# Learning Instrumental Variable from Data Fusion for Treatment Effect Estimation

**Anpeng Wu[1], Kun Kuang[1]*, Ruoxuan Xiong[2], Minqing Zhu[1], Yuxuan Liu[1], Bo Li[3], Furui Liu[4], Zhihua Wang[5,6], Fei Wu[1,5,6]**

[1] Department of Computer Science and Technology, Zhejiang University
[2] Department of Quantitative Theory and Methods, Emory University
[3] School of Economics and Management, Tsinghua University
[4] Huawei Noah's Ark lab
[5] Shanghai AI Laboratory
[6] Shanghai Institute for Advanced Study of Zhejiang University
{anpwu, kunkuang}@zju.edu.cn, ruoxuan.xiong@emory.edu, {minqinzhu, yuxuanliu}@zju.edu.cn, libo@sem.tsinghua.edu.cn, liufurui2@huawei.com, {zhihua.wang, feiwu}@zju.edu.cn

## Abstract

The advent of the big data era brought new opportunities and challenges to draw treatment effect in data fusion, that is, a mixed dataset collected from multiple sources (each source with an independent treatment assignment mechanism). Due to possibly omitted source labels and unmeasured confounders, traditional methods cannot estimate individual treatment assignment probability and infer treatment effect effectively. Therefore, we propose to reconstruct the source label and model it as a Group Instrumental Variable (GIV) to implement IV-based Regression for treatment effect estimation. In this paper, we conceptualize this line of thought and develop a unified framework (Meta-EM) to (1) map the raw data into a representation space to construct Linear Mixed Models for the assigned treatment variable; (2) estimate the distribution differences and model the GIV for the different treatment assignment mechanisms; and (3) adopt an alternating training strategy to iteratively optimize the representations and the joint distribution to model GIV for IV regression. Empirical results demonstrate the advantages of our Meta-EM compared with state-of-the-art methods. The project page with the code and the Supplementary materials is available at https://github.com/causal-machine-learning-lab/meta-em.

## Introduction

Estimating the causal effects of treatment/exposure on the outcome of interest from the observation dataset is crucial for explanatory analysis and decision-making (Pearl 2009; Kuang et al. 2020b; Li et al. 2020; Zhang et al. 2021; Tian et al. 2022). In the presence of unmeasured confounders, assuming a fixed additive noise model (ANM), state-of-the-art (SOTA) approaches use an instrumental variable (IV) to implement a two-stage regression to reduce endogenous confounding bias in treatment effect estimation (Hartford et al. 2017; Lin et al. 2019; Muandet et al. 2020; Wu et al. 2022a). These 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. Under these assumptions, Angrist et al. (1996); Newey and Powell (2003) verify that causal effects can be identified by exogenous IVs. In instrumental variable literature, researchers usually implement Randomized Controlled Trials (RCTs) to obtain exogenous IVs, such as Oregon health insurance experiment (Finkelstein et al. 2012) and effects of military service on lifetime earnings (Angrist 1990), which are too expensive to be universally available.

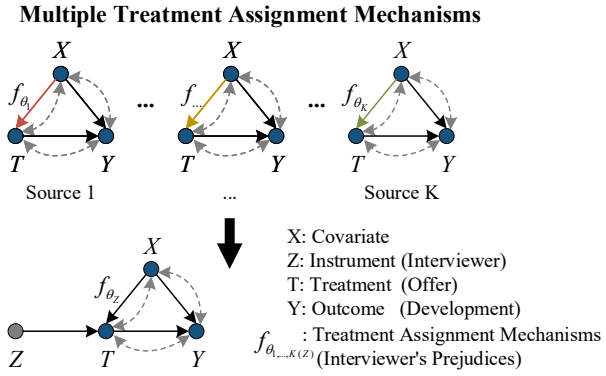


Figure 1: The causal diagram for mixed datasets from multiple sources, each source with an independent treatment assignment mechanism. Blue nodes denote observable variables, and gray nodes indicate latent variables. The arrows with different colors define different causal effects. The bidirected arrows encode unmeasured confounders.

With the advent of the big data era, a variety of observation databases collected from different sources have been established, which may contain the same treatment effect mechanism (from treatment to outcome) but different treatment assignment mechanisms (from covariates to treatment) (Bareinboim and Pearl 2016; Hünermund and Bareinboim 2019). For instance, as shown in Fig. 1, in the study of treatment effect of individual offers (treatment $T$) on enterprise development (outcome $Y$), different human resources (HR)

*Corresponding author.

interviewers (instrument $Z$) may assign different offer decisions to the same individual (covariate $X$) based on different evaluation strategies (assignment $f_{\theta_Z}$). In this case, candidates will be randomly assigned to different interviewers, each with different prejudices or opinions, to decide whether to give an offer or not (Pager and Karafin 2009). Here, the omitted interviewer label (source label) can serve as a latent multi-valued IV, which only affects the outcome through its strong association with offer decisions (Kuang et al. 2020c; Rothenhäusler et al. 2021). Such heterogeneous assignment mechanism is common and widespread in real applications, such as the assessment rules in university admissions or academic title evaluation (Harris et al. 2022), and the environments in Generalized Causal Dantzig (Long et al. 2022).

Nevertheless, due to data privacy and missing data, interviewers' information is rare in public datasets. Besides, the source label is not always available in some scenarios. For example, people tend to consult an expert consultant, and the consultant's emotional state could be a latent IV that cannot be accessed. A large amount of literature for Summary/Selection IVs has attempted to resolve this problem (Burgess, Small, and Thompson 2017; Kuang et al. 2020c; Hartford et al. 2021; Yuan et al. 2022). Two main limitations of these methods are that they require expert knowledge to provide well-predefined IV candidates, and lack metrics to test the validity of IV variables learned by unsupervised methods. Moreover, to obtain valid Summary IVs, these methods assume that at least half of pre-defined IV candidates are valid strong IVs so that they can synthesize an IV through a weighted average (Burgess and Thompson 2013; Davies et al. 2015; Burgess, Small, and Thompson 2017).

