *denoted by* $\boldsymbol{\theta}_0$, *which is an interior point of* $\boldsymbol{\Theta}$.

The sample log-likelihood is not globally concave, thus compactness (Assumption 6) is essential for existence and consistency of MLE (see Newey and McFadden, 1994). To ensure that the sample log-likelihood function is bounded during optimization, we apply some transformation to $\sigma_{\epsilon,q}$, $\sigma_{\upsilon,q}$ and ρ_q to bound them away from their boundary. See Appendix C.

2.3 Maximum likelihood estimator

Since the model is fully parametric, we can estimate the parameters using MLE. Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1^{\top}, ..., \boldsymbol{\beta}_Q^{\top}, \boldsymbol{\delta}_1^{\top}, ..., \boldsymbol{\delta}_Q^{\top}, \boldsymbol{\lambda}_2^{\top}, ..., \boldsymbol{\lambda}_Q^{\top}, \rho_1, ..., \rho_Q, \sigma_{1,\epsilon}...\sigma_{Q,\epsilon}, \sigma_{1,\upsilon}...\sigma_{Q,\upsilon}\right)^{\top}$. Under Assumption 3 it follows that the joint distribution of (y_{i1q}, y_{2iq}), conditional on $\mathbf{z}_{iq}$ is

$$P_{i|q}(\boldsymbol{\zeta}_q) = \frac{1}{\sqrt{\left(1-\rho_q^2\right)\sigma_{q,\epsilon}^2}}\phi\left(\frac{(y_{1iq} - \mathbf{x}_{iq}^{\top}\boldsymbol{\beta}_q) - \frac{\sigma_{q,\epsilon}}{\sigma_{q,\upsilon}}\rho_q(y_{2iq} - \mathbf{z}_{iq}^{\top}\boldsymbol{\delta}_q)}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}\right)\frac{1}{\sigma_{q,\upsilon}}\phi\left(\frac{y_{2iq} - \mathbf{z}_{iq}^{\top}\boldsymbol{\delta}_q}{\sigma_{q,\upsilon}}\right), \tag{13}$$

where $\boldsymbol{\zeta}_q = \left(\boldsymbol{\beta}_1^{\top}, ..., \boldsymbol{\beta}_Q^{\top}, \boldsymbol{\delta}_1^{\top}, ..., \boldsymbol{\delta}_Q^{\top}, \rho_1, ..., \rho_Q, \sigma_{1,\epsilon}...\sigma_{Q,\epsilon}, \sigma_{1,\upsilon}...\sigma_{Q,\upsilon}\right)^{\top}$ and $\phi(\cdot)$ is the probability density function of the standard normal distribution (See Appendix D).

The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \left[\sum_{q=1}^{Q} \pi_{iq}(\boldsymbol{\lambda}_q) P_{i|q}(\boldsymbol{\zeta}_q) \right], \tag{14}$$

where $\pi_{iq}(\boldsymbol{\lambda}_q)$ is given in Equation (11). Thus the log-likelihood function is a mixture of $P_{i|q}(\boldsymbol{\zeta}_q)$ and $\pi_{iq}(\boldsymbol{\lambda}_q)$ is the mixing weight.

The MLE is a value of the parameter vector that maximizes the log-likelihood function:

$$\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}), \tag{15}$$

where $\boldsymbol{\Theta}$ denotes the parameter space in which the parameter vector $\boldsymbol{\theta}$ lies.

Under the assumptions made, the conditions of Theorem 2.5 in Newey and McFadden (1994) are met, so that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}}\right), \tag{16}$$

where $\mathbf{V}_{\boldsymbol{\theta}} = -\left(\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]\right)^{-1}$. See also Galimberti and Soffritti (2020). To estimate $\mathbf{V}_\theta$ we use the outer product of the individual score functions.

$$\widehat{\mathbf{V}}_{\boldsymbol{\theta}} = \left[\sum_{i=1}^{N} \mathbf{s}_i(\widehat{\boldsymbol{\theta}}) \mathbf{s}_i(\widehat{\boldsymbol{\theta}})^\top \right], \tag{17}$$

where $\mathbf{s}_i(\widehat{\boldsymbol{\theta}})$ is given in appendix D.

The conditions for consistency and asymptotic normality of the MLE correspond to high-level assumptions. In practice, when computing MLE for latent class models it is advisable that a number of different initial parameter vectors are considered in the optimization procedure.

After estimating the parameters, we can consistently estimate the ATE as:

$$\widehat{\mathrm{ATE}}_\gamma = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{Q} \widehat{\pi}_{iq} \widehat{\gamma}_q, \tag{18}$$

and its standard error can be estimated using Delta Method or bootstrap.

3 Monte Carlo experiment

3.1 Monte Carlo design

In this section, we aim to evaluate the finite sample properties of the LCIV estimator through two Monte Carlo experiments. For both experiments, we use a just-identified

model represented by the following equations:

$$y_{1iq} = \beta_{0q} + \gamma_q y_{2iq} + \epsilon_{iq},$$
$$y_{2iq} = \delta_{0q} + \delta_{1q} z_{iq} + \upsilon_{iq},$$

where y_{2iq} is the continuous treatment variable, and z_{iq} is the continuous independent instrument which follows a normal distribution with mean 0 and variance $\sigma^2_{zq} = 9$ for all q. The error terms (ϵ, υ) are assumed to follow a bivariate normal distribution with zero mean, and $\sigma_{q,\epsilon} = \sigma_{q,\upsilon} = 1$ for all q. In particular, the error term for the structural equation is created using $\epsilon_{iq} = [(\rho_q \sigma_{q,\epsilon})/\sigma_{q,\upsilon}]\, \upsilon_{iq} + \eta_{iq}$, where $\eta_{iq} \sim \mathrm{N}\left(0, \left[1 - \rho_q^2\right] \sigma^2_{q,\epsilon}\right)$ for all q.

Table 1 presents the parameter values for each of the two experiments. In each experiment, we assume the existence of two classes of individuals in the population $(Q = 2)$.

(Insert Table 1 about here)

In the first experiment, we assume a well-specified model. Since $\rho_q = 0.5$ for $q = 1, 2$, y_{2iq} endogenous in both classes. We also assume that the instrument has sufficient power for both classes. The structural and first-stage parameters, denoted as $\boldsymbol{\psi}_q = (\beta_{0q}, \gamma_q, \delta_{0q}, \delta_{1q})^\top$, take different values in each class and are distributed according to the following mass probability function:

$$g(\boldsymbol{\psi}) = \begin{cases} \boldsymbol{\psi}_1 = (\beta_{01}, \gamma_1, \delta_{01}, \delta_{11}) = (-1, -1, -1, -1) & \text{wp } \pi_1 = 0.3, \\ \boldsymbol{\psi}_2 = (\beta_{02}, \gamma_2, \delta_{02}, \delta_{12}) = (1, 2, 1, 2) & \text{wp } \pi_2 = 0.7. \end{cases}$$

This distribution implies that 30% of the population (the first class) has a negative treatment effect equal to $\gamma_1 = -1$, while the remaining 70% of the population (the second class) has a positive treatment effect equal to $\gamma_2 = 2$. Thus, the ATE in the population, which includes both compliers and non-compliers, is $\mathrm{ATE} = \mathbb{E}(\gamma) = \sum_{q=1}^{Q} \pi_q \gamma_q = (0.3)(-1) + (0.7)(2) = 1.1$. Since the second class represents a larger proportion of the population, its positive treatment effect dominates, resulting in an overall positive ATE.

