



An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies

Qiao Liu^{a,b} , Zhongren Chen^c , and Wing Hung Wong^{a,b,d,1}

Contributed by Wing H. Wong; received December 19, 2023; accepted April 25, 2024; reviewed by Peng Ding and Ker-Chau Li

In this article, we develop CausalEGM, a deep learning framework for nonlinear dimension reduction and generative modeling of the dependency among covariate features affecting treatment and response. CausalEGM can be used for estimating causal effects in both binary and continuous treatment settings. By learning a bidirectional transformation between the high-dimensional covariate space and a low-dimensional latent space and then modeling the dependencies of different subsets of the latent variables on the treatment and response, CausalEGM can extract the latent covariate features that affect both treatment and response. By conditioning on these features, one can mitigate the confounding effect of the high dimensional covariate on the estimation of the causal relation between treatment and response. In a series of experiments, the proposed method is shown to achieve superior performance over existing methods in both binary and continuous treatment settings. The improvement is substantial when the sample size is large and the covariate is of high dimension. Finally, we established excess risk bounds and consistency results for our method, and discuss how our approach is related to and improves upon other dimension reduction approaches in causal inference.

average dose–response function | average treatment effect | potential outcome | dimension reduction | deep generative models

Given data in an observational study, a central problem in causal inference is to estimate the effect of one variable (e.g., treatment) on another variable (e.g., outcome) in the presence of a covariate vector that represents all other variables observed in the study (1–3). Under the well-known “unconfoundedness” condition (4, 5), which assumes that there are no hidden confounding variables beyond the observed covariate vector, valid estimates of the desired effect of treatment on outcome can be obtained by alternative approaches, including matching, weighting, stratification, and regression-based methods (6). Covariate adjustment plays an important role in these methods (7, 8). A common goal in covariate adjustment is to obtain the average dose–response function, which often involves the estimation of the expectation of the outcome conditional on the treatment and the covariate. When the covariate is of high dimension, as is often the case in modern applications (9–11), covariate adjustment becomes difficult because of the “curse of dimensionality” (12).

Various types of dimension reduction approaches have been proposed to alleviate this difficulty. For example, a popular approach is to do adjustment or matching based on the propensity score (4, 13, 14), which is a one-dimensional feature (i.e., a scalar function) of the covariates that captures how the covariates affect the treatment. Of course, the propensity score function must first be learned from the observed data on the treatment and the covariates, which is usually done by logistic regression or other advanced machine learning methods (15). Another type of dimension reduction method is “sufficient dimension reduction” (16, 17) (SDR), which assumes that the treatment assignment is conditionally independent of potential outcomes given the low-dimensional projection of the covariates (18, 19). However, SDR-based causal inference approaches consider only linear dimension reduction which limits the applicability. Furthermore, the dimension reduction is performed separately for each treatment value, which makes it difficult to extend the method to the case when the range of the treatment variable is of high cardinality or is continuous.

The present work develops a covariate adjustment method based on an encoding generative modeling (EGM) approach, called CausalEGM, which simultaneously learns to i) embed the high-dimensional covariates into a low-dimensional latent space where the distribution of the embeddings (latent covariate features) is prespecified. ii) build generative models for treatment given latent features and for outcome given treatment

Significance

Causal inference has been increasingly essential in modern observational studies with rich covariate information. However, it is often challenging to estimate the causal effect with high-dimensional covariates. Here, we introduce an approach by encoding generative modeling (EGM) for handling high-dimensional covariates by a dependency-aware dimension reduction strategy where the key idea is to identify a latent covariate feature set (e.g., latent confounders) that affects both treatment and outcome. EGM provides a flexible and powerful framework for us to develop deep learning-based estimates for the structural equation modeling that describes the causal relations among variables. Comprehensive numerical experiments suggest that the proposed method is effective and scalable in estimating the causal effect of one variable on another under various settings.

Author contributions: Q.L. and W.H.W. designed research; Q.L., Z.C., and W.H.W. performed research; Q.L. and Z.C. analyzed the data; and Q.L., Z.C., and W.H.W. wrote the paper.

Reviewers: P.D., Department of Statistics, University of California Berkeley; and K.-C.L., University of California, Los Angeles.

The authors declare no competing interest.

Copyright © 2024 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹To whom correspondence may be addressed. Email: whwong@stanford.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2322376121/-/DCSupplemental>.

Published May 29, 2024.

and latent features. The key idea of this method is to partition the latent feature vector into different independent components that play different roles in the above two generative models. This partitioning then allows us to identify a minimal latent covariate feature subvector that affects both treatment and outcome. After presenting the model, we will explain how this can be viewed as an approach to constructing good latent covariate features for covariate adjustment. We will also discuss the difference between our method and alternative dimension reduction methods. In particular, the results of our study show that, by adding the generator function to reconstruct the covariate, our approach achieves better performance than approaches that only focus on extracting low-dimensional covariate features to use in the prediction of treatment and outcome.

In implementing our method, we use multilayer neural networks to represent the encoder and generator functions in the model, which allows us to leverage advances in generative AI in the learning of our model. There has been increasing attention on the use of machine learning in causal inference (20–22). However, most of these methods either learn a predictor of the outcome conditional on the joint state of treatment and covariate or build separate outcome prediction models for each fixed value of the treatment. As such, they are different in nature from the dimension reduction approach which tries to learn a low-dimensional covariate feature to replace the original covariate in the adjustment. We present numerical experiments to demonstrate that our dimension reduction approach will lead to better estimates of the causal effect. Finally, we study the theoretical properties of our approach and establish excess risk bounds and consistency results for our estimates.

Methods

Problem Setup. We are interested in the causal effect of a variable X on another variable Y in an observational study based on *i.i.d.* observations of $\{(X_i, Y_i, V_i) | i = 1, \dots, n\}$. X is usually called the treatment (or exposure) variable, and Y is called the response (or outcome) variable. $V \in \mathbb{R}^p$ represents the covariates in a p -dimensional space. Y is real-valued in the $\mathcal{Y}$ outcome space and $X \in \mathcal{X}$, where the support $\mathcal{X}$ is either a finite set or a bounded interval in $\mathbb{R}$.

To investigate causal effects, we aim to determine how the potential outcome will respond to the change of treatment, which is given by the function $Y(\cdot) : \mathcal{X} \rightarrow \mathbb{R}$. We are particularly interested in estimating the population average defined as

$$\mu(x) = \mathbb{E}[Y(x)], \quad [1]$$

which is known as the average dose–response function (ADRF). Note that we only observe the potential outcome indexed by the treatment variable. The random variable $Y(x)$ is not directly observable, and its expectation $\mu(x)$ is generally not identifiable from the joint distribution of the observed (X, Y, V) . Additional assumptions are needed for the identification of $\mu(x)$.

We first assume X and Y are related by the following two unknown equations:

$$\begin{cases} X = b_0(Z_x, U_1), \\ Y = f_0(X, Z_y, U_2), \end{cases} \quad [2]$$

where $U = (U_1, U_2)$ represents the set of all other unobserved variables that may affect X and Y , which include disturbances. $Z_x = t_x(V)$ and $Z_y = t_y(V)$ are two low-dimensional feature sets of V and $t_x(\cdot)$, $t_y(\cdot)$ are the corresponding transformation

functions. We denote $Z_0 = Z_x \cap Z_y$ as the intersection of the two feature sets. We will assume a modified version of the “unconfoundedness” condition.