Since summary IVs require half of the IV candidates to be valid, which rarely happens in practice, the estimation might be unreliable. Therefore, it is highly demanded to model latent IVs and implement a data-driven approach to automatically obtain valid IVs directly from the observed variables $\{X, T, Y\}$, *without pre-defined hand-made IV candidates*. Fortunately, the advent of the big data era brought new opportunities to reconstruct IVs from multiple sources data (each source with an independent treatment assignment mechanism). In the offer case (Fig. 1), the interviewers generate multiple causal relations between the covariates and the treatment, and it can serve as a latent multi-valued IV.
**Motivation**: *Thus, we propose to separate the observational data into multiple groups to reconstruct the source label and then explicitly model the group indicator as a Group Instrumental Variable (GIV) to implement IV-based Regression.*

In this paper, we aim to recover latent IV and estimate the individual treatment effect (ITE) from mixed observational datasets in the presence of unmeasured confounders. Due to possibly omitted source labels and unmeasured confounders, traditional methods cannot estimate individual treatment assignment probability and infer treatment effect effectively (Wu et al. 2022b; Kuang et al. 2017, 2020a). Therefore, we propose to reconstruct the source label and model it as a Group Instrumental Variable (GIV) to implement IV-based Regression for treatment effect estimation. In this paper, we conceptualize this line of thought and develop a unified framework (Meta-EM[1]) to (1) map the raw data into a representation space to construct Linear Mixed Models for the assigned treatment variable; (2) estimate the distribution differences and model the GIV for the different treatment assignment mechanisms using Expectation-Maximization algorithm (EM); and (3) adopt an alternating training strategy to iteratively optimize the representations and the joint distribution to model GIV for IV regression. Empirical results demonstrate the advantages of the GIV compared with SOTA methods.

The contribution of our paper is three-fold:

- We propose a Meta-EM algorithm to reconstruct the source label as GIV directly from the observed variables, i.e., no available IV candidates for learning, which is beyond the capability of existing Summary IV methods. GIV (source label) is effective when there are identifiable differences in mechanisms across groups.
- Meta-EM algorithm uses a shared representation block to learn a nonlinear representation space to EM algorithm, which relaxes the underlying linear regression assumption. Theoretically, Meta-EM can obtain an asymptotic source label as GIV for ITE estimation.
- We empirically demonstrate that the Meta-EM algorithm reconstructs the source label as GIV from the observed variables for accurate treatment effect estimation and gains SOTA performance compared with existing summary IV methods.

## Related Work

### Instrumental Variable Methods

The sufficient identification results for causal effect under the additive noise assumption in *instrumental variable* regression were developed by (Imbens and Angrist 1994; Newey and Powell 2003). For semi-parametric and nonparametric estimation, there are four main research lines about IV methods, including: (1) *The two-Stage Least Squares,* Poly2SLS and NN2SLS; (2) *The Kernel-based Methods,* Kernel IV (Singh, Sahani, and Gretton 2019) and DualIV (Muandet et al. 2020) map X to a reproducing kernel Hilbert space (RKHS); (3) *The Deep Methods,* DeepIV (Hartford et al. 2017), OneSIV (Lin et al. 2019) and DFIV (Xu et al. 2021) adopts deep neural nets and fit a mixture density network; (4) *The Adversarial GMM,* AGMM (Dikkala et al. 2020) and DeepGMM (Bennett et al. 2019) construct a structural function and select moment conditions via adversarial training.

The above methods are reliable only if the pre-defined IVs are valid and strongly correlated with the treatment. In practice, such valid IVs are hardly satisfied due to the untestable exclusion assumption. In this paper, we reconstruct a GIV and plug it into IV methods to predict the treatment effect.

### Summary IV Synthesis

A growing number of works have been proposed to synthesize a Summary IV by combining existing IV candidates. In

[1]Meta means "learn nonlinear mappings to learn EM".

Mendelian Randomization (MR), IV candidates are merged into a summary IV by unweighted/weighted allele scores (UAS/WAS) (Burgess, Small, and Thompson 2017; Kuang et al. 2020c), *UAS* takes the average of IV candidates while *WAS* weights each IV candidate based on the associations with the treatment. Besides, *ModeIV* (Hartford et al. 2021) adopts the tightest cluster center of estimation points as IV to approximate causal effects. Assuming that all IV candidates are independent of the unmeasured confounders, *AutoIV* (Yuan et al. 2022) generates IV representation. Existing Summary IV Methods require a high-quality IV candidates' set with at least half valid IVs, which is unrealistic in practice due to cost issues and lack of expert knowledge. Under a more practical setting, we model latent IVs and implement a data-driven approach to automatically reconstruct valid Group IVs directly from the observed variables, without hand-made IV candidates.

## Problem Setup and Assumptions

In this paper, we aim to learn latent IV and estimate the individual treatment effect (ITE) from mixed datasets in the presence of unmeasured confounders. As shown in Fig. 1, a mixed dataset $D = \{D_1, D_2, \cdots, D_K\}$ collected from $K$ sources $D_k = \{x_i, t_i, y_i, \epsilon_i \mid f_{\theta_k}\}_{i=1}^{n_k}, k = 1, 2, \cdots, K$, each source with an independent treatment assignment mechanism[2] $f_{\theta_k}$, the size of samples from source $k$ is $n_k$ and the total sample size is $n = \sum_{k=1}^{K} n_k$. For unit $i$ from source $z_i = k$, we observe confounders $x_i \in X$ where $X \subset \mathbb{R}^{m_X}$ with dimension $m_X$, a treatment variable $t_i \in T$ from mechanism $f_{\theta_k}$ where $T \subset \mathbb{R}$, and a outcome variable $y_i \in Y$ where $Y \subset \mathbb{R}$. In data fusion, due to data privacy and missing data, the source label $z_i$ and some key confounders may be unrecorded in observational data. We incorporate the unobserved confounders into the term $\epsilon_i$.

Without interactions between unmeasured confounders and treatment, we can represent the effect of infinitely many unmeasured causes as an additive noise $\{\epsilon_T, \epsilon_Y\}$ regardless of how they interact among themselves. The sufficient identification results for causal effect under *the additive noise assumption* in *instrumental variable* regression were developed by (Angrist et al. 1996; Newey and Powell 2003).

**Assumption 0.1** ***Additive Noise Assumption****: Similar to (Hartford et al. 2017; Singh, Sahani, and Gretton 2019; Xu et al. 2021), we assume that the mixed data is generated by:*

$$T = f_{\theta_Z}(X) + \epsilon_T, Y = g(T, X) + \epsilon_Y, \tag{1}$$

**Definition 0.2** ***Individual Treatment Effect (ITE)****:*
$\tau = g(t, X) - g(0, X), g(t, X) = \mathbb{E}[Y \mid do(T = t), X]$.

