THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES, SOKENDAI

DOCTORAL THESIS

Treatment Effect Estimation and Bivariate Causal Discovery via Nonlinear ICA

Author:
Pengzhou WU

Supervisor:
Professor Kenji FUKUMIZU

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Statistical Science

"The one born in the wild knew how to give counsel,
Enkidu spoke to his friend, gave his dream meaning:...."

"O Gilgamesh, where are you wandering?
The life that you seek you never will find."

"He came a far road, was weary but at peace;
all his labours were set on a tablet of stone."

From *The Epic of Gilgamesh*, translated by Andrew George

"Whatever phenomena arise from a cause:
Their cause
& their cessation.
Such is the teaching of the Tathagata,
the Great Contemplative."

From *Mv I.23.5*, translated by Thanissaro Bhikkhu

THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES, SOKENDAI

Abstract

Department of Statistical Science

Doctor of Philosophy

Treatment Effect Estimation and Bivariate Causal Discovery via Nonlinear ICA

by Pengzhou WU

Causality–i.e., asking and answering "Why?"–is fundamental in fields of science. In this era of big data and artificial intelligence, scientists are enthusiastic about exploiting machine learning systems in the causal analysis of scientific datasets at scale. On the other hand, main stream machine learning systems are based on plain statistical associations and focus on prediction and pattern recognition. Thus, causality and machine learning should go hand-in-hand for scientific discovery and decision-making. In this thesis, we develop new machine learning methods for causal effect estimation and causal discovery, the two major problems in causality.

First, we discuss the identification and estimation of treatment effects 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 treatment effects; 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 Intact-VAE—a new type of variational autoencoder (VAE). We derive the treatment effect 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. Moreover, experiments show state-of-the-art performance under diverse settings, including unobserved confounding. We also discuss (possible) theoretical extensions to unobserved confounding.

Second, we address the problem of bivariate causal discovery. Based on recent developments in nonlinear independent component analysis (ICA), we train general nonlinear causal models that are implemented by neural networks and allow non-additive noise. Further, we build an ensemble framework, namely Causal Mosaic, which models a causal pair by a mixture of nonlinear models. We compare this method with other recent methods on artificial and real world benchmark datasets, and our method shows state-of-the-art performance.

Acknowledgements

I am extremely grateful to my supervisor Prof. *Kenji Fukumizu*. Kenji has never hesitated to devote his time when I need discussions. I was always assured by his encouragement of my research ideas, because there was also no lack of critical comments from him. His wise advice came at crucial times: whether it was the right time to start my first research project, whether to write a rushed draft and submit it to NeurIPS for the first time, what and how to write in the replies to reviewers, specifically, honestly, and politely... I am deeply indebted to him for his infinite patience in my bad and tedious writing, and his time and often hands-on effort to better it. He looked tirelessly into the organization, wording, math, and typos, making suggestions, corrections, and once even re-writing a whole paragraph! His deep and broad knowledge saved me from technical mistakes. I cannot imagine our papers would have the current quality without his invaluable help. Moreover, his help is not only professional but also personal–an example is that he sent his WIFI router, by himself and from his home, to me after mine from the institute was expired, and my life was made much easier during the height of COVID. My Ph.D. under Kenji would not have been possible without Prof. *Gang Niu*'s suggestion. I am also grateful to Prof. *Song Liu* and Prof. *Shaogao Lv*, who helped me with my decision.

I would like to extend my sincere thanks to my defense committee: Prof. *Hideitsu Hino*, Prof. *Masaaki Imaizumi*, Prof. *Yoshiyuki Ninomiya* (alphabetical order). I had not thought I could learn so much from revising the thesis, and, thanks to the many constructive suggestions, the current thesis becomes a more coherent whole.

Many thanks to our (former) lab members: Prof. *Masaaki Imaizumi*, Dr. *Yoh-ichi Mototake*, *Hironori Murase*, Dr. *Hideto Nakashima*, *Yuto Tanimoto*, *Shoji Toyota*. Special thanks to Shoji, who is so outgoing and always chats with me in English. Thanks also to Imaizumi-san, to whom I had many face-to-face chats before COVID, I still remember the joke about how to fit four big trays on a small round table! Mototake-san has also amused us with his innate humor. Thanks should also go to our visitors, especially *Jean-Francois Ton* and Dr. *Wenkai Xu*, with whom I had interesting discussions on cause-effect inference, and we are still exchanging ideas on SNSs. I am grateful to Dr. *Krikamol Muandet* who was happy to host my visit to MPI-IS, though, sadly, the visit was stopped by COVID and we could only meet online.

Lastly, I would be remiss in not mentioning my parents' patience and endurance.

Contents

List of Figures

List of Tables

List of Abbreviations

ATE	Average Treatment Effect
BRL	Balanced Representation Learning
CATE	Conditional Average Treatment Effect
CVAE	Conditional Variational AutoEncoder
DGP	Data Generating Process
DAG	Directed Acyclic Graph
ELBO	Evidence Lower BOund
FCM	Functional Causal Model
ICA	Independent Component Analysis
iVAE	identifiable Variational AutoEncoder
MLP	Multi-Layer Perceptron
NN	Neural Network
PEHE	Precision in Estimation of Heterogeneous Effect
RCT	Randomized Controlled Trial
SCM	Structural Causal Model
TCL	Time-Contrastive Learning
VAE	Variational AutoEncoder

For シンテイ

sharing the pain, pleasure, and dullness in all these years

Chapter 1

Introduction

1.1 Causality and Machine Learning

Human knowledge begins with the observation and study of nature. In Ancient Greece, some philosophers, including Aristotle, were also the earliest scientists intrigued by the search for causes of natural phenomena. Even Plato, who was at times very mystical, had enough interest in "inquiry into Nature" and, in his *Phaedo*, stated that it consisted of a quest for "the causes of each thing; why each thing comes into existence, why it goes out of existence, why it exists"[1]. Perhaps due to the eminence of Plato in philosophy, nowadays, the study of causality still focuses on "*Why*?" questions, as indicated by the title of Judea Pearl's widely-received book (Pearl and Mackenzie, 2018).

The difference between statistical correlation and causation can not be too much stressed. Indeed, a well-known mantra is that "correlation does not imply causation". It is a fallacy to conclude causation from correlation because the correlation might be due to omitted data or unobserved links. For example[2], in summer, the owner of an ice cream shop may observe high electricity usage and also high sales. Thus, the owner would observe a strong correlation between the electricity usage and the sales, but the former did not cause the latter–surely, leaving the lights on in the shop over night would have no impact on the sales. In fact, the hot weather is the common cause of both the high electricity usage and the high sales, and we say

[1] I got to know this quote from Nogueira et al., 2022, and this beginning paragraph is also largely inspired by theirs.

[2] This example is taken from Guo et al., 2020.

the weather here is a *confounder* of the relationship between the electricity usage and the ice cream sales.

Perhaps, the less well-known is that causation does not imply correlation, either. There are cases in which, although there is a clear causal relationship between variables, there is no correlation observed. Of course, this might be because of the limited information in a specific sample. However, even if infinite sample size is assumed, a correlation coefficient (e.g., Pearson's) would not capture all kinds of statistical associations between the cause and effect. This is why statistical *(in)dependence* is far more important than simple correlation. It is commonly agreed that, with infinite sample, association (dependence) does imply causation, and this is often known as *Reichenbach Principle* (Schölkopf et al., 2021). Moreover, as we will see later in the thesis, under certain assumptions, statistical independence can be exploited to answer causal questions. Roughly speaking, *the study of causality concerns what causal questions can be reduced to statistical ones, and how (e.g., under what assumptions)*.

Given its inherent multi-disciplinary nature, the study of causality is historically fragmented into several different domains, including epidemiology (biostatistics), economics, statistics, computer science. Due to the scarcity of prior causal knowledge, it is usually hard to making convincing causal assumptions and testing plausibility of them. This is why *randomized controlled trials* (RCTs) becomes the golden standard of studies in causality. For example, to study the efficacy of a new drug, a patient would be randomly assigned to take the drug or not, which would guarantee that, on average, the treated group and the non-treated (control) groups are equivalent in all relevant respects, ruling out the influence of any factors except the treatment. Then, the effect of the drug on a certain health outcome can be measured by comparing the average outcome of the two groups.

While RCTs control biases through randomization, they often have ethical and practical issues, or suffer from expensive costs. Thus, solving causal problems from observational data is important. Recent advance in information collection and storage have made a huge amount of observational data available for researchers and policy makers in those different fields, for example, in the form of electronic health records (Evans, 2016). The scientific communities have considerable interests to exploit the so-called big data to solve causal problems, while they face new challenges

at the same time. Public databases or data collected from the web are unprecedentedly large, people have little intuition about what types of bias a dataset can suffer from–the more plentiful data makes it harder to understand and, consequently, harder to come up and validate causal assumptions.

On the other hand, empowered by the increasing collection of big data and growth in computing power (mainly GPU), machine learning and artificial intelligence (AI), particularly deep learning, have made remarkable progress, surpassing human performance in many tasks such as object recognition, machine translation, and reading comprehension (LeCun, Bengio, and Hinton, 2015). Given its origin in nonparametric statistics (Vapnik, 1999) and connectionism (Rumelhart, McClelland, Group, et al., 1988), main stream machine learning systems are based on plain statistical associations. However, the ability of causal reasoning and learning is considered as a significant ingredient of human-level intelligence and, as argued by some, can serve as the foundation of AI (Pearl, 2018) or help to solve several challenge problems in machine learning such as robustness, reusability, and interpretability (Schölkopf et al., 2021).

The above trends motivate the work contained in this thesis. Specifically, we study the two major problems in causality: first, *causal effect estimation* (Imbens and Rubin, 2015; Pearl, 2009), i.e., quantifying the strength of influence of the cause on the effect, if the cause is intervened; second, *causal discovery* (Spirtes et al., 2000), i.e., finding the causal relationships (which ones are the causes of which ones?) across the a set of variables. These two tasks are complementary: the former is quantitative while the latter is qualitative; the former assumes a causal relationship (which is cause and which is effect) while the latter discovers causal relationships.

Organization of the thesis The organization is rather standard. An exception is that a rather formal Preliminaries chapter is put before Literature Review, because the Preliminaries chapter contains the precise definitions of some concepts used in Literature Review. Nevertheless, Literature Review could be read and understood to a large extent of one prefers. We note that Chapter 4, 5, 6 contains the main contributions of this thesis, with Chapter 4 & 5 addressing treatment effect estimation, and Chapter 6 addressing bivariate causal discovery. Expect these three chapters,

the other chapters/sections usually have two parts discussing the two problems respectively. For example, Sec. 2.2.1 is the a preliminary to Chapter 4 & 5 because iVAE is a basis of our work on treatment effect estimation. Similarly, Sec. 2.2.2 is a preliminary to Chapter 6. These two threads of the thesis are relatively standalone and could be read independently. Finally, sections with titles as "Details/Detailed..." could be omitted on first reading.

Notes on terminology In this section, we have used *causality* to refer to all the studies on causal problems, including causal effect estimation and causal discovery. This also accords with the title of Pearl's encyclopedic book (Pearl, 2009). Traditionally, *causal inference* refers to the studies on causal effect, including both identification and estimation, as in the titles of Rubin, 2005; Hernan and Robins, 2020. However, in recent years, particularly in machine learning and other interdisciplinary contexts, "causal inference" has been extended to refer all causal studies, similar to "causality". This can be seen in the name "Journal of Causal Inference"[3] and in the title of Peters, Janzing, and Schölkopf, 2017 who mainly discuss (bivariate) causal discovery. In the following, we will use the term "causal inference" in the former sense, particularly when we want to include the study of causal effect identification, but *avoid* the latter usage because it would confuse some readers. Finally, we note that bivariate causal discovery is sometimes referred to as *cause-effect inference*, because it aims to distinguish cause and effect.

1.2 Research Problems

The following subsections specify the two important causal problems studied in this thesis and bring up the focuses of our research. Detailed review of previous works can be found in Chapter 3.

1.2.1 Treatment Effect Estimation

In this thesis, we focus on treatment effects based on a set of observations comprising binary labels T for treatment/control (non-treated), outcome Y, and other covariates

[3] https://www.degruyter.com/journal/key/jci/html?lang=en

X. This is arguably the most important situation in causal inference, particularly widely studied in epidemiology and biostatistics. 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 our work contained in Chapter 4, have relied on unconfoundedness; this means that the confounding can be controlled by conditioning on the covariates. The more covariates are collected the more likely unconfoundedness holds; however, more covariates tends to introduce a stronger imbalance between treatment and control.

Chapter 4 studies the issue of imbalance in estimating individualized treatment effects 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., Laan and Rose (2018, TMLE)—which improve finite sample performance, as well as non-parametric methods—e.g., Wager and Athey (2018, CF). Notably, from Johansson, Shalit, and Sontag (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 treatment effects is not

possible at non-overlapping covariate values due to lack of data. There are works that provide robustness to limited overlap (Armstrong and Kolesár, 2021), trim non-overlapping data points (Yang and Ding, 2018), weight data points by overlap (Li and Li, 2019), or study convergence rates depending on overlap (Hong, Leung, and Li, 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 treatment effect 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 treatment effects, 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 treatment effects) 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 and 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., 2020b, iVAE) and conditional VAE (Sohn, Lee, and Yan, 2015, CVAE). Based on model identifiability, we develop two identification results for individualized treatment effects under limited overlap. The current study further provides bounds on individualized treatment effect error, and the bounds justify a conditionally balancing term controlled by hyperparameter β, as an interpolation between the two identifications. This VAE architecture was first proposed by us in Wu and Fukumizu (2020b).

There are a few lines of works that address the difficult but important problem of *unobserved confounding*. Without covariates to adjust for, the naive regression with observed variables introduces bias, if the decision of treatment and the outcome are confounded, as explained in Sec. 2.1.1. Instead, many methods assume special structures among the variables, such as instrumental variables (IVs) (Angrist, Imbens, and Rubin, 1996), proxy variables (Tchetgen et al., 2020), network structure (Veitch, Wang, and Blei, 2019), and multiple causes (Wang and Blei, 2019b). Among them, instrumental and proxy variables are most commonly exploited. *Instrumental variables* are not affected by unobserved confounders, influencing the outcome only through the treatment. On the other hand, *proxy variables* are causally connected to unobserved confounders, but are not confounding the treatment and outcome by themselves. Other methods use restrictive parametric models (Allman, Matias, Rhodes, et al., 2009), or only give interval estimation (Manski, 2009; Kallus, Mao, and Zhou, 2019).

In Chapter 5, we challenge the problem of estimating treatment effects under unobserved confounding. We highlight the promising experimental results of Intact-VAE, under unconfounded, IV, proxy, and networked confounding settings. We also discuss some theoretical ideas under unobserved confounding.

The hallmark of deep neural networks (NNs) is that they can learn representations of data. It is desirable that the learned representations are interpretable, that is, in approximately the same relationship to true latent sources for each down-stream task. A principled approach to interpretable representations is identifiability, that is, when optimizing our learning objective w.r.t. the representation function, only a

unique optimum, which represents the true latent structure, will be returned. Our method provides the stronger identifiability that gives *balanced* representation. VAEs (Kingma, Welling, et al., 2019) are suitable for causal estimation thanks to its probabilistic nature. However, most VAE methods for treatment effects, e.g., Louizos et al., 2017; Zhang, Liu, and Li, 2020, are ad hoc and thus not identifiable. Instead, our goal is to build a VAE that can identify and recover from observational data a sufficient score via the latent variable, which can be seen as a *causal representation* (Schölkopf et al., 2021); recovering the true confounder is not necessary.

1.2.2 Bivariate Causal Discovery

Causal discovery (Spirtes and Zhang, 2016; Peters, Janzing, and Schölkopf, 2017) is a fundamental problem which attracts increasing attention recently. Traditionally, causal discovery algorithms learn the causal structure in the form of a directed acyclic graphical (DAG) model, by searching in the space of possible DAGs. Constraint-based search methods, such as FCI (Spirtes, Meek, and Richardson, 1999), use conditional independence tests to determine the causal structure. Score-based search methods, such as GES (Chickering, 2002), typically search for a graph that optimizes a penalized likelihood score. However, the above methods are not applicable to bivariate case and unable to fully determine edge directions in a DAG.

In recent years, a line of research emerges that is particularly motivated to solve the problem of distinguishing cause from effect in bivariate case, i.e. cause-effect inference. All these methods exploit cause-effect asymmetry to identify causal direction (Mooij et al., 2016). One major approach is to restrict causal mechanism to a certain class of "functional causal models" (FCMs) (Hyvärinen and Zhang, 2016), and the causal direction between C and E is identifiable if $p(E|C)$ can be fitted by this class, while the opposite direction, $p(C|E)$, cannot. Many FCMs are additive noise models, with different types of noises and mean functions. There is another line of work loosely exploit the idea that the process generating cause distribution $p(C)$ is in some way "independent" to the causal mechanism generating conditional distribution $p(E|C)$ (Janzing and Scholkopf, 2010).

We can observe the following limitations in the existing methods. First, FCMs put too strong restrictions on the functional form of causal mechanism. Second,

other works tend to propose simple "principles" that actually reflect the authors' own intuitions on causality. Thus, most methods fail to achieve high accuracy on real world data. Third, there are a few methods–e.g., CGNN (Goudet et al., 2018)–that use more flexible models and achieve better performance, but without theoretical justifications. Fourth, they assume there exist no hidden confounders.

Chapter 6 studies cause-effect inference and address the first three[4] limitations respectively as follows. First, we train nonlinear causal models on cause-effect pairs with (maybe partial) direction information, based on a recent nonlinear ICA method implemented by neural network, without strong restriction on the functional relationship among the variables or the noise structure. Second, the fact that each of the many approaches to causality works to some limited extent suggests us to take a "mosaic" view: causal systems are diverse and heterogeneous, so we should not fit all the different systems at once; instead, study at a time a small number of causal systems that share common aspects, and then build a whole picture. Specifically, we build an ensemble of nonlinear models, which amounts to a Causal Mosaic: a causal pair's mechanism is treated as a mixture of similar mechanisms. It is analogous to constructing a large piece of mosaic from tesserae, which are small blocks of material used in creating a mosaic. Finally, we provide theoretical results on the conditions under which our method will work.

1.3 Contributions

In Chapter 4, we study the identification (Sec. 4.2) and estimation (Sec. 4.3) of individualized treatment effects 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 main contributions of this chapter are:

1. treatment effect identification under limited overlap of X, via prognostic scores and an identifiable model;

2. Bounds on individualized treatment effect error, which justify our conditional BRL;

[4] To deal with confounders, we can combine our method with graphical search methods.

3. A new regularized VAE, β-Intact-VAE, realizing the identification and conditional balance;

4. Experimental comparison to the state-of-the-art methods on (semi-)synthetic datasets.

In Chapter 5, we challenge the problem of estimating treatment effects under unobserved confounding. We highlight the promising experimental results of Intact-VAE, under unconfounded, IV, proxy, and networked confounding settings. We also discuss some theoretical ideas under unobserved confounding. The main contributions of this chapter are:

1. Experimental comparison to state-of-the-art methods under diverse settings;

2. Discussions of further theoretical developments and principled treatment effect estimation using VAEs.

In Chapter 6, we study the problem of cause-effect inference and address the three limitations in previous work mentioned above. The main contributions of this chapter are:

1. Two novel cause-effect inference rules with identifiability proofs;

2. An ensemble framework that works for real world datasets with only limited labeled pairs;

3. A neural network structure designed for causal-effect inference;

4. State-of-the-art performance on a real-world benchmark dataset.

Chapter 2

Preliminaries

2.1 Notations, Formulations, and Definitions

2.1.1 Treatment Effect Estimation

Counterfactuals, Treatment Effects, and Identification

Following Imbens and 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 $\mathcal{D} \coloneqq (X, Y, T) \sim p(\boldsymbol{x}, \boldsymbol{y}, t)$. 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 our work in Chapter 4 & 5 are the conditional ATE (CATE) and average treatment effect (ATE), defined, respectively, by:

$$\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \quad \nu = \mathbb{E}(\tau(X)). \tag{2.1}$$

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

Standard results (Rubin, 2005)(Hernan and Robins, 2020, Ch. 3) show sufficient conditions for treatment effect 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 and 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. 4.2.3, 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 thesis, 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*.

Prognostic Score

Our method aims to recover a prognostic score (Hansen, 2008), adapted to account for both t as in Definition 2. On the other hand, balancing scores (Rosenbaum and 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. 2.1.1 for detail.

Definition 2. A *prognostic score* is $\{\mathfrak{p}(X,t)\}_{t\in\{0,1\}}$ such that $Y(t) \perp\!\!\!\perp X|\mathfrak{p}(X,t)$, where $\mathfrak{p}(\boldsymbol{x},t)$ ($\mathfrak{p}_t(\boldsymbol{x})$ hereafter) is a function defined on $\mathcal{X} \times \{0,1\}$. A prognostic score is called *balanced* (and a *balanced prognostic score*) if $\mathfrak{p}_0 = \mathfrak{p}_1$.

We say a prognostic score is overlapping, if *both* $\mathfrak{p}_0(X)$ and $\mathfrak{p}_1(X)$ are overlapping. Obviously, a balanced prognostic score $\mathfrak{p}(X)$ is a conditionally balanced representation (defined as $Z \perp\!\!\!\perp T|X$ in Introduction) and is thus named. We often write t of the function argument in subscripts.

Details and Relationship to Balancing Score In the fundamental work of (Hansen, 2008), prognostic score is defined equivalently to our $\mathfrak{p}_0$, but it in addition requires no effect modification to work for $Y(1)$. Thus, a useful prognostic score corresponds to our Definition 2.

First, we quote the following three properties of conditional independence (see standard textbooks, e.g., Pearl, 2009, Sec. 1.1.55 for a proof) which will be used repeatedly in the proof of Proposition 2.

Proposition 1 (Properties of conditional independence)**.** *For random variables W, X, Y, Z. We have:*

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

We give main properties of prognostic score as following.

Proposition 2. *If V gives exchangeability, and $\mathfrak{p}_t(V)$ is a prognostic score, then $Y(t) \perp\!\!\!\perp V, T|\mathfrak{p}_t$.*

Proof of Proposition 2. From $Y(t) \perp\!\!\!\perp T|V$ (*exchangeability* of V), and since $\mathfrak{p}_t$ is a *function* of V, we have $Y(t) \perp\!\!\!\perp T|\mathfrak{p}_t, V$ (1).

From (1) and $Y(t) \perp\!\!\!\perp V|\mathfrak{p}_t(V)$ (definition), using contraction rule, we have $Y(t) \perp\!\!\!\perp T, V|\mathfrak{p}_t$ for both t. □

Prognostic scores are closely related to the important concept of balancing score (Rosenbaum and Rubin, 1983). Note particularly, the proposition implies $Y(t) \perp\!\!\!\perp T|\mathfrak{p}_t$ (using decomposition rule). Thus, if $\mathfrak{p}(V)$ is a balanced prognostic score, then $\mathfrak{p}$ also gives weak ignorability (exchangeability and overlap), which is a nice property shared with balancing score, as we will see immediately.

Definition 3 (Balancing score)**.** $\boldsymbol{b}(V)$, a function of random variable V, is a balancing score if $T \perp\!\!\!\perp V | \boldsymbol{b}(V)$.

Proposition 3 (Rosenbaum and Rubin, 1983)**.** *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 as in Sec. 5.1.2) is sufficient for the potential outcomes $Y(t)$ ($Y(t) \perp\!\!\!\perp V | \mathfrak{p}_t(V)$). 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.

2.1.2 Bivariate Causal Discovery

In the following, we first formally introduce our problem setting. In Sec. 2.2.2, we show its connection to nonlinear ICA.

Generally, causal relationships can be formalized by Structural Causal Models (SCMs) (Pearl, 2009), also known as Structural Equation Models (SEMs) (Bollen, 1989). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a causal DAG, where $\mathcal{V}$ is the vertex set and $\mathcal{E}$ is the edge set. Then, the SCM of observed variables $\mathbf{X} = (X_v)_{v \in \mathcal{V}}$ and *independent* hidden variables $\mathbf{E} = (E_v)_{v \in \mathcal{V}}$ is given by the set of equations [1]:

$$X_v = f_v(X_{pa_{\mathcal{G}}(v)}, E_v), v \in \mathcal{V} \tag{2.2}$$

[1] As typical definition of SCM, we rule out *feedback* loops (two-way causal influences) and *confounders* (hidden common causes) here.

f_v represents the *causal mechanism* between effect X_v and its direct causes (parents in the graph) $X_{pa_{\mathcal{G}}(v)}$. And E_v models exogenous (external) influences on X_v and is often treated as an unobserved noise.

FIGURE 2.1: Causal graphs of bivariate SCMs

In our work in Chapter 6, we focus on bivariate cases, where there are only two possibilities: either X_1 or X_2 is the direct cause of the other, as shown in Figure 2. Their SCMs are the following (2.3) for $X_1 \to X_2$, and (2.4) for $X_2 \to X_1$. In cause-effect inference, our goal is to distinguish between these two possibilities, that is, tell cause from effect.

$$X_1 = f_1(E_1), \qquad X_2 = f_2(X_1, E_2) \tag{2.3}$$

$$X_1 = f_1(X_2, E_1), \qquad X_2 = f_2(E_2) \tag{2.4}$$

2.2 Nonlinear ICA

Nonlinear ICA provides a general framework to recover independent components from observed data. Unlike many other representation learning methods, e.g. deep generative networks, it starts from a generative model which is well-defined in the sense that the hidden variables is recoverable.

A straightforward definition of the generative model for nonlinear ICA is that independent hidden variables $\mathbf{Z} = (Z_1, ..., Z_n)$ are mixed by a differentiable and invertible nonlinear function $\mathbf{f}$, and produce observed variables $\mathbf{X} = (X_1, ..., X_n) = \mathbf{f}(\mathbf{Z})$. The goal is to recover the independent components Z_i and the unmixing function $\mathbf{g} = \mathbf{f}^{-1}$, only using observations of $\mathbf{X}$.

2.2.1 VAE from the Viewpoint of Nonlinear ICA

VAEs (Kingma, Welling, 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 and Welling, 2013), the variational lower bound $\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\phi})$ of the log-likelihood is derived as:

$$\begin{aligned}\log p(\boldsymbol{y}) &\geq \log p(\boldsymbol{y}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y})\|p(\boldsymbol{z}|\boldsymbol{y})) \\ &= \mathbb{E}_{z\sim q}\log p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{y})\|p(\boldsymbol{z})),\end{aligned} \tag{2.5}$$

where D_{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 and 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, Lee, and Yan, 2015; Kingma et al., 2014) adds a conditioning variable C, usually a class label, to standard VAE (See Figure 4.1). With the conditioning variable, CVAE can give better reconstruction of each class. The variational lower bound is

$$\log p(\boldsymbol{y}|\boldsymbol{c}) \geq \mathbb{E}_{z\sim q}\log p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{c})\|p(\boldsymbol{z}|\boldsymbol{c})). \tag{2.6}$$

The conditioning on C in the prior is usually omitted (Doersch, 2016), i.e., the prior becomes $Z \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ 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., 2020b) 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

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{x}) &\geq \log p(\boldsymbol{y}|\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\|p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})) \\ &= \mathbb{E}_{z\sim q}\log p_{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{z}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x})\|p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})),\end{aligned} \tag{2.7}$$

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 $Z = \boldsymbol{f}^{-1}(Y)$.

2.2.2 Nonlinear ICA and Causal Discovery

The following definition formally states the connection between SCM and nonlinear ICA:

Definition 4. An SCM (2.2) is **analyzable** if there exists a differentiable and invertible[2] function $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^n$, such that $\mathbf{X} = \mathbf{f}(\mathbf{E})$.

Obviously, an analyzable SCM is a special case of nonlinear ICA's generative model, with particular structure between the variables. For example, in bivariate SCM (2.3), let $f_3(E_1, E_2) = f_2(f_1(E_1), E_2)$ and $\mathbf{f} = (f_1, f_3)$, the SCM can be written as $(X_1, X_2) = \mathbf{f}(E_1, E_2)$. Now if $\mathbf{f}$ is differentiable and invertible on $\mathbf{R}^2$, the SCM is analyzable.

For analyzable SCM, if we can solve the corresponding nonlinear ICA problem, we obtain the hidden variables $\mathbf{E} = \mathbf{g}(\mathbf{X})$. In bivariate case, given E_1 and E_2, under causal Markov and faithfulness assumptions (Spirtes and Zhang, 2016), we can conclude:

$$\begin{aligned}&X_1 \to X_2 \text{ if } X_1 \perp\!\!\!\perp E_2, \\ &X_2 \to X_1 \text{ if } X_2 \perp\!\!\!\perp E_1\end{aligned} \tag{2.8}$$

This criteria was exploited by many classical methods, e.g. LiNGAM and ANM, and can be easily understood as the independence of noise and cause.

[2]This does not imply such a strong restriction as it would seem. See Sec. 6.8.2

Nonlinear ICA violates causal faithfulness assumption Causal Markov and faithfulness assumptions are common in causal discovery literature, and we also require them in our theorem. However, we should note that causal faithfulness assumption is violated for a realized bivariate nonlinear ICA, because $X_1 \not\perp\!\!\!\perp X_2$ and the nonlinear ICA procedure necessarily has one of the following graphical models:

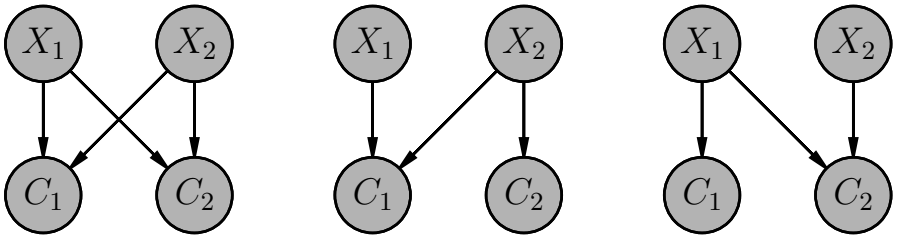

FIGURE 2.2: Graphs of nonlinear ICA procedure.

None of them induce $C_1 \perp\!\!\!\perp C_2$ under causal faithfulness assumption.

Chapter 3

Literature Review

The broadness of causality studies is well represented by Pearl (2009)'s encyclopedic monograph, which spans causal discovery, causal identification, interventional analysis, and counterfactual reasoning. Given the importance and broadness of Pearl's work, a brief review is given here as a general starting point for this section. Readers of historical interest are referred to Geffner, Dechter, and Halpern, 2022, with an annotated bibliography and four introductions by Pearl himself.

Pearl's interests in causality started from his work on Bayesian networks (Pearl, 1988) which later found applications in causality (Verma and Pearl, 1988; Pearl and Verma, 1991). His work started from causal discovery (Verma and Pearl, 1990) and was influenced by the work of some computational-oriented philosophers (Glymour, Scheines, and Spirtes, 1987). Later, his work touched identification of causal effects, i.e., the famous back-door criterion (Pearl, 1993), semantics of counterfactuals (Balke and Pearl, 1994), and mediation analysis (Pearl, 2001), all of which are based on a graphical language. By far, his work has influenced statistics (Drton and Maathuis, 2017), biostatistics (Greenland, Pearl, and Robins, 1999), econometrics (Imbens, 2020), and, of course, machine learning (Schölkopf et al., 2021; Kaddour et al., 2022).

Below, works on the two problems tackled in this thesis are reviewed specifically.

3.1 Treatment Effect Estimation

Under unconfoundedness assumption, the problem of covariate imbalance is traditionally addressed by *balancing methods*, including matching and re-weighting (Stuart, 2010; Rosenbaum, 2020), because adjusting for imbalance in treatment assignment controls bias in treatment effect estimation. In matching methods, similar subjects in the control (treatment) group are found–that is "matched to"–a subject in the treatment (control) group and used as a sample to infer the treated (controlled) subject's potential outcome. There are, to name a few, Mahalanobis matching (Rubin, 1979), propensity score matching (Rosenbaum and Rubin, 1983), full matching (Rosenbaum, 1991), fine balancing (Rosenbaum, Ross, and Silber, 2007), and adaptive hyper-box matching (Morucci et al., 2020). Re-weighting methods balance treatment and control groups by weighting subjects of both groups. The seminal method is inverse propensity weighting (IPW) (Rosenbaum, 1987). To avoid extreme weight values, there are also stabilized weighting (Cole and Hernán, 2008), trimmed weighting (Lee, Lessler, and Stuart, 2011), and overlap weighting (Li, Morgan, and Zaslavsky, 2018).

With the nonparametric statistics and machine learning, there comes *regression methods*. Recall the motivation behind balancing methods is that, in an RCT, or if the propensity score is properly estimated, we can avoid, or relax the assumptions on, modeling the response surfaces $\mu_t(X)$ which might be arbitrary nonlinear and multivariate functions. Conversely, regression methods aim to model the response surfaces precisely, using flexible regression models, without propensity score estimation. This is why nonparametric or machine learning models are considered, e.g., regression trees (Hill, 2011; Athey and Imbens, 2016) and random forests (Wager and Athey, 2018).

