A Proofs

The main part of our model identifiability is essentially the same as that of Theorem 1 in [37], but now adapted to the dependency on t. Here we give an outline of the proof, and the details can be easily filled by referring to [37]. In the proof, subscripts t are omitted for convenience.

Proof of Lemma 1. Using **(M1)** i) and ii) , we transform $p_{\boldsymbol{f},\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x},t) = p_{\boldsymbol{f}',\boldsymbol{\lambda}'}(\mathbf{y}|\mathbf{x},t)$ into equality of noiseless distributions, that is,

$$q_{\boldsymbol{f}',\boldsymbol{\lambda}'}(\mathbf{y}) = q_{\boldsymbol{f},\boldsymbol{\lambda}}(\mathbf{y}) := p_{\boldsymbol{\lambda}}(\boldsymbol{f}^{-1}(\mathbf{y})|\mathbf{x},t)vol(\boldsymbol{J}_{\boldsymbol{f}^{-1}}(\mathbf{y}))\mathbb{I}_{\mathcal{Y}}(\mathbf{y}) \tag{15}$$

where $p_{\boldsymbol{\lambda}}$ is the Gaussian density function of the conditional prior defined in (4) and $vol(A) := \sqrt{\det AA^T}$. $q_{\boldsymbol{f}',\boldsymbol{\lambda}'}$ is defined similarly to $q_{\boldsymbol{f},\boldsymbol{\lambda}}$.

Then, apply model (4) to (15), plug the $2n+1$ points from **(D1)** into it, and re-arrange the resulting $2n+1$ equations in matrix form, we have

$$\mathcal{F}'(\mathbf{y}) = \mathcal{F}(\mathbf{y}) := \boldsymbol{L}^T\boldsymbol{t}(\boldsymbol{f}^{-1}(\mathbf{y})) - \boldsymbol{\beta} \tag{16}$$

where $\boldsymbol{t}(\mathbf{z}) := (\mathbf{z}, \mathbf{z}^2)^T$ is the sufficient statistics of factorized Gaussian, and $\boldsymbol{\beta}_t := (\alpha_t(\boldsymbol{x}_1) - \alpha_t(\boldsymbol{x}_0), ..., \alpha_t(\boldsymbol{x}_{2n}) - \alpha_t(\boldsymbol{x}_0))^T$ where $\alpha_t(\mathbf{x};\boldsymbol{\lambda}_t)$ is the log-partition function of the conditional prior in (4). $\mathcal{F}'$ is defined similarly to $\mathcal{F}$, but with $\boldsymbol{f}', \boldsymbol{\lambda}', \alpha'$

Since $\boldsymbol{L}$ is invertible, we have

$$\boldsymbol{t}(\boldsymbol{f}^{-1}(\mathbf{y})) = \boldsymbol{A}\boldsymbol{t}(\boldsymbol{f}'^{-1}(\mathbf{y})) + \boldsymbol{c} \tag{17}$$

where $\boldsymbol{A} = \boldsymbol{L}^{-T}\boldsymbol{L}'^T$ and $\boldsymbol{c} = \boldsymbol{L}^{-T}(\boldsymbol{\beta} - \boldsymbol{\beta}')$.

The final part of the proof is to show, by following the same reasoning as in Appendix B of [61], that $\boldsymbol{A}$ is a sparse matrix such that

$$\boldsymbol{A} = \begin{pmatrix} \mathrm{diag}(\boldsymbol{a}) & \boldsymbol{O} \\ \mathrm{diag}(\boldsymbol{u}) & \mathrm{diag}(\boldsymbol{a}^2) \end{pmatrix} \tag{18}$$

where $\boldsymbol{A}$ is partitioned into four n-square matrices. Thus

$$\boldsymbol{f}^{-1}(\mathbf{y}) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{f}'^{-1}(\mathbf{y}) + \boldsymbol{b} \tag{19}$$

where $\boldsymbol{b}$ is the first half of $\boldsymbol{c}$. □

Proof of Proposition 2. Under **(G2)**, and **(M3)**, we have

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x},\mathrm{t}) = \mathbb{E}(\mathbf{y}|\mathbf{x},\mathrm{t}) \implies \boldsymbol{f}_t \circ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{j}_t \circ \mathbb{P}(\boldsymbol{x}) \text{ on } (\boldsymbol{x},t) \text{ such that } p(t,\boldsymbol{x}) > 0. \tag{20}$$

We show the solution set of (20) on *overlapped* $\boldsymbol{x}$ is

$$\{(\boldsymbol{f},\boldsymbol{h})|\boldsymbol{f}_t = \boldsymbol{j}_t \circ \Delta^{-1}, \boldsymbol{h} = \Delta \circ \mathbb{P}, \Delta : \mathcal{P} \to \mathbb{R}^n \text{ is injective}\}. \tag{21}$$

By **(G2)(M1)**, and with injective $\boldsymbol{f}_t, \boldsymbol{j}_t$ and $\dim(\mathbf{z}) = \dim(\mathbf{y}) \geq \dim(\mathbb{P})$, for any Δ above, there exists a functional parameter $\boldsymbol{f}_t$ such that $\boldsymbol{j}_t = \boldsymbol{f}_t \circ \Delta$. Thus, set (21) is non-empty, and any element is indeed a solution because $\boldsymbol{f}_t \circ \boldsymbol{h} = \boldsymbol{j}_t \circ \Delta^{-1} \circ \Delta \circ \mathbb{P} = \boldsymbol{j}_t \circ \mathbb{P}$.

Any solution of (20) should be in (21). A solution should satisfy $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{f}_t^{-1} \circ \boldsymbol{j}_t \circ \mathbb{P}(\boldsymbol{x})$ for both t since $\boldsymbol{x}$ is overlapped. This means the *injective* function $\boldsymbol{f}_t^{-1} \circ \boldsymbol{j}_t$ should *not* depend on t, thus it is one of the Δ in (21).

We proved conclusion 1) with $\boldsymbol{v} := \Delta$. And, on overlapped $\boldsymbol{x}$, conclusion 2) is quickly seen from

$$\hat{\mu}_t(\boldsymbol{x}) = \boldsymbol{f}_t(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}(\boldsymbol{v} \circ \mathbb{P}(\boldsymbol{x})) = \boldsymbol{j}_t(\mathbb{P}(\boldsymbol{x})) = \mu_t(\boldsymbol{x}). \tag{22}$$

We rely on overlapped $\mathbb{P}$ to work for non-overlapped $\boldsymbol{x}$. For any $\boldsymbol{x}_t$ with $p(1-t|\boldsymbol{x}_t) = 0$, to ensure $p(1-t|\mathbb{P}(\boldsymbol{x}_t)) > 0$, there should exist $\boldsymbol{x}_{1-t}$ such that $\mathbb{P}(\boldsymbol{x}_{1-t}) = \mathbb{P}(\boldsymbol{x}_t)$ and $p(1-t|\boldsymbol{x}_{1-t}) > 0$. And we also have $\boldsymbol{h}(\boldsymbol{x}_{1-t}) = \boldsymbol{h}(\boldsymbol{x}_t)$ due to **(M2)**. Then, we have

$$\hat{\mu}_{1-t}(\boldsymbol{x}_t) = \boldsymbol{f}_{1-t}(\boldsymbol{h}(\boldsymbol{x}_t)) = \boldsymbol{f}_{1-t}(\boldsymbol{h}(\boldsymbol{x}_{1-t})) = \boldsymbol{j}_{1-t}(\mathbb{P}(\boldsymbol{x}_{1-t})) = \boldsymbol{j}_{1-t}(\mathbb{P}(\boldsymbol{x}_t)) = \mu_{1-t}(\boldsymbol{x}_t). \tag{23}$$

The third equality uses (20) on $(\boldsymbol{x}_{1-t}, 1-t)$. □

Proof of Theorem 1. By **(M1)** and **(G1')**, for any injective function $\Delta : \mathcal{P} \to \mathbb{R}^n$, there exists a functional parameter $\boldsymbol{f}_t^*$ such that $\boldsymbol{j}_t = \boldsymbol{f}_t^* \circ \Delta$. Let $\boldsymbol{h}_t^* = \Delta \circ \mathbb{P}_t$, then, clearly from **(M3')**, such parameters $\boldsymbol{\theta}^* = (\boldsymbol{f}^*, \boldsymbol{h}^*)$ are optimal: $p_{\boldsymbol{\theta}^*}(\mathbf{y}|\mathbf{x},\mathrm{t}) = p(\mathbf{y}|\mathbf{x},\mathrm{t})$.

Since have all assumptions for Lemma 1, we have

$$\Delta \circ \boldsymbol{j}^{-1}(\boldsymbol{y}) = \boldsymbol{f}^{*-1}(\boldsymbol{y}) = \mathcal{A} \circ \boldsymbol{f}^{-1}(\boldsymbol{y})|_t, \text{ on } (\boldsymbol{y}, t) \in \{(\boldsymbol{j}_t \circ \mathbb{P}_t(\boldsymbol{x}), t) | p(t, \boldsymbol{x}) > 0\}, \tag{24}$$

where $\boldsymbol{f}$ is *any* optimal parameter, and "$|_t$" collects all subscripts t. Note, except for Δ, all the symbols should have subscript t.

Nevertheless, using **(D2)**, we can further prove $\mathcal{A}_0 = \mathcal{A}_1$.

We repeat the core quantities from Lemma 1 here: $\boldsymbol{A}_t = \boldsymbol{L}_t^{-T} \boldsymbol{L}_t'^{T}$ and $\boldsymbol{c}_t = \boldsymbol{L}_t^{-T}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t')$.

