A The Convergence Analysis of Meta-EM

Theorem A.1 ***Distribution Learning Step.*** *Meta-EM learns a non-linear representation to construct Linear Mixed Models for the assigned treatment variable. Based on the representation, we model the likelihood function for observational data $c_i = \{t_i, r_i\}, i = 1, 2, \cdots, n$ as $Pr(C_{TR} \mid \theta = \{\pi, \mu, \Sigma\})$. We denote parameter sequence estimated by EM algorithm as $\theta^{[h]}(h = 1, 2, \cdots)$, and denote the corresponding likelihood function sequence as $Pr(C_{TR} \mid \theta^{[h]})(h = 1, 2, \cdots)$. Then $Pr(C_{TR} \mid \theta^{[h]})$ is a monotonic sequence which constantly increase:*

$$Pr(C_{TR} \mid \theta^{[h+1]}) \geq Pr(C_{TR} \mid \theta^{[h]}). \tag{20}$$

Proof The observation:

$$Pr(C_{TR} \mid \theta) = \tfrac{Pr(C_{TR}, Z \mid \theta)}{Pr(Z \mid C_{TR}, \theta)},$$

$$\log Pr(C_{TR} \mid \theta) = \log Pr(C_{TR}, Z \mid \theta) - \log Pr(Z \mid C_{TR}, \theta).$$

Take the expectation of the log likelihood function $\mathcal{Q}(\theta, \theta^{[h]})$:

$$\mathcal{Q}(\theta, \theta^{[h]}) = \Sigma_Z \log Pr(C_{TR}, Z \mid \theta) Pr(Z \mid C_{TR}, \theta^{[h]}), \tag{21}$$

Let

$$\mathcal{H}(\theta, \theta^{[h]}) = \Sigma_Z \log Pr(Z \mid C_{TR}, \theta) Pr(Z \mid C_{TR}, \theta^{[h]}), \tag{22}$$

Then,

$$\begin{aligned} &\log Pr(C_{TR} \mid \theta^{[h+1]}) - \log Pr(C_{TR} \mid \theta^{[h]}) \\ &= [\mathcal{Q}(\theta^{[h+1]}, \theta^{[h]}) - \mathcal{H}(\theta^{[h+1]}, \theta^{[h]})] - [\mathcal{Q}(\theta^{[h]}, \theta^{[h]}) - \mathcal{H}(\theta^{[h]}, \theta^{[h]})] \\ &= [\mathcal{Q}(\theta^{[h+1]}, \theta^{[h]}) - \mathcal{Q}(\theta^{[h]}, \theta^{[h]})] - [\mathcal{H}(\theta^{[h+1]}, \theta^{[h]}) - \mathcal{H}(\theta^{[h]}, \theta^{[h]})] \end{aligned}$$

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

$$\theta^{[h+1]} = \text{argmax}_\theta \mathcal{Q}(\theta, \theta^{[h]}). \tag{23}$$

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

$$\begin{aligned} & \mathcal{H}(\theta^{[h+1]}, \theta^{[h]}) - \mathcal{H}(\theta^{[h]}, \theta^{[h]}) \\ = \;& \Sigma_Z \left(\log \frac{Pr(Z \mid C_{TR}, \theta^{[h+1]})}{Pr(Z \mid C_{TR}, \theta^{[h]})} \right) Pr(Z \mid C_{TR}, \theta^{[h]}) \\ \leq \;& \log \left(\Sigma_Z \frac{Pr(Z \mid C_{TR}, \theta^{[h+1]})}{Pr(Z \mid C_{TR}, \theta^{[h]})} Pr(Z \mid C_{TR}, \theta^{[h]}) \right) \\ = \;& \log \left(\Sigma_Z Pr(Z \mid C_{TR}, \theta^{[h+1]}) \right) = 0. \end{aligned} \tag{24}$$

In conclusion, we obtain $\log Pr(C_{TR} \mid \theta^{[h+1]}) - \log Pr(C_{TR} \mid \theta^{[h]}) \geq 0$, that means $Pr(C_{TR} \mid \theta^{[h]})$ is a monotonic sequence which constantly increase:

$$Pr(C_{TR} \mid \theta^{[h+1]}) \geq Pr(C_{TR} \mid \theta^{[h]}). \tag{25}$$

If there is an upper bound for $Pr(C_{TR} \mid \theta)$, the sequence $\log Pr(C_{TR} \mid \theta^{[h]})(i = 1, 2, \cdots)$ would converge to a specific value L^*.