Assumption 1. (Unconfoundedness) Conditional on the low-dimensional feature set Z_0 , the potential outcomes $Y(x)$ is independent of treatment variable X ,

$$X \perp\!\!\!\perp Y(x) | Z_0. \quad [3]$$

Since $Y(x)$ is a function of Z_y and U_2 , the above assumption is also equivalent to $X \perp\!\!\!\perp \{Z_y, U_2\} | Z_0$. Note that under the conventional unconfoundedness assumption, one must condition on the high-dimensional covariates V . Under our *Assumption 1*, it is sufficient to condition on a low-dimensional feature set of the covariates. Once Z_0 is given, there should be no unobserved confounding variables that drive correlated changes between the treatment and the outcome variables.

Under *Assumption 1*, it is shown that the ADRF is identifiable through the following equation (*SI Appendix, section A*),

$$\mu(x) = \int \mathbb{E}[Y | X = x, Z_0 = z_0] p_{Z_0}(z_0) dz_0. \quad [4]$$

The equation 4 shows that we can replace the original covariate V by a low-dimensional covariate feature $Z_0(V)$, the causal inference problem is transformed into the problem of learning a low-dimensional representation of V from the observational data. To learn this transformation function, we propose a EGM framework, that allows simultaneous learning of an encoder for the high-dimensional V and a generative model for (X, Y, V) . By imposing a suitable constraint on the generative model, one can ensure that certain subsets of the features computed by the encoder can be used as the low-dimensional feature Z_0 in the above condition. In the next subsection, we will illustrate how to use the neural networks to learn the low-dimensional features Z_0 and estimate the $\mu(x)$ in Eq. 4.

EGM. Our model is described in Fig. 1. To handle the high dimension of V , we embed V into a low-dimensional latent space using an encoder function $Z = E(V)$ and a generator/decoder function $V = G(Z)$ to map Z back to the original space. Note that dimension reduction with controllable latent features has been successfully applied in our previous works, including density estimation (23) and clustering (24).

In a standard autoencoder, G and E functions are learned by minimizing the reconstruction error between $G(E(V))$ and V over the observed sample of V . Here, $E(V)$ represents the low-dimensional latent features. However, it is important to consider the complex dependencies of covariates on treatment and outcome and the generation process of treatment and outcome. The proposed EGM framework simultaneously enables a dependency-aware dimension reduction and modeling the generation process of treatment and outcome. It is natural to suppose there are independent covariate feature sets A , B , and C with different roles. A is involved in the generation of both outcome and treatment, B is involved only in outcome generation, and C is involved only in the treatment generation. We can usually find invertible transformations of each to a separate standard multivariate normal vector, denoted as Z_0 , Z_1 , and Z_2 when V has a continuous distribution. We also demonstrated that EGM framework can handle the discrete covariates distribution well. In fact, any one-to-one transformation of A , B , and C can be used for covariates and the form of the generator functions of treatment

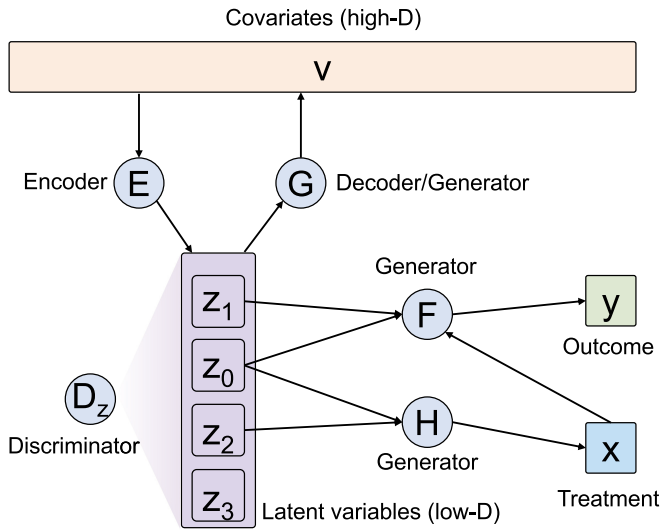


Fig. 1. The overview of CausalEGM model. Variables are in rectangles. Functions are in circles, with incoming arrows indicating inputs to the function and outgoing arrows indicating outputs of the function. Each function is modeled by a neural network. CausalEGM takes triplets of (X, Y, V) as input. E and G networks form a bidirectional transformation between the high-dimensional covariates space and a low-dimensional latent space. D_z network is used to constrain the distribution of latent features. F and H are used for generating/reconstructing the outcome and treatment variables, respectively. As G , F , and H networks take latent variables as input(s) with a desired distribution, they are also known as generators.

and outcome will change depending on which transformation functions are used. However, all of them will lead to the same covariate adjustment because the conditional expectation of the outcome will remain the same.

To achieve the above goal, we impose a “distribution-matching” objective in addition to the reconstruction error. Specifically, we desire that the distribution of $Z = E(V)$ should match a prespecified distribution, which is set to be a standard normal distribution. By the encoding process, the high-dimensional covariates with unknown distribution will be mapped to a low-dimensional latent space with a desired distribution. More importantly, we partition the latent feature vector into different subvectors that play different roles in the generative models for treatment and outcome. This partitioning enables us to identify a minimal covariate feature (e.g., Z_0) that affects both treatment and outcome. In addition to Z_0 , Z_1 , and Z_2 , we find that adding a flexible Z_3 that affects neither treatment nor outcome is useful to improve the learning of the confounding features and the generative models for treatment and outcome. We use deep neural networks to represent the functions $E(\cdot)$ and $G(\cdot)$. We utilized generative adversarial networks (GANs) (25) for distribution match where an adversarial loss (i.e., maximizing the discrimination power between the generated and observed data) is introduced.

Note that the learning of $E(\cdot)$ and $G(\cdot)$ should not be based on V alone. Rather, they must be coupled with the learning of generative models for X and Y , which are the variables of interest in causal inference. To do this, we assume that the $Z = E(V)$ can be partitioned into different subvectors that have different roles in the generators for X and Y . Specifically, $Z = (Z_0, Z_1, Z_2, Z_3)$, $Y = F(X, Z_0, Z_1) + \epsilon_1$ and $X = H(Z_0, Z_2) + \epsilon_2$. For clarity, the additive independent noises ϵ_1 and ϵ_2 are omitted in Fig. 1. Conceptually, Z_0 represents the latent covariate features that affect both treatment and outcome (e.g., confounding features), Z_1 represents the latent features that affect only the outcome,

Z_2 represents the latent features that affect only the treatment, and Z_3 represents the remaining latent features that are also important for the representation of V . By partitioning the latent features Z into four different components, the encoder function decouples the complex dependencies of covariates on treatment and outcome variables in the low-dimensional latent space.

The training details are given in the next subsection for how to jointly learn (F, H, E, G) in an end-to-end fashion given the observational data. Assuming that the four functions (F, H, E, G) are learned, the latent features Z then can be easily extracted through the encoder function $E(\cdot)$. To estimate the average dose-response function $\mu(x)$, the simplest way is to fit a nonparametric regression to estimate the conditional expectation in Eq. 2 and then calculate the empirical expectation given the observational data. In practice, we can also use

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n F(X = x, Z_0 = z_0^{(i)}, Z_1 = z_1^{(i)}), \quad [5]$$

to estimate the ADRF where n is the sample size. In binary treatment settings, the counterfactual (CF) outcome for the i^{th} unit is estimated as

$$y_{CF}^{(i)} = F(X = 1 - x^{(i)}, Z_0 = z_0^{(i)}, Z_1 = z_1^{(i)}). \quad [6]$$

Model Training. The CausalEGM model consists of a bidirectional transformation module (E, G) and two additional networks (F, H) for reconstructing/generating outcome and treatment, respectively (Fig. 1). The bidirectional module is a combination of an autoencoder and a GAN model. In addition to the reconstruction error between $G(E(V))$ and V in the covariates space commonly required by the autoencoder, the encoder network E aims to transform the covariates into latent features, whose distribution matches the standard multivariate normal distribution. A discriminator D_z network tries to distinguish data sampled from the multivariate normal distribution (positives) from data generated by the E network (negatives) where (E, D_z) forms a GAN model. Similarly, it is optional to minimize reconstruction error between $E(G(Z))$ and Z in the latent space and introduce another discriminator D_v in the covariate space to form a GAN model (G, D_v) to match the empirical distribution of the covariates V and the reconstructed data by G . We use the Wasserstein GAN with gradient penalty (26) as the default architecture to improve the generation power and training stability.