From **(D2)**, we immediately have

$$\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{L}_0'^{-1}\boldsymbol{L}_1' = \boldsymbol{C} \iff \boldsymbol{A}_0 = \boldsymbol{A}_1 \tag{25}$$

And also,

$$\begin{aligned} \boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{C} &\iff \boldsymbol{L}_0^{-T}\boldsymbol{C}^{-T} = \boldsymbol{L}_1^{-T} \\ \boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0' - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1' = \boldsymbol{d}/k &\iff \boldsymbol{C}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0') = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1' \end{aligned} \tag{26}$$

Multiply right hand sides of the two lines, we have $\boldsymbol{c}_0 = \boldsymbol{c}_1$. Now we have $\mathcal{A}_0 = \mathcal{A}_1 := \mathcal{A}$. Apply this to (24), we have

$$\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}, \quad \boldsymbol{v} := \mathcal{A}^{-1} \circ \Delta \tag{27}$$

for *any* optimal parameters $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{h})$. Again, from **(M3')**, we have

$$p_{\boldsymbol{\theta}}(\mathrm{y}|\mathrm{x}, \mathrm{t}) = p(\mathrm{y}|\mathrm{x}, \mathrm{t}) \implies p_{\boldsymbol{\epsilon}}(\mathrm{y} - \boldsymbol{f}_{\mathrm{t}}(\boldsymbol{h}_{\mathrm{t}}(\mathrm{x}))) = p_{\mathbf{e}}(\mathrm{y} - \boldsymbol{j}_{\mathrm{t}}(\mathbb{P}_{\mathrm{t}}(\mathrm{x}))) \tag{28}$$

where $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$. And the above is only possible when $\boldsymbol{f}_t \circ \boldsymbol{h}_t = \boldsymbol{j}_t \circ \mathbb{P}_t$. Combined with $\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}$, we have conclusion 1).

And conclusion 2) follows from the same reasoning as Proposition 2, applied to both $\mathbb{P}_0$ and $\mathbb{P}_1$. $\square$

We include an **erratum** here. The definition of the domain of $\boldsymbol{v}$ in Theorem 1 was incorrect. As seen from (24), the domain of $\boldsymbol{v}$ should be $\{\mathbb{P}_t(\boldsymbol{x}) | p(t, \boldsymbol{x}) > 0\}$, which is the support of *factual* PtS $\mathbb{P}_{\mathrm{t}}(\mathbf{x})$. This error was minor for identification of CATE since we assume overlapped $\mathbb{P}_t(\mathbf{x})$ and **(M2)**.

Note, when multiplying the two lines of (26), the effects of $k \to 0$ cancel out, and $\boldsymbol{c}_t$ is finite and well-defined. Also, it is apparent from above proof that **(D2)** is a *necessary and sufficient* condition for $\mathcal{A}_0 = \mathcal{A}_1$, if other conditions of Theorem 1 are given.

Below, we prove the results in Sec. 4.2. To make it apparent that the definitions and results work for the posterior, we replace p_t with q_t and prove the results. The dependence on $\boldsymbol{f}$ and q (or p when repeating the proofs for the prior) prevail, and the sub / superscripts are omitted. The arguments $\boldsymbol{x}$ are sometimes also omitted.

Proof of Lemma 2.

$$\begin{aligned} &\epsilon_{CF} - \sum_t p(1-t|\boldsymbol{x})\epsilon_{F,t} \\ &= p(0|\boldsymbol{x})(\epsilon_{CF,1} - \epsilon_{F,1}) + p(1|\boldsymbol{x})(\epsilon_{CF,0} - \epsilon_{F,0}) \\ &= p(0|\boldsymbol{x}) \int \mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, 1)(q_0(\boldsymbol{z}|\boldsymbol{x}) - q_1(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} + p(1|\boldsymbol{x}) \int \mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, 0)(q_1(\boldsymbol{z}|\boldsymbol{x}) - q_0(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{z} \\ &\leq 2M\mathbb{TV}(q_1, q_0) \leq M\mathbb{D}. \end{aligned}$$

$\square$

$\mathbb{TV}(p, q) := \frac{1}{2}\mathbb{E}|p(\mathbf{z}) - q(\mathbf{z})| = \frac{1}{2}\int |p(\boldsymbol{z}) - q(\boldsymbol{z})|d\boldsymbol{z}$ is the total variance distance between probability density p, q. The last inequality uses Pinsker's inequality $\mathbb{TV}(p, q) \leq \sqrt{D_{\mathrm{KL}}(p\|q)/2}$ twice, to get the symmetric $\mathbb{D}$.

The statement of Theorem 2 in the main text contains typos, and we include an **erratum** below. The typos are minor and all the implications of the result remain the same.

Theorem 3 (Theorem 2, typos fixed)**.** *Assume $|\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, t)| \leq M$ and $|\boldsymbol{g}_t(\boldsymbol{z})| \leq G$, then,*

$$\epsilon_{\boldsymbol{f}}(\boldsymbol{x}) \leq 2[G^2(\epsilon_{F,0}(\boldsymbol{x}) + \epsilon_{F,1}(\boldsymbol{x}) + M\mathbb{D}(\boldsymbol{x})) - \mathbb{V}_{\mathbf{y}}(\boldsymbol{x})]|^p \tag{29}$$

where $\mathbb{V}_{\mathbf{y}}^p(\boldsymbol{x}) := \mathbb{E}_{p(\mathbf{z}|\boldsymbol{x})} \sum_t \mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t = \mathbf{z})}(\mathbf{y}(t) - m_t(\mathbf{z}))^2$, and "$|^p$" collects all superscripts p.

Theorem 2 is a direct corollary of Lemma 2 and the following.

Lemma 3. *Define* $\epsilon_F = \sum_t p(t|\boldsymbol{x})\epsilon_{F,t}$. *We have*

$$\epsilon_{\boldsymbol{f}} \leq 2(G^2(\epsilon_F + \epsilon_{CF}) - \mathbb{V}_{\mathbf{y}}). \tag{30}$$

Simply bound ϵ_{CF} in (30) by Lemma 2, we have Theorem 2. To prove Lemma 3, we first examine a bias-variance decomposition of ϵ_F and ϵ_{CF}.

$$\begin{aligned}\epsilon_{CF,t} &= \mathbb{E}_{q_{1-t}(\mathbf{z}|\boldsymbol{x})}\boldsymbol{g}_t(\mathbf{z})^{-2}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}(\mathbf{y}(t) - \boldsymbol{f}_t(\mathbf{z}))^2 \\ &\geq G^{-2}\mathbb{E}_{q_{1-t}(\mathbf{z}|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}(\mathbf{y}(t) - \boldsymbol{f}_t(\mathbf{z}))^2 \\ &= G^{-2}\mathbb{E}_{q_{1-t}(\mathbf{z}|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}((\mathbf{y}(t) - m_t(\mathbf{z}))^2 + (m_t(\mathbf{z}) - \boldsymbol{f}_t(\mathbf{z}))^2)\end{aligned} \tag{31}$$

The second line uses $|\boldsymbol{g}_t(\mathbf{z})| \leq G$, and the third line is a bias-variance decomposition. Now we can define $\mathbb{V}^q_{CF,t}(\boldsymbol{x}) := \mathbb{E}_{q_{1-t}(\mathbf{z}|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}(\mathbf{y}(t) - m_t(\mathbf{z}))^2$ and $\mathbb{B}^q_{CF,t}(\boldsymbol{x}) := \mathbb{E}_{q_{1-t}(\mathbf{z}|\boldsymbol{x})}(m_t(\mathbf{z}) - \boldsymbol{f}_t(\mathbf{z}))^2$, and we have

$$\epsilon_{CF,t} \geq G^{-2}(\mathbb{V}_{CF,t}(\boldsymbol{x}) + \mathbb{B}_{CF,t}(\boldsymbol{x})) \implies \epsilon_{CF} \geq G^{-2}(\mathbb{V}_{CF}(\boldsymbol{x}) + \mathbb{B}_{CF}(\boldsymbol{x})) \tag{32}$$

where $\mathbb{V}_{CF} := \sum_t p(1-t|\boldsymbol{x})\mathbb{V}_{CF,t} = \sum_t \mathbb{E}_{q(\mathbf{z},\mathrm{t}=1-t|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}(\mathbf{y}(t) - m_t(\mathbf{z}))^2$ and similarly $\mathbb{B}_{CF} = \sum_t \mathbb{E}_{q(\mathbf{z},\mathrm{t}=1-t|\boldsymbol{x})}(m_t(\mathbf{z}) - \boldsymbol{f}_t(\mathbf{z}))^2$. Repeat the above derivation for ϵ_F, we have

$$\epsilon_F \geq G^{-2}(\mathbb{V}_F(\boldsymbol{x}) + \mathbb{B}_F(\boldsymbol{x})) \tag{33}$$

where $\mathbb{V}_F = \sum_t \mathbb{E}_{q(\mathbf{z},\mathrm{t}=t|\boldsymbol{x})}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\mathbf{z})}(\mathbf{y}(t) - m_t(\mathbf{z}))^2$ and $\mathbb{B}_F = \sum_t \mathbb{E}_{q(\mathbf{z},\mathrm{t}=t|\boldsymbol{x})}(m_t(\mathbf{z}) - \boldsymbol{f}_t(\mathbf{z}))^2$. Now, we are ready to prove Lemma 3.