Theorem A.2 ***Representation Learning Step.*** *In s-th iteration, Meta-EM learns a latent source variable as a group instrumental variable (GIV) $Z^{(s)}$ indicating multiple treatment assignment mechanisms. Based on GIV $Z^{(s)}$, We use multiple linear functions indicated by $Z^{(s)}$ to model Linear Mixed Models explicitly to optimize the shared representation $R^{(s)}$. We denote GIV sequence and representation sequence as $Z^{(s)}$ and $R^{(s)}(s = 1, 2, \cdots)$, separately, and let $C_{TR}^{(s)}$ as the concatenation of T and $R^{(s)}$. The corresponding likelihood function sequence as $Pr(Z^{(s+1)} \mid C_{TR}^{(s)})$. Then, $Pr(Z^{(s+1)} \mid C_{TR}^{(s)})$ is a monotonic sequence which constantly increase:*

$$Pr(Z \mid C_{TR}^{(s+1)}) \geq Pr(Z \mid C_{TR}^{(s)}). \tag{26}$$

Proof In the representation learning step of s-th iteration, the Meta-EM algorithm uses a shared representation block to learn a non-linear representation $R = f_R(X), R \in \mathcal{R}^{m_R}$ to regress treatment using Linear Mixed Models indicated by GIV Z. By minimizing the regression error, the representation module will capture the non-linear terms of the raw data on the treatment variables. As missing domain labels, the GIV Z will bring additional information for representation learning. Therefore, the higher the reconstruction accuracy of the instrumental variable Z, the more accurate the non-linear terms from representation learning will be. Further, the performance of the EM algorithm will also be improved by using these non-linear representations to construct linear mixed models of the treatment variables, i.e., $Pr(Z \mid C_{TR}^{(s+1)}) \geq Pr(Z \mid C_{TR}^{(s)})$. It is a Mutual Reinforcement Learning process: learn representation to learn IV, and then learn IV to learn representation at each iteration.

Besides, to ensure that the representation learning does not lose information from the raw data, we also construct a covariate reconstruction loss function and minimize the term. The objective is to minimize $\mathcal{L} = \sum_i^n \left(f_T(z_i^{(s)}, r_i^{(s)}) - t_i \right)^2 + \lambda \sum_i^n \left(f_X(r_i^{(s)}) - x_i \right)^2$.

Remark Theoretically, the Meta-EM algorithm is effective when identifiable differences in treatment assignment mechanisms across groups exist. Overall, the reconstruction accuracy has reached 77% in Section 5.2, and we can estimate the treatment effect function accurately.

B Proof of Theorem

B.1 Asymptotic Results of GIV

In Meta-EM, $\hat{\gamma}_{ik}$ converges to $1_{z_i=k}$ with the rate $o(exp(-(m_R + M)))$, where m_R is the dimension of the representations. This rate comes from bounding the probability of being in the wrong group, which can be shown by using that the density function of each group in EM algorithm follows a normal distribution, and the tail of the normal distribution is exponentially bounded and linear in m_R. Note that this result does not directly involve the sample size n. This is because the estimation error of the representations is not the leading term in the estimation error of $\hat{\gamma}_{ik}$. However, we note that as n gets larger, the representations can be learned more precisely, which may have small effects on $\hat{\gamma}_{ik}$. This implies that the Meta-EM algorithm can asymptotically recover the true group label for each unit i in large samples.

Theorem B.1 ***Asymptotic Results of GIV.***
Suppose each coordinate in the coefficient vector α_k in Eq. (3) *is nonzero for all k. As $(m_R, n) \to \infty$, for each k:*
(1) $\hat{\gamma}_{ik} \xrightarrow{p} 1_{z_i=k}$,
(2) $\hat{z}_i \sim$ Disc$(\hat{\gamma}_{i1}, \hat{\gamma}_{i2}, \cdots, \hat{\gamma}_{iK})$ is an asymptotic IV, i.e., $Pr(\hat{Z}, X, T, Y) \xrightarrow{p} Pr(Z, X, T, Y)$.
where, $\xrightarrow{p}$ denotes convergence in probability.

Proof **(1)** Consider the following representation model (Eq. (3)):

$$t_i = \alpha_k' r_i + \epsilon_i, \text{ if } z_i = k,$$

where α_k is a m_R dimensional vector of coefficients for source k, and each coordinate in the coefficient vector α_k is nonzero for all k. r_i is a m_R dimensional vector of representation for unit x_i, z_t is the source label indicating which unit x_i belongs to, and e_i is the error term allowed to have cross-sectional and heteroskedasticity. $\{x_i, t_i\}$ is observable and $\{\alpha_k, r_i, z_i, \epsilon_i\}$ are unobservable.

Let c_i as the concatenation of t_i and r_i, i.e., $c_i = (t_i, r_i)$:

$$c_i = (\alpha_k, I)' r_i + \epsilon_i, \text{ if } z_i = k,$$

where $I = \{e_j\}_{j=1}^{m_R}$ is an identity matrix of size m_R, and $e_j = (0, \ldots, 0, \underset{j-th}{1}, 0, \ldots, 0)'$:

$$c_i = A_k r_i + \epsilon_i = (\alpha_k, e_1, e_2, \cdots, e_{m_R})' r_i + \epsilon_i, \text{ if } z_i = k.$$

Besides,

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

From Eq. (9), we have:

$$\hat{\gamma}_{ik} = \frac{\pi_k Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j Pr(c_i \mid \hat{\mu}_j, \text{and} \hat{\Sigma}_j)}, \sum_{k=1}^{K} \hat{\gamma}_{ik} = 1.$$

