# $\beta$-INTACT-VAE: IDENTIFYING AND ESTIMATING CAUSAL EFFECTS UNDER LIMITED OVERLAP

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As an important problem in causal inference, we discuss the identification and estimation of treatment effects (TEs) under limited overlap; that is, when subjects with certain features belong to a single treatment group. We use a latent variable to model a prognostic score which is widely used in biostatistics and sufficient for TEs; i.e., we build a generative prognostic model. We prove that the latent variable recovers a prognostic score, and the model identifies individualized treatment effects. The model is then learned as $\beta$-Intact-VAE—a new type of variational autoencoder (VAE). We derive the TE error bounds that enable representations balanced for treatment groups conditioned on individualized features. The proposed method is compared with recent methods using (semi-)synthetic datasets.

## 1 INTRODUCTION

Causal inference (Imbens & Rubin, 2015; Pearl, 2009), i.e, inferring causal effects of interventions, is a fundamental field of research. In this work, we focus on treatment effects (TEs) based on a set of observations comprising binary labels $T$ for treatment/control (non-treated), outcome $Y$, and other covariates $X$. Typical examples include estimating the effects of public policies or new drugs based on the personal records of the subjects. The fundamental difficulty of causal inference is that we never observe *counterfactual* outcomes that would have been if we had made the other decision (treatment or control). While randomized controlled trials (RCTs) control biases through randomization and are ideal protocols for causal inference, they often have ethical and practical issues, or suffer from expensive costs. Thus, causal inference from observational data is important.

Causal inference from observational data has other challenges as well. One is *confounding*: there may be variables, called confounders, that causally affect both the treatment and the outcome, and spurious correlation/bias follows. The other is the systematic *imbalance* (difference) of the distributions of the covariates between the treatment and control groups—that is, $X$ depends on $T$, which introduces bias in estimation. A majority of studies on causal inference, including the current work, have relied on unconfoundedness; this means that appropriate covariates are collected so that the confounding can be controlled by conditioning on the covariates. However, such high-dimensional covariates tend to introduce a stronger imbalance between treatment and control.

The current work studies the issue of imbalance in estimating individualized TEs conditioned on $X$. Classical approaches aim for *covariate balance*, $X$ independent of $T$, by matching and re-weighting (Stuart, 2010; Rosenbaum, 2020). Machine learning methods have also been exploited; there are semi-parametric methods—e.g., Van der Laan & Rose (2018, TMLE)—which improve finite sample performance, as well as non-parametric methods—e.g., Wager & Athey (2018, CF). Notably, from Johansson et al. (2016), there has been a recent increase in interest in *balanced representation learning* (BRL) to learn representations $Z$ of the covariates, such that $Z$ independent of $T$.

The most serious form of imbalance is the *limited (or weak) overlap of covariates*, which means that sample points with certain covariate values belong to a single treatment group. In this case, a straightforward estimation of TEs is not possible at non-overlapping covariate values due to lack of data. Some works have focused on providing robustness to limited overlap (Armstrong & Kolesár, 2021), trimming non-overlapping sample points (Yang & Ding, 2018), or studying convergence rates based on overlap (Hong et al., 2020). Limited overlap is particularly relevant to machine learning methods that exploit high-dimensional covariates. This is because, with higher-dimensional covariates, overlap is harder to satisfy and verify (D'Amour et al., 2020).

To address imbalance and limited overlap, we use a prognostic score (Hansen, 2008); it is a sufficient statistic of outcome predictors and is among the key concepts of sufficient scores for TE estimation. As a function of covariates, it can map some non-overlapping values to an overlapping value in a space of lower-dimensions. For individualized TEs, we consider *conditionally balanced representation* $Z$, such that $Z$ is independent of $T$ given $X$—which, as we will see, is a necessary condition for a balanced prognostic score. Moreover, prognostic score modeling can benefit from methods in predictive analytics and exploit rich literature, particularly in medicine and health (Hajage et al., 2017). Thus, it is promising to combine the predictive power of prognostic modeling and machine learning. With this idea, our method builds on a generative prognostic model that models the prognostic score as a latent variable and factorizes to the score distribution and outcome distribution.

As we consider latent variables and causal inference, *identification* is an issue that must be discussed before estimation is considered. "Identification" means that the parameters of interest (in our case, representation function and TEs) are uniquely determined and expressed using the true observational distribution. Without identification, a consistent estimator is impossible to obtain, and a model would fail silently; in other words, the model may fit perfectly but will return an estimator that converges to a wrong one, or does not converge at all (Lewbel, 2019, particularly Sec. 8). Identification is even more important for causal inference; because, unlike usual (non-causal) model misspecification, causal assumptions are often unverifiable through observables (White & Chalak, 2013). Thus, it is critical to specify the theoretical conditions for identification, and then the applicability of the methods can be judged by knowledge of an application domain.

A major strength of our generative model is that the latent variable is identifiable. This is because the factorization of our model is naturally realized as a combination of identifiable VAE (Khemakhem et al., 2020a, iVAE) and conditional VAE (Sohn et al., 2015, CVAE). Based on model identifiability, we develop two identification results for individualized TEs under limited overlap. A similar VAE architecture was proposed in Wu & Fukumizu (2020b; 2021a); the current study is different in setting, theory, learning objective, and experiments. The previous work studies unobserved confounding but not limited overlap, with different set of assumptions and identification theories. The current study further provides bounds on individualized TE error, and the bounds justify a conditionally balancing term controlled by hyperparameter $\beta$, as an interpolation between the two identifications.

In summary, we study the identification (Sec. 3) and estimation (Sec. 4) of individualized TEs under limited overlap. Our approach is based on recovering prognostic scores from observed variables. To this end, our method exploits recent advances in identifiable representation—particularly iVAE. The code is in Supplementary Material, and the proofs are in Sec. A. Our main contributions are:
1) TE identification under limited overlap of $X$, via prognostic scores and an identifiable model;
2) bounds on individualized TE error, which justify our conditional BRL;
3) a new regularized VAE, $\beta$-Intact-VAE, realizing the identification and conditional balance;
4) experimental comparison to the state-of-the-art methods on (semi-)synthetic datasets.

### 1.1 Related work

**Limited overlap.** Under limited overlap, Luo et al. (2017) estimate the average TE (ATE) by reducing covariates to a linear prognostic score. Farrell (2015) estimates a constant TE under a partial linear outcome model. D'Amour & Franks (2021) study the identification of ATE by a general class of scores, given the (linear) propensity score and prognostic score. Machine learning studies on this topic have focused on finding overlapping regions (Oberst et al., 2020; Dai & Stultz, 2020), or indicating possible failure under limited overlap (Jesson et al., 2020), but not remedies. An exception is Johansson et al. (2020), which provides bounds under limited overlap. To the best of our knowledge, our method is the first machine learning method that provides identification under limited overlap.

**Prognostic scores** have been recently combined with machine learning approaches, mainly in the biostatistics community. For example, Huang & Chan (2017) estimate individualized TE by reducing covariates to a linear score which is a joint propensity-prognostic score. Tarr & Imai (2021) use SVM to minimize the worst-case bias due to prognostic score imbalance. However, in the machine learning community, few methods consider prognostic scores; Zhang et al. (2020a) and Hassanpour & Greiner (2019) learn outcome predictors, without mentioning prognostic score—while Johansson et al. (2020) conceptually, but not formally, connects BRL to prognostic score. Our work is the first to formally connect generative learning and prognostic scores for TE estimation.

**Identifiable representation.** Recently, independent component analysis (ICA) and representation learning—both ill-posed inverse problems—meet together to yield nonlinear ICA and identifiable representation; for example, using VAEs (Khemakhem et al., 2020a), and energy models (Khemakhem et al., 2020b). The results are exploited in causal discovery (Wu & Fukumizu, 2020a) and out-of-distribution generalization (Sun et al., 2020). This study is the first to explore identifiable representations in TE identification.

**BRL and related methods** amount to a major direction. Early BRL methods include BLR/BNN (Johansson et al., 2016) and TARnet/CFR (Shalit et al., 2017). In addition, Yao et al. (2018) exploit the local similarity between data points. Shi et al. (2019) use similar architecture to TARnet, considering the importance of treatment probability. There are also methods that use GAN (Yoon et al., 2018, GANITE) and Gaussian processes (Alaa & van der Schaar, 2017). Our method shares the idea of BRL, and further extends to conditional balance—which is natural for individualized TE.

**Others.** Our work can lay conceptual and theoretical foundations of VAE methods for TEs (e.g., Louizos et al., 2017; Lu et al., 2020). In addition, some studies consider monotonicity, which is injectivity on $\mathbb{R}$, together with overlap (Johansson et al., 2020; Zhang et al., 2020b). See Sec. D.

## 2 Setup and preliminaries

### 2.1 Counterfactuals, treatment effects, and identification

Following Imbens & Rubin (2015), we assume there exist *potential outcomes* $Y(t) \in \mathbb{R}^d, t \in \{0,1\}$. $Y(t)$ is the outcome that would have been observed if the treatment value $T = t$ was applied. We see $Y(t)$ as the hidden variables that give the *factual outcome* $Y$ under *factual assignment* $T = t$. Formally, $Y(t)$ is defined by the *consistency of counterfactuals*: $Y = Y(t)$ if $T = t$; or simply $Y = Y(T)$. The *fundamental problem of causal inference* is that, for a unit under research, we can observe only one of $Y(0)$ or $Y(1)$—w.r.t. the treatment value applied. That is, "factual" refers to $Y$ or $T$, which is *observable*; or estimators built on the observables. We also observe relevant covariate(s) $X \in \mathcal{X} \subseteq \mathbb{R}^m$, which is associated with individuals, with distribution $\mathcal{D} := (X, Y, T) \sim p(\boldsymbol{x}, \boldsymbol{y}, t)$. We use upper-case (e.g. $T$) to denote random variables, and lower-case (e.g. $t$) for realizations.

The expected potential outcome is denoted by $\mu_t(\boldsymbol{x}) = \mathbb{E}(Y(t)|X = \boldsymbol{x})$ conditioned on $X = \boldsymbol{x}$. The estimands in this work are the conditional ATE (CATE) and ATE, defined, respectively, by:

$$\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \quad \nu = \mathbb{E}(\tau(X)). \tag{1}$$

CATE is seen as an *individual-level*, personalized, treatment effect, given highly discriminative $X$.

Standard results (Rubin, 2005)(Hernan & Robins, 2020, Ch. 3) show sufficient conditions for TE identification in general settings. They are *Exchangeability*: $Y(t) \perp\!\!\!\perp T|X$, and *Overlap*: $p(t|\boldsymbol{x}) > 0$ for any $\boldsymbol{x} \in \mathcal{X}$. Both are required for $t \in \{0,1\}$. When $t$ appears in statements without quantification, we always mean "for both $t$". Often, *Consistency* is also listed; however, as mentioned, it is better known as the well-definedness of counterfactuals. Exchangeability means, just as in RCTs, but additionally given $X$, that there is no correlation between factual $T$ and potential $Y(t)$. Note that the popular assumption $Y(0), Y(1) \perp\!\!\!\perp T|X$ is stronger than $Y(t) \perp\!\!\!\perp T|X$ and is not necessary for identification (Hernan & Robins, 2020, pp. 15). Overlap means that the supports of $p(\boldsymbol{x}|t = 0)$ and $p(\boldsymbol{x}|t = 1)$ should be the same, and this ensures that there are data for $\mu_t(\boldsymbol{x})$ on any $(\boldsymbol{x}, t)$.

We rely on consistency and exchangeability, but in Sec. 3.2, will relax the condition of the overlapping covariate to allow some non-overlapping values $\boldsymbol{x}$—that is, covariate $X$ is *limited-overlapping*. In this paper, we also discuss overlapping variables other than $X$ (e.g., prognostic scores), and provide a definition for any random variable $V$ with support $\mathcal{V}$ as follows:

**Definition 1.** $V$ is *Overlapping* if $p(t|V = \boldsymbol{v}) > 0$ for any $t \in \{0,1\}, \boldsymbol{v} \in \mathcal{V}$. If the condition is violated at some value $\boldsymbol{v}$, then $\boldsymbol{v}$ is *non-overlapping* and $V$ is *limited-overlapping*.

### 2.2 Prognostic scores

Our method aims to recover a prognostic score (Hansen, 2008), adapted as a Pt-score in Definition 2. On the other hand, balancing scores (Rosenbaum & Rubin, 1983) $\boldsymbol{b}(X)$ are defined by $T \perp\!\!\!\perp X|\boldsymbol{b}(X)$, of which the propensity score $p(t = 1|X)$ is a special case. See Sec. B.1 for detail.

**Definition 2.** A *Pt-score* (PtS) is two functions $\mathbb{P}_t(X)$ $(t = 0, 1)$ such that $Y(t) \perp\!\!\!\perp X | \mathbb{P}_t(X)$. A PtS is called a *P-score* (PS) if $\mathbb{P}_0 = \mathbb{P}_1$.

