Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Generative adversarial mediation network: A novel generative learning approach to causal mediation analysis

Jiaming Zhang [a], Yiqi Lin [c], Xinyuan Song [c], Hanwen Ning [a,b,*]

[a] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, PR China
[b] Innovation and Talent Base for Digital Technology and Finance, Zhongnan University of Economics and Law, PR China
[c] Department of Statistics, The Chinese University of Hong Kong, Shatin, NT., Hong Kong, China

ARTICLE INFO

Keywords:
Conditional generative adversarial networks
Mediation analysis
Treatment effects
Counterfactuals
Deep neural networks

ABSTRACT

Casual mediation analysis (CMA) plays an essential role in various fields of social sciences. However, traditional models have restrictive parametric settings and strong random assumptions, which can be inflexible due to general nonlinearity, heterogeneity, and complex noise effects in many applications. Motivated by the similarities between the CMA and image-to-image translation that were thought to be unrelated initially, this paper proposes a novel prototype called the Generative Adversarial Mediation Network (GAMN), to explore the generative learning approach in the context of CMA. Thanks to a new encoding scheme for random terms, carefully designed partially linear architecture and inherent advantages of the generative learning framework, GAMN can flexibly handle nonlinear covariate effects and effectively model complex noise and heterogeneous mediating mechanisms with minimal model assumptions. Thus, when encountering intricate data patterns, the counterfactuals relating to treatment effects in CMA can be efficiently inferred, providing more promising mediation results. Experiments conducted on both synthetic and realistic datasets demonstrate that, compared with state-of-the-arts, GAMN can achieve notably more accurate estimations of out-of-sample predictions and treatment/mediation effects, which further illustrate the utility and advantages of our method. With the novel reinterpretations and solid theoretical results, this study also substantially broadens insights into developing mediation models from a machine-learning perspective.

1. Introduction

Causal mediation analysis (CMA) is a powerful tool for investigating causal mechanism of treatment [1,2]. It finds applications in various research fields, such as psychology [3], epidemiology [4] and economics [5]. CMA aims to evaluate treatment or intervention effects on an outcome of interest. This is achieved by disentangling the total treatment effect into an indirect effect operating through one or several observed intermediate variables (mediators) and a direct effect reflecting any impact not captured by the mediator(s).

A typical mediation model comprises a three-variable system in which a treatment variable (T) influences a dependent variable (Y) via one or more intermediary variables (M). This configuration gives rise to both direct ($T \rightarrow Y$) and indirect ($T \rightarrow M \rightarrow Y$) pathways [6]. Linear regression (LR) is a widely used approach to construct the system, wherein linear models are employed to estimate the direct and indirect effects. Corresponding statistical inference methods have also been established to conduct further mediation analysis [2]. It is worth noting that LR-based mediation models primarily focus on single-level mediation and straightforward linear correlations between variables. Alternatively, structural equation models (SEM) are employed to formulate mediation models for more complicated mediation problems [7]. SEMs allow for the creation of complex and interactive structures and are especially valuable for conducting multiple-level mediation analyses [8]. The Bayesian method (BM) represents another significant approach [9], which incorporates prior information into analysis, potentially improving estimation efficiency [10,11]. Bayesian mediation models, by specifying the distribution types of variables, offer straightforward and precise statistical inference. This imparts conceptual simplicity and convenience to the mediation analysis process [12]. Recently, causal mediation models have been investigated with relaxed assumptions based on the counterfactual framework. As unobserved counterfactuals are predicted using a parametric regression model [13], average treatment/mediation effects are evaluated by comparing the observed factual outcomes with the estimated counterfactual

* Corresponding author at: School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, PR China.
E-mail addresses: zjming1994@gmail.com (J. Zhang), yiqiLIN@link.cuhk.edu.hk (Y. Lin), xysong@sta.cuhk.edu.hk (X. Song), ninghanwen@gmail.com (H. Ning).

https://doi.org/10.1016/j.knosys.2023.111117
Received 10 July 2023; Received in revised form 21 October 2023; Accepted 22 October 2023
Available online 28 October 2023

outcomes. These works formulate flexible frameworks and generalize existing results for capturing complex mediating mechanisms. It is noted that conventional mediation models often impose restrictive parametric settings (e.g., linearity) and strong distribution assumptions (e.g., homogeneity and normality) to facilitate estimation and hypothesis testing. However, there is a lack of clear theoretical or intuitive justifications for these convenient assumptions in practice. Conversely, there are compelling reasons to expect that these assumptions are frequently violated. For instance, in fields like labor economics [14], epidemiology [15], and survival analysis [16], the data are sampled from individuals and essentially determined by personal characteristics, resulting in substantial high levels of nonlinearity, complexity and heterogeneity. This holds particularly true in the context of large datasets. Consequently, despite their considerable successes, the conventional methods may be too rigid to accommodate these prevalent intricate patterns and might be inadequate for many real-world applications.

The generative adversarial network (GAN) is an emerging generative learning model [17]. GANs have demonstrated remarkable success in tackling various challenging tasks, primarily within the domain of image processing, such as image generation [18], image-to-image translation [19], image restoration [20] and neural style transfer (NST) [21]. GANs consist of two components: a generative model called a generator and an adversarial model called a discriminator. The generator tries to capture the distribution of the observations and generate new samples, while the discriminator evaluates the new samples and distinguish them from real data. Through a minimax optimization process, wherein the generator and discriminator continually enhance their respective capabilities, GANs drive the generator to accurately learn the underlying data distribution. Despite being implicit generative models, GANs effectively capture distribution characteristics by employing deep neural networks for both the two components. From the generative learning perspective, it is noticed that the LR models can be considered explicit generative models to describe the joint probability involved in the regression. This feature motivates us to explore the feasibility of GANs for addressing CMA problems.

To our knowledge, this study represents the first attempt to investigate CMA using the generative learning approach. We have observed that in a classical image-to-image translation task [22], conditional GAN (CGANs) employs facial features (young/aged) as labels and leverage deep network architectures to learn the distribution of different facial images, generating/predicting the appearance of a young man as he ages. Correspondingly, in CMA, if we conceptualize binary treatment as the "young/aged" label and counterfactual as the desired generated image, CMA can be reinterpreted as an image-to-image translation problem. Inspired by this intriguing connection, we reframe the conventional mediation model and develop a novel prototype named generative adversarial mediation network (GAMN) based on CGAN [23]. Thanks to our new encoding scheme for random terms, carefully designed network architectures, and inherent advantages of the generative learning framework, GAMN can flexibly accommodate nonlinear relationships among variables and effectively model complex noises and heterogeneous mediating mechanisms with minimal assumptions. Therefore, when encountering intricate data patterns, GAMN can efficiently estimate counterfactuals and their associated confidence intervals relating to direct and indirect treatment effects, bringing more promising and adaptive mediation results than state-of-the-art methods. Besides the mediation problem in our study, other benchmark mediation problems can also be addressed following a similar technique line as presented in this paper. With the novel techniques and reinterpretations, this study substantially improves the settings of CMA and represents a significant step toward developing mediation models from a machine-learning perspective. Theoretical analysis and encouraging numerical experiments on synthetic and realistic datasets illustrate the utility and advantages of the proposed model.

The rest of this article is organized as follows. Section 2 briefly reviews the existing benchmark methods for mediation analysis. Section 3 presents the methodology, including the motivations of this study, proposed network models, and some further discussions on our method. The counterfactual framework for CMA problems is also reviewed at the beginning of this section. Section 4 provides the theoretical results for our method. Section 5 conducts extensive numerical experiments on synthetic and realistic datasets to assess the empirical performance of the proposed model. Comparisons between the proposed and existing methods are presented. Section 6 concludes the paper.