From the Convergence Analysis of Meta-EM:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} c_i}{\sum_{i=1}^{n} \hat{\gamma}_{ik}}, \hat{\Sigma}_k = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} \left[c_i - \hat{\mu}_k\right]^2}{\sum_{i=1}^{n} \hat{\gamma}_{ik}}$$

If $n \to \infty$, then the number of samples for each group $n_k = n \times Pr(Z = k) \to \infty$ and we have $\hat{\mu}_k \xrightarrow{p} \mu_k$ and $\hat{\Sigma}_k \xrightarrow{p} \Sigma_k$ for each $k = 1, 2, \cdots, K$.

If the dimension of representations is sufficiently large and the distribution of representations for different groups can be separated with high probability, i.e., $Pr(z_i = k) \xrightarrow{p} 1_{z_i=k}$, then for each unit $z_i = k$ as $(m_R, n) \to \infty$:

$$\begin{aligned} |\hat{\gamma}_{ik} - 1_{z_i=k}| &= \frac{\sum_{j \neq k} \pi_j Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}{\sum_{j=1}^{K} \pi_j Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \\ &\leq \sum_{j \neq k} \frac{\pi_j}{\pi_k} \exp\{\log \frac{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}\} \qquad (27) \end{aligned}$$

If $z_i = v \neq k$, then:

$$\begin{aligned} |\hat{\gamma}_{ik} - 1_{z_i=k}| &= \frac{\pi_k Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \\ &\leq \frac{\pi_k}{\pi_v} \exp\{\log \frac{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{Pr(c_i \mid \hat{\mu}_v, \hat{\Sigma}_v)}\} \qquad (28) \end{aligned}$$

Consider $z_i = k$, then for any $j \neq k$ and for a sufficiently large $M > 0$:

$$Pr\left(\sup_i \left[\log \frac{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}\right] \geq -(m_R + M)\right) \to 0,$$

That is,

$$Pr\left(\min_i \left[\log \frac{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}\right] \leq m_R + M\right) \to 0,$$

Then,

$$\begin{aligned} & \log \frac{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \\ = \;& \log \frac{\left|\hat{\Sigma}_k\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(c_i - \hat{\mu}_k)' \hat{\Sigma}_k^{-1} (c_i - \hat{\mu}_k)}}{\left|\hat{\Sigma}_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(c_i - \hat{\mu}_j)' \hat{\Sigma}_j^{-1} (c_i - \hat{\mu}_j)}} \\ = \;& -\frac{1}{2} \log \frac{\left|\hat{\Sigma}_k\right|}{\left|\hat{\Sigma}_j\right|} - \frac{1}{2} \log \frac{\exp(c_i - \hat{\mu}_k)' \hat{\Sigma}_k^{-1} (c_i - \hat{\mu}_k)}{\exp(c_i - \hat{\mu}_j)' \hat{\Sigma}_j^{-1} (c_i - \hat{\mu}_j)} \end{aligned}$$

Let $B_k = \bar{c}_i^{[k]} + \epsilon_i = c_i - \hat{\mu}_k$:

$$\bar{c}_i^{[k]} + \epsilon_i = A_k r_i - A_k \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} r_i}{\sum_{i=1}^{n} \hat{\gamma}_{ik}} = A_k (r_i - \bar{r}_i[k]) + \epsilon_i \quad (29)$$

where $\bar{r}_i[k] = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik} r_i}{\sum_{i=1}^{n} \hat{\gamma}_{ik}}$ denotes the mean of r_i on source k. Then,

$$\hat{\Sigma}_k = \bar{C}[k] \bar{C}[k]' + \sigma^2 I_{m_R+1} \qquad (30)$$

where $\bar{C}[k]^{-1} = (\bar{c}_{i1}^{[k]}, \bar{c}_{i2}^{[k]}, \cdots, \bar{c}_{i(m_R+1)}^{[k]})'$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ from additive noise assumption.
By Woodbury matrix identity,

$$\hat{\Sigma}_k^{-1} = \sigma^{-2} I - \sigma^{-2} \bar{C}[k] (\sigma^2 I + \bar{C}[k]' \bar{C}[k])^{-1} \bar{C}[k]' \quad (31)$$

and

$$\begin{aligned} (\sigma^2 I + \bar{C}[k]' \bar{C}[k])^{-1} &= (\bar{C}[k]' \bar{C}[k])^{-1} \\ + \sigma^2 (\sigma^2 I + \bar{C}[k]' \bar{C}[k])^{-1} &(\bar{C}[k]' \bar{C}[k])^{-1} \qquad (32) \end{aligned}$$