Proof of Lemma 3.

$$\begin{aligned}\epsilon_{\boldsymbol{f}} &= \mathbb{E}_{q(\mathbf{z}|\boldsymbol{x})}((\boldsymbol{f}_1 - \boldsymbol{f}_0) - (m_1 - m_0))^2 \\ &= \mathbb{E}_q((\boldsymbol{f}_1 - m_1) + (m_0 - \boldsymbol{f}_0))^2 \\ &\leq 2\mathbb{E}_q((\boldsymbol{f}_1 - m_1)^2 + (m_0 - \boldsymbol{f}_0)^2) \\ &= 2\int[(\boldsymbol{f}_1 - m_1)^2 q(\boldsymbol{z},1|\boldsymbol{x}) + (m_0 - \boldsymbol{f}_0)^2 q(\boldsymbol{z},0|\boldsymbol{x}) + \\ &\qquad (\boldsymbol{f}_1 - m_1)^2 q(\boldsymbol{z},0|\boldsymbol{x}) + (m_0 - \boldsymbol{f}_0)^2 q(\boldsymbol{z},1|\boldsymbol{x})]d\boldsymbol{z} \\ &= 2(\mathbb{B}_F + \mathbb{B}_{CF}) \leq 2(G^2(\epsilon_F + \epsilon_{CF}) - \mathbb{V}_{\mathbf{y}})\end{aligned}$$

□

The first inequality uses $(a+b)^2 \leq 2(a^2+b^2)$. The next equality splits $q(\mathbf{z}|\mathbf{x})$ into $q(\mathbf{z},0|\mathbf{x})$ and $q(\mathbf{z},1|\mathbf{x})$ and rearranges to get $\mathbb{B}_F$ and $\mathbb{B}_{CF}$. The last inequality uses the two bias-variance decompositions, and $\mathbb{V}_{\mathbf{y}} = \mathbb{V}_F + \mathbb{V}_{CF}$.

B Additional backgrounds

B.1 Prognostic score and balancing score

In the fundamental work of [22], prognostic score is defined equivalently to our $\mathbb{P}_0$ (P0-score), but it in addition requires no effect modification to work for $\mathbf{y}(1)$. Thus, a useful prognostic score corresponds to our PtS. We give main properties of PtS as following.

Proposition 3. *If* $\mathbf{v}$ *gives exchangeability, and* $\mathbb{P}_{\mathrm{t}}(\mathbf{v})$ *is a PtS, then* $\mathbf{y}(t) \perp\!\!\!\perp \mathbf{v}, \mathrm{t}|\mathbb{P}_t$.

The following three properties of conditional independence will be used repeatedly in proofs.

Proposition 4 (Properties of conditional independence)**.** *[51, Sec. 1.1.55] For random variables* $\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$. *We have:*

$$\begin{aligned}\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \wedge \mathbf{x} \perp\!\!\!\perp \mathbf{w}|\mathbf{y},\mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{w},\mathbf{y}|\mathbf{z} \text{ (Contraction)}. \\ \mathbf{x} \perp\!\!\!\perp \mathbf{w},\mathbf{y}|\mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{w},\mathbf{z} \text{ (Weak union)}. \\ \mathbf{x} \perp\!\!\!\perp \mathbf{w},\mathbf{y}|\mathbf{z} &\implies \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \text{ (Decomposition)}.\end{aligned}$$

Proof of Proposition 3. From $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t}|\mathbf{v}$ (*exchangeability* of $\mathbf{v}$), and since $\mathbb{P}_t$ is a *function* of $\mathbf{v}$, we have $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t}|\mathbb{P}_t, \mathbf{v}$ (1).

From (1) and $\mathbf{y}(t) \perp\!\!\!\perp \mathbf{v}|\mathbb{P}_t(\mathbf{v})$ (definition of Pt-score), using contraction rule, we have $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t}, \mathbf{v}|\mathbb{P}_t$ for both t. □

Prognostic scores are closely related to the important concept of balancing score [54]. Note particularly, the proposition implies $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t}|\mathbb{P}_t$ (using decomposition rule). Thus, if $\mathbb{P}(\mathbf{v})$ is a P-score, then $\mathbb{P}$ also gives weak ignorability (exchangeability and overlap), which is a nice property shared with balancing score, as we will see immediately.

Definition 4 (Balancing score)**.** $\boldsymbol{b}(\mathbf{v})$, a function of random variable $\mathbf{v}$, is a balancing score if $\mathrm{t} \perp\!\!\!\perp \mathbf{v}|\boldsymbol{b}(\mathbf{v})$.

Proposition 5. *Let $\boldsymbol{b}(\mathbf{v})$ be a function of random variable $\mathbf{v}$. $\boldsymbol{b}(\mathbf{v})$ is a balancing score if and only if $f(\boldsymbol{b}(\mathbf{v})) = p(\mathrm{t}=1|\mathbf{v}) := e(\mathbf{v})$ for some function f (or more formally, $e(\mathbf{v})$ is $\boldsymbol{b}(\mathbf{v})$-measurable). Assume further that $\mathbf{v}$ gives weak ignorability, then so does $\boldsymbol{b}(\mathbf{v})$.*

Obviously, the *propensity score* $e(\mathbf{v}) := p(\mathbf{t}=1|\mathbf{v})$, the propensity of assigning the treatment given $\mathbf{v}$, is a balancing score (with f be the identity function). Also, given any invertible function $\boldsymbol{v}$, the composition $\boldsymbol{v} \circ \boldsymbol{b}$ is also a balancing score since $f \circ \boldsymbol{v}^{-1}(\boldsymbol{v} \circ \boldsymbol{b}(\mathbf{v})) = f(\boldsymbol{b}(\mathbf{v})) = e(\mathbf{v})$.

Compare the definition of balancing score and prognostic score, we can say balancing score is sufficient for the treatment t ($\mathrm{t} \perp\!\!\!\perp \mathbf{v}|\boldsymbol{b}(\mathbf{v})$), while prognostic score (Pt-score) is sufficient for the potential outcomes $\mathbf{y}(t)$ ($\mathbf{y}(t) \perp\!\!\!\perp \mathbf{v}|\mathbb{P}_t(\mathbf{v})$). They complement each other; conditioning on either deconfounds the potential outcomes from treatment, with the former focuses on the treatment side, the latter on the outcomes side.

B.2 VAE, Conditional VAE, and iVAE

VAEs [40] are a class of latent variable models with latent variable $\mathbf{z}$, and observable $\mathbf{y}$ is generated by the decoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{z})$. In the standard formulation [39], the variational lower bound $\mathcal{L}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})$ of the log-likelihood is derived as:

$$\begin{aligned} \log p(\mathbf{y}) &\geq \log p(\mathbf{y}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z}|\mathbf{y})) \\ &= \mathbb{E}_{\boldsymbol{z} \sim q} \log p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{y}) \| p(\mathbf{z})), \end{aligned} \tag{34}$$

where D_{KL} denotes KL divergence and the encoder $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{y})$ is introduced to approximate the true posterior $p(\mathbf{z}|\mathbf{y})$. The decoder $p_{\boldsymbol{\theta}}$ and encoder $q_{\boldsymbol{\phi}}$ are usually parametrized by NNs. We will omit the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ in notations when appropriate.

The parameters of the VAE can be learned with stochastic gradient variational Bayes. With Gaussian latent variables, the KL term of $\mathcal{L}$ has closed form, while the first term can be evaluated by drawing samples from the approximate posterior $q_{\boldsymbol{\phi}}$ using the reparameterization trick [39], then, optimizing the evidence lower bound (ELBO) $\mathbb{E}_{\mathbf{y} \sim \mathcal{D}}(\mathcal{L}(\boldsymbol{y}))$ with data $\mathcal{D}$, we train the VAE efficiently.

Conditional VAE (CVAE) [60, 41] adds a conditioning variable $\mathbf{c}$, usually a class label, to standard VAE (See Figure 1). With the conditioning variable, CVAE can give better reconstruction of each class. The variational lower bound is

$$\log p(\mathbf{y}|\mathbf{c}) \geq \mathbb{E}_{\boldsymbol{z} \sim q} \log p(\mathbf{y}|\boldsymbol{z}, \mathbf{c}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}, \mathbf{c}) \| p(\mathbf{z}|\mathbf{c})). \tag{35}$$

The conditioning on $\mathbf{c}$ in the prior is usually omitted [14], i.e., the prior becomes $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as in standard VAE, since the dependence between $\mathbf{c}$ and the latent representation is also modeled in the encoder q. Moreover, unconditional prior in fact gives better reconstruction because it encourages learning representation independent of class, similarly to the idea of beta-VAE [25].

As mentioned, *identifiable* VAE (iVAE) [37] provides the first identifiability result for VAE, using auxiliary variable $\mathbf{x}$. It assumes $\mathbf{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z}$, that is, $p(\mathbf{y}|\mathbf{z}, \mathbf{x}) = p(\mathbf{y}|\mathbf{z})$. The variational lower bound is

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}) &\geq \log p(\mathbf{y}|\mathbf{x}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}, \mathbf{x}) \| p(\mathbf{z}|\mathbf{y}, \mathbf{x})) \\ &= \mathbb{E}_{\boldsymbol{z} \sim q} \log p_{\boldsymbol{f}}(\mathbf{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{y}, \mathbf{x}) \| p_{\boldsymbol{T}, \boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})), \end{aligned} \tag{36}$$

where $\mathbf{y} = \boldsymbol{f}(\mathbf{z}) + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}$ is additive noise, and $\mathbf{z}$ has exponential family distribution with sufficient statistics $\boldsymbol{T}$ and parameter $\boldsymbol{\lambda}(\mathbf{x})$. Note that, unlike CVAE, the decoder does *not* depend on $\mathbf{x}$ due to the independence assumption.

Here, *identifiability of the model* means that the functional *parameters* $(\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ can be identified (learned) up to certain simple transformation. Further, in the limit of $\boldsymbol{\epsilon} \to \mathbf{0}$, iVAE solves the nonlinear ICA problem of recovering $\mathbf{z} = \boldsymbol{f}^{-1}(\mathbf{y})$.

C Expositions

The order of subsections below follows that they are referred in the main text.