Mixed (double) methods combine balancing and regression, because, as we have seen, both are useful for controlling the bias in treatment effect estimation. While many machine learning methods, including ours, fall into this category because they have flexible outcome regressions, the line of work exists in fact long before

the coming of machine learning, for example, in the form of regression with propensity adjustment (Rosenbaum and Rubin, 1983). Also, doubly robust estimators (Cassel, Särndal, and Wretman, 1976; Robins, Rotnitzky, and Zhao, 1994) are consistent if either the propensity estimation or the outcome estimation is consistent and can possibly use machine learning for both estimators (Chernozhukov et al., 2018). Another benefit of the combination is to debias machine learning regressions and get $\sqrt{N}$-consistency, possibly without propensity estimation (Athey, Imbens, and Wager, 2018). Further, double/debiased machine learning (DML) (Chernozhukov et al., 2018) provides a semi-parametric framework, not limited to causal effects as target parameters, exploiting machine learning in estimating nuisance parameters while obtaining $\sqrt{N}$-consistency. We note that, for machine learning methods, balancing is often achieved by a regularization term penalizing the imbalance, and this is true for most of the BRL methods mentioned below, including ours.

Below, we focus on several lines of work that are particularly related to aspects of our method.

Limited overlap. Under limited overlap, Luo, Zhu, and Ghosh (2017) estimate the ATE by reducing covariates to a linear prognostic score. Farrell (2015) estimates a constant treatment effect under a partial linear outcome model. D'Amour and 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 and 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 and Chan (2017) estimate individualized treatment effect by reducing covariates to a linear score which is a joint propensity-prognostic score. Tarr and 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, Liu, and Li (2020) and Hassanpour and 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 treatment effect 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., 2020b), and energy models (Khemakhem et al., 2020a). The results are exploited in causal discovery (Wu and Fukumizu, 2020a) and out-of-distribution (OOD) generalization (Sun et al., 2020). This study is the first to explore identifiable representations in treatment effect identification.

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

Causal inference with auxiliary structures. CEVAE (Louizos et al., 2017) relies on the strong assumption that the true confounder distribution can be recovered from proxies. Our method is quite different in motivation, applicability, architecture. Detailed comparisons are given in Sec. 3.1.1. Also with proxies, Kallus, Mao, and Udell, 2018 use matrix factorization to infer the confounders, and Mastouri et al., 2021 use kernel methods to solve the underlying Fredholm integral equation. IVs are also exploited in machine learning, there are methods using deep NNs (Hartford et al., 2017) and kernels (Singh, Sahani, and Gretton, 2019; Muandet et al., 2019).

Our work lays conceptual and theoretical foundations of VAE methods for treatment effects (e.g., CEVAE Louizos et al., 2017; Lu et al., 2020), see Section 5.3. In Section 3.1.1, we also make detailed comparisons to CFR and CEVAE, which are well-known machine learning methods. In addition, some studies consider monotonicity, which is injectivity on $\mathbb{R}$, together with overlap, and this is discussed in detail below.

3.1.1 Detailed Comparisons

Comparisons with and Criticisms of CEVAE

Motivation CEVAE is motivated by exploiting proxy variables, and its intuition is that the hidden confounder U can be recovered by VAE from proxy variables.

Our method is motivated by prognostic scores (Hansen, 2008), and our model is directly based on equations (5.3) which identifies CATE. There is no need to recover the hidden confounder in our framework.

Architecture Our model is naturally based on (5.3), particularly the independence properties of prognostic score. And as a consequence, our VAE architecture is a natural combination of iVAE and CVAE (see Figure 4.1). Our ELBO (4.3) is derived by the standard variational lower bound.

On the other hand, the architecture of CEVAE is more ad hoc and complex. Its decoder follows the graphical model of descendant proxy mentioned above, but adds an ad hoc component to mimic TARnet (Shalit, Johansson, and Sontag, 2017): it uses separated NNs for the two potential outcomes. We tried this idea on the IHDP dataset, and, as we show in Sec. 4.4.2, it has basically no merits for our method, because we have a principled way for balancing.

The encoder of CEVAE is even more complex. To have post-treatment estimation, $q(T|X)$ and $q(Y|X,T)$ are added into the encoder. As a result, the ELBO of CEVAE has two additional likelihood terms corresponding to the two distributions. However, in our Intact-VAE, post-treatment estimation is given naturally by our standard encoder, thanks to the correspondence between our model and (5.3).

Justification We have given the identifications and bounds of our method in Chapter 4. Moreover, we carefully distinguish assumptions on the DGP and assumptions on our model, and identify the assumptions that are important for causality. There are few theoretical justifications for CEVAE. Their Theorem 1 directly assumes the joint distribution $p(\boldsymbol{x}, \boldsymbol{y}, t, \boldsymbol{u})$ including hidden confounder U is recovered, then identification is trivial by using the standard adjustment equation.

However, the challenge is exactly that the confounder is hidden, unobserved. Many years of work have been done in causal inference to derive conditions under

which hidden confounder can be (partially) recovered (Greenland, 1980; Kuroki and Pearl, 2014; Miao, Geng, and Tchetgen Tchetgen, 2018). In particular, Miao, Geng, and Tchetgen Tchetgen, 2018 gives the most recent identification result for proxy setting, which requires very specific two proxies structure, and other completeness assumptions on distributions. Thus, it is unreasonable to believe that VAE, with simple descendant proxies, can recover the hidden confounder. Indeed, Rissanen and Marttinen, 2021 recently give evidence that the method often fails.

Moreover, the identifiability of VAE itself is a challenging problem. As mentioned in Introduction, Khemakhem et al., 2020b is the first identifiability result for VAE, but it only identifies an equivalence class, not a unique representation function. Thus, it is also unconvincing that VAE can learn a unique latent distribution, without certain assumptions. As we show in Sec. 4.4.1, for relatively simple synthetic datasets, CEVAE can not robustly recover the hidden confounder, even only up to transformation, while our method can (though, again, this is not needed for our method).

Comparisons with CFR

Our method is related to CFR in two ways. Theoretically, our bounds in Sec. 4.3.2 resemble those in Shalit, Johansson, and Sontag, 2017. But we bound CATE error, while CFR bounds PEHE; thus, our bounds give conditional balancing while CFR only has unconditional balancing. See Sec. 4.7.5 for more on the bounds. Conceptually, CFR is loosely related to our method because it also learns a representation as an outcome predictor, as mentioned in the follow-up Johansson et al., 2020. However, CFR does not have a generative model, so their representation is not formally related to prognostic scores. Moreover, CFR does not account the outcome noise, while the uncertainty due to the noise is accounted by our VAE.

3.1.2 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 and Wüthrich, 2018). More related to our method is another line of work based on *monotonicity of outcome*, see (Chernozhukov and Hansen, 2013) and references therein for early results. Some recent works apply monotonicity of outcome to nonparametric IV regression (NPIV) (Freyberger and Horowitz, 2015; Li, Liu, and Li, 2017; Chetverikov and 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 and Wilhelm, 2017) combines monotonicity of both treatment and outcome, and (Freyberger and 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 prognostic score in our method, and the prognostic score is recovered from observables. Finally, as we mentioned in Sec. 4.2.3, monotonicity is a kind of shape restriction which also includes, e.g., concavity and symmetry and attracts recent interests (Chetverikov, Santos, and Shaikh, 2018). However, most of NPIV works focus on identifying f but not directly on treatment effects, and we do not know any works that use monotonicity to address limited overlap.

Recently in machine learning, (Johansson, Sontag, and Ranganath, 2019; Zhang, Bellot, and Schaar, 2020; 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, Sontag, and Ranganath, 2019) explicitly discuss overlap and invertibility, but does not focus on treatment effects. (Zhang, Bellot, and Schaar, 2020) 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 prognostic score and representation function is a sufficient condition, and our outcome model is injective, to *preserve the exchangeability given the prognostic score*. Thus, our method works under more general setting, and arguably under weaker conditions.

3.2 Causal Discovery

3.2.1 Causal Structure Learning

Traditionally, causal discovery algorithms learn causal structure of a directed acyclic graphical (DAG) model, by searching in the space of possible DAGs (Spirtes et al., 2000). Constraint-based search methods, under causal Markov assumption (Spirtes and Zhang, 2016), use conditional independence test to determine causal structure. Among them, IC (Verma and Pearl, 1990) and PC (Spirtes and Glymour, 1991) algorithms are early examples. Later, FCI (Spirtes, Meek, and Richardson, 1999) and its improvements–e.g. RFCI (Colombo et al., 2012), RCI+ (Claassen, Mooij, and Heskes, 2013)–work under the presence of confounders, but can only output Markov equivalent class of DAGs, in which some causal directions are undetermined. Score-based search methods, such as GES (Chickering, 2002), search, usually greedily, for a graph that optimizes a penalized likelihood score. They are fast and robust on small samples and low-dimensional case, but assume no confounder. Hybrid methods, which combine constraint- based and score-based methods, could deal with confounders and be more accurate than constraint-based methods in certain case (Ogarrio, Spirtes, and Ramsey, 2016). However, none of score-based or hybrid methods are able to fully determine causal directions since they assign same scores to all DAGs within the same Markov equivalence class.

3.2.2 Bivariate Causal Discovery

In recent years, a line of research emerges that is particularly motivated to solve the problem of distinguishing cause from effect. Intuitively, cause-effect relationship is asymmetric in the sense that recovering the cause from effect is more complex than modeling the physical process that generates effect from cause. All the methods exploit this asymmetry to identify causal direction (Mooij et al., 2016). For

example, many methods define certain simple, restricted class of functional forms (sometimes referred to as "functional causal models" (FCMs) (Hyvärinen and Zhang, 2016)). Typical FCMs are LiNGAM (Shimizu et al., 2006), ANM (Hoyer et al., 2009) and PNL (Zhang and Hyvärinen, 2009); each has more general functional form than the former, and, thus, is more widely applicable. In particular, ANM is the first to use nonlinear additive noise model in this problem. These methods advance to deal with some harder cases, such as cyclic (Mooij et al., 2011), multivariate (Peters et al., 2014), and discrete (Peters, Janzing, and Scholkopf, 2011).

Other methods exist, and most of them exploits the asymmetry in another way: the so-called "principle of independent causal mechanism (ICM)", which postulates that the process generating cause distribution is in some way "independent" to the causal mechanism generating conditional distribution of the effect given the cause. For example, Janzing et al., 2012, (IGCI) use orthogonality in information space to express the independence between two distributions. It is applicable when the causal relation is deterministic (noise-free). Blöbaum et al., 2018, (RECI) extend IGCI to the setting with small noise, and proceeds by comparing the regression errors in both possible directions. Stegle et al., 2010 do not restrict the class of causal models, and particularly, the noise need not be additive. It explicitly models the "noise" as a latent variable that summarizes unobserved causes, but still assumes the independence of the noise and the observed cause. Both Mitrovic, Sejdinovic, and Teh, 2018, (KCDC) and Budhathoki and Vreeken, 2017 base on the invariance of Kolmogorov complexity on the value of the cause. As computing the Kolmogorov complexity is intractable, Mitrovic, Sejdinovic, and Teh, 2018 use, after kernel mean embedding (KME), the variability in RKHS norm as a proxy for it, while Budhathoki and Vreeken, 2017 leverage stochastic complexity as an approximation. Unfortunately, all these methods assume no confounder, since the asymmetry may become invalid under hidden common causes. Identifiability in the presence of confounders using purely observational data is only well studied in the linear, non-Gaussian noise case (Shimizu, 2014; Hoyer et al., 2008).

The most related to our work in Chapter 6 might be the following. RCC (Lopez-Paz et al., 2015) and its follow-up NCC (Lopez-Paz et al., 2017) also use training data, but they require large numbers of labeled pairs and thus rely on synthetic pairs for

training. There is work which takes related viewpoints: KCDC uses majority voting, the simplest ensemble method; ANM-MM treats mechanism as a mixture. NonSENS (Monti, Zhang, and Hyvärinen, 2019) also employs the same nonlinear ICA method as ours, but needs samples of a causal system available over different environments, which requires interventions or even experiments. We should note that all the above methods neither take a mosaic view explicitly nor use ensemble method as a main building block.

Advances regarding Hidden Confounding

Until recently, there is few work that achieves identifiability under confounders, and this paragraph provides a brief review. Goudet et al., 2018 use deep generative neural networks to learn multivariate causal models, minimizing a maximum mean discrepancy (MMD) loss based on KME. And it models confounders by adding correlated noise between adjacent observed variables. But the loss is to some extent ad-hoc and has a hyperparameter. Zhang, Zhang, and Schölkopf, 2015 test for exogeneity with bootstrap and infers the causal direction, or the existence of confounder if exogeneity holds for neither directions. But it gives non-identifiable result if exogeneity holds for both directions. Moreover, exogeneity is at best a necessary condition of direct causal relation. Chalupka, Eberhardt, and Perona, 2016 work in the discrete bivariate case. In the presence of confounder, it trains a neural network by synthesized data drawn from uninformative Dirichlet prior. Both Rothenhäusler, Bühlmann, and Meinshausen, 2019 and Rothenhäusler et al., 2015 exploit the invariance of causal mechanism under specific type of additive intervention and need data from multiple environments with distinct, but maybe unknown, interventions. And the causal models are both assumed to be linear, with additive noise. They allow non-diagonal covariance matrix of the noise, which indicates possible latent variables (including confounders). Lopez-Paz et al., 2015 do not use common assumptions from causal inference. Instead, it casts causal inference as classification of probability distributions, in RKHS after KME. The train data are samples from joint distributions of pairs of variables, labeled by their true causal relations (which may include "confounded"). Each of these methods is based on at least one of the following: ad-hoc devices, rather than, preferably theoretical, justification by general

principles; too strong or dubious assumptions that limit its application; additional information other than observational data, such as labels of true causal relations, or data from multiple environments involving interventions. In Section 7.2.1 we discuss some ideas to avoid these weaknesses.

Chapter 4

Intact-VAE: Treatment Effect Estimation under Limited Overlap

4.1 Intuition and Data Generating Process

We use balanced prognostic score or prognostic score to construct representations for CATE estimation. **Why not balancing scores?** While balancing scores $\boldsymbol{b}(X)$ have been widely used in causal inference, prognostic scores 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 balanced prognostic score does not imply overlapping $\boldsymbol{b}(X)$. **Example.** 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 balanced prognostic score; and $|X|$ is overlapping but X is not. Moreover, with theoretical and experimental evidence, it is recently conjectured that prognostic scores maximize overlap among a class of sufficient scores, including $\boldsymbol{b}(X)$ (D'Amour and 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 $\mathfrak{p}_t(X)$ satisfies exchangeability.

Proposition 4 (Identification via prognostic score)**.** *If $\mathfrak{p}_t(X)$ is a prognostic score and $Y|\mathfrak{p}_{\hat{t}}(X), T \sim p_{Y|\mathfrak{p}_{\hat{t}},T}(\boldsymbol{y}|P, t)$ where $\hat{t} \in \{0, 1\}$ is a counterfactual assignment, then CATE*

and ATE are identified, using (2.1) *and*

$$\mu_{\hat{t}}(\boldsymbol{x}) = \mathbb{E}(Y(\hat{t})|\mathfrak{p}_{\hat{t}}(X), X = \boldsymbol{x}) = \mathbb{E}(Y|\mathfrak{p}_{\hat{t}}(\boldsymbol{x}), T = \hat{t}) = \int p_{Y|\mathfrak{p}_{\hat{t}},T}(\boldsymbol{y}|\mathfrak{p}_{\hat{t}}(\boldsymbol{x}), \hat{t})\boldsymbol{y}d\boldsymbol{y} \quad (4.1)$$

With the knowledge of $\mathfrak{p}_t$ and $p_{Y|\mathfrak{p}_{\hat{t}},T}$, we choose one of $\mathfrak{p}_0, \mathfrak{p}_1$ 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 consider additive noise models for $Y(t)$, which ensures the existence of prognostic scores.

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

The DGP is causal and defines potential outcomes by $Y(t) := \boldsymbol{f}_t^*(\mathfrak{m}_t(X)) + \mathbf{e}$, and specifies $\mathfrak{m}(X, T)$, T, and $\mathbf{e}$ as the only direct causes of Y. Particularly, $\mathfrak{m}_t(X)$ is a sufficient statistics of X for $Y(t)$. For example, 1) $\mathfrak{m}_t(X)$ can be the component(s) of X that affect $Y(t)$ directly, or 2) if $Y(t)|X$ follows a generalized linear model, then $\mathfrak{m}_t(X)$ can be the linear predictor of $Y(t)$.

Under **(G1)**, 1) $\mathfrak{m}_t(X)$ is a prognostic score; 2) $\mu_t(X) = \boldsymbol{f}_t^*(\mathfrak{m}_t(X))$ is a prognostic score; 3) X is a (trivial) balanced prognostic score; and 4) $\mathfrak{u}(X) := (\mu_0(X), \mu_1(X))$ is a balanced prognostic score. The **essence of our method** is to recover the prognostic score $\mathfrak{m}_t(X)$ as a representation, assuming $\mathfrak{m}_t(X)$ is not higher-dimensional than Y and approximately balanced. Note that $\mu_t(X)$, our final target, is a low-dimensional prognostic score but not balanced, and we estimate it conditioning on the approximate balanced prognostic score $\mathfrak{m}_t(X)$.

4.2 Identification under Generative Prognostic Model

In Sec. 4.2.1, we specify the generative prognostic model $p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}, t)$, and show its identifiability. In Sec. 4.2.3, we prove the identification of CATEs, which is one of our main contributions. The theoretical analysis involves only our generative model

[1] The labels **G**, **M**, or **D** mean Generating process (of Y), probabilistic Model, or Distribution (of X). We introduce assumptions when appropriate but compile them in one place in Sec. 4.7.1.

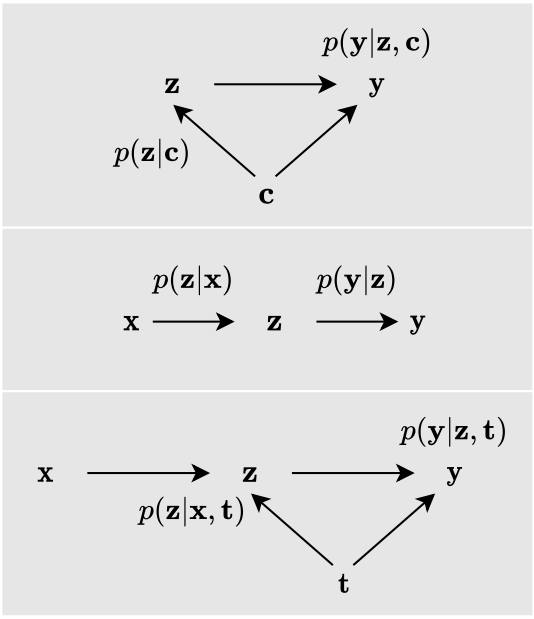

FIGURE 4.1: CVAE, iVAE, and Intact-VAE: Graphical models of the decoders.

(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. 4.3.

4.2.1 Model, Architecture, and Identifiability

Our goal is to build a model that can be learned by VAE from observational data to obtain a prognostic score, or better, a balanced prognostic score, via the latent variable Z. The generative prognostic model of the proposed method is in (4.2), where $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{k})$ contains the functional parameters. The first factor $p_f(\boldsymbol{y}|\boldsymbol{z}, t)$, our decoder, models $p_{Y|\mathfrak{p}_t,T}(\boldsymbol{y}|P, t)$ in (5.3) and is an additive noise model, with $\boldsymbol{\epsilon} \sim p_\epsilon$ as the exogenous noise. The second factor $p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t)$, our conditional prior, models $\mathfrak{p}_T(X)$ and is a factorized Gaussian, with $\boldsymbol{\lambda}_T(X) := \text{diag}^{-1}(\boldsymbol{k}_T(X))(\boldsymbol{h}_T(X), -\frac{1}{2})^T$ as its natural parameter in the exponential family, where diag() gives a diagonal matrix from a vector.

$$\begin{gathered} p_\theta(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}, t) = p_f(\boldsymbol{y}|\boldsymbol{z}, t)p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t), \\ p_f(\boldsymbol{y}|\boldsymbol{z}, t) = p_\epsilon(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z})), \quad p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t) \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{h}_t(\boldsymbol{x}), \text{diag}(\boldsymbol{k}_t(\boldsymbol{x}))). \end{gathered} \tag{4.2}$$

We denote $n := \dim(Z)$. For inference, the ELBO is given by the standard variational lower bound

$$\log p(\boldsymbol{y}|\boldsymbol{x}, t) \geq \mathbb{E}_{z\sim q} \log p_f(\boldsymbol{y}|\boldsymbol{z}, t) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t)\|p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t)). \tag{4.3}$$

Note that the encoder q conditions on all the observables (X, Y, T); this fact plays an important role in Sec. 4.3.1. Full parameterization of the encoder and decoder is also given in Sec. 4.3.1. This architecture is called *Intact-VAE* (*Iden*tifiable *trea*tment-*condi*tional VAE). See Figure 4.1 for comparison in terms of graphical models (which have *not* causal implications here). See Sec. 4.2.2 for more expositions and Sec. 2.2.1 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 $\boldsymbol{f}_t$:

$$\boldsymbol{f}_t^{-1}(\boldsymbol{y}_t) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t) + \boldsymbol{b} =: \mathcal{A}_t(\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t)) \tag{4.4}$$

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. 4.6 for the proof and a relaxation of **(D1)**. In Chapter 4 & 5, symbol $'$ (prime) always indicates another parameter (variable, etc.): $\boldsymbol{\theta}' = (\boldsymbol{f}', \boldsymbol{\lambda}')$.

4.2.2 Details and Explanations on Intact-VAE

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

With the above goal, the generative model of our VAE is built as (4.2). 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:

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{x}, t) &\geq \log p(\boldsymbol{y}|\boldsymbol{x}, t) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) \| p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t)) \\ &= \mathbb{E}_{z\sim q} \log p(\boldsymbol{y}|\boldsymbol{z}, t) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, t) \| p(\boldsymbol{z}|\boldsymbol{x}, t)). \end{aligned} \tag{4.5}$$

We naturally have an identifiable conditional VAE (CVAE), as the name suggests. Note that (4.2) has a similar factorization with the generative model of iVAE (Khemakhem et al., 2020b), 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 (4.2), our VAE architecture is a combination of iVAE and CVAE (Sohn, Lee, and Yan, 2015; Kingma et al., 2014), with T as the conditioning variable. See Figure 4.1 for the comparison in terms of graphical models. The core idea of iVAE is reflected in our model identifiability (see Lemma 1).

Please do not confuse the DGP **(G1)** and the generative model (4.2) of Intact-VAE. The former is the causal model, but the latter is not (at least before we show the treatment effect identifications in Sec. 4.2.3). In our case, the generative model is built as a way to learn the scores through the correspondence to (5.3).

In particular, note that conditionally balanced representation $Z \perp\!\!\!\perp T|X$ is possible under the generative model. This requires a violation of *causal faithfulness*, so that there are other conditional independence relations, which are not generally implied by the graphical model. Our method, based on iVAE, which achieves ICA, performs nonlinear ICA to recover the scores. In fact, ICA procedures often violate causal faithfulness, because it requires finding causes from effects. Also, the violation of causal faithfulness is not caused by the generative model (which is shown in Figure 4.1), because the representation is learned by the encoder, and $Z \perp\!\!\!\perp T|X$ is enforced by β.

4.2.3 Identifications under Limited-overlapping Covariate

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

prognostic scores with dimensionality lower than or equal to $d = \dim(Y)$ are essential to address limited overlapping, as shown below. We set $n = d$ because μ_t is a prognostic score of the same dimension as Y under **(G1)**. In practice, $n = d$ means that we seek a low-dimensional representation of X. We introduce

(G1') (Low-dimensional prognostic score) **(G1)** is true, and $\mu_t = \boldsymbol{j}_t \circ \mathfrak{p}_t$ for some $\mathfrak{p}_t$ and injective $\boldsymbol{j}_t$,

which is equivalent to **(G1)** because $\mu_t = \boldsymbol{j}_t \circ \mathfrak{p}_t$ is trivially satisfied with $\boldsymbol{j}_t$ is identity and $\mathfrak{p}_t = \mu_t$. **(G1')** is used instead in this subsection. First, it explicitly restricts $\dim(\mathfrak{p}_t)$ via injectivity, which ensures that $n = \dim(Y) \geq \dim(\mathfrak{p}_t)$. Second, it reminds us that, possibly, the decomposition is not unique; and, clearly, all $\mathfrak{p}_t$ that satisfy **(G1')** are prognostic scores. For example, if $\boldsymbol{f}_t^*$ is injective, then $\boldsymbol{j}_t = \boldsymbol{f}_t^*$ and $\mathfrak{p}_t = \mathfrak{m}_t$ satisfies $\mu_t = \boldsymbol{j}_t \circ \mathfrak{p}_t$. Finally, it is then natural to introduce

(G2) (Low-dimensional balanced prognostic score) **(G1)** is true, and $\mu_t = \boldsymbol{j}_t \circ \mathfrak{p}$ for some $\mathfrak{p}$ and injective $\boldsymbol{j}_t$,

which is stronger than **(G1)**, gives balanced prognostic score $\mathfrak{p}(X)$, and ensures that $n \geq \dim(\mathfrak{p})$. **(G2)** is satisfied if $\boldsymbol{f}_t^*$ is injective and $\mathfrak{m}_0 = \mathfrak{m}_1$. **(G2)** implies $\mu_1 = \boldsymbol{i} \circ \mu_0$ where $\boldsymbol{i} := \boldsymbol{j}_1 \circ \boldsymbol{j}_0^{-1}$; in words, CATEs are given by μ_0 and an invertible function. See Sec. 4.7.2 for real-world examples and more discussions.

With **(G1')** or **(G2)**, overlapping X can be relaxed to overlapping balanced prognostic score or prognostic score plus the following:

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

Note that **(M2)** is only required for the optimal $\boldsymbol{h}$ specified in Proposition 5 or Theorem 1. The intuition is that $\mathfrak{p}_t$ maps each non-overlapping $\boldsymbol{x}$ to an overlapping value, and $\boldsymbol{h}_t$ preserves this property through learning. This is non-trivial because, for a

given t, some values of X are unobserved due to limited overlap. Thus, **(M2)** can be seen as a weak form of OOD generalization: the NNs for $\boldsymbol{h}$ can learn the OOD score partition. While unnecessary for us, linear $\mathfrak{p}_t$ and $\boldsymbol{h}_t$ trivially imply **(M2)** and are often assumed, e.g., in Huang and Chan, 2017; Luo, Zhu, and Ghosh, 2017; D'Amour and Franks, 2021.

Our first identification, Proposition 5, relies on **(G2)** and our generative model, *without* model identifiability (so differentiable $\boldsymbol{f}_t$ is not needed).

Proposition 5 (Identification via recovery of balanced prognostic score)**.** *Suppose we have DGP* ***(G2)*** *and model* (4.2) *with* $n = d$*. Assume* ***(M1)****-i) and* **(M3)** (PS matching) let $\boldsymbol{h}_0(X) = \boldsymbol{h}_1(X)$ and $\boldsymbol{k}(X) = \mathbf{0}$. *Then, if* $\mathbb{E}_{p_\theta}(Y|X,T) = \mathbb{E}(Y|X,T)$*, we have*

1) *(Recovery of balanced prognostic score)* $\boldsymbol{z}_{\lambda,t} = \boldsymbol{h}_t(\boldsymbol{x}) = \boldsymbol{v}(\mathfrak{p}(\boldsymbol{x}))$ *on overlapping* $\boldsymbol{x}$*, where* $\boldsymbol{v} : \mathcal{P} \to \mathbb{R}^n$ *is an injective function, and* $\mathcal{P} := \{\mathfrak{p}(\boldsymbol{x})|\text{overlapping } \boldsymbol{x}\}$*;*

2) *(CATE identification) if* $\mathfrak{p}(X)$ *in* ***(G2)*** *is overlapping, and* ***(M2)*** *is satisfied, then* $\mu_t(\boldsymbol{x}) = \hat{\mu}_t(\boldsymbol{x}) := \mathbb{E}_{p_\lambda(Z|\boldsymbol{x},t)}\mathbb{E}_{p_f}(Y|Z,t) = \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))$*, for any* $t \in \{0,1\}$ *and* $\boldsymbol{x} \in \mathcal{X}$*.*

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 $\boldsymbol{h} = \boldsymbol{v} \circ \mathfrak{p}$; and ii) $\mathfrak{p}_t$ is recovered up to $\boldsymbol{v}$, and $Y(t) \perp\!\!\!\perp X|\mathfrak{p}_t(X)$ is preserved—with *same* $\boldsymbol{v}$ for both t. Theorem 1 below also achieves the essence i) and ii), under $\mathfrak{p}_0 \neq \mathfrak{p}_1$.

The existence of balanced prognostic score is preferred, because it satisfies overlap and **(M2)** more easily than prognostic score which requires the conditions for each of the two functions of prognostic score. However, the existence of low-dimensional balanced prognostic score 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 prognostic score which generally exists.

Theorem 1 (Identification via recovery of prognostic score)**.** *Suppose we have DGP* ***(G1')*** *and model* (4.2) *with* $n = d$*. For the model, assume* ***(M1)*** *and* **(M3')** (Noise matching) let $p_{\boldsymbol{\epsilon}} = p_{\boldsymbol{\epsilon}}$ 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 (4.4)*. Then, if* $p_\theta(\boldsymbol{y}|\boldsymbol{x},t) = p(\boldsymbol{y}|\boldsymbol{x},t)$*; conclusions 1) and 2) in Proposition 5 hold with* $\mathfrak{p}$ *replaced with* $\mathfrak{p}_t$ *in* ***(G1')****; and the domain of* $\boldsymbol{v}$ *becomes* $\mathcal{P} := \{\mathfrak{p}_t(\boldsymbol{x})|p(t,\boldsymbol{x}) > 0\}$*.*

Theorem 1 implies that, without balanced prognostic score, we need to know or learn the distribution of hidden noise $\boldsymbol{\epsilon}$ to have $p_{\mathbf{e}} = p_{\boldsymbol{\epsilon}}$. Proposition 5 and Theorem 1 achieve recovery and identification in a complementary manner; the former starts from the prior by $\mathfrak{p}_0 = \mathfrak{p}_1$ 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 $\mathfrak{p}_0 = \mathfrak{p}_1$ in Proposition 5. We show in Sec. 4.6 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 (4.4). See Sec. 4.7.3 for more on the complementarity between the two identifications.

4.3 Estimation by β-Intact-VAE

4.3.1 Prior as balanced prognostic score, Posterior as prognostic score, and β as Regularization Strength

In Sec. 4.2.3, we see that the existence of balanced prognostic score (Proposition 5) is preferable in identifying the true DGP up to an equivalent expression—while Theorem 1 allows us to deal with prognostic score by adding other conditions. In learning our model with data, we formally require **(G1)** and further expect that **(G2)** holds approximately; the latter is true when $\boldsymbol{f}_t^*$ is injective and $\mathfrak{m}_0 \approx \mathfrak{m}_1$ ($\mathfrak{m}_t(X)$ is an approximate balanced prognostic score). Instead of the trivial regression $\mu_t(X) = \mathbb{E}(Y|X, T = t)$, we want to recover the approximate balanced prognostic score $\mathfrak{m}_t(X)$. This idea is common in practice. For example, in a real-world nutrition study (Huang and Chan, 2017), a reduction of 11 covariates recovers a 1-dimensional linear balanced prognostic score.

We consider two ways to recover an approximate balanced prognostic score by a VAE. One is to use a prior which does not depend on t, indicating a preference for balanced prognostic score. Namely, we set $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$, denote $\boldsymbol{\Lambda}(X) \coloneqq \boldsymbol{\lambda}(X)$ and have $p_{\boldsymbol{\Lambda}}(\boldsymbol{z}|\boldsymbol{x})$ as the prior in (4.2). The decoder and encoder are factorized Gaussians:

$$p_{\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{y}|\boldsymbol{z},t) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}_t(\boldsymbol{z}), \operatorname{diag}(\boldsymbol{g}_t(\boldsymbol{z}))), \quad q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y},t) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{r}_t(\boldsymbol{x},\boldsymbol{y}), \operatorname{diag}(\boldsymbol{s}_t(\boldsymbol{x},\boldsymbol{y}))), \tag{4.6}$$

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

$$\mathbb{E}_{\mathcal{D}}\{-\beta D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}\|p_{\boldsymbol{\Lambda}}) - \mathbb{E}_{z\sim q_{\boldsymbol{\phi}}}[(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2/2\boldsymbol{g}_t(\boldsymbol{z})] - \mathbb{E}_{z\sim q_{\boldsymbol{\phi}}}\log|\boldsymbol{g}_t(\boldsymbol{z})|\}. \tag{4.7}$$

For convenience, here and in $\mathcal{L}_f$ in Sec. 4.3.2, we omit the summation as if Y is univariate. The encoder $q_{\boldsymbol{\phi}}$ depends on t and can realize a prognostic score. With β, 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 β encourages the conditional balance $Z \perp\!\!\!\perp T|X$ on the posterior. By choosing β appropriately, e.g., by validation, the ELBO can recover an approximate balanced prognostic score while fitting the outcome well. In summary, we base the estimation on Proposition 5 and balanced prognostic score as much as possible, but step into Theorem 1 and noise modeling required by $p_{\boldsymbol{\epsilon}} = p_{\boldsymbol{\epsilon}}$ 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. 4.2.3, but it is more practical and yields better results in our experiments. See Sec. 4.7.4 for more ideas and connections behind the ELBO.