We have,

$$\begin{aligned} & \left(\bar{c}_i^{[k]} + \epsilon_i\right)' \hat{\Sigma}_k^{-1} \left(\bar{c}_i^{[k]} + \epsilon_i\right) = B_k' \hat{\Sigma}_k^{-1} B_k \\ = \;& \sigma^{-2} B_k' (I - \bar{C}[k] (\bar{C}[k]' \bar{C}[k])^{-1} \bar{C}[k]') B_k \\ - \;& B_k' \bar{C}[k] (\sigma^2 I + \bar{C}[k]' \bar{C}[k])^{-1} (\bar{C}[k]' \bar{C}[k])^{-1} \bar{C}[k]' B_k \end{aligned}$$

Let $M_1[k] = \bar{C}[k]' \bar{C}[k]$ and $M_2[k] = \bar{C}[k])^{-1} (\bar{C}[k]' \bar{C}[k])$ and $M_3[k] = (I - \bar{C}[k] (\bar{C}[k]' \bar{C}[k])^{-1} \bar{C}[k]')$:

$$\begin{aligned} & \left(\bar{c}_i^{[k]} + \epsilon_i\right)' \hat{\Sigma}_k^{-1} \left(\bar{c}_i^{[k]} + \epsilon_i\right) \\ = \;& \sigma^{-2} B_k' M_3[k] B_k - B_k' \bar{C}[k] M_2[k]^{-1} M_1[k]^{-1} \bar{C}[k]' B_k \end{aligned}$$

Thus,

$$\begin{aligned}
& \log \frac{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \\
= \ & -\frac{1}{2} \log \frac{\left|\hat{\Sigma}_k\right|}{\left|\hat{\Sigma}_j\right|} - \frac{1}{2} \log \frac{\exp(c_i - \hat{\mu}_k)'\hat{\Sigma}_k^{-1}(c_i - \hat{\mu}_k)}{\exp(c_i - \hat{\mu}_j)'\hat{\Sigma}_j^{-1}(c_i - \hat{\mu}_j)} \\
= \ & -\frac{1}{2} \log |\bar{C}[k]\bar{C}[k]' + \sigma^2 I| + \frac{1}{2} \log |\bar{C}[j]\bar{C}[j]' + \sigma^2 I| \\
- \ & \frac{1}{2}(\sigma^{-2} B_k' M_3[k] B_k + B_k' \bar{C}[k] M_2[k]^{-1} M_1[k]^{-1} \bar{C}[k]' B_k) \\
+ \ & \frac{1}{2}(\sigma^{-2} B_j' M_3[j] B_j + B_j' \bar{C}[j] M_2[j]^{-1} M_1[j]^{-1} \bar{C}[j]' B_j) \\
= \ & -\frac{1}{2} \log |\bar{C}[k]\bar{C}[k]' + \sigma^2 I| + \frac{1}{2} \log |\bar{C}[j]\bar{C}[j]' + \sigma^2 I| \\
- \ & \frac{1}{2}\sigma^{-2} B_k' M_3[k] B_k + \frac{1}{2}\sigma^{-2} B_j' M_3[j] B_j \\
- \ & \frac{1}{2} B_k' \bar{C}[k] M_2[k]^{-1} M_1[k]^{-1} \bar{C}[k]' B_k \\
+ \ & \frac{1}{2} B_j' \bar{C}[j] M_2[j]^{-1} M_1[j]^{-1} \bar{C}[j]' B_j
\end{aligned}$$

$$\begin{aligned}
& \log \frac{Pr(c_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{Pr(c_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \\
\geq \ & -\frac{1}{2} \sup \log |\bar{C}[k]\bar{C}[k]' + \sigma^2 I| \\
- \ & \frac{1}{2}\sigma^{-2} \sup |B_k' M_3[k] B_k - B_k' M_3^*[k] B_k| \\
- \ & \frac{1}{2}\sigma^{-2} \sup |B_j' M_3[j] B_j - B_j' M_3^*[j] B_j| \\
- \ & \frac{1}{2} \sup B_k' \bar{C}[k] M_2[k]^{-1} M_1[k]^{-1} \bar{C}[k]' B_k \\
+ \ & \frac{1}{2} B_j' \bar{C}[j] M_2[j]^{-1} M_1[j]^{-1} \bar{C}[j]' B_j \\
\geq \ & \frac{1}{2} B_j' \bar{C}[j] M_2[j]^{-1} M_1[j]^{-1} \bar{C}[j]' B_j - M
\end{aligned}$$

Thus,

$$Pr(\frac{1}{2} B_j' \bar{C}[j] M_2[j]^{-1} M_1[j]^{-1} \bar{C}[j]' B_j \leq M + m_R) \to 0$$

If $n \to \infty$, then the number of samples for each group $n_k = n \times Pr(Z = k) \to \infty$ and we have $\hat{\mu}_k \xrightarrow{p} \mu_k$ and $\hat{\Sigma}_k \xrightarrow{p} \Sigma_k$ for each $k = 1, 2, \cdots, K$. If the dimension of representations is sufficiently large and the distribution of representations for different groups can be separated with high probability[7], then $|\hat{\gamma}_{ik} - 1_{z_i=k}| = o_p\left(\exp(-(m_R + M))\right)$ for a sufficiently large $M > 0$. That means that $\hat{\gamma}_{ik} \xrightarrow{p} Pr(z_i = k) \xrightarrow{p} 1_{z_i=k}$.