2. Existing benchmark methods and counterfactual framework

In this section, we briefly review the existing benchmark mediation models and present the limitations of traditional methods. Vectors and scalers are denoted by bold and ordinary letters, respectively. Let Y be the dependent variable and T and M be the treatment variable and mediator, respectively. A typical LR-based mediation model is formulated by three linear equations [2]

$$\left\{\begin{array}{lcl} Y & = & \beta_1 + cT + \varepsilon_1, \\ M & = & \beta_2 + aT + \varepsilon_2, \\ Y & = & \beta_3 + bM + c'T + \varepsilon_3, \end{array}\right. \tag{1}$$

where ε_1, ε_2 and ε_3 are random terms with normal distributions. β_1, β_2 and β_3 are intercept terms, which represent the expected values of the dependent variables when all the independent variables are set to zero. The intercept terms are important for improving model fitting and stability in linear regressions. By plugging the first and second equations of (1) into the third one, the indirect effect can be measured by ab or $c' - c$. Ordinary least squares (OLS) [6] and maximum likelihood estimation (MLE) [24] are utilized for estimating a, b and c. The mediation analysis can be conducted by statistical inference on ab with confidence intervals (CI) established in these works. Structural equation model (SEM) is another popular approach. By moving the exogenous variable M to the right side [7,25], the SEM-based mediation model is given by

$$\left\{\begin{array}{lcl} M & = & \beta_2 + aT + \varepsilon_2, \\ Y & = & \beta_3 + bM + c'T + \varepsilon_3 = \beta_3 + \beta_2 b + (c' + ab)T + \varepsilon_3 + b\varepsilon_2 \end{array}\right. \tag{2}$$

where $\varepsilon_2 \sim N(0, \sigma_2^2)$ and $\varepsilon_3 \sim N(0, \sigma_3^2)$. Generalized Least Squares (GLS) and MLE are common estimating methods for (2). Compared with LR-based methods, SEM can be used to design complex and interactive structures and allows for multiple mediators or outcomes, facilitating complex multiple-level mediation analysis. BM-based alternatives utilize the Bayesian rule to estimate the parameters [26,27]. Since the distributions of variables are specified, mediation analysis in the context of (1) or (2) is straightforward.

Despite their effectiveness, the above methods typically impose restrictive parametric settings (e.g., linearity) and strong distribution assumptions (e.g., homogeneity and normality) to benefit estimation and statistical inference. Recent studies, such as [28–30], have highlighted that mediation analysis data often exhibit significant nonlinearity, complexity, and heterogeneity, especially when dealing with large-scale datasets. This suggests that the conventional methods may not be suitable for many real-world applications.

The total, direct, and indirect (or mediated) effects have their own causal interpretations under the counterfactual framework [31]. Counterfactual-based methods have been further developed and proved effective for CMA problems [32]. Following the notations in [33], $\boldsymbol{X}$ denotes pre-treatment covariates, and t^* and t indicate treatments for mediator and outcome, respectively. $Y(t, \boldsymbol{m})$ represents the potential outcome when treatment is set to t under fixed mediator $\boldsymbol{M} = \boldsymbol{m}$. To identify the path-specific direct and indirect effects, the following standard unconfoundedness assumptions are required:

(I) conditional independence of the treatment: $\{Y(t, \boldsymbol{m}), \boldsymbol{M}(t^*)\} \perp T \mid \boldsymbol{X}$,

Table 1
List of notations

Notations	Description	Notations	Description
Y	Outcome variable	$\triangle_{T\to Y}$	Direct effect
T	Treatment variable	$\triangle_{T\to \boldsymbol{M}\to Y}$	Total indirect effect
M ($\boldsymbol{M}$)	Mediator variable (vector)	$\triangle_{T\to M^p\to Y}$	Indirect effect implemented through the pth mediator M^p
$\boldsymbol{X}$	Covariant vector	$\boldsymbol{Z}_M \sim N((0,1)^{d_M})$	d_M-dimensional normally distributed random variable
$\mathbb{E}$	Mathematical expectation	$\boldsymbol{Z}_Y \sim N((0,1)^{d_Y})$	d_Y-dimensional normally distributed random variable
G_M	Generator of the mediator block	D_M	Network parameters of D_M
G_Y	Generator of the outcome block	D_Y	Network parameters of G_M
θ_{G_M}	Network parameters of G_M	θ_{D_M}	Network parameters of D_M
θ_{G_Y}	Network parameters of G_Y	θ_{D_Y}	Network parameters of D_Y
$p_1 \doteq p_2$	two density functions p_1 and p_2 are the same	$A \perp B \mid C$	Independence of A and B conditional on C
p_Z	Density of random variable Z	$\mathbb{D}_{JS}$	Jensen–Shannon divergence
p_g	The density of the generated samples	$\mathbb{D}_{KL}$	Kullback–Leibler divergence
p_{data}	The density of observation data	$\mathbb{D}_{TV}$	Total variation
η	learning rate	$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{T}$	value domains of X, Y, T

(II) conditional independence of the mediator: $Y(t,\boldsymbol{m}) \perp \boldsymbol{M}(t^*) \mid T, \boldsymbol{X}$,

where $A \perp B \mid C$ indicates the independence of A and B conditional on C, and the treatment value t and t^* can be set as different, but not necessary. With unconfoundedness assumptions, suppose the treatment value is chosen from candidates t_0 and t_1, i.e., $T \in \{t_0, t_1\}$. There exist four potential outcomes driven by different paths: $Y(t_0, \boldsymbol{M}(t_0))$, $Y(t_1, \boldsymbol{M}(t_1))$, $Y(t_0, \boldsymbol{M}(t_1))$, and $Y(t_1, \boldsymbol{M}(t_0))$. Obviously, it is impossible to impose both t_0 and t_1 on the same individual simultaneously. Either $Y(t_0, \boldsymbol{M}(t_0))$ or $Y(t_1, \boldsymbol{M}(t_1))$ can be observed. $Y(t_0, \boldsymbol{M}(t_1))$ and $Y(t_1, \boldsymbol{M}(t_0))$ cannot be observed. The observed one is called factual, and the unobserved ones are called counterfactuals. Without loss of generality, the direct effect is formulated by changing t from t_0 to t_1. Accordingly, the direct effect and total indirect effect are given as

$$\triangle_{T\to Y} = \mathbb{E}[Y(t_1, \boldsymbol{M}(t_0)) - Y(t_0, \boldsymbol{M}(t_0))], \quad \triangle_{T\to \boldsymbol{M}\to Y} = \mathbb{E}[Y(t_1, \boldsymbol{M}(t_1)) - Y(t_1, \boldsymbol{M}(t_0))], \tag{3}$$

where $\mathbb{E}$ denotes the mathematical expectation. Moreover, with (2), (3) can be specified as

$$\begin{aligned} \triangle_{T\to Y} &= [\beta_3 + b(\beta_2 + at_0) + c't_1] - [\beta_3 + b(\beta_2 + at_0) + c't_0] = (t_1 - t_0)c', \\ \triangle_{T\to \boldsymbol{M}\to Y} &= [\beta_3 + b(\beta_2 + at_1) + c't_1] - [\beta_3 + b(\beta_2 + at_0) + c't_1] = (t_1 - t_0)ab. \end{aligned} \tag{4}$$

(4) is consistent with the results of the conventional linear regression models, illustrating the effectiveness of the counterfactual framework. In the counterfactual framework, the strict linear and random assumptions are unnecessary. Therefore, deep learning techniques offer the potential to achieve highly accurate counterfactual predictions with minimal assumptions. This study aims to design a novel CMA model using the GAN approach under the counterfactual framework. In Table 1, we also present a list of explanations for the main symbols in this paper.

3. Methodology

3.1. A brief review on benchmark GANs and our key motivations

In this section, we first present a brief overview of benchmark GAN models. We also present our key motivations and elaborate how to solve the problem using CGAN techniques.

3.1.1. A brief review on benchmark GANs

The original version of GAN, initially introduced by Goodfellow, is a type of unsupervised deep learning model [17]. It comprises two components: the generator and the discriminator. The generator is to produce samples that exhibit a statistical similarity to the training data. The generator, denoted as G, is a differentiable function formulated by a neural network that operates on a low-dimensional random variable Z characterized by a density function p_Z. Z is typically referred to as "noise" and is commonly distributed as Gaussian or uniform in most cases. Consequently, the generator G is associated with a natural density, denoted as p_g. To learn the generator over data $\boldsymbol{y}$ (with density denoted by p_{data}), the discriminator D is also formulated by a neural network attempts to distinguish between the observations as "real data" or "generated data" by outputting the probability that $G(Z)$ came from p_{data} rather than p_g. D is trained to maximize the probability of assigning the correct label to both training data and generated data. Simultaneously, G is trained to generate the samples that are following the distribution p_{data} of real data. G and D play a two-player minimax game with objective function $V(G,D)$ to update themselves

$$\min_G \max_D V(G,D) = \mathbb{E}_{\boldsymbol{y}\sim p_{data}}[\log D(\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_Z}[\log(1 - D(G(\boldsymbol{z})))]. \tag{5}$$

The training objective involves minimizing the JS divergence (Jensen–Shannon divergence) between the real data and generated data [34]. Optimal generator and discriminator correspond to the solution of (5). GANs have evolved significantly since their inception. To make GANs applicable to various complex tasks, many important variants have been proposed.

