Estimating Treatment Effect across Heterogeneous Data Sources: An Instrumental Variable Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

To estimate treatment effect in the presence of unmeasured confounders, instrumental variable (IV) approachs have achieved promising advances, but have strict requirements on data collection. To alleviate this issue, the two-sample IV approach is proposed by fusing estimations across two complementary and homogeneous data sources. However, the homogeneous assumption, i.e., data sources share the same joint distribution, is restrictive for realistic cases. Motivated by this, this paper proposes a novel IV problem named Shifted Two-Sample IV (S2IV), which aims to estimate the treatment effect across heterogeneous data sources, i.e., the joint distributions of different data sources are skewed differently. Theoretically, we first show that solving the S2IV problem is equivalent to learning the unbiased treatment-IV relationship from the joint of data sources. To this end, we propose a **R**ecovery-**A**ided **T**ransferable **IV** (**RATIV**) framework by transferring the instruments from one data source and recovering the treatments on the other data source at the same time. Extensive experimental results on both synthetic and real-world datasets verify the effectiveness of our method.

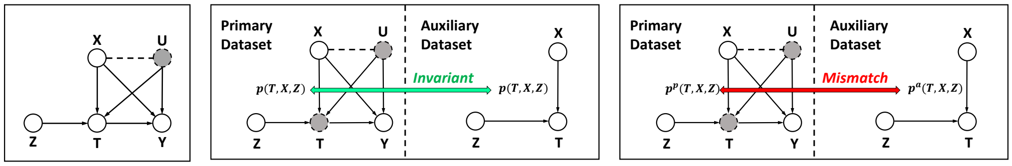
1 Introduction

The development of the instrumental variable (IV) method allows for practical treatment effect estimations in the presence of unobserved confounding [1, 2, 3]. Over the past decades, a bunch of variants has achieved remarkable progress with various linear/non-linear function approximators [2, 4, 5, 6, 7]. A typical example is contributed by [1], where the task is to study the effect of age at school entry (T) on the educational attainment (Y) with some individualized characteristics (X), where the (unobserved) social status of the born family (U) simultaneously affects T and Y (see the Figure 1(a)). To eliminate the (unobserved) confounding effect by U, the quarter of birth (Z) is treated as a valid IV, as Z is simultaneously independent of U and strongly correlated to T.

Despite the success of IV analysis, its requirement on data acquisition becomes restrictive for real-world data acquisition [1]. To be specific, it might be impossible to simultaneously observe the tuple of treatment, outcome, and IV in one sample. Recalling the education-schooling example, [1] points out that a large-scale dataset containing both age at school entry (T) and educational attainment (Y) does not exist. Alternatively, one can access two separate data sources with the day of birth Z recorded in both, while X, T and X, Y are included in only one or the other datasets. We follow [1] to call such IV estimation problem through data fusion as "Two-sample IV", as in Figure 1(b). Notably, since proposed in [1], the two-sample IV has been embraced by researchers across diverse areas, including health-care [8, 9, 10] and economic studies [11, 12].

To fuse estimations from two data sources, the core assumption of the two-sample IV problem [1, 13, 14] is the "structural homogeneity" [13], which states that the joint data distribution, namely $P(T, X, Z)$, is the **same** across data sources, as shown in Figure 1 (b). Such an assumption

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.



(a) Vanilla IV (b) Two-Sample IV (c) Ours S2IV

Figure 1: Causal structure of (a) the vanilla IV problem. (b) the two-sample IV problem. (c) our S2IV problem.Dashed Lines represent correlations between X and U.

corresponds to the case that two data sources share homogeneous data distributions. Nevertheless, real-world cases often exhibit heterogeneous structures across data sources with joint distribution shifts. For instance, one data source might be sampled from populations older than 50 from New York, while the other one might come from populations younger than 30 in Los Angeles. Consequently, the traditional two-sample IV methods fail to address such more practical but challenging cases, as mismatched distributions across data sources lead to biased IV estimations.