Proof **(2)** As $(m_R, n) \to \infty$, for each k and for a sufficiently large $M > 0$: $\hat{\gamma}_{ik} \xrightarrow{p} Pr(z_i = k) \xrightarrow{p} 1_{z_i=k}$. We sample the sub-group indicator z_i as group IV: $\hat{z}_i \sim \text{Disc}(\hat{\gamma}_{i1}, \hat{\gamma}_{i2}, \cdots, \hat{\gamma}_{iK})$. Due to the randomness, it is possible that the estimated source label $\hat{z}_i$ and the true label z_i do not match for each sample $\{x_i, t_i, y_i\}$.

Nevertheless, in probability, the treatment assignment mechanism of the estimated group label is consistent with that of the real label, i.e., $Pr(\hat{Z}, X, T, Y) \xrightarrow{p} Pr(Z, X, T, Y)$. We can view this phenomenon as the fact that when two identical samples swap labels between groups, the treatment assignment mechanism and the joint distribution will remain the same. Therefore, $\hat{z}_i$ still is an asymptotic IV for treatment effect regression.

B.2 Asymptotic Results of ITE Estimation

Identification of ITE Under *the additive noise assumption 3.1*, the identification results for ITE $g(\cdot)$ were developed by (Newey and Powell 2003; Hartford et al. 2017). Therefore, we plug GIV into IV regression methods to estimate ITE $g(\cdot)$.

Theorem B.2 ***Asymptotic Results of ITE Estimation.*** *Taking the expectation of outcome function in Eq. (1) conditional on $\{Z, X\}$ and applying the above assumptions, we establish the relationship:*

$$\begin{aligned}
\mathbb{E}[Y \mid Z, X] &= \mathbb{E}[g(T, X) \mid Z, X] + \mathbb{E}[\epsilon_Y \mid X] \\
&= \int [g(T, X) + C] dF(T \mid Z, X), \qquad (33)
\end{aligned}$$

where $dF(T \mid Z, X)$ is the conditional treatment distribution, C is constant as T is changed. The relationship in Eq. (33) defines an inverse problem for $g(\cdot)$ in terms of two directly observable functions: $\mathbb{E}[Y \mid Z, X]$ and $dF(T \mid Z, X)$.

Proof 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 (Newey and Powell 2003). 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, (Newey and Powell 2003; Hartford et al. 2017) identified the causal effect as the solution of an integral equation:

$$\begin{aligned}
\mathbb{E}[Y \mid Z, X] &= \mathbb{E}[g(X, T) \mid Z, X] + \mathbb{E}[\epsilon_Y \mid X] \\
&= \int [g(T, X) + \mathbb{E}[\epsilon_Y \mid X]] \, dF(T \mid Z, X) \\
&= \int [g(T, X) + C] \, dF(T \mid Z, X) \\
&= \int \hat{g}(T, X) dF(T \mid Z, X) \qquad (34)
\end{aligned}$$

where F denotes the conditional cumulative distribution function of T given $\{Z, X\}$, and $\hat{g}(T, X) = g(T, X) + C$. Given two observable functions $\mathbb{E}[Y \mid Z, X]$ and $F(T \mid Z, X)$, $\hat{g}(T, X)$ is the solution of the inverse problem.

[7] This implicitly assumes that $m_X \to \infty$. Only when the information in the raw data is sufficient, we can obtain as many representations as possible that nonzero contribute (α_k) for all k, i.e., $m_R \to \infty$.

Table 4: The Mean Squared Error $mean(std)$ of Different Synthetic Settings ($Data$-K-m_X)

