Abstract

In the presence of unmeasured confounders, we address the problem of treatment effect estimation from data fusion, that is, multiple datasets collected under different treatment assignment mechanisms. For example, marketers may assign different advertising strategies to the same products at different times/places. To handle the bias induced by unmeasured confounders and data fusion, we propose to separate the observational data into multiple groups (each group with an independent treatment assignment mechanism), and then explicitly model the group indicator as a Latent Group Instrumental Variable (LatGIV) to implement IV-based Regression. In this paper, we conceptualize this line of thought and develop a unified framework to (1) estimate the distribution differences of observed variables across groups; (2) model the LatGIVs from the different treatment assignment mechanisms; and (3) plug LatGIVs to estimate the treatment-response function. Empirical results demonstrate the advantages of the LatGIV compared with state-of-the-art methods.

1 Introduction

Estimating the causal effects of treatment/exposure on outcome of interest is crucial for explanatory analysis and decision-making [19,21,44]. To identify the causal effect of population, the gold standard approach is to perform Randomized Controlled Trials (RCTs) and then make an analysis, which is often unrealistic due to ethical and cost issues. Hence, in real-world scenarios, it is incredibly imperative and highly demanded to develop an automatic statistical approach to infer treatment-response function in observational studies. The main challenges are the confounding bias from the confounders which affect both treatment and outcome, and unmeasured differences hidden in data fusion — the average strategy is different from independent sub-strategies (Fig. 1(a)). Nevertheless, regardless of the approach that one adopts to control confounding in observational studies, there always exists the possibility of bias due to unmeasured confounders [8] and data fusion [4].
Figure 1: The pricing example for data fusion under different treatment regimes over times. (a) As market changes over times, policymakers adjust advertising strategies for different baby doll heights; (b) In causal graph, blue nodes denote observable variables, and gray nodes indicate latent variables. The arrows denote causal relationships. The bidirected arrows encode unmeasured confounders.

A powerful approach to address the unmeasured confounding is to utilize an instrumental variable (IV) [20,30] which are widely used in practice [27,33]. Based on the randomness of exogenous IV, [19,25,29] reformulate nonlinear two-stage approaches to (1) model the deconfounded treatment distribution using IV in stage 1; and (2) estimate the potential outcome [33] from predicted treatment (obtained from the stage 1) in stage 2. These IV methods are reliable when the pre-defined IV is a valid IV that only affects the outcome through its strong association with treatment options, called exclusion assumption. In practice, however, such valid IVs are hardly satisfied due to untestable exclusion association with outcome. In observational mixed data with several groups (each group with an independent treatment assignment mechanism), the group partitioning indicator serves as an exogenous IV related only to treatment. In addition, [19,32,37] denote the exogenous variables which generate mixed data as “anchors/IVs” and provide a solution with the prior knowledge of data.

In real-world scenarios, the observational data collected from different sources are often with different treatment assignment mechanisms. For instance, as market changes over times $Z$, policymakers would set different price $T$ to the same baby doll $X$ for sales $Y$, according to changing circumstances $f$. As shown in Fig. 1(a), the solid line shows the regression (MLR) for the entire population, while dashed lines are regressions for each subgroup. The unmeasured differences hidden in data fusion would fool the strategy regressed in data fusion. Besides, some unrecorded key variables $\epsilon$, e.g., environment and city, may confound the relationship between the observed variables due to the lack of prior knowledge (Fig. 1(b)). In the toy example (Fig. 1(b)), the time variable $Z$ is independent of the potential outcome $Y(T)$ conditional on the price $T$, assuming kids’ passion for doll’s height doesn’t change over time, i.e., $Z \perp X$. Therefore, we can take the latent time variable as a valid multi-valued IV, which only affects the outcome through its strong association with heterogeneous treatment assignment mechanisms [24,32].

Nevertheless, due to irregularities in collecting data and privacy issues, the time or source labels are sometimes missing in available datasets. Besides, the group IV is not available in some scenarios. For example, people tend to consult an expert consultant, and the consultant’s emotional state could be a group IV that can not be accessed. But these missing source labels in mixed data may provide motivation to recover IV. Recent IV-synthesis methods [10,11,14,17,24,42] synthesize an available summary IV from the well-predefined IV candidates, which suffers from the missing data (e.g., source labels) and poor prior for candidates (e.g., unknown genotype). To address the problem, we model latent IVs and implement a data-driven approach to automatically reconstruct valid Group IVs, that worked as well as source labels, directly from the observed variables.

In this paper, we focus on the problem of predicting the treatment-response function (TRF) from combined multiple datasets under different treatment assignment mechanisms, and propose to separate the data into multiple groups indicating the differences, and then explicitly model those heterogeneous groups as a latent IV to implement IV-based regression for estimating the TRF. Inspired by causal graph (Fig. 1(b)), we propose a novel Latent Group Instrumental Variable (LatGIV) reconstruction algorithm to automatically identify the group indicator that induces the different treatment assignment mechanisms. Specifically, we first map the covariates $X$ to representations $R$, then use the distribution
parameters of $R$ to initialize the joint distribution parameters of $T$ & $R$. With the Expectation-Maximization algorithm and correlation measurement, we can automatically and effectively identify the differentiated covariate-treatment distribution across groups and choose the appropriate group indicator as LatGIV by end-to-end learning. Then we can plug LatGIV into IV-based methods for treatment effect estimation.

The contribution of our paper is three-fold: (1) We model latent IVs and propose a data-driven LatGIV algorithm to automatically reconstruct valid Group IVs directly from the observed variables. (2) Theoretically, the identification of latent Group indicators is from the heterogeneous treatment assignment mechanisms across groups. LatGIV is effective when there are identifiable differences in treatment assignment mechanisms across groups. The reconstruction accuracy has reached 77% in Section 5.2. (3) We empirically demonstrate that the proposed algorithm reconstructs valid LatGIVs from the observed variables for accurate treatment-response function prediction, and gains state-of-the-art performance compared with existing IV-synthesis methods.

2 Related Work

Instrumental variable methods The sufficient identification results for causal effect under the homogeneity or monotonicity assumption in instrumental variable regression were developed by [18, 20, 30]. To implement semi-parametric and nonparametric estimation, there are four main research lines about IV methods, including: Two-Stage Least Squares [27, 28], Kernel-based methods [12, 29, 36], Deep-based methods [16, 25, 41], and Moment conditions methods [6, 15]. However, these methods are reliable only if the pre-defined IVs are valid and strongly correlated with the treatment variable. In practice, such valid IVs are hardly satisfied due to the untestable exclusion association. Thus, we aim to reconstruct a LatGIV to replace pre-defined IV in these methods.

Summary IV synthesis Our work is related to Summary IVs [11, 17, 24, 42]. To save the human effort selecting pre-defined IVs, a growing number of works have been proposed to synthesize a Summary IV by combining existing IV candidates. In Mendelian Randomization (MR), IV candidates are merged into a summary IV by unweighted/weighted allele scores (UAS/WAS) [10, 11, 24]. ModelIV [17] adopt the tightest cluster of estimation points from each IV candidate as general IVs. Assuming candidates are independent of the unmeasured confounders, AutoIV [42] generate IV representations by IV independence regularizer. Unlike the summary IV method which relies on candidates from expert knowledge, we model latent IVs and implement a data-driven approach to automatically reconstruct valid Group IVs directly from the observed variables.

Multiple datasets collected under heterogeneous conditions Motivated by latent multiple causal graph in data fusion [5, 43], anchor regression [32] denote the exogenous discrete variables which generate heterogeneity as "anchors/IVs" and provide an interpolation regression algorithm with prior knowledge [3, 40], model K-class estimators to estimate treatment-response function with known sub-populations. To address the limitation of prior knowledge, [19] proposes a unified framework for causal discovery and mechanism-based group identification via a latent causal graph clustering, which only works for linear and cannot adjust unmeasured confounding bias. Motivated by these literatures, we propose to reconstruct the sub-groups (LatGIV) from differentiated treatment mechanisms and using them as instrumental variables to reduce the bias induced by unmeasured confounders and data fusion jointly. Notably, the problem of data fusion differs from heterogeneous treatment effects, and the latter means explainable variability in treatment effects for individuals within a population [22, 38]. The tree-based methods and stratification-based methods [2, 13, 38] cannot adjust the unmeasured confounder. In this paper, we aim to recover IV for two-stage regression [1, 18, 20, 30].

3 Notation and preliminaries

3.1 Notations and assumptions

In this paper, we aim to estimate the treatment-response function from mixed data in the presence of unmeasured confounders. Under heterogeneous treatment assignment mechanisms, we denote the mixed data $D = \{x_i, t_i, y_i, z_i, u_i\}_{i=1}^n$. For each unit $i$, we observe a treatment variable $t_i \in T$ where $T \subset \mathbb{R}$, a outcome variable $y_i \in Y$ where $Y \subset \mathbb{R}$, and observed confounders $x_i \in X$
where \( X \subset \mathbb{R}^{m_X} \). For data fusion, the treatment assignment mechanisms (i.e., the associations between \( X \) and \( T \)) vary across groups of units, where the group ID \( z \in Z = \{1, \ldots, K\} \) is unmeasured/unrecorded, but can act as a latent valid IV as shown in Fig. 1(b). Besides, there are some unmeasured confounders \( u_i \in U, U \subset \mathbb{R}^{m_U} \), where the dimensions of the confounders \( X \) and \( U \), \( K \) is the number of sub-groups in populations. In this paper, we use \( X_{ij}, j = \{1, \ldots, m_X\} \) to denote the \( j \)-th variable in \( X \), \( x_i, i = \{1, \ldots, n\} \) denote the \( i \)-th samples’ feature, and \( x_{ij} \) denote the \( j \)-th dimension in \( i \)-th sample. Besides, \( X_{[k]} \) denotes the corresponding samples in Group \( Z = k \).

**Definition 3.1.** Treatment-Response Function (TRF): \( g(t, X) = \mathbb{E}[Y \mid do(T = t), X] \).

To address the bias induced by unmeasured confounders and data fusion, we propose to reconstruct the group ID \( Z \) that indicates different treatment mechanisms and use it as an instrument variable.