To overcome this gap, we investigate a novel setting named "Shifted Two-Sample IV" (S2IV) problem in this paper, as shown in Figure 1(c), which is the first work allowing for distributional shifts across data sources in IV estimation. More formally, we name the dataset with distributions $P^p(Z, X, Y)$ as the **primary** dataset and the other dataset with $P^a(Z, X, T)$ as the **auxiliary** dataset [14]. Distinct from IV and two-sample IV settings, the unique challenge of our S2IV problem is to correctly learn the primary treatment-IV relationship $P^p(T \mid Z, X)$ with mismatched distributions between primary and auxiliary data[1]. To support this motivation, we develop both lower and upper bounds based on the foundations of non-linear IV estimation theory, which demonstrates that learning $\hat{P}(T \mid Z, X) = P^p(T \mid Z, X)$ is necessary and sufficient for unbiased estimation.

By factorizing the general shift $P^a(T, Z, X) \neq P^p(T, Z, X)$ into the joint of both covariate shift $P^a(Z, X) \neq P^p(Z, X), P^a(T \mid X, Z) = P^p(T \mid X, Z)$ and concept shift $P^a(Z, X) = P^p(Z, X), P^a(T \mid X, Z) \neq P^p(T \mid X, Z)$, we observe that the proposed S2IV problem can be solved, i.e., learning $\hat{P}(T \mid Z, X) = P^p(T \mid Z, X)$, in the case that either the distributions of instrumental variables across two data sources can be **aligned**, or the treatments on the primary data can be **recovered** (see following explanations in (a) and (b)). Therefore, we propose a novel learning framework with two-sample complementarity named **R**ecovery-**A**ided **T**ransferable **IV** (**RATIV**) for S2IV problem.

To be specific, we build RATIV from two aspects: (a) When covariate shift holds, we adapt distributions of Z, X across data sources by developing a transferring framework named **T**ransferable **IV** (**TIV**) based on convention IV estimators; (b) When data exhibits concept shift, we propose to recover the primary treatments T^p[2] by designing a Conditional Bernoulli Variational encoder (CB-VAE) model. By combining solutions designed for covariate and concept shift jointly, our RATIV framework tackles the S2IV problem in the case of the general shift across data sources.

2 Preliminaries

Notations. In this paper, we aim to achieve treatment effect estimation from observational data in the presence of unmeasured confounders. As shown in Figure 1(c), we denote the binary treatment, observed covariates, outcome, IV, and the unobserved confounder as $T \in \{0, 1\}$, X, Y, Z and U, respectively. We follow the potential outcome framework and characterize the potential outcome of Y under the assignment $T = t$ as $Y(t)$. Throughout this paper, we denote the random variables by uppercase letters (e.g., T and Y) and their realizations by lowercase letters (e.g., t and y). Meanwhile, we use superscript, i.e., p (primary) or a (auxiliary), to denote which data source the variable/sample belongs to, and subscript as the sample index (e.g., t_i^p is the i-th sample of primary data). The distribution is denoted as P with the corresponding density function denoted as p.

[1]Two-sample IV assumes that the $P(T, Z, X)$ remains the same across data sources.

[2]We use superscript to denote which data source the variable belongs to.

2.1 The Vanilla IV problem

Data Generation Process. Following the widely adopted separable (additive outcome) assumption, we assume that the structural function of the outcome admits the following expression [15, 2]:

$$Y = h(T, X) + U, \tag{1}$$

where h is the **target** function we aim to recover.

Valid Instruments. An instrumental variable Z is valid if and only if it satisfies the following three principles [15, 2]: (1) **Relevance**: Z is correlated to T, e.g., $T \not\perp\!\!\!\perp Z \mid \mathbf{X}$; (2) **Exclusion**: Z affects Y only through T; (3) **Unconfounded**: $Z \perp\!\!\!\perp U \mid Z$. We also adopt another common assumption to remove the confounding effect: $\mathbb{E}[U|Z, X] = 0$ [2].

Empirical Observations. The vanilla IV problem assumes that one can simultaneously observe (Z, X, T, Y).

Representative Nonlinear Estimators. The representative nonlinear estimators include DeepIV [2] and KIV [5] (see Appendix C for details).

Specific Property of vanilla IV problem. Vanilla IV problem is robust against the misspecification of the treatment-IV relationship estimated [13]. For example, when the outcome structural equation is linear (e.g., h degenerates to the linear coefficient β) without covariates ($Y = \beta T + U$), then the following estimation is unbiased for any function f[3]:

$$\mathbb{E}[f(\mathbf{Z})(Y - \beta T)] = 0 \Rightarrow \hat{\beta} = \frac{\sum_i y_i f(z_i)}{\sum_i t_i f(z_i)},$$

where the asymptotic variance of the estimation is minimized when f identifies the true treatment-IV relationship.