		Linear-2-5				Linear-3-5			
		Poly2SLS	KernelIV	DeepIV	AGMM	Poly2SLS	KernelIV	DeepIV	AGMM
	None	0.203(0.017)	0.292(0.026)	0.371(0.019)	0.116(0.015)	0.312(0.034)	0.415(0.047)	0.492(0.030)	0.129(0.018)
Summary IV	UAS	0.203(0.017)	0.292(0.027)	0.376(0.020)	0.118(0.012)	0.312(0.034)	0.415(0.046)	0.485(0.029)	0.130(0.017)
	WAS	0.204(0.017)	0.288(0.040)	0.368(0.017)	0.127(0.020)	0.314(0.033)	0.417(0.057)	0.489(0.028)	0.186(0.035)
	ModeIV	0.203(0.017)	0.288(0.025)	0.368(0.017)	0.113(0.016)	0.312(0.034)	0.418(0.042)	0.489(0.030)	0.130(0.018)
	AutoIV	12.739(28.272)	0.288(0.028)	0.372(0.021)	0.118(0.017)	> 100	0.416(0.047)	0.486(0.033)	0.130(0.017)
Our Method	GIV$_{KM}$	0.059(0.004)	0.250(0.023)	0.286(0.018)	0.086(0.012)	0.088(0.020)	0.317(0.037)	0.361(0.026)	0.110(0.013)
	GIV$_{KM}$*	0.059(0.004)	0.250(0.023)	0.289(0.023)	0.086(0.012)	0.142(0.087)	0.274(0.031)	0.284(0.029)	0.089(0.007)
	GIV$_{EM}$	**0.058(0.030)**	**0.141(0.016)**	**0.076(0.006)**	**0.057(0.005)**	**0.045(0.004)**	**0.167(0.028)**	**0.104(0.008)**	**0.067(0.007)**
	TrueIV	**0.044(0.006)**	**0.139(0.016)**	**0.078(0.008)**	**0.058(0.006)**	**0.044(0.004)**	**0.169(0.032)**	**0.101(0.010)**	**0.069(0.005)**
		Linear-5-5				Linear-3-10			
		Poly2SLS	KernelIV	DeepIV	AGMM	Poly2SLS	KernelIV	DeepIV	AGMM
	None	0.474(0.044)	0.538(0.093)	0.658(0.042)	0.103(0.022)	0.292(0.023)	0.400(0.096)	0.655(0.051)	0.076(0.015)
Summary IV	UAS	0.474(0.044)	0.540(0.094)	0.651(0.049)	0.102(0.020)	0.292(0.023)	0.400(0.096)	0.656(0.055)	0.077(0.015)
	WAS	0.478(0.045)	0.553(0.100)	0.651(0.045)	0.108(0.020)	0.294(0.024)	0.413(0.098)	0.657(0.048)	0.079(0.015)
	ModeIV	0.474(0.044)	0.547(0.094)	0.657(0.042)	0.103(0.019)	0.293(0.024)	0.403(0.097)	0.653(0.062)	0.079(0.016)
	AutoIV	> 100	0.540(0.093)	0.656(0.055)	0.108(0.026)	> 100	0.399(0.096)	0.654(0.058)	0.077(0.014)
Our Method	GIV$_{KM}$	0.268(0.012)	0.374(0.080)	0.424(0.046)	0.066(0.010)	0.098(0.042)	0.357(0.076)	0.466(0.055)	0.073(0.015)
	GIV$_{KM}$*	0.196(0.290)	0.231(0.063)	**0.208(0.022)**	0.052(0.005)	0.090(0.028)	**0.302(0.063)**	0.258(0.032)	0.062(0.010)
	GIV$_{EM}$	**0.192(0.382)**	**0.207(0.042)**	**0.146(0.032)**	**0.046(0.007)**	**0.062(0.096)**	**0.302(0.061)**	**0.258(0.033)**	**0.051(0.011)**
	TrueIV	**0.130(0.144)**	**0.199(0.041)**	0.239(0.040)	**0.043(0.006)**	**0.032(0.003)**	0.311(0.059)	**0.102(0.016)**	**0.054(0.012)**

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. (34)), the identification/uniqueness 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 ITE under assumption 3.1 and $C = \mathbb{E}[Y - \hat{g}(T, X)]$.

C The Experiments for Stability

We increase the critical level of simulation and set $Data$-K-m_X with different group numbers K and dimensions of covariates m_X to test the stability of our GIV. Comparing with the results of setting Linear-3-3 in Table 1, Linear-3-5 and Linear-3-10 in Table 4, GIV$_{EM}$ 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 4), GIV$_{EM}$ also shows in stability and is in Top2. In the above settings, GIV$_{EM}$ has outstanding performance, which is close to TrueIV.

D The Experiments for Non-linear Cases

To verify the effectiveness of GIV in non-linear cases, we design 5 different treatment functions $f_X(\cdot)$ to evaluate the treatment effect estimation performance of Meta-EM algorithm. We select the SOTA IV-based methods (Poly2SLS, KernelIV, DeepIV, AGMM) in four lines 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. 6) 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.