**Definition 3.2.** An Instrument Variable \( Z \) is an exogenous variable that only affects the outcome through its strong association with the treatment. Besides, an valid instrument variable satisfies the following three assumptions:

**Relevance:** \( Z \) is a cause of \( T \), i.e., \( \mathbb{P}(T \mid Z) \neq \mathbb{P}(T) \).

**Exclusion:** \( Z \) does not directly affect the outcome \( Y \), i.e., \( Z \perp Y \mid T, X, U \).

**Unconfounded:** \( Z \) is independent of all confounders, including \( X \) and \( U \), i.e., \( Z \perp X, U \).

Similar to [16, 25, 36, 41], we assume that the mixed data is generated by:

\[
T = f_{\theta_2}(X) + \epsilon_T^U(U) = f_{\theta_2}(X) + \epsilon_T^Z + g_T^U(U) = g_T(X, U) + \epsilon_Y, \tag{1}
\]

where the IV \( Z \) is independent of all confounders (i.e., \( Z \perp X, U \)) and indicates different treatment assignment mechanisms \( f_{\theta_2}(X) + \epsilon_T^U(U) \); the residuals are \( \epsilon_Y = g^U(U) \) and \( \epsilon_T = f_{\theta_2}(U) \). Heterogeneous treatment assignment mechanism is a common phenomenon in many scenarios, especially precision medicine. To identify the treatment-response function \( g(T, X) = \mathbb{E}[Y \mid Z, X] \), the additive noise assumption is required [18, 20, 30].

**Assumption 3.3. Additive Noise Assumption.** The residual \( \epsilon_Y \) has zero mean and finite variance but independent with the instruments \([6, 15, 16, 36]\). i.e., \( \mathbb{E}[\epsilon_Y] = 0, \mathbb{E}[\epsilon_Y^2] < \infty \) and \( \mathbb{E}[\epsilon_Y \mid Z] = 0 \).

### 3.2 Preliminary propositions

As shown in Fig. 1(b), the multiple datasets collected under heterogeneous conditions (e.g., different times \( Z \)) generate multiple causal relation \( f_{\theta_2}(X) \) between the covariates \( X \) and the treatment \( T \), i.e., different treatment assignment mechanisms \( T = f_{\theta_2}(X) + \epsilon_T^Z \) across groups \( Z = \{1, \ldots, K\} \). The time variable \( Z \) is independent of the potential outcome \( Y(T) \) given the price strategy \( T \), and the heterogeneous instrument effect is strongly related to \( T \). From that, we can take the time variable as a latent Group IV (LatGIV).

Inspired by different treatment assignment mechanisms, we further generate following preliminary propositions to reconstruct the group indicator that induces heterogeneity as LatGIV.

**Proposition 3.4. Representation.** There is a family \( \mathcal{H} \) of function \( h \) that encode the covariates \( X \) to low-dimension representation \( R = h(X) \), in which one encoding variable or a group of encoding variables combines the effect of corresponding covariate \( X_j \) on the treatment variable:

\[
T = f_{\theta_2}(X) + \epsilon_T = \sum_{z=1}^{K} \sum_{j=1}^{m_R} [\alpha_{zj} R_j + \beta_j] + \epsilon_T^z, \tag{2}
\]

where the residual \( \beta_z = \epsilon_T^z - \mathbb{E}[\epsilon_T^z \mid R, Z] \) is independent with \( R \) in group \( Z = \{1, \ldots, K\} \). Besides, \( m_R \) denotes the dimension of representation variable \( R \), and \( 1_{Z = z} \) equals 1 if \( Z = z \), 0 otherwise; \( \alpha_{zj} \) fit a linear (Non)-Gaussian Mixed Model, \( \beta_z \) is the corresponding residual. In linear scenario, \( h(X) = X \) is an identity function.

**Proposition 3.5. Differentiated Representation-Treatment Association across Groups.** The distribution \( \mathcal{P}(R \mid \mathbb{E}[R], Cov(R)) \) is same and fixed in all groups \( Z = \{1, \ldots, K\} \) (Unconfounded Assumption [3.2]), but the Representation-Treatment Association \( T_z = \alpha_{zj} R_j + \beta_z \) vary across Groups \( Z \). By identifying the association differences across groups, we can reconstruct the indicator \( Z \).

**Proposition 3.6. TRF Prediction.** Taking the expectation of outcome function in Eq. (1) conditional on \( \{Z, X\} \), we establish an inverse problem:

\[
\mathbb{E}[Y \mid Z, X] = \mathbb{E}[g(T, X) \mid Z, X] + \mathbb{E}[\epsilon_Y \mid X] = \int [g(T, X) + C] dF(T \mid Z, X), \tag{3}
\]

where \( dF(T \mid Z, X) \) is the conditional treatment distribution, \( C \) is constant as \( T \) is changed.
Proposition 3.4 combines the advantages of previous works \cite{26,41} and learn a representation map, which allows us to fit (non-)linear basis functions to encode the (non-)linear relationship between the covariates and the treatment as a linear (non-)gaussian mixed model of latent representation. Proposition 3.5 shows we can reconstruct the differentiated representation-treatment association to estimate the LatGIV. Proposition 3.6 identifies TRF under additive noise assumption. The proof of the propositions is deferred to Appendix A.

4 Algorithm

Guided by the above preliminary propositions and analyses, we propose a novel Latent Group Instrumental Variable reconstruction algorithm to automatically identify the Group Indicator that induces the different treatment assignment mechanisms. Specifically, we map the covariates \(X\) to low-dimension representations \(R\) by Representation Learning, and then fix the distribution of \(R\) to initialize the joint distribution parameters of \(T & R\). By estimating the latent differentiated covariate-treatment distribution parameters by Expectation-Maximization, we successfully implement the LatGIV Reconstruction and accurately predict treatment-response function by plugging LatGIV into downstream IV-based methods.

4.1 Representation and initialization learning

In this paper, we aim to identify the association differences \(f_{\theta_R}(X) + \epsilon_{TR}\) across groups and reconstruct the group indicator as LatGIV \(Z\). To simplify the reconstruct problem, we can learn and select an independent representation \(R = h(X)\) through PCA, VAE, correlation minimization or prior knowledge selection \cite{23,26}. In this paper, we use prior knowledge \(\text{Cov}(x_i, x_j) = 0\) to select and learn latent representation \(R = h(X)\) by polynomial regression, and set the rest dependent covariates as a part of unmeasured noise \(\epsilon_R\). Then we formulate a linear (non-)gaussian mixed model:

\[
T = f_{\theta_R}(X) + \epsilon_{TR} = \sum_{k=1}^{K} \sum_{r=1}^{R} \left\{ \sum_{j=1}^{m} [a_{rj} h_j(X)] + \beta_{j} \right\} \\
= \sum_{k=1}^{K} 1_{Z=k} \left\{ \sum_{j=1}^{m} [\alpha_{rj}(\xi_{rj} x_j^2 + \xi_{sj} x_j^2 + \cdots) + \beta_{j}] \right\} \\
= \sum_{k=1}^{K} 1_{Z=k} \left\{ \sum_{j=1}^{m} [a_{rj} R_j + \beta_{j}] \right\},
\]

where \(\xi_{rj,d}\) denotes the corresponding expectation coefficient of the \(d\)-th power of variable \(x_j\).

Then, we model the mixed data distribution \(\theta = \{\pi, \mu, \Sigma\} = \{\pi_z, \mu_z, \Sigma_z\}_{z=1}^{K}\) as:

\[
\mathcal{P}(T, R \mid \theta) = \mathcal{P}(T, R \mid \pi_k, \Sigma_k), k = \{1, \cdots, K\},
\]

where \(\pi_k = \mathcal{P}(Z = k)\) denotes the probability mass function of the Latent Group Assignment. Thus, we can estimate the posterior probability \(\gamma_k = \mathcal{P}(Z = k | T, R)\) by given \(\theta\):

\[
\gamma_k = \mathcal{P}(Z = k | T, R) = \frac{\mathcal{P}(Z = k) \mathcal{P}(T, R \mid Z = k)}{\sum_{k=1}^{K} \mathcal{P}(Z = k) \mathcal{P}(T, R)} = \frac{\pi_k \mathcal{P}(T, R \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{P}(T, R \mid \mu_k, \Sigma_k)}, k = \{1, \cdots, K\}.
\]

Then we model the complete-data distribution for heterogeneous treatment regime \(\{t_i, r_i, z_i\}_{i=1}^{n}\). The likelihood function for complete-data:

\[
\mathcal{P}(t, r, z | \theta) = \prod_{i=1}^{n} \mathcal{P}(t_i, r_i, z_i \mid \pi, \mu, \Sigma) = \prod_{k=1}^{K} \pi_k^{n_k} \prod_{i=1}^{n_k} [\mathcal{P}(t_i, r_i \mid \mu_k, \Sigma_k)]^{1_{z_i=k}},
\]

where \(n_k\) is the sample size of the \(k\)-th sub-group, \(k = \{1, \cdots, K\}\).

In the above content, we introduce a latent group data \(\{t_i, r_i, z_i\}_{i=1}^{n}\) and latent group variable’s posterior probability \(\gamma_k\). The Treatment \(T\) can be regard as a linear (non-)gaussian mixed model and the different treatment assignment mechanisms \(T_{Z=k} = \sum_{j=1}^{m} [a_{rj} R_j + \beta_{k}]\) vary across Groups \(Z = k\). Then, we estimate the distributed parameters \(\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}\):

\[
\pi_k = \mathcal{P}(Z = k), \mu_k = (E[T_{[k]}], E[R_{[k]}]), \Sigma_k = \begin{bmatrix} \sigma(T_{[k]}, T_{[k]}) & \sigma(T_{[k]}, R_{[k]}) & \sigma(T_{[k]}, R_{[k]}) \ \sigma(R_{[k]}, T_{[k]}) & \sigma(R_{[k]}, R_{[k]}) & \sigma(R_{[k]}, R_{[k]}) \end{bmatrix},
\]

where \(\{T_{[k]}, R_{[k]}\}\) denotes the corresponding treatment and representation in Group \(Z = k\), \(\sigma(\cdot, \cdot)\) denotes the covariance matrix between two random vectors. We assume the representation distribution is stable across groups, i.e., \(E[R_{[k]}] = E[R], \sigma(R_{[k]}, R_{[k]}) = \sigma(R, R)\). The Covariate-Treatment
Associations vary across Groups $Z = k$, leading to multiple fixed distribution parameters $\{\mathbb{E}[T_{[k]}], \sigma(T_{[k]}), \sigma(T_{[k]}, R_{[k]})\}$ in Groups $Z = k$.