Once the VAE is learned[2] by the ELBO, the estimate of the expected potential outcomes is given by:

$$\hat{\mu}_{\hat{t}}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\boldsymbol{f}_{\hat{t}}(\boldsymbol{z}) = \mathbb{E}_{\mathcal{D}|\boldsymbol{x}\sim p(\boldsymbol{y},t|\boldsymbol{x})}\mathbb{E}_{z\sim q_{\boldsymbol{\phi}}}\boldsymbol{f}_{\hat{t}}(\boldsymbol{z}),\ \hat{t} \in \{0,1\}, \tag{4.8}$$

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 (4.8) (see Sec. 4.3.1 for details and 4.5 for results). Note that $\hat{t}$ in (4.8) indicates a counterfactual

[2] 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. 4.3.3 for formal statements and proofs. Under Gaussian models, it is possible to prove the consistency of the posterior estimation, as shown in Bonhomme and Weidner (2021).

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. 4.3.2). The overall **algorithm** steps are i) train the VAE using (4.7), and ii) infer CATE $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$ by (4.8).

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 (4.8) and get rid of the outer average taken on $\mathcal{D}$; all the others remain the same. We also have sensible pre-treatment prediction even without true low-dimensional prognostic scores, because $p_{\boldsymbol{\Lambda}}$ gives the best balanced approximation of the target prognostic score. The results of pre-treatment prediction are given in the experimental section 4.5.

4.3.2 Conditionally Balanced Representation Learning

We formally justify our ELBO (4.7) from the BRL viewpoint. We show that the conditional BRL via the KL (first) term of the ELBO results from bounding a CATE error; particularly, the error due to the imprecise recovery of $\boldsymbol{j}_t$ in **(G1')** is controlled by the ELBO. Previous works (Shalit, Johansson, and Sontag, 2017; Lu et al., 2020) instead focus on unconditional balance and bound PEHE which is marginalized on X. Sec. 4.4.2 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. 4.7.5 for detailed comparisons to previous works. In Sec. 4.5.3, we empirically validate our bounds, and, particularly, the bounds are more useful under weaker overlap.

We introduce the objective that we bound. Using (4.8) to estimate CATE, $\hat{\tau}_f(\boldsymbol{z}) := \boldsymbol{f}_1(\boldsymbol{z}) - \boldsymbol{f}_0(\boldsymbol{z})$ 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:

$$\tau(\boldsymbol{x}) = \boldsymbol{j}_1(\mathfrak{p}_1(\boldsymbol{x})) - \boldsymbol{j}_0(\mathfrak{p}_0(\boldsymbol{x})), \quad \tau_j(\boldsymbol{z}) = \boldsymbol{j}_1(\boldsymbol{z}) - \boldsymbol{j}_0(\boldsymbol{z}), \tag{4.9}$$

where j_t is associated with an approximate balanced prognostic score $\mathfrak{p}_t$ (say, $\mathfrak{m}_t$) as the target of recovery by our VAE. Accordingly, given $\boldsymbol{x}$, the *error of posterior CATE*, with or without knowing $\mathfrak{p}_t$, is defined as

$$\epsilon_f^*(\boldsymbol{x}) \coloneqq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}(\hat{\tau}_f(\boldsymbol{z}) - \tau(\boldsymbol{x}))^2; \quad \epsilon_f(\boldsymbol{x}) \coloneqq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}(\hat{\tau}_f(\boldsymbol{z}) - \tau_j(\boldsymbol{z}))^2. \tag{4.10}$$

We bound ϵ_f instead of ϵ_f^* because the error between $\tau(X)$ and $\tau_j(Z)$ is small—if the score recovery works well, then $\boldsymbol{z} \approx \mathfrak{p}_0(\boldsymbol{x}) \approx \mathfrak{p}_1(\boldsymbol{x})$ in (4.9). We consider the error between $\hat{\tau}_f$ and τ_j below. We define the risks of outcome regression, into which ϵ_f is decomposed.

Definition 5 (CATE risks)**.** Let $Y(\hat{t})|\mathfrak{p}_{\hat{t}}(X) \sim p_{Y(\hat{t})|\mathfrak{p}_{\hat{t}}}(\boldsymbol{y}|P)$ 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:

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

With $Y(t)$ involved, $\mathcal{L}_f$ 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 $\mathcal{L}_f(\boldsymbol{z}, t)$ 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 $\mathcal{L}_f(\boldsymbol{z}, t)$ since $Y(t) \neq \boldsymbol{y}' = Y(1-t)$.

Thus, *the regression error (second) term in ELBO* (4.7) *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 $M\boldsymbol{D}(\boldsymbol{x})$ in Theorem 2 below—which, in turn, bounds ϵ_f by decomposing it to $\epsilon_{F,t}$, $\epsilon_{CF,t}$, and $\mathbf{V}_Y$.

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

$$\epsilon_f(\boldsymbol{x}) \leq 2[G(\epsilon_{F,0}(\boldsymbol{x}) + \epsilon_{F,1}(\boldsymbol{x}) + M\boldsymbol{D}(\boldsymbol{x})) - \mathbf{V}_Y(\boldsymbol{x})] \tag{4.11}$$

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

$\boldsymbol{D}(\boldsymbol{x})$ measures the imbalance between $q_t(\boldsymbol{z}|\boldsymbol{x})$ and is symmetric for t. Correspondingly, *the KL term in ELBO* (4.7) *is symmetric for* t *and balances* $q_t(\boldsymbol{z}|\boldsymbol{x})$ by encouraging $Z \perp\!\!\!\perp T|X$ for the posterior. $\mathbf{V}_Y(\boldsymbol{x})$ reflects the intrinsic variance in the DGP and can not be controlled. Estimating G, M is nontrivial. Instead, we rely on β in the ELBO (4.7) to weight the terms. We do not need two hyperparameters since G *is implicitly controlled by the third term, a norm constraint, in ELBO*.

4.3.3 Consistency of VAE and Prior Estimation

The following is a refined version of Theorem 4 in Khemakhem et al., 2020b. The result is proved by assuming: i) our VAE is flexible enough to ensure the ELBO is tight (equals to the true log likelihood) for some parameters; ii) the optimization algorithm can achieve the *global* maximum of ELBO (again equals to the log likelihood).

Proposition 6 (Consistency of Intact-VAE)**.** *Given model* (4.2)*&*(4.6)*, 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 $\mathbb{E}_{\mathcal{D}\sim p^*}(\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, t; \boldsymbol{\theta}, \boldsymbol{\phi}))$ (4.3) *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.*

4.4 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. 5.2.2. As in previous works (Shalit, Johansson, and Sontag, 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 (4.7), we use a multilayer perceptron, with $200 * 3$ hidden units (width 200, 3 layers), and ELU activations (Clevert, Unterthiner, and Hochreiter, 2015). $\boldsymbol{\Lambda} = (\boldsymbol{h}, \boldsymbol{k})$ 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.

4.4.1 Synthetic Dataset

We generate synthetic datasets following (5.4).

$$W|X \sim \mathcal{N}(\boldsymbol{h}(X), \boldsymbol{k}(X));\ T|X \sim \text{Bern}(\text{Logi}(\omega l(X)));\ Y|W, T \sim \mathcal{N}(f_T(W), g_T(W)). \tag{4.12}$$

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 balanced prognostic score, but the dimensionality is not low enough to satisfy the injectivity in **(G2)**, when $\dim(W) > 1$. We have 5 different overlap levels controlled

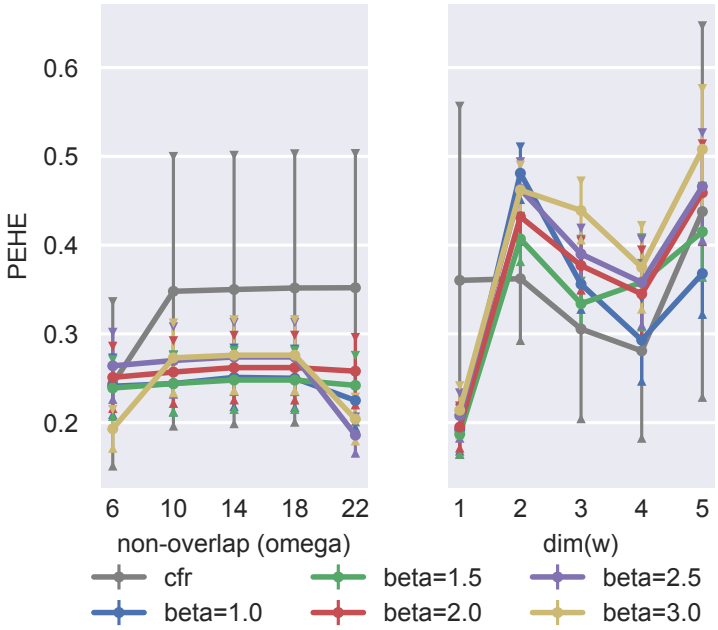

FIGURE 4.2: $\sqrt{\epsilon_{pehe}}$ on synthetic datasets. Each error bar is on 10 random DGPs.

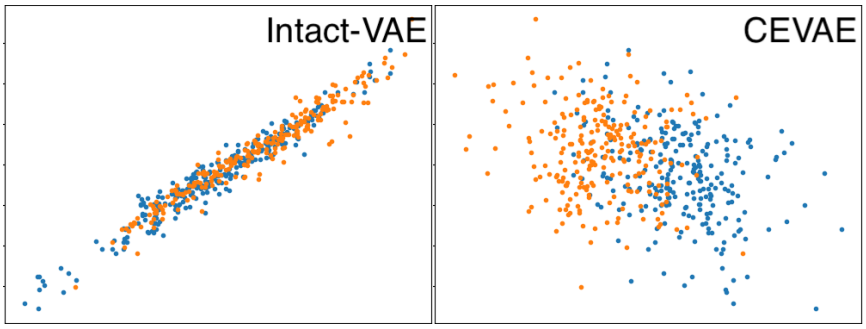

FIGURE 4.3: Plots of recovered - true latent. Blue: $T = 0$, Orange: $T = 1$.

by ω that multiplies the logit value. See Sec. 4.5.1 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 (5.4). 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 β. For CFR, we try different balancing parameters and present the best results (see Sec. 4.5.1 for detail).

In each panel of Figure 4.2, 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 4.3 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. 4.3. When $\dim(W) > 1$, f_t in (4.12) is non-injecitve and learning of prognostic score is necessary, and thus, larger β 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 Figure 4.7). Thus, the performance gap under $\dim(W) > 1$ in Figure 4.2 should be due to the capacity of NNs in β-Intact-VAE. In Figure 4.9 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. 4.4.2 for more explicit comparison).

Experiments for the score recovery When $\dim(W) = 1$, there are no better prognostic scores 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 4.3, and more plots are in Figure A.1. For comparison, we run CEVAE, 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 $p_{\lambda}(\boldsymbol{z}|\boldsymbol{x}, t)$ in Figure A.7.

To show quantitative evidence for the score recovery, we first fit a simple linear regression $W = aZ$ between the standardized true and learned score. Then we examine the linear regression in two ways–by goodness of fit through the *coefficient of determination* R^2 and model specification through the *Ramsey regression equation specification error test (RESET)* (Ramsey, 1969). Specifically, $R^2 = 1 - \Sigma_i(w_i - \hat{w}_i)^2 / \Sigma_i(w_i - \bar{w})^2$ measures how much variation of W is explained by Z in the regression, and the nearer to 1 the R^2 is, the tighter the linear fit is. Moreover, Ramsey RESET tests the null hypothesis of linearity by examine whether the combinations of $\hat{W}^2, ..., \hat{W}^k$

where $\hat{W} = \hat{a}Z$ help explain the response variable W (We set $k = 5$). Linearity is rejected if the p-value of the test is lower than a significant level α.

The Ramsey RESET test can catch the cases where the R^2 is near 1 but a small portion of data causes notable non-linearity, see Figure 4.4 left for an example. In fact, we observe that the RESET test is too sensitive to non-linearity when the R^2 is high. An example is shown in Figure 4.4 right where the non-linearity is barely notable and is possibly due to the several outliers on both sides. However, the RESET test gives a rather low p-values as 0.004. Thus, we decide that $\alpha = 0.01$ is reasonable for our purpose.

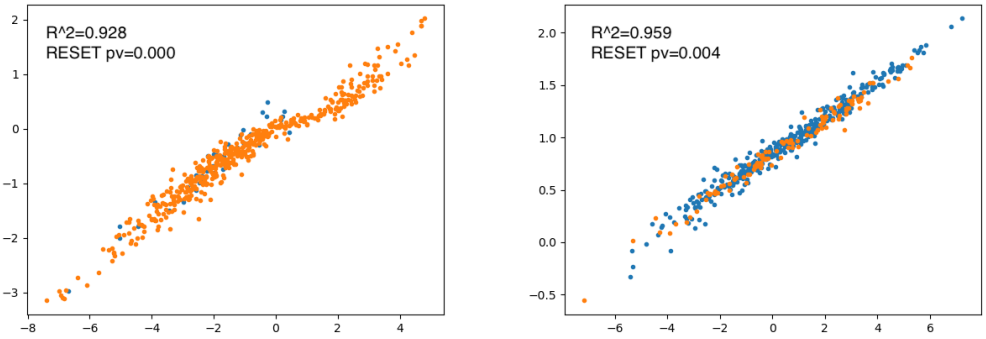

FIGURE 4.4: Examples of low p-values of RESET. Left: a notable non-linearity, and the p-value is practically 0. Right: tiny to no non-linearity, but the p-value is very low.

The histograms of the R^2 values and the RESET p-values on the 100 synthetic datasets are shown in Figure 4.5. We see linear regression often gives good fits and is not misspecified. Specifically, R^2 is higher than 0.75 on 83 datasets and higher than 0.8 on 76 datasets, and RESET p-value is higher than $\alpha = 0.01$ on 82 datasets and higher than 0.05 on 72 datasets. Finally, the two criteria taken together, there are 66 datasets where R^2 is higher than 0.75 and RESET p-value is higher than $\alpha = 0.01$–an impressive result because the two conditions tend to be mutually exclusive and many cases like those in Figure 4.4 are excluded. Thus, we conclude that the experiment quantitatively confirms the theoretical result that Intact-VAE recovers the true score up to an affine transformation.

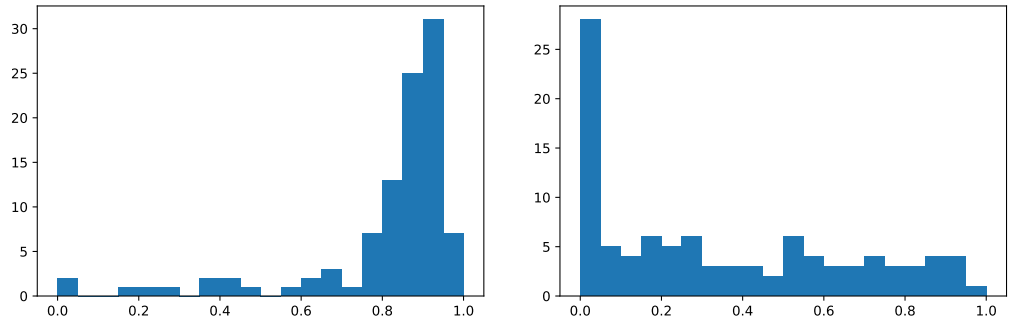

FIGURE 4.5: The histograms of R^2 (left) and RESET p-values (right) for linear regressions between the true and learned score.

4.4.2 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. 4.5.2 for detail and more results.

There is a linear balanced prognostic score (linear combination of the covariates). However, most of the covariates are binary, so the support of the balanced prognostic score is often on small and separated intervals. Thus, the Gaussian latent Z in our model is misspecified. We use higher-dimensional Z to address this, similar to Louizos et al. (2017). Specifically, we set $\dim(Z) = 50$, together with NNs of $50 * 2$ hidden units in the prior and encoder. We set $\beta = 1$ since it works well on synthetic datasets with limited overlap.

As shown in Table 4.1, β-Intact-VAE outperforms or matches the state-of-the-art methods; it has the best performance measured by both ϵ_{ate} and ϵ_{pehe} and matches CF and CFR respectively. Also notably, our method outperforms other generative models (CEVAE and GANITE) by large margins.

To show our conditional balance is preferable, we also modify our method and add two components for *unconditional* balance from CFR (see the Sec. 4.5.1), which

is based on bounding PEHE and is controlled by another hyperparameter γ. In the modified version, the over-focus on PEHE of the unconditional balance is seen clearly–with different γ, it significantly affects PEHE, but barely affects ATE error. In fact, the unconditional balance, with larger γ, only worsens the performance. See also Figure 4.9 where CFR gives larger ATE errors with less overlap.

TABLE 4.1: Errors on IHDP over 1000 random DGPs. "Mod. *" indicates the modified version with unconditional balance of strength $\gamma = *$. *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, Johansson, and Sontag, 2017, except for GANITE and CEVAE, the results of which are taken from original works.

Method	TMLE	BNN	CFR	CF	CEVAE	GANITE	Ours	Mod. 1	Mod. 0.2	Mod. 0.1	Mod. 0.05	Mod. 0.01
ϵ_{ate}	$.30_{\pm.01}$	$.37_{\pm.03}$	$.25_{\pm.01}$	$\mathbf{.18}_{\pm.01}$	$.34_{\pm.01}$	$.43_{\pm.05}$	$\mathbf{.180}_{\pm.007}$	$.185_{\pm.008}$	$.185_{\pm.008}$	$.186_{\pm.009}$	$.183_{\pm.008}$	$.181_{\pm.008}$
$\sqrt{\epsilon_{pehe}}$	$5.0_{\pm.2}$	$2.2_{\pm.1}$	$\mathbf{.71}_{\pm.02}$	$3.8_{\pm.2}$	$2.7_{\pm.1}$	$1.9_{\pm.4}$	$\mathbf{.709}_{\pm.024}$	$1.175_{\pm.046}$	$.797_{\pm.030}$	$.748_{\pm.028}$	$.732_{\pm.028}$	$.719_{\pm.027}$

4.5 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. 4.3.1 the pre/post-treatment distinction.

4.5.1 Synthetic Data

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

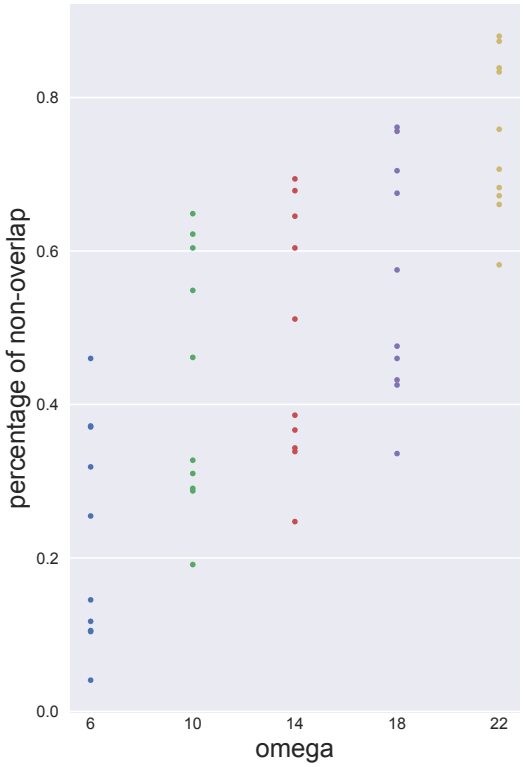

FIGURE 4.6: Degree of limited overlap w.r.t ω.

We use the original implementation of CFR[3]. 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 4.6, non (all) of the 10 DGPs are very limited-overlapping with $\omega = 6$ ($\omega = 22$).

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

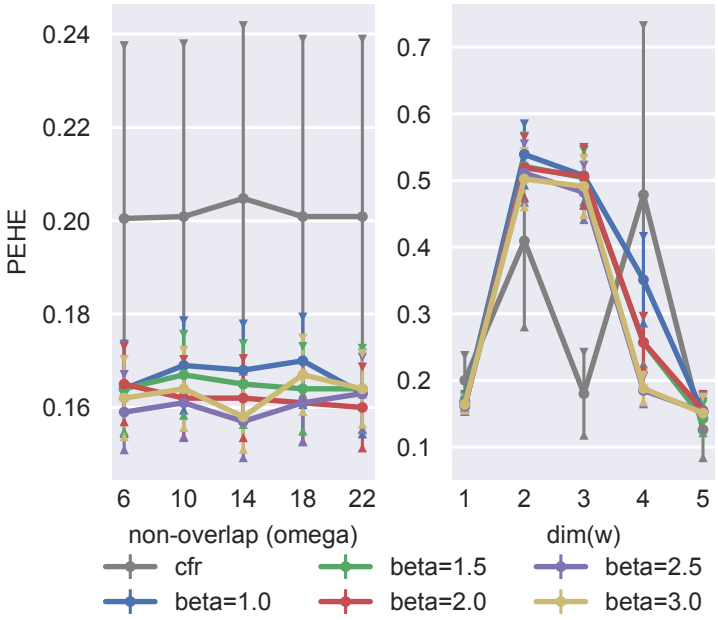

FIGURE 4.7: $\sqrt{\epsilon_{pehe}}$ on synthetic dataset, with $g_t(W) = 1$ in DGPs, and $\dim(Z) = 200$ in our model. Error bar on 10 random DGPs.

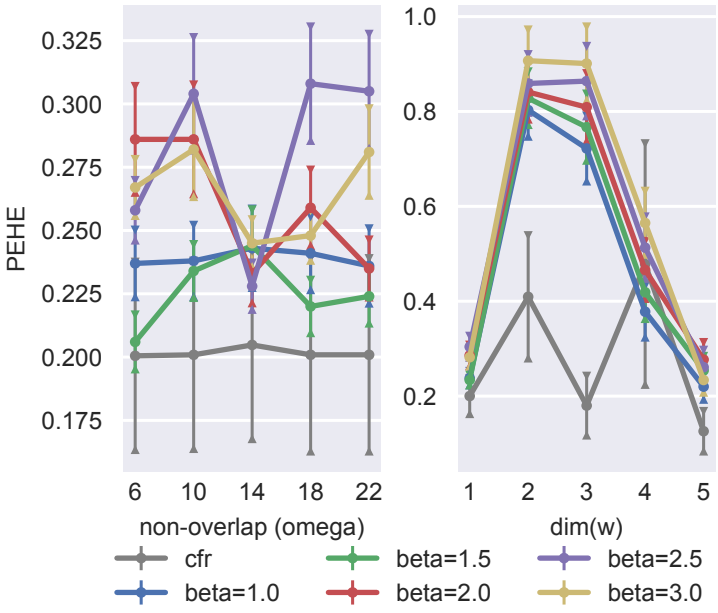

FIGURE 4.8: $\sqrt{\epsilon_{pehe}}$ on synthetic dataset, with $g_t(W) = 1$ in DGPs. Error bar on 10 random DGPs.

For diversity of the datasets, we set $g_t(W) = 1$ in DGPs in Figure 4.7. It 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 β seems to give better results here, this is understandable because β controls the trade-off between fitting and balancing, and the fitting capacity of our decoder is much increased with $\dim(Z) = 200$. Note that the above observations on $\dim(Z)$ are not caused by fixing $g_t(W) = 1$ (compare Figure 4.7 with Figure 5.3 below).

Figure 5.3 shows the importance of noise modeling. Compared to Figure 4.2 in the main text, where $g_t(W)$ in DGPs is not fixed, our method works worse here, particularly for large β, because now noise modeling ($\boldsymbol{g}, \boldsymbol{k}$ in the ELBO) only adds unnecessary complexity. The changes of performance w.r.t different ω should be unrelated to overlap levels, but to the complexity of random DGPs; compare to Figure 4.7, 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 (4.12). In Sec. 2.1.1, we saw that the 2-dimsensional balanced prognostic score $\mathfrak{p} := (\mu_0(X), \mu_1(X))$ always exists under additive noise models. Thus, when $\dim(W) > 2$, our method tries to recover that $\mathfrak{p}$, and generally performs not worse than under $\dim(W) = 2$, but still not better than under $\dim(W) = 1$.

Figure 4.9 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 β ($\beta = 3$). This might be another evidence that CFR and its unconditional balancing overfit to PEHE (see Sec. 4.4.2). 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 β gives better ATE estimation under the existence of balanced prognostic score, but leave this for future investigation.

Figure 4.10 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 balanced prognostic score W with $\dim(W) = 1$, there is no need to learn prognostic score. In the right panel, we also do not see significant drop of performance compared to post-treatment. This might be due to the hardness

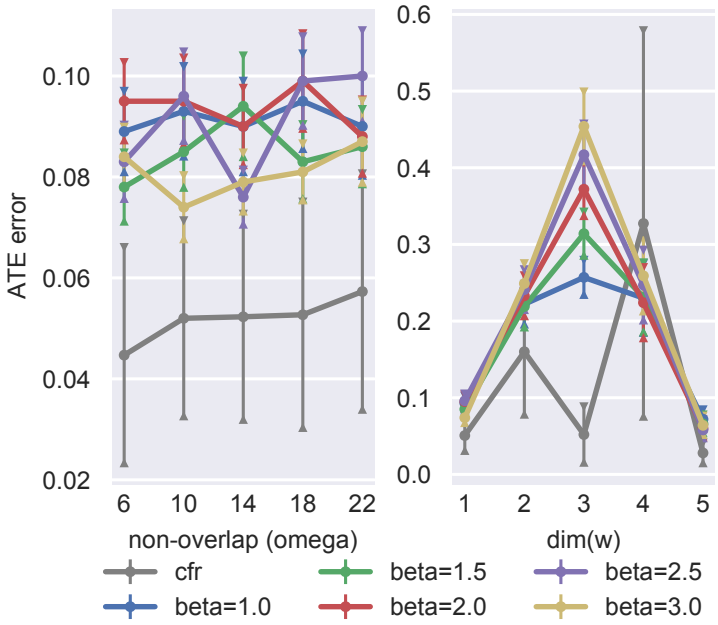

FIGURE 4.9: ϵ_{ate} on synthetic dataset, with $g_t(W) = 1$ in DGPs. Error bar on 10 random DGPs.

of learning approximate balanced prognostic score in this dataset, and posterior estimation does not give much improvements.

You can find more plots for latent recovery in Appendix A.1.

4.5.2 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).

$$Y(0) \sim \mathcal{N}(e^{\boldsymbol{a}^T(X+\boldsymbol{b})}, 1), \quad Y(1) \sim \mathcal{N}(\boldsymbol{a}^T X - c, 1), \tag{4.13}$$

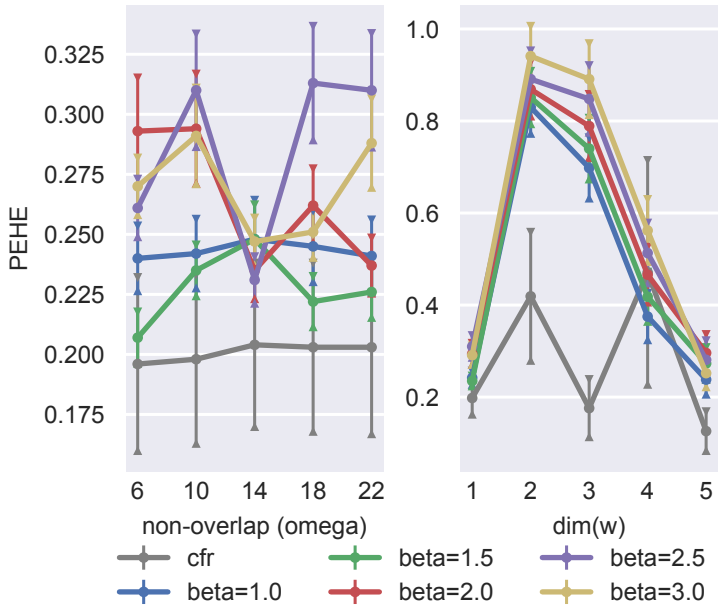

FIGURE 4.10: *Pre-treatment* $\sqrt{\epsilon_{pehe}}$ on synthetic dataset. Error bar on 10 random DGPs.

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 balanced prognostic score. As mentioned in the main text, the balanced prognostic score might be discrete. Thus, this experiment also shows the importance of VAE, even if an apparent balanced prognostic score exists. Under *discrete*prognostic scores, training an regression based on Proposition 5 is hard, but our VAE works well.

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 (4.2), using two separate NNs. Second, we add to our ELBO (4.3) a regularization term, which is the Wasserstein distance (Cuturi, 2013) between $\mathbb{E}_{\mathcal{D}\sim p(X|T=t)} p_\Lambda(Z|X), t \in \{0, 1\}$. As shown in Table 4.2, 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 4.3 shows pre-treatment results, All methods gives reasonable results.

TABLE 4.2: Performance of modified version with different unconditional balancing parameter, the values of which are shown after "Mod.".

Method	Ours	Mod. 1	Mod. 0.2	Mod. 0.1	Mod. 0.05	Mod. 0.01	CFR
ϵ_{ate}	$.177_{\pm.007}$	$.196_{\pm.008}$	$.177_{\pm.007}$	$.167_{\pm.005}$	$.177_{\pm.006}$	$.179_{\pm.006}$	$.25_{\pm.01}$
$\sqrt{\epsilon_{pehe}}$	$.843_{\pm.030}$	$1.979_{\pm.082}$	$1.116_{\pm.046}$	$.777_{\pm.026}$	$.894_{\pm.039}$	$.841_{\pm.029}$	$.71_{\pm.02}$

TABLE 4.3: *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, Johansson, and Sontag, 2017, except GANITE (Yoon, Jordon, and Schaar, 2018) and CEVAE (Louizos et al., 2017).

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

4.5.3 Empirical Validation of the Bounds in Sec. 4.3.2

Here we focus on the $\boldsymbol{D}(X)$ term in Theorem 2 because it is directly related to conditional balance.

In Figure A.9, the rows correspond to 3 overlap levels from strong to weak ($\omega = 6, 14, 22$ respectively). The first column shows the histograms of correlation coefficients between $\boldsymbol{D}(X)$ and $\epsilon_f(X)$ on 100 random DGPs. The vertical bars in the histograms are 5, 25, 50, 75, 95 percentiles (the values are shown in the table below). The other 10 columns show the plots of distributions of $(\boldsymbol{D}(X), \epsilon_f(X))$ for the first 10 DGPs. The correlation coefficient for each DGP is shown as `corrcoef=*` above each histogram. The plots are in log-log scale, because both $\boldsymbol{D}$ and ϵ_f are single-sided, and most data points concentrate near $(0, 0)$, making the plots bad-looking.

We have two important observations from the histograms: 1) on the majority of DGPs, there are positive correlations between $\boldsymbol{D}$ and ϵ_f; 2) the positive correlation is stronger with weaker overlap (the portion of large correlation increases, and the mean `corrcoef` are 0.100, 0.110, and 0.121, respectively).

Thus, our bounds and conditional balance have significance. Not all DGPs have positive correlations, and this is reasonable because our bound (4.11) has three other terms which can obscure the relation between $\boldsymbol{D}$ and ϵ_f. The DGPs 1, 3, 6, 8, 10 show typical situations when there are positive correlations.

TABLE 4.4: Percentiles of correlation coefficients between $\boldsymbol{D}(X)$ and $\epsilon_f(X)$ on 100 random DGPs.

Percentile	5	25	50	75	95
$\omega = 6$	-0.289	-0.086	0.069	0.299	0.609
$\omega = 14$	-0.328	-0.124	0.055	0.337	0.636
$\omega = 22$	-0.274	-0.128	0.067	0.341	0.634

4.6 Proofs

We restate our model identifiability formally.

Lemma 1 (Model identifiability). *Given model* (4.2) *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* $\boldsymbol{f}_t$*,*

$$\boldsymbol{f}_t^{-1}(\boldsymbol{y}_t) = \text{diag}(\boldsymbol{a})\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t) + \boldsymbol{b} =: \mathcal{A}_t(\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t)) \tag{4.14}$$

where $\text{diag}(\boldsymbol{a})$ *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., 2020b. The main part of our model identifiability is essentially the same as that of Theorem 1 in Khemakhem et al., 2020b, 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., 2020b. 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,

$$q_{\boldsymbol{f}',\boldsymbol{\lambda}'}(\boldsymbol{y}) = q_{\boldsymbol{f},\boldsymbol{\lambda}}(\boldsymbol{y}) := p_{\boldsymbol{\lambda}}(\boldsymbol{f}^{-1}(\boldsymbol{y})|\boldsymbol{x}, t)vol(\boldsymbol{J}_{\boldsymbol{f}^{-1}}(\boldsymbol{y}))\mathbb{I}_{\mathcal{Y}}(\boldsymbol{y}) \tag{4.15}$$

where p_λ is the Gaussian density function of the conditional prior defined in (4.2) and $vol(A) := \sqrt{\det AA^T}$. $q_{f',\lambda'}$ is defined similarly to $q_{f,\lambda}$.