Note that a PtS is by definition two functions; thus, overlapping $\mathbb{P}_t(X)$ means *both* $\mathbb{P}_0(X)$ and $\mathbb{P}_1(X)$ are overlapping. **Why not balancing scores?** While balancing scores $\boldsymbol{b}(X)$ have been widely used in causal inference, PtSs are more suitable for discussing overlap. Our purpose is to recover an overlapping score for limited-overlapping $X$. It is known that overlapping $\boldsymbol{b}(X)$ implies overlapping $X$ (D'Amour et al., 2020), which counters our purpose. In contrast, overlapping PS does not imply overlapping $\boldsymbol{b}(X)$ (see Sec. C.1 for an example). Moreover, with theoretical and experimental evidence, it is recently conjectured that PtSs maximize overlap among a class of sufficient scores, including $\boldsymbol{b}(X)$ (D'Amour & Franks, 2021). In general, Hajage et al. (2017) show that prognostic score methods perform better—or as well as—propensity score methods.

Below is a corollary of Proposition 5 in Hansen (2008); note that $\mathbb{P}_t(X)$ satisfies exchangeability.

**Proposition 1** (Identification via PtS)**.** *If $\mathbb{P}_t$ is a PtS and $Y | \mathbb{P}_{\hat{t}}(X), T \sim p_{Y|\mathbb{P}_{\hat{t}},T}(\boldsymbol{y}|P, t)$ where $\hat{t} \in \{0, 1\}$ is a counterfactual assignment, then CATE and ATE are identified, using* (1) *and*

$$\mu_{\hat{t}}(\boldsymbol{x}) = \mathbb{E}(Y(\hat{t}) | \mathbb{P}_{\hat{t}}(X), X = \boldsymbol{x}) = \mathbb{E}(Y | \mathbb{P}_{\hat{t}}(\boldsymbol{x}), T = \hat{t}) = \int p_{Y|\mathbb{P}_{\hat{t}},T}(\boldsymbol{y}|\mathbb{P}_{\hat{t}}(\boldsymbol{x}), \hat{t}) \boldsymbol{y} d\boldsymbol{y} \quad (2)$$

With the knowledge of $\mathbb{P}_t$ and $p_{Y|\mathbb{P}_{\hat{t}},T}$, we choose one of $\mathbb{P}_0, \mathbb{P}_1$ and set $t = \hat{t}$ in the density function, w.r.t the $\mu_{\hat{t}}$ of interest. This counterfactual assignment resolves the problem of non-overlap at $\boldsymbol{x}$. Note that a sample point with $X = \boldsymbol{x}$ may not have $T = \hat{t}$.

We mainly consider additive noise models for $Y(t)$, which ensures the existence of PtSs.

**(G1)**[1] (Additive noise model) the data generating process (DGP) for $Y$ is $Y = \boldsymbol{f}^*(\mathbb{M}(X), T) + \mathbf{e}$ where $\boldsymbol{f}^*, \mathbb{M}$ are functions, and $\mathbf{e}$ denotes a zero-mean exogenous (external) noise.

The potential outcomes can be defined by the DGP as $Y(t) := \boldsymbol{f}^*(\mathbb{M}(X), t) + \mathbf{e}$. The DGP also specifies how other variables causally affect $Y$. For example, $X$ affects $Y$ through $\mathbb{M}$; and thus $\mathbb{M}(X)$ is the effect modifier (Hansen, 2008)—which is often components of $X$ affecting $Y$ directly. Additive noise models are used in nonparametric regression methods for TEs (Caron et al., 2020).

Under **(G1)**, we can find natural examples of PS and PtS. 1) $\mathbb{P}_t(X) := \boldsymbol{f}_t^*(\mathbb{M}(X)) = \mu_t(X)$[2] is a PtS and not PS; 2) $\mathbb{M}$ is a PS ($X$ is a trivial PS); and 3) $\mathbb{P}(X) := (\mu_0(X), \mu_1(X))$ is a PS.

We use PS and PtS to construct a representation for CATE estimation, and their balance is important. Obviously, a PS $\mathbb{P}(X)$ is a conditionally balanced representation (defined as $Z \perp\!\!\!\perp T | X$ in Introduction), because $\mathbb{P}(X)$ does not depend on $T$ given $X$. Note that a PtS is conditionally balanced if and only if it is a PS. Thus, we introduce the notion of *balanced PtS* in a non-rigorous way: a PtS $\mathbb{P}_t$ is called balanced if the value of a measure for the conditional independence $\mathbb{P}_T(X) \perp\!\!\!\perp T | X$ is small.

## 3 Identification under generative prognostic model

In Sec. 3.1, we specify the generative prognostic model $p(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{x}, t)$, and show its identifiability. In Sec. 3.2, we prove the identification of CATEs, which is one of our main contributions. The theoretical analysis involves only our generative model (i.e., prior and decoder), but not the encoder. The encoder is not part of the generative model and is involved as an approximate posterior in the estimation, which is studied in Sec. 4.

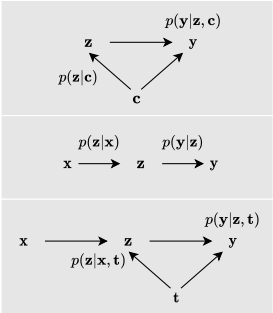


Figure 1: CVAE, iVAE, and Intact-VAE: Graphical models of the decoders.

### 3.1 Model, architecture, and identifiability

We present the necessary definitions and results, and defer more explanations to Sec. C.2. The contents of this subsection are essentially taken from Wu & Fukumizu (2021b), but included here for completeness.

Our goal is to build a model that can be learned by VAE from observational data to obtain a PtS, or better, a PS, via the latent variable $Z$. The generative prognostic model

[1] The symbols **G**, **M**, and **D** in the labels of conditions stand for Generating process, Model, and Data.

[2] We often write $t$ of the function argument in subscripts, indicating possible counterfactual assignments.

of the proposed method is as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}, t) = p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}, t) p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t),$$
$$p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}, t) = p_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z})), \quad p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t) \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{h}_t(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{k}_t(\boldsymbol{x}))). \tag{3}$$

The factor $p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}, t)$ is our decoder, which models $p_{Y|\mathbb{P}_t,T}(\boldsymbol{y}|P, t)$ in (2); and $p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t)$ is the conditional prior, which models $\mathbb{P}_T(X)$. The outcome assumes an additive noise model such that $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}$ denotes the noise model. The prior is a factorized Gaussian, where $\boldsymbol{\lambda}_T(X) := \mathrm{diag}^{-1}(\boldsymbol{k}_T(X))(\boldsymbol{h}_T(X), -\frac{1}{2})^T$ is the natural parameter as in the exponential family. $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}) = (\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{k})$ contains the functional parameters. We denote $n := \dim(Z)$.

For inference, the standard argument derives the ELBO:

$$\log p(\boldsymbol{y}|\boldsymbol{x}, t) \geq \mathbb{E}_{\boldsymbol{z} \sim q} \log p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}, t) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) \| p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t)). \tag{4}$$

Note that the encoder $q$ conditions on all the observables $(X, Y, T)$; this fact plays an important role in Section 4.1. This architecture is called *Intact-VAE* (*I*den*t*ifiable *t*rea*t*ment-con*d*i*t*ional VAE). See Figure 1 for comparison in terms of graphical models. See Sec. B.2 for basics of VAEs.

Our model identifiability extends the theory of iVAE, and the following conditions are inherited.

- **(M1)** i) $\boldsymbol{f}_t$ is injective, and ii) $\boldsymbol{f}_t$ is differentiable.
- **(D1)** $\boldsymbol{\lambda}_t(X)$ is non-degenerate, i.e., the linear hull of its support is $2n$-dimensional.

Under **(M1)** and **(D1)**, we obtain the following identifiability of the parameters in the model: if $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p_{\boldsymbol{\theta}'}(\boldsymbol{y}|\boldsymbol{x}, t)$, we have, for any $\boldsymbol{y}_t$ in the image of $\boldsymbol{f}_t$:

$$\boldsymbol{f}_t^{-1}(\boldsymbol{y}_t) = \mathrm{diag}(\boldsymbol{a}) \boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t) + \boldsymbol{b} =: \mathcal{A}_t(\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t)) \tag{5}$$

where $\mathrm{diag}(\boldsymbol{a})$ is an invertible $n$-diagonal matrix and $\boldsymbol{b}$ is an $n$-vector, both of which depend on $\boldsymbol{\lambda}_t(\boldsymbol{x})$ and $\boldsymbol{\lambda}'_t(\boldsymbol{x})$. The essence of the result is that $\boldsymbol{f}'_t = \boldsymbol{f}_t \circ \mathcal{A}_t$; that is, $\boldsymbol{f}_t$ can be identified (learned) up to an affine transformation $\mathcal{A}_t$. See Sec. A for the proof and a relaxation of **(D1)**. In this paper, symbol $'$ (prime) always indicates another parameter (variable, etc.): $\boldsymbol{\theta}' = (\boldsymbol{f}', \boldsymbol{\lambda}')$.

### 3.2 Identifications under limited-overlapping covariate

In this subsection, we present two results of CATE identification based on the recovery of equivalent PS and PtS, respectively. Since PtSs are functions of $X$, the theory assumes a noiseless prior for simplification, i.e., $\boldsymbol{k}(X) = \boldsymbol{0}$; the prior $Z_{\boldsymbol{\lambda},t} \sim p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t)$ degenerates to function $\boldsymbol{h}_t(\boldsymbol{x})$.

PtSs with dimensionality lower than or equal to $d = \dim(Y)$ are essential to address limited overlapping, as shown below. We set $n = d$ because $\mu_t$ is a PtS of the same dimension as $Y$ under **(G1)**. In practice, $n = d$ means that we seek a low-dimensional representation of $X$. **(G2)** makes the dimensionality explicit and reduces to **(G1)** if the only possibility is that $\mathbb{P}_t = \mu_t$ and $\boldsymbol{j}_t$ is identity.

- **(G2)** (Low-dimensional PtS) Under **(G1)**, $\mu_t(X) = \boldsymbol{j}_t(\mathbb{P}_t(X))$ for some $\mathbb{P}_t$ and injective $\boldsymbol{j}_t$.

We use **(G2)** instead of **(G1)** hereafter. Clearly, $\mathbb{P}_t$ in **(G2)** is PtS. In addition, injectivity and $n = d$ ensure that $n = \dim(Y) \geq \dim(\mathbb{P}_t)$. Similarly, **(G2')** ensures that $n \geq \dim(\mathbb{P})$ for PS.

- **(G2')** (Low-dimensional PS) Under **(G1)**, $\mu_t(X) = \boldsymbol{j}_t(\mathbb{P}(X))$ for some $\mathbb{P}$ and injective $\boldsymbol{j}_t$.

**(G2')** means that CATEs are given by $\mu_0$ and an invertible function $\boldsymbol{i} := \boldsymbol{j}_1 \circ \boldsymbol{j}_0^{-1}$. See Sec. C.3 for more discussions and real-world examples.

With **(G2)** or **(G2')**, overlapping $X$ can be relaxed to overlapping PS or PtS plus the following:

- **(M2)** (Score partition preserving) For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, if $\mathbb{P}_t(\boldsymbol{x}) = \mathbb{P}_t(\boldsymbol{x}')$, then $\boldsymbol{h}_t(\boldsymbol{x}) = \boldsymbol{h}_t(\boldsymbol{x}')$.

Note that **(M2)** is only required for optimal $\boldsymbol{h}$ that satisfies $\mathbb{E}_{p_{\boldsymbol{\theta}}}(Y|X, T) = \mathbb{E}(Y|X, T)$ in Proposition 2, or $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t)$ in Theorem 1. The intuition is that $\mathbb{P}_t$ maps non-overlapping $\boldsymbol{x}$ to an overlapping value, and $\boldsymbol{h}_t$ preserves this property through learning. Linear $\mathbb{P}_t$ and $\boldsymbol{h}_t$ imply **(M2)** and are often assumed, e.g., in Huang & Chan (2017); Luo et al. (2017); D'Amour & Franks (2021). Linear outcome models (Farrell, 2015; Schuler et al., 2020) are also common.

Our first identification, Proposition 2, relies on **(G2')** and our generative model, *without* model identifiability (so differentiable $\boldsymbol{f}_t$ is not needed).