The expectations for the rest of the parameters for the first experiment are $\mathbb{E}(\beta_0) = 0.4$, $\mathbb{E}(\delta_0) = 0.4$ and $\mathbb{E}(\delta_1) = 1.1$. The variance-covariance matrix of $\boldsymbol{\psi}$ is given by:

$$\boldsymbol{\Omega}_{\boldsymbol{\psi}} = \begin{pmatrix} \mathbb{V}(\beta_0) & \mathbb{C}(\beta_0, \gamma) & \mathbb{C}(\beta_0, \delta_0) & \mathbb{C}(\beta_0, \delta_1) \\ \cdot & \mathbb{V}(\gamma) & \mathbb{C}(\gamma, \delta_0) & \mathbb{C}(\gamma, \delta_1) \\ \cdot & \cdot & \mathbb{V}(\delta_0) & \mathbb{C}(\delta_0, \delta_1) \\ \cdot & \cdot & \cdot & \mathbb{V}(\delta_1) \end{pmatrix} = \begin{pmatrix} 0.84 & 1.26 & 0.84 & 1.26 \\ \cdot & 1.89 & 1.26 & 1.89 \\ \cdot & \cdot & 0.84 & 1.26 \\ \cdot & \cdot & \cdot & 1.89 \end{pmatrix}.$$

The causal effect of the instrument is also heterogeneous, as $\delta_{11} = -1$ and $\delta_{12} = 2$. The instrument increases the treatment for class 2 but decreases it for class 1. Consequently, the monotonicity assumption does not hold, and the IV estimator of the treatment will converge to $\widehat{\gamma}_{\text{IV}} \xrightarrow{p} \sum_{q=1}^{Q} \gamma_q \pi_q \left(\frac{\delta_{1q}}{\sum_{q=1}^{Q} \delta_{1q} \pi_q} \right) = 2.818$ (see Equation (4)). Note that this causal estimator is not representative of any group as it contains negative weights. Thus, the bias of the IV estimator with respect to ATE is $\text{bias}(\widehat{\gamma}_{\text{IV}}) = \frac{\mathbb{C}(\gamma,\delta_1)}{\mathbb{E}(\delta_1)} = \frac{1.89}{1.1} = 1.718$ (see Equation (5)).

In the second experiment, we investigate the finite sample properties of the LCIV estimator in a scenario where the instrument has no power for the first class. That is, condition (i) of Assumption 5 fails. Specifically, we assume:

$$g(\boldsymbol{\psi}) = \begin{cases} \boldsymbol{\psi}_1 = (\beta_{01}, \gamma_1, \delta_{01}, \delta_{11}) = (-1, -1, -1, 0) & \text{wp } \pi_1 = 0.3, \\ \boldsymbol{\psi}_2 = (\beta_{02}, \gamma_2, \delta_{02}, \delta_{12}) = (1, 2, 1, 2) & \text{wp } \pi_2 = 0.7. \end{cases}$$

Since the structural parameters in the first and second experiment are the same, then ATE = 1.1 in the population. Note that under this setup, the monotonicity assumption holds since $\delta_{1q} \geq 0$ for all q. In fact, since $\mathbb{E}(\delta_1) = 1.4$, the IV estimator will converge $\widehat{\gamma}_{IV} \xrightarrow{p} (-1)(0.3)\left(\frac{0}{1.4}\right) + (2)(0.7)\left(\frac{2}{1.4}\right) = 2$, which is the treatment effect for the second class.

In each experiment, we generate $S = 1000$ simulated samples, each with $N = \{100, 500, 1000, 5000\}$ observation. In each simulated sample, we compute the IV estimator, the LCIV estimator under correct specification $Q = 2$, and the LCIV estimator under a misspecified model $Q = 3$. The ATE for the LCIV models (both correct and misspecified model) in each sample $s = 1, ..., S$ is computed as

$$\text{ATE}_{\widehat{\gamma}} = \widehat{\pi}_1 \widehat{\gamma}_1 + \widehat{\pi}_2 \widehat{\gamma}_2, \tag{19}$$

where $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are the MLE for the treatment effect in each class, and

$$\begin{aligned} \widehat{\pi}_1 &= \frac{\exp(0)}{\exp(0) + \exp(\widehat{\lambda}_2)}, \\ \widehat{\pi}_2 &= \frac{\exp(\widehat{\lambda}_2)}{\exp(0) + \exp(\widehat{\lambda}_2)}. \end{aligned} \tag{20}$$

The MLE estimator for the LCIV model was implemented in R and is available for download at https://github.com/mauricio1986/ivhetLc. To speed-up the computational time, we used the BHHH optimization algorithm, and the variance-covariance matrix of the parameters, $\mathbf{V}_{\boldsymbol{\theta}}$, was estimated using Equation (17). The starting values for the optimization came from a two-equation LC model, assuming no endogeneity.

3.2 Monte Carlo results

3.2.1 Experiment 1

In this section, we present the findings from Experiment 1 (E1), which aimed to assess the finite sample properties of the LCIV estimator within the framework of a well-defined model. In this context, a well-defined model implies that both latent classes are considered as endogenous, and the instrument has power in both classes.

Figure 1 summarizes the distribution of estimated treatment effects (γ) and instrument effects (δ_1) based on 1000 Monte Carlo samples for varying sample sizes (N). Similarly, Figure 2 provides results for ρ and π. Each graph displays results for both class 1 (left column) and class 2 (right column). The red horizontal line represents the true population parameters detailed in Table 1. Table 2 complements these figures by offering insights into bias and Root Mean Square Error (RMSE).

(Insert Figure 1 about here)

(Insert Figure 2 about here)

(Insert Table 2 about here)

The examination of Figure 1 demonstrates unbiased estimates for both treatment and instrument effects, even with a relatively modest sample size of 100. This conclusion is corroborated by the bias statistics presented in Table 2. For example, the bias for γ_1 reduces from -0.0048 to approximately 0.0001 as the sample size increases from 100 to 5000, representing a substantial 102% reduction. Overall, bias tends to approach zero, and estimate variability significantly diminishes with larger sample sizes, as indicated by the RMSE values in Table 2.

A similar pattern is observed in Figure 2 for ρ and π. Generally, bias and RMSE for ρ are lower for the class with a higher proportion, as confirmed by Table 2.

Overall, the well-specified LCIV model consistently produces reliable estimates of treatment and instrument effect parameters for each class when the instrumental variable exhibits the desired properties.

Table 3 provides the mean and standard deviation for the causal parameters for each sample size.[10] ATE is the average treatment effect computed for the well-specified model, $Q = 2$. ATEM is the ATE computed for the miss-specified model, $Q = 3$. Both estimates are computed using Equation (19).

According to our previous discussion in Section 3.1, the ATE should converge to 1.1 as the sample size (N) increases without bound. Remarkably, both the LCIV models, one with the correct number of classes and the other with an incorrect

[10]The distribution of the causal parameters are reported in Figure A.1.

number of classes, produce unbiased ATE estimates. ATEM estimates show slightly more variability, but remain close to the true value. In contrast, the IV estimator exhibits significant bias toward the theoretical value of 2.818 due to the positive covariance between the treatment and instrument effects and the average instrument effect.

(Insert Table 3 about here)

Experiment 1 demonstrates the robustness and accuracy of the LCIV estimator in a well-specified model. It consistently produces unbiased estimates of treatment and instrument effect parameters, even with small sample sizes. These findings show the utility of the LCIV estimator in estimating causal effects when latent classes are endogenous and the instrument has power.

3.2.2 Experiment 2

In this section, we evaluate the behavior of the LCIV estimator and the causal parameters for Experiment 2 (E2). The setup in this experiment assumes that the instrument in the first class has no power, $\delta_{11} = 0$.

Table 4 reveals a striking disparity in the estimation of treatment effects for the two latent classes. Notably, the treatment effect for the first class exhibits significant positive bias that does not diminish with increasing sample size. The identification failure of γ_1 is a direct consequence of the lack of power of the instrument.

As expected, the estimates for ρ_1 are also biased. In fact, these estimates tend toward zero, leading to a negative bias. This observation underscores that the estimation of the degree of endogeneity is influenced by the instrument's strength. Thus, the validity of testing exogeneity for each class, $H_0 : \rho_q = 0$, hinges on the instrument's power.