Information Maximizing Generative Adversarial Network (InfoGAN) extends the basic GAN framework by explicitly modeling and controlling the latent representations of generated data [35]. InfoGAN introduces an additional objective during training, which encourages the generator to learn disentangled and interpretable representations in the latent space. InfoGAN has been applied to the tasks where understanding and controlling specific attributes or characteristics of generated data are crucial, for example, facial expression manipulation [36] and data augmentation [37]. Another popular variant is Cycle-Consistent Adversarial Network (CycleGAN). CycleGAN is first designed for image-to-image translation tasks, where there are no paired training examples [38]. Traditional methods typically require a one-to-one correspondence between images in the source and target domains, which can be impractical or expensive in many cases [19]. In contrast, CycleGAN introduces a cycle consistency constraint, ensuring that the identity of images is preserved during translation, and can perform image translation between two different domains without such pairs. This makes it highly versatile for tasks, such as style transfer [39], domain adaptation [40] and creative image transformations [41].

While GANs offer a promising framework for generating data, the training of GANs is difficult and often unstable due to the complexities of their min–max optimization problems that cannot be easily resolved by simply altering the network architecture. To address this issue, many benchmark variants have been proposed by redefining the objective function. In stead of JS divergence, Wasserstein GAN (WGAN) adopts the Wasserstein distance to quantify the dissimilarity between the real data distribution and the generated data distribution [42]. The Wasserstein distance offers a continuous and smooth measure of dissimilarity. This smoothness ensures that meaningful gradients are available throughout the training process, resulting in more stable training and helping to prevent issues like mode collapse. To further ensure that

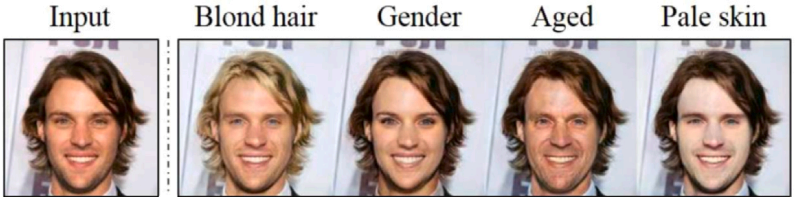


Fig. 1. Application of CGAN in image-to-image translation task: changing the facial attributes.

the Wasserstein distance is well-defined, WGAN with gradient penalty (WGAN-GP) adds a gradient penalty term to the loss function [43]. This encourages smoother behavior in the discriminator, promoting more reliable and consistent training. Instead of binary cross-entropy loss, Least Square GAN (LSGAN) utilizes least squares loss [44]. This adjustment yields several benefits, such as enhanced training stability and the generation of visually appealing samples [45].

The aforementioned GAN variants introduce diverse innovations and adaptations to tackle specific challenges in generative modeling. The choice of variant depends on the task requirements and the desired characteristics of the generated output. [46,47].

3.1.2. CGAN and our key motivations

When both the generator and discriminator are conditioned on auxiliary information $\boldsymbol{x}$, the GAN framework can be extended to a conditional version, referred to as CGAN. $\boldsymbol{x}$ can take various forms, including class labels or data from other modalities [23]. Specifically, the generator uses the noise vector $\boldsymbol{z}$ and auxiliary information $\boldsymbol{x}$ to synthesize fake sample $G(\boldsymbol{z}, \boldsymbol{x})$ with distribution $p_g(\boldsymbol{y}|\boldsymbol{x})$, while the discriminator classifies whether the extended observation $(\boldsymbol{y}, \boldsymbol{x})$ comes from real data $p_{data}(\boldsymbol{y}|\boldsymbol{x})$. For CGAN, the minimax optimization for training CGAN is given by

$$\min_G \max_D \mathbb{E}_{\boldsymbol{y}\sim p_{data}}[\log D(\boldsymbol{y}|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_Z}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{x})))]. \tag{6}$$

As auxiliary information is incorporated into the learning process, samples with different class labels can be generated. CGAN and its variants have been widely used for text-to-image synthesis [48], image-to-image translation [22] and neural style transfer tasks [38], where predictions of image data are needed. We remark that WGAN, WGAN-GP and LSGAN have their conditional versions, and we shall develop our GAMN using CGAN techniques.

The key motivation for this paper. For the learning tasks relating to complex conditional distributions, recent studies have shown that CGAN can achieve remarkable success in image-to-image translations. The image-to-image translation is to change a particular aspect of a given image to another, e.g., changing a landscape photograph into paintings of famous artists [38], or changing the facial appearance of a person [22]. As shown in Fig. 1, when given an image of a young man, the task is to generate an image of his aged appearance, for which a young-to-old mapping need to be learned. For this task, there are two main difficulties. First, it involves estimating high-dimensional and complex conditional distributions, a flexible enough network must be designed to learn the patterns of the images. Second, the young man has not become old yet, we need to "predict" his aged appearance. With deep network structures, CGAN can efficiently overcome the difficulties by setting the facial feature young/aged as a label in training. Fig. 1 also shows the intuitively amazing translation results by CGAN.

We emphasize that there are interesting similarities between image-to-image translation and counterfactual estimation in CMA. In image-to-image translation, the goal is to generate images of aged faces according to "young/aged" label, while in counterfactual estimation, the objective is to characterize the distribution of a dependent variable given a specific treatment. In essence, both tasks involve estimating conditional distributions of the variables of interest. If the binary treatment, counterfactual, and covariants are regarded as analogs to the "young/aged" label, desired generated image, and the codes for other facial characteristics (such as hair color), respectively, the counterfactual estimation problem can be reinterpreted as an image-to-image translation problem. Therefore, the GAN approach can be reasonably employed to develop novel CMA models, which implies the advantages of GAN such as flexibility to describe nonlinearity and high capacity to approximate complex distributions, can be leveraged to achieve more promising mediation results, thereby overcoming the limitations of traditional methods.

3.2. GAMN

3.2.1. GAMN for single mediator case

We first consider the model with one mediator. It is noted that $\boldsymbol{X}$, Y, M, and T are utilized to represent the variables involved in the general mediation models (1) and (2). To avoid any potential confusions, we emphasize that $(Y_i, M_i, T_i, \boldsymbol{X}_i)$ denote the values of $(Y, M, T, \boldsymbol{X})$ corresponding to the ith individual within the sample dataset used for modeling, and these values are specifically employed for calculation and estimation purposes. Assume that n samples $(Y_i, M_i, T_i, \boldsymbol{X}_i)$ $(i = 1, 2, \ldots, n)$ are used for training, where $\boldsymbol{X}_i \in \mathbb{R}^d$, $T_i \in \{t_0, t_1\}$, Y_i and M_i are in real space. For given i and fixed $\boldsymbol{X}_i$, either $Y_i(t_0, M_i(t_0))$ or $Y_i(t_1, M_i(t_1))$ can be observed, while $Y_i(t_0, M_i(t_1))$ and $Y_i(t_1, M_i(t_0))$ cannot be observed. In the following, without loss of generality, we always assume that t_0 and t_1 are the treatments corresponding to the factual and counterfactual, respectively, which implies $Y_i(t_1, M_i(t_1))$ and $Y_i(t_1, M_i(t_0))$ are the counterfactuals to be predicted. We propose to learn the distributions of potential outcomes conditional on $\boldsymbol{X}$ and T. Once the conditional distributions of M and Y are obtained, the counterfactuals can be estimated, and the individual direct and indirect effects can be calculated accordingly. The proposed CGAN for the single mediator case is named as GAMN-S.