Then, apply model (4.2) to (4.15), plug the $2n+1$ points from **(D1')** into it, and re-arrange the resulting $2n+1$ equations in matrix form, we have

$$\mathcal{F}'(Y) = \mathcal{F}(Y) := \boldsymbol{L}^T \boldsymbol{t}(\boldsymbol{f}^{-1}(Y)) - \boldsymbol{\beta} \tag{4.16}$$

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 (4.2). $\mathcal{F}'$ is defined similarly to $\mathcal{F}$, but with $\boldsymbol{f}', \boldsymbol{\lambda}', \alpha'$

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

$$\boldsymbol{t}(\boldsymbol{f}^{-1}(Y)) = \boldsymbol{A}\boldsymbol{t}(\boldsymbol{f}'^{-1}(Y)) + \boldsymbol{c} \tag{4.17}$$

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

The final part of the proof is to show, by following the same reasoning as in Appendix B of Sorrenson, Rother, and Köthe, 2019, that $\boldsymbol{A}$ is a sparse matrix such that

$$\boldsymbol{A} = \begin{pmatrix} \operatorname{diag}(\boldsymbol{a}) & \boldsymbol{O} \\ \operatorname{diag}(\boldsymbol{u}) & \operatorname{diag}(\boldsymbol{a}^2) \end{pmatrix} \tag{4.18}$$

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

$$\boldsymbol{f}^{-1}(Y) = \operatorname{diag}(\boldsymbol{a})\boldsymbol{f}'^{-1}(Y) + \boldsymbol{b} \tag{4.19}$$

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

Proof of Proposition 5. Under **(G2)**, and **(M3)**, we have

$$\mathbb{E}_{p_\theta}(Y|X,T) = \mathbb{E}(Y|X,T) \implies \boldsymbol{f}_t \circ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{j}_t \circ \mathfrak{p}(\boldsymbol{x}) \text{ on } (\boldsymbol{x}, t) \text{ such that } p(t, \boldsymbol{x}) > 0. \tag{4.20}$$

We show the solution set of (4.20) on *overlapping* $\boldsymbol{x}$ is

$$\{(\boldsymbol{f}, \boldsymbol{h}) | \boldsymbol{f}_t = \boldsymbol{j}_t \circ \Delta^{-1}, \boldsymbol{h} = \Delta \circ \mathfrak{p}, \Delta : \mathcal{P} \to \mathbb{R}^n \text{ is injective}\}. \tag{4.21}$$

By **(G2)(M1)**, and with injective f_t, j_t and $\dim(Z) = \dim(Y) \geq \dim(\mathfrak{p})$, for any Δ above, there exists a functional parameter f_t such that $j_t = f_t \circ \Delta$. Thus, set (4.21) is non-empty, and any element is indeed a solution because $f_t \circ \boldsymbol{h} = j_t \circ \Delta^{-1} \circ \Delta \circ \mathfrak{p} = j_t \circ \mathfrak{p}$.

Any solution of (4.20) should be in (4.21). A solution should satisfy $\boldsymbol{h}(\boldsymbol{x}) = f_t^{-1} \circ j_t \circ \mathfrak{p}(\boldsymbol{x})$ for both t since $\boldsymbol{x}$ is overlapping. This means the *injective* function $f_t^{-1} \circ j_t$ should *not* depend on t, thus it is one of the Δ in (4.21).

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

$$\hat{\mu}_t(\boldsymbol{x}) = f_t(\boldsymbol{h}(\boldsymbol{x})) = j_t \circ v^{-1}(v \circ \mathfrak{p}(\boldsymbol{x})) = j_t(\mathfrak{p}(\boldsymbol{x})) = \mu_t(\boldsymbol{x}). \tag{4.22}$$

We rely on overlapping $\mathfrak{p}$ to work for non-overlapping $\boldsymbol{x}$. For any $\boldsymbol{x}_t$ with $p(1-t|\boldsymbol{x}_t) = 0$, to ensure $p(1-t|\mathfrak{p}(\boldsymbol{x}_t)) > 0$, there should exist $\boldsymbol{x}_{1-t}$ such that $\mathfrak{p}(\boldsymbol{x}_{1-t}) = \mathfrak{p}(\boldsymbol{x}_t)$ 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

$$\hat{\mu}_{1-t}(\boldsymbol{x}_t) = f_{1-t}(\boldsymbol{h}(\boldsymbol{x}_t)) = f_{1-t}(\boldsymbol{h}(\boldsymbol{x}_{1-t})) = j_{1-t}(\mathfrak{p}(\boldsymbol{x}_{1-t})) = j_{1-t}(\mathfrak{p}(\boldsymbol{x}_t)) = \mu_{1-t}(\boldsymbol{x}_t). \tag{4.23}$$

The third equality uses (4.20) 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')**, $\boldsymbol{\beta}_t := (\alpha_t(\boldsymbol{x}_1) - \alpha_t(\boldsymbol{x}_0), ..., \alpha_t(\boldsymbol{x}_{2n}) - \alpha_t(\boldsymbol{x}_0))^T$, *and* $\alpha_t(X; \boldsymbol{\lambda}_t)$ *is the log-partition function of the prior in* (4.2).

(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. 4.3.1, which corresponds to $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{d} = \boldsymbol{0}$.

Proof of Theorem 1. By **(M1)** and **(G1')**, for any injective function $\Delta : \mathcal{P} \to \mathbb{R}^n$, there exists a functional parameter $\boldsymbol{f}_t^*$ such that $\boldsymbol{j}_t = \boldsymbol{f}_t^* \circ \Delta$. Let $\boldsymbol{h}_t^* = \Delta \circ \mathfrak{p}_t$, 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

$$\Delta \circ \boldsymbol{j}^{-1}(\boldsymbol{y}) = \boldsymbol{f}^{*-1}(\boldsymbol{y}) = \mathcal{A} \circ \boldsymbol{f}^{-1}(\boldsymbol{y})|_t, \text{ on } (\boldsymbol{y}, t) \in \{(\boldsymbol{j}_t \circ \mathfrak{p}_t(\boldsymbol{x}), t) | p(t, \boldsymbol{x}) > 0\}, \quad (4.24)$$

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

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

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

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

$$\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{L}_0'^{-1}\boldsymbol{L}_1' = \boldsymbol{C} \iff \boldsymbol{A}_0 = \boldsymbol{A}_1 \quad (4.25)$$

And also,

$$\begin{aligned} \boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{C} &\iff \boldsymbol{L}_0^{-T}\boldsymbol{C}^{-T} = \boldsymbol{L}_1^{-T} \\ \boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0' - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1' = \boldsymbol{d}/k &\iff \boldsymbol{C}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0') = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1' \end{aligned} \quad (4.26)$$

Multiply right hand sides of the two lines, we have $\boldsymbol{c}_0 = \boldsymbol{c}_1$. Now we have $\mathcal{A}_0 = \mathcal{A}_1 := \mathcal{A}$. Apply this to (4.24), we have

$$\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}, \quad \boldsymbol{v} := \mathcal{A}^{-1} \circ \Delta \quad (4.27)$$

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

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t) \implies p_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))) = p_{\mathbf{e}}(\boldsymbol{y} - \boldsymbol{j}_t(\mathfrak{p}_t(\boldsymbol{x}))) \tag{4.28}$$

where $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$. And the above is only possible when $\boldsymbol{f}_t \circ \boldsymbol{h}_t = \boldsymbol{j}_t \circ \mathfrak{p}_t$. 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 5, applied to both $\mathfrak{p}_0$ and $\mathfrak{p}_1$. □

Note, when multiplying the two lines of (4.26), the effects of $k \to 0$ cancel out, and $\boldsymbol{c}_t$ is finite and well-defined. Also, it is apparent from above proof that **(D2')** is a necessary and sufficient condition for $\mathcal{A}_0 = \mathcal{A}_1$, if other conditions of Theorem 1 are given.

Below, we prove the results in Sec. 4.3.2. The definitions and results work for the prior; simply *replace* $q_t(\boldsymbol{x}|\boldsymbol{x})$ *with* $p_t(\boldsymbol{z}|\boldsymbol{x}) \coloneqq p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t)$ *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* $|\mathcal{L}_f(\boldsymbol{z}, t)| \leq M$*, we have*

$$\epsilon_{CF}(\boldsymbol{x}) \leq \textstyle\sum_t q(1-t|\boldsymbol{x})\epsilon_{F,t}(\boldsymbol{x}) + M\boldsymbol{D}(\boldsymbol{x}) \tag{4.29}$$

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

Proof of Lemma 2.

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

□

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

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*

$$\epsilon_f \leq 2(G(\epsilon_F + \epsilon_{CF}) - \mathbf{V}_Y). \tag{4.30}$$

Simply bound ϵ_{CF} in (4.30) by Lemma 2, we have Theorem 2. To prove Lemma 3, we first examine a bias-variance decomposition of ϵ_F and ϵ_{CF}.

$$\begin{aligned}
\epsilon_{CF,t} &= \mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\boldsymbol{g}_t(\boldsymbol{z})\mathbb{E}_{p_{Y(t)|\mathbb{p}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2 \\
&\geq G\mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{p}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2 \\
&= G\mathbb{E}_{q_{1-t}(\boldsymbol{z}|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{p}_t}(\boldsymbol{y}|\boldsymbol{z})}((\boldsymbol{y} - \boldsymbol{j}_t(\boldsymbol{z}))^2 + (\boldsymbol{j}_t(\boldsymbol{z}) - \boldsymbol{f}_t(\boldsymbol{z}))^2)
\end{aligned} \tag{4.31}$$

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

$$\epsilon_{CF,t} \geq G(\mathbf{V}_{CF,t}(\boldsymbol{x}) + \mathbb{B}_{CF,t}(\boldsymbol{x})) \implies \epsilon_{CF} \geq G(\mathbf{V}_{CF}(\boldsymbol{x}) + \mathbb{B}_{CF}(\boldsymbol{x})) \tag{4.32}$$

where $\mathbf{V}_{CF} := \sum_t p(1-t|\boldsymbol{x})\mathbf{V}_{CF,t} = \sum_t \mathbb{E}_{q(\boldsymbol{z},1-t|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{p}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{j}_t(\boldsymbol{z}))^2$ 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 ϵ_F, we have

$$\epsilon_F \geq G(\mathbf{V}_F(\boldsymbol{x}) + \mathbb{B}_F(\boldsymbol{x})) \tag{4.33}$$

where $\mathbf{V}_F = \sum_t \mathbb{E}_{q(\boldsymbol{z},t|\boldsymbol{x})}\mathbb{E}_{p_{Y(t)|\mathbb{p}_t}(\boldsymbol{y}|\boldsymbol{z})}(\boldsymbol{y} - \boldsymbol{j}_t(\boldsymbol{z}))^2$ 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.

$$\begin{aligned}
\epsilon_f &= \mathbb{E}_{q(z|x)}((f_1 - f_0) - (j_1 - j_0))^2 \\
&= \mathbb{E}_q((f_1 - j_1) + (j_0 - f_0))^2 \\
&\leq 2\mathbb{E}_q((f_1 - j_1)^2 + (j_0 - f_0)^2) \\
&= 2\int [(f_1 - j_1)^2 q(z,1|x) + (j_0 - f_0)^2 q(z,0|x) + \\
&\qquad (f_1 - j_1)^2 q(z,0|x) + (j_0 - f_0)^2 q(z,1|x)] dz \\
&= 2(\mathbb{B}_F + \mathbb{B}_{CF}) \leq 2(G(\epsilon_F + \epsilon_{CF}) - \mathbf{V}_Y)
\end{aligned}$$

□

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

4.7 Detailed Explanations and Discussions

The order of subsections below follows that they are referred in the previous chapters.

4.7.1 List of Assumptions

The following is a list of assumptions required by our identification theory, with comments on their roles and subtleties.

(G1) additive noise model is needed to ensure the existence of PtSs. **(G1')** is equivalent to **(G1)**, and is introduced for better presentation, e.g., it connects to **(G2)** and **(M1)** through injectivity.

(M1) and **(D1)** are inherited from iVAE and are required for model (parameter) identifiability (identifying f_t up to affine mapping), which does not imply CATE identification in general. Arguably here the most important is that the mapping f_t from latent Z to outcome Y is injective, or else some information of Z is in principle unrecoverable. These two conditions are not required by Proposition 5 which does not need model identifiability.

(M2), together with overlapping PtSs, is important to address limited overlap of X and can be seen as a weak form of OOD generalization.

(M3') means 1) we need to know or learn the distribution of hidden noise e and 2) noiseless prior. This simplifies the proof of identification, but when implementing the VAE as an estimation method, both noises are learned.

(D2), or in fact **(D2')**, strengthens the model identifiability to determine both f_0 and f_1 up to the *same* affine mapping, which replaces the balance ofprognostic score.

(G2) is required by Proposition 5 but not Theorem 1. It is no less important than **(G1')**, because the core intuition of our method is that **(G2)** should hold approximately. Sec. 4.7.2 contains several detailed real-world examples on **(G2)**.

4.7.2 Discussions and Examples of (G2)

We focus on univariate outcome on $\mathbb{R}$ which is the most practical case and the intuitions apply to more general types of outcomes. Then, $\boldsymbol{i}$, the mapping between μ_0 and μ_1, is monotone, i.e, either increasing or decreasing. The increasing $\boldsymbol{i}$ means, if a change of the value of 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 and 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 μ_0 and μ_1 are both monotone increasing. This, and also that both μ_t are monotone decreasing, are natural and sufficient conditions for increasing $\boldsymbol{i}$, though not necessary. The first example is form Health. (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 μ_t should be monotone increasing in gestational age. The next example is from economics. (Gan and 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, Hsu, and Lieli, 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.

4.7.3 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)** ($\mathfrak{p}_0 = \mathfrak{p}_1$) is replaced by **(D2')** which instead makes $\mathcal{A}_0 = \mathcal{A}_1 := \mathcal{A}$ in (4.4). 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, $p_\epsilon = p_{\mathbf{e}}$ in **(M3')** is impractical, but it ensures that $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}, t) = p(\boldsymbol{y}|\boldsymbol{x}, t)$ so that (4.4) can be used. In Sec. 4.3.1, we consider practical estimation method and introduce the *regularization* that encourages learning a prognostic score similar to balanced prognostic score so that $p_\epsilon = p_{\mathbf{e}}$ can be relaxed.

4.7.4 Ideas and Connections behind the ELBO (4.7)

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

In fact, β connects our VAE to β-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 $\mathfrak{p}_T$ 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 and 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 β implicitly encourages $\boldsymbol{g}$ *smaller* than the scale of $\mathbf{e}$, through stressing the third term in ELBO (4.7). 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 β. 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 5.

Now, it is clear that β naturally controls at the same time noise scale and balancing. And the regularization can also be understood as an interpolation between Proposition 5 and Theorem 1: relying on balanced prognostic score, 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 j. Sec. 4.3.2 shows that this error and balancing form a trade-off, which is adjusted by β.

Importance of balancing from misspecification view. If we must learn an unapproximate balanced prognostic score, 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 and 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 balanced prognostic score, so that we can explore the *middle ground*: relatively low-dimensional $\mathfrak{p}$, or relatively small $\mathbf{e}$.

Example. Assume the true outcome noise is (near) zero. By setting $\epsilon \to \mathbf{0}$ in our model, the posterior $p_\theta(z|x,y,t) = p_\theta(y,z|x,t)/p_\theta(y|x,t)$ degenerates to $f_T^{-1}(Y) = f_T^{-1}(j_T(\mathfrak{p}_T)) = v^{-1}(\mathfrak{p}_T)$, a *factual* prognostic score. However, $f_{1-T}^{-1}(Y) = f_{1-T}^{-1}(j_T(\mathfrak{p}_T)) = v^{-1}(j_{1-T}^{-1} \circ j_T(\mathfrak{p}_T)) \neq v^{-1}(\mathfrak{p}_{1-T})$, *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).

4.7.5 Additional Notes on Novelties of the Bounds in Sec. 4.3.2

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 (4.7), without ad hoc regularization; moreover, we consider outcome noise modeling which is a strength of VAE and relate it to hyperparameter β. Particularly, in (Lu et al., 2020), latent variable Z is confused with the true representation ($\mathfrak{p}_t$ up to invertible mapping in our case). *Without* identification, the method in fact has unbounded error. Note that Shalit, Johansson, and Sontag, 2017 do not consider connection to identification and noise modeling as well. The error between $\hat{\tau}_f$ and τ_j, which we

bound, is due to the unknown outcome noise that is not accounted by our Theorem 1; thus, the theory in Sec. 4.3.2 is complementary to that in Sec. 4.2.3. Finally, β is a trade-off between the conditional balance of learned prognostic score (affected by f_t), and precision/effective sample size of outcome regression—and can be seen as the probabilistic counterpart of Tarr and Imai (2021) and Kallus, Pennicooke, and Santacatterina (2018).

Chapter 5

Intact-VAE: Theoretical Ideas and Experiments under Unobserved Confounding

5.1 Unobserved Confounding

In this section, we extend the framework in Section 2.1.1 to include unobserved confounding. Although there are mostly no original results in this section, we put it here because the formulation is not standard.

5.1.1 Identification

Adapting standard identification results (Rubin, 2005)(Hernan and Robins, 2020, Ch. 3), we start with the following conditions, denoted by **(A)**: there exists a (possibly unobserved) variable $U \in \mathbb{R}^n$ such that together with X, it gives (i) (Exchangeability) $Y(t) \perp\!\!\!\perp T|U, X$ and (ii) (Overlap, or Positivity) $p(T|U, X) > 0$; and (iii) (Consistency of counterfactuals) $Y = Y(t)$ if $T = t$. All of them are satisfied for *both* t, which is our convention when t appears in a statement without quantification. Intuitively, exchangeability means all confounders are in essence contained in (U, X), and overlap means each possible value of (U, X) is observed for both treatment groups. Note that, joint exchangeability $Y(0), Y(1) \perp\!\!\!\perp T|X, U$ is stronger than exchangeability and is not necessary for identification (Hernan and Robins, 2020, pp. 15).

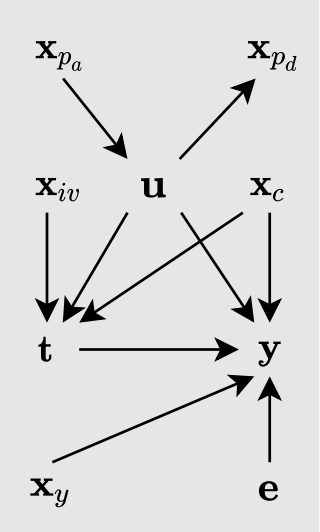

FIGURE 5.1: A possible causal graph of unobserved confounding.

A general example of causal structure that satisfies the three conditions is shown in Figure 5.1, although further structural constraints might be necessary for theoretical developments (see Sec. 5.4). Here, X_c,X_{iv},X_{p_a},X_{p_d},X_y are covariates that are: (observed) confounder, IV, antecedent proxy (that is antecedent of Z), descendant proxy, and antecedent of Y, respectively. The covariate(s) X may *not* have subsets in any categories in the graph. **e** is unobserved exogenous noise on Y. Assumption **(A)** may hold otherwise, e.g., X is a child of T.

CATE can be given by (5.1), using assumption **(A)** in the second equality.

$$\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathbb{E}(Y(t)|U,\boldsymbol{x})) = \mathbb{E}(\mathbb{E}(Y|U,\boldsymbol{x},T=t)) = \int(\int p(y|\boldsymbol{u},\boldsymbol{x},t)ydy)p(\boldsymbol{u}|\boldsymbol{x})d\boldsymbol{u}. \tag{5.1}$$

If the variable U is observed, then (5.1) identifies CATE. However, if U is an *unobserved confounder*, the naive regression $\mathbb{E}[Y|X=\boldsymbol{x},T=t]$ based on observable variables is not equal to $\mu_t(\boldsymbol{x})$. In fact, if an unknown factor correlates with T positively and tends to give higher value for Y, the naive regression $\mathbb{E}[Y|X=\boldsymbol{x},T=1]$ should be higher than $\mathbb{E}[Y(1)|X=\boldsymbol{x}]$.

5.1.2 Prognostic Score with U

Our method models prognostic scores (Hansen, 2008), adapted as *Pt-scores* in this chapter, closely related to the important concept of balancing score (Rosenbaum and

Rubin, 1983). Both are sufficient scores for identification; prognostic scores are sufficient statistics of outcome predictors and balancing score is for the treatment.

Definition 6. A *Pt-score* (PtS) is two functions $\mathfrak{p}_t(U, X)$ $(t = 0, 1)$ such that $Y(t) \perp\!\!\!\perp U, X | \mathfrak{p}_t(U, X)$. A PtS is called a *P-score* (PS) if $\mathfrak{p}_0 = \mathfrak{p}_1$.

The identity function is a trivial PS. If the true data generating process (DGP) satisfies additive noise model, i.e., $Y = \boldsymbol{f}^*(U, X, T) + \mathbf{e}$, then $\boldsymbol{f}_t^*$ is a PtS (Hansen, 2008); and it is a causal representation (Schölkopf et al., 2021) of the direct cause on Y, summarizing the effects of (U, X). The independence property of PtS (Proposition 2 in Sec. 2.1.1),

$$Y(t) \perp\!\!\!\perp T, U, X | \mathfrak{p}_t(U, X), \tag{5.2}$$

is used in second equality of (5.3) in Theorem 3 which extends Proposition 5 in Hansen, 2008.

Theorem 3 (CATE by PtS)**.** *If $\mathfrak{p}_t$ is a PtS, then CATE can be given by*

$$\begin{aligned} \mu_t(\boldsymbol{x}) &= \mathbb{E}(\mathbb{E}(Y(t)|\mathfrak{p}_t(U, \boldsymbol{x}), \boldsymbol{x})) = \mathbb{E}(\mathbb{E}(Y|\mathfrak{p}_t(U, \boldsymbol{x}), t)) \\ &= \int (\int p_{Y|\mathfrak{p}_t,T}(y|P, t) y dy) p_{\mathfrak{p}_t|X}(P|\boldsymbol{x}) dP, \end{aligned} \tag{5.3}$$

where $Y|\mathfrak{p}_t(U, X), T \sim p_{Y|\mathfrak{p}_t,T}(\boldsymbol{y}|P, t)$ and $\mathfrak{p}_t(U, X)|X \sim p_{\mathfrak{p}_t|X}(P|\boldsymbol{x})$.

Compared to (5.1), $P = \mathfrak{p}_t(\boldsymbol{u}, \boldsymbol{x})$ plays the role of $\boldsymbol{u}$, and $p_{Y|\mathfrak{p}_t,T}$ conditions on P instead of $(\boldsymbol{u}, \boldsymbol{x})$. In general, information from U is needed to determine $p_{Y|\mathfrak{p}_t,T}$ and $p_{\mathfrak{p}_t|X}$. In Sec. 5.4, we discuss how our model is connected to and might learn relaxations of PtS when U is unobserved.

5.2 Experiments

We use the proposed Intact-VAE for three types of data, and compare it with existing methods.

Unless otherwise indicated, for each function $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{h}, \boldsymbol{k}, \boldsymbol{r}, \boldsymbol{s}$ in our VAE, we use a multilayer perceptron (MLP) that has 3*200 hidden units with ReLU activation, and $\boldsymbol{\lambda} = (\boldsymbol{h}, \boldsymbol{k})$ 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. We test post-treatment results on training and validation set jointly. The treatment and (factual) outcome should not be observed for pre-treatment predictions, so we report them on a testing set. More details on hyper-parameters and settings are given in each experiment.

As in previous works (Shalit, Johansson, and Sontag, 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 the square root of empirical PEHE (Hill, 2011) $\epsilon_{pehe} := \mathbb{E}_{\mathcal{D}}((y(1) - y(0)) - \hat{\tau}(\boldsymbol{x}))^2$ for individual-level treatment effects.

5.2.1 Synthetic Dataset

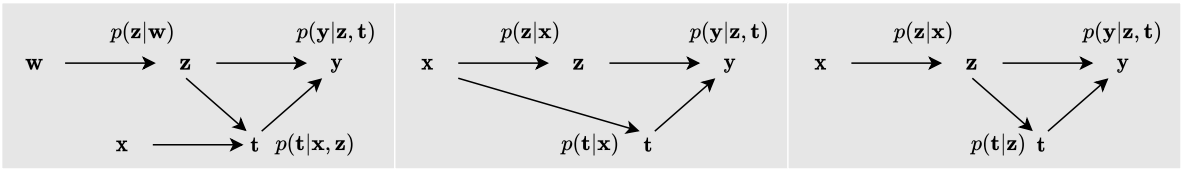

FIGURE 5.2: Graphical models for generating synthetic datasets. From left: IV X, ignorability given X, and proxy X. Note that in the latter two cases, reversing the arrow between X, Z does not change any independence relationships, and causal interpretations of the graphs remain the same.

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}); Z|X \sim \mathcal{N}(h(X), \beta k(X)); T|X, Z \sim \text{Bern}(\text{Logi}(l(X, Z))); Y|Z, T \sim \mathcal{N}(f(Z, T), \alpha). \tag{5.4}$$

We generate synthetic datasets by (5.4). The parameters are different between DGPs: μ_i and σ_i are randomly generated; the functions h, k, l are linear with random coefficients; and f_0, f_1 is built by separated randomly initialized (then fixed) NNs. We generate two kinds of outcome models, depending on the type of f: linear and nonlinear outcome models use random linear functions and NNs with invertible activations and random weights, respectively. We set the outcome and proxy noise level by α, β respectively.

We experiment on three different causal structures as shown in graphical models of Figure 5.2, by variation on (5.4). Instead of taking inputs X, Z in l, we consider two special cases: $l := l(X)$, then X fully adjusts for confounding, we are in fact

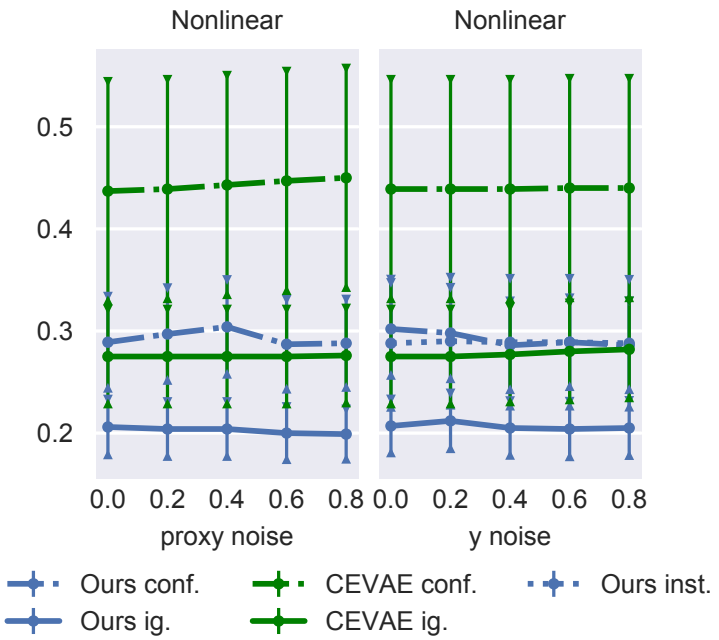

FIGURE 5.3: Pre-treatment $\sqrt{\epsilon_{pehe}}$ on nonlinear synthetic dataset. Error bar on 100 random DGPs. We adjust one of the noise levels α, β in each panel, with another fixed to 0.2.

unconfounded; and $l := l(Z)$, then we have unobserved confounder Z and *proxy* X of Z. To introduce X as *instrumental variable*, we generate another 1-dimensional random source W in the same way as X, and use W instead of X to generate $Z|W \sim \mathcal{N}(h(W), \beta k(W))$; except indicated above, other aspects of the models are specified by (5.4).

For each causal structure, and with the same kind of outcome models, and the same noise levels (α, β), we evaluate Intact-VAE and CEVAE on 100 random DGPs, with different sets of parameters in (5.4). For each DGP, we sample 1500 data points, and split them into 3 equal sets for training, validation, and testing. Both the methods use 1-dimensional latent variable in VAE. For fair comparison, all the hyperparameters, including type and size of NNs, learning rate, and batch size, are the same for both the methods.

Figure 5.3 shows our method significantly outperforms CEVAE on all cases Both methods work the best under unconfoundedness ("ig."), as expected. The performances of our method on IV ("inst.") and proxy ("conf.") settings match that of CEVAE under unconfoundedness, showing the effective deconfounding. Results for ATE and post-treatment are similar.

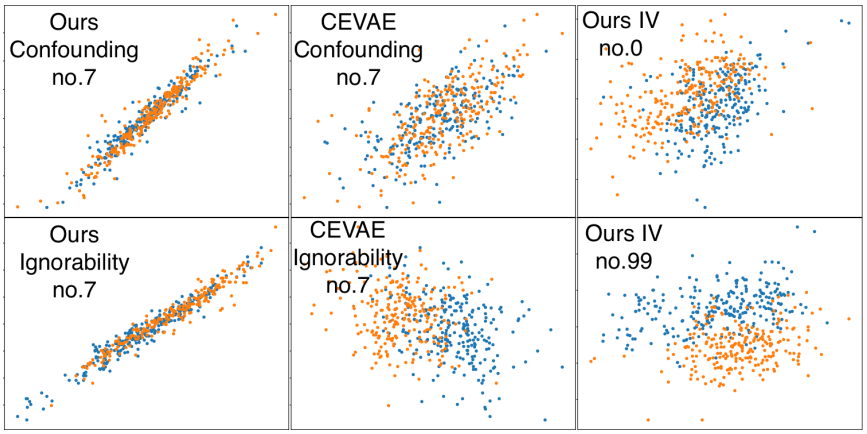

FIGURE 5.4: Plots of recovered - true latent on the nonlinear outcome. Blue: $t = 0$, Orange: $t = 1$. $\alpha, \beta = 0.4$. "no." indicates index among the 100 random models.

Here, the true latent Z is a PS, and there are no better candidate PSs than Z, because f_t is invertible and no information can be dropped from Z. Thus, as shown in Figure 5.4, our method learns representation as an approximate affine transformation of the true latent value, as a result of our model identifiability. More latent plots are in Appendix A.1. As expected, both recovery and estimation are better with unconditional prior $p(Z|X)$, and we see examples of bad recovery using conditional $p(Z|X,T)$ in Appendix Figure A.7. CEVAE shows much lower quality of recovery, particularly with large noises. Under the IV setting, while treatment effects are estimated as well as for the proxy setting, the relationship to the true latent is significantly obscured, because the true latent is correlated to IV X only given T, while we model it by $p(Z|X)$. This confirms that our method does not need to recover the true confounder distribution.

We see our method is robust to the unknown noise level. This indicates that noises are learned by our VAE. Appendix A.1 shows that the noise level affects how well we recover the latent variable.

5.2.2 Pokec Social Network Dataset

We show our method is the best compared with the methods specialized for networked deconfounding, a challenging problem in its own right. Pokec (Leskovec and Krevl, 2014) is a real world social network dataset. We experiment on a semi-synthetic dataset based on Pokec, introduced in Veitch, Wang, and Blei, 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 that are within the same region. The network structure is expressed by binary adjacency matrix $\boldsymbol{G}$.

Each user has 12 attributes, among which `district`, `age`, or `join date` is used as a confounder Z to build 3 different datasets, with remaining 11 attributes used as covariate X. Treatment T and outcome Y are synthesised as following:

$$T \sim \text{Bern}(g(Z)), \quad Y = T + 10(g(Z) - 0.5) + \epsilon, \text{ where } \epsilon \text{ is standard normal.} \quad (5.5)$$

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

As in Veitch, Wang, and Blei, 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. Table 5.1 shows the results. In addition, 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. Veitch, Wang, and Blei, 2019 do not give individual-level prediction.

TABLE 5.1: Pre-treatment ATE on Pokec. Ground truth ATE is 1, as we can see in (5.5). "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 and Blei, 2013). "Embed-" denotes the best alternatives given by (Veitch, Wang, and Blei, 2019). **Bold** indicates method(s) that are *significantly* better than all the others. 20-dimensional latent variable in Intact-VAE works better, and its result is reported. The results of the other methods are taken from (Veitch, Wang, and Blei, 2019).

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

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 Z can be roughly seen as a deterministic function of $\boldsymbol{G}, X$. This idea is comparable to Assumptions 2 and 4 in Veitch, Wang, and Blei, 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 and Welling, 2017) in the prior and encoder of Intact-VAE. Note that GCN cannot be trained by mini-batch, instead, we perform batch gradient decent using all data for each iteration, with initial learning rate 10^{-2}. We use dropout (Srivastava et al., 2014) with rate 0.1 to prevent overfitting.

GCN need to take as inputs the network matrix $\boldsymbol{G}$ and the covariates matrix $\boldsymbol{X} := (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_M^T)^T$ of *all* users, where M is user number, regardless of whether it is in training, validation, or testing phase; and it outputs a representation matrix $\boldsymbol{R}$, again for all users. To enable sample separation, we need to make sure 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. 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 covariate 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, except that they take $(\boldsymbol{R}_{m,:})^T$ where $\boldsymbol{x}_m$ was expected as input. And we have respective selection operations for validation and testing. We can still train Intact-VAE including GCN by Adam, simply setting the gradients of non-seleted rows of $\boldsymbol{R}$ to 0.

5.3 VAEs for Treatment Effect Estimation: a Critical Examination

Most VAE methods for treatment effects, e.g., Louizos et al., 2017; Zhang, Liu, and Li, 2020; Vowels, Camgoz, and Bowden, 2020; Lu et al., 2020, add ad hoc heuristics into the VAEs, and thus break down probabilistic modeling, not to mention model identifiability. Moreover, the methods learn representations from proxy variables, leading to either impractical assumptions or conceptual inconsistency, in treatment effect identification. As a case study, you can find detailed comparisons and criticisms of CEVAE in Section 3.1.1.

On identification. First, as to treatment effect identification, CEVAE assumes unobserved confounder can be recovered, which is rarely possible even under further structural assumptions (Tchetgen et al., 2020). Indeed, Rissanen and Marttinen, 2021 recently give evidence that the method often fails. Other methods (Zhang, Liu,

and Li, 2020; Vowels, Camgoz, and Bowden, 2020; Lu et al., 2020) assume unconfoundedness but still rely on proxy at least intuitively; for example, Lu et al., 2020 factorize the decoder as if 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 model identifiability, the empirical results of the methods lack solid ground; under settings not covered by their experiments, the methods would silently fail to learn proper representations, as we show in Sec. 4.4.1.