C.1 Discussions and examples of (G2)

We focus on univariate outcome on $\mathbb{R}$ which is the most practical case and the intuitions apply to more general types of outcomes. Then, $\boldsymbol{i}$, the mapping between μ_0 and μ_1, is monotone, i.e, either increasing or decreasing. The increasing $\boldsymbol{i}$ means, if a change of the value of $\mathbf{x}$ increases (decreases) the outcome in the treatment group, then it is also the case for the controlled group. This is often true because the treatment does *not* change the mechanism how the covariates affect the outcome, under the principle of "independence of causal mechanisms (ICM)" [31]. The decreasing $\boldsymbol{i}$ corresponds to another common interpretation when ICM does not hold. Now, the treatment does change the way covariates affect $\mathbf{y}$, but in a *global* manner: it acts like a "switch" on the mechanism: the same change of $\mathbf{x}$ always has *opposite* effects on the two treatment groups.

We support the above reasoning by real world examples. First we give two examples where μ_0 and μ_1 are both monotone increasing. This, and also that both μ_t are monotone decreasing, are natural and sufficient conditions for increasing $\boldsymbol{i}$, though not necessary. The first example is form Health. [63] mentions that gestational age (length of pregnancy) has a monotone increasing effect on babies' birth weight, regardless of many other covariates. Thus, if we intervene on one of the other binary covariates (say, t = receive healthcare program or not), both μ_t should be monotone increasing in gestational age. The next example is from economics. [18] shows that job-matching probability is monotone increasing in market size. Then, we can imagine that, with t = receive training in job finding or not, the monotonicity is not changed. Intuitively, the examples corresponds to two common scenarios: the causal effects are accumulated though time (the first example), or the link between a covariate and the outcome is direct and/or strong (the second example).

Examples for decreasing $\boldsymbol{i}$ are rarer and the following is a bit deliberate. This example is also about babies' birth weight as the outcome. [1] shows that, with t = mother smokes or not and $\mathbf{x}$ = mother's age, the CATE $\tau(\boldsymbol{x})$ is monotone decreasing for $20 < \boldsymbol{x} < 26$ (smoking decreases birth weight, and the absolute causal effect is larger for older mother). On the other hand, it is shown that birth weight slightly increases (by about 100g) in the same age range in a surveyed population [73]. Thus, it is convince that, smoking changes the the tendency of birth weight w.r.t mother's age from increasing to decreasing, and gives the large decreasing of birth weight (by about 300g) as its causal effect. This could be understood: the negative effects of smoking on mother's heath and in turn on birth weight are accumulated during the many years of smoking.

C.2 Complementarity between the two identifications

We examine the complementarity between the two identifications more closely. The conditions **(M3) / (M3')** and **(G2) / (D2)** form two pairs, and are complementary inside each pair. The first pair matches model and truth, while the second pair restricts the discrepancy between the treatment groups. In Theorem 1, **(G2)** ($\mathbb{P}_0 = \mathbb{P}_1$) is replaced by **(D2)** which instead makes $\mathcal{A}_0 = \mathcal{A}_1 := \mathcal{A}$ in (6). And **(D2)** is easily satisfied with high-dimensional $\mathbf{x}$, even if the possible values of $\boldsymbol{C}, \boldsymbol{d}$ are restricted to $\boldsymbol{C} = c\boldsymbol{I}$ and $\boldsymbol{d} = \mathbf{0}$ (see below). On the other hand, $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ in **(M3')** is impractical, but it ensures that $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = p(\mathbf{y}|\mathbf{x}, \mathrm{t})$ so that (6) can be used. In Sec. 4.1, we consider practical estimation method and introduce the *regularization* that encourages learning a PtS similar to PS so that $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ can be relaxed.

(D2) is general despite (or thanks to) the involved formulation. Let us see its generality even under a highly special case: $\boldsymbol{C} = c\boldsymbol{I}$ and $\boldsymbol{d} = \mathbf{0}$. Then, $\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = c\boldsymbol{I}$ requires that, $\boldsymbol{h}_1(\boldsymbol{x}_k) - c\boldsymbol{h}_0(\boldsymbol{x}_k)$ is the same for $2n+1$ points $\boldsymbol{x}_k$. This is easily satisfied except for $n \gg m$ where m is the dimension of $\mathbf{x}$, which *rarely* happens in practice. And, $\boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{d}$ becomes just $\boldsymbol{\beta}_1 = c\boldsymbol{\beta}_0$. This is equivalent to $\alpha_1(\boldsymbol{x}_k) - c\alpha_0(\boldsymbol{x}_k)$ same for $2n+1$ points, again fine in practice. However, the high generality comes with price. Verifying **(D2)** using data is challenging, particularly with high-dimensional covariate and latent variable. Although we believe fast algorithms for this purpose could be developed, the effort would be nontrivial. This is another motivation to use the extreme case $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$, which corresponds to $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{d} = \mathbf{0}$.

C.3 Ideas and connections behind the ELBO (8)

Bayesian approach is favorable to express the prior belief that balanced PtSs exist and the preference for them, and to still have reasonable posterior estimation when the belief fails and learning general PtS is necessary. This is the causal importance of VAE as an estimation method for us. By the unconditional but still flexible $\boldsymbol{\lambda}$, and also the identifications, the ELBO encourages the discovery of an equivalent DGP with a balanced PtS and the recovery of it as the posterior, which still learns the dependence on t if necessary. Moreover, β expresses our additional knowledge (or, inductive bias) about whether or not there exist balanced PtSs (e.g., from domain expertise).

In fact, β connects our VAE to β-VAE [25], which is closely related to noise and variance control [14, Sec. 2.4][49].

Considerations on noise modeling. In Theorem 1, with large and mismatched *noises* (then **(M3')** is easily violated), the identification of outcome model $\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}$ would fail, and, in turn, the prior would learn confounding bias, by confusing the causal effect of t on $\mathbb{P}_{\mathrm{t}}$ and the correlation between t and $\mathbf{x}$. This is another reason to prefer $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$, besides balancing. On the other hand, the posterior conditioning on $\mathbf{y}$ provides information of noise $\mathbf{e}$, and it is shown in [5] that posterior effect estimation has *minimum worst-case error* under model misspecification (of the noise and prior, in our case).

Under large $\mathbf{e}$, a relatively small β implicitly encourages $\boldsymbol{g}$ *smaller* than the scale of $\mathbf{e}$, through stressing the third term in ELBO (8). And the the model as a whole would still learn $p(\mathbf{y}|\mathbf{x}, \mathrm{t})$ well, because the randomness of $\mathbf{e}$ can be moved to and modeled by the prior. This is why $\boldsymbol{k}$ is *not* set to zero because learnable prior noise (variance) allows us to implicitly control $\boldsymbol{g}$ via β. Intuitively, smaller $\boldsymbol{g}$ strengthens the correlation between $\mathbf{y}$ and $\mathbf{z}$ in our model, and this naturally reflects that posterior conditioning on $\mathbf{y}$ is more important under larger $\mathbf{e}$. Hopefully, precise learning of outcome noise **(M3')** is not required, as in Proposition 2.

Now, it is clear that β naturally controls at the same time noise scale and balancing. And the regularization can also be understood as an interpolation between Proposition 2 and Theorem 1: relying on PS, or on model identifiability; learning loosely, or precisely, the outcome regression. When the noise scale is different from truth, there would be error due to imperfect recovery of $\boldsymbol{j}$. Sec. 4.2 shows that this error and balancing form a trade-off, which is adjusted by β.

Importance of balancing from misspecification view. If we must learn an unbalanced PtS, we have larger misspecification under a balanced prior and rely more on $\mathbf{y}$ in the posterior. Both are bad because it is shown in [5] that posterior only helps under bounded (small) misspecification, and posterior estimator has higher variance than prior estimator (see below for an extreme case). Again, we want a regularizer to encourage learning of PS, so that we can explore the *middle ground*: relatively low-dimensional $\mathbb{P}$, or relatively small $\mathbf{e}$.

Example. Assume the true outcome noise is (near) zero. By setting $\boldsymbol{\epsilon} \to \mathbf{0}$ in our model, the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) = p_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}|\mathbf{x}, \mathrm{t})/p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t})$ degenerates to $\boldsymbol{f}_{\mathrm{t}}^{-1}(\mathbf{y}) = \boldsymbol{f}_{\mathrm{t}}^{-1}(\boldsymbol{j}_{\mathrm{t}}(\mathbb{P}_{\mathrm{t}})) = \boldsymbol{v}^{-1}(\mathbb{P}_{\mathrm{t}})$, a *factual* PtS. However, $\boldsymbol{f}_{1-\mathrm{t}}^{-1}(\mathbf{y}) = \boldsymbol{f}_{1-\mathrm{t}}^{-1}(\boldsymbol{j}_{\mathrm{t}}(\mathbb{P}_{\mathrm{t}})) = \boldsymbol{v}^{-1}(\boldsymbol{j}_{1-\mathrm{t}}^{-1} \circ \boldsymbol{j}_{\mathrm{t}}(\mathbb{P}_{\mathrm{t}})) \neq \boldsymbol{v}^{-1}(\mathbb{P}_{1-\mathrm{t}})$, *the score recovered by posterior does not work for counterfactual assignment*! The problem is, unlike $\mathbf{x}$, the outcome $\mathbf{y} = \mathbf{y}(\mathrm{t})$ is affected by t, and, the degenerated posterior disregards the information of $\mathbf{x}$ from the prior and depends exclusively on factual $(\mathbf{y}, \mathrm{t})$.

C.4 Consistency of VAE and prior estimation

The following is a refined version of Theorem 4 in [37]. The result is proved by assuming: i) our VAE is flexible enough to ensure the ELBO is tight (equals to the true log likelihood) for some parameters; ii) the optimization algorithm can achieve the *global* maximum of ELBO (again equals to the log likelihood).