2.2 Two-sample IV Problem

Empirical Observations. Distinct from the vanilla IV problem, two-sample IV and our S2IV assume that the empirical data consists of two separate sources:**(a). The primary dataset** includes the instruments, the outcomes, and the covariates: $\mathcal{D}^p = \{Z_i, Y_i, X_i\}_{i=1}^m$. Meanwhile, Z is required to be a valid instrument in $\mathcal{D}^p$ [14]. **(b). The auxiliary dataset** encodes the instruments, the treatments, and the covariates: $\mathcal{D}^a = \{Z_i, T_i, X_i\}_{i=1}^n$.

Homogeneous Populations. To fuse estimations from two sources, a core assumption for the two-sample IV problem is that the $\mathcal{D}^p$ and $\mathcal{D}^a$ should be sampled from the same (homogeneous) population [1, 13, 14]: $P^a(X, Z, T) = P^p(X, Z, T)$[4].

3 Problem Definition and Motivating Analysis

3.1 Our Problem: Shifted Two-sample IV

Heterogeneous Populations. However, the principle of homogeneous populations introduced above is unrealistic in real-world scenarios. Sampling $\mathcal{D}^a$ and $\mathcal{D}^p$ from different locations or times easily tends to cause distributional shifts between $P^a(X, Z, T)$ and $P^p(X, Z, T)$. Hence, we relax the principle of two-sample IV and propose the Shifted Two-sample IV (S2IV) problem with mismatched joint distributions across datasets: $P^p(T, Z, X) \neq P^a(T, Z, X)$.

Impact of biased treatment-IV relationship. We note that the specific property of vanilla IV does not hold for either two-sample IV or our S2IV. A direct consequence is that the biased estimation of the treatment-IV relationship will further bias the total estimation:

Example 1. *Suppose the binary treatment with $\mathcal{T} = \{0, 1\}$ and the covariate shift from $\mathcal{D}^a$ to $\mathcal{D}^p$: $P^a(T \mid Z, X) = P^p(T \mid Z, X)$ while $P^a(Z, X) \neq P^p(Z, X)$. As $P^a(Z, X) \neq P^p(Z, X)$, the $\hat{P}^a(T \mid Z, X)$ learned by DeepIV from $\mathcal{D}^a$ is biased with the underlying $P^p(T \mid Z, X)$ on the*

[3]The GMM methods [7] is based on this formulation.

[4]Although previous work only assumes that $P(T \mid X, Z)$ is learnable and invariant, this implies that $P^a(X, Z) = P^p(X, Z)$ based on covariate shift theory.

primary dataset. Thus, when DeepIV plugs the biased $\hat{P}^a(T \mid Z, X)$ into the second stage on $\mathcal{D}^p$, the solution of the integral equation $\mathbb{E}[Y|Z,X] = \int \hat{h}(T,X)d\hat{P}(T \mid Z,X)$ will be biased.

Target of S2IV. To solve such negative impact brought by heterogeneous populations of S2IV, we first claim that **learning $P^p(T \mid Z, X)$ from $\mathcal{D}^a \cup \mathcal{D}^p$ is necessary and sufficient to solve our S2IV problem.** To support such claim, we present theoretical analysis based on the theory of non-linear IV estimations [5, 16, 17].

3.2 Deriving Bounds for Motivation

We first introduce some basic notations and definitions of non-linear IV estimations[5]. We use $\hat{h}$ to denote the estimation of h. Meanwhile, we introduce the excess risk of h and $\hat{h}$ as $\mathcal{E}(\hat{h}) := \mathbb{E}_{P^p(Y,X,Z)}\|Y - \hat{h}(\mu(Z,X))\|^2_{\mathcal{Y}}$ and $\mathcal{E}(h) := \mathbb{E}_{P^p(Y,X,Z)}\|Y - h(\mu(Z,X))\|^2_{\mathcal{Y}}$, where $\mu(Z,X)$ represents the embedding of $P(T \mid Z,X)$ in the kernel space. Intuitively, $\mathcal{E}(\hat{h})$ and $\mathcal{E}(h)$ represents expected error of h and $\hat{h}$ compared with ground truth on the primary data. In addition, all the norm w.r.t. functions, e.g., $\|\hat{h} - h\|$, is defined as the operator norm. We then show the necessity of our claim, i.e., by deriving a lower bound on the performance of non-linear IV estimation on the S2IV problem:

Theorem 1. *The error of $\hat{H}$ from h is lower bounded by divergence between $\hat{P}(T \mid Z,X)$ and $P^p(T \mid Z,X)$:*

$$\|\hat{h} - h\| \geq \frac{C}{K}\textit{CMMD}\left(\hat{P}(T \mid Z,X), P^p(T \mid Z,X)\right), \tag{2}$$

where C, K are constants, and the term CMMD is the conditional MMD divergence [18] between the estimated $\hat{P}(T \mid Z,X)$ and primary treatment-IV distribution $P^p(T \mid Z,X)$.

Remark The above theorem shows that the divergence (CMMD) (see Appendix 2 for details) between learned $\hat{P}(T \mid Z,X)$ and $P^p(T \mid Z,X)$ definitely induces estimation error in the right side of Eq. (2). In other words, it indicates the necessity of learning correct treatment-IV relationship. On the other hand, to show the sufficiency, we first derive a population-level upper bound as follows:

Afterwards, we present the last upper bound to show that estimation of $\hat{P}(T \mid Z,X) = P^p(T \mid Z,X)$ is sufficient to identify the underlying h:

Theorem 2. *The following inequality holds w.r.t to $\mathcal{E}(\hat{h})$ and $\mathcal{E}(h)$:*

$$\mathcal{E}(\hat{h}) \leq \mathcal{E}(h) + \kappa^2\hat{K}^2\textit{CMMD}(\hat{P}(T \mid Z,X), P^p(T \mid Z,X)),$$

where κ and $\hat{K}$ are constants.

The above theorem immediately leads to following result.

Corollary 1. *If CMMD$(\hat{P}(T \mid Z,X), P^p(T \mid Z,X)) = 0$, then $\hat{h} = h$.*

Remark. The Upper bound in Theorem 2 indicates that learning correct treatment-IV relationship such that $\hat{P}(T \mid Z,X) = P^p(T \mid Z,X)$ is also sufficient for unbiased IV estimation (see Appendix D.1 for proofs).

4 Learning Treatment Effects with Shifted Two-Sample Complementarity

To learn $P^p(T \mid Z,X)$ from $\mathcal{D}^a \cup \mathcal{D}^p$, we build a unified learning framework in this section.

4.1 Aligning Instruments across Data Sources

We first consider the case that the covariate shift holds such that $P^p(T \mid Z,X)$ is learnable from $\mathcal{D}_a$ by aligning $P^p(Z,X)$ with $P^a(Z,X)$. Inspired by a domain adaptation literature [19], we propose to migrate the distributional shift between $P^a(Z,X)$ and $P^p(Z,X)$ such that the $\hat{P}(T \mid Z,X)$ learned

[5] In this paper, we characterize the non-linearity using kernel tricks [5].

on $\mathcal{D}^a$ correctly estimates $P^p(T \mid Z, X)$. To this end, we propose a joint **Transferable IV (TIV)** framework by mapping the instruments from $\mathcal{D}^a$ to the $\mathcal{D}^p$ based on the optimal transport (OT) [19, 20]. We choose the OT-based adaptation based on two **advantages**: (a) it supports measuring distributional divergence in both kernel [21] and Euclidean feature space [19]; (b) it is compatible with both categorical and continuous outcomes [20]. Suppose f is the learning model for $P(T \mid Z, X)$[6], the first stage of our TIV framework follows the objective:

$$\min_f \mathcal{W}_p(P^a(Z, X, T), P^p(Z, X, f(Z, X))) + \lambda\Omega(f),$$

where $f(Z, X)$ is the proxy of underlying T^p, $\mathcal{W}_p$ refers to the p-order Wasserstein distance, and Ω is the regularization term. Following protocols in OT-based transferring frameworks [20, 19], we let $p = 1$ and the objective reduces to the inner product between the Kantorovitch's coupling matrix $\gamma \in \mathcal{R}^{m \times n}$ and the cost matrix $C \in \mathcal{R}^{m \times n}$ ($C_{ij} = d(x_i^a, x_j^p) + d(z_i^a, z_j^p) + \mathcal{L}(t_i^a, f(z_j^p, x_j^p))$): $\min_{f, \gamma \in \Delta} \mathrm{Tr}(\gamma^{\mathrm{T}} C) + \lambda\Omega(f)$, where Δ is the transportation polytope, d is the distance metric (e.g., Euclidean distance) and $\mathcal{L}$ is the loss function (e.g., squared loss for KIV). Moreover, the joint optimization on f and γ can be decomposed into alternative optimization as:

$$\begin{cases} \min_f \sum_{i,j} \gamma_{i,j} \mathcal{L}(t_i^a, f(z_j^p, x_j^p)) + \lambda\Omega(f), \\ \min_\gamma \sum_{i,j} \gamma_{i,j} C_{i,j}. \end{cases} \tag{3}$$

We then instantiate our proposed TIV framework with two representative non-linear estimators:

T-KIV. Based on the TIV framework in Eq. (3), we derive the following closed solution and propose the corresponding Transferable KIV (T-KIV) algorithm.

Proposition 1. *Let K_{TT}^a be the kernel matrix of treatments on the auxiliary data. Meanwhile, $K_{T_a, \dot{t}}$ represents the kernel vector between T_a and the testing $\dot{t}$. Suppose the coupling matrix computed from the first stage is γ_f, then the solution of T-KIV is:*

$$\begin{aligned} &K_R^p = K_{ZZ}^p \odot K_{XX}^p, \\ &W = (m^2 \gamma_f^T K_{TT}^a \gamma_f) \left(K_R^p + m\lambda I\right)^{-1} K_R^p, \\ &\alpha = (WW^T + \xi m^3 \gamma_f^T K_{tt}^a \gamma_f)^{-1} W Y_p, \\ &\hat{h}(\dot{t}, \dot{x}) = m \gamma_f^T K_{tt}^a K_{T_a, \dot{t}} \odot K_{X_p, \dot{x}}, \end{aligned} \tag{4}$$

where $\odot$ means element-wise multiplication.

T-DeepIV. Similarly, we propose the Transferable DeepIV (T-DeepIV) by (a) specializing $\mathcal{L}$ in Eq. (3) to be the squared loss; (b) computing the coupling matrix γ and the cost matrix C on the mini-batch (which is a standard OT problem and can be via network simplex algorithm [21]). When the first stage finishes, the second stage of the T-DeepIV remains the same as in [2], which is implemented with an outcome regression network. Details on the algorithm of T-KIV and T-DeepIV are present in Appendix E.1 and E.2.

4.2 Recovering Treatments via Generative Models

However, the solutions remain still unclear when the covariate shift principle is violated: $P^p(T \mid Z, X) \neq P^a(T \mid Z, X)$ and $P^a(Z, X) \neq P^p(Z, X)$. In general, this problem is ill-posed based on the transfer learning theory [22], due to the arbitrary shift on $P(T \mid Z, X)$ and missing T^p (Consider Z, X and T as features and labels in unsupervised domain adaptation, where auxiliary and primary data are the source and target domains).

Fortunately, as we restrict on binary primary treatments, recent advances in unsupervised representation learning bring us the opportunity to recover underlying T^p based on $\mathcal{D}^p$, and further identify $P^p(T \mid Z, X)$. To be specific, two facts connect our problem and iVAE: (a) The separable Assumption 1 corresponds to the additive noise in [23]; (b) Z^p, X^p play as a similar but weaker version of the auxiliary variables in iVAE [23, 24]. However, two obstacles prevent us from directly adopting the iVAE to our S2IV problem.

- The noise term in S2IV, e.g., the confounder U, is not exogenous such that one cannot derive noise-free identifications [23].

[6] f could be either specialized as the treatment network in DeepIV or the ridge kernel regression in KIV

Algorithm 1 Training framework of RATIV

1: **Input:** The primary and auxiliary datasets $\mathcal{D}^p = \{z_i^p, y_i^p, x_i^p\}_{i=1}^m$ and $\mathcal{D}^a = \{z_i^a, t_i^a, x_i^a\}_{i=1}^n$, the hyper-parameter β, Maximum number of iterations $\mathcal{I}$.
2: **Recovery procedure**:
3: Train CBVAE model $\{q_\psi, p_\psi\}$ by optimizing Eq. (5).
4: Output the recovered primary treatments as $\tilde{T}^p$.
5: Train T-DeepIV with regularized objective in Eq. (6).