On ad hoc heuristics. Ad hoc heuristics break down probabilistic modeling and/or give ELBOs that do not optimally estimate the models. For example, in CEVAE, $q(T|X)$ and $q(Y|X,T)$ are added into the encoder to have pre-treatment estimation, and the ELBO has two additional likelihood terms respectively. The VAE in Zhang, Liu, and Li, 2020 is even more ad hoc; it splits the latent variable Z into three components, and applies the ad hoc tricks of CEVAE to each of the component. Particularly, when constructing the encoder, they implicitly assume the three components of Z are conditional independent give X, which violates the intended graphical model.

Compared to the above methods, our Intact-VAE is simpler and more principled, and often has better performance. It models a prognostic score as the latent variable and is based on the identification equation (5.3), while not compromised by ad hoc heuristics. Our ELBO is derived by standard variational lower bound (4.3). Moreover, our pre-treatment prediction is given naturally by the prior, thanks to the correspondence between our model and (5.3). We show in the following subsections how our model and its identifiability inspire theoretical developments in treatment effect identification.

5.4 Theoretical Ideas under Unobserved Confounding

The positive experimental results motivate us to consider the theory under unobserved confounding. Moreover, the prior in (4.2) is even more natural with U unobserved, since $p_{\mathfrak{p}_T|X}$ is not degenerated due to the uncertainty of U. Thus, we conjecture that, in our VAE framework, unobserved confounding is treated as a source of uncertainty of scores and is handled in a Bayesian way. We give more considerations for future theoretical work below.

Identification. Auxiliary structures (e.g., IVs) can give treatment effect identifications via *control functions* $\mathbb{C}(T, X)$, conditioning on which the treatment becomes exogenous, that is, $Y(t) \perp\!\!\!\perp T | \mathbb{C}(t, X)$ (Matzkin, 2007; Wooldridge, 2015). Control functions can be stochastic, as in Puli and Ranganath, 2020. Consistent treatment effect estimation can be given by a regression of outcome on the treatment and a control function. Our model (4.2) can be seen as a two-stage procedure: first, $p_\lambda(\boldsymbol{z}|\boldsymbol{x}, t)$ gives a stochastic control function; second, $p_f(\boldsymbol{y}|\boldsymbol{z}, t)$ regresses the outcome. We need to specify the control function learned by Intact-VAE and the required structural constraints. Control functions are recently found under the proxy setting (Nagasawa, 2021), or in the presence of both proxies and IVs (Tien, 2021).

Estimation. In causal inference, many models, including nonparametric IV regression (NPIV), are stated as *conditional moment restrictions* (CMRs) (Newey, 1993). Optimizing the ELBO of our VAE, given by (4.3), can be seen as finding functions f and $\mathbb{C}$, subject to the CMR $\mathbb{E}_{p_\theta}(Y|X, T) = \mathbb{E}(Y|X, T)$. We believe our Intact-VAE framework, possibly with modifications, can be shown to give optimal estimation under the CMR. There are formal connections between CMRs and *quasi-Bayesian* analysis using KL divergence (Zhang, 2006; Jiang and Tanner, 2008; Kim, 2002). For example, Kato, 2013 uses a quasi-likelihood from the CMR of NPIV to set the prior, and the Gibbs posterior (Zhang, 2006; Jiang and Tanner, 2008) is a minimizer of an information complexity which has a variational characterization similar to an ELBO. For general CMR models, Liao and Jiang, 2011 extend Kim, 2002 and give the best approximation to the true likelihood function under the CMR by minimizing a KL divergence. Very recently, Wang et al., 2021 employ quasi-Bayesian analysis to kernel-based IV methods, but only consider unconditional moments.

Chapter 6

Causal Mosaic: Bivariate Causal Discovery via Nonlinear ICA and Ensemble Method

6.1 Intuition of Shared Mechanisms

As mentioned in the Introduction, we encounter a large diversity of causal relationships in nature. And causality might only be studied and learned piecemeal. Our idea is to extract the common mechanism shared by a small number of causal systems. We should note that, systems that seem to have different mechanisms can actually share the same mechanism. When all we have at hand is observational data, the sample, it would be true that two systems sharing the same mechanism, but by looking at the samples, they seem very different, to the extent that we would be tempted to model them by different functional forms. As an example, we give some pairs we used in experiment.

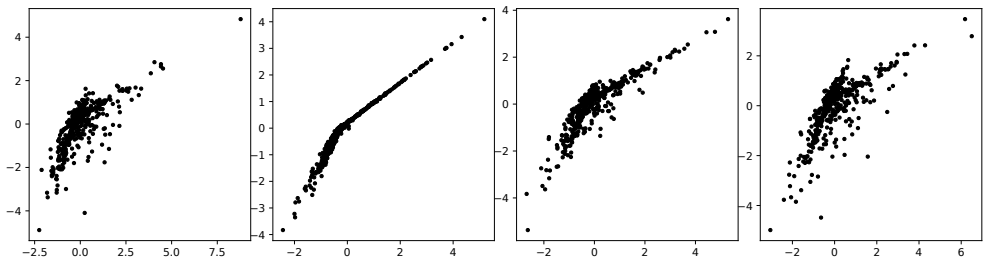

FIGURE 6.1: Artificial causal pairs sharing same mechanism. The pairs have significant diversity though still show some regularity.

6.2 Learning the Shared Mechanism by TCL

We review the nonlinear ICA method which we exploit to learn shared mechanism. This section is not placed in Preliminaries because we change the original use case and adapt it to our own setting.

Lately, Time-Contrastive Learning (TCL) (Hyvärinen and Morioka, 2016) provided the first general identifiability result for nonlinear ICA. The method depends on learning the different distributions of time series through time, and hence the name. After artificially dividing time series into segments, it trains a classification task to tell which segment each sample point belongs to. In fact, we can go a step further as indicated in Hyvärinen, Sasaki, and Turner, 2019: we only need the so-called *auxiliary variable*, conditioned on which the hidden components $\mathbf{Z}$ are jointly independent. The segment index is an example of the auxiliary variable because the two hidden variables are independent if there are no hidden confounders.

With the intuition that different causal pairs in real world could share similar mechanisms, we can develop a method to learn and invert the mechanisms by TCL. We feed the pairs into TCL as if they are the time segments and replace the segment index with the pair index as auxiliary variable. For now, let us assume the causal pairs share exactly the same mechanism and restate the theory of TCL under our own setting:

Theorem 4 (Hyvärinen and Morioka, 2016)**.** *Assume the following:*

A1. We observe causal pairs $\mathcal{X}(P) := \{\mathbf{X}_p\}_{p=1}^{P}$ *which satisfy the same analyzable SCM* $\mathbf{X}_p = \mathbf{f}(\mathbf{E}_p)$*, and the hidden variables* $E_{i,p}, i = 1, 2$ *are of exponential family distribution* $p_{E_{i,p}}(e) = \exp[T_i(e)\eta_i(p) - A(\eta_i(p))]$ *where* $T_i(e)$ *is the sufficient statistic.*

A2. The matrix $\mathbf{L}$*, with elements* $[\mathbf{L}]_{p,i} = \eta_i(p) - \eta_i(1), p = 1, ..., P, i = 1, 2$*, has full column rank 2.*

A3. We train a feature extractor $\mathbf{h} : \mathbf{R}^2 \to \mathbf{R}^2$ *with universal approximation capability, followed by a final softmax layer to classify all sample points of the pairs, with pair index used as class label.*

Then, in the limit of infinite data, for each p, $\mathbf{T}(\mathbf{E}_p) := (T_1(E_{1,p}), T_2(E_{2,p}))^T = \mathbf{A}\mathbf{h}(\mathbf{X}_p; \boldsymbol{\theta}) + \mathbf{b}$ *where* $\mathbf{A}, \mathbf{b}$ *are unknown constants, and* $\mathbf{A}$ *is invertible.*

Unlike the time contrast exploited in the original TCL, the contrast here is among the pairs. But, by convention, we will still use the word *"TCL"* when referring to the method trained on causal pairs. By a slight abuse of terminology, the produced $\mathbf{h}$ is also called TCL in this thesis.

In practice, a multilayer perceptron (MLP) is used as the feature extractor. The theorem implies that the identification (recovery) of $\mathbf{T}(\mathbf{E}_\mathrm{p})$ can be achieved by first performing TCL, and then linear ICA on $\mathbf{h}(\mathbf{X}_\mathrm{p})$. Denoting the composition of $\mathbf{h}$ and linear ICA as $\mathbf{hICA}$, we have $\mathbf{T}(\mathbf{E}_\mathrm{p}) = \mathbf{hICA}(\mathbf{X}_\mathrm{p})$. In this sense, we say that $\mathbf{h}$ is successfully learned and the nonlinear ICA of $\mathbf{X}_p$ is *realized* by $\mathbf{hICA}$. Here we learn the shared mechanism $\mathbf{f}$ (or precisely its inverse) as part of $\mathbf{h}$, along with $\mathbf{T}$.

While we can recover only the sufficient statistics $T_i(E_{i,p})$, not $E_{i,p}$, they are sufficient for building a method for cause-effect inference; $T_i(E_{i,p})$ generally has the same independence relationships with other variables as $E_{i,p}$. In practice, under the assumption that there exist direct causal effects, we can just compare values of an independence measure, as we will detail in Sec. 6.3.

Now, thoughtful readers would counter that we cannot expect many causal pairs satisfy A1 and A2 of Theorem 4. However, it is reasonable that there are small sets of (say, a handful) pairs among the many pairs on which the theoretical conditions are at least loosely satisfied. It is particularly the case for the real-world dataset we will consider in the experiments, where there are sets of pairs from the same causal scenario, e.g., {`altitude`, `temperature`} and {`altitude`, `rainfall`} could share similar mechanisms.

In Sec. 6.3, the theoretical results are derived when the assumptions of TCL are satisfied. In Sec. 6.4, we use ensemble method to exploit the imperfect TCLs trained on those loosely satisfactory sets mentioned in the previous paragraph.

6.3 Theoretical Results

6.3.1 Separation of Training and Testing

It should be clear from the above that we want to learn causal mechanism via TCL. However, to successfully learn TCL, we at least need to know that the pairs indeed share causal mechanism! To address the above dilemma, our idea is to learn causal

mechanism from some training pairs that we have good causal knowledge (e.g. we might know their SCMs and causal directions), and then predict the causal directions for unseen pairs. The following corollary of Theorem 4 makes this separation possible:

Corollary 2 (**Transferability** of TCL)**.** *Assume:*

A1. Pairs $\mathcal{X}^{tr}(P)$ *satisfy A1 and A2 of Theorem 4.*

A2. A pair $\mathbf{X}^{te}$ *satisfy A1 of Theorem 4, with the same* $\mathbf{f}$ *and* $\mathbf{T}$ *as* $\mathcal{X}^{tr}(P)$*, but different parameter* η_i*.*

A3. Let $\mathcal{R}_X$ *denote the support of a random variable* X*. We have* $\mathcal{R}_{E_i^{te}} \subseteq \cup_{p=1}^{P} \mathcal{R}_{E_{i,p}^{tr}}, i = 1, 2$*.*

A4. We learn a feature extractor $\mathbf{h}$ *on* $\mathcal{X}^{tr}(P)$ *as in A3 of Theorem 4 and have* $\mathbf{T}(\mathbf{E}_p^{tr}) = \mathbf{A}\mathbf{h}(\mathbf{X}_p^{tr}) + \mathbf{b}$*. Then, we have* $\mathbf{T}(\mathbf{E}^{te}) = \mathbf{A}\mathbf{h}(\mathbf{X}^{te}) + \mathbf{b} = \mathbf{hICA}(\mathbf{X}^{te})$*.*

Intuitively, after we successfully learned TCL $\mathbf{h}$, we can re-use it to analyze other unseen pairs that have the same SCM and sufficient statistics as the training pairs. We should note that, as in transfer learning, training and testing pairs do *not* have the same distribution, and hence the name of this corollary. From now on, we will also refer to the learning of TCL and analysis of new pairs on it as training and testing, respectively.

6.3.2 Inference Methods and Identifiability

We first present a general procedure (Algorithm 1) as the common basis, before detailing the two inference rules (`inferule`) with their identifiability results (and also *Direction*^{tr} and `align`). In the following, $\alpha_0 = (1, 2)$ and $\alpha_1 = (2, 1)$ denotes the two permutations on $\{1, 2\}$, and $\alpha_i(\mathbf{X}) := (X_{\alpha_i(1)}, X_{\alpha_i(2)})$.

Algorithm 1: Inferring causal direction

input : $\mathcal{X}^{tr}(P)$, $\mathbf{X}^{te}$, $Direction^{tr}$, `align`, `inferule`

output: $Cause^{te}$

1 Align training set, exploiting $Direction^{tr}$:
$\mathcal{X}^{al}(P) = \texttt{align}(\mathcal{X}^{tr}(P), Direction^{tr})$

2 Learn TCL $\mathbf{h}$ on $\mathcal{X}^{al}(P)$

3 **foreach** $\alpha = \alpha_0, \alpha_1$ **do**

4 $\quad (C_1, C_2)^T_\alpha = \mathbf{hICA}(\alpha_i(\mathbf{X}^{te}))$

5 Run inference rule: $Cause^{te} = \texttt{inferule}(\mathbf{C}_{\mathbf{ff}_0}, \mathbf{C}_{\mathbf{ff}_1}, \mathbf{X}^{\mathrm{te}})$

With $\mathbf{T}(\mathbf{E}^{\mathrm{te}})$ recovered, we can find ways to infer a causal direction for $\mathbf{X}^{te}$. To find the asymmetry between the two possible causal directions, we use the fact that, when testing, if we *flip* input direction to **hICA** and try nonlinear ICA for each (line 3,4 Algorithm 1), there will be one and only one trial that is realized by the **hICA**. This information will be exploited in `inferule` (line 5 Algorithm 1).

A remaining issue is that, to apply Theorem 4 and in turn Corollary 2, we need to at least partially know the directions of training pairs. More precisely, $\mathcal{X}^{tr}(P)$ must be *aligned*, as in the following definition. (This is implied by $\forall p\,(\mathbf{X}_{\mathrm{p}} = \mathbf{f}(\mathbf{E}_{\mathrm{p}}))$ in A1 of Theorem 4.)

Definition 7. Causal pairs $\mathcal{X}^{al}(P) := \{\mathbf{X}_p\}_{p=1}^P$ are **aligned** if $\forall p\,(X_{1,p} \to X_{2,p})$ or $\forall p\,(X_{2,p} \to X_{1,p})$.

In the first inference rule, it is assumed that we know the causal direction for each of the training pairs so that they can be trivially aligned. For a test pair, a realized (successful) nonlinear ICA among the two trials should output independent components, and this in turn tells us the direction of the pair, because we know which input of $\mathbf{h}$ corresponds to the cause. This leads to the following theorem:

Theorem 5 (**Identifiability** by independence of hidden components)**.** *In Algorithm 1, let:*

$Direction^{tr} = \{c_p\}_{p=1}^P$ *where* $c_p \in \{1, 2\}$ *is the cause index:* $X^{tr}_{c_p,p} \to X^{tr}_{3-c_p,p}$,

$\texttt{align} = \{X^{tr}_{c_p,p}, X^{tr}_{3-c_p,p}\}_{p=1}^P$,

$\texttt{inferule} = \alpha^*(1), \alpha^* = \arg\max_{\alpha\in\{\alpha_0,\alpha_1\}} \texttt{dindep}(\mathbf{C}_\alpha)$ *where* `dindep` *measures degree of independence.*

And assume:

A1. Causal Markov assumption and causal faithfulness assumption hold for data generating SCMs and analysis procedure except[1] *for a realized nonlinear ICA.*

A2. $\mathcal{X}^{tr}(P)$ *and* $\mathbf{X}^{te}$ *satisfy A1–A3 of Corollary 2.*

Then, the `inferule` *defined above (*`inferule1` *afterwards) identifies the true cause variable.*

The second inference rule only assumes we know how to align the training pairs. In fact, under certain practical scenarios, we know the training pairs *are* aligned; for example, 1) pairs from multiple environments (per environment per pair), as in many domain adaptation problems and in Monti, Zhang, and Hyvärinen, 2019, and 2) pairs from stratified sampling (per sample per pair).

The `inferule` determines the realized trial and identifies causal directions, *without* the directions of training pairs. We examine the independence of the pair $\{T_j(E_j^{te}), X_i^{te}\}$, as in the relation (2.8). Note, however, that as described in Monti, Zhang, and Hyvärinen, 2019, the outputs of a realized nonlinear ICA are equivalent to hidden variables only up to a permutation, i.e. $\mathbf{T}(\mathbf{E}^{\text{te}}) = (C_{\alpha(1)}, C_{\alpha(2)})^T$, with α unknown. This requires us to evaluate the degree of independence for four pairs at each trial, as in the following theorem:

Theorem 6 (**Identifiability** by independence of noise and cause)**.** *In Algorithm 1, let:*

$\mathit{Direction}^{tr} = \{i_p\}_{p=1}^{P}$ *where* $i_p \in \{1,2\}$ *such that* $\forall p\,(X_{i_p,p} \to X_{3-i_p,p})$ *or* $\forall p\,(X_{3-i_p,p} \to X_{i_p,p})$

$\texttt{align} = \{X_{i_p,p}^{tr}, X_{3-i_p,p}^{tr}\}_{p=1}^{P}$,

$\texttt{inferule} = i^*, (i^*, ., .) = \arg\max_{i,j,\alpha} \texttt{dindep}(X_i^{te}, C_{j,\alpha})$.

And assume the same as Theorem 5.

Then, the `inferule` *defined above (*`inferule2` *afterwards) identifies true cause variable.*

Since we can use the causal directions to recover an aligned training set, so in Theorem 5, letting `inferule` = `inferule2`, the true causal index can also be identified. However, as we will see in the experiments, `inferule1` will outperform `inferule2` if the former is applicable in practice.

[1] See Sec. 2.2.2 on this.

Finally, we will employ distance correlation (dCor) (Székely, Rizzo, Bakirov, et al., 2007) as our main choice of `dindep`.

6.3.3 Choice of Independence Test

HSIC is a widely used independence test in causal discovery literature, but it has several drawbacks. First, its test statistic is not normalized for different testing pairs, and thus not comparable[2]. Second, although p-value of the test is comparable, it does not directly measure the degree of independence. Most importantly, as mentioned in Mooij et al. (2016, sec. 2.2), standard threshold of the test would be too tight for our purpose. This is because in causal discovery we often want to test the independence between an observed variable and an estimation *from* observed data, and there always exists small dependence with finite sample and other real world limitations. For the same reason, the flexibility of HSIC to detect dependence can do harm, not benefit, to causal discovery.

Unlike HSIC[3], dCor value is always in [0, 1], and equals to 0 if and only if the pair under test are independent. Thus, the value 1 - dCor works as a comparable degree of independence. As a bonus, dCor is much faster than HSIC when testing independence between univariate real-valued variables, particularly when sample size is large [4].

Hence, we suggest dCor rather than HSIC as the default choice to measure degree of independence for cause-effect inference, and try HSIC when you can afford the time, both for tuning and running.

6.3.4 Structural MLP

We discuss an MLP structure to improve TCL's performance on bivariate analyzable SCMs. We first study the form of the inverse SCM, since this is what the MLP should learn.

[2] If we use the default Gaussian kernel and median heuristic for kernel bandwidth (Gretton et al., 2005). And this is also the most common way it is used in bivariate causal discovery (Mooij et al., 2016; Hu et al., 2018)

[3] We noticed that distance covariance is an instance of HSIC for certain choice of kernels (Sejdinovic et al., 2013). But again, this is not default for HSIC.

[4] We use Huo and Székely, 2016 for dCor and Zhang et al., 2018 for HSIC, the implementation can be found at https://github.com/vnmabus/dcor and https://github.com/oxmlcs/kerpy, respectively.

Proposition 7 (Inverse of bivariate analyzable SCM)**.** *For any analyzable SCM as shown in* (2.3)*, denote the whole system* $\mathbf{X} = \mathbf{f}(\mathbf{E})$*, if the Jacobian matrix of* $\mathbf{f}$ *is invertible, then* f_1 *is invertible.*

Denote $g_1 = f_1^{-1}$, then $E_1 = g_1(X_1)$. And we have $E_2 = g_2(X_1, X_2)$ in general. This implies the inverse SCM has the graph as shown in Figure 6.2 (left):

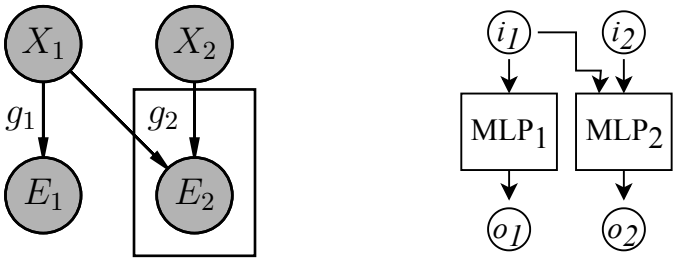

FIGURE 6.2: Inverse bivariate analyzable SCM (left) and the indicated MLP structure (right).

Building an MLP for TCL with this asymmetric structure will help TCL learn the inverse SCM. This can be easily implemented as shown in Figure 6.2 (right): we build an MLP with one output node for g_1 and g_2 respectively, and then concatenate the outputs together.

Caveats on Structural MLP

1) While one might think that we need to make MLP1 invertible since g_1 is invertible, we should *not* impose it; the sufficient statistics $\mathbf{T}$ are also learned as part of MLP, and they are in general non-invertible. 2) The structural MLP works only when there is a direct causal effect, as required by SCM (2). 3) Since node i_1 corresponds to the cause, we need to input the cause variable to i_1 for training the asymmetric MLP properly. This requires knowledge on the causal directions of training pairs, and thus, we can only apply it with `inferule1`.

6.4 Assembling Causal Mosaic

In the following, we will refer to training pairs that satisfy A1 (same SCM and exponential family) and A2 (enough variability among parameters) of Theorem 4 as *tessera pairs*, because they form the small portion of causal pairs that can be easily

modeled together, and thus a small block of the whole mosaic. Also, we will refer to a TCL learned on tessera pairs as a *tessera*.

We have so far assumed that we have tessera pairs, under the ideal situation that we have well-studied systems. However, for many real world applications, it is unlikely that most training pairs amount to tessera pairs. Our idea for handling real world problems is to train many TCLs on random selections of pairs, and then choose from these TCLs the (imperfect) tesserae that are trained on *approximate* tessera pairs, in the sense that they have similar SCMs and are approximately in the same family. We further develop an ensemble method to effectively exploit imperfect tesserae.

In this section, Let S be the set of all labeled causal pairs we have at hand, and c_s be the true cause index for $s \in S$.

6.4.1 Preparing Materials

As in Algorithm 2, by training a large number (N) of TCLs on randomly chosen pairs, we hope some of these TCLs amount to tesserae. To ensure TCL is trained properly on each set of pairs, we train MLP M times with different hyperparameters (See experiment for details).

Algorithm 2: Random training of TCLs

input : S, M, N

output: $\{(\mathbf{h}_n, T_n)\}_{n=1}^{N}$

foreach *n in 1,...,N* **do**
- Randomly choose training pairs $T_n \subset S$
- Split the sample points of each training pair by half, and build training set *Tr* and testing set *Te*
- **foreach** *m in 1,...,M* **do**
  - Randomly choose a set of hyperparameters and train TCL on *Tr*
  - Evaluate classification accuracy ($Cacc_m$) for pair index on *Te*.
- Use the trained TCL with the highest $Cacc_m$ for this set of training pair, denote it $\mathbf{h}_n$

6.4.2 Choosing Tesserae

Because our goal is to infer causal directions, we choose TCLs that perform well on this task. First, we can use each TCL to infer the causal directions of its own training pairs (Algorithm 3, line 2,3), and choose TCLs that produce accuracy higher than a threshold $ThreT$. Second, for each TCL, we also input unseen validation pairs and infer their directions, and we choose TCLs that produce accuracy higher than $ThreV$. The good training accuracy indicates the success of training and TCL indeed learned to infer causal directions. The good validation accuracy shows that the learning generalizes to unseen pairs.

To efficiently use S for training and validation, and still be able to test on all the pairs in S, we use the idea of leave-one-out cross validation (LOOCV). That is, each pair l not used in training a TCL is left out once when validating that TCL (line 5,6). As we can see, every pair in S is not used as a training pair or validating pair for its tessera (line 10,11). On the other hand, in training (T_n) and validation ($(S \setminus T_n) \setminus \{l\}$), every trained TCL exploits all the pairs except the left out one l.

Algorithm 3: Selecting TCLs

input : S, $ThreT$, $ThreV$, $\{(\mathbf{h}_n, T_n)\}_{n=1}^{N}$

output: $\{TSR_s : s \in S\}$

foreach *n in 1,...,N* **do**

  // Training accuracy $Tacc_n$ for $\mathbf{h}_n$ on T_n

  foreach *t in* T_n **do**

    Use **hICA**$_n$, run line 3–5 of Algorithm 1 on t, get inferred direction $\hat{c}_t$

  $Tacc_n = |\{t : \hat{c}_t = c_t\}|/|T_n|$

  // LOOCV

  foreach *l in* $S \setminus T_n$ **do**

    As line 2–4, get validation accuracy for $\mathbf{h}_n$ on $(S \setminus T_n) \setminus \{l\}$, denote it as $Vacc_n(l)$

// Select TCLs by accuracy thresholds

foreach *s in S* **do**

  Initialize tessera index set for s: $TSR_s = \varnothing$

  foreach *n in 1,...,N* **do**

    if $s \notin T_n$ **and** $Tacc_n > ThreT$ **and** $Vacc_n(s) > ThreV$ **then**

      Add n to TSR_s

By the identifiability theorems, if TCL $\mathbf{h}_n$ has high training accuracy, it is likely that the training pairs T_n are approximate tessera pairs (required by A1 & A2 of Theorem 4). Similarly, if $\mathbf{h}_n$ gives high validation accuracy, the evidence for tessera pairs T_n is strengthened (required by A1 of Corollary 2), and further it is likely that pairs T_n are similar to many of pairs in $S \setminus T_n$ (required by A2 of Corollary 2).

6.4.3 From Tesserae to Causal Mosaic

We employ an ensemble method for making effective use of each imperfect tessera, and construct a whole piece of mosaic, in the same way as we will obtain a strong classifier from weaker ones by ensemble methods. Put simply, for each testing pair, ensemble method will take the causal direction predicted by tesserae, and produce a final, weighted average. We introduce two levels of weighting as follows.

First, as Algorithm 4, line 3, we weight a TCL $\mathbf{h}_n$ by the average $\texttt{dindep}(\mathbf{hICA}_n(.))$ for the training pairs T_n. This is to address the problem that, even if we have selected

TCLs as in Algorithm 3, it is very possible that the chosen tesserae would not be perfect, e.g., the mechanisms of training pairs are not exactly the same. Thus, we use this weight to measure how well T_n fit together (by Theorem 4, if we get more independent components, A1 & A2 are more likely to hold), and in turn how likely the causal direction will be correctly inferred if we use this $\mathbf{h}_n$.

Second, we weight by the $\texttt{dindep}(\mathbf{hICA}_n(.))$ for a particular testing pair s. Again, even if w_n is large, it is possible that s and T_n do not satisfy A2 of Corollary 2, so we need to weight each tessera for *each* testing pair. Similarly to the reasoning for w_n, if we get independent components for s, A2 of Corollary 2 is likely to hold. Note that, as in Algorithm 1, in theory only realized nonlinear ICA outputs independent components, so we weight by the larger `dindep` of the two trials (line 4–7). We multiply the two weights as the final pair-specified weight.

Algorithm 4: Ensemble method

input : S, $\{TSR_s : s \in S\}$, $\{(\mathbf{h}_n, T_n)\}_{n=1}^N$

output: $\{Direction_s : s \in S\}$

1 **foreach** s *in* S **do**

2 **foreach** n *in* TSR_s **do**

3 $w_n = \sum_{t \in T_n} (\texttt{dindep}(\mathbf{hICA}_n(t)))/|T_n|$

4 **foreach** $i = 0, 1$ **do**

5 $\mathbf{C}_{\alpha_i} = \mathbf{hICA}_n(\alpha_i(s))$

6 $w_{ns,i+1} = \texttt{dindep}(\mathbf{C}_{\alpha_i})$

7 $w_{ns} = \max(w_{ns,1}, w_{ns,2})$

8 $\hat{c}_s = \texttt{inferule}(\mathbf{C}_{\mathbf{ff}_0}, \mathbf{C}_{\mathbf{ff}_1}, s)$

9 $Direction_{ns} = 1$ if $\hat{c}_s = 1$, -1 if $\hat{c}_s = 2$

10 Calculate weighted prediction $Score_s = \sum_{n \in TSR_s} w_n w_{ns} Direction_{ns}$

11 $$Direction_s = \begin{cases} X_1 \to X_2 & Score_s > 0 \\ X_2 \to X_1 & Score_s < 0 \\ ? & Score_s = 0 \end{cases}$$

6.4.4 Alternative Ensemble Scorings

Without loss of generality, assume X_1 is input to the same node when calculating $w_{ns,1}$, as cause variable is when training. Then we have $Direction_{ns} = \mathbb{I}(w_{ns,1} > w_{ns,2}) - \mathbb{I}(w_{ns,1} < w_{ns,2})$ where indicator function $\mathbb{I}$ maps $true/false$ to $1/0$.

Now the ensemble score in Algorithm 4 line 10 becomes:

$$\begin{aligned} Score_s = & \sum_{n \in TSR_s} w_{ns,1} w_n \mathbb{I}(w_{ns,1} > w_{ns,2}) \\ & - \sum_{n \in TSR_s} w_{ns,2} w_n \mathbb{I}(w_{ns,1} < w_{ns,2}) \end{aligned} \tag{6.1}$$

But since $\mathbb{I}(w_{ns,1} > w_{ns,2})$ and $\mathbb{I}(w_{ns,1} < w_{ns,2})$ just reflect the relative value of $w_{ns,1}$ and $w_{ns,2}$, the following simplification is reasonable:

$$Score_s = \sum_{n \in TSR_s} w_n (w_{ns,1} - w_{ns,2}) \tag{6.2}$$

And on the same line of reasoning, we can alternatively disregard $w_{ns,1}, w_{ns,2}$ and have:

$$\begin{aligned} Score_s = & \sum_{n \in TSR_s} w_n \mathbb{I}(w_{ns,1} > w_{ns,2}) \\ & - \sum_{n \in TSR_s} w_n \mathbb{I}(w_{ns,1} < w_{ns,2}) \\ = & \sum_{n \in TSR_s} w_n Direction_{ns} \end{aligned} \tag{6.3}$$

This is just the weighted average of prediction by each $\mathbf{h}_n$. And finally, since $\mathbf{h}_n$ with small w_n is unlikely to produce large $w_{ns,i}$, we can further disregard w_n in (6.2). This gives:

$$Score_s = \sum_{n \in TSR_s} (w_{ns,1} - w_{ns,2}) \tag{6.4}$$

We compared these scoring equations and found the last one is stably the best.

6.5 Experiments

None of the methods compared in this section use an ensemble. Among them, Non-SENS is the most related because it is also based on TCL, but it does not exploit the

known direction of training pairs in the multi-pair setting detailed below, related to our Theorem 5. NCC also uses labeled pairs as training data but requires very large numbers of training pairs because it does not split and re-use the pair for each testing pair, as in our Sec. 6.4.2. ANM uses an additive noise model, but our model does not assume an additive noise. IGCI and RECI propose simple but somewhat ad-hoc criteria and thus limit their performance; please refer to Literature Review for details.

6.5.1 Artificial Data

We compare NonSENS to variations of our method with different inference rules, independence measures, and MLP types on artificial data. To see the comparisons with other recent methods on similar artificial data, we refer readers to Monti, Zhang, and Hyvärinen, 2019.

Multi-environment setting This is the setting under which NonSENS works. Mathematically, our tessera pairs $\{\mathbf{X}_p^{tr}\}$ are equivalent to the samples $\mathcal{X}^{en} := \{\mathbf{X}_p^{en}\}$ of a *same* causal system under P different "environments" in their interpretation. That is, they define different environments by different parameter $\boldsymbol{\eta}$ of hidden variables, and $\forall p(\mathbf{X}_p^{en} = \mathbf{f}(\mathbf{E}_p^{en}))$ is by definition satisfied. Moreover, there is no separate testing pairs here. Our goal is to distinguish between two possibilities, $\forall p\,(X_{1,p}^{en} \to X_{2,p}^{en})$ or $\forall p\,(X_{2,p}^{en} \to X_{1,p}^{en})$, for $\mathcal{X}^{en}$ themselves (note the pairs (environments) are *aligned*), rather than 2^P possibilities for individual pairs $\mathcal{X}(P)$.

Our Algorithm 1 can reduce to this setting, as shown in Algorithm 5. Both training and testing pairs are $\mathcal{X}^{en}$ themselves. Note that $Direction^{tr}$, `align` and the input permutation (Algorithm 1, line 3,4) are not needed, since $\mathcal{X}^{en}$ is already aligned. We apply a simplified version of `inferule2` to infer direction for each environment without input permutation, but still need to deal with the output permutation.

Finally, we use majority voting to combine the results of all environments and give the final decision, and this is an important difference between our method and NonSENS under this setting. NonSENS treats the samples of environments as coming from a mixture, runs `dindep` on pooled sample and output, and gives

$c^{en} = i^*, (i^*, j^*) = \text{argmax}_{i,j}\, \texttt{dindep}(\{X^{en}_{i,p}\}, \{C_{j,p}\})$[5]. In practice, as we will see, majority voting often outperforms NonSENS since it uses information from each environment and thus is more robust.