Since the instrument is correlated with the endogenous variable in class 2, its estimates for the treatment effect remain unbiased. Thus, we are able at least to correctly identify the treatment effect for the class for which the instrument is strongly correlated with the endogenous variable.

(Insert Table 4 about here)

Turning our attention to Table 5, we observe the results for the causal parameters in Experiment 2. Both the Average Treatment Effect (ATE) and the Average Treatment Effect for the misspecified model (ATEM) estimates display an upward bias. This bias primarily stems from the substantial upward bias exhibited by γ_1. However, it's worth noting that the bias of the ATE estimate is somewhat mitigated by the fact that class 1 represents only 30% of the sample. In addition, the variability

of the ATE estimates in E2 is noticeably higher than that in E1, reflecting the challenging estimation conditions imposed by limited instrumental power.

As expected, the IV estimator converges to the treatment effect for the compliers. That is, it converges to the treatment of the second class, $\gamma_2 = 2$ even for moderate sample sizes.

In summary, the LCIV estimator faces difficulties in recovering the treatment effect for classes where endogeneity exists, and the instrument lacks power. As anticipated, this limitation also impacts the bias observed in the ATE estimator.

(Insert Table 5 about here)

4 Conclusion

References

Abadie, A., Gu, J., and Shen, S. (2023). Instrumental variable estimation with first-stage heterogeneity. *Journal of Econometrics*.

Abrevaya, J. and Xu, H. (2021). Estimation of treatment effects under endogenous heteroskedasticity. *Journal of Econometrics*.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The economic journal*, 114(494):C52–C83.

Benini, G. and Sperlich, S. (2022). Modeling heterogeneous treatment effects in the presence of endogeneity. *Econometric Reviews*, 41(3):359–372.

Breen, R. and Ermisch, J. (2021). Instrumental variable estimation in demographic studies: The late interpretation of the iv estimator with heterogenous effects. Technical report, Center for Open Science.

Ebenstein, A. (2009). When is the local average treatment close to the average? evidence from fertility and labor supply. *Journal of Human Resources*, 44(4):955–975.

Florens, J.-P., Heckman, J. J., Meghir, C., and Vytlacil, E. (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206.

Galimberti, G. and Soffritti, G. (2020). A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models. *Statistics & Probability Letters*, 157:108630.

Gao, Z. and Pesaran, M. H. (2023). Identification and estimation of categorical random coefficient models. *arXiv preprint arXiv:2302.14380*.

Heckman, J. and Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, pages 974–987.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3):389–432.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of econometrics*, 6:4875–5143.

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In *Handbook of choice modelling*, pages 311–330. Edward Elgar Publishing.

Huntington-Klein, N. (2020). Instruments with heterogeneous effects: Bias, monotonicity, and localness. *Journal of Causal Inference*, 8(1):182–208.

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

Klein, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *Journal of Econometrics*, 155(2):99–116.

Masten, M. A. and Torgovitsky, A. (2016). Identification of instrumental variable correlated random coefficients models. *Review of Economics and Statistics*, 98(5):1001–1005.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–98.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Stouli, S. (2022). Heterogeneous coefficients, control variables and identification of multiple treatment effects. *Biometrika*, 109(3):865–872.

Ravallion, M. (2015). On the implications of essential heterogeneity for estimating causal impacts using social experiments. *Journal of Econometric Methods*, 4(1):145–151.

Sarrias, M. (2021). A two recursive equation model to correct for endogeneity in latent class binary probit models. *Journal of Choice Modelling*, page 100301.

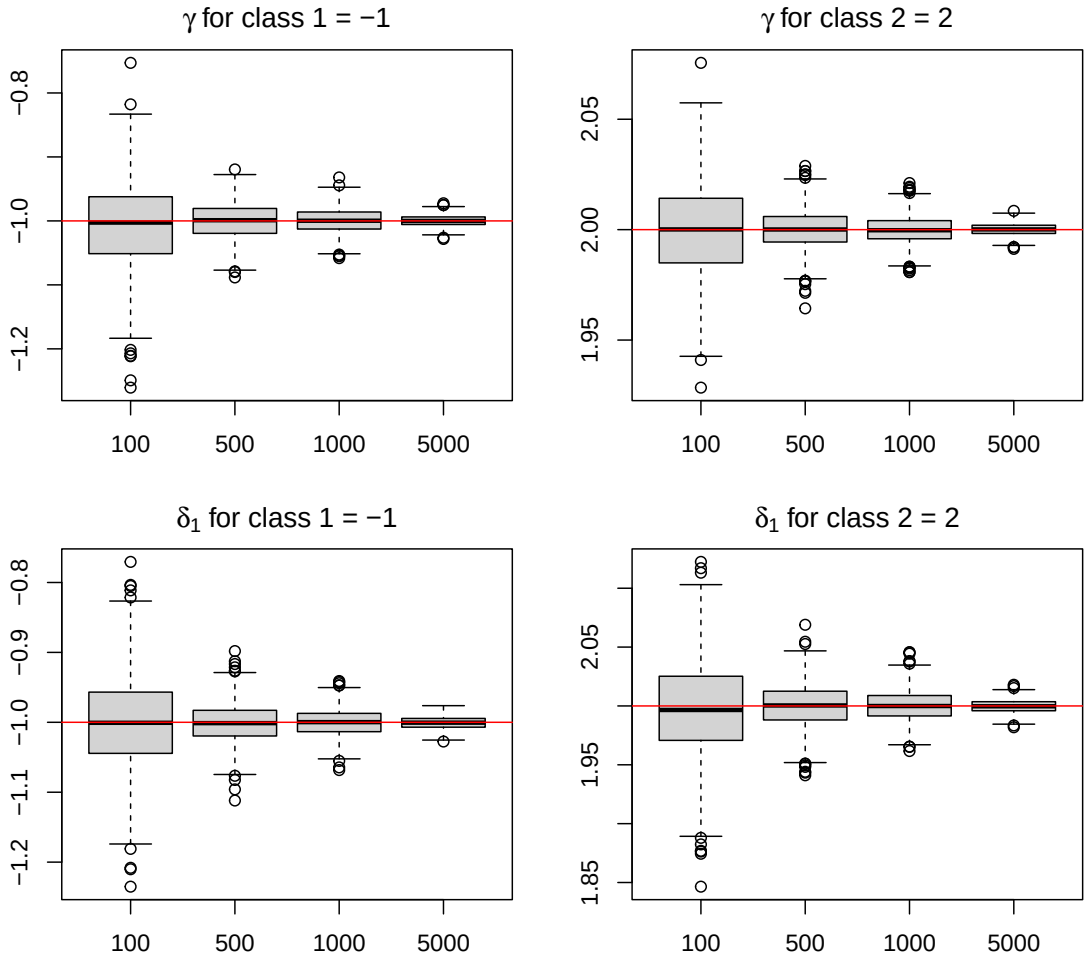
Wooldridge, J. M. (1997). On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics letters*, 56(2):129–133.

Wooldridge, J. M. (2003). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics letters*, 79(2):185–191.

Wooldridge, J. M. (2005). Unobserved heterogeneity and estimation of average partial effects. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, pages 27–55.