**Definition 0.3** ***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:*** *interviewers $Z$ assign treatments $T$ to each unit, i.e., $\mathbb{P}(T \mid Z) \neq \mathbb{P}(T)$.*
***Exclusion:*** *interviewers $Z$ does not directly affect the outcome $Y$, i.e., $Z \perp Y \mid T, X, \epsilon$.*
***Unconfounded:*** *the offer-seekers will be randomly assigned to different interviewers, so $Z$ is independent of all confounders, including $X$ and $\epsilon$, i.e., $Z \perp X, \epsilon$.*

With the advent of the big data era, a variety of observation databases collected from different sources have been established, which may contain the same treatment effect mechanism (from treatment to outcome) but different treatment assignment mechanisms (from covariates to treatment). Such heterogeneous assignment mechanism is common and widespread in real applications(Harris et al. 2022; Long et al. 2022). Plausible settings include: the socioeconomic status influences treatment but not outcomes, and admissions assessment rules affect students' SAT scores but do not determine their success in college. In addition, there are many subtle factors that are easily overlooked in real-world applications that may be latent assignment variables, such as weather, holiday, mood, dresses, travel style, lunch, etc. All of them may only affect the treatment choice without directly changing the outcome, but they are often ignored. In the presence of such assignment variables, we propose to separate the observational data into multiple groups to reconstruct the assignment variables and then explicitly model the group indicator as a Group Instrumental Variable (GIV) to implement IV-based Regression.

## Algorithm

In this section, we propose a Meta-EM algorithm to automatically identify the latent source label $Z$, inducing the different treatment assignment mechanisms, as group indicator to separate data into multiple groups. Specifically, the overall Meta-EM architecture (Fig. 2) of our model consists of the following components: (1) Meta-EM uses a **Shared Network Layer** to map the covariates $X$ to non-linear representations $R$, and then uses latent variable $Z$ (obtained from EM algorithm) to regress the treatment variables and optimize the representation; (2) Meta-EM estimates the distribution differences across sources and models latent variable as a GIV for the different treatment assignment mechanisms using **Expectation-Maximization** algorithm (EM); (3) Meta-EM adopts an alternating training strategy to iteratively optimize the **Representations** and the joint distribution for **GIV Reconstruction**. Theoretically, Meta-EM achieve an asymptotic IV and accurately predict ITE by plugging GIV into downstream IV-based methods.

### Representation Learning Step

Let $z_i = k$ denotes the latent source label ($k = 1, 2, \cdots, K$) for unit $i$, and source number $K$ is a hyper-parameter. To construct Linear Mixed Models for the assigned treatment variable $T$, we use a representation function $f_R$ maps the covariates $X \in \mathbb{R}^{m_X}$ into a representation $R \in \mathbb{R}^{m_R}$. Consider the following representation model (Fig. 2(a)):

$$t_i = f_{\theta_{z_i}}(x_i) + \epsilon_i = \alpha'_{z_i} f_R(x_i) + \epsilon_i = \alpha'_{z_i} r_i + \epsilon_i, \tag{2}$$

where $f_R$ is a shared representation block which can be learned from polynomial functions, kernel functions or a

[2] In causal inference, we assume the causal effect of treatment on the outcome is invariant across sources. If the treatment effect varies across sources, then we will not identify which treatment effect mechanism the individual's outcome came from in testing.

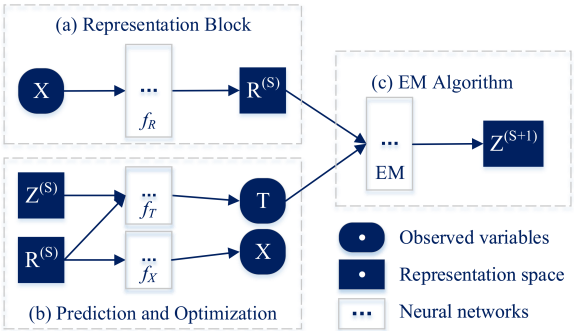


Figure 2: Overview of Meta-EM Architecture.

neural network, $r_i$ is a (non-)linear representation of $x_i$, and $\alpha_{z_i}$ is the corresponding coefficients for Linear Mixed Models. Then we formulate a linear (non-)gaussian mixed model:

$$t_i = \sum_{k=1}^{K} 1_{z_i=k} \left(\alpha_k' r_i + \epsilon_i\right), \tag{3}$$

where $1_{z_i=k}$ denote the indicator function. Specifically, in polynomial from, we expect to obtain $t_i = \sum_{k=1}^{K} 1_{z_i=k} \left\{\sum_{j=1}^{m_R} \left[\alpha_{kj}(\xi_{kj,1} x_{ij}^1 + \xi_{kj,2} x_{ij}^2 + \cdots)\right] + \epsilon_i\right\}$, where $\xi_{kj,d}$ denotes the corresponding expectation coefficient of the $d$-th power of $j$-th variable $x_{ij}$.

We design two prediction networks $f_T$ and $f_X$ to regress treatment $T$ and covariates $X$, and adopt an alternating training strategy to optimize the representations iteratively (Fig. 2(b)), and the superscript $(s)$ denotes the $s$-th iteration:

$$\begin{aligned} \mathcal{L} &= \sum_i^n \left(f_T(z_i^{(s)}, r_i^{(s)}) - t_i\right)^2 + \lambda \sum_i^n \left(f_X(r_i^{(s)}) - x_i\right)^2 \\ &= \sum_i^n \left(\alpha'_{z_i^{(s)}} r_i^{(s)} - t_i\right)^2 + \lambda \sum_i^n \left(f_X(r_i^{(s)}) - x_i\right)^2, \end{aligned} \tag{4}$$

In the term $f_T(z_i^{(s)}, r_i^{(s)})$, $z_i^{(s)}$ is a latent function indicator to activate the corresponding linear coefficients $\alpha_{z_i}$ for treatment regression, and the representation $r_i$ is shared in all sources. The second term is a regularization term to ensure that the representation contains as much information as possible from the original data. Besides, $\lambda$ is a trade-off parameter to control the relative importance of treatment regression and covariate regression. We let $\lambda = 1/m_X$, representing that we adopt mean square of $L_2$ norm ($\lambda(f_X(r_i) - x_i)^2 = \sum_j (f_j - x_{ij})^2/m_X$, and $j = 1, \cdots, m_X$) in covariates regression and treat it as important as treatment regression. By minimizing $\mathcal{L}$, our model can map the raw data into a representation space to construct Linear Mixed Models for the assigned treatment variable.