Algorithm 5: Algorithm 1 on multi-env. setting

input : $\mathcal{X}^{en}$

output: c^{en}

1 Learn TCL **h** on $\mathcal{X}^{en}$

2 $\mathcal{C} = \textbf{hICA}(\mathcal{X}^{en})$

3 **foreach** $\mathbf{X}^{en}_p$ *in* $\mathcal{X}^{en}$, $\mathbf{C}_p$ *in* $\mathcal{C}$ **do**

4 $\quad c_p = i^*, (i^*, j^*) = \arg\max_{i,j} \texttt{dindep}(X^{en}_{i,p}, C_{j,p})$

// Majority voting

5 $c^{en} = \arg\max_i |\{c_p : c_p = i\}|$

Multi-pair setting If we know the *directions* of training pairs, we separate training and testing, and both Theorem 5(`inferule1`) and Theorem 6 (`inferule2`) can apply. Here, we infer the direction for each individual testing pair. NonSENS cannot apply here, so we compare different variations of our method. We name this multi-pair setting, to contrast the multi-environment setting, although the main difference is the direction information of training pairs (our method *can* also infer for each environment as in Algorithm 5, line 3,4).

Data generation As in Hyvärinen and Morioka, 2016 and Monti, Zhang, and Hyvärinen, 2019, we use 5-layer randomly initialized MLPs as mixing functions, with leaky ReLU activation and 2 units in each layer to ensure invertibility. To simulate the independent relationships of a direct causal graph, we use a lower-triangle weight matrix for each layer of the MLP. We use Laplace distribution for both hidden components, and their variance parameters are i.i.d. generated across different pairs. Multi-environment setting can be easily simulated by aligning all the pairs and then perform nonlinear ICA.

[5]Originally, NonSENS uses independence tests with a threshold. We write it here using `dindep` for easy comparison, because we will use this modified rule for NonSENS in experiment.

We generate 100 mixing functions and same number of training/testing pairs for each mixing function. To observe how the pair number affect results, we try 5 different number ranging from 10 to 50. Please see Sec. 6.6 for more details.

Hyperparameters To make fair comparisons, for both our method and NonSENS, we keep all the hyperparameters the same, including the parameters for training and independent tests. Please see Sec. 6.6 for details.

Assuming direct causal effect Our method and NonSENS[6] formally requires direct causal effects exist between pairs, and this is our main experiment setting. Please see Sec. 5 for the experiment without this assumption.

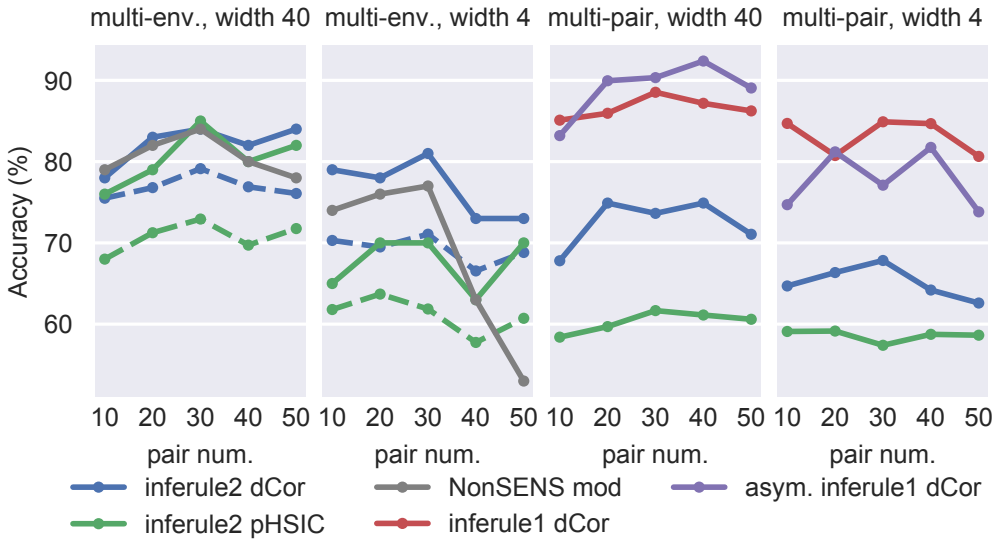

FIGURE 6.3: Performance assuming direct causal effect. "width" means MLP width. In the legend, "dCor/pHSIC" indicates the independence measure, and "asym." means asymmetric MLP in TCL. Dashed lines are intended to show transferability of TCL, see Sec. 6.6.

As shown in Figure 4, in multi-environment setting, our method outperforms NonSENS, particularly when the pair number is large. The decreasing performance of NonSENS is consistent with the results when not assuming pure causal effects and is due to the unwanted dependence between estimated noise and the cause, as explained in detail in Sec. 6.6.

[6]We cannot reproduce the likelihood ratio based NonSENS proposed for this setting. Instead, we use a slightly modified version of NonSENS originally proposed for may-not-direct-causal setting, see the previous footnote.

In multi-pair setting, `inferule1` is applicable and performs much better than `inferule2`. The main reason is that the independence between two output components is much easier to realize than the independence between estimated noise and observed cause. And this is in turn because of the direct dependence between observed variables and outputs (see Figure 1 in Sec. 6.6). Note that Theorem 5required known causal directions of training pairs, and thus cannot be used in multi-environment setting.

Moreover, when the MLP width is 40, `inferule1` achieves near-optimal results when applied with asymmetry MLP. This is also the best result we have obtained with artificial data. While the asymmetry MLP with width 4 performs worse than the fully-connected one, this is due to the limited fitting capacity (see Sec. 6.6 for details).

When inferring by Theorem 6, we try both dCor and the p-value of HSIC (Gretton et al., 2005) as `dindep`. dCor constantly outperforms HSIC (See Sec. 6.6 for details).

Experiments without Assuming Direct Causal Effect

We also experiment without assuming direct causal effect necessarily exists, and allow "inconclusive" outputs when the assumption is possibly violated. The purpose here is mainly to conform the problem mentioned in S.3 above, and to show how our method can address it to a large extent. When applying the inference rules, now we need to set a threshold or alpha value for the independence tests. For clearer comparisons, we apply Theorem 6 and also use HSIC, though Theorem 5or other independence tests can also be applied. Then our method only differs with NonSENS by inferring for each environment and then majority voting.

Similarly to Monti, Zhang, and Hyvärinen, 2019, we evaluate on two datasets: 1) all pairs are direct causal (1st row). 2) all pairs are purely confounded (simply use a fully connected MLP) (2nd row). On direct causal pairs, we can see NonSENS' accuracy decreases drastically w.r.t pair number and is nearly always below 10% when MLP width is 4. On the other hand, on purely confounded pairs, it always reports 100% inconclusive.

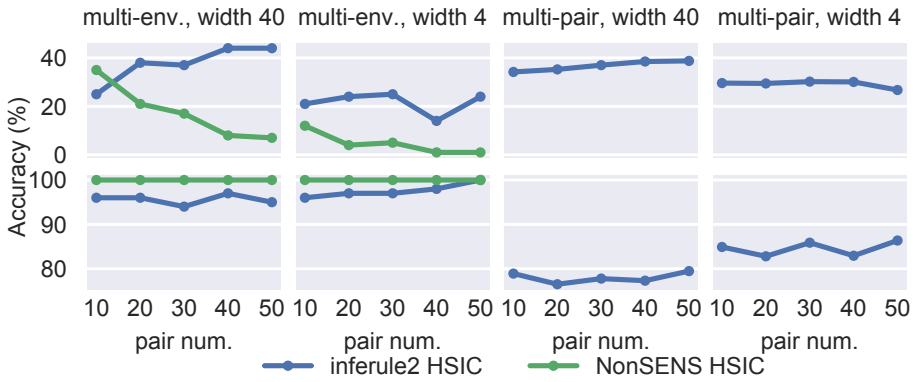

FIGURE 6.4: Performance without assuming direct causal effect. 1st/2nd row is results on direct causal data/purely confounded data respectively.

Here the results conform that the default alpha value (0.05) for independence test is way too tight. Specifically, the problem here is that, with more pairs (which means more sample points for NonSENS), HSIC is more sensitive to small dependence between estimated noise and observed cause. This means we must train TCL very optimally to avoid the unwanted dependence.

Our method performs much better than NonSENS, especially with large pair number. The reason is that, it is easier to get rid of unwanted dependence by looking at each environment, since if any one of the environments shows dependence, then the pooled data tested in NonSENS will be dependent.

6.5.2 Real World Dataset

Tuebingen cause-effect pairs (TCEP) dataset (Mooij et al., 2016, dataset version December 2017) is a commonly used benchmark for cause-effect inference tasks. Causal Mosaic can be suitably applied here because of the very diverse scenarios of the pairs. Each pair is assigned a weight in order to account for the possible correlation with other pairs that are selected from the same multivariate scenario. Currently, the dataset contains 108 real-world cause-effect pairs with true causal directions labeled by human experts. We exclude 6 multivariate pairs in our evaluation.

Implementation We use Theorem 5 with asymmetric MLP since it already shows much better results on artificial data. Unlike on artificial data with Laplace hidden variables, we use maxout activation for the output layer. Since the sample sizes of

TCEP pairs range wildly from a hundred to several thousands, we fix this imbalance in classification by under-sampling using imbalanced-learn package (Lemaître, Nogueira, and Aridas, 2017). When implementing Algorithm 4 line 10, we use a simplified version $Score_s = \sum_{n \in TSR_s}(w_{ns,1} - w_{ns,2})$, since this works the best. See Sec. 6.4.4.

Hyperparameters We train TCL on 300 (N) sets of randomly picked pairs, which are of size ranging from 4 to 32. For selecting TCLs, we randomly search 100 pairs of accuracy thresholds $(ThreT, ThreV)$ in $[65\%, 75\%]^2$ and rule out too large thresholds that give 0 or only 1 tessera for more than 10 TCEP pairs. We train 10 (M) TCLs on each pair set and choose the best, and the following hyperparameters are randomly searched from uniform distributions: depth and width of MLP, learning rate, decay factor, max step (decay step is 10% of max step), momentum, and batch size. Among them, the depth of MLP larger than 10 might lead to divergence in training, but the ranges of other parameters seem to have few impacts if we do not use some extreme values. To save training time, we change the ranges of MLP width and max. step according to training pair number (small width and step for small pair number).

TABLE 6.1: Accuracy (%) on TCEP. "A/B" means with/without applying pair weight.

ANM	IGCI	RECI	NCC	OURS
52.5/52.0	60.4/60.8	70.5/62.8	51.8/56.9	**81.5**$_{\pm 4.1}$/**83.3**$_{\pm 5.2}$

We compare our method to ANM, IGCI, RECI and NCC, using implementations from CDT package (Kalainathan and Goudet, 2019). The results are shown in Table 1. We report the *median* and std-error of accuracies of our method calculated on all the 83 pairs of thresholds. And this already shows state-of-the-art performance. The best result on all thresholds is 86.3% without pair weight and might overfit TCEP dataset. For NCC, we infer each pair by training the method on rest of the pairs. The accuracy is much worse than the reported 79% in Lopez-Paz et al., 2017, the most possible reason is that NCC requires much more training data (320,000 artificial pairs in the original paper). The performance of ANM is worse than reported in Mooij et al., 2016, possibly because of the different implementation of independence test.

6.6 Details and Notes for Artificial Experiments

Training and testing data As mentioned, under multi-environment setting, the pairs are for both training and testing. Under multi-pair setting, these same pairs are again used for testing. But for training, we generate another set of pairs with random parameters, while the mixing functions and pair number for each mixing function are the same as testing pairs. For each pair, we always generate 512 data points.

Hyperparameters For the MLP in TCL, we use the same number of layers as data-generating MLP, and each hidden layer has same number of units (4 or 40 in the experiments) with the maxout activation. The two output units have the absolute value function as activation. For the asymmetric MLP (Figure 3, right), we use same width for both sub-MLPs, and keep the sum of the widths the same as fully-connected MLP. Note that the asymmetric MLP has much less parameters than the fully-connected one, since the sub-MLPs are disconnected. We use Momentum optimizer with momentum 0.9 and initial learning rate 0.01, and the batch size is 32.

MLP width The experimental results show that we need large enough MLP to fit more pairs. Note in particular that the MLP of width 4 performs almost always worse than that of width 40. If we use asymmetric MLP, this tendency is more drastic since it has much less parameters. When the MLP width is 4, the accuracy often decreases w.r.t the number of training pairs. When the MLP width is 40, the accuracy usually increases w.r.t the number of training pairs, but when the pair size is larger than 30, it increases slowly or even slightly drops.

Training pair number We observe better performance as the pair size grows (under the MLP width 40). Under the multi-pair setting, this implies that TCL learns more thoroughly the shared mechanism. Under multi-environment setting, we have one more reason: majority voting performs better with more voters (pairs).

Transferability To confirm the transferability of TCL, we also try inferring directions for individual pairs without voting under multi-environment setting (Figure

4, dashed lines). The results from the two settings are similar, meaning the transferability. The slight drop of performance under multi-pair setting should come from the two input trials needed.

6.7 Proofs

Corollary 2

Proof. From A4, and substitute $\mathbf{X}_{\mathrm{p}}^{\mathrm{tr}} = \mathbf{f}(\mathbf{E}_{\mathrm{p}}^{\mathrm{tr}})$, we have $\mathbf{T}(\mathbf{E}_{\mathrm{p}}^{\mathrm{tr}})^{\mathrm{T}} = \mathbf{A}\mathbf{h}(\mathbf{f}(\mathbf{E}_{\mathrm{p}}^{\mathrm{tr}})) + \mathbf{b}$.

From A3, we know each $\mathbf{X}^{te}$'s support is contained in the support of $\mathbf{h}$. Thus, we can replace $\mathbf{E}_p^{tr}$ with $\mathbf{E}^{te}$ and the equality still holds, we get: $\mathbf{T}(\mathbf{E}^{\mathrm{te}}) = \mathbf{A}\mathbf{h}(\mathbf{f}(\mathbf{E}^{\mathrm{te}})) + \mathbf{b} = \mathbf{A}\mathbf{h}(\mathbf{X}^{\mathrm{te}}) + \mathbf{b}$.

□

Theorem 5

Proof. Without loss of generality, assume after alignment cause variable for each training pair is input to $\mathbf{h}$ as the first argument. By A2 and Theorem 4, we will successfully learn $\mathbf{h}$ (Algorithm 1, line 1,2).

By A2 and Corollary 2, if the cause variable of $\mathbf{X}^{te}$ is input to **hICA** as the first argument, then its nonlinear ICA is realized (Algorithm 1, line 3,4). Denote the respective input permutation as α_r, then $C_{\alpha_r(1)} \perp\!\!\!\perp C_{\alpha_r(2)}$. While for the other input direction α_{1-r}, by A1, $C_{\alpha_{1-r}(1)} \not\perp\!\!\!\perp C_{\alpha_{1-r}(2)}$

Thus, we have $\texttt{dindep}(\mathbf{C}_{\alpha_r}) > \texttt{dindep}(\mathbf{C}_{\alpha_{1-r}})$, and $\alpha^* = \alpha_r$. □

Theorem 6

Proof. Similarly to the proof of Theorem 5, we know there is one and only one input direction α_r where nonlinear ICA is realized. We have $\mathbf{T}(\mathbf{E}^{\mathrm{te}}) = (C_{\alpha(1)}, C_{\alpha(2)})^T_{\alpha_r}$ where α is the unknown output permutation.

By A1 (which also implies rule (4)), we have $X_c^{te} \perp\!\!\!\perp C_{3-c,\alpha_r}$ where c is the cause index, but $X_i^{te} \not\perp\!\!\!\perp C_{j,\alpha}$ for all other i, j, α. Thus, $(i^*, j^*, \alpha^*) = (c, 3-c, \alpha_r)$ □

Proposition 7

Proof. From Definition 4, we write $\mathbf{X} = \mathbf{f}(\mathbf{E})$ and denote $\mathbf{g} = \mathbf{f}^{-1}$. And we have the relation of Jacobians $\mathbf{J_g} = \mathbf{J_f^{-1}}$, and:

$$\begin{aligned}\mathbf{J_f^{-1}} &= \begin{pmatrix} \frac{df_1}{dE_1} & 0 \\ \frac{\partial f_2}{\partial X_1}\frac{\partial f_1}{\partial E_1} & \frac{\partial f_2}{\partial E_2} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\frac{df_1}{dE_1})^{-1} & 0 \\ -(\frac{\partial f_2}{\partial E_2})^{-1}\frac{\partial f_2}{\partial X_1}\frac{df_1}{dE_1}(\frac{df_1}{dE_1})^{-1} & (\frac{\partial f_2}{\partial E_2})^{-1} \end{pmatrix}\end{aligned}$$

By comparing the 1st row of $\mathbf{J_g}$ and $\mathbf{J_f^{-1}}$, we have $\frac{\partial g_1}{\partial X_2} = 0$ which indicates g_1 is not a function of X_2, and $\frac{dg_1}{dX_1} = (\frac{df_1}{dE_1})^{-1}$ which, by inverse function theorem, implies f_1 is invertible and $g_1 = f_1^{-1}$. □

6.8 Discussions

6.8.1 Combining Graphical Search Methods

There are already some studies that successfully combine cause-effect inference methods with graphical search methods; for example, cause-effect inference methods can be directly employed to infer the undirected edges output by search methods (Monti, Zhang, and Hyvärinen, 2019; Zhang and Hyvärinen, 2009), and overlapping datasets can be integrated using bivariate causal discovery to give more precise output class (Dhir and Lee, 2020). Our method can easily be applied in the same way to help multivariate causal discovery under confounding.

6.8.2 Invertibility Requirement in Definition 4

Our method is still valid if there exists a transformation $\boldsymbol{\tau}(\mathbf{E}) := (\tau_1(E_1), \tau_2(E_2))$ such that the transformed SCM satisfies the assumptions of Theorem 1 (e.g., $\mathbf{X} = \mathcal{F}(\boldsymbol{\tau}(\mathbf{E}))$ and $\mathcal{F}$ is invertible). By Theorem 1, the TCL followed by linear ICA can successfully output the sufficient statistics of $\tau_i(E_i)$, which plays the same role as E_i when testing independence. Note that now the mixing function $\mathbf{f} = \mathcal{F} \circ \boldsymbol{\tau}$ can be *non*-invertible. We believe that the existence of such $\boldsymbol{\tau}$ should prevail in practice, and the results on real world benchmark datasets suggest this. We can go a step further

to say $\boldsymbol{\tau}(\mathbf{E})$ are *the* exogenous variables, since, by definition, exogenous variables are unknown and the only requirement is that they are independent of each other.

Another note is that, Definition 1 does *not* mean that the function relating X_1 and X_2 should be invertible. Quite oppositely, take analyzable SCM (1), f_2 is a function from $\mathbf{R}^2$ to $\mathbf{R}$, which is always non-invertible. Moreover, even if $E_2 = e_2$ is given, the deterministic relation $X_2 = f_2^{e_2}(X_1) := f_2(X_1, e_2)$ could still be non-invertible.

Chapter 7

Conclusion

Our work in this thesis shows promising adaptations and applications of recent advances in nonlinear ICA and more generally probabilistic generative learning, a subfield of machine learning which is arguably most relevant to causality (see Section 7.3 below). However, before going into larger perspectives, the summaries of the two lines of our work are given first.

7.1 On Intact-VAE

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 have evidence that Intact-VAE works under unobserved confounding and believe that VAEs are suitable for *principled* causal inference owing to their probabilistic nature, if not compromised by ad hoc heuristics (Wu and Fukumizu, 2021).

The advantage of VAE approach can also be related to the fact that posterior effect estimation has minimum worst-case error under model misspecification (Bonhomme and Weidner, 2021). in our case, of the outcome noise and prior are possibly misspecified. We believe it is possible to extend the bounds in Sec. 4.3.2 to limited overlap, just as (Johansson et al., 2020) extends (Shalit, Johansson, and Sontag, 2017) to limited overlap, and leave this for future. To avoid potential negative societal impact (e.g, bad prescriptions), practitioners should judge the conditions of the

proposed method by their domain expertise, and careful trials are always recommended.

Our method outperforms or matches state-of-the-art methods under diverse settings including unobserved confounding. In Sec. 5.3, we explained why the current VAE methods are unsatisfactory from a more "causal" viewpoint. We discussed future theoretical work–approaches to identification and optimal estimation under unobserved confounding. We believe this series of work will also pave the way towards principled causal effect estimation by other deep architectures, given the fast advances in deep identifiable models. For example, recently, Khemakhem et al., 2020a provide identifiability to deep energy models, and Roeder, Metz, and Kingma, 2020 extend the result to a wide class of state-of-the-art deep discriminative models. We hope this work will inspire other methods based on deep identifiable models.

7.1.1 Future Work

As we see in the estimator (4.8), our representation Z is in fact capable of *counterfactual inference*: $\hat{t}$ can be different to factual $T = t$. Experiments on counterfactual generation, like those in (Kocaoglu et al., 2018, CausalGAN) and (Yang et al., 2020, CausalVAE), are on the way.

Since our method works without the recovery of either hidden confounder or true score distribution, we often cannot see apparent relationships between recovered latent representation and the true hidden confounder/scores. It would be nice to directly see the learned representation preserves causal properties, for example, by some causally-specialized metrics, e.g. Suter et al., 2019.

Despite the formal requirement in Theorem 1 of fixed distribution of noise on Y, inherited from Khemakhem et al., 2020b, the experiments show evidence that our method can learn the outcome noise. We observed that, in most cases, allowing the noise distribution to be learned depending on Z, T improves performance. Theoretical analysis of this phenomenon is an interesting direction for future work.

We conjecture that, it is possible to extend model identifiability to conditional noise models $g_t(Z)$. And we expect that the noise on Y can also be identified up to some eq. class (or joint eq. class together with f). In that case, the model identifiability may also be sufficient for causal inference, under some respective assumptions

on true generating process and our current assumptions in Theorem 1 can be *relaxed* to large extent. Similarly to current f, we may have identification for a general class of noises.

Also, our causal theory does not in principle require continuous latent distributions, though in Theorem 1, differentiability of f is inherited from iVAE. Given the fact that currently all nonlinear ICA based identifiability requires differentiable mapping between the latent and observables, directly based on it, theoretical extensions to *discrete* latent variable would be challenging. However, what is essential for CATE identification is the *same* transformation between true and recovered score distribution for both t, but the transformation needs *not* to be affine, and, possibly, neither injective. This opens directions for future extensions, based not necessarily on nonlinear ICA.

7.2 On Causal Mosaic

In this work, we proposed a highly flexible cause-effect inference method that learns a mixture of general nonlinear causal models, with proof of identifiability. We exploited TCL to extract the common mechanism shared by different causal pairs, and transferred the causal knowledge to unseen pairs. More specifically, our method learns how to distinguish cause from effect, from some training pairs, and predicts the causal direction on testing pairs. We gave two inference rules with identifiability proofs and an ensemble framework that works on real world cause-effect pairs with limited labeled causal directions. We compared our method to recent methods on artificial and real world benchmark datasets, and it showed state-of-the-art results.

Hence, we justified the “mosaic" perspective of causal discovery, which proposes to learn causality piecemeal, and then build a whole picture by the pieces. Here, shared mechanism learned by TCL forms a tessera of the whole causal mosaic, and many tesserae are learned and further combined into a whole picture by ensemble method. We believe this new perspective would promote other novel methods for bivariate and also more general causal discovery problems.

7.2.1 Future Work on Hidden Confounding

Tell Exactly Where the Correlations Come From

Generally, the relationship between two variables can be categorized into one of the four cases: 1) purely causal (no confounder between them), 2) totally confounded (none of them causes another), 3) causal relation and confounder both exist, 4) neither causal nor confounded. By the existence of statistical dependence, we can eliminate the last case. Before we could determine the causal direction, the question naturally arises: which case we are confronting? However, to the author's knowledge, no work has addressed this question explicitly. Most research only asks whether it is purely causal or not and, consequentially, cannot distinguish between 2) and 3). For example, as mentioned before, Zhang, Zhang, and Schölkopf, 2015 infer the existence of confounder if exogeneity holds for neither directions. While this is reasonable, exogeneity might be invalidated because of the confounder, and causal relation might exist at the same time. On the other hand, it is noteworthy that some, though much less, work assumes dependence is purely due to confounders, and derives necessary condition (Chaves et al., 2014) or infers the latent causal structure (Kela et al., 2019). Under similar lines of reasoning, they would mistake the above case 1) and 3). Therefore, a possible solution would be to combine the two approaches, and we might know it is the mixed case if test for purely causal relation and test for purely confounding both fail.

Extend FCMs to Confounded Case

Perhaps this is the most obvious approach pointed out by current research. Ideally, it would be a remarkable contribution to make ANMs work under confounders. However, over the years, there is still only LiNGAM that can handle confounders. This fact possibly suggests that we should take an entirely different path from FCMs, which would be a great endeavor. Other types of constraint that work under confounders (see Peters, Janzing, and Schölkopf, 2017, Chapter 9) could be explored and possibly exploited. A more achievable goal might be to work mainly under linear SEM. First, we could relax the assumptions on noise. Some special cases of Gaussian noise could be considered (see e.g. Peters and Bühlmann, 2014, but without

confounders). And we might also consider non-additive noise. Second, we might extend the functional class to some extent, such as allowing GLMs with deliberately defined basis functions.

Follow the Path of Distribution Classification

The main difficulty is how to extend the method (e.g., Lopez-Paz et al., 2015) to multivariate case since the class number grows super-exponentially w.r.t variable number. A possible approach is to embed graph into RKHS (e.g. using graph kernel (Ghosh et al., 2018)), then exploit distribution regression methods (Szabó et al., 2016). Training data is another problem, since human labelling of causal structure involving hundreds of variables will be too expensive, if not impossible. To address it, we could resort to some data synthesis method. Another, perhaps more practical, research direction might be to introduce some recent advance of distribution learning into causal discovery and improve accuracy, efficiency and scalability. For example, it would be interesting to see if Bayesian learning (Law et al., 2018) could bring up something new, e.g. the integration of prior knowledge.

Leverage Implicit Generative Models

Confounders could be treated as hidden variables from which the observed distribution is generated. In Goudet et al., 2018, we have already seen that 1) the loss does not really penalize anti-causal learning, and 2) the hill-climbing-like procedure is separated into artificial phases and has no guarantee to reach the global optimum. For the former, we ask the question: how to design a loss from first principles regarding causality? For example, can we define a discrepancy metric, that could also take into account the complexity of conditional distributions? Using KME, we might combine distance of distributions (such as MMD) with a complexity metric (like in Chen et al., 2014). Another possible way is to explicitly penalize anti-causal learning, by integrate the result of causal detection, like in the 'Causal Regularization' (Janzing, 2019; Bahadori et al., 2017). For the latter, a research question would be how to design a coherent training procedure driving the discovery of underlying causal structure? Here we may try hierarchical implicit models Tran and Blei, 2018; Tran, Ranganath, and Blei, 2017, which is more powerful than deep generative models,

in that they are more scalable, and can place prior on parameters and quantify the uncertainty of causal relations. Combining graph structure learning into generative models is also a possible solution.

7.3 Prospects at the Intersection of Causality and Machine Learning

Here, we focus on the causality side of the intersection, that is, "machine learning for causality". Arguably, the current thesis lies more on this side, because the problems that we aim to solve are causal but not plainly predictive, although we rely heavily on recent advances in machine learning models.

Causal (effect) inference has been studied heavily in economy and particularly econometrics as the identification and estimation of causal effects. Here, we comment on the future of machine learning for causal inference from the econometrics angle, while we touch on a bit of history first.

At the beginning of this century, Leo Breiman, a statistician and pioneer machine learner, wrote his famous "two cultures" paper, stating that "[t]he statistical community has been committed to the almost exclusive use of data [generating] models". Things changed largely in less than 10 years, as indicated by the seminal machine learning textbook (Hastie et al., 2009) written from a statistical perspective, and, even taken for granted by most people of my generation, the statistics community had accepted the algorithmic modeling culture as named by Breiman, referring to machine learning.

Now, after another 10 years, around 2020, two leading econometricians claim the further merge of the two cultures in their field (Athey and Imbens, 2019; Imbens and Athey, 2021). The slower acceptance of machine learning in econometrics is again due to the more prevalence of data modeling culture. This is not without good reason: the goal of many econometrics studies is parameter estimation, and often structural and causal parameters, while machine learning often focuses on prediction of an output variable (Mullainathan and Spiess, 2017). For example, economists are interested in the evaluation of the impact of interventions on a covariate, thus they

dig into the causal structure, say $Y = \mu(X) + \epsilon$, and try to understand the parameter μ, with interest in its identification, consistency, and uncertainty quantification. On the other hand, machine learning methods, especially in supervised learning, would focus on predicting Y and are evaluated by the performance on this task. Although machine learning at times fits a function f, it is often not treated as the target of statistical parameter estimation. Instead, the strengths of machine learning, besides prediction accuracy, are mainly flexible models (e.g., NNs) and data-driven model selection (e.g., regularization and cross-validation).

Thus, the challenge for econometricians is to 1) find the right places in their models, where economic theories are silent (regarding the requirements for identification etc.), so that machine learning can help in selecting the correct functional forms, and 2) tame the bias introduced by machine learning (often due to both overfitting and regularization). Athey, Imbens, and Wager, 2018 and Chernozhukov et al., 2018, (DML) provide recent examples in this line. From the machine learning side, we notice that there are sub-fields that are less known to outsiders but in fact more relevant to causal inference; for example, probabilistic (Bayesian) learning (Murphy, 2022) excels at uncertainty quantification, and generative learning (Murphy, 2023, Part IV) focuses on data generating models. Our Intact-VAE lies exactly here–that is, *probabilistic generative learning*. Indeed, in machine learning community, this sub-field has attracted much attention to both identifiability (Roeder, Metz, and Kingma, 2021; Wang, Blei, and Cunningham, 2021; Reizinger et al., 2022) and uncertainty (Jesson et al., 2021; Seitzer et al., 2022).

A future direction is to combine the techniques from econometrics literature to analyze the consistency and convergence of probabilistic generative learning for causal inference. Take VAEs for example, due to their incorporation of NNs, general consistency and convergence analysis cannot be expected in the recent future. Although there are results on variational Bayes at large (Wang and Blei, 2019a; Zhang and Gao, 2020), it would be hard to apply them to VAEs because there are no explicit ways to separate the NNs as nuisance parameters. Nevertheless, econometrics literature, e.g., DML, provides the wanted separation. We would need to specify a causal setting where 1) (a certain type of) VAEs could invert the mechanism and identify

the true latent variables, and 2) the functional parameters in the encoder and decoder could be treated as nuisance parameters in the econometric models. The idea in Reizinger et al., 2022 would be a possible bridge; it roughly states that the ELBO of a plain VAE can be seen as encouraging a certain kind of orthogonalization that indicates the latent components influence the observations "independently". Many econometric methods obtain balancing weights by encouraging a variety of orthogonal properties, and this is the case for DML. A related direction is to obtain double robustness; Intact-VAE does not model the propensity score, and we need to design a new VAE architecture.

Causal discovery as a field is special; it is intrinsically causal while could be seen as a sub-field of machine learning due to its origin in Bayesian networks (Verma and Pearl, 1988) and the work of some computational-oriented philosophers (Glymour, Scheines, and Spirtes, 1987). However, if we approach it from a methodological instead of historical perspective, we would see it as an early effort on machine learning for causality–using machine learning techniques, mainly Bayesian structure learning, to find the causal structure among variables, and focusing much on identification. Indeed, there are recent trends in re-labeling "causal discovery" as, or merge it into, "causal inference", as mentioned in the **Notes on terminology** in Introduction.

In the technical side, the challenges of the problem is at two levels: 1) some learning problems required by causal discovery is hard; for example, see the nonexistence of general purpose conditional independence tests (Shah and Peters, 2020); and 2) the combinatorial nature itself makes score-based causal discovery HP-hard, even if we are given oracles of independence, inference, and information (Chickering, Heckerman, and Meek, 2004). This is why there come advances in new machine learning based approaches, other than the searching methods based on conditional independence constraints or penalized likelihood scores. For example, score[1] matching is an important method in machine learning to workaround intractable normalizing constants, and Rolland et al., 2022 recently use it for a novel purpose–to determine

[1]Note that, the "score" in "score matching" is $\nabla \log p(\boldsymbol{x})$, i.e., the gradient of the log density function, and do not confuse it with the "score" in "score-based searching methods", which is a penalized likelihood function encouraging the simplest causal structure.

the topological ordering of nodes in a causal graph. The method outperforms or matches state-of-the-art methods and is much faster, sometimes by 10 times. Regarding the combinatorial hardness, there is a line of work that casts the problem to a continuous optimization problem by providing a continuous constraint for acyclicity (Zheng et al., 2018) and theoretical properties such as convergence of this continuous optimization are also studied (Ng et al., 2022).