The formulation of GAMN-S. Analogous to the benchmark model (2), our GAMN-S is composed of two blocks, a mediator block and an outcome block. Each block is designed as a CGAN. Corresponding to the first equation of (2), let G_M be the generator of the mediator block

$$\hat{M} = G_M(\boldsymbol{Z}_M, T, \boldsymbol{X}; \theta_{G_M}), \tag{7}$$

where $\hat{M}$ is the generated variable for mediator M, $\boldsymbol{Z}_M \sim N((0,1)^{d_M})$ is the random noise. $N((0,1)^{d_M})$ denotes a multivariate normal distribution with mean vector $(0, 0, \ldots, 0)$ and covariance matrix equal to the $d_M \times d_M$ identity matrix. θ_{G_M} represents the network parameters. G_M is designed as a feedforward network, for which $(\boldsymbol{Z}_M, T, \boldsymbol{X})$ forms the input layer. $\boldsymbol{Z}_M$ and $\boldsymbol{X}$ are associated with a deep fully-connected network, while T is associated with a linear structure. The architecture of G_M is demonstrated in Fig. 2. The sample set used for training the mediator block is denoted as $S_M = \{(M_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$. Corresponding to S_M, $\widehat{S_M} = \{(\hat{M}_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ represents the data set generated by G_M. Let $D_M(M, T, \boldsymbol{X}; \theta_{D_M})$ be the discriminator of the mediator block. D_M is used to measure the similarity between $\widehat{S_M}$ and S_M, and designed as a deep fully-connected forward neural network (FNN) with parameters θ_{D_M}. D_M takes sigmoid function as its output layer, with outputs ranged in $[0, 1]$. G_M and D_M constitute the mediator block of GAMN-S. The optimal θ_{G_M} and θ_{D_M} are obtained by the following minimax

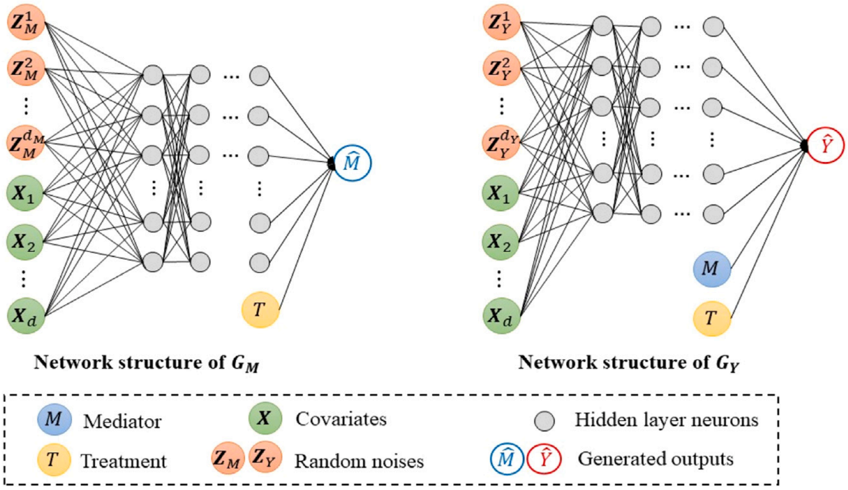


Fig. 2. The network structure of G_M and G_Y for GAMN-S.

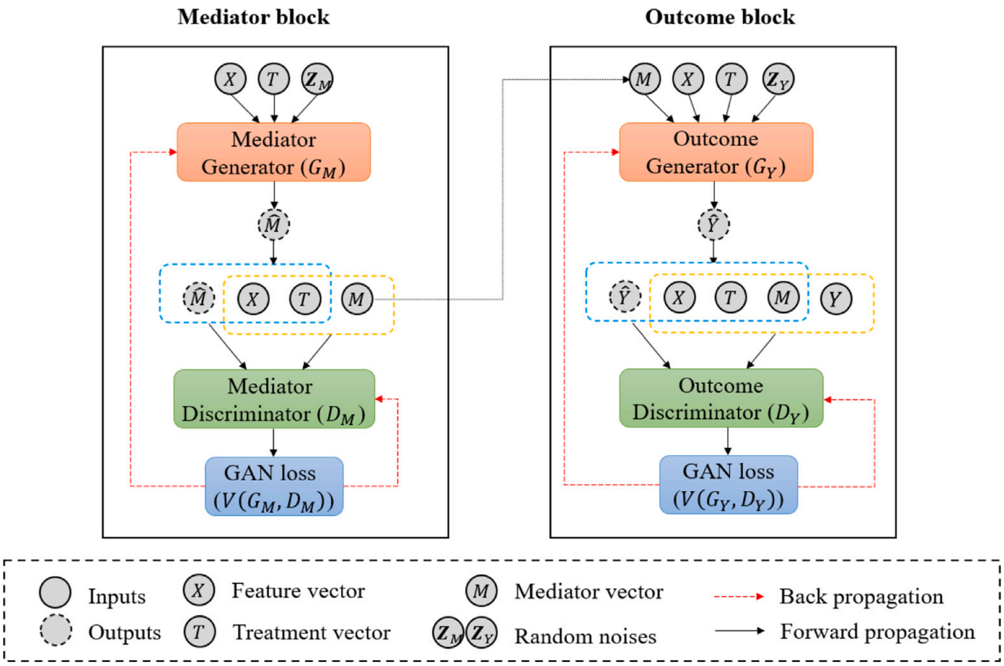


Fig. 3. Block diagram of GAMN-S.

optimization

$$\min_{\theta_{G_M}} \max_{\theta_{D_M}} \mathbb{E}_{M\sim P_{data}}[\log D_M(M,T,\boldsymbol{X};\theta_{D_M})]$$
$$+\mathbb{E}_{Z_M\sim N((0,1)^{d_M})}[\log(1-D_M(G_M(\boldsymbol{Z}_M,T,\boldsymbol{X};\theta_{G_M}),T,\boldsymbol{X};\theta_{D_M}))], \quad (8)$$

where P_{data} denotes the distribution of the observations. For the outcome block, the generator G_Y is given as

$$\hat{Y} = G_Y(\boldsymbol{Z}_Y, T, \boldsymbol{X}, M; \theta_{G_Y}), \quad (9)$$

where $\hat{Y}$ is the generated outcome, $\boldsymbol{Z}_Y$ is d_Y-dimensional normally distributed random noise, i.e. $\boldsymbol{Z}_Y \sim N((0,1)^{d_Y})$. θ_{G_Y} denotes the network parameters. Corresponding to the output Y, G_Y is designed as a particular FNN, setting M and T in the last layer but one of the network. The architecture of G_Y is also presented in Fig. 2. The designment of G_M and G_Y is further illustrated in Section 3.3. The sample set used for training the outcome block is denoted as $S_Y = \{(Y_i, M_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$. Corresponding to S_Y, $\widehat{S_Y} = \{(\hat{Y}_i, M_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ represents the data set generated by G_Y. $D_Y(Y, M, T, \boldsymbol{X}; \theta_{D_Y})$ is the discriminator of the outcome block, also designed as a fully-connected FNN with parameter θ_{D_Y}. D_Y also takes the sigmoid function as its output layer and is used to measure the similarity between $\widehat{S_Y}$ and S_Y. G_Y and D_Y form the CGAN for the outcome block and are trained by the following minimax optimization problem

$$\min_{\theta_{G_Y}} \max_{\theta_{D_Y}} \mathbb{E}_{Y\sim P_{data}}[\log D_Y(Y,M,T,\boldsymbol{X};\theta_{D_Y})]$$
$$+\mathbb{E}_{Z_Y\sim N((0,1)^{d_Y})}[\log(1-D_Y(G_Y(\boldsymbol{Z}_Y,T,\boldsymbol{X},M;\theta_{G_Y}),M,T,\boldsymbol{X};\theta_{D_Y}))] \quad (10)$$

We use Adaptive Momentum (Adam) algorithm as the optimizer to perform the training. The optimization schemes for our GAMN-S are presented in Algorithm 1. The block diagram of GAMN-S is summarized and presented in Fig. 3. If GAMN-S were well trained, the conditional densities $P(\hat{M}|\boldsymbol{X},T)$ and $P(\hat{Y}|T,\boldsymbol{X},M)$ can closely approximate $P(M|\boldsymbol{X},T)$ and $P(Y|T,\boldsymbol{X},M)$, respectively. Then, GAMN-S can be utilized to generate the desired counterfactuals with a high degree of accuracy.