Therefore, we can use the prior knowledge $\mathcal{P}(\mathbb{E}[R], \sigma(R, R))$ to re-initialize the random parameter $\theta^{(0)} = \{\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}\}$:

\[
\mu_k^{(0)} = \{\mu_k^{(0)}(T), \mathbb{E}[R]\}, \Sigma_k^{(0)} = \begin{bmatrix} \sigma_k^{(0)}(T, T) & \sigma_k^{(0)}(T, R) \\ \sigma_k^{(0)}(T, R) & \sigma_k^{(0)}(R, R) \end{bmatrix},
\]

where $\{\mu_k^{(0)}(T), \sigma_k^{(0)}(T, T), \sigma_k^{(0)}(T, R)\}$ are the random initialization of mean of $T$, the covariance of $T$, and the covariance of $T$ and $R$ in the group $Z = k$, respectively.

### 4.2 Expectation-Maximization

Based on the traditional Expectation-Maximization (EM) algorithm with group number $K$, we seek to find the Maximum Likelihood Estimate (MLE) of the marginal likelihood (Eq. (7)) by iteratively applying Expectation step and Maximization step.

**Expectation step.** In the expectation step of the $s$-th iteration, given the observation data $\{T, R\}$ and current parameter estimation $\theta^{(s)}$, we calculate the log expectation of likelihood function Eq. (7):

\[
Q(\theta, \theta^{(s)}) = \mathbb{E}[\log \mathcal{P}(T, R, Z \mid \theta) \mid T, R, \theta^{(s)}] = \sum_{k=1}^{K} \{\sum_{i=1}^{n}\hat{\gamma}_{ik}\log \pi_k + \sum_{i=1}^{n}\hat{\gamma}_{ik}\log \mathcal{P}(t_i, r_i \mid \mu_k, \Sigma_k)\}. \tag{9}
\]

The Expectation $\hat{\gamma}_{ik}$ is the conditional probability distribution that the $j$-th sample comes from the $k$-th sub-group given $\theta^{(s)}$:

\[
\hat{\gamma}_{jk} = \mathbb{E}[\gamma_{jk} \mid T, R, \theta^{(s)}] = \mathcal{P}(z_j = k \mid T, R, \theta^{(s)}) = \frac{\pi_k \mathcal{P}(T, R_{\theta^{(s)}})}{\sum_{i=1}^{K}\pi_i \mathcal{P}(T, R_{\theta^{(s)}})}. \tag{10}
\]

**Maximization step.** In the maximization step of the $s$-th iteration, given the observation data $\{T, R\}$ and the current parameter estimation $\theta^{(s)}$, we maximize the expectation of the log likelihood function $Q(\theta, \theta^{(s)})$ (Eq. (9)) to obtain the parameter estimation $\theta^{(s+1)}$ of next iteration:

\[
\theta^{(s+1)} = \text{argmax}_\theta Q(\theta, \theta^{(s)}). \tag{11}
\]

The solution is:

\[
\mu_k^{(s+1)} = \frac{\sum_{k=1}^{n}\hat{\gamma}_{jk}(T_{\oplus R_k})}{\sum_{j=1}^{n}\hat{\gamma}_{jk}}, k = \{1, \cdots, K\} \tag{12}
\]

\[
\Sigma_k^{(s+1)} = \frac{\sum_{k=1}^{n}\hat{\gamma}_{jk}(T_{\oplus R_k} - \mu_k^{(s+1)})^2}{\sum_{j=1}^{n}\hat{\gamma}_{jk}}, k = \{1, \cdots, K\} \tag{13}
\]

\[
\sigma_{ik}^{(s+1)} = \frac{\sum_{k=1}^{n}\hat{\gamma}_{jk}(T_{\oplus R_k})}{N}, k = \{1, \cdots, K\} \tag{14}
\]

where $T \oplus R$ concatenates $T$ and $R$. $M^2 = MM^T$, if $M$ is a matrix.

Then, the EM algorithm would obtain the convergent parameters $\theta^* = \{\pi^*, \mu^*, \Sigma^*\}$ by iteratively applying Expectation step and Maximization step. The convergence property of the EM algorithm is elaborated in Appendix B.

### 4.3 LatGIV reconstruction

Given the cluster number $K$, we use Expectation Maximization (EM) algorithm to get the convergent distribution parameters $\theta^* = \{\pi^*, \mu^*, \Sigma^*\}$ of complete-data $\{T, R, Z\}$. Then we identify the subgroup indicator $Z$ based on the distribution parameter differences $\{\mu_k^*(T), \sigma_k^*(T, T), \sigma_k^*(T, R)\}$, which corresponds to different treatment assignment mechanisms (Eq. 2):

\[
z_i = \text{argmax}_k \mathcal{P}(t_i, h(x_i) \mid \mu_k^*, \Sigma_k^*). \tag{15}
\]

In the EM algorithm, group number $K$ is a hyper-parameter. To implement an end-to-end algorithm, we use Maximum Mean Discrepancy (MMD) to measure the correlation between discrete variable $Z$
and the distribution of high-dimensional features $R$, and automatically select the most appropriate group number $K^*$ by the minimum correlation (Unconfounded Assumption 3.2):

\[
MMD_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} ||\bar{R}_{Z=i} - \bar{R}_{Z=j}||_2^2, \\
K^* = \arg\min_K MMD_K(Z, R), K = \{2, 3, 5, 10, \ldots\}.
\]

where $\bar{R}_{Z=i}/\bar{R}_{Z=j}$ denotes the mean of the representation $R$ in the $i$/j-sub-group according to the EM algorithm with group number $K$. Then, we estimate the heterogeneous $R$-$T$ distribution parameters, and use LatGIV to reconstruct the group indicator that satisfies IV’s three assumptions. More discussion about MMD and the hyper-parameter $K$ is given in Appendix $C$. The Pseudo-Code of LatGIV is shown in Algorithm $1$ & $2$ in Appendix $D$.

**Discussion:** The latent sub-group indicator is an exogenous variable that generates different treatment assignment mechanisms $f_{\theta_x}(X) + \epsilon_{TZ}$ but does not affect the distribution of the covariates $X$ and the unmeasured confounders $U$. In addition, the group indicator only affects the outcome through its strong association with the treatment. Thus, we can take it as a **strong valid IV** that satisfies Assumptions 3.2. We reconstruct the latent sub-group indicator as LatGIV $Z$ (the reconstruction accuracy has reached 77% in Section 5.2) by modeling the difference of potential distribution parameters $\{\mu_0(T), \sigma_0^2(T), \sigma_0^2(T, R)\}$. The greater the heterogeneity between parameter distributions of Covariate-Treatment, the stronger the IV effect and higher the reconstruction accuracy. Apart from that, the residual $\beta_z = \epsilon_{TZ} - \mathbb{E}[\epsilon_{TZ} | R_z]$ can be regarded as a random perturbation to the distribution of the data $T_{[k]} \sim \mathbb{P}(E[T_{[k]}], \sigma(T_{[k]}), T_{[k]})$. Therefore, smaller residual variance positively support better reconstruction performance. As noted, although $\beta_z$ affects the distribution of $T_{[k]}$, $\{\mathbb{E}[T_{[k]}], \sigma(T_{[k]}, T_{[k]})\}$ is fixed for each Group $Z = k$ and LatGIV still could be reconstructed.

### 4.4 Treatment-response function prediction

A reconstructed LatGIV can be used as an strong valid IV. In the presence of the unmeasured confounders, we plug it into the IV-based methods to estimate the causal effect with additive noise assumption $f_{\theta_y}(X) + \epsilon_{TX}$, i.e., treatment-response function $g(T, X)$. We use four lines of nine **IV-based methods** as listed: (1) **The two-Stage Least Squares**, Poly2SLS and NN2SLS: 2SLS with polynomial regression and Neural network; (2) **The Kernel-based Methods**, Kernel IV $[36]$ and DualIV $[29]$ map $X$ to a reproducing kernel Hilbert space (RKHS) and performing kernel ridge regression in two-stage; (3) **The Deep Methods**, DeepIV $[16]$, OneSIV $[25]$ and DFIV $[41]$ adopts deep neural nets and fit a mixture density network to estimate the conditional probability distribution of treatments $T$ in stage 1 and performs a joint mapping from resampled treatments $\hat{T}$ and confounders $X$ to the outcomes $Y$ in stage 2; (4) **The Adversarial GMM**, AGMM $[15]$ and DeepGMM $[6]$ construct a structural function and select moment conditions via adversarial training, based on the optimally weighted Generalized Method of Moments (GMM).

## 5 Experiments

### 5.1 Baselines

In this paper, we plug LatGIV into the different downstream IV-based methods for treatment-response function prediction. Then, we compare our algorithm LatGIV with the following baseline Summary IV methods: (1) **NoneIV** use a full-zeros vector as IV; (2) **UAS** $[14]$ takes the average of IV candidates as IV; (3) **WAS** $[9]$ weights each candidate based on the associations as IV; (4) **ModeIV** $[17]$ takes the tightest center of estimation points as IV; (5) **AutoIV** $[42]$ learn a disentangled representation as IV, but they all need an available IV candidates set. Besides, we adopt K-Means with MMD to find the most appropriate group number $K$ and take the cluster results as LatGIV$_{KM}$ (with K-Means algorithm), and take K-Means with known group number $K$ as LatGIV$_{KM}$. We adopt EM with MMD to take the cluster results as LatGIV$_{EM}$ (with EM algorithm). Finally, we also evaluate the performance of **TrueIV**, i.e., known group indicators, in the downstream regression tasks.

### 5.2 Experiments on synthetic datasets

**Datasets** In this paper, we focus on the mixed data under different treatment assignment mechanisms. For simulation, we first generate the confounders $\{X, \epsilon\}: X, \epsilon_Y \sim \mathcal{N}(0, \Sigma_{m_X+1}), \Sigma_{m_X+1} =$
The instruments $Z$ and treatments $T$ are generated by:

$$\begin{align*}
T &= \sum_{z=1}^{K} 1[z=z] \left[ \sum_{i=1}^{m} w_{zi}^{1} X_i + \sum_{i=1}^{m} w_{zi}^{2} f(X_i) + f(z) \right] + \delta_T, \quad \delta_T \sim \mathcal{N}(0, 0.1) (18) \\
Z &= \mathcal{P}(Z = z) = 1/K, w_{zi}^{1}, w_{zi}^{2} \sim \mathcal{U}(-1, 1), z = 1, \ldots, K, (19)
\end{align*}$$

where $f(x) = X_i$ and $f_2(x) = 0.2x + c_z, c_z \sim \mathcal{U}(-1, 1)$ (More discussion about the stability of LatGIV with various unmeasured confounding settings $f_2(x)$ is place in Appendix E). The mixed data derives from $K$ different sub-groups, meaning that there are $K$ independent potential treatment assignment models, i.e., $z \in \{1, \ldots, K\}$. $Z$ is the indicator of the potential causal model, which can be regarded as an instrumental variable. Besides, we design 5 different treatment functions $f_X(\cdot)$ to discuss the performance of LatGIV algorithm: (1) linear scenario $f_X(X_i) = X_i$; (2) poly scenario $f_X(X_i) = X_i^2$; (3) sin scenario $f_X(X_i) = \sin(X_i)$; (4) sigmoid scenario $f_X(X_i) = 1/(1 + \exp(-X_i))$; (5) abs scenario $f_X(X_i) = \text{abs}(X_i)$.