E The Full Results in IHDP & PM-CMR

E.1 Real-world datasets

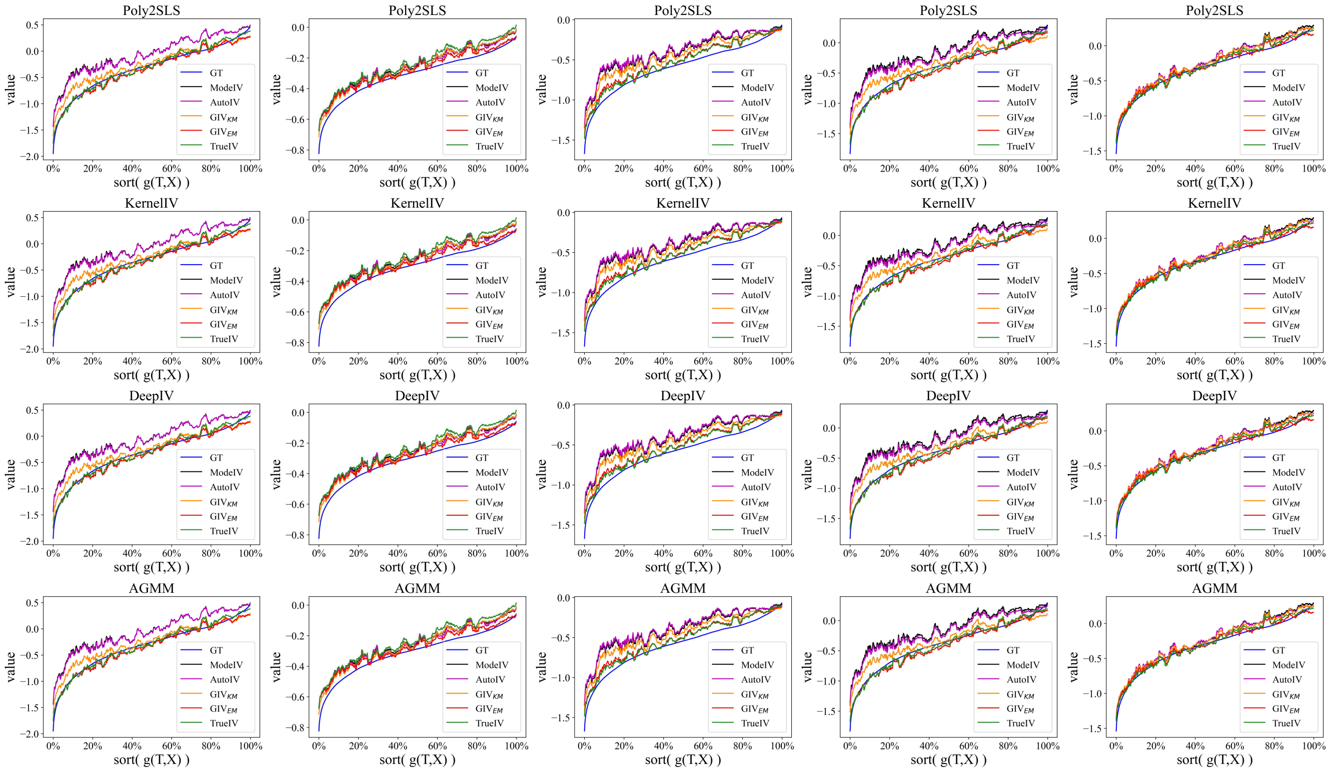
Similar to previous methods(Nie et al. 2020; Hartford et al. 2017; Bica, Jordon, and van der Schaar 2020; Schwab et al. 2020), we perform experiments on two semi-synthetic real-world datasets **IHDP** (Shalit, Johansson, and Sontag 2017) & **PM-CMR** (Wyatt et al. 2020), as the counterfactual outcomes are rarely available for real-world data. 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 estimation.

IHDP (Shalit, Johansson, and Sontag 2017). 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. (17)&(19) to generate the treatment T and corresponding outcome Y.

PM-CMR (Wyatt et al. 2020). The PM-CMR study the impact of $PM_{2.5}$ partical 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 (Wyatt et al. 2020). Then we use 6 continuous variables about CMR in each city as the confounders to replace the covariates X in Eq. (17)&(19) to generate the treatment T and the corresponding outcome Y.

E.2 Results

By estimating the latent differentiated covariate-treatment distribution parameters across groups, Meta-EM reconstructs the latent IV. From the results in Fig. 8) & 7, we have the following observation: (1) the optimal parameter ($K = 2$) identified by Meta-EM is consistent with the



(a) Linear-3-3 (b) Poly-3-3 (c) Sin-3-3 (d) Sigmoid-3-3 (e) Abs-3-3

Figure 6: Treatment Effect Estimation (sorted by Ground-Truth $g(T, X)$) in Different Scenario $Data$-3-3.

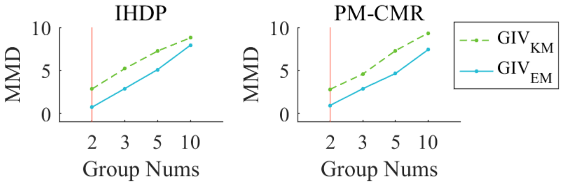


Figure 7: MMD for Selection of Group Number in IHDP & PM-CMR Dataset.

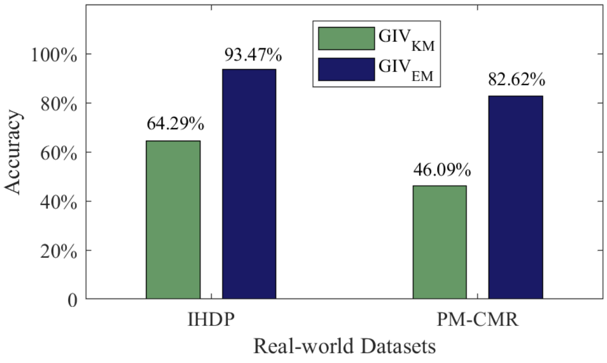


Figure 8: Reconstruction Accuracy of the Group IV in IHDP & PM-CMR Dataset.

ground truth, meaning Meta-EM can find the optimal parameter K; (2) the MMD of Meta-EM is still significantly smaller than that of Meta-KM and (3) the reconstruction accuracy reaches 93.47% and 82.62% on **IHDP** and **PM-CMR**, however, K-Means is only 64.29% and 46.09%. (2-3) demonstrate Meta-EM can automatically find the optimal IV, but K-Means cannot.