Proposition 6 (Consistency of Intact-VAE). *Given model* (4)&(7)*, and let* $p^*(\mathbf{x}, \mathbf{y}, \mathrm{t})$ *be the true observational distribution, assume*

i) there exists $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}})$ *such that* $p_{\bar{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = p^*(\mathbf{y}|\mathbf{x}, \mathrm{t})$ *and* $p_{\bar{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) = q_{\bar{\boldsymbol{\phi}}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})$*;*

ii) the ELBO $\mathbb{E}_{\mathcal{D} \sim p^*}(\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathrm{t}; \boldsymbol{\theta}, \boldsymbol{\phi}))$ (5) *can be optimized to its global maximum at* $(\boldsymbol{\theta}', \boldsymbol{\phi}')$*;*

Then, in the limit of infinite data, $p_{\boldsymbol{\theta}'}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = p^*(\mathbf{y}|\mathbf{x}, \mathrm{t})$ *and* $p_{\boldsymbol{\theta}'}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) = q_{\boldsymbol{\phi}'}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})$.

Proof. From i), we have $\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathrm{t}; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}) = \log p^*(\mathbf{y}|\mathbf{x}, \mathrm{t})$. But we know $\mathcal{L}$ is upper-bounded by $\log p^*(\mathbf{y}|\mathbf{x}, \mathrm{t})$. So, $\mathbb{E}_{\mathcal{D}\sim p^*}(\log p^*(\mathbf{y}|\mathbf{x}, \mathrm{t}))$ should be the global maximum of the ELBO (even if the data is finite).

Moreover, note that, for any $(\boldsymbol{\theta}, \boldsymbol{\phi})$, we have $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \| q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \geq 0$ and, in the limit of infinite data, $\mathbb{E}_{\mathcal{D}\sim p^*}(\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t})) \leq \mathbb{E}_{\mathcal{D}\sim p^*}(\log p^*(\mathbf{y}|\mathbf{x}, \mathrm{t}))$. Thus, the global maximum of ELBO is achieved *only* when $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = p^*(\mathbf{y}|\mathbf{x}, \mathrm{t})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) = q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})$. □

Consistent prior estimation of CATE follows directly from the identifications. The following is a corollary of Theorem 1.

Corollary 1. *Under the conditions of Theorem 1, further require the consistency of Intact-VAE. Then, in the limit of infinite data, we have* $\mu_t(\boldsymbol{x}) = \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))$ *where* $\boldsymbol{f}, \boldsymbol{h}$ *are the optimal parameters learned by the VAE.*

C.5 Pre / Post-treatment prediction

Sampling posterior requires *post-treatment* observation $(\boldsymbol{y}, t)$. Often, it is desirable that we can also have *pre-treatment* prediction for a new subject, with only the observation of its covariate $\mathbf{x} = \boldsymbol{x}$. To this end, we use prior as a pre-treatment predictor for $\mathbf{z}$: replace $q_{\boldsymbol{\phi}}$ with $p_{\boldsymbol{\lambda}}$ in (9) and all the others remain the same. We also have sensible pre-treatment prediction even without true low-dimensional PSs, because $p_{\boldsymbol{\lambda}}$ gives the best balanced approximation of the target PtS. The results of pre-treatment prediction are given in the experimental section below.

C.6 Novelties of the bounds in Sec. 4.2

We summarize the novelties of our bounds compared to those in [58, 47]. Most importantly, our bounds and balancing are *conditional* on $\mathbf{x}$. The previous works are based on bound and balancing among the whole population, and thus *overfit* the PEHE error, a population version of the CATE error (See Experiments, particularly Sec. 6.2). Focusing on VAE, our method strengthens [47], in a simpler and principled way: we distinguish true score and latent $\mathbf{z}$ and show that identification is the link; considering both prior and posterior, we show the symmetric nature of the balancing term and relate it to our KL term in (8), without ad hoc regularization; moreover, we consider outcome noise modeling which is a strength of VAE and relate it to hyperparameter β. Particularly, in [47], latent variable $\mathbf{z}$ is confused with the true representation ($\mathbb{P}_t$ up to invertible mapping in our case). *Without* identification, the method in fact has unbounded error.

C.7 Prior / Posterior CATE error as surrogates of the truth

Note that, $\epsilon^*_{\boldsymbol{f}} = \epsilon_{\boldsymbol{f}}$ if $\tau(\mathbf{x}) = \tau_m(\mathbf{z}), \mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$, and we have $\mathbf{z}_t = \mathbb{P}_t(\mathbf{x}) \implies \tau(\mathbf{x}) = \tau_m(\mathbf{z}_t), \mathbf{z}_t \sim p_t(\mathbf{z}|\mathbf{x})$ under the *recovery of scores* in Sec. 3.2 (the invertible $\boldsymbol{v}$ is omitted; replace $\mathbb{P}_t = \boldsymbol{z}$ with $\mathbb{P}_t = \boldsymbol{v}(\boldsymbol{z})$ in the definitions, and others remain the same). Thus, we have $\tau(\mathbf{x}) = \tau_m(\mathbf{z})$ if $\mathbb{P}_t$ is a PS. Generally, if $\mathbb{P}_t$ is well balanced and recovered, the error between $\tau(\mathbf{x})$ and $\tau_m(\mathbf{z})$ is expected to be small and, thus, is not considered in Sec. 4.2. Instead, by bounding $\epsilon^p_{\boldsymbol{f}}$ (or, $\epsilon^q_{\boldsymbol{f}}$ for posterior), we consider the error between $\hat{\tau}_{\boldsymbol{f}}$ and τ_m, *due to the unknown outcome noise*, which is not accounted by our Theorem 1.

D Other related work

D.1 Injectivity, invertibility, monotonicity, and overlap

Let us note that *any injective mapping defines an invertible mapping*, by restrict the domain of the inverse function to the range of the injective mapping. Also note that injectivity is weaker than monotonicity; a monotone mapping can be defined by an injective and *order-preserving* mapping between ordered sets. Particularly, *an injective and continuous mapping on* $\mathbb{R}$ *is monotone*, and many works in econometrics give examples of this case.

Many classical and recent works (with many real world applications, see C.1) in econometrics are based on monotonicity. Particularly, there is a long line of work based on *monotonicity of treatment* [29]. More related to our method is another line of work based on *monotonicity of outcome*, see

[8] and references therein for early results. Some recent works apply monotonicity of outcome to nonparametric IV regression (NPIV) [17, 45, 10], where the structural equation of the outcome is assumed to be $\mathrm{y} = f(\mathrm{t}) + \epsilon$, and f is monotone and t (the treatment) is often continuous. Particularly, [10] combines monotonicity of both treatment and outcome, and [17] considers *discrete* treatment (note continuity or differentiability is not necessary for monotonicity). NPIV with monotone f is closely related to our method, but the difference is that t is replaced by a PtS in our method, and the PtS is recovered from observables. Finally, as we mentioned in Sec. 3.2, monotonicity is a kind of shape restriction which also includes, e.g., concavity and symmetry and attracts recent interests [9]. However, most of NPIV works focus on identifying f but not directly on TEs, and we do not know any works that use monotonicity to address weak overlap.

Recently in machine learning, [35, 80, 34] note the relationship between invertibility and overlap. As mentioned, [34] gives bounds without overlap, but the relationship between invertibility and overlap is not explicit in their theory. [35] explicitly discuss overlap and invertibility, but does not focus on TEs. [80] assumes overlap so that identification is given, and then focuses on learning overlapped representation that preserves the overlap of the covariate. However, it does not relate invertibility and overlap, but uses invertible representation function to *preserve exchangeability given the covariate*, and linear outcome regression to simply the model. Related, our identifications required **(M2)**, of which linearity of PtS and representation function is a sufficient condition, and our outcome model is injective, to *preserve the exchangeability given the PtS*. Thus, our method works under more general setting, and arguably under weaker conditions.

D.2 VAEs for TE estimation

VAEs are suitable for causal estimation thanks to its probabilistic nature. However, most VAE methods for TEs, e.g. [46, 79, 71, 47], add ad hoc heuristics into their VAEs, and thus break down probabilistic modeling, not to mention identifiable representation. Moreover, the methods rely on learning sufficient representations from *proxy* variables, leading to either impractical assumptions or conceptual inconsistency, in causal identification.

On identification. First, as to causal identification, [46] assumes unobserved confounder can be recovered, which is rarely possible even under further structural assumptions [68], and [52] recently gives evidence that the method often fails. Other methods [79, 71, 47] assume unconfoundedness but still rely on proxy at least intuitively; particularly, [47] factorizes the decoder as in the proxy setting. However, *unconfoundedness and proxy should not be put together*. The conceptual inconsistency is that, by definition, unconfoundedness means covariates *fully* control confounding, while the motivation for proxy is that unconfoundedness is often *not* satisfied in practice and covariates are at best proxies of confounding, which are non-confounders causally connected to confounders [68]. Second, without identifiable representation, the empirical results of the methods lacks solid ground; under settings not covered by their experiments, the methods would silently fail to learn proper representations, as we show in Sec. 6.1.

On ad hoc heuristics. Ad hoc heuristics break down probabilistic modeling and / or give ELBOs that do not estimate the probabilistic models. For example, [46] uses separated NNs for the two POs to mimic TARnet [58]. And, to have pre-treatment estimation, $q(\mathrm{t}|\mathbf{x})$ and $q(\mathbf{y}|\mathbf{x}, \mathrm{t})$ are added into the encoder. As a result, the ELBO of [46] has two additional likelihood terms corresponding to the two distributions. [79] is even more ad hoc because it splits the latent variable $\mathbf{z}$ into three components, and applies the ad hoc tricks of [46] to each of the component. Particularly, when constructing the encoder, [79] implicitly assumes the three components of $\mathbf{z}$ are conditional independent give $\mathbf{x}$, which violates the intended graphical model.