## Distribution Learning Step

Based on the traditional Expectation-Maximization (EM) algorithm with group number $K$ (Fig. 2(c)), we seek to find the Maximum Likelihood Estimate (MLE) of the marginal likelihood by iteratively applying the Expectation step and Maximization step. Consider the following log-likelihood function for Gaussian Mixture with $\theta = \{\pi, \mu, \Sigma\}$:

$$\begin{aligned} &\log Pr(C_{TR} \mid \theta) = \log Pr(C_{TR} \mid \pi, \mu, \Sigma) \\ = &\sum_{i=1}^{n} \sum_{k=1}^{K} \log\left(\pi_k Pr(c_i \mid \mu_k, \Sigma_k)\right)^{1_{z_i=k}}, \end{aligned} \tag{5}$$

where $C_{TR}$ denotes the concatenation of $T$ and $R$, $c_i = \{t_i, r_i\}$, and $\pi_k = Pr(z_i = k), k = 1, 2, \cdots, K$. $\mu_k$ and $\Sigma_k$ are the mean vector and covariance matrix of samples $\{c_i\}_{i:z_i=k}$ for group $k$. $Pr(c_i \mid \mu_k, \Sigma_k)$ is the density of $c_i$ conditional on $z_i = k$:

$$Pr(c_i \mid \mu_k, \Sigma_k) = (2\pi)^{-\frac{m_R}{2}} |\Sigma_j|^{-\frac{1}{2}} e^{-\frac{1}{2}(c_i - \mu_k)' \Sigma_k^{-1} (c_i - \mu_k)}.$$

**Initialization.** $Pr(R^{(s)})$ and $Pr(\epsilon)$ should be fixed among all groups since the population does not change according to treatment assignment (Fig. 1), i.e., $Pr(R|Z = i) = Pr(R|Z = j)$ for any groups $Z = i$ and $Z = j$. Therefore, we can use $\mathbb{E}[R]$ and $Cov(R, R)$ to initialize the distribution parameters $\theta^{[0]} = \{\pi^{[0]}, \mu^{[0]}, \Sigma^{[0]}\} = \{\{\pi_k^{[0]}, \mu_k^{[0]}, \Sigma_k^{[0]}\}_{k=1,2,\cdots,K}\}$ with $\pi_k^{[0]} = 1/K$:

$$\mu_k^{[0]} = \{\mu_k^{[0]}(T), \mathbb{E}[R]\}, \Sigma_k^{[0]} = \begin{bmatrix} \sigma_k^{[0]}(T,T) & \sigma_k^{[0]}(T,R)^T \\ \sigma_k^{[0]}(T,R) & Cov(R,R) \end{bmatrix}, \tag{6}$$

where $\{\mu_k^{[0]}(T), \sigma_k^{[0]}(T,T), \sigma_k^{[0]}(T,R)\}$ are the random initialization of the mean of $T$, the covariance of $T$, and the covariance matrix of $T$ and $R$ in the group $Z = k$, respectively. The superscript $[h]$ denotes $h$-th iteration

**Expectation Step.** In the expectation step of the $v$-th iteration, given the observation data $C_{TR}^{(s)}$ and current parameter estimation $\theta^{[h]} = \{\pi^{[h]}, \mu^{[h]}, \Sigma^{[h]}\}$, we calculate the log expectation of likelihood function Eq. (5):

$$\mathcal{Q}(\theta^{[h]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_{ik} \log\left(\pi_k Pr(c_i \mid \mu_k, \Sigma_k)\right)^{1_{z_i=k}}, \tag{7}$$

where $\hat{\gamma}_{ik}$ is the conditional probability distribution that the $i$-th unit comes from the $k$-th group given $\theta^{[h]}$:

$$\hat{\gamma}_{ik} = Pr(z_i = k \mid \theta^{[h]}) = \frac{\pi_k^{[h]} Pr(c_i | \mu_k^{[h]}, \Sigma_k^{[h]})}{\sum_{j=1}^{K} \pi_j^{[h]} Pr(c_i | \mu_j^{[h]}, \Sigma_j^{[h]})}. \tag{8}$$

**Maximization Step.** In the maximization step of the $h$-th iteration, given the observational data $C_{TR}^{(s)}$ and the current parameter estimation $\theta^{[h]} = \{\pi^{[h]}, \mu^{[h]}, \Sigma^{[h]}\}$, we maximize the expectation of the log-likelihood function $\mathcal{Q}(\{\pi^{[h]}, \mu^{[h]}, \Sigma^{[h]}\})$ (Eq. (7)) to obtain the parameter estimation $\theta^{[h+1]}$ of next iteration:

$$\theta^{[h+1]} = \text{argmax}_\theta \mathcal{Q}(\{\pi^{[h]}, \mu^{[h]}, \Sigma^{[h]}\}). \tag{9}$$

The solution is: for any $k = 1, 2, \cdots, K$,

$$\mu_k^{[h+1]} = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} c_i}{\sum_{i=1}^{n} \hat{\gamma}_{ik}}, \tag{10}$$

$$\Sigma_k^{[h+1]} = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} \left[c_i - \mu_k^{[h+1]}\right]^2}{\sum_{i=1}^{n} \hat{\gamma}_{ik}}, \tag{11}$$

$$\pi_k^{[h+1]} = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik}}{n}. \tag{12}$$

Then, the EM algorithm would obtain the convergent parameters $\theta^* = \{\pi^*, \mu^*, \Sigma^*\}$ by iteratively applying the Expectation and Maximization steps. We can sample/identify the sub-group indicator $Z^{(s+1)}$ from the estimated distribution parameters $\{\pi^*, \mu^*, \Sigma^*\}$:

$$\gamma_{ik}^{(s+1)} = Pr(z_i = k) = \frac{\pi_k^* Pr(c_i^{(s)} | \mu_k^*, \Sigma_k^*)}{\sum_{j=1}^{K} \pi_j^* Pr(c_i^{(s)} | \mu_j^*, \Sigma_j^*)},$$

$$z_i^{(s+1)} \sim \text{Disc}(\gamma_{i1}^{(s+1)}, \gamma_{i2}^{(s+1)}, \cdots, \gamma_{iK}^{(s+1)}),$$

where $\text{Disc}(\cdot)$ denotes discrete distribution with $\{\gamma_{ik}\}_{k=1}^{K}$.