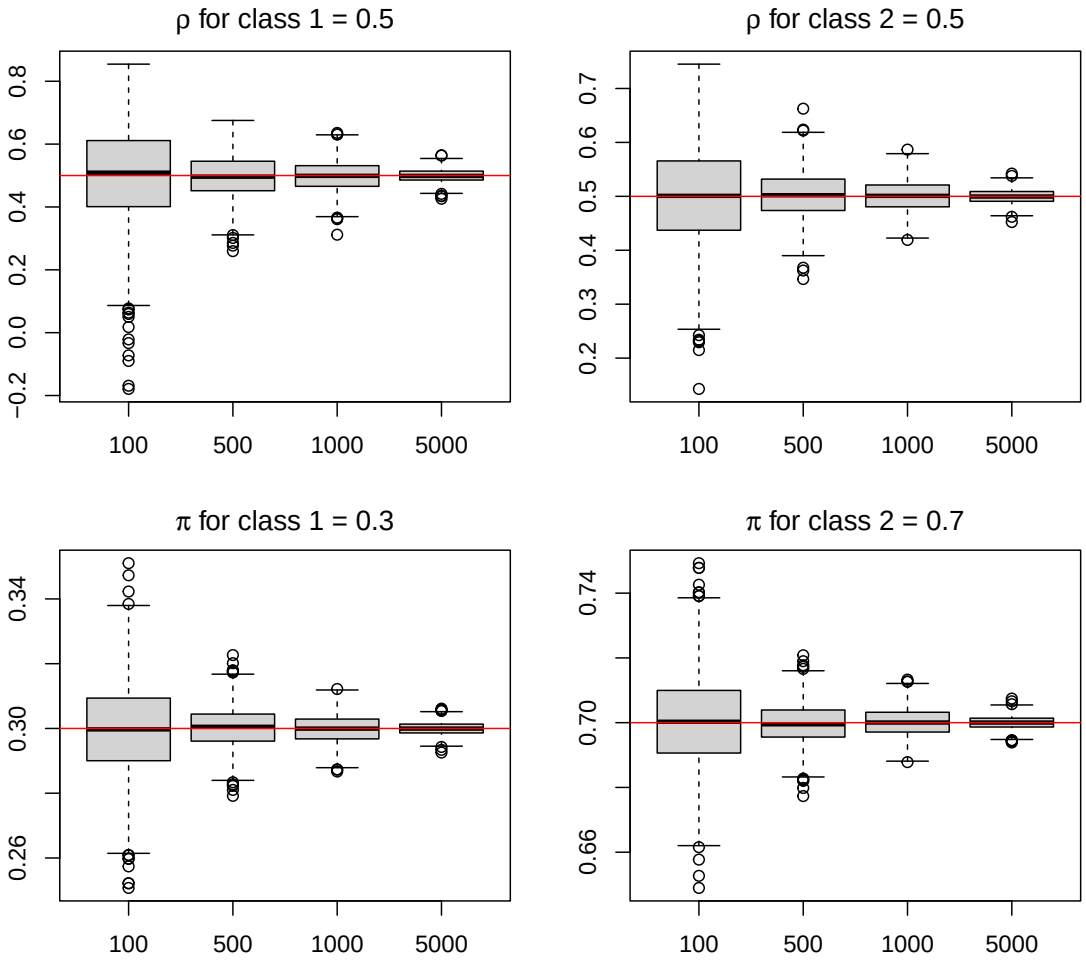
Figures

Figure 1: E1: Distribution of γ and δ_1 for both classes by sample size



Notes: Each figure shows the distribution of the parameters across the 1,000 MC samples. The red horizontal line represents the true value of the parameter.

Figure 2: E1: Distribution of ρ and π for both classes by sample size



Notes: Each figure shows the distribution of the parameters across the 1,000 MC samples. The red horizontal line represents the true value of the parameter.

Tables

Table 1: True values of the parameters for MC experiments

E1: *Endogeneity and power in both classes*								
	β_0	γ	δ_0	δ_1	ρ	σ_ϵ	σ_υ	π
Class 1	-1	-1	-1	-1	0.5	1	1	0.3
Class 2	1	2	1	2	0.5	1	1	0.7
E2: *Endogeneity, but no power in class 1*								
	β_0	γ	δ_0	δ_1	ρ	σ_ϵ	σ_υ	π
Class 1	-1	-1	-1	0	0.5	1	1	0.3
Class 2	1	2	1	2	0.5	1	1	0.7

Notes: Value of the true parameters in each class and experiment.

Table 2: E1: Simulation results for LCIV parameters

	$N = 100$		$N = 500$		$N = 1000$		$N = 5000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
γ_1	-0.0048	0.0657	-0.0003	0.0281	0.0004	0.0197	0.0001	0.0085
γ_2	-0.0001	0.0206	0.0000	0.0089	-0.0001	0.0063	0.0001	0.0028
δ_{11}	0.0012	0.0671	-0.0008	0.0277	0.0000	0.0200	-0.0007	0.0090
δ_{12}	-0.0023	0.0414	0.0003	0.0183	0.0001	0.0129	-0.0003	0.0056
ρ_1	-0.0026	0.1577	-0.0033	0.0718	-0.0014	0.0491	-0.0011	0.0215
ρ_2	-0.0017	0.0933	0.0016	0.0421	0.0003	0.0282	-0.0002	0.0134
π_1	-0.0004	0.0144	0.0003	0.0063	-0.0002	0.0044	-0.0000	0.0021
π_2	0.0004	0.0144	-0.0003	0.0063	0.0002	0.0044	0.0000	0.0021

Notes: Results for experiment 2. The number of MC samples is $S = 1000$. The bias for each parameter θ_k is computed as $(1/S)\sum_{s=1}^{S}(\widehat{\theta}_{ks} - \theta_{0k})$, where θ_{0k} is the true value of the parameter. The Means Squared Error is computed as $Bias_k^2 + \mathbb{V}(\widehat{\theta}_k)$, where $\mathbb{V}(\widehat{\theta}_k)$ is the variance of the estimates across the MC samples.

Table 3: E1: Simulation results for causal parameters

	$N = 100$		$N = 500$		$N = 1000$		$N = 5000$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ATE	1.0998	0.0499	1.0992	0.0221	1.1005	0.0150	1.1002	0.0072
ATEM	1.0986	0.0589	1.0996	0.0223	1.1003	0.0147	1.1001	0.0073
IV	2.8966	0.4008	2.8252	0.1487	2.8252	0.1070	2.8213	0.0465

Notes: ATE is the average treatment effect computed for the well-specified model, ($Q = 2$). ATEM is the ATE for computed for the miss-specified model, ($Q = 3$). In both cases, ATE are computed using Equation (19). The mean and standard deviation are computed for the $S = 1,000$ MC samples.

Table 4: E2: Simulation results for LCIV parameters

	$N = 100$		$N = 500$		$N = 1000$		$N = 5000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
γ_1	0.6626	9.9453	0.3207	8.2491	0.4009	6.9113	0.8328	5.6113
γ_2	0.0010	0.0206	0.0006	0.0088	0.0006	0.0065	0.0004	0.0032
δ_{11}	0.0004	0.0674	0.0001	0.0270	0.0004	0.0193	0.0005	0.0085
δ_{12}	0.0037	0.0407	0.0006	0.0180	0.0006	0.0131	0.0005	0.0059
ρ_1	-0.5119	0.8873	-0.4634	0.8537	-0.4579	0.8507	-0.5182	0.8924
ρ_2	-0.0208	0.1162	-0.0125	0.0660	-0.0127	0.0597	-0.0145	0.0583
π_1	0.0003	0.0193	-0.0002	0.0086	-0.0001	0.0064	0.0001	0.0052
π_2	-0.0003	0.0193	0.0002	0.0086	0.0001	0.0064	-0.0001	0.0052

Notes: Results for experiment 2. The number of MC samples is $S = 1000$. The bias for each parameter θ_k is computed as $(1/S)\sum_{s=1}^{S}(\widehat{\theta}_{ks} - \theta_{0k})$, where θ_{0k} is the true value of the parameter. The Means Squared Error is computed as $Bias_k^2 + \mathbb{V}(\widehat{\theta}_k)$, where $\mathbb{V}(\widehat{\theta}_k)$ is the variance of the estimates across the MC samples.