Estimating direct effects. Let $G_M(\cdot;\hat{\theta}_{G_M})$ and $G_Y(\cdot;\hat{\theta}_{G_Y})$ be the well-trained generators. Let N_g be a large positive integer. To obtain filtered factual and counterfactual relating to M_i, two groups of random noises

$\boldsymbol{Z}_{M_i}(j_0)$ and $\boldsymbol{Z}_{M_i}(j_1)$ $(j_0, j_1 = 1, 2, \ldots, N_g)$ are independently drawn from $N((0,1)^{d_M})$. M_i under the treatment t_0 can be given as

$$\hat{M}_i(t_0, \boldsymbol{Z}_{M_i}(j_0)) = G_M(\boldsymbol{Z}_{M_i}(j_0), t_0, \boldsymbol{X}_i; \hat{\theta}_{G_M}), \tag{11}$$

where $\hat{M}_i(t_0, \boldsymbol{Z}_{M_i}(j_0))s$ are used to mimic the data generation process of the factual $M_i(t_0)$. Similarly, the possible individual counterfactuals for M_i can be generated by

$$\hat{M}_i(t_1, \boldsymbol{Z}_{M_i}(j_1)) = G_M(\boldsymbol{Z}_{M_i}(j_1), t_1, \boldsymbol{X}_i; \hat{\theta}_{G_M}). \tag{12}$$

$\hat{M}_i(t_1, \boldsymbol{Z}_{M_i}(j))s$ denote the generated individual counterfactuals, and essentially provide predictions in terms of the empirical distribution. By averaging these generated values, the mean of filtered factual can be calculated as

$$\hat{M}_i(t_0) = \frac{1}{N_g} \sum_{j_0=1}^{N_g} \hat{M}_i(t_0, \boldsymbol{Z}_{M_i}(j_0)), \tag{13}$$

and the mean estimation of individual counterfactual can be calculated as

$$\hat{M}_i(t_1) = \frac{1}{N_g} \sum_{j_1=1}^{N_g} \hat{M}_i(t_1, \boldsymbol{Z}_{M_i}(j_1)). \tag{14}$$

With $\hat{M}_i(t_0)$, the individual direct effects for the ith individual can be generated as

$$\triangle_{i,T\to Y}(j) = \hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^1_{Y_i}(j)) - \hat{Y}_i(t_0, \hat{M}_i(t_0), \boldsymbol{Z}^2_{Y_i}(j)), \tag{15}$$

where $\triangle_{i,T\to Y}(j)$ is the individual direct effect associated with $\boldsymbol{Z}^1_{Y_i}(j)$ and $\boldsymbol{Z}^2_{Y_i}(j)$. $\boldsymbol{Z}^1_{Y_i}(j)$ and $\boldsymbol{Z}^2_{Y_i}(j)$ $(j = 1, 2, \ldots, N_g)$ are independently drawn from $N((0,1)^{d_Y})$. $\hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^1_{Y_i}(j))s$ and $\hat{Y}_i(t_0, \hat{M}_i(t_0), \boldsymbol{Z}^2_{Y_i}(j))s$ are the generated values of Y_i with different treatments, and calculated as

$$\begin{aligned} \hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^1_{Y_i}(j)) &= G_Y(\boldsymbol{Z}^1_{Y_i}(j), t_1, \boldsymbol{X}(i), \hat{M}_i(t_0); \hat{\theta}_{G_Y}), \\ \hat{Y}_i(t_0, \hat{M}_i(t_0), \boldsymbol{Z}^2_{Y_i}(j)) &= G_Y(\boldsymbol{Z}^2_{Y_i}(j), t_0, \boldsymbol{X}(i), \hat{M}_i(t_0); \hat{\theta}_{G_Y}). \end{aligned} \tag{16}$$

Accordingly, the average direct effects corresponding to different noises can be calculated as

$$\triangle_{T\to Y}(j) = \frac{1}{n} \sum_{i=1}^{n} \triangle_{i,T\to Y}(j). \tag{17}$$

The direct effect can be estimated by

$$\begin{aligned} \triangle_{T\to Y} &= \frac{1}{N_g} \sum_{j=1}^{N_g} \triangle_{T\to Y}(j) \\ &= \frac{1}{N_g} \frac{1}{n} \sum_{j=1}^{N_g} \sum_{i=1}^{n} \Big(\hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^1_{Y_i}(j)) - \hat{Y}_i(t_0, \hat{M}_i(t_0), \boldsymbol{Z}^2_{Y_i}(j)) \Big). \end{aligned} \tag{18}$$

Estimating indirect effects. With $\hat{M}_i(t_1)$, the individual indirect effects for the ith individual can be generated as

$$\triangle_{i,T\to M\to Y}(j) = \hat{Y}_i(t_1, \hat{M}_i(t_1), \boldsymbol{Z}^3_{Y_i}(j)) - \hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^4_{Y_i}(j)), \tag{19}$$

where $\boldsymbol{Z}^3_{Y_i}(j)$ and $\boldsymbol{Z}^4_{Y_i}(j)$ $(j = 1, 2, \ldots, N_g)$ are also two groups of noises independently drawn from $N((0,1)^{d_Y})$. $\hat{Y}_i(t_1, \hat{M}_i(t_1), \boldsymbol{Z}^3_{Y_i}(j))s$ and $\hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^4_{Y_i}(j))s$ denote the generated counterfactuals of Y_i, which are calculated as

$$\begin{aligned} \hat{Y}_i(t_1, \hat{M}_i(t_1), \boldsymbol{Z}^3_{Y_i}(j)) &= G_Y(\boldsymbol{Z}^3_{Y_i}(j), t_1, \boldsymbol{X}(i), \hat{M}_i(t_1); \hat{\theta}_{G_Y}), \\ \hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^4_{Y_i}(j)) &= G_Y(\boldsymbol{Z}^4_{Y_i}(j), t_1, \boldsymbol{X}(i), \hat{M}_i(t_0); \hat{\theta}_{G_Y}). \end{aligned} \tag{20}$$

$\triangle_{i,T\to M\to Y}(j)$ is the individual indirect effects corresponding to $\boldsymbol{Z}^3_{Y_i}(j)$ and $\boldsymbol{Z}^4_{Y_i}(j)$. Based on (19), the average indirect effect corresponding to different noises can be calculated as

$$\triangle_{T\to M\to Y}(j) = \frac{1}{n} \sum_{i=1}^{n} \triangle_{i,T\to M\to Y}(j). \tag{21}$$

Then, the indirect effect can be calculated as

$$\begin{aligned} \triangle_{T\to M\to Y} &= \frac{1}{N_g} \sum_{j=1}^{N_g} \triangle_{T\to M\to Y}(j) \\ &= \frac{1}{N_g} \frac{1}{n} \sum_{j=1}^{N_g} \sum_{i=1}^{n} \Big(\hat{Y}_i(t_1, \hat{M}_i(t_1), \boldsymbol{Z}^3_{Y_i}(j)) - \hat{Y}_i(t_1, \hat{M}_i(t_0), \boldsymbol{Z}^4_{Y_i}(j)) \Big). \end{aligned} \tag{22}$$

We emphasize that if the state of any individual under t_0 is counterfactual, the counterfactual can also be generated by $G_M(\cdot; \hat{\theta}_{G_M})$ and $G_Y(\cdot; \hat{\theta}_{G_Y})$ following the method presented in this subsection. Thus, all the states under t_1 are set as counterfactuals for consistency.