- In S2IV, Z^p, X^p is weaker than auxiliary variables in [23] as $Y^p \not\perp\!\!\!\perp (Z^p, X^p) \mid T^p$.

Therefore, we re-design a conditional Bernoulli VAE (CBVAE) model on $\mathcal{D}^p$ to recover T^p. Letting q be the posterior distribution modeled by CBVAE, we minimize the evidence lower bound (ELBO) of $P(Y^p \mid X^p, Z^p)$ as follows:

$$\mathbb{E}_{\mathcal{D}^p}[\mathbb{E}_{q_\psi(T^p|X^p,Z^p)} \log p_\psi(Y^p \mid X^p, T^p, Z^p) - \mathrm{KL}\left(q_\psi(T^p \mid X^p, Z^p) \| p(T^p \mid X^p, Z^p)\right)], \quad (5)$$

where q_ψ and p_ψ refers to the posterior and likelihood that are parameterized by ψ, KL is the Kullback-Leibler divergence, and we follow [23] to model $p_\psi(Y^p \mid X^p, T^p, Z^p)$ as a Gaussian distribution.

Notably, as the T^p is a binary variable, it is reasonable to model the posterior $q_\psi(T^p \mid X^p, Z^p)$ and the prior $p(T^p \mid X^p, Z^p)$ as the Bernoulli distribution: $q_\psi(T^p \mid X^p, Z^p) = \prod_{j=1}^m \mathcal{B}(t_j^p \mid \theta_j(x_j^p, z_j^p))$ and $p(T^p \mid X^p, Z^p) = \prod_{j=1}^m \mathcal{B}(t_j^p \mid \rho)$, where θ is predicted by the encoder q_ψ and ρ is a fixed prior parameter. Meanwhile, we follow the concrete reparameterization trick in [25] and re-sample the latent T^p as $T_i^p = \sigma\left(\ln \epsilon - \ln(1-\epsilon) + \ln \theta_i(y) - \ln\left(1 - \theta_i(x_i^p, z_i^p)\right)\right)$, where $\epsilon \sim \mathcal{U}(0,1)$ following the uniform distribution and σ is the sigmoid function. We present the following theorem to state the reliability of our CBVAE model (see Appendix E.3.2 for algorithmic details with proofs).

Theorem 3. *Let $R = (Z, X)$ be the joint of instruments and covariates. Assume: (a) The binary treatments T^p are conditionally exponential of given instruments and covariates R^p with differentiable parameters and normalizing factors; (b) The effect function h is injective; (c) There exists some realizations $\{r_l^p\}_{l=0}^K$ ($r_l^p = (x_l^p, z_l^p)$) such that the parameter-difference matrix L is invertible. Then the conditional density estimated by CBVAE, $\tilde{T}^p$, identifies the true T^p up to a linear transformation, where $\tilde{T}^p$ is the recovered treatments. With specific constraints on $\{r_l^p\}_{l=0}^K$ and the parameter space of θ, $\tilde{T}^p$ exactly identifies T^p.*

4.3 RATIV: Treatment Effect Estimator with Two-sample Complementarity

Combined with the T-DeepIV baseline and the CBVAE model, we obtain the Recovery-aided Transferable IV (RATIV) estimator to solve the S2IV problem. RATIV achieves accurate estimation in the sense that it performs well if at least one of the two data sources is reliable[7]. To be specific, we regularize the first stage of the T-DeepIV baseline[8] based on the recovered primary treatments:

$$\mathcal{L}_{\mathrm{T1-DeepIV}} + \lambda \mathcal{L}(\hat{t}^p, \tilde{t}^p), \quad (6)$$

where $\mathcal{L}_{\mathrm{T1-DeepIV}}$ refers to the first-stage objective of T-DeepIV baseline (see Appendix 3 for details), $\mathcal{L}$ is the BCE loss, λ is the hyper-parameter, $\hat{t}_j^k$ and $\tilde{t}_j^p$ refers to primary treatments predicted by T-DeePIV and the recovered by CBVAE, respectively. We note that the two-sample complementarity property of RATIV stems from the fact that RATIV achieves accurate estimation in the case that either distributions of instruments across data sources follow the covariate shift principle or the primary treatments can be recovered.