**Proposition 2** (Identification via recovery of PS)**.** *Suppose we have DGP **(G2')** and model* (3) *with* $n = d$*. Assume **(M1)**-i) and* **(M3)** (PS matching) $\boldsymbol{h}_0(X) = \boldsymbol{h}_1(X)$ and $\boldsymbol{k}(X) = \mathbf{0}$*. Then, if* $\mathbb{E}_{p_{\boldsymbol{\theta}}}(Y|X,T) = \mathbb{E}(Y|X,T)$*, we have*

1) *(Recovery of PS)* $\boldsymbol{z}_{\boldsymbol{\lambda},t} = \boldsymbol{h}_t(\boldsymbol{x}) = \boldsymbol{v}(\mathbb{P}(\boldsymbol{x}))$ *on overlapping* $\boldsymbol{x}$*, where* $\boldsymbol{v} : \mathcal{P} \to \mathbb{R}^n$ *is an injective function, and* $\mathcal{P} := \{\mathbb{P}(\boldsymbol{x}) | \text{overlapping } \boldsymbol{x}\}$*.*
2) *(CATE Identification) if* $\mathbb{P}$ *in **(G2')** is overlapping, and **(M2)** is satisfied, then* $\mu_t(\boldsymbol{x}) = \hat{\mu}_t(\boldsymbol{x})$ *for any* $t \in \{0,1\}$ *and* $\boldsymbol{x} \in \mathcal{X}$*, where* $\hat{\mu}_t(\boldsymbol{x}) := \mathbb{E}_{p_{\boldsymbol{\lambda}}(Z|\boldsymbol{x},t)}\mathbb{E}_{p_{\boldsymbol{f}}}(Y|Z,t) = \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))$*.*

In essence, i) the true DGP is identified up to an invertible mapping $\boldsymbol{v}$, such that $\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}$ and $\boldsymbol{h}_t = \boldsymbol{v} \circ \mathbb{P}_t$; and ii) $\mathbb{P}_t$ is recovered up to $\boldsymbol{v}$, and $Y(t) \perp\!\!\!\perp X | \mathbb{P}_t(X)$ is preserved—with *same* $\boldsymbol{v}$ for both $t$. Theorem 1 below also achieves the essence i) and ii), under $\mathbb{P}_0 \neq \mathbb{P}_1$.

The existence of PS is more preferred, because it satisfies overlap and **(M2)** more easily than PtS which requires the conditions for each of the two functions of PtS. However, the existence of low-dimensional PS is uncertain in practice when our knowledge of the DGP is limited. Thus, we depend on Theorem 1 based on the model identifiability to work under PtS which generally exists.

**Theorem 1** (Identification via recovery of PtS)**.** *Suppose we have DGP **(G2)** and model* (3) *with* $n = d$*. For the model, assume **(M1)** and* **(M3')** (Noise matching) $p_{\mathbf{e}} = p_{\boldsymbol{\epsilon}}$ and $\boldsymbol{k}(X) = k\boldsymbol{k}'(X), k \to 0$*. Assume further that **(D1)** and* **(D2)** (Balance from data) $\mathcal{A}_0 = \mathcal{A}_1$ in (5)*. Then, if* $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},t) = p(\boldsymbol{y}|\boldsymbol{x},t)$*; conclusions 1) and 2) in Proposition 2 hold with* $\mathbb{P}$ *replaced with* $\mathbb{P}_t$ *in **(G2)**; and the domain of* $\boldsymbol{v}$ *becomes* $\mathcal{P} := \{\mathbb{P}_t(\boldsymbol{x}) | p(t,\boldsymbol{x}) > 0\}$*.*

Theorem 1 implies that, without PS, we need to know or learn the distribution of hidden noise $\boldsymbol{\epsilon}$ to have $p_{\mathbf{e}} = p_{\boldsymbol{\epsilon}}$. Proposition 2 and Theorem 1 achieve recovery and identification in a complementary manner; the former starts from the prior by $\mathbb{P}_0 = \mathbb{P}_1$ and $\boldsymbol{h}_0 = \boldsymbol{h}_1$, while the latter starts from the decoder by $\mathcal{A}_0 = \mathcal{A}_1$ and $p_{\mathbf{e}} = p_{\boldsymbol{\epsilon}}$. We see that $\mathcal{A}_0 = \mathcal{A}_1$ acts as a kind of balance because it replaces $\mathbb{P}_0 = \mathbb{P}_1$ (balanced PtS) in Proposition 2. We show in Sec. A a sufficient and necessary condition **(D2')** on data that ensures $\mathcal{A}_0 = \mathcal{A}_1$. Note that the singularities due to $k \to 0$ (e.g., $\boldsymbol{\lambda} \to \mathbf{0}$) cancel out in (5). See Sec. C.4 for more on the complementarity between the two identifications.

# 4 Estimation by $\beta$-Intact-VAE

## 4.1 Prior as PS, posterior as PtS, and $\beta$ as regularization strength

In Sec. 3.2, we see that the existence of PS (Proposition 2) is preferable in identifying the true DGP up to an equivalent expression—while Theorem 1 allows us to deal with PtS by adding other conditions. In learning our model with data, we assume that there is a PtS, and the decomposition of **(G2)** holds. However, such decompositions are not unique in general, and they are equivalent for CATE identification; and those with more balanced equivalent PtS are preferable. In this sense, we want to not only recover PtS, but also *discover* equivalent PS, if possible. This idea is common in practice. For example, in a real-world nutrition study (Huang & Chan, 2017), a reduction of 11 covariates discovers a 1-dimensional linear PS.

We consider two ways to discover and recover a equivalent PS (or balanced PtS) by a VAE. One is to use a prior which does not depend on $t$, indicating a preference for PS. Namely, we set $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1 =: \boldsymbol{\Lambda}$ and have $p_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$ as the prior in (3). The decoder and encoder are factorized Gaussians:

$$p_{\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{y}|\boldsymbol{z},t) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}_t(\boldsymbol{z}), \mathrm{diag}(\boldsymbol{g}_t(\boldsymbol{z}))), \quad q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y},t) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{r}_t(\boldsymbol{x},\boldsymbol{y}), \mathrm{diag}(\boldsymbol{s}_t(\boldsymbol{x},\boldsymbol{y}))), \quad (6)$$

where $\boldsymbol{\phi} = (\boldsymbol{r}, \boldsymbol{s})$. The other is to introduce a hyperparameter $\beta$ in the ELBO as in $\beta$-VAE (Higgins et al., 2017). The modified ELBO with $\beta$, up to the additive constant, is derived as:

$$\mathbb{E}_{\mathcal{D}}\{-\beta D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \| p_{\boldsymbol{\Lambda}}) - \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}[(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2/2\boldsymbol{g}_t^2(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}} \log |\boldsymbol{g}_t(\boldsymbol{z})|\}. \quad (7)$$

For convenience, here and in $\mathcal{L}_{\boldsymbol{f}}$ in Sec. 4.2, we omit the summation as if $Y$ is univariate. The encoder $q_{\boldsymbol{\phi}}$ depends on $t$ and can realize a PtS. With $\beta$, we control the trade-off between the first and second terms: the former is the divergence of the posterior from the balanced prior, and the latter is the reconstruction of the outcome. Note that a larger $\beta$ encourages the conditional balance $Z \perp\!\!\!\perp T | X$ on the posterior. By choosing $\beta$ appropriately, e.g., by validation, the ELBO can recover a balanced PtS while fitting the outcome well. In summary, we base the estimation on Proposition 2 and PS as much as possible, but step into Theorem 1 and noise modeling required by $p_{\mathbf{e}} = p_{\boldsymbol{\epsilon}}$ when necessary.

Note also that the parameters $\boldsymbol{g}$ and $\boldsymbol{k}$, which model the outcome noise and express the uncertainty of the prior, respectively, are both learned by the ELBO. This deviates from the theoretical conditions described in Sec. 3.2, but it is more practical and yields better results in our experiments. See Sec. C.5 for more ideas and connections behind the ELBO.

Once the VAE is learned[3] by the ELBO, the estimate of the expected potential outcomes is given by:

$$\hat{\mu}_{\hat{t}}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \boldsymbol{f}_{\hat{t}}(\boldsymbol{z}) = \mathbb{E}_{\mathcal{D}|\boldsymbol{x} \sim p(\boldsymbol{y},t|\boldsymbol{x})} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\phi}}} \boldsymbol{f}_{\hat{t}}(\boldsymbol{z}),\ \hat{t} \in \{0,1\}, \tag{8}$$

where $q(\boldsymbol{z}|\boldsymbol{x}) \coloneqq \mathbb{E}_{p(\boldsymbol{y},t|\boldsymbol{x})} q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y},t)$ is the aggregated posterior. We mainly consider the case where $\boldsymbol{x}$ is observed in the data, and the sample of $(Y,T)$ is taken from the data given $X = \boldsymbol{x}$. When $\boldsymbol{x}$ is not in the data, we replace $q_{\boldsymbol{\phi}}$ with $p_{\boldsymbol{\Lambda}}$ in (8) (see Sec. C.7 for details and E for results). Note that $\hat{t}$ in (8) indicates a counterfactual assignment that may not be the same as the factual $T = t$ in the data. That is, we set $T = \hat{t}$ in the decoder. The assignment is not applied to the encoder which is learned from factual $X, Y, T$ (see also the explanation of $\epsilon_{CF,t}$ in Sec. 4.2). The overall **algorithm** steps are i) train the VAE using (7), and ii) infer CATE $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$ by (8).

### 4.2 Conditionally balanced representation learning

We formally justify our ELBO (7) from the BRL viewpoint. We show that the conditional BRL via the first term of the ELBO results from bounding a CATE error; particularly, the error due to the imprecise recovery of $\boldsymbol{j}_t$ in **(G2)** is controlled by the ELBO. Previous works (Shalit et al., 2017; Lu et al., 2020) instead focus on unconditional balance and bound PEHE which is marginalized on $X$. Sec. 5.2 experimentally shows the advantage of our bounds and ELBO. Further, we connect the bounds to identification and consider noise modeling through $\boldsymbol{g}_t(\boldsymbol{z})$. Sec Sec. C.8 for detail.

We introduce the objective that we bound. Using (8) to estimate CATE, $\hat{\tau}_{\boldsymbol{f}}(\boldsymbol{z}) \coloneqq \boldsymbol{f}_1(\boldsymbol{z}) - \boldsymbol{f}_0(\boldsymbol{z})$ is marginalized on $q(\boldsymbol{z}|\boldsymbol{x})$. On the other hand, the *true CATE*, given the covariate $\boldsymbol{x}$ or score $\boldsymbol{z}$, is:

$$\tau(\boldsymbol{x}) = \boldsymbol{j}_1(\mathbb{P}_1(\boldsymbol{x})) - \boldsymbol{j}_0(\mathbb{P}_0(\boldsymbol{x})), \quad \tau_{\boldsymbol{j}}(\boldsymbol{z}) = \boldsymbol{j}_1(\boldsymbol{z}) - \boldsymbol{j}_0(\boldsymbol{z}), \tag{9}$$

where $\boldsymbol{j}_t$ is associated with a balanced PtS $\mathbb{P}_t$ discovered as the target of recovery by our VAE. Accordingly, given $\boldsymbol{x}$, the *error of posterior CATE*, with or without knowing $\mathbb{P}_t$, is defined as

$$\epsilon^*_{\boldsymbol{f}}(\boldsymbol{x}) \coloneqq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} (\hat{\tau}_{\boldsymbol{f}}(\boldsymbol{z}) - \tau(\boldsymbol{x}))^2; \quad \epsilon_{\boldsymbol{f}}(\boldsymbol{x}) \coloneqq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} (\hat{\tau}_{\boldsymbol{f}}(\boldsymbol{z}) - \tau_{\boldsymbol{j}}(\boldsymbol{z}))^2. \tag{10}$$

We bound $\epsilon_{\boldsymbol{f}}$ instead of $\epsilon^*_{\boldsymbol{f}}$ because the error between $\tau(X)$ and $\tau_{\boldsymbol{j}}(Z)$ is small—if the balanced $\mathbb{P}_t$ is recovered, then $\boldsymbol{z} \approx \mathbb{P}_0(\boldsymbol{x}) \approx \mathbb{P}_1(\boldsymbol{x})$ in (9). We consider the error between $\hat{\tau}_{\boldsymbol{f}}$ and $\tau_{\boldsymbol{j}}$ below. We define the risks of outcome regression, into which $\epsilon_{\boldsymbol{f}}$ is decomposed.