The response function $Y$:

$$Y = -1.5T + 0.9T^2 + \sum_{i=1}^{m} \frac{X_i}{m} + |X_1X_2| - \sin(10 + X_2X_3) + 2\epsilon + \delta_Y$$

where $\epsilon$ is an unmeasured confounder and $\delta_Y \sim \mathcal{N}(0, 0.1)$.

For synthetic datasets, we sample 3,000 units and perform 10 replications to report the mean squared error (MSE) and its standard deviations of the treatment-response function estimation, and the MSE is calculated over the testing data (3000 units) that we intervene the treatment as $T = \text{do}(t)$. To verify the effectiveness of LatGIV$_{EM}$ in different scenarios with different $m_X$-dimension and different group numbers $K$, we use DataK-mX to denote the different dimensions of covariates and group number and perform 10 replications.

### Table 1: The Mean Squared Error $\text{mean}(std)$ of linear Experiments (Linear-3-3)

|               | Poly2SLS | NNS2SLS | KernelIV | DualIV$^{(1)}$ | DeepIV | OneSIV | DFIV$^{(1)}$ | DeepGMM | AutoIV |
|---------------|----------|---------|-----------|---------------|--------|--------|-------------|---------|--------|
| NoneIV        | 0.330(0.073) | 1.905(0.155) | 0.351(0.075) | 1.924(0.479) | 0.376(0.011) | 0.316(0.028) | 1.350(0.130) | 0.334(0.068) | 0.214(0.046) |
| UAS           | 0.330(0.073) | 2.289(1.461) | 0.350(0.074) | 0.974(0.340) | 0.372(0.024) | 0.340(0.030) | 1.299(0.090) | 0.324(0.037) | 0.210(0.047) |
| WAG           | 0.333(0.073) | 1.595(0.241) | 0.360(0.054) | 2.164(0.439) | 0.369(0.023) | 0.370(0.027) | 1.290(0.123) | 0.321(0.086) | 0.234(0.041) |
| ModelIV       | 0.330(0.073) | 2.247(1.300) | 0.355(0.076) | 1.899(0.509) | 0.367(0.021) | 0.312(0.025) | 1.392(0.120) | 0.307(0.070) | 0.199(0.041) |
| AutoIV        | $>100^{(2)}$ | 2.096(1.088) | 0.352(0.075) | 0.798(0.324) | 0.367(0.021) | 0.310(0.031) | 1.280(0.113) | 0.306(0.090) | 0.214(0.046) |

$^{(1)}$ DualIV and DFIV don't perform well on LatGIV and even on TrueIV because they require continuous IVs rather than discrete IVs. $^{(2)}$ $>100$ means "MSE<100".

**The results of response function estimation**

As shown in Table 1 (The top-2 performance is highlighted in bold for all tables), we compare the performance of LatGIV with other Summary IVs and TrueIV on various downstream IV methods in linear scenario (Linear-3-3) with $T = \text{do}(t)$. Following observations are identified from the results: (1) Without valid IV candidates, Summary IVs are not reliable and fail to synthesize a valid IV, and plugging them into the IV methods can hardly improve the estimation performance, which is close to the NoneIV; (2) Through clustering, we reconstruct the latent exogenous IV that generate different treatment mechanisms, LatGIVs (with K-means or EM) bring higher accuracy on response function estimation by comparing with Summary IV methods in various IV-based methods except DualIV; (3) By estimating the latent differentiated covariate-treatment distribution parameters across groups and reconstructing the latent IV, LatGIV$_{EM}$ significantly improves the performance of clustering methods with LatGIV$_{K|M}$ and achieves SOTA performance to predict treatment-response function, even comparable with TrueIV. (4) DualIV and DFIV do not perform well on LatGIV and fail to predict the treatment-response function, even with TrueIV, because they require continuous IVs rather than discrete IVs.

Then, to verify the effectiveness of LatGIV in nonlinearity cases, we design 5 different treatment functions $f_X(\cdot)$ to evaluate the response function estimation performance of LatGIV algorithm. Due to limited space, we select the SOTA IV-based methods (Poly2SLS, KernelIV, DeepIV, AGMM) in four lines to evaluate LatGIV. We plot the estimated value of response function with $T=\text{do}(0)$ and sort it by Ground-Truth (GT) for different synthetic scenarios. The results (Fig. 3) show LatGIVs...
Figure 2: Reconstruction Accuracy of the Group IV with Different Group Number.

Table 2: The Mean Squared Error \( \text{mean}(\text{std}) \) of IHDP & PM-CMR Dataset

| IHDP Dataset | Poly2SLS | KernelIV | DeepIV | AGMM | Poly2SLS | KernelIV | DeepIV | AGMM |
|--------------|----------|----------|--------|------|----------|----------|--------|------|
| None         | 0.238(0.132) | 0.456(0.243) | 0.583(0.240) | 0.140(0.063) | 0.181(0.044) | 0.352(0.198) | 0.409(0.160) | 0.130(0.064) |
| UAS          | 0.250(0.133) | 0.470(0.244) | 0.576(0.245) | 0.342(0.061) | 0.181(0.044) | 0.352(0.198) | 0.409(0.160) | 0.130(0.064) |
| ModIV        | 0.248(0.133) | 0.450(0.241) | 0.567(0.226) | 0.144(0.059) | 0.181(0.044) | 0.372(0.207) | 0.417(0.164) | 0.137(0.080) |
| AußIV        | >100      | 0.457(0.241) | 0.583(0.250) | 0.142(0.060) | 0.179(0.044) | 0.351(0.198) | 0.409(0.180) | 0.129(0.064) |

| IHDP Dataset | Poly2SLS | KernelIV | DeepIV | AGMM | Poly2SLS | KernelIV | DeepIV | AGMM |
|--------------|----------|----------|--------|------|----------|----------|--------|------|
| LatGIV\( EM \) | 0.078(0.029) | 0.354(0.179) | 0.505(0.203) | 0.112(0.050) | 0.088(0.044) | 0.320(0.202) | 0.381(0.165) | 0.117(0.053) |
| LatGIV\( EM \) | 0.034(0.011) | 0.202(0.173) | 0.482(0.228) | 0.095(0.035) | 0.048(0.012) | 0.308(0.210) | 0.339(0.184) | 0.085(0.045) |
| TrueIV       | 0.033(0.009) | 0.151(0.066) | 0.458(0.166) | 0.095(0.035) | 0.028(0.007) | 0.140(0.074) | 0.141(0.054) | 0.054(0.023) |

(with K-means or EM) achieve SOTA performance, especially LatGIV\( EM \) achieves comparable results with TrueIV and the estimated response curves from LatGIV\( EM \) and TrueIV approximates the Ground-Truth curve.

The results of reconstruction accuracy of the Group IV As shown in Fig. 2, LatGIV\( EM \) algorithm successfully reconstructs the latent group IV, and the average reconstruction accuracy has reached 77% under various group numbers, especially exceeding 90% accuracy on Two Groups setting. In contrast, the identification accuracy of LatGIV\( KM \) is basically below 60%. LatGIV gains from the reconstruction of latent IV, LatGIV\( EM \) with higher reconstruction accuracy achieves better performance to predict treatment-response function compared with other Summary IVs, even comparable with TrueIV. Besides, using correlation measured by MMD, LatGIV\( EM \) always find the proper group number (red-line) automatically in different synthetic settings (Linear-\( K \)-m\( X \)), as shown in Fig. 4. However, LatGIV\( KM \) with MMD fails to choose the group number as hype-parameter, and the correlation between LatGIV\( KM \) and covariates \( X \) and group number \( K \) have a positive correlation.

We increase the critical level of simulation to test the stability of our LatGIV in Data-\( K \)-m\( X \) with different group number \( K \) and dimensions of covariates \( m_X \) as shown in Table 3 in Appendix. Besides, we perform experiments with various unmeasured confounding \( f_x(\epsilon) \) (Eq. 44), as shown in Table in Appendix. We elaborate the results in Appendix E.2. LatGIV\( EM \) still shows in stability and is in Top2, i.e., LatGIV\( EM \) is second only to TrueIV.

5.3 Experiments on real-world datasets

Real-world datasets Similar to previous methods [7, 16, 31, 34], we perform experiments on two semi-synthetic real-world datasets IHDP [35] & PM-CMR [39], as the true response function is rarely available for real-world data. Then we use the continuous variables from IHDP & PM-CMR
to replace the covariates $X$ in Eq. (18) & (20) to generate treatment $T$ and outcome $Y$, respectively. Both two datasets are randomly split into training (63%), validation (27%) and testing (10%). We perform 10 replications to report the mean squared error (MSE) and its standard deviations (std) of the treatment-response function estimation. Due to limited space, we will elaborate the real-word data in Appendix F.1.

The results of response function estimation Finally, we assess all Summary IVs’ performance in treatment-response function estimation with the covariates from the real-world data IHDP & PM-CMR. We perform 10 replications and report the mean and standard deviations of MSE in the treatment-response function estimation. In Table 2, LatGIV EM shows consistent and robust performance, always maintaining the performance of top-2 and almost achieving the same effect as TrueIV on IHDP & PM-CMR Datasets. This means that LatGIV EM can also reconstruct the latent group IV with the data distribution from the real scenarios. The full results of response function estimation of LatGIV in all IV-based methods are shown in Table 5 & 6 in Appendix F.