To verify that GIV$_{EM}$ with higher reconstruction accuracy achieves better performance to predict treatment effect, we assess GIV and Summary IVs' performance in treatment effect 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 effect estimation here. The full results of MSE $mean(std)$ of IHDP & PM-CMR Dataset with T=do(t) are shown in Table 5, 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 GIV$_{EM}$, the performance of GIV$_{KM}$ exceeds most baselines in downstream tasks, but it is still inferior to GIV$_{EM}$ and TrueIV.

Table 5: The Full Results of MSE $mean(std)$ of IHDP & PM-CMR Dataset

IHDP Dataset

		Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
	NoneIV	0.238(0.132)	2.127(1.967)	0.456(0.243)	0.762(0.279)	0.583(0.240)	0.444(0.186)	0.979(0.391)	0.239(0.186)	0.140(0.063)
Summary IV	UAS	0.238(0.133)	2.712(2.795)	0.457(0.244)	0.659(0.254)	0.574(0.245)	0.447(0.185)	1.024(0.426)	0.243(0.129)	0.142(0.061)
	WAS	0.239(0.134)	1.954(1.836)	0.455(0.241)	0.746(0.254)	0.567(0.226)	0.456(0.185)	1.010(0.408)	0.195(0.102)	0.144(0.059)
	ModeIV	0.240(0.133)	2.511(2.048)	0.460(0.246)	0.752(0.277)	0.572(0.245)	0.468(0.193)	0.989(0.387)	0.220(0.085)	0.149(0.060)
	AutoIV	> 100	2.392(2.052)	0.457(0.243)	0.682(0.257)	0.583(0.250)	0.458(0.199)	0.998(0.405)	0.196(0.087)	0.142(0.069)
Our Method	GIV$_{KM}$	0.078(0.029)	1.009(0.964)	0.354(0.179)	**0.651(0.266)**	0.505(0.203)	0.383(0.155)	0.967(0.343)	**0.131(0.037)**	0.112(0.050)
	GIV$_{EM}$	**0.034(0.011)**	**0.585(1.342)**	**0.202(0.173)**	**0.653(0.255)**	**0.482(0.228)**	**0.283(0.163)**	**0.967(0.333)**	**0.137(0.041)**	**0.095(0.035)**
	TrueIV	**0.033(0.009)**	**0.146(0.044)**	**0.151(0.060)**	0.654(0.256)	**0.458(0.166)**	**0.227(0.079)**	**0.948(0.342)**	0.152(0.040)	**0.093(0.035)**

PM-CMR Dataset

		Poly2SLS	NN2SLS	KernelIV	DualIV	DeepIV	OneSIV	DFIV	DeepGMM	AGMM
	NoneIV	0.181(0.044)	1.241(0.646)	0.352(0.198)	0.727(0.230)	0.409(0.160)	0.369(0.182)	0.995(0.166)	0.145(0.043)	0.130(0.064)
Summary IV	UAS	0.181(0.044)	1.439(0.733)	0.352(0.198)	0.656(0.205)	0.404(0.164)	0.365(0.181)	0.994(0.177)	0.213(0.096)	0.128(0.064)
	WAS	0.181(0.044)	0.939(0.465)	0.372(0.207)	0.906(0.251)	0.417(0.164)	0.417(0.166)	**0.967(0.202)**	0.200(0.088)	0.157(0.080)
	ModeIV	0.181(0.044)	1.515(0.687)	0.359(0.201)	0.749(0.203)	0.406(0.150)	0.391(0.190)	1.035(0.186)	0.204(0.108)	0.131(0.070)
	AutoIV	0.179(0.044)	1.224(0.719)	0.351(0.198)	0.706(0.220)	0.409(0.180)	0.379(0.204)	0.984(0.195)	0.227(0.193)	0.129(0.064)
Our Method	GIV$_{KM}$	0.088(0.044)	0.719(0.452)	0.329(0.202)	**0.658(0.208)**	0.381(0.165)	0.341(0.178)	1.049(0.199)	0.174(0.062)	0.117(0.053)
	GIV$_{EM}$	**0.048(0.012)**	**0.624(0.395)**	**0.308(0.210)**	**0.666(0.203)**	**0.339(0.184)**	**0.306(0.164)**	**0.982(0.205)**	**0.125(0.052)**	**0.085(0.045)**
	TrueIV	**0.028(0.007)**	**0.190(0.054)**	**0.140(0.074)**	0.678(0.202)	**0.141(0.054)**	**0.154(0.071)**	0.993(0.207)	**0.112(0.048)**	**0.054(0.023)**