Thus, the loss functions of the adversarial training for distribution matching in latent space are represented as

$$\begin{cases} \mathcal{L}_E = -\mathbb{E}_{v \sim \hat{p}(v)} [D_{z,-1}(E(v))], \\ \mathcal{L}_{D_z} = -\mathbb{E}_{z \sim p(z)} [D_{z,-1}(z)] + \mathbb{E}_{v \sim \hat{p}(v)} [D_{z,-1}(E(v))] \\ \quad + \lambda \mathbb{E}_{z \sim \bar{p}(z)} [(\nabla D_{z,-1}(z) - 1)^2], \end{cases} \quad [7]$$

where $p(z)$ and $\hat{p}(\cdot)$ denote the standard normal distribution and the empirical distribution, respectively. To make the output of discriminator D_z differentiable, we use $D_{z,-1}(\cdot)$ to denote the output before binarization, which is achieved by a sigmoid function. $\bar{p}(z)$ denotes the uniform sampling from the straight lines between the two points sampled from the standard normal distribution $p(z)$ and the empirical distribution of latent features $\hat{p}(z)$. The network E and D_z are competing with each other during the adversarial training until reaching a Nash equilibrium. λ is the gradient penalty coefficient, which is set to 10 in all experiments.

In addition to adversarial training for distribution match, the reconstruction losses for (X, Y, V) are denoted as

$$\begin{cases} \mathcal{L}_{rec}^x = \mathbb{E}_{x \sim \hat{p}(x), z_0 \sim \hat{p}_{e_0(V)}, z_2 \sim \hat{p}_{e_2(V)}} [(x - F(z_0, z_2))^2], \\ \mathcal{L}_{rec}^y = \mathbb{E}_{y \sim \hat{p}(y), z_0 \sim \hat{p}_{e_0(V)}, z_1 \sim \hat{p}_{e_1(V)}} [(y - H(z_0, z_1, x))^2], \\ \mathcal{L}_{rec}^v = \mathbb{E}_{v \sim \hat{p}(v)} [\|v - G(E(v))\|_2^2]. \end{cases} \quad [8]$$

where $\hat{p}_{e_k(V)}$ ($k = 0, 1, 2, 3$) is the empirical distribution of the k^{th} component of the encoder $e(V)$ and $\|\cdot\|_2^2$ denotes the squared l^2 -norm. The total loss for CausalEGM can be summarized into two parts $\mathcal{L}_{FHEG} = \mathcal{L}_E + \mathcal{L}_{rec}^x + \mathcal{L}_{rec}^y + \mathcal{L}_{rec}^v$ and $\mathcal{L}_{D_z}$, which correspond to the major four networks (F, H, E, G) and the auxiliary discriminator network D_z , respectively. To train the CausalEGM model in an end-to-end fashion, we iteratively update the parameters (weights) in one of (F, H, E, G) or D_z given the value of the other. Each iteration contains the following two steps. In the first step, a minibatch of data is randomly sampled and the parameters of discriminator D_z are updated by minimizing $\mathcal{L}(D_z)$ while fixing the parameters in (F, H, E, G) . In the second step, a minibatch of data is randomly sampled and the parameters of (F, H, E, G) are updated by minimizing $\mathcal{L}_{FHEG}$ while fixing the parameters in D_z .

Model Architecture. For all numerical examples below, we use fully connected layers for all networks. Specifically, the (F, H, E, G) networks contain five fully connected layers, and each layer has 64 hidden nodes. The D_z network contains three fully connected layers with 64, 32, and 8 hidden nodes, respectively. The leaky-ReLU activation function is deployed as a nonlinear transformation in each hidden layer. We use Sigmoid as the activation function in the last layer of H network when the treatment is binary. For continuous treatments, we do not use any nonlinear activation function in the last layer of H network. Batch normalization (27) is applied in discriminator networks. Adam optimizer (28) with initial learning rate as 2×10^{-4} is used. The model parameters were updated in a minibatch manner with batch size 32. The default number of training iterations is 30,000.

Theoretical Analysis

GAN background. Let P and Q be two probability measures and $\mathcal{A}$ be a class of measurable subsets of the space $\mathcal{U}$. Then define $d(P, Q; \mathcal{A}) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$. Note that the function $d(\cdot)$ defines a pseudodistance function between two probability measures. For example, if we let $\mathcal{B}$ be the Borel sets, $d(P, Q; \mathcal{B})$ would become the variation distance between P and Q . Suppose P and Q have densities p and q , we then have $d(P, Q; \mathcal{B}) = \frac{1}{2} \|p - q\|_{L_1}$.

Let $\mathcal{A}_M := \{A \in \mathcal{A} : \exists D \in \mathcal{D}_M \text{ s.t. } \forall u \in A, D(u) = 1\}$, where $D : \mathcal{U} \rightarrow \{0, 1\}$ indicates a discriminator (classifier) and $\mathcal{D}_M$ is the set of discriminators constructed by deep neural networks with complexity parameter M (M can represent the number of layers, numbers of hidden nodes, etc). In general, M increases when sample size n goes large. If Q_G is the probability measure induced by the generative model $G(Z)$, the adversarial training is then equivalent to minimizing the pseudodistance between the induced distribution and the empirical distribution

$$\inf_G \sup_{A \in \mathcal{A}_M} |Q_G(A) - \hat{P}(A)| = \inf_G d(Q_G, \hat{P}; \mathcal{A}_M), \quad [9]$$

where $\hat{P}(u)$ is the empirical distribution given by $\frac{1}{n} \sum_{i=1}^n \delta_{u_i}(u)$, where $\{u_i | i = 1, \dots, n\}$ is the observed data and δ_u is the Dirac measure.

Empirical Risk Minimization. The empirical risk terms are represented as

$$\begin{cases} L_1 = \mathbb{E}_n[(Y - f(X, e_0(V), e_1(V)))^2], \\ L_2 = \mathbb{E}_n[(X - h(e_0(V), e_2(V)))^2], \\ L_3 = \sup_{A \in \mathcal{A}_M} |P(A; Z^0) - \hat{P}(A; e(V))| = d(\hat{P}_{e(V)}, P_{Z^0}; \mathcal{A}_M), \\ L_4 = \mathbb{E}_n[\|V - g(e(V))\|_2^2], \end{cases} \quad [10]$$

where $\mathbb{E}_n$ denotes the empirical expectation based on the observed data with sample size n . P_{Z^0} is the probability measures of $Z^0 \sim N(\mathbf{0}, \mathbf{I})$ and $\hat{P}_{e(V)}$ is the empirical distribution of $e(V)$. The empirical risk is denoted as $R_{emp}(f, h, e, g) = L_1 + L_2 + L_3 + L_4$.