The empirical distributions and confidence intervals of direct/indirect effects. Since $\forall i$ $(i = 1, 2, \ldots, n)$ and $\forall j$ $(j = 1, 2, \ldots, N_g)$, $\boldsymbol{Z}^1_{Y_i}(j)$s and $\boldsymbol{Z}^2_{Y_i}(j)$s are independently sampled from $N((0,1)^{d_Y})$, $\triangle_{i,T\to Y}(j)s$ and $\triangle_{T\to Y}(j)s$ essentially provide the projected empirical distributions of individual and average direct effects, respectively. Similarly, $\triangle_{i,T\to M\to Y}(j)s$ and $\triangle_{T\to M\to Y}(j)s$ formulate the projected empirical distributions of individual and average indirect effects from the perspective of Monte Carlo, respectively. The corresponding confidence intervals for direct effects and indirect effects can be naturally established by computing the quantiles of $\triangle_{T\to Y}(j)s$ and $\triangle_{T\to M\to Y}(j)s$. Then, with (17) and (21), a more promising mediation analysis can be conducted using the proposed GAMN-S.

3.2.2. GAMN for multiple mediators case

Assume that there are P mediators, for the ith sample $(Y_i, \boldsymbol{M}_i, T_i, \boldsymbol{X}_i)$ $(i = 1, 2, \ldots, n)$, $\boldsymbol{M}_i$ is set as a P-dimensional mediator vector rather than a scaler in GAMN-S. The pth element of $\boldsymbol{M}_i$ is denoted as M_i^p, which implies $\boldsymbol{M}_i = (M_i^1, M_i^2, \ldots, M_i^P)$. The proposed CGAN for the multiple-mediator case is named GAMN-M.

The formulation of GAMN-M. The proposed GAMN-M also comprises two blocks, a mediator block and an outcome block. With P mediators, the mediator block consists of P CGANs. For $\forall p = 1, 2, \ldots, P$, let G_M^p be the generator for the pth mediator, defined as

$$\hat{M}^p = G_M^p(\boldsymbol{Z}_M^p, T, \boldsymbol{X}; \theta_{G_M^p}), \tag{23}$$

where $\boldsymbol{Z}_M^p \sim N((0,1)^{d_{M^p}})$ is set to be a d_{M^p}-dimensional white noise (corresponding to the pth mediator). $\theta_{G_M^p}$ represents the parameters of G_M^p. Let $D_M^p(M^p, T, \boldsymbol{X}; \theta_{D_M^p})$ be the discriminator for G_M^p, where $\theta_{D_M^p}$ represents the parameters of D_M^p. G_M^p and D_M^p construct the CGAN for M^p. We remark that their structures and corresponding minimax optimization problem are similar to G_M and D_M developed for the single mediator case. For the outcome block, the generator is still denoted as G_Y, and given as

$$\hat{Y} = G_Y(\boldsymbol{Z}_Y, T, \boldsymbol{X}, \boldsymbol{M}; \theta_{G_Y}), \tag{24}$$

where $\hat{Y}$ is the generated outcome, $\boldsymbol{Z}_Y$ is d_Y-dimensional normally distributed random noise, and θ_{G_Y} represents the network parameters. $D_Y(Y, \boldsymbol{M}, T, \boldsymbol{X}; \theta_{D_Y})$ is the discriminator of the outcome block. The optimization problem and network structure of CGAN for the outcome block is the same as GAMN-S, except that $\boldsymbol{M}$ in (24) is a P-dimensional vector instead of a scalar in (9). The training is also performed using Adam algorithm based on (10). The block diagram and training schemes are summarized in Fig. 4 and Algorithm 2, respectively.

Estimating direct effects. Let $G_M^p(\cdot; \hat{\theta}_{G_{M^p}})$s and $G_Y(\cdot; \hat{\theta}_{G_Y})$ be the well-trained generators. For $\forall i$ $(i = 1, 2, \ldots, n)$ and $\forall p$ $(p = 1, 2, \ldots, P)$, with a large N_g, two groups of random noises $\boldsymbol{Z}^p_{M_i}(j_0)$ and $\boldsymbol{Z}^p_{M_i}(j_1)$ $(j_0, j_1 = 1, 2, \ldots, N_g)$ are independently drawn from $N((0,1)^{d_{M^p}})$. The individual factuals and counterfactuals can be generated by

$$\begin{aligned} \hat{M}_i^p(t_0, \boldsymbol{Z}^p_{M_i}(j_0)) &= G_M^p(\boldsymbol{Z}^p_{M_i}(j_0), t_0, \boldsymbol{X}_i; \hat{\theta}_{G_M^p}), \\ \hat{M}_i^p(t_1, \boldsymbol{Z}^p_{M_i}(j_1)) &= G_M^p(\boldsymbol{Z}^p_{M_i}(j_1), t_1, \boldsymbol{X}_i; \hat{\theta}_{G_M^p}). \end{aligned} \tag{25}$$

Then, the mean of the filtered factuals and counterfactuals can be calculated as

$$\hat{M}_i^p(t_0) = \frac{1}{N_g} \sum_{j_0=1}^{N_g} \hat{M}_i^p(t_0, \boldsymbol{Z}^p_{M_i}(j_0)), \quad \hat{M}_i^p(t_1) = \frac{1}{N_g} \sum_{j_1=1}^{N_g} \hat{M}_i^p(t_1, \boldsymbol{Z}^p_{M_i}(j_1)). \tag{26}$$

Algorithm 1: Training schemes of GAMN-S.

```
Initialization: parameters $\theta_{G_M}, \theta_{D_M}, \theta_{G_Y}, \theta_{D_Y}$ and learning rate $\eta$.
 while training loss $V_1$ and $V_2$ has not converged do
    Receiving $\{(Y_i, M_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$; Drawing $\{\boldsymbol{Z}_{M_i}\}_{i=1}^n$; Drawing $\{\boldsymbol{Z}_{Y_i}\}_{i=1}^n$
    for $i = 1, 2, \ldots, n$ do
        $\hat{M}_i \leftarrow G_M(\boldsymbol{Z}_{M_i}, T_i, \boldsymbol{X}_i; \theta_{G_M})$
        $\hat{Y}_i \leftarrow G_Y(\boldsymbol{Z}_{Y_i}, M_i, T_i, \boldsymbol{X}_i; \theta_{G_Y})$
    end for
    Discriminator optimization
    Fixed $\theta_{G_M}$ and $\theta_{G_Y}$
    Maximize $V_1 = \frac{1}{n}\sum_{i=1}^n \left[\log D_M(M_i, T_i, \boldsymbol{X}_i; \theta_{D_M}) + \log D_Y(Y_i, M_i, T_i, \boldsymbol{X}_i; \theta_{D_Y})\right]$
        $+\frac{1}{n}\sum_{i=1}^n \left[\log(1 - D_M(\hat{M}_i, T_i, \boldsymbol{X}_i; \theta_{D_M})) + \log(1 - D_Y(\hat{Y}_i, M_i, T_i, \boldsymbol{X}_i; \theta_{D_Y}))\right]$
    Update $\theta_{D_M}$ and $\theta_{D_Y}$ by Adam
    Generator optimization
    Fixed $\theta_{D_M}$ and $\theta_{D_Y}$
    Minimize $V_2 = \frac{1}{n}\sum_{i=1}^n \left[\log(1 - D_M(G_M(\boldsymbol{Z}_{M_i}, T_i, \boldsymbol{X}_i; \theta_{G_M}), T_i, \boldsymbol{X}_i; \theta_{D_M}))\right.$
        $\left.+\log(1 - D_Y(G_Y(\boldsymbol{Z}_{Y_i}, M_i, T_i, \boldsymbol{X}_i; \theta_{G_Y}), M_i, T_i, \boldsymbol{X}_i; \theta_{D_Y}))\right]$
    Update $\theta_{G_M}$ and $\theta_{G_Y}$ by Adam
 end while
Output: $\theta_{G_M}, \theta_{D_M}, \theta_{G_Y}$ and $\theta_{D_Y}$
```