6 Conclusion

By estimating the differentiated covariate-treatment distribution across groups (each group with an independent treatment assignment mechanism), we propose a novel LatGIV EM algorithm for reconstructing latent Group IVs and predicting treatment-response function from data fusion with unmeasured confounders. The average reconstruction accuracy has reached 77%. Empirical results demonstrate the advantages of the LatGIV algorithm compared with Summary IV methods. One limitation is that LatGIV requires the treatment-response function is stable in mixed data.

References

[1] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.
[2] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[3] S. Athey and G. W. Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

[4] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[5] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, pages 7345–7352, 2016.

[6] A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. In *NeurIPS 2019*, 2019.

[7] I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*, 2020.

[8] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6):S114, 2010.

[9] S. Burgess, F. Dudbridge, and S. G. Thompson. Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906, 2016.

[10] S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[11] S. Burgess and S. G. Thompson. Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, 42(4):1134–1144, 2013.

[12] X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

[13] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[14] N. Davies, S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. D. Smith. The many weak instruments problem and mendelian randomization. *Statistics in medicine*, 34(3):454–468, 2015.

[15] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. In *NeurIPS 2020*, 2020.

[16] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. In *ICML 2017*, 2017.

[17] J. S. Hartford, V. Veitch, D. Sridhar, and K. Leyton-Brown. Valid causal inference with (some) invalid instruments. In *International Conference on Machine Learning*, pages 4096–4106. PMLR, 2021.

[18] M. A. Hernan and J. M. Robins. Causal inference, 2010.

[19] B. Huang and K. Zhang. Specific and shared causal relation modeling and mechanism-based clustering. In *NeurIPS 2019*, 2019.

[20] G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, pages 467–475, 1994.

[21] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[22] D. M. Kent, P. M. Rothwell, J. P. Ioannidis, D. G. Altman, and R. A. Hayward. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 11(1):1–11, 2010.

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Z. Kuang, F. Sala, N. Sohoni, S. Wu, A. Córdova-Palomera, J. Dunnmon, J. Priest, and C. Ré. Ivy: Instrumental variable synthesis for causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 398–410. PMLR, 2020.

[25] A. Lin, J. Lu, J. Xuan, F. Zhu, and G. Zhang. One-stage deep instrumental variable method for causal inference from observational data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 419–428. IEEE, 2019.

[26] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *NeurIPS 2017*, 2017.

[27] J. Y. Millwood, R. G. Walters, X. W. Mei, Y. Guo, L. Yang, Z. Bian, D. A. Bennett, Y. Chen, C. Dong, R. Hu, et al. Conventional and genetic evidence on alcohol and vascular disease aetiology. *The Lancet*, 393(10183):1831–1842, 2019.
[28] L. E. Mokry, O. Ahmad, V. Forgetta, G. Thanassoulis, and J. B. Richards. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of medical genetics*, 52(2):71–79, 2015.

[29] K. Muandet, A. Mehrjou, S. Le Kai, and A. Raj. Dual instrumental variable regression. In *NeurIPS 2020*, 2020.

[30] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[31] L. Nie, M. Ye, D. Nicolae, et al. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2020.

[32] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[33] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[34] P. Schwab, L. Linhardt, S. Bauer, J. M. Buhmann, and W. Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *The AAAI Conference on Artificial Intelligence*, 2020.

[35] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML 2017*, 2017.

[36] R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *NeurIPS 2019*, pages 4593–4605, 2019.

[37] N. Sokolovska and P.-H. Wuillemin. The role of instrumental variables in causal inference based on independence of cause and mechanism. *Entropy*, 23(8):928, 2021.

[38] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[39] L. H. Wyatt, G. C. L. Peterson, T. J. Wade, L. M. Neas, and A. G. Rappold. Annual pm2. 5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in brief*, 30:105–318, 2020.

[40] Y. Xie. Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences*, 110(16):6262–6268, 2013.

[41] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

[42] J. Yuan, A. Wu, K. Kuang, B. Li, R. Wu, F. Wu, and L. Lin. Auto iv: Counterfactual prediction via automatic instrumental variable decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–20, 2022.

[43] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI 2017*, 2017.

[44] S. Zhang, D. Yao, Z. Zhao, T.-S. Chua, and F. Wu. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *ACM SIGIR on Research and Development in Information Retrieval*, 2021.
A Proof of propositions

In machine learning, representation is a set of feature learning techniques that automatically map the high-dimensional redundant covariates to low dimensional manifold and discover the representations needed for feature detection or classification from raw dat, such as supervised dictionary learning, neural networks embedding, autoencoder, principal component analysis, and independent component analysis. In causal inference, representation is receiving increasing attention from economists, statisticians, epidemiologists and social scientists, such as CFR [35], CEVAE [26] and DFIV [41]. After reviewing these works, we combine the advantages of previous works and propose a representation proposition for Expectation-Maximization algorithm.

Proposition. Representation. There is a family $\mathcal{H}$ of function $h$ that encode the covariates $X$ to low-dimension representation $R = h(X)$, in which one encoding variable or a group of encoding variables combines the effect of corresponding covariate $X_j$ on the treatment variable:

$$T = f_{\theta_Z}(X) + \epsilon_{TZ} = \sum_{z=1}^{K} 1_{[Z = z]} \sum_{j=1}^{m_R} (\alpha_{zj} R_j + \beta_z), R \in \mathbb{R}^{m_R},$$

(21)

where the residual $\beta_z = \epsilon_{TZ} - \mathbb{E}[\epsilon_{TZ} \mid R, Z]$ is independent with $R$ in group $Z = \{1, \ldots, K\}$. Besides, $m_R$ denotes the dimension of representation variable $R$, and $1_{[Z = z]}$ equals 1 if $Z = z$, 0 otherwise; $\alpha_{zj}$ fit a linear (Non)-Gaussian Mixed Model, $\beta_z$ is the corresponding residual. In linear scenario, $h(X) = X$ is an identity function.

Proof. In linear scenario, the treatment variable $T$ is a linear combination of high-dimensional covariates. Without loss of generality, we denote the corresponding coefficients $\alpha_{zj}$ for covariate $X_j$ and denote the intercept $\beta_z$ in source/group $z \in Z$. The ground-truth treatment assignment mechanism is:

$$T = f_{\theta_Z}(X) + \epsilon_{TZ} = \sum_{z=1}^{K} 1_{[Z = z]} \sum_{j=1}^{m_X} [\alpha_{zj} X_j + \beta_z],$$

(22)

where $\beta_z$ is a constant, independent with $R$. Then, we can use identity function to set $R = h(X) = X$:

$$T = f_{\theta_Z}(X) + \epsilon_{TZ} = \sum_{z=1}^{K} 1_{[Z = z]} \sum_{j=1}^{m_R} (\alpha_{zj} R_j + \beta_z),$$

(23)

where $R_j = X_j$ and $m_R = m_X$.

In non-linear scenario, we can learn and select a representation $R = h(X)$ through Principal Component Analysis (PCA), Variational Auto-Encoder (VAE), Correlation Minimization or prior knowledge selection [23][26].

Next, we propose a Variational Auto-Encoder Variant to elaborate the proof in non-linear scenario.

In each source/group $Z$ of data fusion, there exists a set of latent variables $R \in \mathbb{R}^{m_R}$ that generates the observed variables $X, T$ by the decoder $p_h(X, T \mid R)$, assuming that all observed data is generated by a set of high-level cause variables $R$ in real-world. That means that we can figure out the observed variables iff we know the decoder $p_h(X, T \mid R)$ and high-level latent variable $R$. Thus, we use a tractable distribution $q_h(R \mid X, T)$ to approximate the true posterior $p_h(R \mid X, T)$, the distribution

$^{2}\mathbb{E}[\epsilon_{TZ} \mid R, Z = z]$ is constant for the specified $R$ given $Z$, and $\beta_z$ will not affect $\alpha_{zj}$.
distance can be measured by:
\[ KL(q_\phi(R \mid X, T) \Vert p_\theta(R \mid X, T)) \]
\[ = \int q_\phi(R \mid X, T) \ln \frac{q_\phi(R \mid X, T)}{p_\theta(R \mid X, T)} dR \]
\[ = E_{R \sim q_\phi(R \mid X, T)} \left[ \ln \frac{q_\phi(R \mid X, T)}{p_\theta(R \mid X, T)} \right] \]
\[ = E_{R \sim q_\phi(R \mid X, T)} \left[ \ln q_\phi(R \mid X, T) - \ln p_\theta(R \mid X, T) \right] \]
\[ = E_{R \sim q_\phi(R \mid X, T)} \left[ \ln q_\phi(R \mid X, T) - \ln \left( \frac{p(X, T \mid R)p(R)}{p(X, T)} \right) \right] \]
\[ = KL(q_\phi(R \mid X, T) \Vert p(R)) - E_{R \sim q_\phi(R \mid X, T)} \left[ \ln p(X, T \mid R) \right] + \ln p(X, T). \quad (24) \]

That is:
\[ \ln p(X, T) \geq ELBO = E_{R \sim q_\phi(R \mid X, T)} \left[ \ln p(X, T \mid R) \right] - KL(q_\phi(R \mid X, T) \Vert p(R)), \quad (26) \]

where \( KL(q_\phi(R \mid X, T) \Vert p(R \mid X, T)) \geq 0 \), which equals to 0 iff \( q_\phi(R \mid X, T) = p_\theta(R \mid X, T) \).

Then, we can obtain the following evidence lower bound (ELBO):
\[ \ln p(X, T) \geq ELBO = E_{R \sim q_\phi(R \mid X, T)} \left[ \ln p(X, T \mid R) \right] - KL(q_\phi(R \mid X, T) \Vert p(R)), \quad (26) \]

where we can model \( p_\phi(X, T \mid R) \) as a normal distribution \( \mathcal{N}(\mu(R), \Lambda(R)) \), the density function is
\[ \frac{1}{\sqrt{|2\pi|^{m} \cdot |2\lambda|}} \exp\left(\frac{1}{2} (X + T - \mu(R))^T \Lambda(R)^{-1} (X + T - \mu(R)) \right) \]

and \( \Lambda(R) \). We fixed \( \Lambda(R) \) as identity matrix \( I \) in Eq. (26):
\[ E_R [\ln p_\phi(X, T | R)] \]
\[ = E_R \left[ \ln \left( \frac{1}{(2\pi)^{m} \cdot |2\lambda|} \right) + \left( -\frac{1}{2} (X + T - \mu(R))^T \Lambda(R)^{-1} (X + T - \mu(R)) \right) \right] \]
\[ = E_R \left[ \ln \left( (2\pi)^{-(m+1)/2} \right) - \frac{1}{2} \ln |(X + T - \mu(R))|^2 \right]. \quad (27) \]

where \( X + T \) concatenates \( X \) and \( T \). VAE achieves \( q_\phi(R \mid X, T) = p_\theta(R \mid X, T) \) by maximizing ELBO. To learn the latent representation \( R \) for \( T = f_{\theta_Z}(X) + \varepsilon_{TZ} \), we can re-formulate the minimization loss for neural network as:
\[ L = \| (X + T - \mu(R)) \|^2 + KL(q_\phi(R \mid X, T) \Vert p(R)) \]
\[ = \| (X - \mu(R)) \|^2 + (T - \sum_{j=1}^{m_R} [\alpha_{zj} R_j + \beta_{zj}])^2 + KL(q_\phi(R \mid X, T) \Vert p(R)). \quad (28) \]

where, we set \( p(R) \) as standard multivariate normal distribution \( \mathcal{N}(0, I) \). Then we can learn an independent representation \( R \) by minimizing the objective \( L \) (Eq. (28)).