The corresponding true risk is $R^0(f, h, e, g) = R_1^0 + R_2^0 + R_3^0 + R_4^0$, where

$$\begin{cases} R_1^0 = \mathbb{E}_0[(Y - f(X, Z_0, Z_1))^2], \\ R_2^0 = \mathbb{E}_0[(X - h(Z_0, Z_2))^2], \\ R_3^0 = d(P_{e(V)}, P_{Z^0}; \mathcal{A}_M), \\ R_4^0 = \mathbb{E}_0[\|V - g(e(V))\|_2^2]. \end{cases} \quad [11]$$

where $\mathbb{E}_0$ stands for the population expectation with respect to the underlying distribution of the random variables and $P_{e(V)}$ is the probability measure induced by $e(V)$. We denote $\mathcal{F}_M$ and $\mathcal{D}_M$ as the function classes of the neural network generator/encoder and discriminator with complexity M , respectively. By minimizing empirical risk (MER), we obtain the empirical solution as

$$\hat{f}_{M,n}, \hat{h}_{M,n}, \hat{e}_{M,n}, \hat{g}_{M,n} = \arg \min_{f, h, e, g \in \mathcal{F}_M} R_{emp}(f, h, e, g). \quad [12]$$

Rademacher Complexity. We use Rademacher complexity to measure the richness of a function class *w.r.t.* a probability distribution. The empirical Rademacher complexity term is defined as

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{F \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i F(X_i) \right], \quad [13]$$

where σ_i is *i.i.d.* drawn from the Rademacher distribution with $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Excess Risk. We can now define the excess risk within the function class $\mathcal{F}_M$ as

$$R^0(\hat{f}_{M,n}, \hat{h}_{M,n}, \hat{e}_{M,n}, \hat{g}_{M,n}) - \inf_{f, h, e, g \in \mathcal{F}_M} R^0(f, h, e, g). \quad [14]$$

We then characterize the convergence rate of the excess risk in terms of the Rademacher complexity of the function classes from the empirical risk terms.

Theorem 1. (Bound of Excess Risk) Under our problem setup with bounded input domain, $\mathcal{F}_M$ and $\mathcal{D}_M$ are the classes of neural network encoder/generator and discriminator with complexity M used in our model, respectively. We also add the composition

functions of encoder and generator, such as $g \circ e$, to $\mathcal{F}_M$. Let L_λ be the λ -Lipschitz squared loss function used in the empirical risk terms. Assume that $\mathcal{F}_M$ and $\mathcal{D}_M$ are uniformly equi-continuous and b -uniformly bounded, respectively. $\mathcal{R}_n(\cdot)$ denotes the empirical Rademacher complexity. For any $\delta > 0$, we have

$$\begin{aligned} R^0(\hat{f}_{M,n}, \hat{h}_{M,n}, \hat{e}_{M,n}, \hat{g}_{M,n}) - \inf_{f, h, e, g \in \mathcal{F}_M} R^0(f, h, e, g) \\ \leq (8 + 4p)\mathcal{R}_n(L_\lambda \circ \mathcal{F}_M) + 4\mathcal{R}_n(\mathcal{D}_M) + \delta \end{aligned} \quad [15]$$

with probability at least $1 - 4e^{-\frac{n\delta^2}{32b^2(3+p)^2}}$,

where the operator $\circ$ denotes the function composition and p is the number of covariates. The detailed proof is given in (SI Appendix, section B). Theorem 1 gives a high probability bound of the excess risk, which involves the calculation of Rademacher complexity. The Rademacher complexity can be further upper bounded in terms of the complexity of the class, such as covering number or Vapnik–Chervonenkis dimension. See refs. 29 and 30 for details from the viewpoint of the empirical process. Bounding the Rademacher complexity of deep neural networks has also been widely explored. For example, Truong (31) provided an order $O(1/\sqrt{n})$ for bounding the Rademacher complexity of feedforward neural network with finite depth and width (see its theorem 5). Li et al. (32) further showed that given a λ -Lipschitz continuous loss function L_λ , then $\mathcal{R}_n(L_\lambda \circ \mathcal{F}_M)$ has the order of $\tilde{O}(\lambda\sqrt{DWr}/\sqrt{n})$, where $\tilde{O}(\cdot)$ represents the rate by ignoring logarithmic factors and D, W , and r represent the depth, width, and rank of weight matrices in the neural network, respectively (see its theorem 1). Since generators, encoder, and discriminator in CausalEGM are all fully connected neural networks, the excess risk from the left-hand side of Eq. 15 converges to zero almost surely as long as the Rademacher complexity terms are in the order of $o(1)$.

Consistency. Under an assumption on the encoder–decoder networks related to dimension reduction, we can show the consistency of our empirical solution in formula 12.

Assumption 2. There exists $\tilde{e}_3, \tilde{g}$, and $\delta > 0$ s.t.

$$(e_0^0, e_1^0, e_2^0, \tilde{e}_3) \stackrel{D}{=} Z^0. \quad [16]$$

For any function e and g , we have

$$\mathbb{E}_0[\|V - \tilde{g}((e_0^0, e_1^0, e_2^0, \tilde{e}_3)(V))\|_2^2] \leq \mathbb{E}_0[\|V - g((e)(V))\|_2^2] + \delta. \quad [17]$$

The left-hand side of the inequality in Eq. 17 denotes the reconstruction error with the “distribution match” constraint while the first term on the right-hand side of the inequality in Eq. 17 denotes the reconstruction error in a dimension reduction framework without any constrain. The constant delta is the “price” for adding the distribution match constraint in the latent space. This assumption is expected to hold with a small delta when the distribution of V satisfies a certain “dimension reduction” property. We provide a concrete example in SI Appendix, section C to demonstrate the rationale of this assumption.

Based on the above assumption, we can derive the consistency theorem as follows.

Theorem 2. (Consistency) Under the same setting as Theorem 1, suppose $\cup_M \mathcal{F}_M$ and $\cup_M \mathcal{D}_M$ are uniformly equi-continuous and

b -uniformly bounded, respectively. Let (f^*, h^*, e^*, g^*) be any limit point of MER solution $(\hat{f}, \hat{h}, \hat{e}, \hat{g})_n$ when $n \rightarrow \infty$. If Assumption 1,2 hold and the Rademacher complexity terms in Theorem 1 go to zero as the sample size n and the model complexity M increase, we have

$$\begin{aligned} \mathbb{E}_0[(f^0 - f^*)(X, Z_0, Z_1)]^2 + \mathbb{E}_0[(h^0 - h^*)(Z_0, Z_2)]^2 \\ + d(P_{Z^0}, P_{e^*(V)}; \mathcal{A}_M) \leq 2\delta. \end{aligned} \quad [18]$$

Theorem 2 suggests that if V can be encoded effectively s.t. the Assumption 2 is satisfied with $\delta \approx 0$, we would have approximately

$$f^* \approx f^0, h^* \approx h^0, e^*(V) \stackrel{D}{\approx} Z^0. \quad [19]$$

This holds for any limit points of $\{(\hat{f}, \hat{h}, \hat{e}, \hat{g})_n\}$. The detailed proof is given in (SI Appendix, section D).

Results

We performed a series of experiments to evaluate the performance of CausalEGM against state-of-the-art methods under different settings. In the continuous treatment setting, we test the performance of CausalEGM in learning the average dose–response function (ADRF). In the binary treatment setting, we aim to verify the ability of CausalEGM to estimate both the average treatment effect (ATE) and the individual treatment estimation (ITE).

Datasets. For the continuous treatment setting, four different datasets from previous publications (14, 33, 34) were used, including three simulated datasets and one semisynthetic dataset. The semisynthetic dataset was collected from 71,345 twins where the weight is the continuous treatment variable and we simulate the risk of death (outcome) under a model in which higher weight leads to a lower death rate in general.