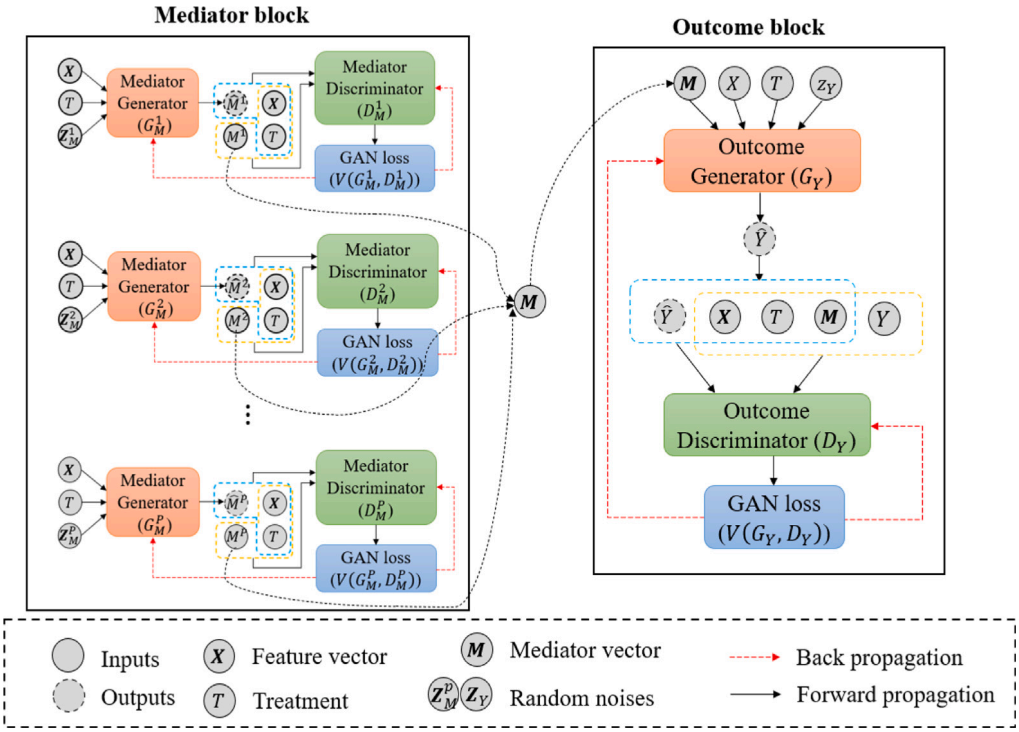

Mediator block
X
T
Z_M^1
Mediator Generator (G_M^1)
M̂^1
X
M^1
T
Mediator Discriminator (D_M^1)
GAN loss (V(G_M^1, D_M^1))
X
T
Z_M^2
Mediator Generator (G_M^2)
M̂^2
X
M^2
T
Mediator Discriminator (D_M^2)
GAN loss (V(G_M^2, D_M^2))
X
T
Z_M^P
Mediator Generator (G_M^P)
M̂^P
X
M^P
T
Mediator Discriminator (D_M^P)
GAN loss (V(G_M^P, D_M^P))
M
Outcome block
M
X
T
Z_Y
Outcome Generator (G_Y)
Ŷ
Ŷ
X
T
M
Y
Outcome Discriminator (D_Y)
GAN loss (V(G_Y, D_Y))
Inputs
Outputs
X Feature vector
T Treatment
M Mediator vector
Z_M^P Z_Y Random noises
Back propagation
Forward propagation

Fig. 4. Block Diagram of GAMN-M.

As P mediators are involved, let $\hat{\boldsymbol{M}}_i(t_0) = [\hat{M}_i^1(t_0), \ldots, \hat{M}_i^P(t_0)]$ and $\hat{\boldsymbol{M}}_i(t_1) = [\hat{M}_i^1(t_1), \ldots, \hat{M}_i^P(t_1)]$. We also have $\hat{\boldsymbol{M}}_i^{(-p)}(t_1) = [\hat{M}_i^1(t_1), \ldots, \hat{M}_i^{p-1}(t_1), \hat{M}_i^p(t_0), \hat{M}_i^{p+1}(t_1) \ldots, \hat{M}_i^P(t_1)]$, where $\hat{\boldsymbol{M}}_i^{(-p)}(t_1)$ indicates the pth dimension of $\hat{\boldsymbol{M}}_i(t_1)$ replaced by $\hat{\boldsymbol{M}}_i^p(t_0)$ while the other $P-1$ dimensions are kept unchanged. With $\hat{\boldsymbol{M}}_i(t_0)$, the factuals and counterfactuals for Y_i can be generated by

$$\begin{aligned}\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^1(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^1(j), t_1, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i(t_0); \hat{\theta}_{G_Y}),\\ \hat{Y}_i(t_0, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^2(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^2(j), t_0, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i(t_0); \hat{\theta}_{G_Y}).\end{aligned} \tag{27}$$

where $\boldsymbol{Z}_{Y_i}^1(j)$ and $\boldsymbol{Z}_{Y_i}^2(j)$ $(j = 1, 2, \ldots, N_g)$ are drawn from $N((0,1)^{d_Y})$. Then, the individual direct effects can be generated by

$$\triangle_{i,T\to Y}(j) = \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^1(j)) - \hat{Y}_i(t_0, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^2(j)). \tag{28}$$

Accordingly, the average direct effect can be calculated as

$$\triangle_{T\to Y}(j) = \frac{1}{n}\sum_{i=1}^n \triangle_{i,T\to Y}(j). \tag{29}$$

The direct effect can be estimated as

$$\triangle_{T\to Y} = \frac{1}{N_g}\frac{1}{n}\sum_{j=1}^{N_g}\sum_{i=1}^n \left(\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^1(j)) - \hat{Y}_i(t_0, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^2(j))\right). \tag{30}$$

Estimating total indirect effects. With $\hat{\boldsymbol{M}}_i(t_0)$ and $\hat{\boldsymbol{M}}_i(t_1)$, for $\forall i$, the individual indirect effects can be generated by

$$\triangle_{i,T\to \boldsymbol{M}\to Y}(j) = \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^3(j)) - \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^4(j)), \tag{31}$$

where $\boldsymbol{Z}_{Y_i}^3(j)s$ and $\boldsymbol{Z}_{Y_i}^4(j)s$ are drawn from $N((0,1)^{d_Y})$. $\triangle_{i,T\to \boldsymbol{M}\to Y}(j)$ denotes the total individual indirect effects corresponding to $\boldsymbol{Z}_{Y_i}^3(j)$ and

Algorithm 2 Training schemes of GAMN-M

```
Initialization: parameters $\theta_{G_M^p}$, $\theta_{D_M^p}$, $(p = 1, 2, \ldots, P)$, $\theta_{G_Y}$, $\theta_{D_Y}$, and learning rate $\eta$.
  while training loss $V_1$ and $V_2$ has not converged do
    Receiving $\{(Y_i, \boldsymbol{M}_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$; Drawing $\{\boldsymbol{Z}_{Y_i}\}_{i=1}^n$ independently
    for $p = 1, 2, \ldots, P$ do
      Drawing $\{\boldsymbol{Z}_{M_i}^p\}_{i=1}^n$ independently
    end for
    for $i = 1, 2, \ldots, n$ do
      for $p = 1, 2, \ldots, P$ do
        $\hat{M}_i^p \leftarrow G_M^p(\boldsymbol{Z}_{M_i}^p, T_i, \boldsymbol{X}_i; \theta_{G_M^p})$
      end for
      $\hat{Y}_i \leftarrow G_Y(\boldsymbol{Z}_{Y_i}, \boldsymbol{M}_i, T_i, \boldsymbol{X}_i; \theta_{G_Y})$
    end for
    Discriminator optimization
    Fixed $\theta_{G_M^1}, \ldots, \theta_{G_M^p}$ and $\theta_{G_Y}$
    Maximize $V_1 = \frac{1}{n}\sum_{i=1}^n \left[\sum_{p=1}^P \log D_M^p(M_i^p, T_i, \boldsymbol{X}_i; \theta_{D_M^p}) + \log D_Y(Y_i, \boldsymbol{M}_i, T_i, \boldsymbol{X}_i; \theta_{D_Y})\right]$
        $+\frac{1}{n}\sum_{i=1}^n \left[\sum_{p=1}^P \log(1 - D_M^p(\hat{M}_i^p, T_i, \boldsymbol{X}_i; \theta_{D_M^p})) + \log(1 - D_Y(\hat{Y}_i, \boldsymbol{M}_i, T_i, \boldsymbol{X}_i; \theta_{D_Y}))\right]$
    Update $\theta_{D_M^p}$ $(p = 1, 2, \ldots, P)$ and $\theta_{D_Y}$ by Adam
    Generator optimization
    Fixed $\theta_{D_M^1}, \ldots, \theta_{D_M^p}$ and $\theta_{D_Y}$
    Minimize $V_2 = \frac{1}{n}\sum_{i=1}^n \big[\sum_{p=1}^P \log(1 - D_M^p(G_M^p(\boldsymbol{Z}_{M,i}^p, T_i, \boldsymbol{X}_i; \theta_{G_M^p}), T_i, \boldsymbol{X}_i; \theta_{D_M^p}))$
        $+\log(1 - D_Y(G_Y(\boldsymbol{Z}_{Y_i}, T_i, \boldsymbol{X}_i, \boldsymbol{M}_i; \theta_{G_Y}), \boldsymbol{M}_i, T_i, \boldsymbol{X}_i; \theta_{D_Y}))\big]$
    Update $\theta_{G_M^p}$ $(p = 1, 2, \ldots, P)$ and $\theta_{G_Y}$ by Adam
  end while
Output: $\theta_{G_M^p}$ $(p = 1, 2, \ldots, P)$, $\theta_{D_M^p}$ $(p = 1, 2, \ldots, P)$, $\theta_{G_Y}$ and $\theta_{D_Y}$
```