## Optimization and MMD

Specifically, our Meta-EM is composed of two phases: in representation learning phase, we fix GIV z to optimize representation network and corresponding coefficients ($f_R$ and $\alpha$) by minimizing objective in Eq. (4); in distribution learning phase, we use representation R to re-devide group and update GIV $z$ using EM algorithm. As shown in Fig 2, the Meta-EM algorithm would learn an optimal representation $R^*$ to learn $\gamma_{ik}^{(*)}$ and $z_i^{(*)}$ by iteratively applying the representation learning step and distribution learning step:

$$\gamma_{ik}^{(*)} = Pr(z_i = k) = \frac{\pi_k^* Pr(c_i^{(*)}|\mu_k^*, \Sigma_k^*)}{\sum_{j=1}^K \pi_j^* Pr(c_i^{(*)}|\mu_j^*, \Sigma_j^*)},$$
$$z_i^{(*)} \sim \text{Disc}(\gamma_{i1}^{(*)}, \gamma_{i2}^{(*)}, \cdots, \gamma_{iK}^{(*)}).$$

Suppose each coordinate in the coefficient vector $\alpha_k$ in Eq. (3) is nonzero for all $K = k$. As $(m_R, n) \to \infty$, $\hat{\gamma}_{ik}$ converges to $1_{z_i=k}$ with the rate $o(exp(-(m_R + M)))$ for each $k$, where $m_R$ is the dimension of the representations. For theorems and proofs, see the Supplementary material.

As shown in Fig. 1, $Pr(X)$ should be fixed among all groups since the population does not change according to treatment assignment, which means the instrumental variable should be independent of all confounders (Unconfounded Assumption of IV), i.e., $Pr(X|Z = i) = Pr(X|Z = j)$ for any groups $Z = i$ and $Z = j$. To implement an end-to-end algorithm, we use Maximum Mean Discrepancy (MMD) to measure the correlation between discrete variable $Z$ and observed confounder $X$: $MMD = \|\mathbb{E}[R|Z = i] - \mathbb{E}[X|Z = j]\|_2^2$. Furthermore, we can automatically select the most appropriate group number $K^*$ by the minimum correlation:

$$MMD_K = \frac{2}{n(n-1)} \sum_{i=1}^K \sum_{j=i+1}^K ||\bar{X}_{Z=i} - \bar{X}_{Z=j}||_2^2,$$
$$K^* = \text{argmin}_K MMD_K, K = \{2, 3, ...\}.$$

where $\bar{X}_{Z=i}$ denotes the mean of the covariates $X$ in the $i$-th sub-group according to the EM algorithm.

# Experiments

## Baselines

**IV Generation** In this paper, we adopt Meta-EM with MMD to find the most appropriate group number $K$ and take the cluster results as $\mathbf{GIV}_{EM}$. We compare our algorithm Meta-EM with the Summary IV methods: (1) **NoneIV** uses a full-zeros vector as IV; (2) **UAS** (Davies et al. 2015) takes the average of IV candidates as IV; (3) **WAS** (Burgess, Dudbridge, and Thompson 2016) weights each candidate based on the associations as IV; (4) **ModeIV** (Hartford et al. 2021) takes the tightest center of estimation points as IV; (5) **AutoIV**(Yuan et al. 2022) learns a disentangled representation as IV. Besides, we adopt Meta-KM[3] to generate $\mathbf{GIV}_{KM}$ and use the superscript $*$ to represent the priori of the number of groups, i.e., $\mathbf{GIV}_{KM}*$. **TrueIV** denotes the known ground-truth source label.

[3]Meta-KM is the K-means replacement of Meta-EM.

**IV Regression** To evaluate the performance of Meta-EM for IV generation, we plug synthetic IVs, obtained from Meta-EM and other IV generation baselines, into IV regression methods (as listed in Related Work) for ITE estimation.

## Experiments on Synthetic Datasets

Similar to DeepIV (Hartford et al. 2017), DFIV (Xu et al. 2021), DeepGMM (Bennett et al. 2019) and AutoIV (Yuan et al. 2022), due to lack of the prior of latent outcome function and instrumental variable in existing real-world datasets, we evaluate and compare our algorithm Meta-EM with the above baselines on the synthetic and semi-synthetic data. To simulate real-world data as much as possible, we adjust the difficulty of the simulation and expand experiments to various non-linear scenarios (Fig. 3), increase the number of sub-groups and the dimension of covariates (Table 4 in Supplementary material).

**Datasets** We generate the synthetic datasets as follows:

- **The confounders $\{X, \epsilon\}$:**

$$X, \epsilon \sim \mathcal{N}(0, \Sigma_{m_X+1}), \Sigma_{m_X+1} = \begin{bmatrix} \mathrm{I}_{m_X} & \sigma \\ \sigma & 1 \end{bmatrix}, \quad (13)$$

where $m_X$ is the dimensions of observed confounders $X$, $\mathrm{I}_{m_X}$ denotes $m_X$ order identity matrix, and $\sigma$ denotes the covariance between confounders $X$ and unmeasured confounder $\epsilon$. In this paper, we let $\sigma = 0.1$.