Our method is motivated by the important concept of PGS, and is naturally based on (2). As a consequence, our VAE architecture is a natural combination of iVAE and CVAE (see Figure 1). Our ELBO (5) is derived by standard variational lower bound. Moreover, in our Intact-VAE, pre-treatment prediction is given naturally by our conditional prior, thanks to the correspondence between our model and (2).

E Details and additions of experiments

E.1 Synthetic data

We detail how the random parameters in the DGPs are sampled. μ_i and σ_i are uniformly sampled in range $(-0.2, 0.2)$ and $(0, 0.2)$, respectively. The weights of linear functions $\boldsymbol{h}, \boldsymbol{k}, l$ are sampled from standard normal distributions. The NNs f_0, f_1 use leaky ReLU activation with $\alpha = 0.5$ and are of 3 to 8 layers randomly, and the weights of each layer are sampled from $(-1.1, -0.9)$. To have a large but still reasonable outcome variance, the output of f_t is divided by $C_t := \mathrm{Var}_{\{\mathcal{D}|\mathrm{t}=t\}}(f_t(\mathbf{z}))$. When generating DGPs with dependent noise, the variance parameter for the outcome is generated by adding a softplus layer after respective f_t, and then normalized to range $(0, 2)$.

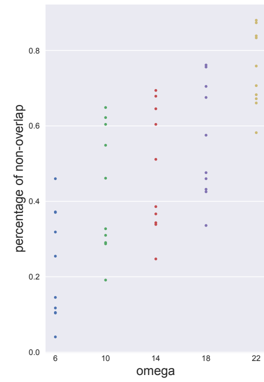

Figure 4: Degree of weak overlap w.r.t ω.

We use the original implementation of CFR[6]. Very possibly due to bugs in implementation, the CFR version using Wasserstein distance has error of TensorFlow type mismatch on our synthetic dataset, and the CFR version using MMD diverges with very large loss value often on one or two of the 10 random DGPs. We use MMD version, and, when the divergence of training happens, report the results from trained models before divergence, which still give reasonable results. We search the balancing parameter alpha in [0.16, 0.32, 0.64, 0.8, 1.28], and fix other hyperparameters as they were in the default config file.

We characterize the degree of weak overlap by examining the percentage of observed values $\boldsymbol{x}$ that give probability less than 0.001 for one of $p(t|\boldsymbol{x})$. The threshold is chosen so that all sample points near those values $\boldsymbol{x}$ almost certainly belong to a single group since we have 500 sample point in total. If we regard a DGP as very weakly overlapped when the above percentage is larger than 50%, then, as shown in Figure 4, non (all) of the 10 DGPs are very weakly overlapped with $\omega = 6$ ($\omega = 22$).

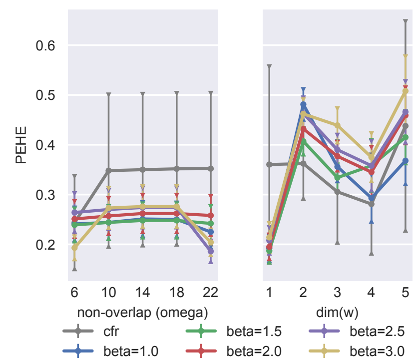

Figure 5: $\sqrt{\epsilon_{PEHE}}$ on synthetic dataset with *dependent noise*. Error bar on 10 random DGPs.

Figure 5 shows the importance of noise modeling under DGP of dependent noise. Compared to Figure 2 in the main text, our method works better here, particularly for large β, while CFR works worse. In the left panel, notably, we see our method is better than CFR even with only 1-dimensional $\mathbf{z}$. Interestingly, learning $\boldsymbol{g}$ in the model (the results of which are not shown) does not improve performance even under this setting, this might imply that learning $\boldsymbol{k}$ in the prior is enough, and the VAE can focus more on balancing with $\boldsymbol{g}$ fixed (see also the exposition C.3 on the ELBO).

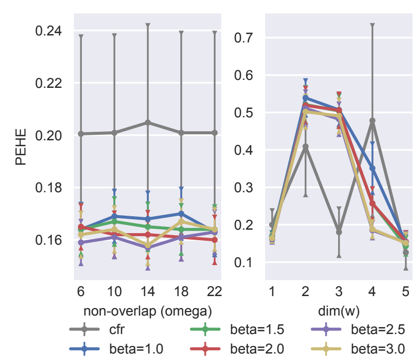

Figure 6: $\sqrt{\epsilon_{PEHE}}$ on synthetic dataset, with $\dim(\mathbf{z}) = 200$ in our model. Error bar on 10 random DGPs.

Figure 6 shows, with $\dim(\mathbf{z}) = 200$, our method works better than CFR under $\dim(\mathbf{w}) = 1$ and as well as CFR under $\dim(\mathbf{w}) > 1$. As mentioned in Conclusion, this indicates that the theoretical requirement of injective $\boldsymbol{f}_t$ in our model might be relaxed. Interestingly, larger β seems to give better results here, this is understandable because β controls the trade-off between fitting and balancing, and the fitting capacity of our decoder is much increased with $\dim(\mathbf{z}) = 200$.

Figure 7 shows results of ATE estimation. Notably, CFR drops performance w.r.t degree of weak overlap. Our method does not show this tendency except for very large β ($\beta = 3$). This might be another evidence that CFR and its unconditional balancing overfit to PEHE (see Sec. 6.2). Also note that, under $\dim(\mathbf{w}) = 1$, $\beta = 3$ gives the best results for ATE although it does not work well for PEHE, and we do not know if this generalizes to the conclusion that large β gives better ATE estimation under the existence of PS, but leave this for future investigation.

[6] `https://github.com/clinicalml/cfrnet`

Figure 8 shows results of pre-treatment prediction. In left panel, both our method and CFR perform only slightly worse than post-treatment. This is reasonable because here we have PS $\mathbf{w}$ with $\dim(\mathbf{w}) = 1$, there is no need to learn PtS. In the right panel, we also do not see significant drop of performance compared to post-treatment. This might be due to the hardness of learning balanced PtS in this dataset, and posterior estimation does not give much improvements.

You can find more plots for latent recovery at the end of the paper.

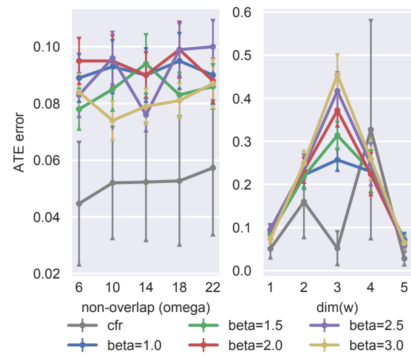

Figure 7: ϵ_{ATE} on synthetic dataset. Error bar on 10 random DGPs.

E.2 IHDP

IHDP is based on an RCT where each data point represents a child with 25 features (6 continuous, 19 binary) about their birth and mothers. `Race` is introduced as a confounder by artificially removing all treated children with nonwhite mothers. There are 747 subjects left in the dataset. The outcome is synthesized by taking the covariates (features excluding `Race`) as input, hence *unconfoundedness* holds given the covariates. Following previous work, we split the dataset by 63:27:10 for training, validation, and testing. Note, there is no ethical concerns here, because the treatment assignment mechanism is artificial by processing the data. Also our results are only quantitative and we make no ethical conclusions.

The generating process is as following [26, Sec. 4.1].

$$\mathrm{y}(0) \sim \mathcal{N}(e^{\boldsymbol{a}^T(\mathbf{x}+\boldsymbol{b})}, 1), \quad \mathrm{y}(1) \sim \mathcal{N}(\boldsymbol{a}^T\mathbf{x} - c, 1), \tag{37}$$

where $\boldsymbol{a}$ is a random coefficient, $\boldsymbol{b}$ is a constant bias with all elements equal to 0.5, and c is a random parameter adjusting degree of overlapping between the treatment groups. As we can see, $\boldsymbol{a}^T\mathbf{x}$ is a true PS. As mentioned in the main text, the PS might be discrete. Thus, this experiment also shows the importance of VAE, even if an apparent PS exists. Under *discrete* PSs, training an regression based on Proposition 2 is hard, but our VAE works well.

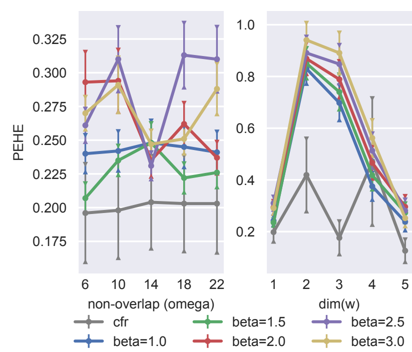

Figure 8: *Pre-treatment* $\sqrt{\epsilon_{PEHE}}$ on synthetic dataset. Error bar on 10 random DGPs.

The two added components in the modified version of our method are as following. First, we build the two outcome functions $\boldsymbol{f}_t(\mathbf{z}), t = 0, 1$ in our learning model (4), using two separate NNs. Second, we add to our ELBO (5) a regularization term, which is the Wasserstein distance [11] between $\mathbb{E}_{\mathcal{D}\sim p(\mathbf{x}|\mathrm{t}=t)} p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}), t \in \{0, 1\}$. As shown in Table 2, best unconditional balancing parameter is 0.1, the results of which is reported in the main text. Larger parameters gives much worse PEHE and does not improve ATE estimation. Smaller parameters are more reasonable but still do not improve the results. The overall tendency is clear. Compared to ours, CFR with its unconditional balancing does not improve ATE estimation, it may improve PEHE results with fine tuned parameter, but possibly at the price of worse ATE estimation.

Table 2: Performance of modified version with different unconditional balancing parameter, the values of which are shown after "Mod.".