**Definition 3** (CATE risks)**.** Let $Y(\hat{t})|\mathbb{P}_{\hat{t}}(X) \sim p_{Y(\hat{t})|\mathbb{P}_{\hat{t}}}(\boldsymbol{y}|P)$ and $q_t(\boldsymbol{z}|\boldsymbol{x}) \coloneqq q(\boldsymbol{z}|\boldsymbol{x},t) = \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},t)} q_{\boldsymbol{\phi}}$. The *potential outcome loss* at $(\boldsymbol{z},t)$, *factual risk*, and *counterfactual risk* are:

$$\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},\hat{t}) \coloneqq \mathbb{E}_{p_{Y(\hat{t})|\mathbb{P}_{\hat{t}}}(\boldsymbol{y}|P=\boldsymbol{z})} (\boldsymbol{y} - \boldsymbol{f}_{\hat{t}}(\boldsymbol{z}))^2 / \boldsymbol{g}_{\hat{t}}(\boldsymbol{z})^2 = \boldsymbol{g}_{\hat{t}}(\boldsymbol{z})^{-2} \int (\boldsymbol{y} - \boldsymbol{f}_{\hat{t}}(\boldsymbol{z}))^2 p_{Y(\hat{t})|\mathbb{P}_{\hat{t}}}(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{y};$$
$$\epsilon_{F,t}(\boldsymbol{x}) \coloneqq \mathbb{E}_{q_t(\boldsymbol{z}|\boldsymbol{x})} \mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t); \quad \epsilon_{CF,t}(\boldsymbol{x}) \coloneqq \mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})} \mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t).$$

With $Y(t)$ involved, $\mathcal{L}_{\boldsymbol{f}}$ is a potential outcome loss on $\boldsymbol{f}$, weighted by $\boldsymbol{g}$. The factual and counterfactual counterparts, $\epsilon_{F,t}$ and $\epsilon_{CF,t}$, are defined accordingly. In $\epsilon_{F,t}$, unit $\boldsymbol{u} = (\boldsymbol{x},\boldsymbol{y},t)$ is involved in the learning of $q_t(\boldsymbol{z}|\boldsymbol{x})$, as well as in $\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t)$ since $Y(t) = \boldsymbol{y}$ for the unit. In $\epsilon_{CF,t}$, however, unit $\boldsymbol{u}' = (\boldsymbol{x},\boldsymbol{y}',1-t)$ is involved in $q_{1-t}(\boldsymbol{z}|\boldsymbol{x})$, but not in $\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t)$ since $Y(t) \neq \boldsymbol{y}' = Y(1-t)$.

Thus, *the regression error (second) term in ELBO* (7) *controls* $\epsilon_{F,t}$ *via factual data*. On the other hand, $\epsilon_{CF,t}$ is not estimable due to the unobservable $Y(1-T)$, but is bounded by $\epsilon_{F,t}$ plus $M\mathbb{D}(\boldsymbol{x})$ in Theorem 2 below—which, in turn, bounds $\epsilon_{\boldsymbol{f}}$ by decomposing it to $\epsilon_{F,t}$, $\epsilon_{CF,t}$, and $\mathbb{V}_Y$.

**Theorem 2** (CATE error bound)**.** *Assume* $|\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t)| \leq M$ *and* $|\boldsymbol{g}_t(\boldsymbol{z})| \leq G$*, then:*

$$\epsilon_{\boldsymbol{f}}(\boldsymbol{x}) \leq 2[G^2(\epsilon_{F,0}(\boldsymbol{x}) + \epsilon_{F,1}(\boldsymbol{x}) + M\mathbb{D}(\boldsymbol{x})) - \mathbb{V}_Y(\boldsymbol{x})] \tag{11}$$

*where* $\mathbb{D}(\boldsymbol{x}) \coloneqq \sum_t \sqrt{D_{\mathrm{KL}}(q_t \| q_{1-t})/2}$*, and* $\mathbb{V}_Y(\boldsymbol{x}) \coloneqq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \sum_t \mathbb{E}_{p_{Y(t)|\mathbb{P}_t}(\boldsymbol{y}|\boldsymbol{z})} (\boldsymbol{y} - \boldsymbol{j}_t(\boldsymbol{z}))^2$.

[3] As usual, we expect the variational inference and optimization procedure to be (near) optimal; that is, consistency of VAE. *Consistent estimation* using the prior is a direct corollary of the consistent VAE. see Sec. C.6 for formal statements and proofs. Under Gaussian models, it is possible to prove the consistency of the posterior estimation, as shown in Bonhomme & Weidner (2021).

$\mathbb{D}(\boldsymbol{x})$ measures the imbalance between $q_t(Z|\boldsymbol{x})$ and is symmetric for $t$. Correspondingly, *the KL term in ELBO* (7) *is also symmetric for* $t$ *and balances* $q_t(\boldsymbol{z}|\boldsymbol{x})$ by encouraging $Z \perp\!\!\!\perp T|X$ for the posterior. $\mathbb{V}_Y(\boldsymbol{x})$ reflects the intrinsic variance in the DGP and can not be controlled.

Estimating $G, M$ is nontrivial. Instead, we rely on $\beta$ in the ELBO to weight the terms in (11). We do not need two hyperparameters since $G$ *is implicitly controlled by the third term in ELBO* (7), which is a norm constraint. $\beta$ is a trade-off between the conditional balance of learned PtS (affected by $\boldsymbol{f}_t$), and precision/effective sample size of outcome regression—and can be seen as the probabilistic counterpart of Tarr & Imai (2021) and Kallus et al. (2018).

## 5 Experiments

We compare our method with existing methods on three types of datasets. Here, we present two experiments; the remaining one on the Pokec dataset is deferred to Sec. E.3. As in previous works (Shalit et al., 2017; Louizos et al., 2017), we report the absolute error of ATE $\epsilon_{ate} := |\mathbb{E}_{\mathcal{D}}(y(1) - y(0)) - \mathbb{E}_{\mathcal{D}}\hat{\tau}(\boldsymbol{x})|$ and, as a surrogate of square CATE error $\epsilon_{cate}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}|\boldsymbol{x}}[(y(1)-y(0)) - \hat{\tau}(\boldsymbol{x})]^2$, the empirical PEHE $\epsilon_{pehe} := \mathbb{E}_{\mathcal{D}}\epsilon_{cate}(\boldsymbol{x})$ (Hill, 2011), which is the average square CATE error.

Unless otherwise indicated, for each function $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{k}, \boldsymbol{r}, \boldsymbol{s}$ in ELBO (7), we use a multilayer perceptron, with $3 * 200$ hidden units, and ReLU activations. Further, $\boldsymbol{\Lambda} = (\boldsymbol{h}, \boldsymbol{k})$ depends only on $X$. The Adam optimizer with initial learning rate $10^{-4}$ and batch size 100 is employed. All experiments use early-stopping of training by evaluating the ELBO on a validation set. More details on hyper-parameters and settings are given in each experiment.

### 5.1 Synthetic Dataset

$$W|X \sim \mathcal{N}(\boldsymbol{h}(X), \boldsymbol{k}(X));\ T|X \sim \text{Bern}(\text{Logi}(\omega l(X)));\ Y|W, T \sim \mathcal{N}(f_T(W), g_T(W)). \quad (12)$$

We generate synthetic datasets following (12). Both $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $W$ are factorized Gaussians. $\boldsymbol{\mu}, \boldsymbol{\sigma}$ are randomly sampled. The functions $\boldsymbol{h}, \boldsymbol{k}, l$ are linear. Outcome models $f_0, f_1$ are built by NNs with invertible activations. $Y$ is univariate, $\dim(X) = 30$, and $\dim(W)$ ranges from 1 to 5. $W$ is a PS, but the dimensionality is not low enough to satisfy the injectivity in **(G2')**, when $\dim(W) > 1$. We have 5 different overlap levels controlled by $\omega$ that multiplies the logit value. See Sec. E.1 for details and more results on synthetic datasets.

With the same $(\dim(W), \omega)$, we evaluate our method and CFR on 10 random DGPs, with different sets of functions $f, g, \boldsymbol{h}, \boldsymbol{k}, l$ in (12). For each DGP, we sample 1500 data points, and split them into 3 equal sets for training, validation, and testing. We show our results for different hyperparameter $\beta$. For CFR, we try different balancing parameters and present the best results (see the Appendix for detail).

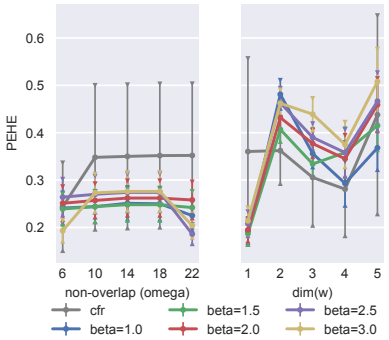


Figure 2: $\sqrt{\epsilon_{pehe}}$ on synthetic datasets. Error bar on 10 random DGPs.

In each panel of Figure 2, we adjust one of $\omega, \dim(W)$, with the other fixed to the lowest. As implied by our theory, our method, with only *1-dimensional* $Z$, performs much better in the left panel (where $\dim(W) = 1$ satisfies **(G2')**) than in the right panel (when $\dim(W) > 1$). Although CFR uses 200-dimensional representation, in the left panel our method performs much better than CFR; moreover, in the right panel CFR is not much better than ours. Further, our method is much more robust against different DGPs than CFR (see the error bars). Thus, the results indicate the power of identification and recovery of scores. (see Figure 3 also).

Under the lowest overlap level ($\omega = 22$), large $\beta (= 2.5, 3)$ shows the best results, which accords with the intuition and bounds in Sec. 4. When $\dim(W) > 1$, $f_t$ in (12) is non-injecitve and learning of PtS is necessary, and thus, larger $\beta$ has a negative effect. In fact, $\beta = 1$ is significantly better than $\beta = 3$ when $\dim(W) > 2$. We note that our method, with a higher-dimensional $Z$, outperforms or matches CFR also under $\dim(W) > 1$ (see Appendix Figure 5). Thus,

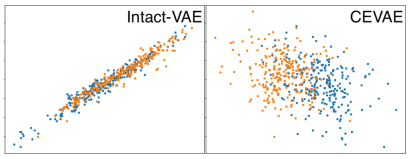


Figure 3: Plots of recovered - true latent. Blue: $t = 0$, Orange: $t = 1$.

the performance gap under $\dim(W) > 1$ in Figure 2 should be due to the capacity of NNs in $\beta$-Intact-VAE. In Appendix Figure 7 for ATE error, CFR drops performance w.r.t overlap levels. This is evidence that CFR and its unconditional balance overly focus on PEHE (see Sec. 5.2 for detail).

When $\dim(W) = 1$, there are no better PSs than $W$, because $f_t$ is invertible and no information can be dropped from $W$. Thus, our method stably learns $Z$ as an approximate affine transformation of the true $W$, showing identification. An example is shown in Figure 3, and more plots are in Appendix Figure 9. For comparison, we run CEVAE (Louizos et al., 2017), which is also based on VAE but without identification; CEVAE shows much lower quality of recovery. As expected, both recovery and estimation are better with the balanced prior $p_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$, and we can see examples of bad recovery using $p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t)$ in Appendix Figure 10 (see Appendix for more plots).

### 5.2 IHDP BENCHMARK DATASET

This experiment shows our conditional BRL matches state-of-the-art BRL methods and does not overly focus on PEHE. The IHDP (Hill, 2011) is a widely used benchmark dataset; while it is less known, its covariates are limited-overlapping, and thus it is used in Johansson et al. (2020) which considers limited overlap. The dataset is based on an RCT, but `Race` is artificially introduced as a confounder by removing all treated babies with nonwhite mothers in the data. Thus, `Race` is highly limited-overlapping, and other covariates that have high correlation to `Race`, e.g, `Birth weight` (Kelly et al., 2009), are also limited-overlapping. See Sec. E.2 for detail and more results.

There is a linear PS (linear combination of the covariates). However, most of the covariates are binary, so the support of the PS is often on small and separated intervals. Thus, the Gaussian latent $Z$ in our model is misspecified. We use 10-dimensional $Z$ to address this, similar to Louizos et al. (2017). We set $\beta = 1$ since it works well on synthetic datasets with limited overlap.

As shown in Table 1, $\beta$-Intact-VAE outperforms or matches the state-of-the-art methods. Notably, our method outperforms other generative models (CEVAE and GANITE) by large margins. Our method has the best ATE estimation and is only slightly worse than CFR for $\epsilon_{pehe}$. This fact reflects that $\epsilon_{pehe}$ is not a good criterion for CATE estimation because it is the marginalized CATE error—one expects less ATE error with overall less CATE error on individual-level, while PEHE focuses on those $X = \boldsymbol{x}$ with high probability and/or large $\epsilon_{cate}(\boldsymbol{x})$. Indeed, the unconditional balance in Shalit et al. (2017) is based on bounding PEHE, thus results in sub-optimal ATE estimation (see also Appendix Figure 7 where CFR gives larger ATE errors with less overlap).

Table 1: Errors on IHDP over 1000 random DGPs. "Mod. *" indicates the modified version with unconditional balancing hyperparameter of value "*". *Italic* indicates where the modified version is significantly worse than the original. **Bold** indicates method(s) which is significantly better than others. The results of other methods are taken from Shalit et al. (2017), GANITE (Yoon et al., 2018), and CEVAE (Louizos et al., 2017).

| Method | TMLE | BNN | CFR | CF | CEVAE | GANITE | Ours | Mod. 1 | Mod. 0.2 | Mod. 0.1 | Mod. 0.05 | Mod. 0.01 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon_{ate}$ | $.30_{\pm.01}$ | $.37_{\pm.03}$ | $.25_{\pm.01}$ | $\mathbf{.18}_{\pm.01}$ | $.34_{\pm.01}$ | $.43_{\pm.05}$ | $\mathbf{.177}_{\pm.007}$ | $\mathit{.196}_{\pm.008}$ | $\mathbf{.177}_{\pm.007}$ | $\mathbf{.167}_{\pm.005}$ | $\mathbf{.177}_{\pm.006}$ | $\mathbf{.179}_{\pm.006}$ |
| $\sqrt{\epsilon_{pehe}}$ | $5.0_{\pm.2}$ | $2.2_{\pm.1}$ | $\mathbf{.71}_{\pm.02}$ | $3.8_{\pm.2}$ | $2.7_{\pm.1}$ | $1.9_{\pm.4}$ | $.843_{\pm.030}$ | $\mathit{1.979}_{\pm.082}$ | $\mathit{1.116}_{\pm.046}$ | $.777_{\pm.026}$ | $.894_{\pm.039}$ | $.841_{\pm.029}$ |

To examine conditional v.s unconditional balance clearly, we modify our method and add two components for unconditional balance from Shalit et al. (2017) (see the Appendix), and compare the modified version to the original. The over-focus on PEHE of the unconditional balance from CFR is seen more clearly in the modified version. With different values of the hyperparameter, unconditional balance does not improve (and barely affects) ATE estimation; it does affect PEHE more significantly, but often gives worse PEHE unless the hyperparameter is fine-tuned (with value 0.1).