- **The treatments $T$ collected from multiple sources $Z$:**

$$T = \sum_{z=1}^K 1_{[Z=z]} [\sum_{i=1}^{m_X} w_{zi}[X_i + f_X(X_i)] + f_z(\epsilon)] + \delta_T, \quad (14)$$
$$Z \sim Pr(Z = z) = 1/K, w_{zi} \sim \text{Unif}(-1, 1), z = 1, \cdots, K, \quad (15)$$

where $X_i, i = \{1, \cdots, m_X\}$ denotes the $i$-th variable in $X$, $\delta_T \sim \mathcal{N}(0, 0.1)$, Unif means we draw $w_{zi}$ from the parameterized uniform distribution, and $f_z(\epsilon) = 0.2\epsilon$. The mixed data derives from $K$ different sources, meaning that there are $K$ independent potential treatment assignment models. $Z$ is the indicator of the potential treatment assignment model, which can be regarded as an instrumental variable. To simulate real-world data as much as possible, we design 5 different treatment functions $f_X(\cdot)$ to discuss the performance of Meta-EM 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 latent outcome function $Y$:**

$$Y = - 1.5T + 0.9T^2 + \sum_{i=1}^m \frac{X_i}{m} + |X_1 X_2|$$
$$- \sin(10 + X_2 X_3) + 2\epsilon + \delta_Y. \quad (16)$$

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 independent replications to report mean squared error (MSE) and standard deviations of the individual treatment effect estimation over the testing data (3000 units) that we intervene the treatment as $T = do(t)$. To verify the effectiveness of $\text{GIV}_{EM}$ in different scenarios with different dimensions of covariates $m_X$ and different group numbers $K$, we use $Data$-$K$-$m_X$ to denote the different scenarios. In this paper, we set the representation dimension as $m_R = m_X$.

| | Poly2SLS | NN2SLS | KernelIV | DualIV$^{(1)}$ | DeepIV | OneSIV | DFIV$^{(1)}$ | DeepGMM | AGMM |
|---|---|---|---|---|---|---|---|---|---|
| NoneIV | 0.33(0.04) | 1.90(1.25) | 0.35(0.07) | 1.92(0.48) | 0.37(0.01) | 0.31(0.03) | 1.33(0.14) | 0.33(0.07) | 0.21(0.05) |
| UAS | 0.33(0.04) | 2.30(1.46) | 0.35(0.07) | **0.98(0.34)** | 0.37(0.02) | 0.31(0.03) | 1.30(0.10) | 0.32(0.04) | 0.21(0.05) |
| WAS | 0.31(0.04) | 1.59(0.92) | 0.36(0.05) | 2.16(0.46) | 0.37(0.02) | 0.34(0.03) | 1.29(0.12) | 0.32(0.06) | 0.23(0.04) |
| ModeIV | 0.33(0.04) | 2.25(1.30) | 0.35(0.08) | 1.90(0.56) | 0.37(0.02) | 0.31(0.02) | 1.29(0.12) | 0.31(0.07) | 0.20(0.04) |
| AutoIV | $> 100^{(2)}$ | 2.10(1.01) | 0.35(0.07) | **0.79(0.32)** | 0.37(0.02) | 0.31(0.03) | 1.29(0.11) | 0.31(0.09) | 0.21(0.05) |
| GIV$_{KM}$ | 0.27(0.13) | 0.65(0.36) | 0.22(0.04) | 1.47(0.26) | 0.28(0.01) | 0.23(0.02) | 1.25(0.11) | 0.14(0.01) | 0.12(0.02) |
| GIV$_{KM}$* | 0.19(0.09) | 0.37(0.22) | 0.21(0.03) | 1.60(0.36) | 0.26(0.04) | 0.22(0.03) | 1.24(0.10) | 0.14(0.04) | 0.10(0.01) |
| GIV$_{EM}$ | **0.05(0.00)** | **0.12(0.01)** | **0.11(0.02)** | 1.99(0.41) | **0.08(0.00)** | **0.12(0.01)** | **0.79(0.08)** | **0.08(0.01)** | **0.06(0.00)** |
| TrueIV | **0.05(0.00)** | **0.08(0.01)** | **0.11(0.02)** | 1.93(0.39) | **0.08(0.00)** | **0.11(0.01)** | **0.79(0.06)** | **0.06(0.01)** | **0.06(0.01)** |

(1) DualIV and DFIV don't perform well on GIV, because they require continuous IVs rather than discrete IVs. (2) "$> 100$" means "MSE $> 100$".

Table 1: The Mean Squared Error $mean(std)$ on Linear Experiments ($Linear$-3-3)

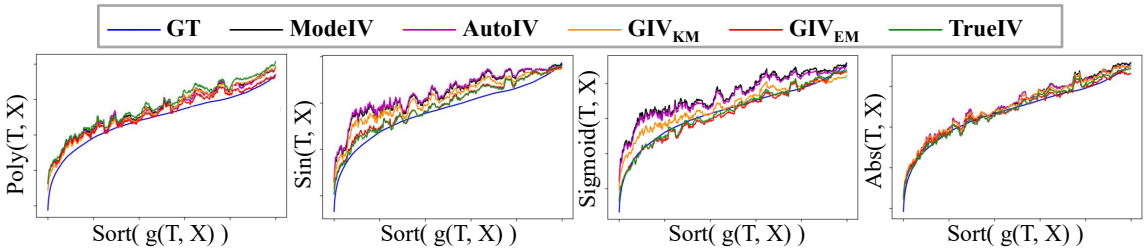


Figure 3: Treatment Effect Estimation (sorted by Ground-Truth $g(T, X)$) in Non-linear Scenario $Data$-3-3.

| | Linear | Poly | Sin | Sigmoid | Abs |
|---|---|---|---|---|---|
| EM | 86.1% | 82.9% | 87.2% | 89.4% | 85.2% |
| Meta-EM | **86.1%** | **92.7%** | **90.3%** | **94.7%** | **92.3%** |
| Fun-EM | **86.1%** | **98.1%** | **96.3%** | **98.4%** | **96.5%** |

Table 2: Ablation Experiments for Meta-EM in $Data$-2-5

**The Results of Individual Treatment Effect Estimation** As shown in Table 1 (The top-2 is highlighted in bold for all tables), 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) DualIV and DFIV do not perform well on GIV and fail to estimate treatment effect, even with TrueIV, because they require continuous IVs rather than discrete IVs. (3) Through clustering, we reconstruct the latent exogenous IV that generates different treatment mechanisms, GIVs (with Meta-KM or Meta-EM) bring higher accuracy on individual treatment effect estimation by comparing with Summary IV methods in various IV-based methods except for DualIV; (4) By estimating the latent differentiated covariate-treatment distribution parameters across groups and reconstructing the source label, GIV$_{EM}$ significantly improves the performance of clustering methods compared with GIV$_{KM}$ and achieves SOTA performance for individual treatment effect estimation, even comparable with TrueIV. Empirically, this demonstrates that our Meta-EM successfully reconstructs the GIV, and it converges to the TrueIV, i.e., source label.