Broader perspective. Finally, we take a step back and look briefly at the intersection of causality and machine learning as a whole. First, in the above review and prospects, the "causality for machine learning" side is basically omitted. At least in Pearl's eyes, the data-centric thinking and data-fitting culture in machine learning are still too strong (Pearl, 2021). A symptom is a heavy reliance on (usually a few) benchmark datasets for performance evaluation, and this has caused some problems for causal inference applications. For example, on the IHDP dataset for evaluation of CATE estimation, certain kinds of algorithms could easily achieve good performance, but by exploiting artificial properties in the dataset that is not quite relevant in the real-world (Curth et al., 2021). While Pearl is certainly right in encouraging the machine learning community to have more patience to learn lessons from causality research, there is a bright prospect of intensive incorporation of causal ideas into machine learning, indicated by the manifesto by two prominent machine learning researchers (Schölkopf et al., 2021). Now, we even have an emerging sub-field called "causal machine learning" (Kaddour et al., 2022). In summary, for research at the intersection, from both machine learning and causality sides, Mullainathan and Spiess, 2017 state succinctly that the challenge is to "make sense of the estimated prediction function without making strong assumptions about the underlying true world." For scientists interested in understanding cause and effect, this opens the way from predictive power to a model of "the underlying true world." For machine learners, this gives good properties such as robustness, reusability, and interpretability (Schölkopf et al., 2021) and perhaps the way to artificial general intelligence (Pearl, 2018).

Appendix A

Full-page Figures

This Appendix contains full-page figures that would be hard to fit into the main text.

A.1 Additional Plots of Latent Recovery by Intact-VAE

See the next pages for full-page figures. Please refer to Sec. 4.4.1 for detailed explanations.

FIGURE A.1: Plots of recovered-true latent. Rows: first 10 nonlinear random models, columns: outcome noise level.

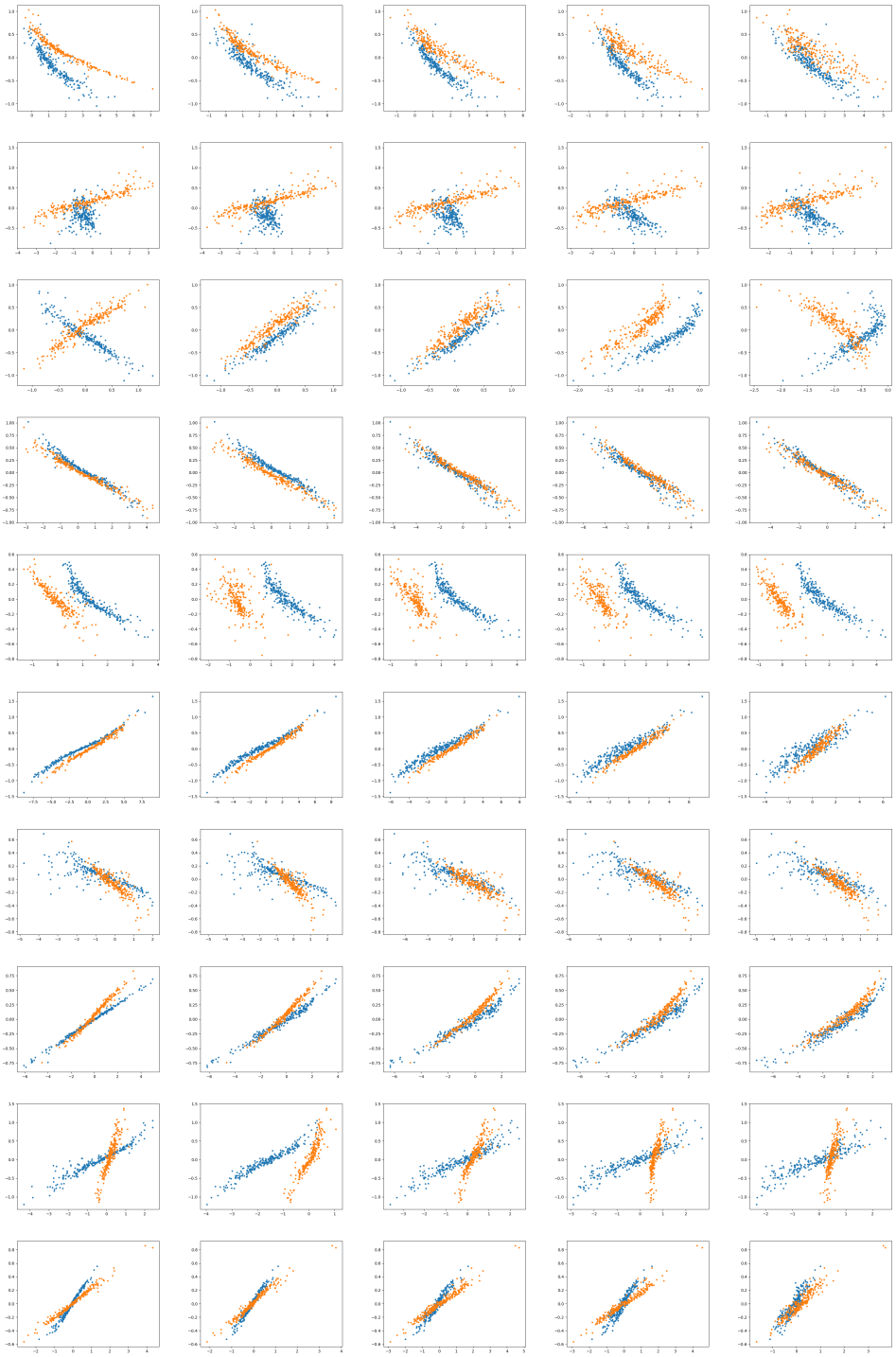

FIGURE A.2: Plots of recovered-true latent. Conditional prior *depends* on t. Rows: first 10 nonlinear random models, columns: outcome noise level. Compare to the previous figure, we can see the transformations for $t = 0, 1$ are *not* the same, confirming the importance of balanced prior.

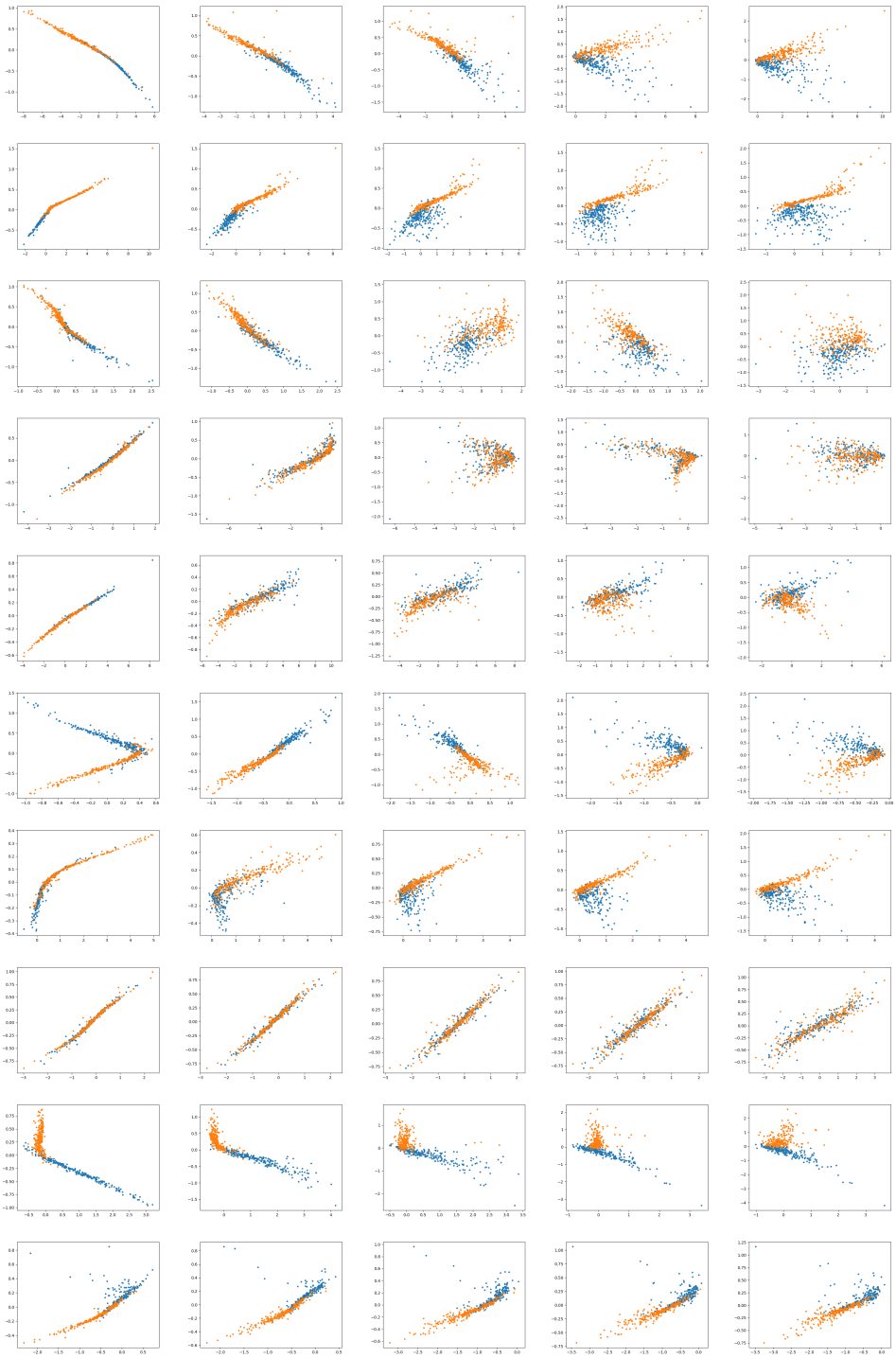

FIGURE A.3: Plots of recovered-true latent under *unobserved confounding*. Rows: first 10 nonlinear random models, columns: *proxy* noise level.

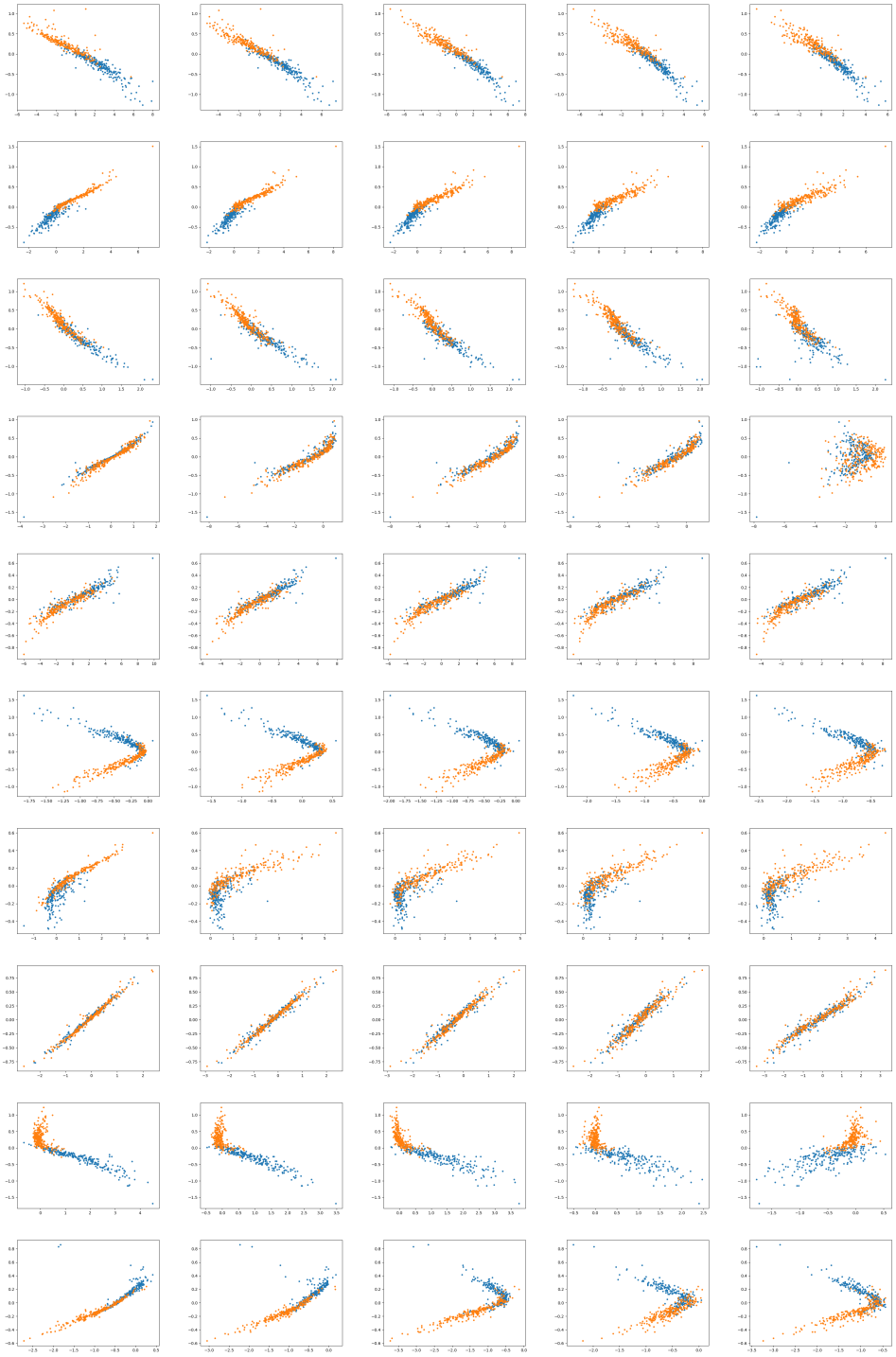

FIGURE A.4: Plots of recovered-true latent under *unobserved confounding*. Rows: first 10 nonlinear random models, columns: *outcome* noise level.

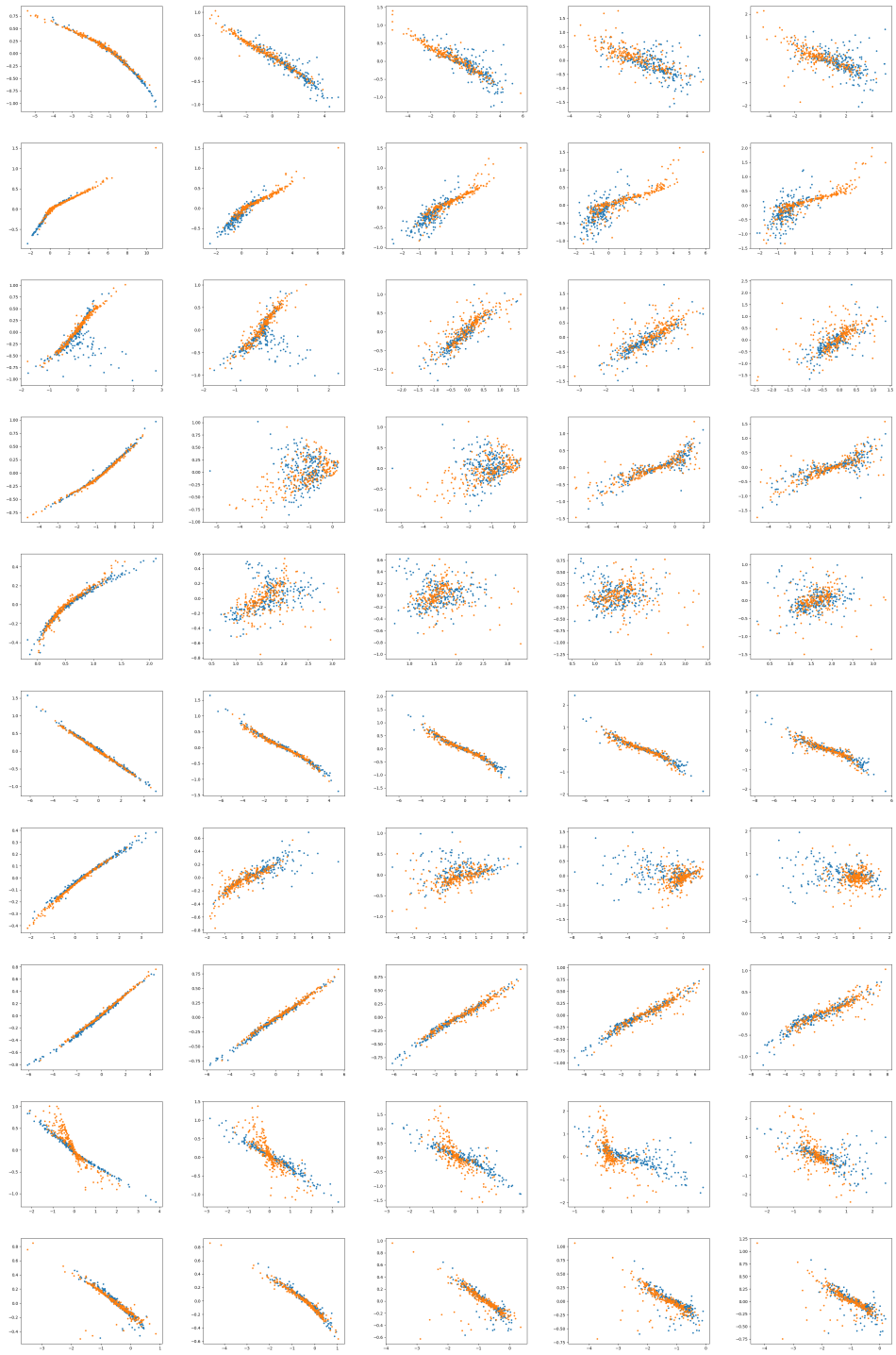

FIGURE A.5: Plots of recovered-true latent when *ignorability* holds. Rows: first 10 nonlinear random models, columns: *proxy* noise level.

FIGURE A.6: Plots of recovered-true latent when *ignorability* holds. Rows: first 10 nonlinear random models, columns: *outcome* noise level.

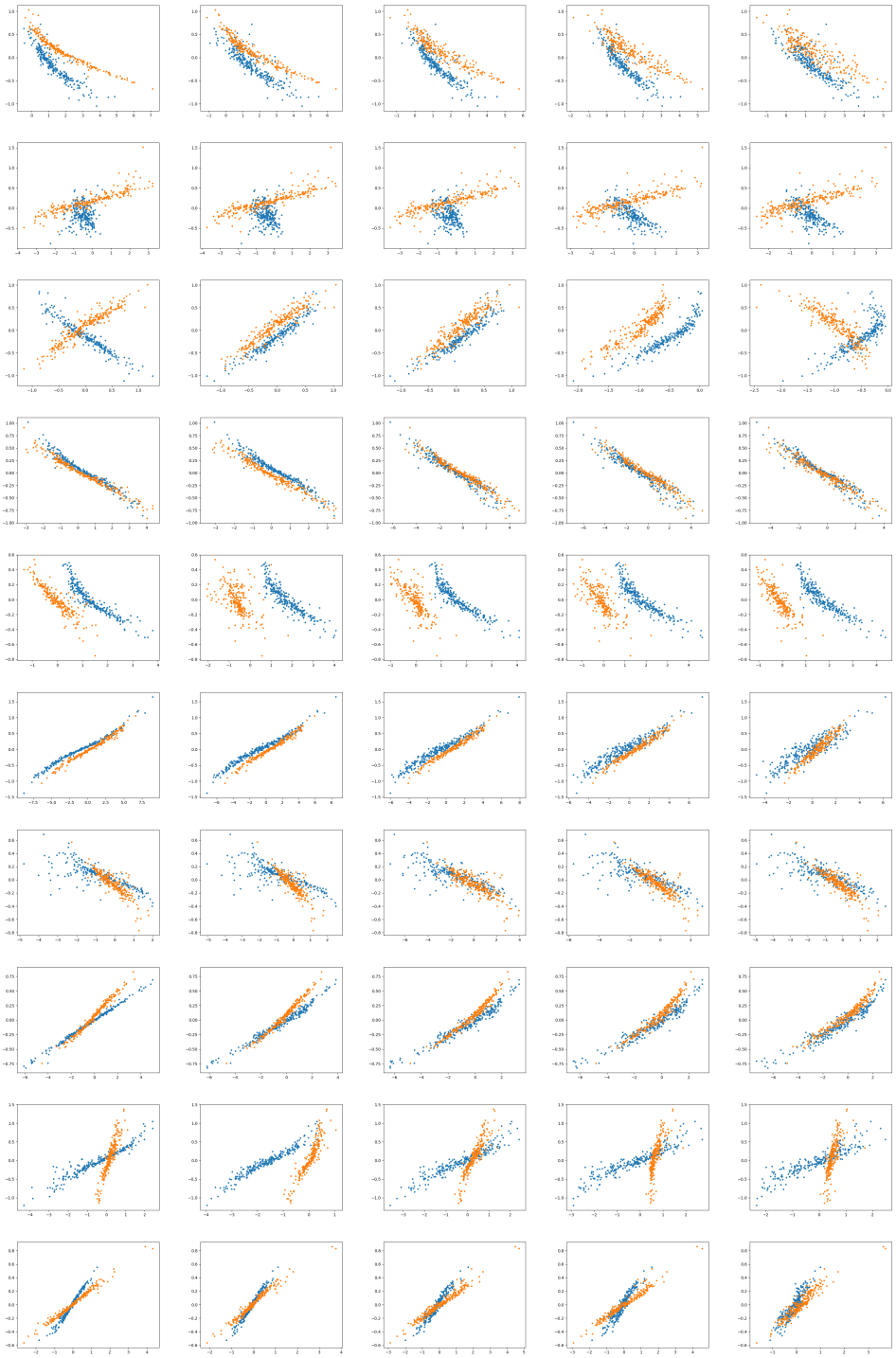

FIGURE A.7: Plots of recovered-true latent when *ignorability* holds. Conditional prior *depends* on t. Rows: first 10 nonlinear random models, columns: *outcome* noise level. Compare to the previous figure, we can see the transformations for $t = 0, 1$ are *not* the same.

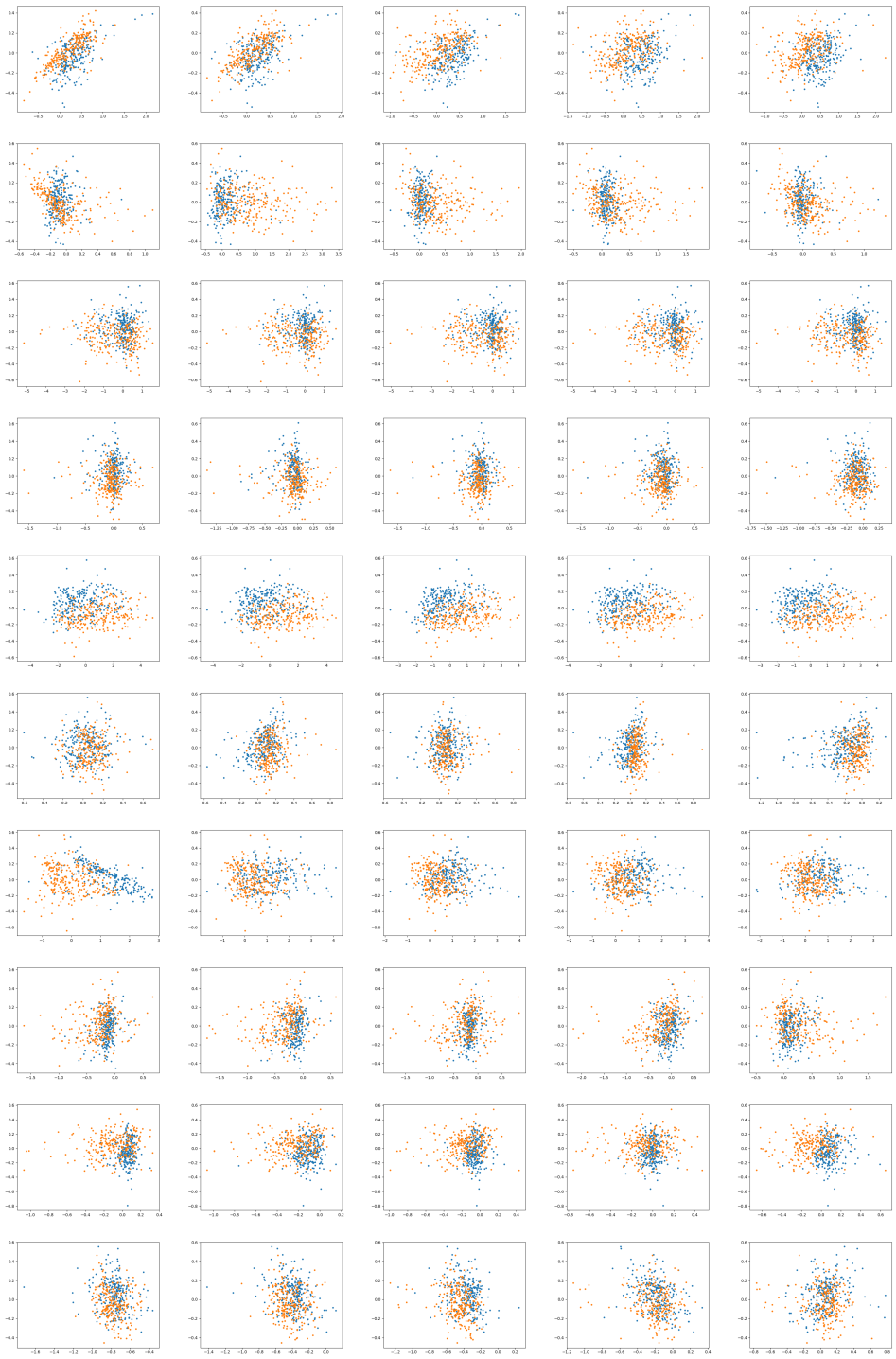

FIGURE A.8: Plots of recovered-true latent on *IVs*. Rows: first 10 nonlinear random models, columns: *outcome* noise level.

A.2 Empirical Validation of the Error Bound of Intact-VAE

See the next page for a full-page figure. Please refer to Sec. 4.5.3 for detailed explanations.

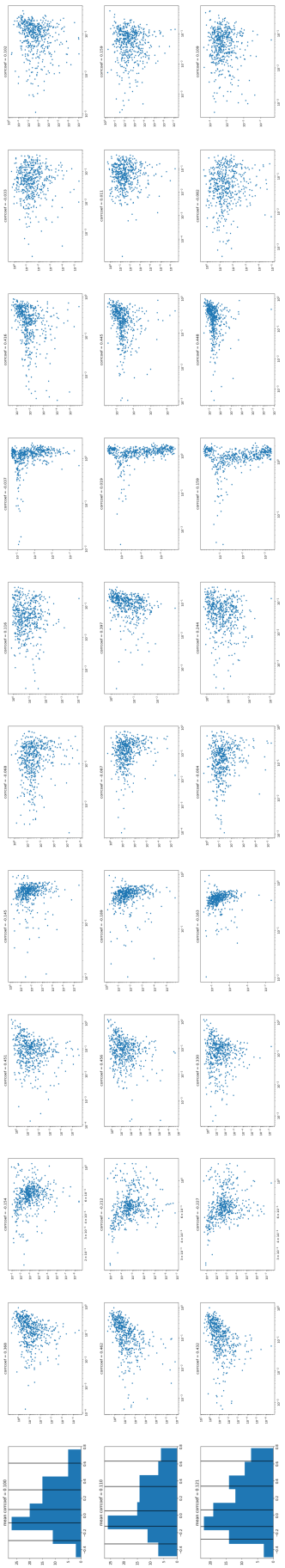

FIGURE A.9: Empirical validation of the error bound of Intact-VAE.

Appendix B

Old Lessons on Intact-VAE

Nowhere in the main text refers this section, so you can omit it if not interested. However, if reading, you may gain insight of how we came to our final theoretical formulation.

B.1 Identifiability of Representation (Is Not Enough)

Here we explain that the model identifiability given in Theorem 1 alone is, albeit interesting, not enough for estimation of TEs.

The importance of model identifiability can be seen clearly in the following corollary. That is, given $T = t$, the latent representation can be identified up to an invertible element-wise affine transformation. It can be easily understood by noting that, with the small noise and the injective $\boldsymbol{f}$, the decoder degenerates to deterministic function and the latent representation $Z = \boldsymbol{f}^{-1}(Y)$.

Corollary 3. *In Theorem 1, let* $\boldsymbol{\sigma}_{Y,t} = \mathbf{0}$, *then* $Z = \mathcal{A}_t(Z')$.

The good news is that, all the possible latent representations in our model are equivalent if we consider their independence relationships with any random variables, because any two of them are related by an *invertible* mapping. However, the bad news is that, this holds only given $T = t$, while the definition of B/P-score involves both t.

Consider how the *recovered* Z' would be used. For a control group ($t = 0$) data point $(\boldsymbol{x}, y, 0)$, the real challenge under finite sample is to predict the counterfactual outcome $y(1)$. Taking the observation, the encoder will output a posterior sample point $\boldsymbol{z}_0' = \boldsymbol{f}_0'^{-1}(y) = \mathcal{A}_0^{-1}(\boldsymbol{z}_0)$ (with zero outcome noise, the encoder degenerates to

a delta function: $q(Z|\boldsymbol{x}, y, 0) = \delta(Z - \boldsymbol{f}_0'^{-1}(y)))$. Then, we should do *counterfactual inference*, using decoder with counterfactual assignment $t = 1$: $y_1' = \boldsymbol{f}_1'(\boldsymbol{z}_0') = \boldsymbol{f}_1 \circ \mathcal{A}_1(\mathcal{A}_0^{-1}(\boldsymbol{z}_0))$. This prediction can be arbitrary far from the truth $y(1) = \boldsymbol{f}_1(\boldsymbol{z}_0)$, due to the difference between $\mathcal{A}_1$ and $\mathcal{A}_0$. More concretely, this is because when learning the decoder, only the posterior sample of the treatment group ($t = 1$) is fed to $\boldsymbol{f}_1'$, and the posterior sample is different to the true value by the affine transformation $\mathcal{A}_1$, while it is $\mathcal{A}_0$ for $\boldsymbol{z}_0'$.

Now we know what we need: $\mathcal{A}_0 = \mathcal{A}_1$ so that the equivalence of independence holds unconditionally; and, there exists at least one representation that is indeed a B-score. Then, *any* representation in our model will be a B-score. These indeed are what we have in Anonymous, 2021.

Proof of Corollary 1. In this proof, all equations and variables should condition on t, and we omit the conditioning in notation for convenience.

When $\boldsymbol{\sigma}_Y = \mathbf{0}$, the decoder degenerates to a delta function: $p(Y|Z) = \delta(Y - \boldsymbol{f}(Z))$, we have $Y = \boldsymbol{f}(Z)$ and $Y' = \boldsymbol{f}'(Z')$. For any $\boldsymbol{y}$ in the common support of Y, Y', there exist a *unique* $\boldsymbol{z}$ and a *unique* $\boldsymbol{z}'$ satisfy $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}'(\boldsymbol{z}')$ (use injectivity). Substitute $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z})$ into the l.h.s of (4.4), and $\boldsymbol{y} = \boldsymbol{f}'(\boldsymbol{z}')$ into the r.h.s, so we get $Z = \mathcal{A}(Z')$. The result for $\boldsymbol{f}$ follows. □

A technical detail is that, $\boldsymbol{z}, \boldsymbol{z}'$ might not always be related by $\mathcal{A}$, because we used the *common* support of Y, Y' in the proof. Thus, the relation holds for partial supports of Z, Z' correspond to the common support of Y, Y'. This problem disappears if we have the a consistent learning method (see Proposition 6).

B.2 Balancing Covariate and its Two Special Cases

Here we demonstrate part of our old, limited, theoretical formulation, and extract some insights from it.

The following definition was used in the old theory. The importance of this definition is immediate from the definition of balancing score, that is, if a balancing *covariate* is also a function of V, then it is a balancing *score*.

Definition 8 (Balancing covariate)**.** Random variable X is a *balancing covariate* of random variable V if $T \perp\!\!\!\perp V|X$. We also simply say X is *balancing* (or *non*-balancing if it does not satisfy this definition).

Given that a balancing score of the true (hidden or not) confounder is sufficient for weak ignorability, a natural and interesting question is that, does a balancing covariate of the true confounder also satisfies weak ignorability? The answer is *no*. To see why, we give the next Proposition indicating that a balancing covariate of the true confounder might *not* satisfy *exchangeability*.

Proposition 8. *Let X be a balancing covariate of V. If V satisfies exchangeability* and $Y(t) \perp\!\!\!\perp X|V, T$, *then so does X.*

The proof will use the properties of conditional independence (Proposition 1).

Proof. Let $W := Y(t)$ for convenience. We first write our assumptions in conditional independence, as *A1*. $T \perp\!\!\!\perp V|X$ (balancing covariate), *A2*. $W \perp\!\!\!\perp T|V$ (exchangeability given V), and *A3*. $W \perp\!\!\!\perp X|V, T$.

Now, from *A2* and *A3*, using contraction, we have $W \perp\!\!\!\perp X, T|V$, then using weak union, we have $W \perp\!\!\!\perp T|X, V$. From this last independence and *A1*, using contraction, we have $T \perp\!\!\!\perp V, W|X$. Then $T \perp\!\!\!\perp W|X$ follows by decomposition. □

Given this proposition, we know assumptions

$$\begin{aligned} &\text{i) } Y(t) \perp\!\!\!\perp T|V \text{ (exchangeability given } V), \\ &\text{ii) } T \perp\!\!\!\perp V|X \text{ (X is a balancing covariate of V), and} \\ &\text{iii) } Y \perp\!\!\!\perp X|V, T \end{aligned} \tag{B.1}$$

do not imply exchangeability given X, thus seem to be reasonable. Note the independence $Y(t) \perp\!\!\!\perp X|V, T$ assumed in the above proposition implies, but is not implied by, $Y \perp\!\!\!\perp X|V, T$. This is because, in general, $Y(0) \perp\!\!\!\perp X|V, T = 1$ and $Y(1) \perp\!\!\!\perp X|V, T = 0$ do not hold.