Besides, another Correlation Minimization Algorithm also can be used to elaborate the proof in non-linear scenario. In this paper, we propose to use mutual information to measure the correlation \( Cov(X_A, X_B) \) between two random variables \( X_A \) and \( X_B \): firstly, we use variational distribution \( q_\psi(X_A \mid X_B) = \mathcal{N}(\mu_\psi(X_B), \sigma_\psi(X_B)) \) parameterized by neural networks \( \{\mu_\psi, \sigma_\psi\} \) to approximate the true conditional distribution \( P(X_A \mid X_B) \); then, we minimize the log-likelihood loss function of variational approximation \( q_\psi(X_A \mid X_B) \) with \( n \) samples to estimate MI, i.e., \( Cov_{CLUB}(x_i, x_j) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log q_\phi(X_A, i \mid X_B, i) - \log q_\phi(X_A, i \mid X_B, i) \).

Then, we use the correlation minimization knowledge (\( Cov(x_i, x_j) = 0 \)) to select independent covariates \( X \) and learn latent representation \( R = h(X) \) by polynomial regression, and set the rest dependent covariates as a part of unmeasured noise \( \varepsilon_Y \). We formulate a linear (non-)gaussian mixed model:
\[ T = f_{\theta_Z}(X) + \varepsilon_{TZ} = \sum_{z=1}^{K} 1[z=x] \left\{ \sum_{j=1}^{m_R} [\alpha_{zj} h_j(X)] + \beta_{zj} \right\} \]
\[ = \sum_{z=1}^{K} 1[z=x] \left\{ \sum_{j=1}^{m_R} [\alpha_{zj} \xi_{zj} x^2 + \xi_{zj} x^2 + \cdots] + \beta_{zj} \right\} \]
\[ = \sum_{z=1}^{K} 1[z=x] \left\{ \sum_{j=1}^{m_R} [\alpha_{zj} R_j + \beta_{zj}] \right\}. \quad (30) \]
where \( \xi_{zj,d} \) denotes the corresponding expectation coefficient of the \( d \)-th power of variable \( x_j \). Besides, all information irrelevant to \( R \) is summed up in one residual term \( \beta_z \).

In a conclusion, to simplify the reconstruct problem, we can learn and select an independent representation \( R = h(X) \) through PCA, VAE, correlation minimization or prior knowledge selection \([23, 26]\). In this paper, we use Maximum Mean Discrepancy (MMD) to measure the correlation between LatGIV \( Z \) and the covariates \( R \), and automatically select the most appropriate group number \( K^* \) by the minimum correlation:

\[
MMD_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} ||R_{Z=i} - R_{Z=j}||_2^2, \\
K^* = \arg\min_K MMD_K(Z, R), \ K = \{2, 3, 5, 10, \ldots \}.
\]

According to the convergence property of the EM algorithm \([3] \), we can identify the association differences across groups, and get the convergent distribution parameters \( \theta^* = \{\pi^*, \mu^*, \Sigma^*\} \) of complete-data \( \{T, R, Z\} \). Then we reconstruct the Latent Group IV \( Z \).

Similar to \([16,30]\), we establish an inverse problem proposition as follows:

**Proposition. Differentiated Representation-Treatment Association across Groups.** The distribution \( P(R \mid \mathbb{E}[R], \text{Cov}[R]) \) is same and fixed in all groups \( Z = \{1, \cdots, K\} \) (Unconfounded Assumption \([3,2]\)), but the Representation-Treatment Association \( T_z = \alpha_{zj}R_j + \beta_z \) vary across Groups \( Z \). By identifying the association differences across groups, we can reconstruct the indicator \( Z \).

**Proof.** Based on the unconfounded condition and the initialization distribution \( P(\mathbb{E}[R], \sigma(R, R)) \), we use Maximum Mean Discrepancy (MMD) to measure the correlation between LatGIV \( Z \) and the covariates \( R \), and automatically select the most appropriate group number \( K^* \) by the minimum correlation:

\[
MMD_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} ||R_{Z=i} - R_{Z=j}||_2^2, \\
K^* = \arg\min_K MMD_K(Z, R), \ K = \{2, 3, 5, 10, \ldots \}.
\]

In the non-linear scenario, the relationship between the outcome process and reduced form belongs a 1st Fredholm integral equation and leads an ill-posed inverse problem \([30]\). Considering the identification of a general outcome model in Eq. (1):

\[
Y = g(X, T) + \epsilon_Y; \mathbb{E}[\epsilon_Y \mid Z] = \mathbb{E}[\epsilon_Y] = 0.
\]

where \( g(\cdot) \) denotes a true, unknown structural function of interest. For a consistency estimation, \([16,30]\) identified the causal effect as the solution of an integral equation:

\[
\mathbb{E}[Y \mid Z, X] = \mathbb{E}[g(X, T) \mid Z, X] + \mathbb{E}[\epsilon_Y \mid X] = \int [g(T, X) + C] dF(T \mid Z, X),
\]

where \( F \) denotes the conditional cumulative distribution function of \( T \) given \( \{Z, X\} \), and \( \hat{g}(T, X) = \int [g(T, X) + C] dF(T \mid Z, X) \) is the solution of the inverse problem. Therefore, we characterize the identification of structural functions as completeness of certain conditional distributions \( \mathbb{E}[\epsilon_Y \mid Z] = 0 \).

In the parametric/nonparametric model (Eq. (33)), the identification/uniquness of \( \hat{g}(T, X) \) is equivalent to the nonexistence of any function \( \delta(X, T) := \hat{g}(T, X) - g(T, X) \neq 0 \) such that \( \mathbb{E}[\delta(X, T) \mid Z] = 0 \). Plugging the LatGIV into IV-based methods, we can predict TRF under assumption \([3,3]\) and \( C = \mathbb{E}[Y - \hat{g}(T, X)] \).
B The convergence analysis of the EM algorithm

Theorem. Through representation, we model the likelihood function for observational data \( \{t_i, r_i\}_{i=1}^{n} \) as \( P(T, R \mid \theta) \). We denote parameter sequence estimated by EM algorithm as \( \theta^{(i)}(i = 1, 2, \cdots) \), and denote the corresponding likelihood function sequence as \( P(T, R \mid \theta^{(i)})(i = 1, 2, \cdots) \). Then \( P(T, R \mid \theta^{(i)}) \) is a monotonic sequence which constantly increase:

\[
P(T, R \mid \theta^{(i+1)}) \geq P(T, R \mid \theta^{(i)}). \tag{35}
\]

Proof. The observation:

\[
P(T, R \mid \theta) = \frac{P(T, R, Z \mid \theta)}{P(Z \mid T, R, \theta)}, \tag{36}
\]

\[
\ln P(T, R \mid \theta) = \ln P(T, R, Z \mid \theta) - \ln P(Z \mid T, R, \theta). \tag{37}
\]

The expectation of the log likelihood function \( Q(\theta, \theta^{(i)}) \) (Eq. (41)):

\[
Q(\theta, \theta^{(i)}) = \Sigma Z \ln P(T, R, Z \mid \theta)P(Z \mid T, R, \theta^{(i)}), \tag{38}
\]

Let

\[
H(\theta, \theta^{(i)}) = \Sigma Z \ln P(Z \mid T, R, \theta)P(Z \mid T, R, \theta^{(i)}), \tag{39}
\]

Then,

\[
\ln P(T, R \mid \theta^{(i+1)}) - \ln P(T, R \mid \theta^{(i)}) = [Q(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i+1)}, \theta^{(i)})] - [Q(\theta^{(i)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})] = [Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})] - [H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})] \tag{40}
\]

where, the term \([Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})] \geq 0 \), because we maximize the expectation of the log likelihood function to obtain the parameter estimation \( \theta^{(i+1)} \) of next iteration:

\[
\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)}). \tag{41}
\]

The key component of EM algorithm is the use of Jensen’s inequality:

\[
H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) \leq \ln \left( \Sigma Z \frac{P(Z \mid T, R, \theta^{(i+1)})}{P(Z \mid T, R, \theta^{(i)})} \right) \leq \ln \left( \Sigma Z \frac{P(Z \mid T, R, \theta^{(i+1)})}{P(Z \mid T, R, \theta^{(i)})} \right) \tag{42}
\]

In conclusion, we obtain \( \ln P(T, R \mid \theta^{(i+1)}) - \ln P(T, R \mid \theta^{(i)}) \geq 0 \), that means \( P(T, R \mid \theta^{(i)}) \) is a monotonic sequence which constantly increase:

\[
P(T, R \mid \theta^{(i+1)}) \geq P(T, R \mid \theta^{(i)}). \tag{43}
\]

If there is an upper bound for \( P(T, R \mid \theta) \), the sequence \( \ln P(T, R \mid \theta^{(i)})(i = 1, 2, \cdots) \) would converge to a specific value \( L^* \).

Theoretically, EM algorithm is effective when there are identifiable differences in treatment assignment mechanisms across groups. Overall, the reconstruction accuracy has reached 77% in Section 5.2 and we can estimate the treatment-response function accurately. However, one limitation is that EM typically converges to a local optimum, not necessarily the global optimum, with no bound on the convergence rate in general.
C The discussion for MMD and hyperparameter

In real-world scenarios, the observational data collected from different sources or different times are often with different treatment assignment mechanisms. To handle the bias induced by data fusion — multiple datasets collected under multiple different treatment assignment mechanisms, we propose to separate the observational data into multiple groups (each group with an independent treatment assignment mechanism), and then explicitly model the group indicator as a Latent Group Instrumental Variable (LatGIV) to implement IV-based Regression. Nevertheless, due to irregularities in collecting data and privacy issues, the time or source labels are sometimes missing in available datasets. To reconstruct the latent group/source indicator, we propose to separate the observational data into $K$ groups using EM algorithm with representation.