For the binary treatment setting, we used the datasets from the 2018 Atlantic Causal Inference Conference (ACIC) competition. This dataset utilizes the linked births and infant deaths database (LBIDD) based on real-world medical measurements. The LBIDD data are semisynthetic, where 117 measured covariates are given, and the treatment and outcome are simulated based on different data-generating processes. We selected nine datasets by using the most complicated generation process (e.g., the highest degree of generation function) with sample sizes ranging from 1,000 to 50,000. The details of all datasets used in this paper were provided at (SI Appendix, section E).

Model Evaluation. In the continuous treatment setting, we aim to evaluate whether the estimated dose–response function $\mu(x)$ can well approximate the true dose–response function. Two commonly used metrics, including RMSE and mean absolute percentage error (MAPE) are used where

$$\begin{cases} RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu(x_i) - \hat{\mu}(x_i))^2}, \\ MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\mu(x_i) - \hat{\mu}(x_i)}{\mu(x_i)} \right|. \end{cases} \quad [20]$$

In the binary treatment setting, we use absolute error of average treatment effect (ϵ_{ATE}) and mean squared error of precision in

estimation of heterogeneous effect (ϵ_{PEHE}) for evaluating the performance where

$$\begin{cases} \epsilon_{ATE} = |\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i(1) - \hat{Y}_i(0)) - \frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0))|, \\ \epsilon_{PEHE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i(1) - \hat{Y}_i(0) - (Y_i(1) - Y_i(0)))^2. \end{cases} \tag{21}$$

Note that $\hat{Y}_i(\cdot)$ denotes the predicted/imputed value of potential outcome.

Baselines. For the continuous treatment setting, three different baselines, including ordinary least squares regression (OLS), regression prediction estimator (REG) (35, 36), and double debiased machine learning estimators (34) were used. For the binary treatment setting, we compared CausalEGM to neural network-based methods [CFR (37), Dragonnet (38), CEVAE (39) and GANITE (40)], tree-based methods CausalForest (41), and sufficient dimension reduction based method SDRcausal (42). A detailed introduction of these competing methods was provided in *SI Appendix, section F*.

Continuous Treatment Experiments. We first evaluate the performance of the CausalEGM model in the continuous setting where the treatment $x \in \mathcal{X}$ and $\mathcal{X}$ is a bounded interval in $\mathbb{R}$. We set the sample size and number of covariates to be 20,000 and 200 in the simulated datasets, respectively. The latent dimensions (dimension of z_i where $i = 0, 1, 2, 3$) of the four datasets are set to be (1, 1, 1, 7), (2, 2, 2, 4), (5, 5, 5, 5), and (1, 1, 1, 7), respectively. It is shown that CausalEGM demonstrates superior results over the existing methods, including two linear regression-based methods OLS and REG, and a kernel-based machine learning approach with two different machine

learning algorithms (lasso and neural network). We first evaluate whether the dose–response function can be well estimated by different competing methods (Fig. 2). It is observed that OLS and Reg result in relatively large estimation errors. The dose–response curves estimated by the DML methods have spikes and fluctuations. In contrast, the curves estimated by CausalEGM are smooth and the estimation errors are small.

In terms of the quantitative measurements, CausalEGM achieves the lowest RMSE and MAPE in all three simulated datasets compared to baseline methods (Table 1). We also note that DML method performs much better than linear regression-based methods (OLS and REG) in Hiranos and Imbens and Twins datasets while performing less well in the other two. CausalEGM reduces the RMSE, MAPE by 24.2% to 63.4%, 6.9% to 55.2% compared to the best baseline method across different datasets, respectively. The results of both simulated data and real data illustrate that CausalEGM offers significant improvement for estimating the causal effect in continuous settings.

Binary Treatment Experiments. In the binary treatment settings where the treatment $x \in \{0, 1\}$, we aim to evaluate whether CausalEGM could estimate an accurate treatment effect. The latent dimensions were set to be (3, 6, 3, 6). CausalEGM was benchmarked against a number of state-of-the-art methods on the ACIC 2018 benchmark datasets, which provide various simulation settings and sample sizes. We chose three datasets from each of three different sample sizes (1, 10, and 50 K) with the most complicated generation process (e.g., the generation functions are of the highest order/degree). CausalEGM is compared to six baseline methods on each of these datasets. As shown in Table 2, CausalEGM achieves the smallest ϵ_{ATE} in six out of nine datasets. CausalEGM performs especially well in datasets with large sample sizes (e.g., 50 K). For example, the ϵ_{ATE} is reduced by 16.7% to 98.7% in the three largest datasets compared to the

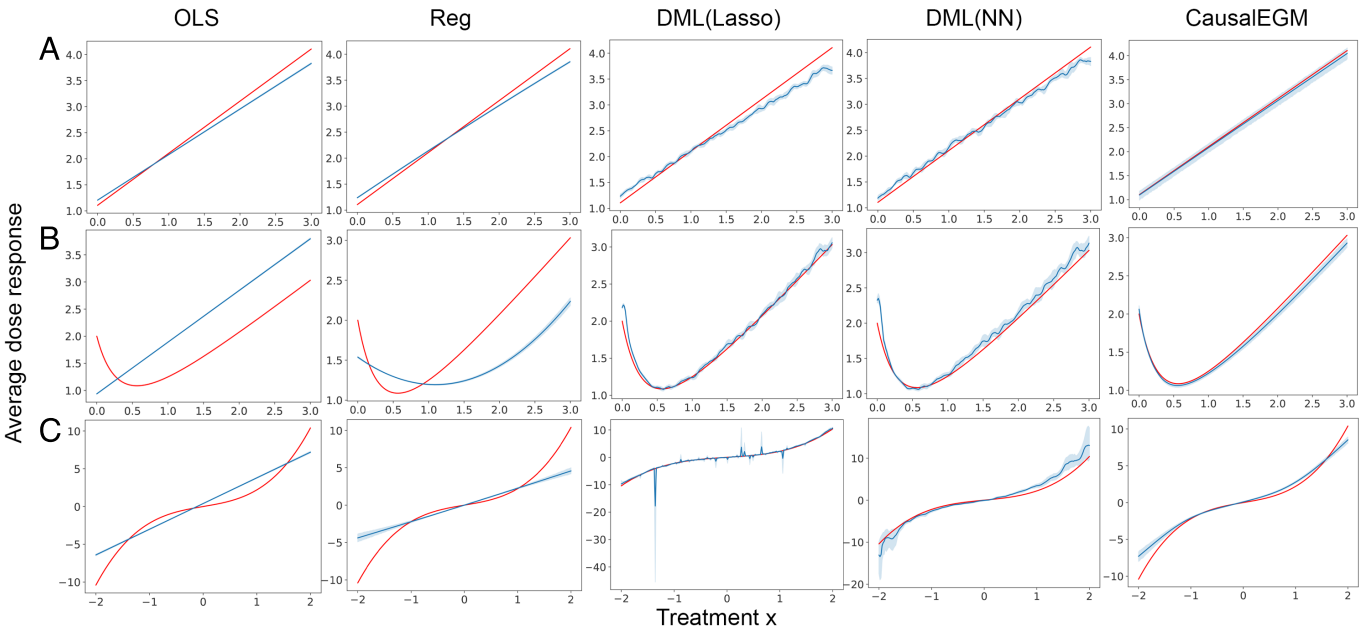


Fig. 2. The performance of CausalEGM and baseline methods (OLS, Reg, DML with Lasso or neural network) under continuous treatment settings across three benchmark datasets. (A) Hiranos and Imbens dataset. (B) Sun et al. dataset. (C) Colangelo and Lee dataset. The red curves are the ground truth while the blue curves are the estimated average dose–response with 95% CI based on 10 independent simulations.