The assumptions in (B.1) were assumed by our old theory, with V is hidden confounder U plus observed confounder X_c. And also note that, iii) is the independence shared by PGS.

We examine two important special cases of balancing covariate, which provide further evidence that balancing covariate does not make the problem trivial.

Definition 9 (Noiseless proxy)**.** Random variable X is a noiseless proxy of random variable V if V is a function of X ($V = \boldsymbol{\omega}(X)$).

Noiseless proxy is a special case of balancing covariate because if $X = \boldsymbol{x}$ is given, we know $\boldsymbol{v} = \boldsymbol{\omega}(\boldsymbol{x})$ and $\boldsymbol{\omega}$ is a deterministic function, then $p(V|X = \boldsymbol{x}) = p(V|X = \boldsymbol{x}, T) = \delta(V - \boldsymbol{\omega}(\boldsymbol{x}))$. Also note that, a noiseless proxy always has higher dimensionality than V, or at least the same.

Intuitively, if the value of X is given, there is no further uncertainty about $\boldsymbol{v}$, so the observation of $\boldsymbol{x}$ may work equally well to adjust for confounding. But, as we will see soon, a noiseless proxy of the true confounder does *not* satisfy positivity.

Definition 10 (Injective proxy)**.** Random variable X is an injective proxy of random variable V if X is an injective function of V ($X = \boldsymbol{\chi}(V)$, $\boldsymbol{\chi}$ is injective).

Injective proxy is again a special case of noiseless proxy, since, by injectivity, $V = \boldsymbol{\chi}^{-1}(X)$, i.e. V is also a function of X.

Under this very special case, that is, if X is an injective proxy of the true confounder V, we finally have X is a balancing score and satisfies weak ignorability, since X is a balancing covariate and a function of V. To see this in another way, let $f = e \circ \boldsymbol{\chi}^{-1}$ and $\boldsymbol{\beta} = \boldsymbol{\chi}$ in Proposition 3, then $f(X) = f(\boldsymbol{\beta}(V)) = e(V)$. By weak ignorability of X, (5.1) has a simpler counterpart $\mu_t(\boldsymbol{x}) = \mathbb{E}(Y(t)|X = \boldsymbol{x}) = \mathbb{E}(Y|X = \boldsymbol{x}, T = t)$. Thus, a naive regression of Y on (X, T) will give a valid estimator of CATE and ATE.

However, a noiseless but *non*-injective proxy is *not* a balancing score, in particular, positivity might *not* hold. Here, a naive regression will not do. This is exactly because $\boldsymbol{\omega}$ is non-injective, hence multiple values of X that cause non-overlapped supports of $p(T = t|X = \boldsymbol{x}), t = 0, 1$ might be mapped to the same value of V. An extreme example would be $T = \mathbb{I}(X > 0), Z = |X|$. We can see $p(T = t|X)$ are totally non-overlapped, but $\forall t, z \neq 0 : p(T = t|Z = z) = 1/2$.

So far, so good. In the end, what is the problem of balancing covariate? Here it is. *If* the we have the positivity of X ($p(T|X) > 0$ always), then, using the positivity

and balancing to get $p(\boldsymbol{u}|\boldsymbol{x}) = p(\boldsymbol{u}|\boldsymbol{x}, T = t)$ for all $\boldsymbol{x}$, we follow (5.1),

$$\begin{aligned}\mu_t(\boldsymbol{x}) &= \int (\int p(y|\boldsymbol{u}, \boldsymbol{x}, t) y dy) p(\boldsymbol{u}|\boldsymbol{x}) d\boldsymbol{u} \\ &= \int (\int p(y|\boldsymbol{u}, \boldsymbol{x}, t) y dy) p(\boldsymbol{u}|\boldsymbol{x}, T = t) d\boldsymbol{u} \\ &= \int (\int p(y, \boldsymbol{u}|\boldsymbol{x}, t) d\boldsymbol{u}) y dy = \mathbb{E}(Y|\boldsymbol{x}, t).\end{aligned} \tag{B.2}$$

Naive estimator just works! Thus, *if* X indeed was a balancing covariate of true confounder, we gave a better method than naive estimator only in the sense that it works without positivity of X. It seems what our old theory really addressed was lack of positivity, another important issue in causal inference (D'Amour et al., 2020), but not confounding.

There are several lessons learned from the old formulation. First, there may exist cases that exchangeability given X fails to hold even when positivity of X holds, but the naive estimator still works. This is related to the fact that the conditional independence based on which balancing score/covariate are defined is not necessary for identification. And we should be able to find weaker but still sufficient conditions for identification. Second, balancing covariate assumption in (B.1) is strong, though may not make a trivial problem. It basically means that X, only one of the observables, gives sufficient information for treatment assignment. This inspires us to consider both X, Y, as in the latent variables given by our posterior and encoder.

Bibliography

Abrevaya, Jason, Yu-Chin Hsu, and Robert P Lieli (2015). "Estimating conditional average treatment effects". In: *Journal of Business & Economic Statistics* 33.4, pp. 485–505.

Alaa, Ahmed M and Mihaela van der Schaar (2017). "Bayesian inference of individualized treatment effects using multi-task gaussian processes". In: *Advances in Neural Information Processing Systems*, pp. 3424–3432.

Allman, Elizabeth S, Catherine Matias, John A Rhodes, et al. (2009). "Identifiability of parameters in latent structure models with many observed variables". In: *The Annals of Statistics* 37.6A, pp. 3099–3132.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). "Identification of causal effects using instrumental variables". In: *Journal of the American statistical Association* 91.434, pp. 444–455.

Anonymous (2021). "\beta-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap". In: *arXiv preprint arXiv:2110.05225*.

Armstrong, Timothy B and Michal Kolesár (2021). "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness". In: *Econometrica* 89.3, pp. 1141–1177.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Athey, Susan and Guido W Imbens (2019). "Machine Learning Methods That Economists Should Know About". In: *Annual Review of Economics* 11, pp. 685–725.

Athey, Susan, Guido W Imbens, and Stefan Wager (2018). "Approximate residual balancing: debiased inference of average treatment effects in high dimensions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4, pp. 597–623.

Bahadori, Mohammad Taha et al. (2017). "Causal regularization". In: *arXiv preprint arXiv:1702.02604*.

Balke, Alexander and Judea Pearl (1994). "Probabilistic Evaluation of Counterfactual Queries". In: *AAAI*.

Blöbaum, Patrick et al. (2018). "Cause-effect inference by comparing regression errors". In: *International Conference on Artificial Intelligence and Statistics*, pp. 900–909.

Bollen, Kenneth A (1989). *Structural equations with latent variables*. New York John Wiley and Sons.

Bonhomme, Stéphane and Martin Weidner (2021). "Posterior average effects". In: *Journal of Business & Economic Statistics* just-accepted, pp. 1–38.

Budhathoki, Kailash and Jilles Vreeken (2017). "MDL for causal inference on discrete data". In: *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 751–756.

Cassel, Claes M, Carl E Särndal, and Jan H Wretman (1976). "Some results on generalized difference estimation and generalized regression estimation for finite populations". In: *Biometrika* 63.3, pp. 615–620.

Chalupka, Krzysztof, Frederick Eberhardt, and Pietro Perona (2016). "Estimating causal direction and confounding of two discrete variables". In: *arXiv preprint arXiv:1611.01504*.

Chaves, R et al. (2014). "Inferring latent structures via information inequalities". In: *30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*. AUAI Press, pp. 112–121.

Chen, Zhitang et al. (2014). "Causal discovery via reproducing kernel Hilbert space embeddings". In: *Neural computation* 26.7, pp. 1484–1517.

Chernozhukov, Victor and Christian Hansen (2013). "Quantile models with endogeneity". In: *Annu. Rev. Econ.* 5.1, pp. 57–81.

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning". In: *The Econometrics Journal* 21.1.

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

Chetverikov, Denis and Daniel Wilhelm (2017). "Nonparametric instrumental variable estimation under monotonicity". In: *Econometrica* 85.4, pp. 1303–1320.

Chickering, David Maxwell (2002). "Optimal Structure Identification With Greedy Search". In: *Journal of Machine Learning Research* 3, pp. 507–552.

Chickering, Max, David Heckerman, and Chris Meek (2004). "Large-sample learning of Bayesian networks is NP-hard". In: *Journal of Machine Learning Research* 5, pp. 1287–1330.

Claassen, Tom, Joris M Mooij, and Tom Heskes (2013). "Learning Sparse Causal Models is not NP-hard". In: *Uncertainty in Artificial Intelligence*, p. 172.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2015). "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289*.

Cole, Stephen R and Miguel A Hernán (2008). "Constructing inverse probability weights for marginal structural models". In: *American journal of epidemiology* 168.6, pp. 656–664.

Colombo, Diego et al. (2012). "Learning high-dimensional directed acyclic graphs with latent and selection variables". In: *The Annals of Statistics*, pp. 294–321.

Curth, Alicia et al. (2021). "Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

Dai, Wangzhi and Collin M Stultz (2020). "Quantifying Common Support between Multiple Treatment Groups Using a Contrastive-VAE". In: *Machine Learning for Health*. PMLR, pp. 41–52.

D'Amour, Alexander and Alexander Franks (2021). "Deconfounding Scores: Feature Representations for Causal Effect Estimation with Weak Overlap". In: *arXiv preprint arXiv:2104.05762*.

Dhir, Anish and Ciarán M Lee (2020). "Integrating overlapping datasets using bivariate causal discovery". In: *Thirty-Fourth AAAI conference on artificial intelligence*.

Doersch, Carl (2016). "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908*.

Drton, Mathias and Marloes H Maathuis (2017). "Structure learning in graphical modeling". In: *Annual Review of Statistics and Its Application* 4, pp. 365–393.

D'Amour, Alexander et al. (2020). "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics*.

Evans, R Scott (2016). "Electronic health records: then, now, and in the future". In: *Yearbook of medical informatics* 25.S 01, S48–S61.

Farrell, Max H (2015). "Robust inference on average treatment effects with possibly more covariates than observations". In: *Journal of Econometrics* 189.1, pp. 1–23.

Freyberger, Joachim and Joel L Horowitz (2015). "Identification and shape restrictions in nonparametric instrumental variables estimation". In: *Journal of Econometrics* 189.1, pp. 41–53.

Gan, Li and Qi Li (2016). "Efficiency of thin and thick markets". In: *Journal of Econometrics* 192.1, pp. 40–54.

Geffner, Hector, Rina Dechter, and Joseph Halpern, eds. (2022). *Probabilistic and Causal Inference: The Works of Judea Pearl*. Morgan & Claypool.

Ghosh, Swarnendu et al. (2018). "The journey of graph kernels through two decades". In: *Computer Science Review* 27, pp. 88–111.

Glymour, Clark, Richard Scheines, and Peter Spirtes (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Gopalan, Prem K and David M Blei (2013). "Efficient discovery of overlapping communities in massive networks". In: *Proceedings of the National Academy of Sciences* 110.36, pp. 14534–14539.

Goudet, Olivier et al. (2018). "Learning functional causal models with generative neural networks". In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, pp. 39–80.

Greenland, Sander (1980). "The effect of misclassification in the presence of covariates". In: *American journal of epidemiology* 112.4, pp. 564–569.

Greenland, Sander, Judea Pearl, and James M Robins (1999). "Causal diagrams for epidemiologic research". In: *Epidemiology*, pp. 37–48.

Gretton, Arthur et al. (2005). "Measuring statistical dependence with Hilbert-Schmidt norms". In: *International conference on algorithmic learning theory*. Springer, pp. 63–77.

Guo, Ruocheng et al. (2020). "A survey of learning causality with data: Problems and methods". In: *ACM Computing Surveys (CSUR)* 53.4, pp. 1–37.

Hajage, David et al. (2017). "Estimation of conditional and marginal odds ratios using the prognostic score". In: *Statistics in medicine* 36.4, pp. 687–716.

Hansen, Ben B (2008). "The prognostic analogue of the propensity score". In: *Biometrika* 95.2, pp. 481–488.

Hartford, Jason et al. (2017). "Deep IV: A flexible approach for counterfactual prediction". In: *International Conference on Machine Learning*, pp. 1414–1423.

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

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hernan, Miguel A. and James M. Robins (2020). *Causal Inference: What If*. 1st. CRC Press. 352 pp. ISBN: 978-1-4200-7616-5.

Higgins, Irina et al. (2017). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *5th International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hill, Jennifer L (2011). "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hong, Han, Michael P Leung, and Jessie Li (2020). "Inference on finite-population treatment effects under limited overlap". In: *The Econometrics Journal* 23.1, pp. 32–47.

Hoyer, Patrik O et al. (2008). "Estimation of causal effects using linear non-Gaussian causal models with hidden variables". In: *International Journal of Approximate Reasoning* 49.2, pp. 362–378.

Hoyer, Patrik O et al. (2009). "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems*, pp. 689–696.

Hu, Shoubo et al. (2018). "Causal inference and mechanism clustering of a mixture of additive noise models". In: *Advances in Neural Information Processing Systems*, pp. 5206–5216.

Huang, Ming-Yueh and Kwun Chuen Gary Chan (2017). "Joint sufficient dimension reduction and estimation of conditional and average treatment effects". In: *Biometrika* 104.3, pp. 583–596.

Huber, Martin and Kaspar Wüthrich (2018). "Local average and quantile treatment effects under endogeneity: a review". In: *Journal of Econometric Methods* 8.1.

Huo, Xiaoming and Gábor J Székely (2016). "Fast computing for distance covariance". In: *Technometrics* 58.4, pp. 435–447.

Hyvärinen, Aapo and Hiroshi Morioka (2016). "Unsupervised feature extraction by time-contrastive learning and nonlinear ICA". In: *Advances in Neural Information Processing Systems*, pp. 3765–3773.

Hyvärinen, Aapo, Hiroaki Sasaki, and Richard Turner (2019). "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868.

Hyvärinen, Aapo and Kun Zhang (2016). "Nonlinear Functional Causal Models for Distinguishing Cause from Effect". In: *Statistics and Causality: Methods for Applied Empirical Research*. John Wiley, pp. 185–201.

Imbens, Guido and Susan Athey (2021). "Breiman's two cultures: A perspective from econometrics". In: *Observational Studies* 7.1, pp. 127–133.

Imbens, Guido W (2020). "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics". In: *Journal of Economic Literature* 58.4, pp. 1129–79.

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

Janzing, Dominik (2019). "Causal regularization". In: *Advances in Neural Information Processing Systems* 32.

Janzing, Dominik and Bernhard Scholkopf (2010). "Causal inference using the algorithmic Markov condition". In: *IEEE Transactions on Information Theory* 56.10, pp. 5168–5194.

Janzing, Dominik et al. (2012). "Information-geometric approach to inferring causal directions". In: *Artificial Intelligence* 182, pp. 1–31.

Jesson, Andrew et al. (2020). "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models". In: *Advances in Neural Information Processing Systems* 33.

Jesson, Andrew et al. (2021). "Quantifying ignorance in individual-level causal-effect estimates under hidden confounding". In: *International Conference on Machine Learning*. PMLR, pp. 4829–4838.

Jiang, Wenxin and Martin A Tanner (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining". In: *The Annals of Statistics* 36.5, pp. 2207–2231.

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

Johansson, Fredrik D, David Sontag, and Rajesh Ranganath (2019). "Support and invertibility in domain-invariant representations". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 527–536.

Johansson, Fredrik D et al. (2020). "Generalization bounds and representation learning for estimation of potential outcomes and causal effects". In: *arXiv preprint arXiv:2001.07426*.

Kaddour, Jean et al. (2022). "Causal Machine Learning: A Survey and Open Problems". In: *arXiv preprint arXiv:2206.15475*.

Kalainathan, Diviyan and Olivier Goudet (2019). "Causal Discovery Toolbox: Uncover causal relationships in Python". In: *arXiv preprint arXiv:1903.02278*.

Kallus, Nathan, Xiaojie Mao, and Madeleine Udell (2018). "Causal inference with noisy and missing covariates via matrix factorization". In: *Advances in neural information processing systems*, pp. 6921–6932.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou (2019). "Interval estimation of individual-level causal effects under unobserved confounding". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2281–2290.

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

Kato, Kengo (2013). "Quasi-Bayesian analysis of nonparametric instrumental variables models". In: *The Annals of Statistics* 41.5, pp. 2359–2390.

Kela, Aditya et al. (2019). "Semidefinite tests for latent causal structures". In: *IEEE Transactions on Information Theory* 66.1, pp. 339–349.

Kelly, Yvonne et al. (2009). "Why does birthweight vary among ethnic groups in the UK? Findings from the Millennium Cohort Study". In: *Journal of public health* 31.1, pp. 131–137.

Khemakhem, Ilyes et al. (2020a). "ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA". In: *Advances in Neural Information Processing Systems* 33.

Khemakhem, Ilyes et al. (2020b). "Variational autoencoders and nonlinear ica: A unifying framework". In: *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217.

Kim, Jae-Young (2002). "Limited information likelihood and Bayesian analysis". In: *Journal of Econometrics* 107.1-2, pp. 175–193.

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

Kingma, Diederik P, Max Welling, et al. (2019). "An Introduction to Variational Autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4, pp. 307–392.

Kingma, Durk P et al. (2014). "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems*, pp. 3581–3589.

Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJU4ayYgl.

Kocaoglu, Murat et al. (2018). "CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJE-4xW0W.

Kuroki, Manabu and Judea Pearl (2014). "Measurement bias and effect restoration in causal inference". In: *Biometrika* 101.2, pp. 423–437.

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

Law, Ho Chung Leon et al. (2018). "Bayesian approaches to distribution regression". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1167–1176.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Lee, Brian K, Justin Lessler, and Elizabeth A Stuart (2011). "Weight trimming and propensity score weighting". In: *PloS one* 6.3, e18174.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.

Leskovec, Jure and Andrej Krevl (2014). *SNAP Datasets: Stanford large network dataset collection*.

Lewbel, Arthur (2019). "The identification zoo: Meanings of identification in econometrics". In: *Journal of Economic Literature* 57.4, pp. 835–903.

Li, Fan and Fan Li (2019). "Propensity score weighting for causal inference with multiple treatments". In: *The Annals of Applied Statistics* 13.4, pp. 2389–2415.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky (2018). "Balancing covariates via propensity score weighting". In: *Journal of the American Statistical Association* 113.521, pp. 390–400.

Li, Zheng, Guannan Liu, and Qi Li (2017). "Nonparametric Knn estimation with monotone constraints". In: *Econometric Reviews* 36.6-9, pp. 988–1006.

Liao, Yuan and Wenxin Jiang (2011). "Posterior consistency of nonparametric conditional moment restricted models". In: *The Annals of Statistics* 39.6, pp. 3003–3031.

Lopez-Paz, David et al. (2015). "Towards a learning theory of cause-effect inference". In: *International Conference on Machine Learning*, pp. 1452–1461.

Lopez-Paz, David et al. (2017). "Discovering causal signals in images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6979–6987.

Louizos, Christos et al. (2017). "Causal effect inference with deep latent-variable models". In: *Advances in Neural Information Processing Systems*, pp. 6446–6456.

Lu, Danni et al. (2020). "Reconsidering Generative Objectives For Counterfactual Reasoning". In: *Advances in Neural Information Processing Systems* 33.

Luo, Wei, Yeying Zhu, and Debashis Ghosh (2017). "On estimating regression-based causal effects using sufficient dimension reduction". In: *Biometrika* 104.1, pp. 51–65.

Manski, Charles F (2009). *Identification for prediction and decision*. Harvard University Press.

Mastouri, Afsaneh et al. (2021). "Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction". In: *ICML 2021: 38th International Conference on Machine Learning*, pp. 7512–7523.

Mathieu, Emile et al. (2019). "Disentangling disentanglement in variational autoencoders". In: *International Conference on Machine Learning*. PMLR, pp. 4402–4412.

Matzkin, Rosa L (2007). "Nonparametric identification". In: *Handbook of econometrics* 6, pp. 5307–5368.

Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen (2018). "Identifying causal effects with proxy variables of an unmeasured confounder". In: *Biometrika* 105.4, pp. 987–993.

Mitrovic, Jovana, Dino Sejdinovic, and Yee Whye Teh (2018). "Causal inference via kernel deviance measures". In: *Advances in Neural Information Processing Systems*, pp. 6986–6994.

Monti, Ricardo Pio, Kun Zhang, and Aapo Hyvärinen (2019). "Causal Discovery with General Non-Linear Relationships using Non-Linear ICA". In: *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019*, p. 45.

Mooij, Joris M et al. (2011). "On causal discovery with cyclic additive noise models". In: *Advances in neural information processing systems* 24.

Mooij, Joris M et al. (2016). "Distinguishing cause from effect using observational data: methods and benchmarks". In: *The Journal of Machine Learning Research* 17.1, pp. 1103–1204.

Morucci, Marco et al. (2020). "Adaptive hyper-box matching for interpretable individualized treatment effect estimation". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR, pp. 1089–1098.

Muandet, Krikamol et al. (2019). "Dual instrumental variable regression". In: *arXiv preprint arXiv:1910.12358*.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Murphy, Kevin P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press. URL: probml.ai.

— (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press. URL: probml.ai.

Nagasawa, Kenichi (2021). "Treatment Effect Estimation with Noisy Conditioning Variables". In: *arXiv preprint arXiv:1811.00667v3*.

Newey, Whitney K (1993). "Efficient estimation of models with conditional moment restrictions". In.

Ng, Ignavier et al. (2022). "On the convergence of continuous constrained optimization for structure learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 8176–8198.

Nogueira, Ana Rita et al. (2022). "Methods and tools for causal discovery and causal inference". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.2, e1449.

Oberst, Michael et al. (2020). "Characterization of overlap in observational studies". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 788–798.

Ogarrio, Juan Miguel, Peter Spirtes, and Joe Ramsey (2016). "A hybrid causal search algorithm for latent variable models". In: *Conference on probabilistic graphical models*. PMLR, pp. 368–379.

Pearl, J (1988). "Probabilistic Reasoning in Intelligent Systems; Network of Plausible Inference". In: *Morgan Kaufmann, 1988*.

Pearl, J (2001). "Direct and indirect effects". In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*. Morgan Kaufman, pp. 411–420.

Pearl, Judea (1993). "Graphical models, causality and intervention". In: *Statistical Science* 8.3, pp. 266–269.

— (2009). *Causality: models, reasoning and inference*. Cambridge University Press.

— (2018). "Theoretical impediments to machine learning with seven sparks from the causal revolution". In: *arXiv preprint arXiv:1801.04016*.

— (2021). "Radical empiricism and machine learning research". In: *Journal of Causal Inference* 9.1, pp. 78–82.

Pearl, Judea and Dana Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic books.

Pearl, Judea and Thomas Verma (1991). "A Theory of Inferred Causation". In: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*. KR'91. Cambridge, MA, USA: Morgan Kaufmann Publishers Inc., 441–452. ISBN: 1558601651.

Peters, Jonas and Peter Bühlmann (2014). "Identifiability of Gaussian structural equation models with equal error variances". In: *Biometrika* 101.1, pp. 219–228.

Peters, Jonas, Dominik Janzing, and Bernhard Scholkopf (2011). "Causal inference on discrete data using additive noise models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.12, pp. 2436–2450.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Peters, Jonas et al. (2014). "Causal Discovery with Continuous Additive Noise Models". In: *Journal of Machine Learning Research* 15, pp. 2009–2053.

Puli, Aahlad and Rajesh Ranganath (2020). "General Control Functions for Causal Effect Estimation from Instrumental Variables". In: *Advances in neural information processing systems* 33, p. 8440.

Ramsey, James Bernard (1969). "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 31.2, pp. 350–371.

Reizinger, Patrik et al. (2022). "Embrace the Gap: VAEs Perform Independent Mechanism Analysis". In: *arXiv preprint arXiv:2206.02416*.

Rissanen, Severi and Pekka Marttinen (2021). "A Critical Look At The Identifiability of Causal Effects with Deep Latent Variable Models". In: *NeurIPS 2021, to appear*.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American statistical Association* 89.427, pp. 846–866.

Roeder, Geoffrey, Luke Metz, and Diederik P Kingma (2020). "On Linear Identifiability of Learned Representations". In: *arXiv preprint arXiv:2007.00810*.

Roeder, Geoffrey, Luke Metz, and Durk Kingma (2021). "On linear identifiability of learned representations". In: *International Conference on Machine Learning*. PMLR, pp. 9030–9039.

Rolland, Paul et al. (2022). "Score matching enables causal discovery of nonlinear additive noise models". In: *International Conference on Machine Learning*. PMLR, pp. 18741–18753.

Rosenbaum, Paul R (1987). "Model-based direct adjustment". In: *Journal of the American statistical Association* 82.398, pp. 387–394.

— (1991). "A characterization of optimal designs for observational studies". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3, pp. 597–610.

— (2020). "Modern algorithms for matching in observational studies". In: *Annual Review of Statistics and Its Application* 7, pp. 143–176.

Rosenbaum, Paul R, Richard N Ross, and Jeffrey H Silber (2007). "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer". In: *Journal of the American Statistical Association* 102.477, pp. 75–83.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

Rothenhäusler, Dominik, Peter Bühlmann, and Nicolai Meinshausen (2019). "Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions". In: *The Annals of Statistics* 47.3, pp. 1688–1722.

Rothenhäusler, Dominik et al. (2015). "BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions". In: *Advances in Neural Information Processing Systems* 28.

Rubin, Donald B (1979). "Using multivariate matched sampling and regression adjustment to control bias in observational studies". In: *Journal of the American Statistical Association* 74.366a, pp. 318–328.

— (2005). "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469, pp. 322–331.

Rumelhart, David E, James L McClelland, PDP Research Group, et al. (1988). *Parallel distributed processing*. Vol. 1. IEEE New York.

Schölkopf, Bernhard et al. (2021). "Toward Causal Representation Learning". In: *Proceedings of the IEEE*.

Seitzer, Maximilian et al. (2022). "On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks". In: *International Conference on Learning Representations*.

Sejdinovic, Dino et al. (2013). "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". In: *The Annals of Statistics* 41.5, pp. 2263–2291.

Shah, Rajen D and Jonas Peters (2020). "The hardness of conditional independence testing and the generalised covariance measure". In: *The Annals of Statistics* 48.3, pp. 1514–1538.

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

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

Shimizu, Shohei (2014). "LiNGAM: Non-Gaussian methods for estimating causal structures". In: *Behaviormetrika* 41.1, pp. 65–98.

Shimizu, Shohei et al. (2006). "A linear non-Gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* 7.Oct, pp. 2003–2030.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2019). "Kernel instrumental variable regression". In: *arXiv preprint arXiv:1906.00232*.

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

Sorrenson, Peter, Carsten Rother, and Ullrich Köthe (2019). "Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN)". In: *International Conference on Learning Representations*.

Spirtes, Peter and Clark Glymour (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9.1, pp. 62–72.

Spirtes, Peter, Christopher Meek, and Thomas Richardson (1999). "An algorithm for causal inference in the presence of latent variables and selection bias". In: *Computation, causation, and discovery* 21, pp. 211–252.

Spirtes, Peter and Kun Zhang (2016). "Causal discovery and inference: concepts and recent methodological advances". In: *Applied informatics*. Vol. 3. 1. SpringerOpen, p. 3.

Spirtes, Peter et al. (2000). *Causation, prediction, and search*. MIT press.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Starling, Jennifer E et al. (2019). "Monotone function estimation in the presence of extreme data coarsening: Analysis of preeclampsia and birth weight in urban Uganda". In: *arXiv preprint arXiv:1912.06946*.

Stegle, Oliver et al. (2010). "Probabilistic latent variable models for distinguishing between cause and effect". In: *Advances in neural information processing systems* 23.

Stuart, Elizabeth A. (2010). "Matching Methods for Causal Inference: A Review and a Look Forward". In: *Statistical Science* 25.1, pp. 1 –21. DOI: 10.1214/09-STS313. URL: https://doi.org/10.1214/09-STS313.

Sun, Xinwei et al. (2020). "Latent Causal Invariant Model". In: *arXiv preprint arXiv:2011.02203*.

Suter, Raphael et al. (2019). "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness". In: *International Conference on Machine Learning*. PMLR, pp. 6056–6065.

Szabó, Zoltán et al. (2016). "Learning theory for distribution regression". In: *The Journal of Machine Learning Research* 17.1, pp. 5272–5311.

Székely, Gábor J, Maria L Rizzo, Nail K Bakirov, et al. (2007). "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6, pp. 2769–2794.

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

Tchetgen, Eric J Tchetgen et al. (2020). "An Introduction to Proximal Causal Learning". In: *arXiv preprint arXiv:2009.10982*.

Tien, Christian (2021). "Instrumental Common Confounding". In: URL: https://www.christiantien.com/publication/preprint/preprint.pdf.

Tran, Dustin and David M. Blei (2018). "Implicit Causal Models for Genome-wide Association Studies". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SyELrEeAb.

Tran, Dustin, Rajesh Ranganath, and David Blei (2017). "Hierarchical implicit models and likelihood-free variational inference". In: *Advances in Neural Information Processing Systems* 30.

Vapnik, Vladimir (1999). *The nature of statistical learning theory*. Springer science & business media.

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

Verma, T. and J. Pearl (1988). "Causal networks: Semantics and expressiveness". In: *Proc. Workshop on Uncertainty in Artificial Intelligence (UAI-88)*, pp. 352–359.

Verma, Thomas and Judea Pearl (1990). "Equivalence and Synthesis of Causal Models". In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. UAI '90. USA: Elsevier Science Inc., 255–270. ISBN: 0444892648.

Vowels, Matthew James, Necati Cihan Camgoz, and Richard Bowden (2020). "Targeted VAE: Structured Inference and Targeted Learning for Causal Parameter Estimation". In: *arXiv preprint arXiv:2009.13472*.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wang, Shanshan et al. (2020). "Changing trends of birth weight with maternal age: a cross-sectional study in Xi'an city of Northwestern China". In: *BMC Pregnancy and Childbirth* 20.1, pp. 1–8.

Wang, Yixin, David Blei, and John P Cunningham (2021). "Posterior collapse and latent variable non-identifiability". In: *Advances in Neural Information Processing Systems* 34, pp. 5443–5455.

Wang, Yixin and David M Blei (2019a). "Frequentist consistency of variational Bayes". In: *Journal of the American Statistical Association* 114.527, pp. 1147–1161.

— (2019b). "The blessings of multiple causes". In: *Journal of the American Statistical Association* 114.528, pp. 1574–1596.

Wang, Ziyu et al. (2021). "Scalable Quasi-Bayesian Inference for Instrumental Variable Regression". In: *NeurIPS 2021, to appear*.

White, Halbert and Karim Chalak (2013). "Identification and identification failure for treatment effects using structural systems". In: *Econometric Reviews* 32.3, pp. 273–317.

Wooldridge, Jeffrey M (2015). "Control function methods in applied econometrics". In: *Journal of Human Resources* 50.2, pp. 420–445.

Wu, Pengzhou and Kenji Fukumizu (2020a). "Causal Mosaic: Cause-Effect Inference via Nonlinear ICA and Ensemble Method". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1157–1167. URL: http://proceedings.mlr.press/v108/wu20b.html.

— (2021). "Towards Principled Causal Effect Estimation by Deep Identifiable Models". In: *arXiv preprint arXiv:2109.15062*. arXiv: 2109.15062 [stat.ML].

Wu, Pengzhou Abel and Kenji Fukumizu (2020b). "Identifying Treatment Effects under Unobserved Confounding by Causal Representation Learning". In: *submitted to ICLR 2021*. URL: https://openreview.net/forum?id=D3TNqCspFpM.

Yang, Mengyue et al. (2020). "CausalVAE: Structured Causal Disentanglement in Variational Autoencoder". In: *arXiv preprint arXiv:2004.08697*.

Yang, S and P Ding (Mar. 2018). "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores". In: *Biometrika* 105.2, pp. 487–493. ISSN: 0006-3444. DOI: 10.1093/biomet/asy008. eprint: https://academic.oup.com/biomet/article-pdf/105/2/487/24821002/asy008_supp.pdf. URL: https://doi.org/10.1093/biomet/asy008.

Yao, Liuyi et al. (2018). "Representation learning for treatment effect estimation from observational data". In: *Advances in Neural Information Processing Systems*, pp. 2633–2643.

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar (2018). "GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ByKWUeWA-.

Zhang, Fengshuo and Chao Gao (2020). "Convergence rates of variational posterior distributions". In: *The Annals of Statistics* 48.4, pp. 2180–2207.

Zhang, K, J Zhang, and B Schölkopf (2015). "Distinguishing Cause from Effect Based on Exogeneity". In: *Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK), 2015*. Carnegie Mellon University, pp. 261–271.

Zhang, Kun and Aapo Hyvärinen (2009). "On the identifiability of the post-nonlinear causal model". In: *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, pp. 647–655.

Zhang, Qinyi et al. (2018). "Large-scale kernel methods for independence testing". In: *Statistics and Computing* 28.1, pp. 113–130.

Zhang, Tong (2006). "From epsilon-entropy to KL-entropy: Analysis of minimum information complexity density estimation". In: *The Annals of Statistics* 34.5, pp. 2180–2210.

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

Zhang, Yao, Alexis Bellot, and Mihaela Schaar (2020). "Learning overlapping representations for the estimation of individualized treatment effects". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1005–1014.

Zheng, Xun et al. (2018). "Dags with no tears: Continuous optimization for structure learning". In: *Advances in Neural Information Processing Systems* 31.