Then, to verify the effectiveness of GIV in non-linear cases, we design 4 different non-linear treatment functions $f_X(\cdot)$ to evaluate the treatment effect estimation performance of Meta-EM algorithm. We select the SOTA IV-based method (AGMM) to evaluate GIV. We plot the estimated value of effect function with T=do(t) and sort it by Ground-Truth (GT) for different synthetic scenarios. The results (Fig. 3) show GIVs (with Meta-KM or Meta-EM) achieve SOTA performance, especially GIV$_{EM}$ achieves comparable results with TrueIV and estimated outcome curves from GIV$_{EM}$ approximate the true curve. For the detailed results of non-linear cases, see the Supplementary material.

**The Ablation Study for Reconstruction Accuracy of GIV** To demonstrate that Meta-EM can automatically find the proper group number and implement end-to-end train for IV generation, we plot MMD line for each group number in different synthetic settings (Linear-$K$-$m_X$). As shown in Fig. 4, Meta-EM always find the proper group number (red-line) automatically, but Meta-KM fails to do it. Besides, as an ablation experiment, we compare the accuracy of Meta-KM and Meta-EM for GIV reconstruction on data fusion with different group numbers. As shown in Fig. 5, Meta-EM algorithm successfully reconstructs the GIV, 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 Meta-KM is basically below 60%.

Meta-EM algorithm uses a shared representation block to learn a nonlinear representation space to EM algorithm, which relaxes the underlying linear regression assumption. To verify it, in the ablation experiments (Table 2), we compare the accuracy of EM, Meta-EM and Fun-EM, where Fun-EM implements EM algorithm with known non-linear functions $f_X(X)$ (Eq. (14)). The results show that Meta-EM improves the reconstruction accuracy by 6.3% than EM algorithm, bus still below the GT Fun-EM.

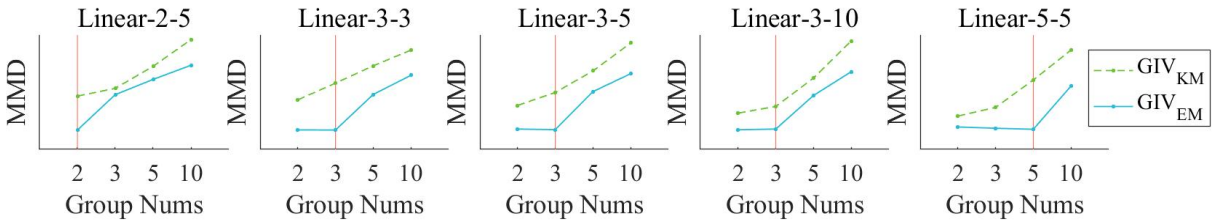


Figure 4: MMD for Selection of Group Number with Different Synthetic Setting ($Data$-$K$-$m_X$).

| | IHDP Dataset | | | | PM-CMR Dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | Poly2SLS | KernelIV | DeepIV | AGMM | Poly2SLS | KernelIV | DeepIV | AGMM |
| NoneIV | 0.24(0.13) | 0.46(0.24) | 0.58(0.24) | 0.14(0.06) | 0.18(0.04) | 0.35(0.20) | 0.41(0.16) | 0.13(0.06) |
| UAS | 0.24(0.133) | 0.46(0.24) | 0.57(0.24) | 0.14(0.06) | 0.18(0.04) | 0.35(0.20) | 0.40(0.16) | 0.13(0.06) |
| WAS | 0.24(0.13) | 0.45(0.24) | 0.57(0.23) | 0.14(0.06) | 0.18(0.04) | 0.37(0.21) | 0.42(0.16) | 0.16(0.08) |
| ModeIV | 0.24(0.13) | 0.46(0.25) | 0.57(0.24) | 0.15(0.06) | 0.18(0.04) | 0.36(0.20) | 0.41(0.15) | 0.13(0.07) |
| AutoIV | $> 100$ | 0.46(0.24) | 0.58(0.25) | 0.14(0.07) | 0.18(0.04) | 0.35(0.20) | 0.41(0.18) | 0.13(0.06) |
| $\text{GIV}_{KM}$ | 0.05(0.03) | 0.35(0.18) | 0.50(0.20) | 0.11(0.05) | 0.09(0.04) | 0.33(0.20) | 0.38(0.16) | 0.12(0.05) |
| $\text{GIV}_{EM}$ | **0.03(0.01)** | **0.20(0.17)** | **0.48(0.23)** | **0.09(0.03)** | **0.05(0.01)** | **0.31(0.21)** | **0.34(0.18)** | **0.08(0.04)** |
| TrueIV | **0.03(0.01)** | **0.15(0.06)** | **0.46(0.17)** | **0.09(0.03)** | **0.03(0.01)** | **0.14(0.07)** | **0.14(0.05)** | **0.05(0.02)** |

Table 3: The Mean Squared Error $mean(std)$ on IHDP & PM-CMR Dataset

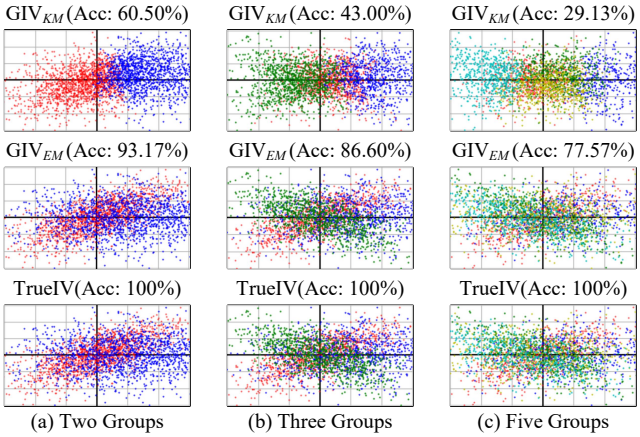


Figure 5: GIV Reconstruction Accuracy of Meta-EM.

## Experiments on Real-World Datasets

**Real-World Datasets** Similar to previous methods(Nie et al. 2020; Hartford et al. 2017; Bica, Jordon, and van der Schaar 2020; Schwab et al. 2020), we perform experiments on two real-world datasets **IHDP**[4] (Shalit, Johansson, and Sontag 2017) & **PM-CMR**[5] (Wyatt et al. 2020), as the true effect 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. (14)&(16) 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 effect function estimation. We select four SOTA IV-based methods to evaluate the performance of GIV.