Method	Ours	Mod. 1	Mod. 0.2	Mod. 0.1	Mod. 0.05	Mod. 0.01	CFR
ϵ_{ATE}	.178$\pm$.006	.196$\pm$.008	.177$\pm$.007	.167$\pm$.005	.177$\pm$.006	.179$\pm$.006	.25$\pm$.01
$\sqrt{\epsilon_{PEHE}}$	.859$\pm$.033	1.979$\pm$.082	1.116$\pm$.046	.777$\pm$.026	.894$\pm$.039	.841$\pm$.029	.71$\pm$.02

Table 3 shows pre-treatment results, All methods gives reasonable results.

E.3 Pokec Social Network Dataset

This experiment shows our method is the best compared with the methods specialized for networked deconfounding, a challenging problem in its own right. Thus, our method has the potential to work under *unobserved confounding*, but we leave detailed experimental and theoretical investigation to future.

Table 3: *Pre-treatment* Errors on IHDP over 1000 random DGPs. We report results with $\dim(\mathbf{z}) = 10$. **Bold** indicates method(s) that are *significantly* better. The results are taken from [58], except GANITE [78] and CEVAE [46].

Method	TMLE	BNN	CFR	CF	CEVAE	GANITE	Ours
pre- ϵ_{ATE}	NA	$.42_{\pm.03}$	$.27_{\pm.01}$	$.40_{\pm.03}$	$.46_{\pm.02}$	$.49_{\pm.05}$	$\mathbf{.211}_{\pm.011}$
pre- $\sqrt{\epsilon_{PEHE}}$	NA	$2.1_{\pm.1}$	$\mathbf{.76}_{\pm.02}$	$3.8_{\pm.2}$	$2.6_{\pm.1}$	$2.4_{\pm.4}$	$.946_{\pm.048}$

Pokec [43] is a real world social network dataset. We experiment on a semi-synthetic dataset based on Pokec, which was introduced in [70], and use exactly the same pre-processing and generating procedure. The pre-processed network has about 79,000 vertexes (users) connected by 1.3 $\times 10^6$ undirected edges. The subset of users used here are restricted to three living districts that are within the same region. The network structure is expressed by binary adjacency matrix $\boldsymbol{G}$. Following [70], we split the users into 10 folds, test on each fold and report the mean and std of pre-treatment ATE predictions. We further separate the rest of users (in the other 9 folds) by $6:3$, for training and validation.

Each user has 12 attributes, among which `district`, `age`, or `join date` is used as a confounder u to build 3 different datasets, with remaining 11 attributes used as covariate $\mathbf{x}$. Treatment t and outcome $\mathbf{y}$ are synthesised as following:

$$\mathrm{t} \sim \mathrm{Bern}(g(\mathrm{u})), \quad \mathbf{y} = \mathrm{t} + 10(g(\mathrm{u}) - 0.5) + \epsilon, \tag{38}$$

where ϵ is standard normal. Note that `district` is of 3 categories; `age` and `join date` are also discretized into three bins. $g(\mathrm{u})$, which is a PS, maps these three categories and values to $\{0.15, 0.5, 0.85\}$.

Intact-VAE is expected to learn a PS from $\boldsymbol{G}, \mathbf{x}$, if we can exploit the network structure effectively. Given the huge network structure, most users can practically be identified by their attributes and neighborhood structure, which means u can be roughly seen as a deterministic function of $\boldsymbol{G}, \mathbf{x}$. This idea is comparable to Assumptions 2 and 4 in [70], which postulate directly that a balancing score can be learned in the limit of infinite large network. To extract information from the network structure, we use Graph Convolutional Network (GCN) [42] in conditional prior and encoder of Intact-VAE. The implementation details are given at the end of this subsection.

Table 4 shows the results. The pre-treatment $\sqrt{\epsilon_{PEHE}}$ for `Age`, `District`, and `Join date` confounders are 1.085, 0.686, and 0.699 respectively, practically the same as the ATE errors. Note that, [70] does not give individual-level prediction.

Table 4: Pre-treatment ATE on Pokec. Ground truth ATE is 1, as we can see in (38). "Unadjusted" estimates ATE by $\mathbb{E}_{\mathcal{D}}(y_1) - \mathbb{E}_{\mathcal{D}}(y_0)$. "Parametric" is a stochastic block model for networked data [19]. "Embed-" denotes the best alternatives given by [70]. **Bold** indicates method(s) that are *significantly* better than all the others. We report results with 20-dimensional latent $\mathbf{z}$. The results of the other methods are taken from [70].

	`Age`	`District`	`Join Date`
Unadjusted	4.34 ± 0.05	4.51 ± 0.05	4.03 ± 0.06
Parametric	4.06 ± 0.01	3.22 ± 0.01	3.73 ± 0.01
Embedding-Reg.	2.77 ± 0.35	**1.75** ± 0.20	2.41 ± 0.45
Embedding-IPW	3.12 ± 0.06	**1.66** ± 0.07	3.10 ± 0.07
Ours	**2.08** ± 0.32	**1.68** ± 0.10	**1.70** ± 0.13

To extract information from the network structure, we use Graph Convolutional Network (GCN) [42] in conditional prior and encoder of Intact-VAE. A difficulty is that, the network $\boldsymbol{G}$ and covariates $\boldsymbol{X}$ of *all* users are always needed by GCN, regardless of whether it is in training, validation, or testing phase. However, the separation can still make sense if we take care that the treatment and outcome are used only in the respective phase, e.g., (y_m, t_m) of a testing user m is only used in testing.

GCN takes the network matrix $\boldsymbol{G}$ and the *whole* covariates matrix $\boldsymbol{X} := (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_M^T)^T$, where M is user number, and outputs a representation matrix $\boldsymbol{R}$, again for all users. During training, we *select* the rows in $\boldsymbol{R}$ that correspond to users in training set. Then, treat this *training representation matrix* as if it is the covariates matrix for a non-networked dataset, that is, the downstream networks in conditional prior and encoder are the same as in the other two experiments, but take $(\boldsymbol{R}_{m,:})^T$ where $\boldsymbol{x}_m$ was expected as input. And we have respective selection operations for validation and testing.

We can still train Intact-VAE including GCN by Adam, simply setting the gradients of non-seleted rows of $\boldsymbol{R}$ to 0.

Note that GCN cannot be trained using mini-batch, instead, we perform batch gradient decent using full dataset for each iteration, with initial learning rate 10^{-2}. We use dropout [62] with rate 0.1 to prevent overfitting.

E.4 Additional plots on synthetic datasets

See next pages.

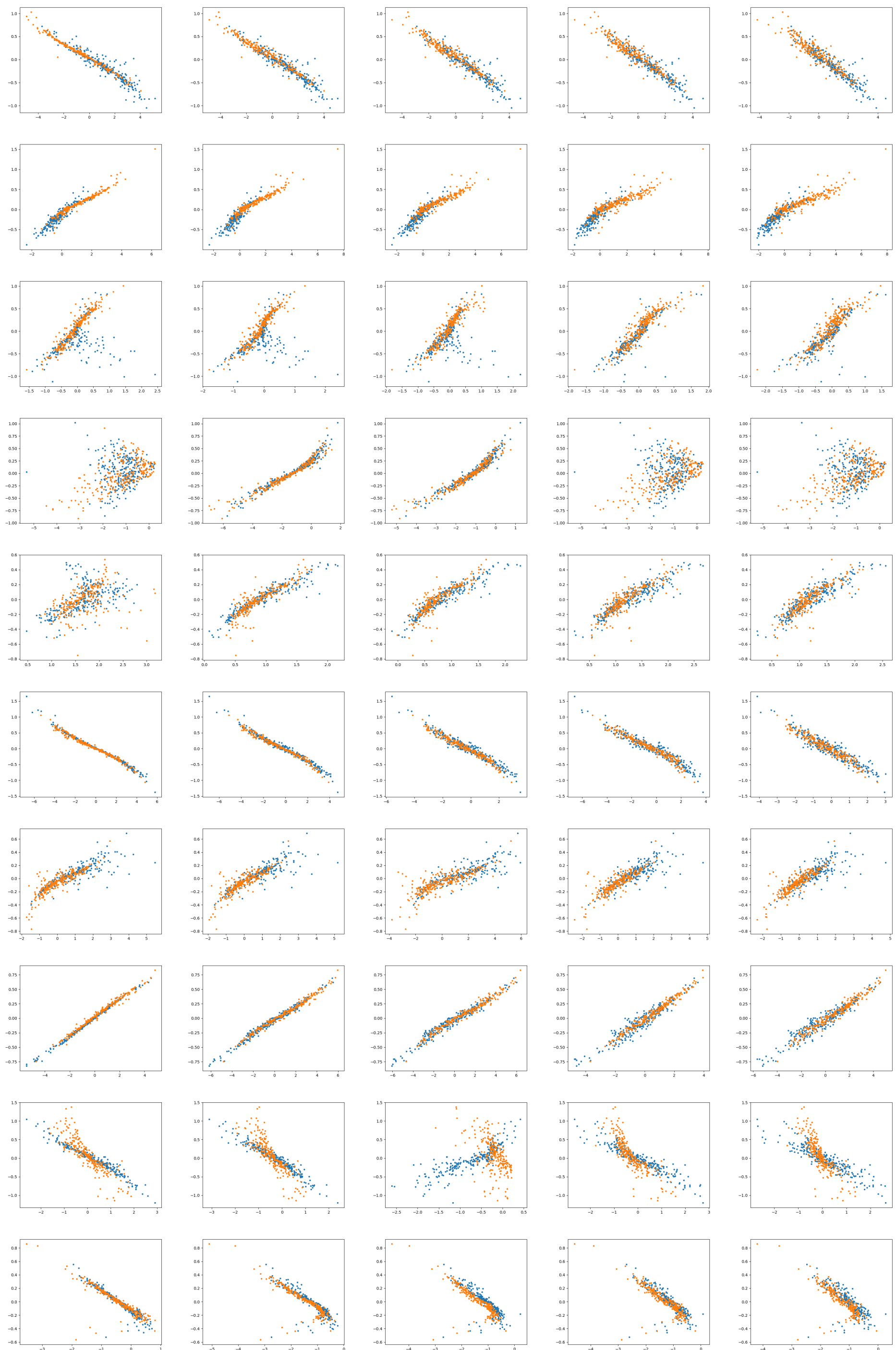

Figure 9: Plots of recovered-true latent. Rows: first 10 nonlinear random models, columns: outcome noise level.