Table 5: E2: Simulation results for causal parameters

	$N = 100$		$N = 500$		$N = 1000$		$N = 5000$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ATE	1.2961	2.9173	1.1983	2.4607	1.2203	2.0548	1.3516	1.6582
ATEM	1.2495	2.6833	1.1821	2.5962	1.1729	2.0715	1.2541	1.5848
IV	2.0000	0.0451	2.0001	0.0187	2.0002	0.0133	1.9999	0.0059

Notes: ATE is the average treatment effect computed for the well-specified model, ($Q = 2$). ATEM is the ATE for computed for the miss-specified model, ($Q = 3$). In both cases, ATE are computed using Equation (19). The mean and standard deviation are computed for the $S = 1,000$ MC samples.

A Additional Figures

Figure A.1: E1: Distribution of causal parameters by sample size

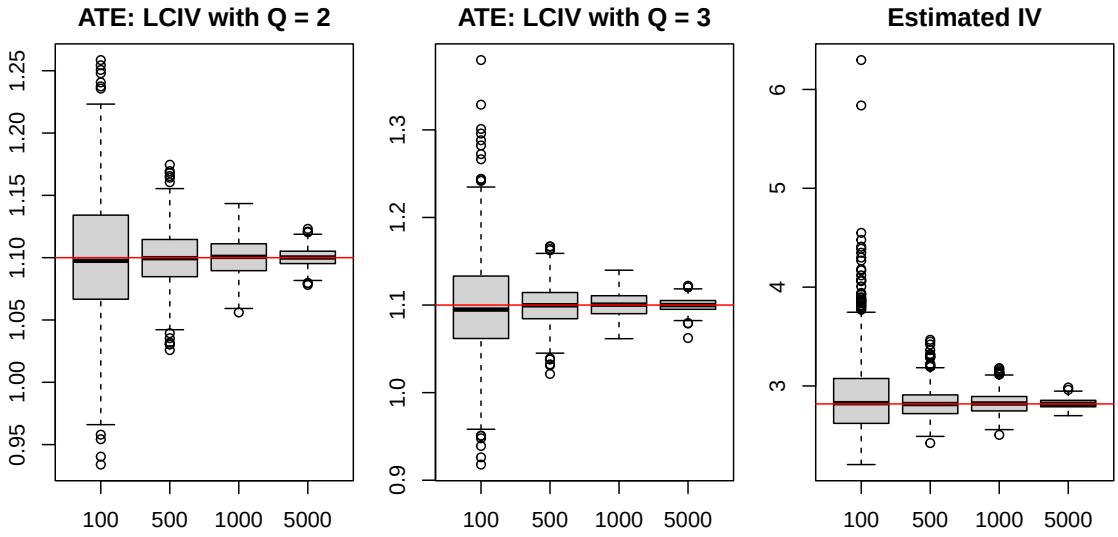


Figure A.2: E2: Distribution of γ and δ_1 for both classes by sample size

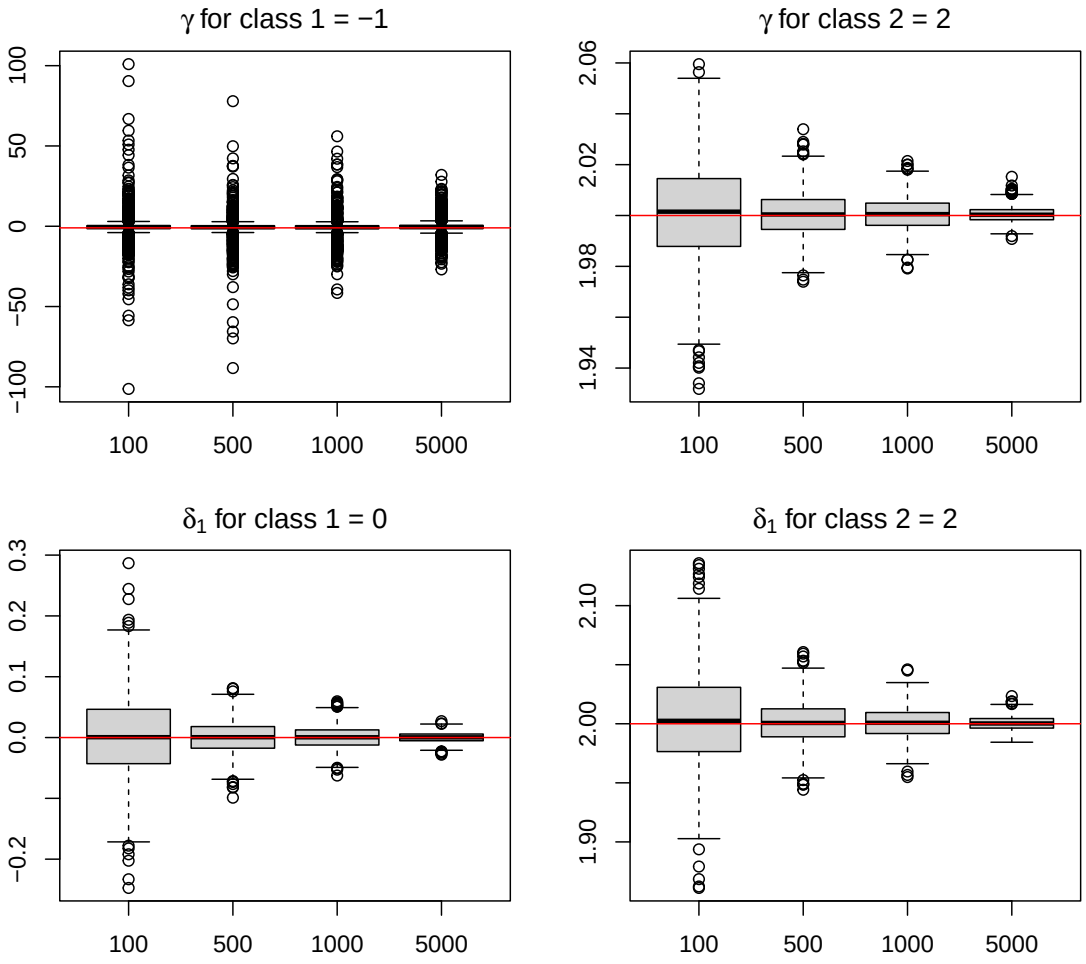


Figure A.3: E2: Distribution of ρ and π for both classes by sample size

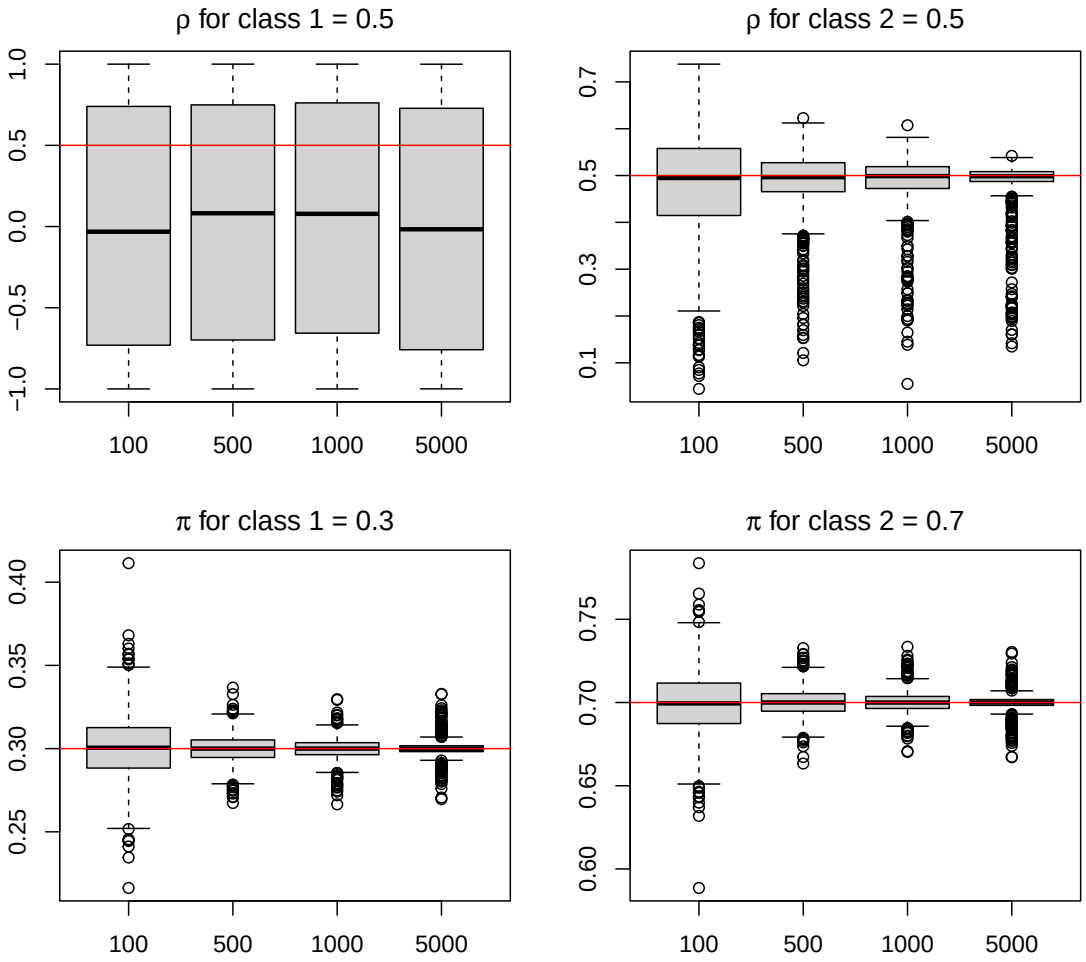
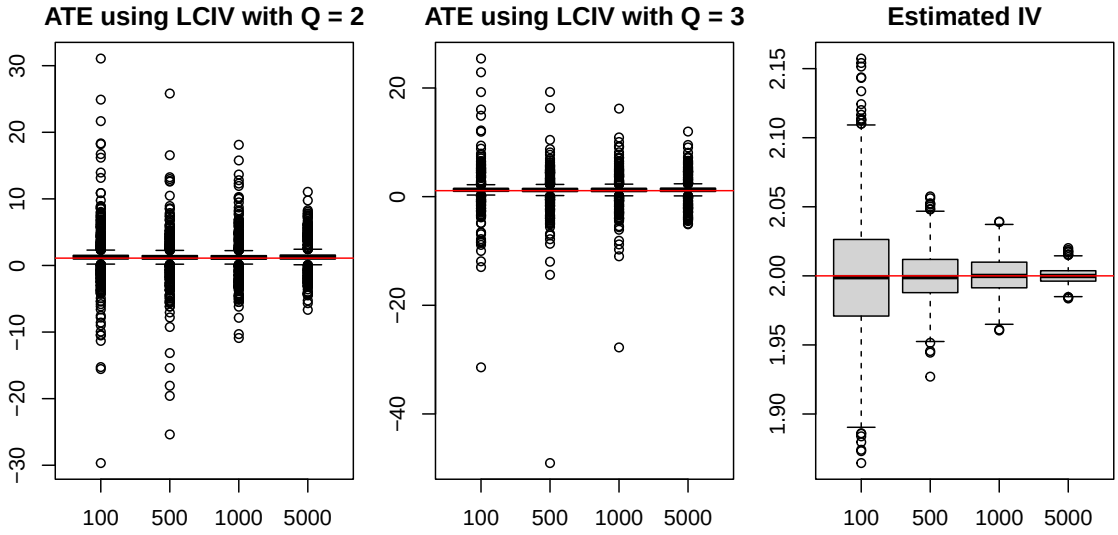


Figure A.4: E2: Distribution of causal parameters by sample size



B Appendix: IV estimator under treatment and instrument heterogeneity

For simplicity, assume a model with continuous heterogeneity:

$$\begin{aligned} y_{1i} &= \beta_{0i} + \gamma_i y_{2i} + \epsilon_i, \\ y_{2i} &= \delta_{0i} + \delta_{1i} z_i + \upsilon_i. \end{aligned} \tag{B.1}$$

It is well known that the IV estimator will converge to

$$\widehat{\gamma}_{\text{IV}} \xrightarrow{p} \frac{\mathbb{C}(y_{1i}, z_i)}{\mathbb{C}(y_{2i}, z_i)}. \tag{B.2}$$

Thus, using the model in Equation (B.1) and assuming that the random parameters are distributed independently of ϵ_i, υ_i, and z_i, the denominator of (B.2) is expressed as