## 6 CONCLUSION

We proposed a method for CATE estimation under limited overlap. Our method exploits identifiable VAE, a recent advance in generative models, and is fully motivated and theoretically justified by causal considerations: identification, prognostic score, and balance. Experiments show evidence that the injectivity of $\boldsymbol{f}_t$ in our model is possibly unnecessary because $\dim(Z) > \dim(Y)$ yields better results. A theoretical study of this is an interesting future direction. We believe that VAEs are suitable for *principled* causal inference owing to their probabilistic nature, if not compromised by ad hoc heuristics (see Sec. D.2).

## REFERENCES


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

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*, pp. 3424–3432, 2017.

Timothy B Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177, 2021.

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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

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

Miguel A. Hernan and James M. Robins. *Causal Inference: What If.* CRC Press, 1st edition, 2020. ISBN 978-1-4200-7616-5.

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. URL https://openreview.net/forum?id=Sy2fzU9gl.

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

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.

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.

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

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

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

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.

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

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*, pp. 527–536. PMLR, 2019.

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.

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.

Yvonne Kelly, Lidia Panico, Mel Bartley, Michael Marmot, James Nazroo, and Amanda Sacker. Why does birthweight vary among ethnic groups in the uk? findings from the millennium cohort study. *Journal of public health*, 31(1):131–137, 2009.

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*, pp. 2207–2217, 2020a.

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, 2020b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. URL http://arxiv.org/abs/1312.6114.

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

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*, pp. 3581–3589, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

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

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

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

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*, pp. 6446–6456, 2017.

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.

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

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

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*, pp. 788–798. PMLR, 2020.

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

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.

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

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.

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

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.

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

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

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

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

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.

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.

Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010. doi: 10.1214/09-STS313. URL https://doi.org/10.1214/09-STS313.

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.

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

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.

Mark J Van der Laan and Sherri Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer, 2018.

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

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.

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.

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.

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

Pengzhou Wu and Kenji Fukumizu. Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*, pp. 1157–1167. PMLR, 2020a. URL http://proceedings.mlr.press/v108/wu20b.html.

Pengzhou Wu and Kenji Fukumizu. Intact-vae: Estimating treatment effects under unobserved confounding. *arXiv preprint arXiv:2101.06662v2*, 2021a.

Pengzhou Wu and Kenji Fukumizu. Towards principled causal effect estimation by deep identifiable models. *arXiv preprint arXiv:2109.15062*, 2021b.

Pengzhou Abel Wu and Kenji Fukumizu. Identifying treatment effects under unobserved confounding by causal representation learning. *submitted to ICLR 2021*, 2020b. URL https://openreview.net/forum?id=D3TNqCspFpM.

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. ISSN 0006-3444. doi: 10.1093/biomet/asy008. URL https://doi.org/10.1093/biomet/asy008.

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*, pp. 2633–2643, 2018.

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. URL https://openreview.net/forum?id=ByKWUeWA-.

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

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*, pp. 1005–1014. PMLR, 2020b.

# A PROOFS

We restate our model identifiability formally.

**Lemma 1** (Model identifiability)**.** *Given model* (3) *under* ***(M1)****, for* $T = t$*, assume*

**(D1')** (Non-degenerated data for $\boldsymbol{\lambda}$) there exist $2n+1$ points $\boldsymbol{x}_0, ..., \boldsymbol{x}_{2n} \in \mathcal{X}$ such that the $2n$-square matrix $\boldsymbol{L}_t := [\boldsymbol{\gamma}_{t,1}, ..., \boldsymbol{\gamma}_{t,2n}]$ is invertible, where $\boldsymbol{\gamma}_{t,k} := \boldsymbol{\lambda}_t(\boldsymbol{x}_k) - \boldsymbol{\lambda}_t(\boldsymbol{x}_0)$.

*Then, given* $T = t$*, the family is* identifiable *up to an equivalence class. That is, if* $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p_{\boldsymbol{\theta}'}(\boldsymbol{y}|\boldsymbol{x}, t)$*, we have the relation between parameters: for any* $\boldsymbol{y}_t$ *in the image of* $\boldsymbol{f}_t$*,*

$$\boldsymbol{f}_t^{-1}(\boldsymbol{y}_t) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t) + \boldsymbol{b} =: \mathcal{A}_t(\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t)) \tag{13}$$

*where* $\mathrm{diag}(\boldsymbol{a})$ *is an invertible* $n$*-diagonal matrix and* $\boldsymbol{b}$ *is a* $n$*-vector, both depend on* $\boldsymbol{\lambda}_t$ *and* $\boldsymbol{\lambda}'_t$*.*

Note, **(D1)** in the main text implies **(D1')**, see Sec. B.2.3 in Khemakhem et al. (2020a). The main part of our model identifiability is essentially the same as that of Theorem 1 in Khemakhem et al. (2020a), but now adapted to include the dependency on $t$. Here we give an outline of the proof, and the details can be easily filled by referring to Khemakhem et al. (2020a). In the proof, subscripts $t$ are omitted for convenience.

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

$$q_{\boldsymbol{f}',\boldsymbol{\lambda}'}(\boldsymbol{y}) = q_{\boldsymbol{f},\boldsymbol{\lambda}}(\boldsymbol{y}) := p_{\boldsymbol{\lambda}}(\boldsymbol{f}^{-1}(\boldsymbol{y})|\boldsymbol{x}, t) vol(\boldsymbol{J}_{\boldsymbol{f}^{-1}}(\boldsymbol{y}))\mathbb{I}_{\mathcal{Y}}(\boldsymbol{y}) \tag{14}$$

where $p_{\boldsymbol{\lambda}}$ is the Gaussian density function of the conditional prior defined in (3) and $vol(A) := \sqrt{\det AA^T}$. $q_{\boldsymbol{f}',\boldsymbol{\lambda}'}$ is defined similarly to $q_{\boldsymbol{f},\boldsymbol{\lambda}}$.

Then, apply model (3) to (14), plug the $2n+1$ points from **(D1')** into it, and re-arrange the resulting $2n+1$ equations in matrix form, we have

$$\mathcal{F}'(Y) = \mathcal{F}(Y) := \boldsymbol{L}^T\boldsymbol{t}(\boldsymbol{f}^{-1}(Y)) - \boldsymbol{\beta} \tag{15}$$

where $\boldsymbol{t}(Z) := (Z, Z^2)^T$ 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 $\alpha_t(X; \boldsymbol{\lambda}_t)$ is the log-partition function of the conditional prior in (3). $\mathcal{F}'$ is defined similarly to $\mathcal{F}$, but with $\boldsymbol{f}', \boldsymbol{\lambda}', \alpha'$

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

$$\boldsymbol{t}(\boldsymbol{f}^{-1}(Y)) = \boldsymbol{A}\boldsymbol{t}(\boldsymbol{f}'^{-1}(Y)) + \boldsymbol{c} \tag{16}$$

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 Sorrenson et al. (2019), that $\boldsymbol{A}$ is a sparse matrix such that

$$\boldsymbol{A} = \begin{pmatrix} \mathrm{diag}(\boldsymbol{a}) & \boldsymbol{O} \\ \mathrm{diag}(\boldsymbol{u}) & \mathrm{diag}(\boldsymbol{a}^2) \end{pmatrix} \tag{17}$$

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

$$\boldsymbol{f}^{-1}(Y) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{f}'^{-1}(Y) + \boldsymbol{b} \tag{18}$$

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

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

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}(Y|X, T) = \mathbb{E}(Y|X, T) \implies \boldsymbol{f}_t \circ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{j}_t \circ \mathbb{P}(\boldsymbol{x}) \text{ on } (\boldsymbol{x}, t) \text{ such that } p(t, \boldsymbol{x}) > 0. \tag{19}$$

We show the solution set of (19) on *overlapping* $\boldsymbol{x}$ is

$$\{(\boldsymbol{f}, \boldsymbol{h}) | \boldsymbol{f}_t = \boldsymbol{j}_t \circ \Delta^{-1}, \boldsymbol{h} = \Delta \circ \mathbb{P}, \Delta : \mathcal{P} \to \mathbb{R}^n \text{ is injective}\}. \tag{20}$$

By **(G2')(M1)**, and with injective $\boldsymbol{f}_t, \boldsymbol{j}_t$ and $\dim(Z) = \dim(Y) \geq \dim(\mathbb{P})$, for any $\Delta$ above, there exists a functional parameter $\boldsymbol{f}_t$ such that $\boldsymbol{j}_t = \boldsymbol{f}_t \circ \Delta$. Thus, set (20) 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 (19) should be in (20). 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 overlapping. This means the *injective* function $\boldsymbol{f}_t^{-1} \circ \boldsymbol{j}_t$ should *not* depend on $t$, thus it is one of the $\Delta$ in (20).

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

$$\hat{\mu}_t(\boldsymbol{x}) = \boldsymbol{f}_t(\boldsymbol{h}(\boldsymbol{x})) = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}(\boldsymbol{v} \circ \mathbb{P}(\boldsymbol{x})) = \boldsymbol{j}_t(\mathbb{P}(\boldsymbol{x})) = \mu_t(\boldsymbol{x}). \tag{21}$$

We rely on overlapping $\mathbb{P}$ to work for non-overlapping $\boldsymbol{x}$. 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

$$\hat{\mu}_{1-t}(\boldsymbol{x}_t) = \boldsymbol{f}_{1-t}(\boldsymbol{h}(\boldsymbol{x}_t)) = \boldsymbol{f}_{1-t}(\boldsymbol{h}(\boldsymbol{x}_{1-t})) = \boldsymbol{j}_{1-t}(\mathbb{P}(\boldsymbol{x}_{1-t})) = \boldsymbol{j}_{1-t}(\mathbb{P}(\boldsymbol{x}_t)) = \mu_{1-t}(\boldsymbol{x}_t). \tag{22}$$

The third equality uses (19) on $(\boldsymbol{x}_{1-t}, 1-t)$. □

Below we prove Theorem 1 with **(D2)** replaced by

**(D2')** *(Spontaneous balance) there exist $2n+1$ points $\boldsymbol{x}_0, ..., \boldsymbol{x}_{2n} \in \mathcal{X}$, $2n$-square matrix $\boldsymbol{C}$, and $2n$-vector $\boldsymbol{d}$, such that $\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{C}$ and $\boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{d}/k$ for optimal $\boldsymbol{\lambda}_t$ (see below), where $\boldsymbol{L}_t$ is defined in **(D1')**, $\boldsymbol{\beta}_t \coloneqq (\alpha_t(\boldsymbol{x}_1) - \alpha_t(\boldsymbol{x}_0), ..., \alpha_t(\boldsymbol{x}_{2n}) - \alpha_t(\boldsymbol{x}_0))^T$, and $\alpha_t(X; \boldsymbol{\lambda}_t)$ is the log-partition function of the prior in* (3).

**(D2')** restricts the discrepancy between $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1$ on $2n+1$ values of $X$, thus is relatively easy to satisfy with high-dimensional $X$. **(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 $\boldsymbol{d} = \boldsymbol{0}$. 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 $X$, 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 $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$ in Sec. 4.1, which corresponds to $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{d} = \boldsymbol{0}$.

*Proof of Theorem 1.* By **(M1)** and **(G2)**, 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: $p_{\boldsymbol{\theta}^*}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t)$.