Table 1. Result on continuous treatment setting

Dataset	Method	RMSE	MAPE
Imbens et al.	OLS	0.680 ± 0.0	0.367 ± 0.0
	REG	0.525 ± 0.0	0.214 ± 0.0
	DML(lasso)	0.090 ± 0.0	0.037 ± 0.0
	DML(nn)	0.133 ± 0.022	0.052 ± 0.011
	CausalEGM	0.041 ± 0.014	0.019 ± 0.006
Sun et al.	OLS	0.140 ± 0.0	0.041 ± 0.0
	REG	0.117 ± 0.0	0.039 ± 0.0
	DML(lasso)	0.163 ± 0.0	0.050 ± 0.0
	DML(nn)	0.0970 ± 0.0190	0.0346 ± 0.006
	CausalEGM	0.0738 ± 0.0399	0.0345 ± 0.0170
Lee et al.	OLS	1.3 ± 0.0	1.2 ± 0.0
	REG	1.5 ± 0.0	0.565 ± 0.0
	DML(lasso)	0.487 ± 0.0	0.168 ± 0.0
	DML(nn)	1.3 ± 0.581	0.494 ± 0.181
	CausalEGM	0.125 ± 0.040	0.119 ± 0.080
Twins	OLS	0.109 ± 0.0	0.260 ± 0.0
	REG	11 ± 0.0	64 ± 0.0
	DML(lasso)	0.075 ± 0.0	0.165 ± 0.0
	DML(nn)	0.059 ± 0.002	0.158 ± 0.006
	CausalEGM	0.0339 ± 0.020	0.090 ± 0.053

Each method was run for 10 times and the SD was also shown. The best performance is marked in bold.

second-best method. For another metric, CausalEGM achieves the smallest ϵ_{PEHE} in five out of nine datasets and the second-best performance in the remaining four datasets. To sum up, our model shows superior performance in estimating both average treatment effect and individual treatment effect and is especially powerful when the sample size is large.

Next, we evaluate whether the EGM framework can learn a better low-dimensional representation compared to sufficient dimension reduction (SDR). Note that all SDR-based methods for causal inference use the linear SDR which is restrictive and may not be able to capture the nonlinear relationship in complex datasets. We made a comprehensive comparison of CausalEGM with SDRcausal under experimental settings either satisfying or violating the SDR assumption. Note that SDRcausal implements

several different variants from the original paper (42) and we always choose the best result to report. CausalEGM shows great improvement over SDRcausal in both settings, especially in the nonlinear dataset where a linear SDR failed to work (*SI Appendix, section G*).

Impact of Discrete Covariates. CausalEGM has demonstrated superior empirical performance with discrete covariates, such as the ACIC 2018 where all 117 covariates are discrete. One natural question to ask is whether we can always construct standard normal variables in the latent space when some of the covariates are discrete.

From the theoretical perspective, under the condition that there are enough covariates that are independent of both treatment and response, we can construct an approximated standard normal distribution even when the confounders are discrete variables. For example, suppose covariates $V_1, V_2, \dots, V_p$ are *i.i.d.* binary variables with probability $\frac{1}{2}$ on +1 and -1. Assume V_1 is the true confounder that affects both treatment X and outcome Y while other covariates are not involved in the generative models for X and Y . Then the statistic $W = V_1 \sqrt{\frac{\sum_{i=2}^p V_i^2}{p-1}}$ will satisfy the following two conditions. 1) $W \sim N(0, 1)$ approximately when p is large. 2) W contains all the information in the confounder as V_1 can be recovered by taking the sign of W . We added this example in *SI Appendix, section H*.

From the empirical perspective, we conducted the following experiment to show that the proposed CausalEGM framework can learn to construct such a statistic approximately. In this experiment, 1) $V_1, V_2, \dots, V_p$ are *i.i.d.* binary variables distributed as above. 2) V_1, V_2 , and V_3 are used for generating treatment variable and V_1 . 3) V_4, V_5 are used for generating outcome variable. Thus V_1 is the true confounder and $V_{6:p}$ provides independent randomness that is not needed in the generative modeling of treatment and outcome. We set $p = 15$ and the dimension of the latent confounding variable Z_0 to be 1 (see details in *SI Appendix, section H*). Our numerical results in this setting showed that Z_0 learned by CausalEGM still follows a standard normal distribution approximately while also preserving most of the information in V_1 (Fig. 3). We conclude that

Table 2. The performance of CausalEGM and comparison methods in ACIC 2018 dataset with various sample sizes

Metric	Dataset	TARNET	CFRNET	CEVAE	GANITE	Dragonnet	CausalForest	CausalEGM
ϵ_{ATE}	Datasets-1k	0.022 ± 0.015	0.018 ± 0.015	0.035 ± 0.021	0.27 ± 0.08	0.010 ± 0.004	0.021 ± 0.001	0.0097 ± 0.0075
		0.038 ± 0.029	0.041 ± 0.027	0.12 ± 0.10	2.0 ± 0.3	0.012 ± 0.007	0.017 ± 0.003	0.032 ± 0.020
		0.10 ± 0.06	0.095 ± 0.079	0.38 ± 0.27	2.0 ± 1.4	0.16 ± 0.10	0.23 ± 0.02	0.26 ± 0.07
	Datasets-10k	6.4 ± 3.5	12 ± 7	204 ± 58	2.7 ± 1.2	124 ± 11	2.5 ± 1.1	1.3 ± 0.6
		0.056 ± 0.001	0.056 ± 0.001	0.070 ± 0.031	1.2 ± 0.2	0.0097 ± 0.069	0.0057 ± 0.0004	0.0043 ± 0.0025
		0.034 ± 0.023	0.060 ± 0.002	0.018 ± 0.011	0.12 ± 0.09	0.078 ± 0.057	0.013 ± 0.003	0.039 ± 0.016
	Datasets-50k	0.038 ± 0.021	0.085 ± 0.105	0.59 ± 0.31	1.4 ± 0.5	0.89 ± 0.53	0.024 ± 0.003	0.020 ± 0.013
		0.044 ± 0.003	0.045 ± 0.004	0.66 ± 0.59	2.3 ± 0.2	0.027 ± 0.028	0.010 ± 0.001	0.0098 ± 0.0089
		0.30 ± 0.01	0.30 ± 0.01	0.64 ± 0.45	1.9 ± 0.3	0.16 ± 0.08	0.12 ± 0.01	0.0016 ± 0.0010
ϵ_{PEHE}	Datasets-1k	0.11 ± 0.02	0.00069 ± 0.00075	0.012 ± 0.005	0.14 ± 0.04	0.038 ± 0.003	0.00080 ± 0.00005	0.0069 ± 0.0016
		0.35 ± 0.03	0.29 ± 0.04	0.27 ± 0.04	4.34 ± 1.24	0.34 ± 0.01	0.27 ± 0.01	0.25 ± 0.01
		0.31 ± 0.14	0.28 ± 0.23	7.6 ± 5.3	12 ± 6	1.7 ± 0.4	0.075 ± 0.006	0.20 ± 0.03
	Datasets-10k	433 ± 106	662 ± 288	46200 ± 15500	78.7 ± 26.8	22200 ± 4130	483.72 ± 31.68	7.2 ± 2.6
		0.024 ± 0.005	0.022 ± 0.006	0.091 ± 0.019	2.08 ± 0.45	0.042 ± 0.003	0.015 ± 0.001	0.014 ± 0.001
		0.012 ± 0.005	0.0040 ± 0.0028	0.0034 ± 0.0013	0.14 ± 0.08	0.036 ± 0.015	0.0016 ± 0.0008	0.0028 ± 0.0013
	Datasets-50k	0.88 ± 0.04	0.90 ± 0.08	1.1 ± 0.5	3.4 ± 1.4	1.84 ± 0.83	0.65 ± 0.01	0.55 ± 0.01
		0.031 ± 0.006	0.030 ± 0.011	0.84 ± 0.76	5.454 ± 0.65	0.039 ± 0.007	0.020 ± 0.002	0.022 ± 0.001
		0.22 ± 0.07	0.27 ± 0.05	0.67 ± 0.61	3.8 ± 1.1	0.14 ± 0.06	0.022 ± 0.001	0.0054 ± 0.0013

Each method was run 10 times and the SD are shown. The best performance is marked in bold.