In the EM algorithm, group number $K$ is a hyper-parameter. To implement an end-to-end algorithm, we use Maximum Mean Discrepancy (MMD) to measure the correlation between discrete variable $Z$ and the distribution of high-dimensional features $R$, and automatically select the most appropriate group number $K^*$ by the minimum correlation (Unconfounded Assumption 3.2):

$$MMD_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \| \bar{R}_{Z=i} - \bar{R}_{Z=j} \|^2_2,$$

$$K^* = \text{argmin}_K MMD_K(Z, R), K = \{2, 3, 5, 10, \ldots \}.$$

where $\bar{R}_{Z=i}/\bar{R}_{Z=j}$ denotes the mean of the representation $R$ in the $i/j$-th sub-group according to the EM algorithm with group number $K$.

The hyperparameter $K$ In this paper, the identification of LatGIV is from the multiple treatment assignment mechanisms across groups (each group with an independent treatment assignment mechanism). LatGIV is effective when there are identifiable differences in treatment assignment mechanisms across groups. That means that there exists a latent IV indicating multiple treatment assignment mechanisms hidden in data fusion, but the covariates’ distribution is same and fixed in all groups $Z = k$, i.e., IV’s unconfounded condition $Z \perp X, U$. Based on the unconfounded condition and representation, we use covariates’ distribution $P(\mathbb{E}[R], \sigma(R, R))$ from all samples to re-initialize the random parameter $\theta^{(0)} = \{\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}\}$ for EM algorithm (Eq. 8):

$$\mu_k^{(0)} = \{\mu_k^{(0)}(T), \mathbb{E}[R]\}, \Sigma_k^{(0)} = \begin{bmatrix} \sigma_k^{0}(T, T) & \sigma_k^{0}(T, R)^T \\ \sigma_k^{0}(T, R) & \sigma_k(R, R) \end{bmatrix}.$$

To hold the unconfounded condition and the initialization distribution $P(\mathbb{E}[R], \sigma(R, R))$, we need enough samples in each group - at least 40 samples, the more the better - to empirically estimated covariates distributions. When the number of samples in some groups are too small or two treatment assignment mechanisms are very similar, we allow different groups to be merged into one. Therefore, we recommend the candidate set $\{2, 3, 5, 10, \ldots , n/40\}$ for hyperparameter $K$ to hold the unconfounded condition, where $n$ is the sample size. In addition, we also allow other more candidate set options.

However, due to the lack of prior knowledge of the data source and no supervised information about the counterfactual evaluation, in observational data, we have to establish a metric to help us choose hyperparameter $K$.

The choice for MMD Based on the unconfounded condition and the initialization distribution $P(\mathbb{E}[R], \sigma(R, R))$, one natural idea is to measure the correlation between LatGIV $Z$ and the covariates $R$, and automatically select the most appropriate group number $K^*$ by the minimum correlation. Inspired by discrete instrumental variables being a group partitioning indicator, therefore, we adopt a simple but effective method to figure out the mean of the representation $R$ in the $i$-th sub-group, denoted by $\bar{R}_{Z=i}(i = 1, 2, \ldots , K)$. Then we measure the maximum mean of the representation discrepancy $\bar{R}_Z$ across groups $Z = 1, 2, \ldots , K$ to measure the correlation between LatGIV $Z$ and the covariates $R$, and automatically select the most appropriate group number $K^*$ by the minimum correlation:

$$MMD_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \| \bar{R}_{Z=i} - \bar{R}_{Z=j} \|^2_2,$$

$$K^* = \text{argmin}_K MMD_K(Z, R), K = \{2, 3, 5, 10, \ldots , n/40\}.$$
D Pseudo-Code and configuration

In this paper, we aim to identify the association differences \( f_{\theta_Z}(X) + \epsilon_TZ \) across groups and reconstruct the group indicator as LatGIV \( Z \). To simplify the reconstruct problem, we can learn and select an independent representation \( R = h(X) \) through PCA, VAE, correlation minimization or prior knowledge selection \([23, 26]\). In this paper, we use prior knowledge \( \text{Cov}(x_i, x_j) = 0 \) to select and learn latent representation \( R = h(X) \) by polynomial regression, and set the rest dependent covariates as a part of unmeasured noise \( \epsilon_Y \). Then we combine the EM algorithm and MMD measure to reconstruct latent LatGIVs from data fusion, and then plug LatGIVs into IV Methods to predict the response function. The Pseudo-Code of LatGIV\( ^3 \) is shown in Algorithm 1 & 2.

Algorithm 1 LatGIV(\( K \)): Recover Instrumental Variables by Expectation-Maximization algorithm with Clustering Number \( K \)

**Input:** Observational data \( \{x_i, t_i, y_i\}_{i=1}^n \); Clustering Number \( K \); A family of function \( h \in \mathcal{H} \); Init Paramters \( \theta^{(0)} = \{\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}\} \); Threshold \( \tau \).

**Output:** Latent IVs \( Z(K) \).

1: **Representation:** \( R = h(X) \).
2: \( \mathbb{E}[R] = \frac{1}{n}\sum_{i=1}^n r_i \), \( \mathbb{E}[R] = \frac{1}{n}\sum_{i=1}^n r_i \).
3: \( \text{Cov}[R] = \mathbb{E}[(R - \mathbb{E}[R])(R - \mathbb{E}[R])^T] \).
4: \( \mu^{(1)} = \{\mu^{(0)}(T), \mathbb{E}[R]\} \), \( \sigma^{(1)} = \left[ \begin{array}{cc} \sigma^2_{(T, T)} & \sigma^2_{(T, R)} \\ \sigma^2_{(R, T)} & \text{Cov}[R] \end{array} \right] \).
5: \( \theta^{(1)} = \{\pi^{(0)}, \mu^{(1)}, \Sigma^{(1)}\} \), \( s = 1 \).
6: while \( ||\theta^{(s)} - \theta^{(s-1)}||^2 < \tau \) do
7: \( \hat{\gamma}_{jk} = \frac{\mathbb{E}[P(T|R,\mu^{(s-1)}_k)]}{\mathbb{E}[P(T|R,\mu^{(s-1)}_k)]} \), \( j = \{1, \ldots, n\} \).
8: \( \mu^{(s+1)}_k = \frac{\sum_{j=1}^n \hat{\gamma}_{jk} \mu^{(s)}_j}{\sum_{j=1}^n \hat{\gamma}_{jk}} \), \( k = \{1, \ldots, K\} \).
9: \( \Sigma^{(s+1)} = \frac{\sum_{j=1}^n \hat{\gamma}_{jk} (\mu^{(s+1)}_j - \mu^{(s+1)}_k)^2}{\sum_{j=1}^n \hat{\gamma}_{jk}} \), \( k = \{1, \ldots, K\} \).
10: \( \pi^{(s+1)} = \frac{\sum_{j=1}^n \hat{\gamma}_{jk}}{N} \), \( k = \{1, \ldots, K\} \).
11: \( s = s + 1 \).
12: end while
13: \( z_i = \arg\max \mathcal{P}(t_i, h(x_i) | \mu^{(s)}, \Sigma^{(s)} \) .
14: \( Z(K) = \{z_i\}_{i=1}^n \).

Algorithm 2 LatGIV with Response Function Estimation

**Input:** Observational data \( \{x_i, t_i, y_i\}_{i=1}^n \); A family of function \( h \in \mathcal{H} \); Init Paramters \( \theta^{(0)} = \{\pi^{(0)}, \mu^{(0)}, \Sigma^{(0)}\} \); Threshold \( \tau \).

**Output:** Latent IVs \( Z \); Response Function \( \hat{g}(T, X) \).

1: \( Z(K) = \text{LatGIV}(K), K = \{2, 3, 5, 10\} \).
2: \( \text{MD}_K(Z, R) = \frac{2}{n(n-1)} \sum_{i=1}^K \sum_{j=i+1}^K ||R_{Z=i} - R_{Z=j}||_2^2 \), \( K = \{2, 3, 5, 10\} \).
3: \( K^* = \arg\min_K \text{MD}_K(Z, R), K = \{2, 3, 5, 10\} \).
4: \( Z = \text{LatGIV}(K^*) \).
5: Plug \( Z \) into get Response Function \( \hat{g}(T, X) \).

Hardware configuration: Ubuntu 16.04.5 LTS operating system with 2 * Intel Xeon E5-2678 v3 CPU, 384GB of RAM, and 4 * GeForce GTX 1080Ti GPU with 44GB of VRAM.

Software configuration: Python with Pytorch 1.2.0, TensorFlow 1.15.0, Keras 2.0.6, Scikit-learn 0.24.2, NumPy 1.17.4, and Matplotlib 3.1.1.

\( ^3 \) The code is available at: https://www.dropbox.com/sh/qaoga6rklzlbes1o/AABokKu_Ry1L3C2Pfgqreq5a7di=0.
E The experiments for testing stability

E.1 Synthetic datasets

For simulation, we generate the confounders \( \{X, \epsilon_Y\} \): \( X, \epsilon_Y \sim \mathcal{N}(0, \Sigma_{m_X+1}) \), \( \Sigma_{m_X+1} = \begin{bmatrix} I_{m_X} & \sigma \\ \sigma & 1 \end{bmatrix} \), where \( m_X \) is the dimensions of observed confounders \( X \), \( I_{m_X} \) denotes \( m_X \) order identity matrix, and \( \sigma \) denotes the covariance between confounders \( X \) and unmeasured confounder \( \epsilon_Y \). Without loss of generality, in this paper, we let \( \sigma = 0.1 \).