**The Results of Individual Treatment Effect Estimation** By estimating the latent differentiated covariate-treatment distribution parameters across groups, Meta-EM reconstructs the latent IV and the reconstruction accuracy reaches 93.47% and 82.62% on **IHDP** and **PM-CMR**, however, K-Means is only 64.29% and 46.09%. This demonstrates Meta-EM can automatically find the optimal IV, but K-Means cannot. In Table 3, comparing the two optimal combinations (AutoIV in Poly2SLS & UAS in AGMM) in effect estimation in Table 3, Meta-EM further reduced the errors by 0.131(↓73%) and 0.043(↓33%), which well eliminated the unmeasured confounding bias. Besides, $\text{GIV}_{EM}$ 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 $\text{GIV}_{EM}$, the performance of $\text{GIV}_{KM}$ exceeds most baselines in downstream tasks, but it is still inferior to $\text{GIV}_{EM}$ and TrueIV. This means that $\text{GIV}_{EM}$ can reconstruct the latent group IV with the data distribution in the real scene and obtain asymptotically unbiased causal effect estimation.

## Conclusion

In this paper, by estimating the differentiated covariate-treatment distribution across groups, we propose a novel Meta-EM, a tool for reconstructing latent Group IVs and predicting treatment effect function from data fusion. To the best of our knowledge, using representation learning to reconstruct Group Instrumental Variables by Meta-EM algorithm in data fusion is the first work for IV generation without expert knowledge. Theoretically and empirically, we address a vital problem in causal inference: how to learn valid IVs from observational data for ITE.

[4]IHDP: https://www.fredjo.com/

[5]PM-CMR:https://pasteur.epa.gov/uploads/10.23719/1506014/SES_PM25_CMR_data.zip

## Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 62006207, U20A20387, No. 62037001), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Project by Shanghai AI Laboratory (P22KS00111), the StarryNight Science Fund of ZhejiangUniversity Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131); the Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11); Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.

## References

Angrist, J. D. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *The american economic review*, 313–336.

Angrist, J. D.; Imbens, G. W.; Rubin, D. B.; and . 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455.

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

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

Bica, I.; Jordon, J.; and van der Schaar, M. 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33: 16434–16445.

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

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

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

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

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

Finkelstein, A.; Taubman, S.; Wright, B.; Bernstein, M.; Gruber, J.; Newhouse, J. P.; Allen, H.; Baicker, K.; and Group, O. H. S. 2012. The Oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3): 1057–1106.

Harris, K.; Ngo, D. D. T.; Stapleton, L.; Heidari, H.; and Wu, S. 2022. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, 8502–8522. PMLR.

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

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

Hünermund, P.; and Bareinboim, E. 2019. Causal Inference and Data Fusion in Econometrics. arXiv:1912.09104.

Imbens, G. W.; and Angrist, J. D. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica: Journal of the Econometric Society*, 467–475.

Kuang, K.; Cui, P.; Li, B.; Jiang, M.; and Yang, S. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 265–274.

Kuang, K.; Cui, P.; Zou, H.; Li, B.; Tao, J.; Wu, F.; and Yang, S. 2020a. Data-driven variable decomposition for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*.

Kuang, K.; Li, L.; Geng, Z.; Xu, L.; Zhang, K.; Liao, B.; Huang, H.; Ding, P.; Miao, W.; and Jiang, Z. 2020b. Causal inference. *Engineering*, 6(3): 253–263.

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

Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Miao, J.; Zhang, W.; Tan, W.; Wang, J.; Wang, P.; et al. 2022. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. *ACL*.

Li, X.-H.; Cao, C. C.; Shi, Y.; Bai, W.; Gao, H.; Qiu, L.; Wang, C.; Gao, Y.; Zhang, S.; Xue, X.; et al. 2020. A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*.

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

Long, J. P.; Zhu, H.; Do, K.-A.; and Ha, M. J. 2022. The Generalized Causal Dantzig: A Unified Approach to Instruments and Environments. arXiv:2207.14753.

Muandet, K.; Mehrjou, A.; Le Kai, S.; and Raj, A. 2020. Dual Instrumental Variable Regression. In *NeurIPS 2020*.

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

Nie, L.; Ye, M.; Nicolae, D.; et al. 2020. VCNet and Functional Targeted Regularization For Learning Causal Effects

of Continuous Treatments. In *International Conference on Learning Representations*.

Pager, D.; and Karafin, D. 2009. Bayesian bigot? Statistical discrimination, stereotypes, and employer decision making. *The Annals of the American Academy of Political and Social Science*, 621(1): 70–93.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

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

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

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

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

Tian, Q.; Kuang, K.; Jiang, K.; Liu, F.; Wang, Z.; and Wu, F. 2022. ConfounderGAN: Protecting Image Data Privacy with Causal Confounder. In *Advances in Neural Information Processing Systems*.

Wu, A.; Kuang, K.; Li, B.; and Wu, F. 2022a. Instrumental Variable Regression with Confounder Balancing. In *International Conference on Machine Learning*, 24056–24075. PMLR.

Wu, A.; Yuan, J.; Kuang, K.; Li, B.; Wu, R.; Zhu, Q.; Zhuang, Y. T.; and Wu, F. 2022b. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*.

Wyatt, L. H.; Peterson, G. C. L.; Wade, T. J.; Neas, L. M.; and Rappold, A. G. 2020. 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.

Xu, L.; Chen, Y.; Srinivasan, S.; de Freitas, N.; Doucet, A.; and Gretton, A. 2021. Learning Deep Features in Instrumental Variable Regression. In *International Conference on Learning Representations*.

Yuan, J.; Wu, A.; Kuang, K.; Li, B.; Wu, R.; Wu, F.; and Lin, L. 2022. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4): 1–20.

Zhang, S.; Jiang, T.; Wang, T.; Kuang, K.; Zhao, Z.; Zhu, J.; Yu, J.; Yang, H.; and Wu, F. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4373–4382.

Zhang, S.; Yao, D.; Zhao, Z.; Chua, T.-S.; and Wu, F. 2021. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 367–377.