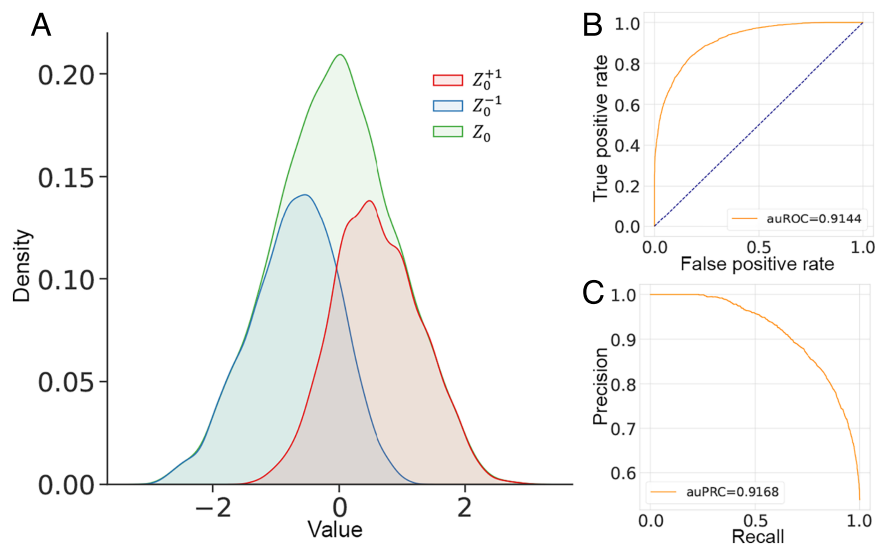


Fig. 3. Latent confounding variable given discrete covariates. (A) Distribution of latent confounding variable (Z_0), latent confounding variable selected by positive sign of V_1 (Z_0^{+1}), and latent confounding variable selected by negative sign of V_1 (Z_0^{-1}). (B) The receiver operating characteristic (ROC) curve when using Z_0 to predict the sign of V_1 . (C) The precision–recall (PR) curve when using Z_0 to predict the sign of V_1 .

as long as the high-dimensional covariates assumption holds, the randomness from a sufficient number of nonconfounding covariates can be used for constructing the latent normal variables needed for the conditioning. We also explored the alternative strategy of adding random noise to discrete covariates to directly transform the discrete distribution to a continuous distribution in *SI Appendix, section H*.

Ablation Study. Since CausalEGM consists of different modules. It is important to investigate the contribution of each component. First, we test the performance gain brought by the EGM framework. To do this, we removed the G network for reconstructing V and the discriminator network D_z for “distribution match” and denoted the model as “CausalEGM*.” Taking the continuous treatment setting for an example, we note that the performance of CausalEGM without the EGM framework has a noticeable decline in all datasets. The RMSE, MAPE increased by 32.6% to 164.9%, and 14.55% to 377.0%, respectively (Table 3). Such experimental results imply that the adversarial training and the reconstruction error are essential for learning a good low-dimensional representation of the high-dimensional covariates.

Next, we investigate whether the adversarial training for the covariates and the reconstruction for the latent features are necessary. In our model design, adversarial training in latent space

is necessary to guarantee the independence of latent variables. The reconstruction of V is also required to ensure the latent features contain all the information possessed by the original covariates. So we designed experiments to quantitatively evaluate the contribution of the adversarial training in covariate space and the reconstruction in latent space. The experiments show that the reconstruction of latent features could benefit the model training and achieve slightly better performance while the adversarial training in covariate space is not that helpful (*SI Appendix, section I*).

Robustness and Scalability. We conduct comprehensive experiments to examine the robustness and scalability of CausalEGM (*SI Appendix, section J*). Specifically, we first verify whether CausalEGM is sensitive to the choice of latent feature dimensions, which includes the total dimension of latent space and the dimension of common latent confounders Z_0 . The experimental results show that CausalEGM is quite robust to the choice of latent feature dimensions. For the scalability test, we demonstrate that CausalEGM is capable of handling datasets with a large number of covariates (e.g., >50K) and a large sample size (e.g., >5M) while many competing methods fail.

Discussion

In this paper, we developed a causal inference model named CausalEMG, which applies a dependency-aware dimension reduction to the high-dimensional covariates and extracts the latent confounding features that are used for covariate adjustment. The proposed EGM framework is shown to be effective in unraveling the dependencies of covariates on treatment and outcome and constructing the generative models for covariates, treatment, and outcome. A wide range of experiments demonstrate the superiority of our approach compared to existing methods. To sum up, CausalEGM is a flexible, scalable, and powerful approach for estimating the causal effect of a variable (e.g., treatment) on another (e.g., outcome), which provides a new perspective to analyze modern observational data in various domains with a large number of covariates and a large sample size.

Table 3. Ablation study on evaluating the contribution of EGM framework

Dataset	Method	RMSE	MAPE
Imbens et al	CausalEGM*	0.0936 ± 0.0579	0.0434 ± 0.0293
	CausalEGM	0.0706 ± 0.0445	0.0352 ± 0.0210
Sun et al	CausalEGM*	0.106 ± 0.0473	0.0438 ± 0.0224
	CausalEGM	0.0436 ± 0.0085	0.0180 ± 0.0038
Lee et al	CausalEGM*	1.28 ± 0.129	0.488 ± 0.0950
	CausalEGM	0.886 ± 0.232	0.426 ± 0.124
Twins	CausalEGM*	0.0641 ± 0.0252	2.38 ± 6.64
	CausalEGM	0.0242 ± 0.0132	0.499 ± 1.39

CausalEGM* represents CausalEGM approach without distribution match in latent space. Each method was run for 10 times and the SD are shown.

Several extensions and refinements of the CausalEGM model are left open. First, $H(\cdot)$ function can be further used and adapted to learn the propensity score in both binary and continuous treatment settings. It may benefit the development of new methods where an accurate propensity score model is required. Second, it is worth investigating the interpretable mechanism of latent features. For example, Z_2 naturally serves as a “latent instrumental variable” where it will only affect the outcome through the treatment variable. How to utilize the latent structure under the EGM framework to help identify (in)valid instruments is an open problem. Third, because of the extreme nonlinearity and complexity of neural network models, mathematically derived statistical properties such as valid CI and convergence rate are almost always missing for deep learning methods. It is helpful to use conformal prediction approaches (43–45) to further study the uncertainty of the estimate.

Materials and Methods

All the simulated datasets were generated through the generation processes provided by the original papers. Twins dataset was downloaded

from <https://www.nber.org/research/data/linked-birthinfant-death-cohort-data>. The data preprocess was followed by ref. 46. The ACIC 2018 benchmark datasets were downloaded from <https://www.synapse.org/#!/Synapse:syn11294478/wiki/486304>.