$$\begin{aligned} \mathbb{C}(y_{2i}, z_i) &= \mathbb{C}(\delta_{0i} + \delta_{1i} z_i + \upsilon_i, z_i), \\ &= \mathbb{C}(\delta_{0i}, z_i) + \mathbb{C}(\delta_{1i} z_i, z_i) + \mathbb{C}(\upsilon_i, z_i), \\ &= \mathbb{C}(\delta_{1i} z_i, z_i), \\ &= \mathbb{E}(\delta_{1i})\, \mathbb{V}(z_i), \end{aligned} \tag{B.3}$$

where the third equality follows because $\mathbb{C}(\upsilon_i, z_i) = 0$ by assumption, and because $\mathbb{C}(\delta_{0i}, z_i) = \mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, \delta_{0i}\right] = \mathbb{E}\{\mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, \delta_{0i}\right] | z_i\} = \mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, \mathbb{E}(\delta_{0i} | z_i)\right] = \mathbb{E}\left(z_i - \mathbb{E}(z_i)\right) \mathbb{E}(\delta_{0i}) = 0$ by law of iterated expectations. The last equality follows because $\mathbb{C}(\delta_{1i} z_i, z_i) = \mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, \delta_{1i} z_i\right] = \mathbb{E}\{\mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, \delta_{1i} z_i\right] | z_i\} = \mathbb{E}\left[(z_i - \mathbb{E}(z_i))\, z_i\right] \mathbb{E}(\delta_{1i})$.

Similarly, for the numerator of (B.2) we obtain

$$\begin{aligned} \mathbb{C}(y_{1i}, z_i) &= \mathbb{C}(\beta_{0i} + \gamma_i y_{2i} + \epsilon_i, z_i), \\ &= \mathbb{C}(\beta_{0i}, z_i) + \mathbb{C}(\gamma_i y_{2i}, z_i) + \mathbb{C}(\epsilon_i, z_i), \\ &= \mathbb{C}(\gamma_i \delta_{1i} z_i, z_i), \\ &= \mathbb{E}(\gamma_i \delta_{1i})\, \mathbb{V}(z_i). \end{aligned} \tag{B.4}$$

Then, $\widehat{\gamma}_{\text{IV}} \xrightarrow{p} \frac{\mathbb{C}(y_{1i}, z_i)}{\mathbb{C}(y_{2i}, z_i)} = \frac{\mathbb{E}(\gamma_i \delta_{1i})}{\mathbb{E}(\delta_{1i})}$. Since in our setup the parameters are distributed assuming a discrete distribution, then

$$\widehat{\gamma}_{\text{IV}} \xrightarrow{p} \frac{\mathbb{E}(\gamma_i \delta_{1i})}{\mathbb{E}(\delta_{1i})} = \frac{\sum_{q=1}^{Q} \pi_q \gamma_q \delta_{1q}}{\sum_{q=1}^{Q} \pi_q \delta_{1q}}. \tag{B.5}$$