Since have all assumptions for Lemma 1, we have

$$\Delta \circ \boldsymbol{j}^{-1}(\boldsymbol{y}) = \boldsymbol{f}^{*-1}(\boldsymbol{y}) = \mathcal{A} \circ \boldsymbol{f}^{-1}(\boldsymbol{y})|_t, \text{ on } (\boldsymbol{y}, t) \in \{(\boldsymbol{j}_t \circ \mathbb{P}_t(\boldsymbol{x}), t) | p(t, \boldsymbol{x}) > 0\}, \tag{23}$$

where $\boldsymbol{f}$ is *any* optimal parameter, and "$|_t$" collects all subscripts $t$. Note, except for $\Delta$, 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

$$\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{L}_0'^{-1}\boldsymbol{L}_1' = \boldsymbol{C} \iff \boldsymbol{A}_0 = \boldsymbol{A}_1 \tag{24}$$

And also,

$$\begin{gathered} \boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{C} \iff \boldsymbol{L}_0^{-T}\boldsymbol{C}^{-T} = \boldsymbol{L}_1^{-T} \\ \boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0' - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1' = \boldsymbol{d}/k \iff \boldsymbol{C}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0') = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1' \end{gathered} \tag{25}$$

Multiply right hand sides of the two lines, we have $\boldsymbol{c}_0 = \boldsymbol{c}_1$. Now we have $\mathcal{A}_0 = \mathcal{A}_1 \coloneqq \mathcal{A}$. Apply this to (23), we have

$$\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}, \quad \boldsymbol{v} \coloneqq \mathcal{A}^{-1} \circ \Delta \tag{26}$$

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

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t) \implies p_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))) = p_{\mathbf{e}}(\boldsymbol{y} - \boldsymbol{j}_t(\mathbb{P}_t(\boldsymbol{x}))) \tag{27}$$

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 $\mathbb{P}_1$. □

Note, when multiplying the two lines of (25), 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. The definitions and results work for the prior; simply *replace $q_t(\boldsymbol{x}|\boldsymbol{x})$ with $p_t(\boldsymbol{z}|\boldsymbol{x}) \coloneqq p_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}, t)$ in definitions and statements, and the proofs below hold as the same*. The dependence on $\boldsymbol{f}$ prevail, and the superscripts are omitted. The arguments $\boldsymbol{x}$ are sometimes also omitted.

**Lemma 2** (Counterfactual risk bound)**.** *Assume $|\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, t)| \leq M$, we have*

$$\epsilon_{CF}(\boldsymbol{x}) \leq \textstyle\sum_t q(1-t|\boldsymbol{x})\epsilon_{F,t}(\boldsymbol{x}) + M\mathbb{D}(\boldsymbol{x}) \tag{28}$$

*where $\epsilon_{CF}(\boldsymbol{x}) \coloneqq \sum_t p(1-t|\boldsymbol{x})\epsilon_{CF,t}(\boldsymbol{x})$, and $\mathbb{D}(\boldsymbol{x}) \coloneqq \sum_t \sqrt{D_{\mathrm{KL}}(q_t \| q_{1-t})/2}$.*

*Proof of Lemma 2.*

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

□

$\mathbb{TV}(p, q) \coloneqq \frac{1}{2}\mathbb{E}|p(\boldsymbol{z}) - q(\boldsymbol{z})|$ 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 $\mathbb{D}$.

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*

$$\epsilon_{\boldsymbol{f}} \leq 2(G^2(\epsilon_F + \epsilon_{CF}) - \mathbb{V}_Y). \tag{29}$$

*Simply bound $\epsilon_{CF}$ in (29) by Lemma 2, we have Theorem 2.* To prove Lemma 3, we first examine a bias-variance decomposition of $\epsilon_F$ and $\epsilon_{CF}$.

$$\begin{aligned}
\epsilon_{CF,t} &= \mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\boldsymbol{g}_t(\boldsymbol{z})^{-2}\mathbb{E}_{p_{Y(t)|\mathbb{P}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2 \\
&\geq G^{-2}\mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{P}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2 \\
&= G^{-2}\mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{P}_t}(\boldsymbol{y}|\boldsymbol{z})}((\boldsymbol{y} - \boldsymbol{j}_t(\boldsymbol{z}))^2 + (\boldsymbol{j}_t(\boldsymbol{z}) - \boldsymbol{f}_t(\boldsymbol{z}))^2)
\end{aligned} \tag{30}$$

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

$$\epsilon_{CF,t} \geq G^{-2}(\mathbb{V}_{CF,t}(\boldsymbol{x}) + \mathbb{B}_{CF,t}(\boldsymbol{x})) \implies \epsilon_{CF} \geq G^{-2}(\mathbb{V}_{CF}(\boldsymbol{x}) + \mathbb{B}_{CF}(\boldsymbol{x})) \tag{31}$$

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

$$\epsilon_F \geq G^{-2}(\mathbb{V}_F(\boldsymbol{x}) + \mathbb{B}_F(\boldsymbol{x})) \tag{32}$$

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

*Proof of Lemma 3.*

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

□

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

# B ADDITIONAL BACKGROUNDS

## B.1 PROGNOSTIC SCORE AND BALANCING SCORE

In the fundamental work of (Hansen, 2008), prognostic score is defined equivalently to our $\mathbb{P}_0$ (P0-score), but it in addition requires no effect modification to work for $Y(1)$. Thus, a useful prognostic score corresponds to our PtS. We give main properties of PtS as following.

**Proposition 3.** *If $V$ gives exchangeability, and $\mathbb{P}_t(V)$ is a PtS, then $Y(t) \perp\!\!\!\perp V, T | \mathbb{P}_t$.*

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

**Proposition 4** (Properties of conditional independence)**.** *(Pearl, 2009, Sec. 1.1.55) For random variables $W, X, Y, Z$. We have:*

$$\begin{aligned} X \perp\!\!\!\perp Y | Z \wedge X \perp\!\!\!\perp W | Y, Z &\implies X \perp\!\!\!\perp W, Y | Z \text{ (Contraction)}. \\ X \perp\!\!\!\perp W, Y | Z &\implies X \perp\!\!\!\perp Y | W, Z \text{ (Weak union)}. \\ X \perp\!\!\!\perp W, Y | Z &\implies X \perp\!\!\!\perp Y | Z \text{ (Decomposition)}. \end{aligned}$$

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

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

Prognostic scores are closely related to the important concept of balancing score (Rosenbaum & Rubin, 1983). Note particularly, the proposition implies $Y(t) \perp\!\!\!\perp T | \mathbb{P}_t$ (using decomposition rule). Thus, if $\mathbb{P}(V)$ 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)**.** $\boldsymbol{b}(V)$, a function of random variable $V$, is a balancing score if $T \perp\!\!\!\perp V | \boldsymbol{b}(V)$.

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

Obviously, the *propensity score* $e(V) := p(T=1|V)$, the propensity of assigning the treatment given $V$, 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 $f \circ \boldsymbol{v}^{-1}(\boldsymbol{v} \circ \boldsymbol{b}(V)) = f(\boldsymbol{b}(V)) = e(V)$.

Compare the definition of balancing score and prognostic score, we can say balancing score is sufficient for the treatment $T$ ($T \perp\!\!\!\perp V | \boldsymbol{b}(V)$), while prognostic score (Pt-score) is sufficient for the potential outcomes $Y(t)$ ($Y(t) \perp\!\!\!\perp V | \mathbb{P}_t(V)$). 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 (Kingma et al., 2019) are a class of latent variable models with latent variable $Z$, and observable $Y$ is generated by the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z})$. In the standard formulation (Kingma & Welling, 2013), the variational lower bound $\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\phi})$ of the log-likelihood is derived as:

$$\begin{aligned} \log p(\boldsymbol{y}) &\geq \log p(\boldsymbol{y}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y}) \| p(\boldsymbol{z}|\boldsymbol{y})) \\ &= \mathbb{E}_{\boldsymbol{z}\sim q} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y}) \| p(\boldsymbol{z})), \end{aligned} \tag{33}$$

where $D_{\mathrm{KL}}$ denotes KL divergence and the encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})$ is introduced to approximate the true posterior $p(\boldsymbol{z}|\boldsymbol{y})$. 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 (Kingma & Welling, 2013), then, optimizing the evidence lower bound (ELBO) $\mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}}(\mathcal{L}(\boldsymbol{y}))$ with data $\mathcal{D}$, we train the VAE efficiently.

Conditional VAE (CVAE) (Sohn et al., 2015; Kingma et al., 2014) adds a conditioning variable $C$, 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

$$\log p(\boldsymbol{y}|\boldsymbol{c}) \geq \mathbb{E}_{\boldsymbol{z}\sim q} \log p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{c})\|p(\boldsymbol{z}|\boldsymbol{c})). \tag{34}$$

The conditioning on $C$ in the prior is usually omitted (Doersch, 2016), i.e., the prior becomes $Z \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$ as in standard VAE, since the dependence between $C$ 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 (Higgins et al., 2017).

As mentioned, *identifiable* VAE (iVAE) (Khemakhem et al., 2020a) provides the first identifiability result for VAE, using auxiliary variable $X$. It assumes $Y \perp\!\!\!\perp X|Z$, that is, $p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{z})$. The variational lower bound is

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{x}) &\geq \log p(\boldsymbol{y}|\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}))\\ &= \mathbb{E}_{\boldsymbol{z}\sim q}\log p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\|p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})),\end{aligned} \tag{35}$$

where $Y = \boldsymbol{f}(Z) + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}$ is additive noise, and $Z$ has exponential family distribution with sufficient statistics $\boldsymbol{T}$ and parameter $\boldsymbol{\lambda}(X)$. Note that, unlike CVAE, the decoder does *not* depend on $X$ 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 $Z = \boldsymbol{f}^{-1}(Y)$.

# C EXPOSITIONS

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

## C.1 PROGNOSTIC SCORE IS MORE APPLICABLE THAN BALANCING SCORE

Proposition 3 in D'Amour et al. (2020) shows that overlapping balancing score implies overlapping $X$, and Footnote 5 in D'Amour et al. (2020) shows that overlapping $X$ implies overlapping PS.

Here is a simple example showing overlapping PS does not imply overlapping balancing score. Let $T = \mathbb{I}(X + \boldsymbol{\epsilon} > 0)$ and $Y = \boldsymbol{f}(|X|, T) + \mathbf{e}$, where $\mathbb{I}$ is the indicator function, $\boldsymbol{\epsilon}$ and $\mathbf{e}$ are exogenous zero-mean noises, and the support of $X$ is on the entire real line while $\boldsymbol{\epsilon}$ is bounded. Now, $X$ itself is a balancing score and $|X|$ is a PS; and $|X|$ is overlapping but $X$ is not.

## C.2 DETAILS AND EXPLANATIONS ON INTACT-VAE

Generative models are useful to solve the inverse problem of recovering Pt-score. Our goal is to build a model that can be learned by VAE from observational data to obtain a PtS, or more ideally PS, via the latent variable $Z$. That is, a generative prognostic model.

With the above goal, the generative model of our VAE is built as (3). Conditioning on $X$ in the joint model $p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x},t)$ reflects that our estimand is CATE given $X$. Modeling the score by a conditional distribution rather than a deterministic function is more flexible.

The ELBO of our model can be derived from standard variational lower bound as following:

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

We naturally have an identifiable conditional VAE (CVAE), as the name suggests. Note that (3) has a similar factorization with the generative model of iVAE (Khemakhem et al., 2020a), that is

$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})$; the first factor does not depend on $X$. Further, since we have the conditioning on $T$ in both the factors of (3), our VAE architecture is a combination of iVAE and CVAE (Sohn et al., 2015; Kingma et al., 2014), with $T$ as the conditioning variable. See Figure 1 for the comparison in terms of graphical models. The core idea of iVAE is reflected in our model identifiability (see Lemma 1).

### C.3 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 $\mu_0$ and $\mu_1$, is monotone, i.e, either increasing or decreasing. The increasing $\boldsymbol{i}$ means, if a change of the value of $X$ 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)" (Janzing & Scholkopf, 2010). The decreasing $\boldsymbol{i}$ corresponds to another common interpretation when ICM does not hold. Now, the treatment does change the way covariates affect $Y$, but in a *global* manner: it acts like a "switch" on the mechanism: the same change of $X$ always has *opposite* effects on the two treatment groups.

We support the above reasoning by real world examples. First we give two examples where $\mu_0$ and $\mu_1$ are both monotone increasing. This, and also that both $\mu_t$ are monotone decreasing, are natural and sufficient conditions for increasing $\boldsymbol{i}$, though not necessary. The first example is form Health. (Starling et al., 2019) 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 $\mu_t$ should be monotone increasing in gestational age. The next example is from economics. (Gan & Li, 2016) 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. (Abrevaya et al., 2015) shows that, with t = mother smokes or not and $X$ = 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 (Wang et al., 2020). 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.4 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 (5). And **(D2')** is easily satisfied with high-dimensional $X$, even if the possible values of $\boldsymbol{C}, \boldsymbol{d}$ are restricted to $\boldsymbol{C} = c\boldsymbol{I}$ and $\boldsymbol{d} = \boldsymbol{0}$ (see below). On the other hand, $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ in **(M3')** is impractical, but it ensures that $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t)$ so that (5) 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.

### C.5 Ideas and connections behind the ELBO (7)

**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, $\beta$ expresses our additional knowledge (or, inductive bias) about whether or not there exist balanced PtSs (e.g., from domain expertise).

In fact, $\beta$ connects our VAE to $\beta$-VAE (Higgins et al., 2017), which is closely related to noise and variance control (Doersch, 2016, Sec. 2.4)(Mathieu et al., 2019).