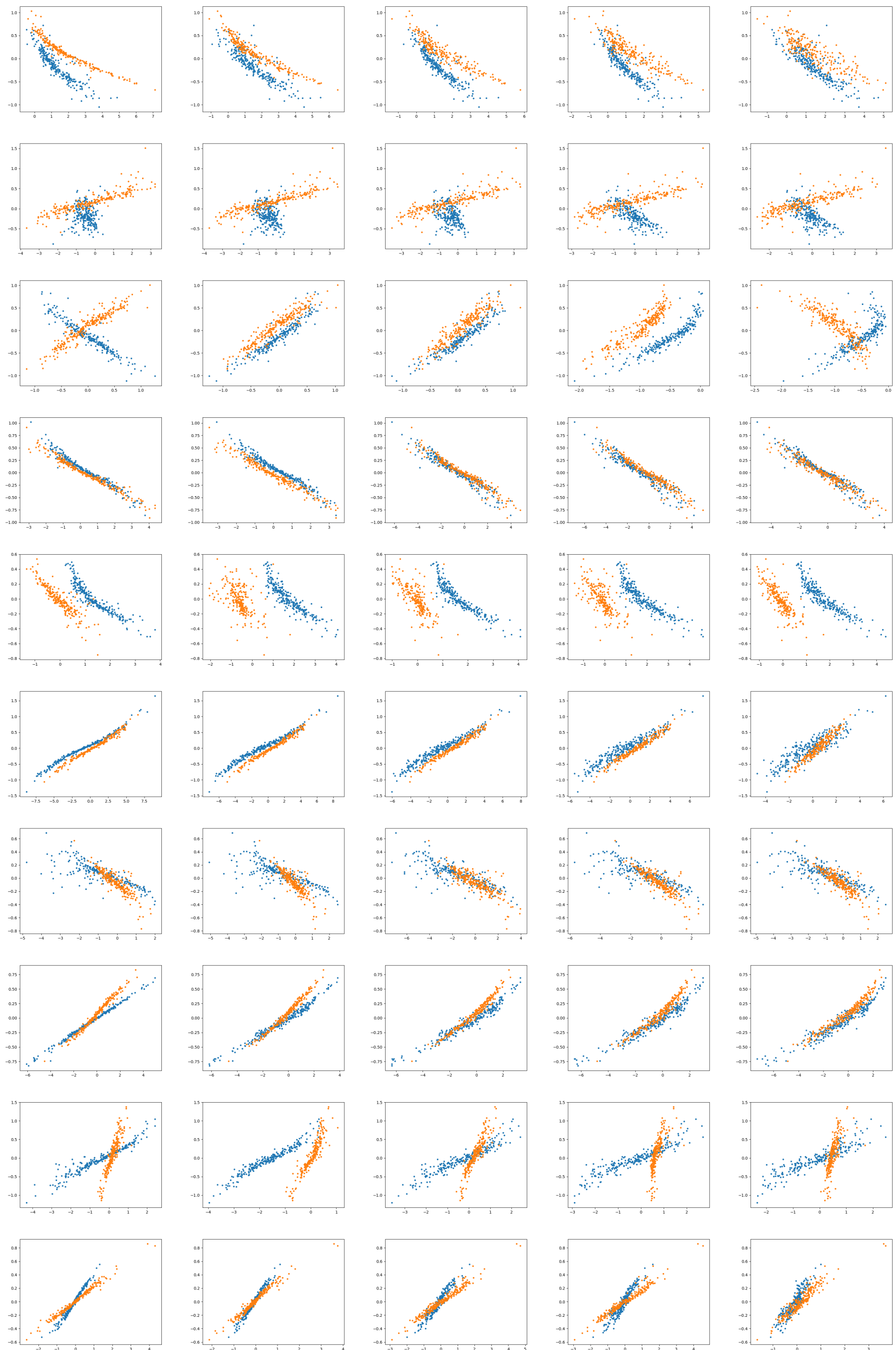

Figure 10: Plots of recovered-true latent. Conditional prior *depends* on t. Rows: first 10 nonlinear random models, columns: outcome noise level. Compare to the previous figure, we can see the transformations for $t = 0, 1$ are *not* the same, confirming the importance of balanced prior.

References

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[3] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[4] Timothy B. Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *arXiv preprint arXiv:1712.04594v5*, 2021.

[5] Stéphane Bonhomme and Martin Weidner. Posterior average effects. *arXiv preprint arXiv:1906.06360v5*, 2021.

[6] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[7] Alberto Caron, Ioanna Manolopoulou, and Gianluca Baio. Estimating individual treatment effects using non-parametric regression models: a review. *arXiv preprint arXiv:2009.06472*, 2020.

[8] Victor Chernozhukov and Christian Hansen. Quantile models with endogeneity. *Annu. Rev. Econ.*, 5(1):57–81, 2013.

[9] Denis Chetverikov, Andres Santos, and Azeem M Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10:31–63, 2018.

[10] Denis Chetverikov and Daniel Wilhelm. Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320, 2017.

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[12] Wangzhi Dai and Collin M Stultz. Quantifying common support between multiple treatment groups using a contrastive-vae. In *Machine Learning for Health*, pages 41–52. PMLR, 2020.

[13] Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

[14] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[15] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 2020.

[16] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[17] Joachim Freyberger and Joel L Horowitz. Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, 189(1):41–53, 2015.

[18] Li Gan and Qi Li. Efficiency of thin and thick markets. *Journal of Econometrics*, 192(1):40–54, 2016.

[19] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[20] P Richard Hahn, Jared S Murray, Carlos M Carvalho, et al. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[21] David Hajage, Yann De Rycke, Guillaume Chauvet, and Florence Tubach. Estimation of conditional and marginal odds ratios using the prognostic score. *Statistics in medicine*, 36(4):687–716, 2017.

[22] Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

[23] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[24] Miguel A. Hernan and James M. Robins. *Causal Inference: What If*. CRC Press, 1st edition, 2020.

[25] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

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

[27] Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):32–47, 2020.

[28] Ming-Yueh Huang and Kwun Chuen Gary Chan. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 104(3):583–596, 2017.

[29] Martin Huber and Kaspar Wüthrich. Local average and quantile treatment effects under endogeneity: a review. *Journal of Econometric Methods*, 8(1), 2018.

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

[31] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[32] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

[33] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[34] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[35] Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

[36] Nathan Kallus, Brenton Pennicooke, and Michele Santacatterina. More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *arXiv preprint arXiv:1811.04274*, 2018.

[37] 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, 2020.

[38] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33, 2020.

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[40] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[41] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[42] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*, 2017.

[43] Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection, 2014.

[44] Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

[45] Zheng Li, Guannan Liu, and Qi Li. Nonparametric knn estimation with monotone constraints. *Econometric Reviews*, 36(6-9):988–1006, 2017.

[46] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[47] Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. Reconsidering generative objectives for counterfactual reasoning. *Advances in Neural Information Processing Systems*, 33, 2020.

[48] Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

[49] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019.

[50] Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

[51] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

[52] Severi Rissanen and Pekka Marttinen. A critical look at the identifiability of causal effects with deep latent variable models. *arXiv preprint arXiv:2102.06648*, 2021.

[53] Paul R Rosenbaum. Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7:143–176, 2020.

[54] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[55] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[56] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[57] Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *arXiv preprint arXiv:2012.09935*, 2020.

[58] 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.

[59] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517, 2019.

[60] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[61] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2019.

[62] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[63] Jennifer E Starling, Catherine E Aiken, Jared S Murray, Annettee Nakimuli, and James G Scott. Monotone function estimation in the presence of extreme data coarsening: Analysis of preeclampsia and birth weight in urban uganda. *arXiv preprint arXiv:1912.06946*, 2019.

[64] Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010.

[65] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S84–S90, 2013.

[66] Xinwei Sun, Botong Wu, Chang Liu, Xiangyu Zheng, Wei Chen, Tao Qin, and Tie-yan Liu. Latent causal invariant model. *arXiv preprint arXiv:2011.02203*, 2020.

[67] Alexander Tarr and Kosuke Imai. Estimating average treatment effects with support vector machines. *arXiv preprint arXiv:2102.11926*, 2021.

[68] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

[69] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[70] Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. In *Advances in Neural Information Processing Systems*, pages 13792–13802, 2019.

[71] Matthew James Vowels, Necati Cihan Camgoz, and Richard Bowden. Targeted vae: Structured inference and targeted learning for causal parameter estimation. *arXiv preprint arXiv:2009.13472*, 2020.

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

[73] Shanshan Wang, Liren Yang, Li Shang, Wenfang Yang, Cuifang Qi, Liyan Huang, Guilan Xie, Ruiqi Wang, and Mei Chun Chung. Changing trends of birth weight with maternal age: a cross-sectional study in xi'an city of northwestern china. *BMC Pregnancy and Childbirth*, 20(1):1–8, 2020.

[74] Halbert White and Karim Chalak. Identification and identification failure for treatment effects using structural systems. *Econometric Reviews*, 32(3):273–317, 2013.

[75] 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.

[76] S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018.

[77] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[78] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[79] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. *arXiv preprint arXiv:2001.10652*, 2020.

[80] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.