C Appendix: Parameter transformation

During the optimization procedure of the log-likelihood function the parameters $\sigma_{\epsilon,q}$, $\sigma_{\upsilon,q}$, and ρ_q $\forall q = 1, ..., Q$, might tend to the boundary points of the parameter space generating identifiability problems of the MLE. To avoid this issue, we make some re-parametrization of the parameters. First, to ensure $\sigma_{\epsilon,q} > 0$ and $\sigma_{\upsilon,q} > 0$, we rather estimate $\ln \nu_{\epsilon,q}$ and $\ln \nu_{\upsilon,q}$ such that

$$\begin{aligned} \sigma_{\upsilon,q} &= \exp(\ln \nu_{\epsilon,q}), \\ \sigma_{\epsilon,q} &= \exp(\ln \nu_{\epsilon,q}). \end{aligned} \tag{C.1}$$

Second, we force the correlation to remain in the $(-1, +1)$ interval by using the inverse hyperbolic tangent:

$$\operatorname{atanh}(\rho_q) = \tau_q = \frac{1}{2} \log \left(\frac{1 + \rho_q}{1 - \rho_q} \right), \tag{C.2}$$

where τ_q is unrestricted. Therefore, ρ_q can be obtained using the inverse of τ_q:

$$\tau_q^{-1} = \rho_q = \tanh(\tau_q). \tag{C.3}$$

D Appendix: Score function

In this section, we provide the gradient of the log-likelihood function using the transformations presented in Appendix C.

Given the class q that individual i belongs to, the joint distribution of (y_{1iq}, y_{2iq}), conditional on $\mathbf{z}_{iq}$, is $P_{i|q} = f_q(y_{1iq}|y_{2iq}, \mathbf{z}_{iq}) f_q(y_{2iq}|\mathbf{z}_{iq})$. Under Assumption 3, we can write $\epsilon_{iq}|\upsilon_{iq} = [(\rho_q \sigma_{q,\epsilon})/\sigma_{q,\upsilon}]\, \upsilon_{iq} + \eta_{iq}$, where $\eta_{iq} \sim \mathrm{N}\left(0, \left[1 - \rho_q^2\right] \sigma_{q,\epsilon}^2\right)$. Then, replacing the error term into Equation (7) yields:

$$y_{1iq}|y_{2iq} = \mathbf{x}_{iq}^\top \boldsymbol{\beta}_q + [(\rho_q \sigma_{q,\epsilon})/\sigma_{q,\upsilon}]\, \upsilon_{iq} + \eta_{iq}.$$

Since $\upsilon_{iq} = y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q$, then the probability of individual i conditional on y_{2iq} and $\mathbf{z}_{iq}$ is

$$\begin{aligned} f_q(y_{1iq}|y_{2iq}, \mathbf{z}_{iq}) &= \mathrm{N}\left(\mathbf{x}_{iq}^\top \boldsymbol{\beta}_q + \frac{\sigma_{q,\epsilon}}{\sigma_{q,\upsilon}} \rho_q (y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q), (1 - \rho_q^2) \sigma_{q,\epsilon}^2 \right), \\ &= \frac{1}{\sqrt{\left(1 - \rho_q^2\right) \sigma_{q,\epsilon}^2}} \phi \left(\frac{y_{1iq} - \mathbf{x}_{iq}^\top \boldsymbol{\beta}_q - \frac{\sigma_{q,\epsilon}}{\sigma_{q,\upsilon}} \rho_q (y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q)}{\sqrt{(1 - \rho_q^2) \sigma_{q,\epsilon}^2}} \right), \end{aligned} \tag{D.1}$$

where $\phi(\cdot)$ is the standard normal density function.

Since $y_{2i}|\mathbf{z}_{iq} \sim \mathrm{N}(\mathbf{z}_{iq}^\top \boldsymbol{\delta}_q, \sigma_{q,\upsilon}^2)$, the conditional (on q) marginal distribution is:

$$f_q(y_{2iq}|\mathbf{z}_{iq}) = \frac{1}{\sigma_{q,\upsilon}} \phi \left(\frac{y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q}{\sigma_{q,\upsilon}} \right). \tag{D.2}$$

Using Equations (D.1) and (D.2), we obtain Equation (13).

The contribution of each individual to the log-likelihood function is:

$$\ell_i(\boldsymbol{\theta}) = \ln P_i(\boldsymbol{\theta}) = \ln \left[\sum_{q=1}^{Q} \pi_{iq}(\boldsymbol{\lambda}_q) P_{i|q}(\boldsymbol{\zeta}_q) \right], \tag{D.3}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\zeta}_1^\top, \boldsymbol{\zeta}_2^\top, ..., \boldsymbol{\zeta}_Q^\top, \boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top, ..., \boldsymbol{\lambda}_Q^\top\right)^\top$, $\boldsymbol{\lambda}_1 = \mathbf{0}$ for identification, $\boldsymbol{\zeta}_q = (\boldsymbol{\beta}_q^\top, \boldsymbol{\delta}_q^\top, \sigma_{q,\epsilon}, \sigma_{q,\upsilon}, \rho_q)^\top$ is an $(K+P+3)$-dimensional vector, and $\boldsymbol{\lambda}_q$ is an L-dimensional vector for $q = 1, ..., Q$.

Taking the derivative of (D.3) with respect to $\boldsymbol{\zeta}_q$ yields

$$\begin{aligned} \underset{(K+P+3)\times 1}{\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\zeta}_q}} &= \left[\frac{1}{\sum_{q=1}^{Q} \pi_{iq} P_{i|q}} \right] \left[\pi_{iq} \frac{\partial P_{i|q}}{\partial \boldsymbol{\zeta}_q} \right], \\ &= \left[\frac{\pi_{iq} P_{i|q}}{\sum_{q=1}^{Q} \pi_{iq} P_{i|q}} \right] \left[\frac{1}{P_{i|q}} \frac{\partial P_{i|q}}{\partial \boldsymbol{\zeta}_q} \right], \\ &= \left[w_{iq} \frac{\partial \ln P_{i|q}}{\partial \boldsymbol{\zeta}_q} \right], \quad q = 1, ..., Q, \end{aligned} \tag{D.4}$$

where:

$$w_{iq} = \frac{\pi_{iq} P_{i|q}}{\sum_{q=1}^{Q} \pi_{iq} P_{i|q}} = \frac{\pi_{iq} P_{i|q}}{P_i},$$

is a weight so that $0 < w_{iq} < 1$ and $\sum_{q=1}^{Q} w_{iq} = 1$.

Similarly, taking the derivative of (D.3) with respect to $\boldsymbol{\lambda}_q$ yields

$$\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}_q} = \left[\frac{1}{P_i} \sum_{q=1}^{Q} P_{i|q} \pi_{iq} \frac{\partial \ln \pi_{iq}}{\partial \boldsymbol{\lambda}_q} \right],$$

where $\ln \pi_{iq} = \mathbf{h}_i^\top \boldsymbol{\lambda}_q - \ln \left(\sum_{c=1}^{Q} \exp \left(\mathbf{h}_i^\top \boldsymbol{\lambda}_c \right) \right)$. Thus

$$\frac{\partial \ln \pi_{iq}}{\partial \boldsymbol{\lambda}_q} = \mathbb{1}\left(q = c\right) \mathbf{h}_i - \pi_{iq} \mathbf{h}_i.$$

To derive $\partial \ln P_{i|j}(\boldsymbol{\zeta}_q) / \partial \boldsymbol{\zeta}_q$ in Equation (D.4), note that the logarithm of Equation