The instruments \( Z \) and treatments \( T \) are generated by Eqs. (18) & (19):
\[
T = \sum_{z=1}^{K} 1_{[Z=z]} \left[ \sum_{i=1}^{m} w_{zi}^l X_i + \sum_{i=1}^{m} w_{zi}^m f_X(X_i) + f_z(\epsilon) \right] + \delta_T, \delta_T \sim \mathcal{N}(0, 0.1)
\]
where \( f_X(X_i) = X_i \). In these section, we will discuss the clustering performance and stability of LatGIV algorithm for mixed data under different treatment assignment mechanisms with various unmeasured confounding \( f_z(\epsilon) \):
\[
f_z(\epsilon) = \begin{cases} 
0.2\epsilon + c_z, & \text{setting 1, expectations vary according to } Z; \\
0.5\epsilon + c_z, & \text{setting 2, increase variance;} \\
0.2\epsilon, & \text{setting 3, stable with the change of } Z; \\
w_{z}\epsilon, & \text{setting 4, variance vary according to } Z.
\end{cases}
\]
where \( c_z, w_z \sim \mathcal{U}(-1, 1) \), and \( z = \{1, \cdots, K\} \).

The response function \( Y \) is generated by Eq. (20):
\[
Y = -1.5T + 0.9T^2 + \sum_{i=1}^{m} \frac{X_i}{m} + |X_1 X_2| - \sin(10 + X_2 X_3) + 2\epsilon + \delta_Y
\]
where \( \epsilon \) is an unmeasured confounder and \( \delta_Y \sim \mathcal{N}(0, 0.1) \).

For synthetic datasets, we sample 3,000 units and perform 10 replications to report the mean squared error (MSE) and its standard deviations of the treatment-response function estimation, and the MSE is calculated over the testing data that we intervene the treatment as \( T = do(t) \). To verify the effectiveness of LatGIV\textsubscript{EM} in different scenarios with different \( m_X \)-dimension and different group numbers \( K \), we use Data-K-m\textsubscript{X} to denote the different dimensions of covariates and group number and perform 10 replications.

E.2 The results of stability

We increase the critical level of simulation and set Data-K-m\textsubscript{X} with different group number \( K \) and dimensions of covariates \( m_X \) to test the stability of our LatGIV. Comparing with the results of setting Linear-3-3 in Table 1, Linear-3-5 and Linear-3-10 in Table 3, LatGIV\textsubscript{EM} consistently achieves Top2 performance as the dimensions of covariates change. Adjusting the number of latent groups (Linear-2-5, Linear-3-5, Linear-5-5 in Table 3), LatGIV\textsubscript{EM} also shows in stability and is in Top2. In the above settings, LatGIV\textsubscript{EM} has outstanding performance, which is close to TrueIV.

Without loss of generality, we perform experiments with various unmeasured confounding \( f_z(\epsilon) \) (Eq. (44)) in Linear-3-3. Table 4 shows that LatGIV\textsubscript{EM} achieve the top-2 performance robustly with different settings in downstream IV-based methods, except Poly2SLS. In fact, the response function \( g(T, X) \) (Eq. (20)) is nonlinear and Poly2SLS may magnify arbitrary bias. In general, LatGIV\textsubscript{EM} is robust to varying critical simulation.

F The full results in IHDP & PM-CMR

F.1 Real-world datasets

Similar to previous methods [7,16,31,34], we perform experiments on two semi-synthetic real-world datasets IHDP [35] & PM-CMR [39], as the true response function is rarely available for real-world
We perform 10 replications to report the mean squared error (MSE) and its standard deviations (std) of different synthetic settings. Then we use 6 continuous variables about CMR in each city as mortality rate (CMR) in 2132 counties in the US using the data provided by the National Studies PM-CMR: https://pasteur.epa.gov/uploads/10.23719/1506014/SES_PM25_CMR_data.zip. From the original data, we select all 6 continuous variables as the confounders to replace the covariates aiming at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. From the original data, we select all 6 continuous variables as the confounders to replace the covariates X in Eq. (18)&(20) to generate the treatment T and the corresponding outcome Y.

**IHDP**[35]. The Infant Health and Development Program (IHDP) comprises 747 units with 6 pre-treatment continuous variables and 19 discrete variables related to the children and their mothers, aiming at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. From the original data, we select all 6 continuous variables as the confounders to replace the covariates X in Eq. (18)&(20) to generate the treatment T and the corresponding outcome Y.

**PM-CMR**[39]. The PM-CMR study the impact of $PM_{2.5}$ partial level on the cardiovascular mortality rate (CMR) in 2132 counties in the US using the data provided by the National Studies on Air Pollution and Health [39]. Then we use 6 continuous variables about CMR in each city as the confounders to replace the covariates X in Eq. (18)&(20) to generate the treatment T and the corresponding outcome Y.

---

### Table 3: The Mean Squared Error ($mean(\epsilon_d)$) of Different Synthetic Settings (Data-$K$-$m_X$)

| Setting1 | Setting2 | Setting3 | Setting4 |
|----------|----------|----------|----------|
| None     | Poly2SLS | Poly2SLS | Poly2SLS |
| UAS      | 0.203(0.017) | 0.200(0.017) | 0.200(0.017) |
| WAS      | 0.200(0.017) | 0.200(0.017) | 0.200(0.017) |
| ModelIV  | 0.200(0.017) | 0.200(0.017) | 0.200(0.017) |
| AutoIV   | 0.200(0.017) | 0.200(0.017) | 0.200(0.017) |
| LatGIV   | 0.059(0.004) | 0.059(0.004) | 0.059(0.004) |

---

### Table 4: The Mean Squared Error ($mean(\epsilon_d)$) of Different $\epsilon$ Experiments

| Setting1 | Setting2 | Setting3 | Setting4 |
|----------|----------|----------|----------|
| None     | Poly2SLS | Poly2SLS | Poly2SLS |
| UAS      | 0.330(0.037) | 0.330(0.037) | 0.330(0.037) |
| WAS      | 0.330(0.037) | 0.330(0.037) | 0.330(0.037) |
| ModelIV  | 0.330(0.037) | 0.330(0.037) | 0.330(0.037) |
| AutoIV   | >100 | >100 | >100 |
| LatGIV   | 0.260(0.128) | 0.260(0.128) | 0.260(0.128) |

---

**IHDP[35]**. The Infant Health and Development Program (IHDP) comprises 747 units with 6 pre-treatment continuous variables and 19 discrete variables related to the children and their mothers, aiming at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. From the original data, we select all 6 continuous variables as the confounders to replace the covariates X in Eq. (18)&(20) to generate the treatment T and the corresponding outcome Y.

**PM-CMR[39]**. The PM-CMR study the impact of $PM_{2.5}$ partial level on the cardiovascular mortality rate (CMR) in 2132 counties in the US using the data provided by the National Studies on Air Pollution and Health [39]. Then we use 6 continuous variables about CMR in each city as the confounders to replace the covariates X in Eq. (18)&(20) to generate the treatment T and the corresponding outcome Y.
F.2 Results

By estimating the latent differentiated covariate-treatment distribution parameters across groups, LatGIV$_{EM}$ successfully reconstructs the latent IV and the average reconstruction accuracy (Fig. 5) on IHDP and PM-CMR reaches 93.47% and 82.62% respectively. However, LatGIV$_{KM}$ fails to reconstruct it and achieves only 64.29% and 46.09% accuracy (Fig. 6), respectively. In addition, the correlation measure MMD (Fig. 5) between the observed covariates and the IV reconstructed by LatGIV$_{EM}$ is still lower than that of LatGIV$_{KM}$. This means that LatGIV$_{EM}$ can reconstruct the latent group IV more accurately with the data distribution from the real scenarios.

To verify that LatGIV$_{EM}$ with higher reconstruction accuracy achieves better performance to predict treatment-response function, we assess LatGIV and Summary IVs’ performance in treatment-response function estimation with the covariates from the real-world data IHDP & PM-CMR. We perform 10 replications and report the mean and standard deviations of MSE in the treatment-response function estimation here. The full results of MSE mean(std) of IHDP & PM-CMR Dataset with $T=do(t)$ and $T=do(0)$ are shown in Table 5 & 6 on Page 11, LatGIV$_{EM}$ shows consistent and robust performance, always maintaining the performance of top-2 and almost achieving the same effect as TrueIV on IHDP & PM-CMR Datasets. Compared with LatGIV$_{EM}$, the performance of LatGIV$_{KM}$ exceeds most baselines in downstream tasks, but it is still inferior to LatGIV$_{EM}$ and TrueIV.
Table 5: The Full Results of MSE mean(std) of IHDP & PM-CMR Dataset with T=do(t)

|                | IHDP Dataset |                | PM-CMR Dataset |                |
|----------------|--------------|----------------|----------------|--------------|
|                | Poly2SLS     | NW2LS          | KernelIV       | DualIV        |
| NoneIV         | 0.238(0.132) | 0.217(0.167)   | 0.456(0.243)   | 0.372(0.279)  |
| UAS            | 0.181(0.044) | 0.149(0.046)   | 0.372(0.303)   | 0.270(0.239)  |
| LatGIV         | 0.078(0.029) | 0.099(0.056)   | 0.147(0.069)   | 0.020(0.018)  |
| LatGIV**       | 0.205(0.159) | 0.253(0.141)   | 0.192(0.066)   | 0.303(0.069)  |
| LatGIV**       | 0.091(0.041) | 0.247(0.058)   | 0.170(0.079)   | 0.303(0.068)  |
| LatGIV**       | 0.024(0.012) | 0.120(0.037)   | 0.075(0.019)   | 0.220(0.046)  |
|                | 0.150(0.075) | 0.100(0.056)   | 0.217(0.067)   | 0.100(0.021)  |

Table 6: The Full Results of MSE mean(std) of IHDP & PM-CMR Dataset with T=do(0)

|                | IHDP Dataset |                | PM-CMR Dataset |                |
|----------------|--------------|----------------|----------------|--------------|
|                | Poly2SLS     | NW2LS          | KernelIV       | DualIV        |
| NoneIV         | 0.350(0.129) | 0.379(0.179)   | 0.322(0.121)   | 0.370(0.067)  |
| UAS            | 0.349(0.130) | 0.443(0.152)   | 0.323(0.121)   | 0.395(0.090)  |
| LatGIV         | 0.160(0.069) | 0.202(0.098)   | 0.203(0.058)   | 0.411(0.078)  |
| LatGIV**       | 0.206(0.054) | 0.450(0.123)   | 0.200(0.072)   | 0.232(0.046)  |
| LatGIV**       | 0.091(0.041) | 0.247(0.058)   | 0.170(0.079)   | 0.303(0.068)  |
| LatGIV**       | 0.024(0.012) | 0.120(0.037)   | 0.075(0.019)   | 0.220(0.046)  |
|                | 0.150(0.075) | 0.100(0.056)   | 0.217(0.067)   | 0.100(0.021)  |

22