**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 $\mathbb{P}_T$ and the correlation between $T$ and $X$. This is another reason to prefer $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$, besides balancing. On the other hand, the posterior conditioning on $Y$ provides information of noise $\mathbf{e}$, and it is shown in (Bonhomme & Weidner, 2021) 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 $\beta$ implicitly encourages $\boldsymbol{g}$ *smaller* than the scale of $\mathbf{e}$, through stressing the third term in ELBO (7). And the the model as a whole would still learn $p(\boldsymbol{y}|\boldsymbol{x}, t)$ well, because the uncertainty 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 $\beta$. Intuitively, smaller $\boldsymbol{g}$ strengthens the correlation between $Y$ and $Z$ in our model, and this naturally reflects that posterior conditioning on $Y$ 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 $\beta$ 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 $\beta$.

**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 $Y$ in the posterior. Both are bad because it is shown in (Bonhomme & Weidner, 2021) 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 $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) = p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}, t)/p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t)$ degenerates to $\boldsymbol{f}_T^{-1}(Y) = \boldsymbol{f}_T^{-1}(\boldsymbol{j}_T(\mathbb{P}_T)) = \boldsymbol{v}^{-1}(\mathbb{P}_T)$, a *factual* PtS. However, $\boldsymbol{f}_{1-T}^{-1}(Y) = \boldsymbol{f}_{1-T}^{-1}(\boldsymbol{j}_T(\mathbb{P}_T)) = \boldsymbol{v}^{-1}(\boldsymbol{j}_{1-T}^{-1} \circ \boldsymbol{j}_T(\mathbb{P}_T)) \neq \boldsymbol{v}^{-1}(\mathbb{P}_{1-T})$, *the score recovered by posterior does not work for counterfactual assignment*! The problem is, unlike $X$, the outcome $Y = Y(T)$ is affected by $T$, and, the degenerated posterior disregards the information of $X$ from the prior and depends exclusively on factual $(Y, T)$.

### C.6 Consistency of VAE and prior estimation

The following is a refined version of Theorem 4 in Khemakhem et al. (2020a). 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* (3)&(6)*, and let* $p^*(\boldsymbol{x}, \boldsymbol{y}, t)$ *be the true observational distribution, assume*

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

*ii) the ELBO* $\mathbb{E}_{\mathcal{D}\sim p^*}(\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, t; \boldsymbol{\theta}, \boldsymbol{\phi}))$ (4) *can be optimized to its global maximum at* $(\boldsymbol{\theta}', \boldsymbol{\phi}')$*;*

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

*Proof.* From i), we have $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, t; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}) = \log p^*(\boldsymbol{y}|\boldsymbol{x}, t)$. But we know $\mathcal{L}$ is upper-bounded by $\log p^*(\boldsymbol{y}|\boldsymbol{x}, t)$. So, $\mathbb{E}_{\mathcal{D}\sim p^*}(\log p^*(\boldsymbol{y}|\boldsymbol{x}, t))$ 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 $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) \| q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) \geq 0$ and, in the limit of infinite data, $\mathbb{E}_{\mathcal{D}\sim p^*}(\log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t)) \leq \mathbb{E}_{\mathcal{D}\sim p^*}(\log p^*(\boldsymbol{y}|\boldsymbol{x}, t))$. Thus, the global maximum of ELBO is achieved *only* when $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p^*(\boldsymbol{y}|\boldsymbol{x}, t)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t)$. □

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 $\mu_t(X) = \boldsymbol{f}_t(\boldsymbol{h}_t(X))$ where $\boldsymbol{f}, \boldsymbol{h}$ are the optimal parameters learned by the VAE.*

### C.7 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 $X = \boldsymbol{x}$. To this end, we use the prior as a pre-treatment predictor for $Z$: replace $q_{\boldsymbol{\phi}}$ with $p_{\boldsymbol{\Lambda}}$ in (8) and get rid of the average taken on $\mathcal{D}$; 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 E.

### C.8 Additional notes on novelties of the bounds in Sec. 4.2

We give details and additional points regarding the novelties. Lu et al. (2020) also use a VAE and derive bounds most related to ours. Still, our method strengthens Lu et al. (2020), in a simpler and principled way: we distinguish true score and latent $Z$ 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 (7), without ad hoc regularization; moreover, we consider outcome noise modeling which is a strength of VAE and relate it to hyperparameter $\beta$. Particularly, in (Lu et al., 2020), latent variable $Z$ 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. Note that Shalit et al. (2017) do not consider connection to identification and noise modeling as well. The error between $\hat{\tau}_{\boldsymbol{f}}$ and $\tau_{\boldsymbol{j}}$, which we bound, is due to the unknown outcome noise that is not accounted by our Theorem 1; thus, the theory in Sec. 4.2 is complementary to that in Sec. 3.2.

## 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* (Huber & Wüthrich, 2018). More related to our method is another line of work based on *monotonicity of outcome*, see (Chernozhukov & Hansen, 2013) and references therein for early results. Some recent works apply monotonicity of outcome to nonparametric IV regression (NPIV) (Freyberger & Horowitz, 2015; Li et al., 2017; Chetverikov & Wilhelm, 2017), where the structural equation of the outcome is assumed to be $Y = f(T) + \epsilon$, and $f$ is monotone and $T$ (the treatment) is often continuous. Particularly, (Chetverikov & Wilhelm, 2017) combines monotonicity of both treatment and outcome, and (Freyberger & Horowitz, 2015) 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 (Chetverikov et al.,

2018). 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 limited overlap.

Recently in machine learning, (Johansson et al., 2019; Zhang et al., 2020b; Johansson et al., 2020) note the relationship between invertibility and overlap. As mentioned, (Johansson et al., 2020) gives bounds without overlap, but the relationship between invertibility and overlap is not explicit in their theory. (Johansson et al., 2019) explicitly discuss overlap and invertibility, but does not focus on TEs. (Zhang et al., 2020b) assumes overlap so that identification is given, and then focuses on learning overlapping representation that preserves the overlapping 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. (Louizos et al., 2017; Zhang et al., 2020a; Vowels et al., 2020; Lu et al., 2020), 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, (Louizos et al., 2017) assumes unobserved confounder can be recovered, which is rarely possible even under further structural assumptions (Tchetgen et al., 2020), and (Rissanen & Marttinen, 2021) recently gives evidence that the method often fails. Other methods (Zhang et al., 2020a; Vowels et al., 2020; Lu et al., 2020) assume unconfoundedness but still rely on proxy at least intuitively; particularly, (Lu et al., 2020) 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 (Tchetgen et al., 2020). 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. 5.1.

**On ad hoc heuristics.** Ad hoc heuristics break down probabilistic modeling and / or give ELBOs that do not estimate the probabilistic models. For example, (Louizos et al., 2017) uses separated NNs for the two potential outcomes to mimic TARnet (Shalit et al., 2017). And, to have pre-treatment estimation, $q(T|X)$ and $q(Y|X,T)$ are added into the encoder. As a result, the ELBO of (Louizos et al., 2017) has two additional likelihood terms corresponding to the two distributions. (Zhang et al., 2020a) is even more ad hoc because it splits the latent variable $Z$ into three components, and applies the ad hoc tricks of (Louizos et al., 2017) to each of the component. Particularly, when constructing the encoder, (Zhang et al., 2020a) implicitly assumes the three components of $Z$ are conditional independent give $X$, which violates the intended graphical model.

Our method is motivated by the important concept of prognostic score, 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 (4) is derived by standard variational lower bound. Moreover, in our $\beta$-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

We evaluate the post-treatment performance on training and validation set jointly (This is non-trivial. Recall the fundamental problem of causal inference). The treatment and (factual) outcome should not be observed for pre-treatment predictions, so we report them on a testing set. See also Sec. C.7 the pre/post-treatment distinction.

### E.1 SYNTHETIC DATA

We detail how the random parameters in the DGPs are sampled. $\mu_i$ and $\sigma_i$ 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 $C_t := \mathrm{Var}_{\{\mathcal{D}|T=t\}}(f_t(Z))$. When generating DGPs with dependent noise, the variance parameter $g_t$ for the outcome is generated by adding a softplus layer after respective $f_t$, and then normalized to range $(0, 2)$.

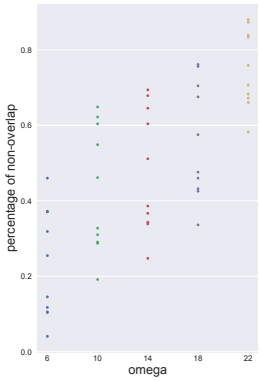


Figure 4: Degree of limited overlap w.r.t $\omega$.

We use the original implementation of CFR[4]. 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 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 limited 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 limited-overlapping when the above percentage is larger than 50%, then, as shown in Figure 4, non (all) of the 10 DGPs are very limited-overlapping with $\omega = 6$ ($\omega = 22$).

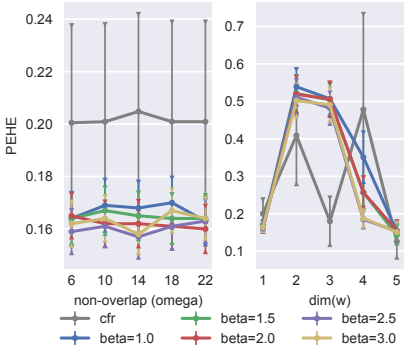


Figure 5: $\sqrt{\epsilon_{pehe}}$ on synthetic dataset, with $g_t(W) = 1$ in DGPs, and $\dim(Z) = 200$ in our model. Error bar on 10 random DGPs.

For diversity of the datasets, we set $g_t(W) = 1$ in DGPs in Appendix. Figure 5 shows, with $\dim(Z) = 200$, our method works better than CFR under $\dim(W) = 1$ and as well as CFR under $\dim(W) > 1$. As mentioned in Conclusion, this indicates that the theoretical requirement of injective $\boldsymbol{f}_t$ in our model might be relaxed. Interestingly, larger $\beta$ seems to give better results here, this is understandable because $\beta$ controls the trade-off between fitting and balancing, and the fitting capacity of our decoder is much increased with $\dim(Z) = 200$. Note that the above observations on $\dim(Z)$ are not caused by fixing $g_t(W) = 1$ (compare Figure 5 with Figure 6 below).

Figure 6 shows the importance of noise modeling. Compared to Figure 2 in the main text, where $g_t(W)$ in DGPs is not fixed, our method works worse here, particularly for large $\beta$, because now noise modeling ($\boldsymbol{g}, \boldsymbol{k}$ in the ELBO) only adds unnecessary complexity. The changes of performance w.r.t different $\omega$ should be unrelated to overlap levels, but to the complexity of random DGPs; compare to Figure 5, with larger NNs in our VAE, the changes become much insignificant. The drop of error for $\dim(W) > 3$ is due to the randomness of $f$ in (36). In Sec. 2.2, we saw that the 2-dimsensional PS $\mathbb{P} := (\mu_0(X), \mu_1(X))$ always exists under ANMs. Thus, when $\dim(W) > 2$, our method tries to recover that $\mathbb{P}$, and generally performs not worse than under $\dim(W) = 2$, but still not better than under $\dim(W) = 1$.

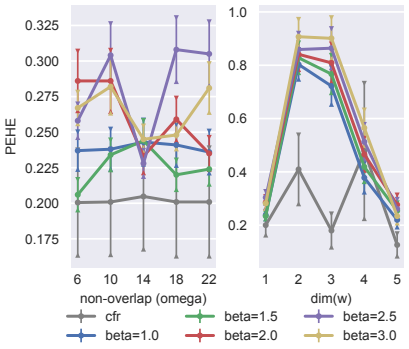


Figure 6: $\sqrt{\epsilon_{pehe}}$ on synthetic dataset, with $g_t(W) = 1$ in DGPs. Error bar on 10 random DGPs.

[4] https://github.com/clinicalml/cfrnet

Figure 7 shows results of ATE estimation. Notably, CFR drops performance w.r.t degree of limited overlap. Our method does not show this tendency except for very large $\beta$ ($\beta = 3$). This might be another evidence that CFR and its unconditional balancing overfit to PEHE (see Sec. 5.2). Also note that, under $\dim(W) = 1$, $\beta = 3$ 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 $\beta$ gives better ATE estimation under the existence of PS, but leave this for future investigation.

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 $W$ with $\dim(W) = 1$, 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.

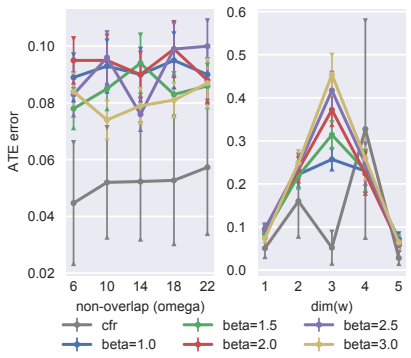


Figure 7: $\epsilon_{ate}$ on synthetic dataset, with $g_t(W) = 1$ in DGPs. Error bar on 10 random DGPs.

## E.2 IHDP

IHDP is based on an RCT where each data point represents a baby 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 (Hill, 2011, Sec. 4.1).