(13) is

$$\ln P_{i|q} = \ln(1) - \ln\left(\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right) + \ln\left[\phi(a_{iq})\right] + \ln(1) - \ln(\sigma_{q,\upsilon}) + \ln\left[\phi(b_{iq})\right],$$

where we use the following notation and transformations introduced in Appendix C:

$$\begin{aligned}
P_i &= \sum_{q=1}^{Q} \left(\pi_{iq} P_{i|q}\right), \\
P_{i|q} &= \left[\frac{1}{\sqrt{\left(1-\rho_q^2\right)\sigma_{q,\epsilon}^2}}\phi\left(a_{iq}\right)\right]\left[\frac{1}{\sigma_{q,\upsilon}}\phi\left(b_{iq}\right)\right], \\
a_{iq} &= \frac{y_{1iq} - \mathbf{x}_{iq}^\top \boldsymbol{\beta}_q - \frac{\sigma_{q,\epsilon}}{\sigma_{q,\upsilon}}\rho_q(y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q)}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}, \\
b_{iq} &= \frac{y_{2iq} - \mathbf{z}_{iq}^\top \boldsymbol{\delta}_q}{\sigma_{\upsilon,q}}, \\
\sigma_{\epsilon,q} &= \exp(\ln \nu_{\epsilon,q}), \\
\sigma_{\upsilon,q} &= \exp(\ln \nu_{\upsilon,q}), \\
\rho_q &= \tanh(\tau_q) = \frac{\exp(2\tau_q) - 1}{\exp(2\tau_q) + 1}.
\end{aligned}$$

Using this notation, the derivative $\partial \ln P_{i|q} / \partial \boldsymbol{\zeta}_q$ is

$$\begin{aligned}
\underset{(K\times 1)}{\frac{\partial \ln P_{i|q}}{\partial \boldsymbol{\beta}_q}} &= \left[\frac{\phi'(a_{iq})}{\phi(a_{iq})}\right]\left[\frac{\partial a_{iq}}{\partial \boldsymbol{\beta}_q}\right] = -a_{iq}\left[\frac{\partial a_{iq}}{\partial \boldsymbol{\beta}_q}\right], \\
\underset{(P\times 1)}{\frac{\partial \ln P_{i|q}}{\partial \boldsymbol{\delta}_q}} &= \left[\frac{\phi'(a_{iq})}{\phi(a_{iq})}\right]\left[\frac{\partial a_{iq}}{\partial \boldsymbol{\delta}_q}\right] + \left[\frac{\phi'(b_{iq})}{\phi(b_{iq})}\right]\left[\frac{\partial b_{iq}}{\partial \boldsymbol{\delta}_q}\right] = -a_{iq}\left[\frac{\partial a_{iq}}{\partial \boldsymbol{\delta}_q}\right] - b_{iq}\left[\frac{\partial b_{iq}}{\partial \boldsymbol{\delta}_q}\right], \\
\underset{(1\times 1)}{\frac{\partial \ln P_{i|q}}{\partial \ln \nu_{\epsilon,q}}} &= -1 + \left[\frac{\phi'(a_{iq})}{\phi(a_{iq})}\right]\left[\frac{\partial a_{iq}}{\partial \ln \nu_{\epsilon,q}}\right] = -1 - a_{iq}\left[\frac{\partial a_{iq}}{\partial \ln \nu_{\epsilon,q}}\right], \\
\underset{(1\times 1)}{\frac{\partial \ln P_{i|q}}{\partial \ln \nu_{\upsilon,q}}} &= -1 + \left[\frac{\phi'(a_{iq})}{\phi(a_{iq})}\right]\left[\frac{\partial a_{iq}}{\partial \ln \nu_{\upsilon,q}}\right] + \left[\frac{\phi'(b_{iq})}{\phi(b_{iq})}\right]\left[\frac{\partial b_{iq}}{\partial \ln \nu_{\upsilon,q}}\right], \\
&= -1 - a_{iq}\left[\frac{\partial a_{iq}}{\partial \ln \nu_{\upsilon,q}}\right] - b_{iq}\left[\frac{\partial b_{iq}}{\partial \ln \nu_{\upsilon,q}}\right], \\
\underset{(1\times 1)}{\frac{\partial \ln P_{i|q}}{\partial \tau_q}} &= \rho_q + \left[\frac{\phi'(a_{iq})}{\phi(a_{iq})}\right]\left[\frac{\partial a_{iq}}{\partial \tau_q}\right] = \rho_q - a_{iq}\left[\frac{\partial a_{iq}}{\partial \tau_q}\right].
\end{aligned}$$

where we use the fact that $\phi'(z) = -z\phi(z)$ so that $\phi'(z)/\phi(z) = -z$. The derivatives

of a_{iq} and b_{iq} are

$$\begin{aligned}
\frac{\partial a_{iq}}{\partial \boldsymbol{\beta}_q} &= -\left(\frac{1}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}\right)\mathbf{x}_{iq}, \\
\frac{\partial a_{iq}}{\partial \boldsymbol{\delta}_q} &= \left(\frac{1}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}\right)\left(\frac{\sigma_{q,\epsilon}\rho_q}{\sigma_{\upsilon,q}}\right)\mathbf{z}_{iq}, \\
\frac{\partial a_{iq}}{\partial \ln \nu_{q,\epsilon}} &= -\frac{\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}\right)}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}, \\
\frac{\partial a_{iq}}{\partial \ln \nu_{q,\upsilon}} &= \frac{\sigma_{q,\epsilon}\rho_q}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}b_{iq}, \\
\frac{\partial a_{iq}}{\partial \tau_q} &= \frac{\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q\right)\rho_q - b_{iq}\sigma_{q,\epsilon}}{\sqrt{(1-\rho_q^2)\sigma_{q,\epsilon}^2}}, \\
\frac{\partial b_{iq}}{\partial \boldsymbol{\delta}_q} &= -\left(\frac{1}{\sigma_{q,\upsilon}}\right)\mathbf{z}_{iq}, \\
\frac{\partial b_{iq}}{\partial \ln \nu_{q,\upsilon}} &= -b_{iq},
\end{aligned}$$

where $d\tanh(\tau_q)/d\tau_q = \operatorname{sech}^2(\tau_q) = 1-\tanh^2(\tau_q)$.

Finally, the contribution of each individual to the score function for each $q = 1, ..., Q$ is

$$\mathbf{s}_{iq}(\boldsymbol{\theta}) = \begin{pmatrix}
w_{iq}\left[\frac{(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q)-\eta_q(y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)}{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right]\mathbf{x}_{iq} \\
w_{iq}\left(-\left[\frac{\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q\right)-\eta_q(y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)}{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right]\eta_q + \left[\frac{y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q}{\sigma_{\upsilon,q}^2}\right]\right)\mathbf{z}_{iq} \\
w_{iq}\left(-1+\left[\frac{\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}\right)^2-\eta_q(y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}\right)}{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right]\right) \\
w_{iq}\left[-1-\left[\frac{(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q)(y_{2qi}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)-\eta_q(y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)^2}{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right]\eta_q + \left(\frac{y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q}{\sigma_{\upsilon,q}}\right)^2\right] \\
w_{iq}\left[\rho_q - \left[\frac{(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q)-\eta_q(y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q)}{(1-\rho_q^2)\sigma_{q,\epsilon}^2}\right]\left[\left(y_{1iq}-\mathbf{x}_{iq}^\top\boldsymbol{\beta}_q\right)\rho_q - \left(\frac{y_{2iq}-\mathbf{z}_{iq}^\top\boldsymbol{\delta}_q}{\sigma_{\upsilon,q}}\right)\sigma_{q,\epsilon}\right]\right] \\
\sum_{c=1}^{Q}\frac{P_{i|q}\pi_{ic}}{P_i}\left[\mathbb{1}\left(q=c\right)-\pi_{ic}\right]\mathbf{h}_i
\end{pmatrix} \tag{D.5}$$

where $\eta_q = (\sigma_{\epsilon,q}/\sigma_{\upsilon,q})\rho_q$. Then $\mathbf{s}_i(\boldsymbol{\theta}) = (\mathbf{s}_{iq}^\top(\boldsymbol{\theta}), ..., \mathbf{s}_{iQ}^\top(\boldsymbol{\theta}))^\top$