5 Experiment Results

5.1 Baselines and Metric

Baselines. We compare our T-KIV, T-DeepIV, and RATIV methods with a bunch of two-stage IV baselines: (1) the DeepIV method [2]; (2) Ploy-2SLS (P-2SLS) method [1, 13]; (3) KernelIV (KIV) [5]; (4)

[7] Here we use useful to mean that the data source suffices to learn $P(T^p \mid X^p, Z^p)$.

[8] We choose the T-DeepIV baseline to build our RATIV due to its flexibility.

DualIV [26]. Due to the missing data in the S2IV problem, some one-stage IV baselines are not implementable [27]. A bunch of semi-parametric two-sample IV baselines [14, 28] cannot be applied to our S2IV problem, as they require binary instruments.

Metrics. We evaluate our model using in-sample performance and out-of-sample performance, respectively. In-sample results estimate treatment effects for units where the factual outcome is observed, and out-of-sample results estimate on units with no observed outcomes. For synthetic data, we evaluate each method by measuring its capability of recovering the structural function h by the mean squared error (MSE): MSE $= \sum_i (h(t_i, x_i) - \hat{h}(t_i, x_i))^2$, where t_i, x_i are testing data. For real-world data, as the underlying h is inaccessible, we evaluate the estimation error of ATE as $\epsilon_{\text{ATE}} = |\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^1 - \hat{y}_i^0) - \frac{1}{n}\sum_{i=1}^{n}(m_i^1 - m_i^0)|$, where $\hat{y}_i^1, \hat{y}_i^0$ are estimated outcomes, and m_i^1, m_i^0 are noiseless responses of sample i [29, 30]. We also evaluate the error of Conditional Average Treatment Effect (CATE) by measuring the Precision in Estimation of Heterogeneous Effect (PEHE) error [30] as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}((\hat{y}_i^1 - \hat{y}_i^0) - (m_i^1 - m_i^0))^2$ (see Appendix F for experiment details).

5.2 Synthetic Experiments

Data Generation. We simulate two settings of our S2IV problem, where the first setting follows the covariate shift and the second one has the general shift. For each simulation setting, we fix its generation and vary the true response function h between the following cases: (a) $h(t) = \sin(t)$; (b) $h(t) = \mathbf{1}(t \geq 0)$; (c) $h(t) = |t|$; (d) $h(t) = t^2 + t$. We set the size of training samples as $m = n = 2000$ for primary and auxiliary data sources, and report the MSE error on 2000 testing samples. To be specific, we simulate our S2IV problem with both covariate and general shifts across data sources (see Appendix F.2.1 for detailed protocols).

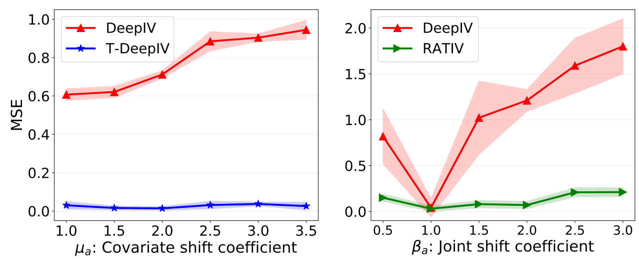


Figure 2: Left: The influence of covariate distributional shift on MSE of IV estimation, where the $[\mu_a]$ in X-axis refers to the covariate shift setup with $Z_a \sim \mathcal{N}(-\mu_a, 0.25) \cup \mathcal{N}(\mu_a, 0.25)$. Right: The influence of joint distributional shift on MSE of IV estimation, where the $[\beta_a]$ in the X-axis refers to the general shift setup with $T^a \sim \mathcal{B}(\sigma(\beta_a Z^a + U^a + 0.1\eta^a))$ generated by varying coefficient of Z_a, i.e., β_a, while the generation of Z_p keep invariant. The shade region presents the interval [mean-std,mean+std] of MSE under 10 repeated experiments.

Results. Corresponding results on structural function recovery in Table **??** verify the effectiveness of our proposed method in a synthetic setting, which also matches the upper bound of our motivating analysis. In addition, to strengthen both our motivations and the effectiveness of our methods, we investigate the influence of the distributional shift on the performance of conventional IV estimators (see details in the Appendix). It is unsurprising to see a drop of DeepIV on the right side of Figure 2, as the treatment assignments across data sources coincide and the distributional shift vanishes.