Data, Materials, and Software Availability. The project of CausalEGM is maintained at the website <https://causalegm.readthedocs.io/> (47), which provides detailed instructions and tutorials. We provide both Pypi Python package CausalEGM (48) and CRAN R package RcausalEGM (49). The source code of CausalEGM is provided at <https://github.com/SUwonglab/CausalEGM> (50). All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We wish to thank Balasubramanian Narasimhan for his help in releasing the CRAN R package. The work of Q.L. and W.H.W. was supported by the NSF Grant DMS 1952386 and NIH Grants (HG010359 and HG007735).

Author affiliations: ^aDepartment of Statistics, Stanford University, Stanford, CA 94305; ^bBio-X Program, Stanford University, Stanford, CA 94305; ^cDepartment of Statistics and Data Science, Yale University, New Haven, CT 06520; and ^dDepartment of Biomedical Data Science, Stanford University, Stanford, CA 94305

- K. J. Rothman, S. Greenland, Causation and causal inference in epidemiology. *Am. J. Public Health* **95**, S144–S150 (2005).
- J. Pearl, Causal inference in statistics: An overview. *Stat. Surv.* **3**, 96–146 (2009).
- G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).
- D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688 (1974).
- J. Neyman, D. M. Dabrowska, T. P. Speed, On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.*, 465–472 (1990).
- P. Ding, A first course in causal inference. arXiv [Preprint] (2023). <https://arxiv.org/abs/2305.18793> (Accessed 2 June 2023).
- I. Shpitser, T. VanderWeele, J. M. Robins, “On the validity of covariate adjustment for estimating causal effects” in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI’10* (AUAI Press, Arlington, VA, 2010), pp. 527–536.
- T. J. VanderWeele, Principles of confounder selection. *Eur. J. Epidemiol.* **34**, 211–219 (2019).
- M. Prosperi *et al.*, Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* **2**, 369–375 (2020).
- G. Davey Smith, M. V. Holmes, N. M. Davies, S. Ebrahim, Mendel’s laws, mendelian randomization and causal inference in observational data: Substantive and nomenclatural issues. *Eur. J. Epidemiol.* **35**, 99–111 (2020).
- L. Forastiere, E. M. Airolidi, F. Mealli, Identification and estimation of treatment and interference effects in observational studies on networks. *J. Am. Stat. Assoc.* **116**, 901–918 (2021).
- A. D’Amour, P. Ding, A. Feller, L. Lei, J. Sekhon, Overlap in observational studies with high-dimensional covariates. *J. Econ.* **221**, 644–654 (2021).
- P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
- K. Hirano, G. W. Imbens, The propensity score with continuous treatments. *Appl. Bayes. Mod. Causal Infer. Incomplete-data Perspect.* **226164**, 73–84 (2004).
- B. K. Lee, J. Lessler, E. A. Stuart, Improving propensity score weighting using machine learning. *Stat. Med.* **29**, 337–346 (2010).
- K. C. Li, Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316–327 (1991).
- K. C. Li, On principal hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma. *J. Am. Stat. Assoc.* **87**, 1025–1039 (1992).
- T. Ghosh, Y. Ma, X. De Luna, Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Stat. Sin.* **31**, 821 (2021).
- W. Luo, Y. Zhu, D. Ghosh, On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika* **104**, 51–65 (2017).
- J. Pearl, The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62**, 54–60 (2019).
- S. Athey, G. W. Imbens, Machine learning methods for estimating heterogeneous causal effects. *STAT* **1050**, 1–26 (2015).
- J. E. Brand, X. Zhou, Y. Xie, Recent developments in causal inference and machine learning. *Annu. Rev. Sociol.* **49**, 81–110 (2023).
- Q. Liu, J. Xu, R. Jiang, W. H. Wong, Density estimation using deep generative neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101344118 (2021).
- Q. Liu, S. Chen, R. Jiang, W. H. Wong, Simultaneous deep generative modelling and clustering of single-cell genomic data. *Nat. Mach. Intell.* **3**, 536–544 (2021).
- I. Goodfellow *et al.*, “Generative adversarial nets” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger, Eds. (Curran Associates, Inc., 2014), vol. 27.
- I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, “Improved training of Wasserstein GANs” in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), vol. 30.
- S. Ioffe, C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift” in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach, D. Blei, Eds. (PMLR, Lille, France, 2015), vol. 37, pp. 448–456.
- D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, May 7–9, 2015, Conference Track Proceedings*, Y. Bengio, Y. LeCun, Eds. (2015).
- Avd. Vaart, J. A. Wellner, “Empirical processes” in *Weak Convergence and Empirical Processes: With Applications to Statistics* (Springer, 2023), pp. 127–384.
- G. R. Shorack, J. A. Wellner, *Empirical Processes with Applications to Statistics* (SIAM, 2009).
- L. V. Truong, On rademacher complexity-based generalization bounds for deep learning. arXiv [Preprint] (2022). <https://arxiv.org/abs/2208.04284> (Accessed 15 October 2023).
- X. Li, J. Lu, Z. Wang, J. Haupt, T. Zhao, On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond. arXiv [Preprint] (2018). <https://arxiv.org/abs/1806.05159> (Accessed 15 October 2023).
- W. Sun, P. Wang, D. Yin, J. Yang, Y. Chang, “Causal inference via sparse additive models with application to online advertising” in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
- K. Colangelo, Y. Y. Lee, Double debiased machine learning nonparametric inference with continuous treatments. arXiv [Preprint] (2020). <https://arxiv.org/abs/2004.03036> (Accessed 3 October 2023).
- D. Galagate, *Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response Functions with Applications* (University of Maryland, College Park, MD, 2016).
- K. Imai, D. A. Van Dyk, Causal inference with general treatment regimes: Generalizing the propensity score. *J. Am. Stat. Assoc.* **99**, 854–866 (2004).
- U. Shalit, F. D. Johansson, D. Sontag, “Estimating individual treatment effect: Generalization bounds and algorithms” in *International Conference on Machine Learning* (PMLR, 2017), pp. 3076–3085.
- C. Shi, D. Blei, V. Veitch, Adapting neural networks for the estimation of treatment effects. *Adv. Neural Inf. Process. Syst.* **32**, 2507–2517 (2019).
- C. Louizos *et al.*, Causal effect inference with deep latent-variable models. *Adv. Neural Inf. Process. Syst.* **30**, 6449–6459 (2017).
- J. Yoon, J. Jordon, M. Van Der Schaar, “Ganite: Estimation of individualized treatment effects using generative adversarial nets” in *International Conference on Learning Representations* (2018).
- S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
- T. Ghosh, Y. Ma, X. De Luna, Sufficient dimension reduction for feasible and robust estimation of average causal effect. *Stat. Sin.* **31**, 821–842 (2021).
- G. Shafer, V. Vovk, A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**, 371–421 (2008).
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
- L. Guan, Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika* **110**, 33–50 (2022).
- Y. Li *et al.*, “Continuous treatment effect estimation via generative adversarial de-confounding” in *Proceedings of the 2020 KDD Workshop on Causal Discovery* (PMLR, 2020), vol. 127, pp. 4–22.
- Q. Liu, Z. Chen, W. Wong, Website: causal inference with encoding generative modeling. Read the Docs. <https://causalegm.readthedocs.io>. Deposited 6 June 2022.
- Q. Liu, Z. Chen, W. Wong, Python software: causal inference with encoding generative modeling. The Python Package Index. <https://pypi.org/project/CausalEGM>. Deposited 5 October 2022.
- Q. Liu, Z. Chen, W. Wong, R software: causal inference with encoding generative modeling. The Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/RcausalEGM/index.html>. Deposited 28 March 2023.
- Q. Liu, Z. Chen, W. Wong, Code: causal inference with encoding generative modeling. Github. <https://github.com/SUwonglab/CausalEGM>. Deposited 28 December 2022.