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

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, $\boldsymbol{a}^T X$ 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.

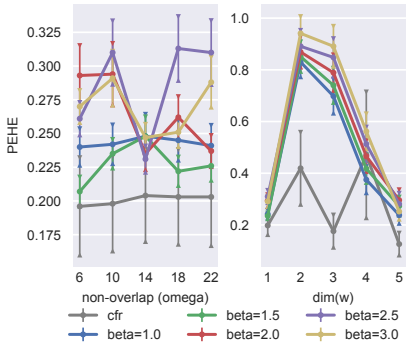


Figure 8: *Pre-treatment* $\sqrt{\epsilon_{pehe}}$ on synthetic dataset. Error bar on 10 random DGPs.

The two added components in the modified version of our method are as following. First, we build the two outcome functions $\boldsymbol{f}_t(Z), t = 0, 1$ in our learning model (3), using two separate NNs. Second, we add to our ELBO (4) a regularization term, which is the Wasserstein distance (Cuturi, 2013) between $\mathbb{E}_{\mathcal{D}\sim p(X|T=t)} p_{\boldsymbol{\Lambda}}(Z|X), t \in \{0, 1\}$. As shown in Table 2, best unconditional balancing parameter is 0.1. 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 3 shows pre-treatment results, All methods gives reasonable results.

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 |
|---|---|---|---|---|---|---|---|
| $\epsilon_{ate}$ | $.177_{\pm .007}$ | $.196_{\pm .008}$ | $.177_{\pm .007}$ | $.167_{\pm .005}$ | $.177_{\pm .006}$ | $.179_{\pm .006}$ | $.25_{\pm .01}$ |
| $\sqrt{\epsilon_{pehe}}$ | $.843_{\pm .030}$ | $1.979_{\pm .082}$ | $1.116_{\pm .046}$ | $.777_{\pm .026}$ | $.894_{\pm .039}$ | $.841_{\pm .029}$ | $.71_{\pm .02}$ |

Table 3: *Pre-treatment* Errors on IHDP over 1000 random DGPs. We report results with $\dim(Z) = 10$. **Bold** indicates method(s) which is *significantly* better. The results are taken from Shalit et al. (2017), except GANITE (Yoon et al., 2018) and CEVAE (Louizos et al., 2017).

| Method | TMLE | BNN | CFR | CF | CEVAE | GANITE | Ours |
|---|---|---|---|---|---|---|---|
| pre-$\epsilon_{ate}$ | 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 | $2.1_{\pm .1}$ | $\mathbf{.76}_{\pm .02}$ | $3.8_{\pm .2}$ | $2.6_{\pm .1}$ | $2.4_{\pm .4}$ | $.946_{\pm .048}$ |

### 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.

Pokec (Leskovec & Krevl, 2014) is a real world social network dataset. We experiment on a semi-synthetic dataset based on Pokec, which was introduced in (Veitch et al., 2019), 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 which are within the same region. The network structure is expressed by binary adjacency matrix $\boldsymbol{G}$. Following (Veitch et al., 2019), 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 $X$. Treatment $T$ and outcome $Y$ are synthesised as following:

$$T \sim \mathrm{Bern}(g(U)), \quad Y = T + 10(g(U) - 0.5) + \epsilon, \tag{38}$$

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

$\beta$-Intact-VAE is expected to learn a PS from $\boldsymbol{G}, X$, 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 $\boldsymbol{G}, X$. This idea is comparable to Assumptions 2 and 4 in (Veitch et al., 2019), 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) (Kipf & Welling, 2017) in conditional prior and encoder of $\beta$-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, Veitch et al. (2019) does not give individual-level prediction.

To extract information from the network structure, we use Graph Convolutional Network (GCN) (Kipf & Welling, 2017) in conditional prior and encoder of $\beta$-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.

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 (Gopalan & Blei, 2013). "Embed-" denotes the best alternatives given by (Veitch et al., 2019). **Bold** indicates method(s) which is *significantly* better than all the others. We report results with 20-dimensional latent $Z$. The results of the other methods are taken from (Veitch et al., 2019).

| | Age | District | Join Date |
|---|---|---|---|
| Unadjusted | $4.34 \pm 0.05$ | $4.51 \pm 0.05$ | $4.03 \pm 0.06$ |
| Parametric | $4.06 \pm 0.01$ | $3.22 \pm 0.01$ | $3.73 \pm 0.01$ |
| Embedding-Reg. | $2.77 \pm 0.35$ | $\mathbf{1.75} \pm 0.20$ | $2.41 \pm 0.45$ |
| Embedding-IPW | $3.12 \pm 0.06$ | $\mathbf{1.66} \pm 0.07$ | $3.10 \pm 0.07$ |
| Ours | $\mathbf{2.08} \pm 0.32$ | $\mathbf{1.68} \pm 0.10$ | $\mathbf{1.70} \pm 0.13$ |

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 $\beta$-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 (Srivastava et al., 2014) with rate 0.1 to prevent overfitting.

### E.4 Additional plots on synthetic datasets

See next pages.

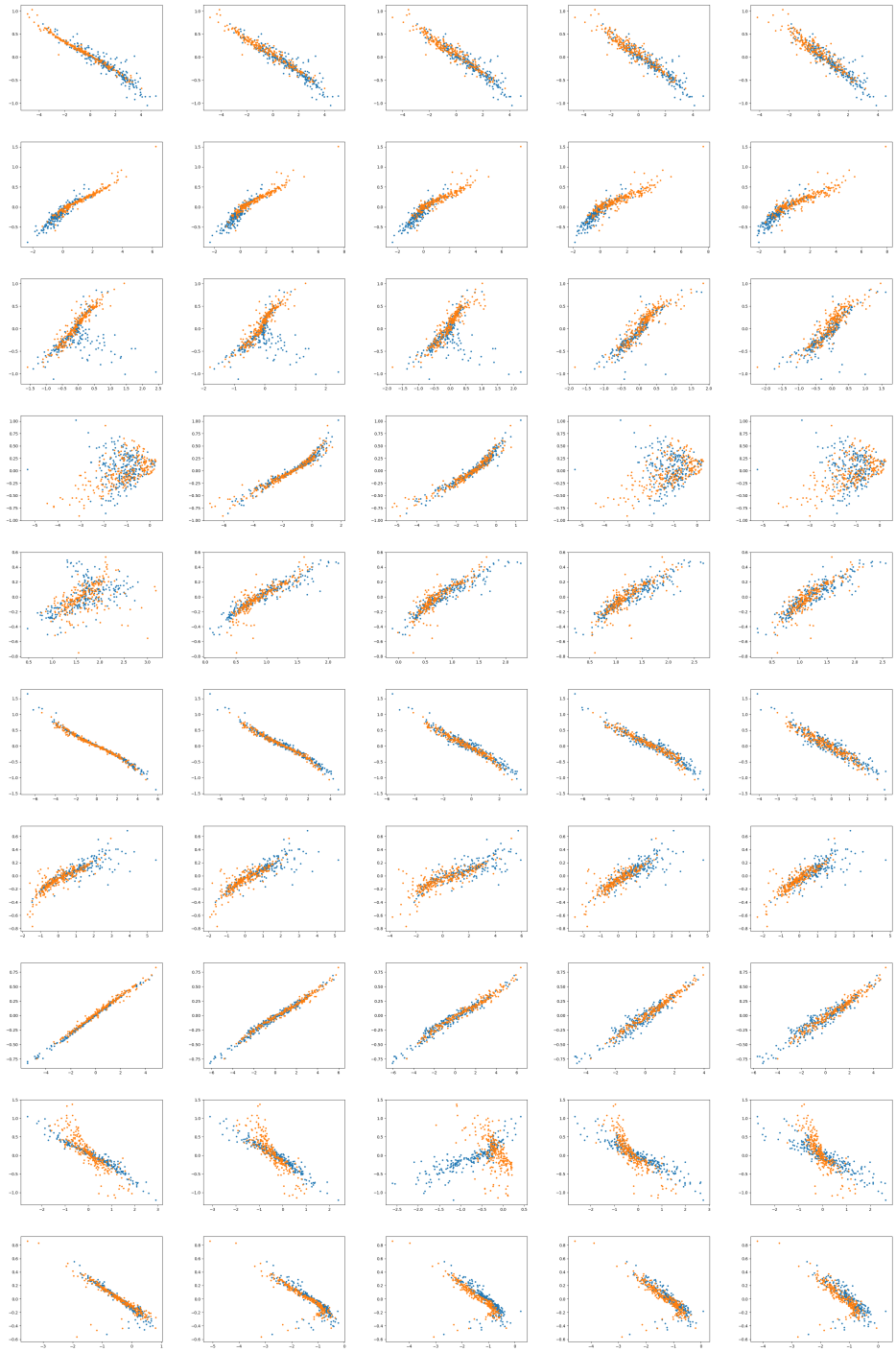

Figure 9: Plots of recovered-true latent. Rows: first 10 nonlinear random models, columns: outcome noise level.

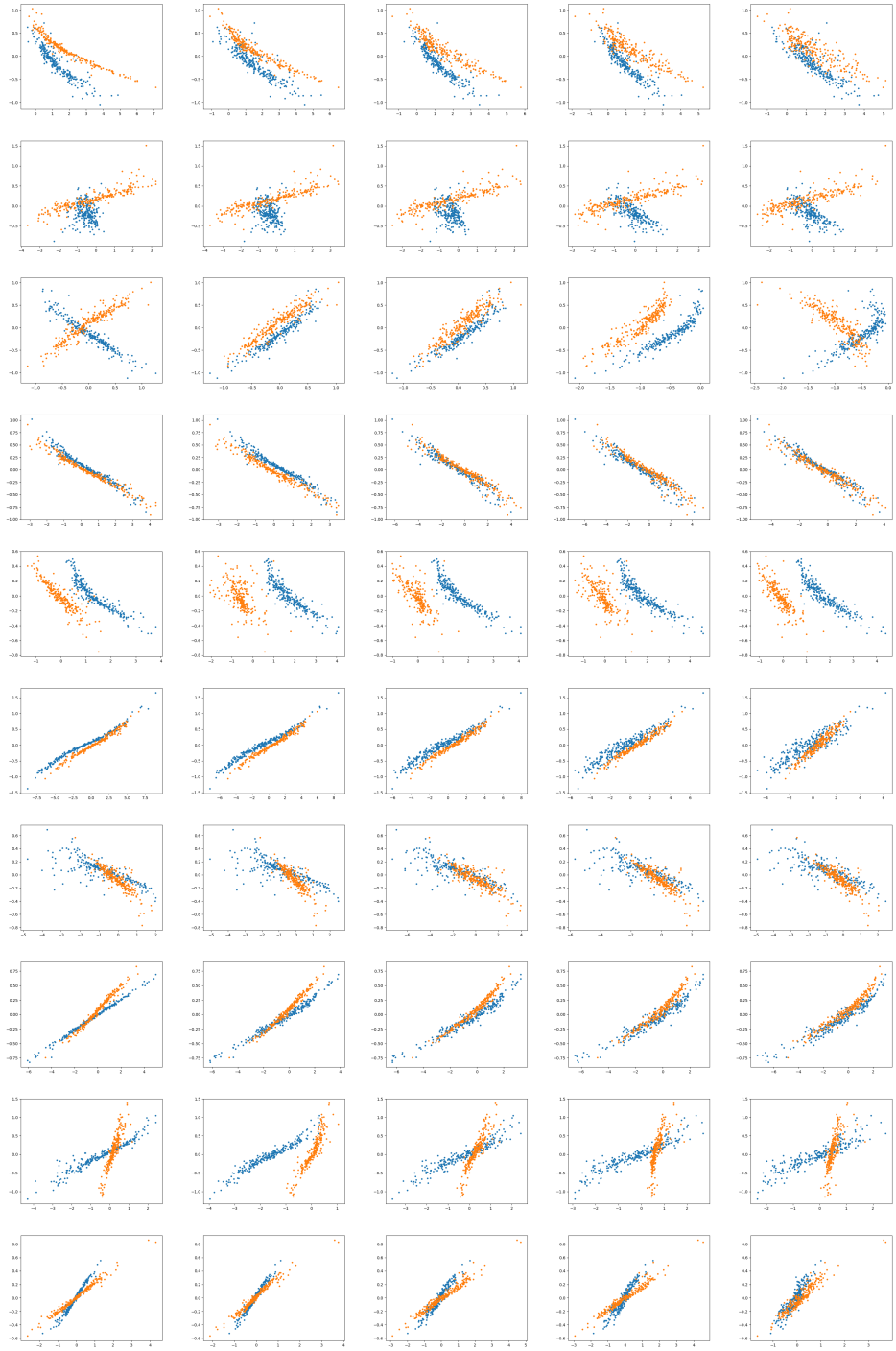

Figure 10: Plots of recovered-true latent. Conditional prior *depends* on $t$. Rows: first 10 nonlinear random models, columns: outcome noise level. Compare to the previous figure, we can see the transformations for $t = 0, 1$ are *not* the same, confirming the importance of balanced prior.