6 Conclusion, Limitations, and Future Work

Conclusion. This paper contributes the Shifted Two-sample IV (S2IV) problem with tight bounds for motivation. By transferring instruments and recovering treatments, we design RATIV as a distributionally robust with a two-sample complementarity framework. Extensive experiments show the effectiveness of the proposed RATIV.

Limitations. However, there are still weak points remaining for further efforts: (a) regarding the joint optimization of our transfer and recovery modules. Although our experimental results verify the effectiveness of our RATIV model, a deeper insight from the theoretical perspective is considered in future work. (b) The error analysis on other IV branches, such as control functions [31]. As the intrinsic logic of two-stage methods differs from the control functions, it requires additional efforts.

References

[1] Joshua D Angrist and Alan B Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418):328–336, 1992.

[2] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

[3] Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, pages 24056–24075. PMLR, 2022.

[4] Joshua D Angrist and Alan B Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

[5] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Pengzhou Wu and Kenji Fukumizu. Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*, pages 1157–1167. PMLR, 2020.

[7] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32:3564–3574, 2019.

[8] Stephen Burgess, Neil M Davies, and Simon G Thompson. Bias due to participant overlap in two-sample mendelian randomization. *Genetic epidemiology*, 40(7):597–608, 2016.

[9] Fernando Pires Hartwig, Neil Martin Davies, Gibran Hemani, and George Davey Smith. Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique, 2016.

[10] Debbie A Lawlor. Commentary: Two-sample mendelian randomization: opportunities and challenges. *International journal of epidemiology*, 45(3):908–915, 2016.

[11] Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

[12] Thomas S Dee and William N Evans. Teen drinking and educational attainment: evidence from two-sample instrumental variables estimates. *Journal of Labor Economics*, 21(1):178–209, 2003.

[13] Qingyuan Zhao, Jingshu Wang, Wes Spiller, Jack Bowden, and Dylan S Small. Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34(2):317–333, 2019.

[14] BaoLuo Sun and Wang Miao. On semiparametric instrumental variable estimation of average treatment effects through data fusion. *Statistica Sinica*, 32:569–590, 2022.

[15] Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

[16] Yuta Kawakami, Manabu Kuroki, and Jin Tian. Instrumental variable estimation of average partial causal effects. In *International Conference on Machine Learning*, pages 16097–16130. PMLR, 2023.

[17] Diego Martinez Taboada, Aaditya Ramdas, and Edward Kennedy. An efficient doubly-robust test for the kernel treatment effect. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Yong Ren, Jun Zhu, Jialian Li, and Yucen Luo. Conditional generative moment-matching networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[19] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 447–463, 2018.

[20] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

[21] Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1741–1754, 2019.

[22] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

[23] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[24] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[25] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2016.

[26] Krikamol Muandet, Arash Mehrjou, Si Le Kai, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 2020.

[27] Adi Lin, Jie Lu, Junyu Xuan, Fujin Zhu, and Guangquan Zhang. One-stage deep instrumental variable method for causal inference from observational data. In *International Conference on Data Mining*, pages 419–428. IEEE, 2019.

[28] Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34(2):288–301, 2016.

[29] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[30] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[31] Aahlad Puli and Rajesh Ranganath. General control functions for causal effect estimation from instrumental variables. *Advances in Neural Information Processing Systems*, 33:8440, 2020.

[32] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[33] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2020.

[34] Joshua D Angrist and Alan B Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

[35] Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110, 1974.

[36] Manuel Arellano and Costas Meghir. Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. *The Review of Economic Studies*, 59(3):537–559, 1992.

[37] Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

[38] Debbie A Lawlor. Commentary: Two-sample mendelian randomization: opportunities and challenges. *International Journal of Epidemiology*, 45(3):908, 2016.

[39] Joshua D Angrist and Alan B Krueger. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235, 1995.

[40] Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

[41] Pengzhou Abel Wu and Kenji Fukumizu. β-Intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2021.

[42] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029. PMLR, 2016.

[43] Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *Advances in Neural Information Processing Systems*, 35:37184–37198, 2022.

[44] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022.

[45] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[46] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

[47] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, pages 1823–1830. PMLR, 2012.

[48] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

[49] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.