$\boldsymbol{Z}_{Y_i}^4(j)$, i.e., the sum of the individual indirect effects through all the M^ps. $\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^3(j))$ and $\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^4(j))$ can be calculated by

$$\begin{aligned}\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^3(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^3(j), t_1, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i(t_1); \hat{\theta}_{G_Y}),\\ \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^4(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^4(j), t_1, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i(t_0); \hat{\theta}_{G_Y}).\end{aligned} \tag{32}$$

Then, the average total indirect effects can be calculated as

$$\triangle_{T\to\boldsymbol{M}\to Y}(j) = \frac{1}{n}\sum_{i=1}^{n} \triangle_{i,T\to\boldsymbol{M}\to Y}(j). \tag{33}$$

The total indirect effect can be calculated as

$$\triangle_{T\to\boldsymbol{M}\to Y} = \frac{1}{N_g}\frac{1}{n}\sum_{j=1}^{N_g}\sum_{i=1}^{n}\left(\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^3(j)) - \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_0), \boldsymbol{Z}_{Y_i}^4(j))\right). \tag{34}$$

Estimating indirect effects through a given mediator. For $\forall p$ $(p = 1, 2, \ldots, P)$, the individual indirect effects implemented through the pth mediator M^p can be generated by

$$\triangle_{i,T\to M^p\to Y}(j) = \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^{p(1)}(j)) - \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i^{(-p)}(t_1), \boldsymbol{Z}_{Y_i}^{p(2)}(j)), \tag{35}$$

where $\boldsymbol{Z}_{Y_i}^{p(1)}(j)s$ and $\boldsymbol{Z}_{Y_i}^{p(2)}(j)s$ are drawn from $N((0,1)^{d_Y})$. $\triangle_{i,T\to M^p\to Y}(j)$ denotes the individual indirect effects. $\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^{p(1)}(j))s$ and $\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i^{(-p)}(t_1), \boldsymbol{Z}_{Y_i}^{p(2)}(j))s$ denote the generated counterfactuals, calculated by

$$\begin{aligned}\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^{p(1)}(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^{p(1)}(j), t_1, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i(t_1); \hat{\theta}_{G_Y}),\\ \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i^{(-p)}(t_1), \boldsymbol{Z}_{Y_i}^{p(2)}(j)) &= G_Y(\boldsymbol{Z}_{Y_i}^{p(2)}(j), t_1, \boldsymbol{X}(i), \hat{\boldsymbol{M}}_i^{(-p)}(t_1); \hat{\theta}_{G_Y}).\end{aligned} \tag{36}$$

Then, the average indirect effects corresponding to different noises can be calculated as

$$\triangle_{T\to M^p\to Y}(j) = \frac{1}{n}\sum_{i=1}^{n} \triangle_{i,T\to M^p\to Y}(j). \tag{37}$$

The indirect effect implemented through the pth mediator M^p can be estimated as

$$\triangle_{T\to M^p\to Y} = \frac{1}{N_g}\frac{1}{n}\sum_{j=1}^{N_g}\sum_{i=1}^{n}\left(\hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i(t_1), \boldsymbol{Z}_{Y_i}^{p(1)}(j)) - \hat{Y}_i(t_1, \hat{\boldsymbol{M}}_i^{(-p)}(t_1), \boldsymbol{Z}_{Y_i}^{p(2)}(j))\right). \tag{38}$$

The empirical distributions and confidence intervals. The empirical distributions and confidence intervals of the direct and indirect effects can be obtained using $\triangle_{T\to Y}(j)$s, $\triangle_{T\to\boldsymbol{M}\to Y}(j)$s and $\triangle_{T\to M^p\to Y}(j)$s, similar to the one mediator case.

3.3. Further theoretical discussions

In this section, we present some further theoretical discussions on our GAMN. Compared with the benchmark methods, our GAMN provides several significant improvements.

Describing complex noise. In the benchmark mediation model (2), ε_2 and ε_3 are set to be normally distributed with zero means and appear as additive terms in the model. The only parameters that require estimation for ε_2 and ε_3 are their respective standard errors, i.e., σ_2 and σ_3. From the perspective of machine learning, model (2) uses a one-dimensional code to handle stochastic factors, i.e., σ_2 for ε_2 and σ_3 for ε_3. However, a single code insufficient to accurately capture the underlying noise characteristics if the normality setting is violated or a complex noise is present. It is noted that $\boldsymbol{Z}_M$ and $\boldsymbol{Z}_Y$ in GAMN are used to model the stochastic factors given by ε_2 and ε_3. In contrast to the normality setting, according to GAN theory, $\boldsymbol{Z}_M$ and $\boldsymbol{Z}_Y$ are multi-dimensional stochastic vectors and can be mapped by a proper deep neural network to approximate arbitrarily complex density functions, which implies $\boldsymbol{Z}_M$ and $\boldsymbol{Z}_Y$ and the associated deep network essentially provide a multi-dimensional encoding scheme to efficiently describe the complex noise, making our GAMN a more flexible alternative to the existing benchmarks.

Handling heterogeneity. The architecture of GAMN, as depicted in Fig. 2, shows that the covariate vectors X and $\boldsymbol{Z}_M$ are input into G_M, while X and $\boldsymbol{Z}_Y$ are input into G_Y, all within the same input layer of the network. These vectors interact with each other through the associated deep structure. In terms of the system (2), G_M defined by (7) and G_Y defined by (9) can be viewed as two equations from a nonlinear regression perspective, generally given as

$$M = g_M(\boldsymbol{X}, T, \varepsilon_2) \quad \text{and} \quad Y = g_Y(M, \boldsymbol{X}, T, \varepsilon_3), \tag{39}$$

where g_M and g_Y are nonlinear regression functions corresponding to G_M and G_Y, and ε_2 and ε_3 correspond to $\boldsymbol{Z}_M$ and $\boldsymbol{Z}_Y$, respectively. Since ε_2 and ε_3 can appear nonlinearly and are not mandatory to be additive, the couplings between the random terms and covariates are allowed in (39). If heterogeneity exists, e.g., $\boldsymbol{X}$, ε_2, and ε_3 are coupled as $g_1(\boldsymbol{X})\varepsilon_2$ and

$g_2(\boldsymbol{X})\varepsilon_3$ (g_1 and g_2 are unknown functions), the complex patterns can be well specified by GAMN with proper deep network structures. Moreover, unlike traditional regression models that impose strict parametric assumptions on the random terms to address heterogeneity, our method relaxes these apriori restrictions and can efficiently learn complex heterogeneous patterns.

Modeling nonlinearity. As shown in Fig. 2, $\boldsymbol{X}$ and $\boldsymbol{Z}_M$ in G_M are located in the input layer of the network. Using a deep feedforward structure, GAMN is capable of capturing the nonlinear relationships among the variables involved in G_M. For the network structure associated with T, on the one hand, in many benchmark mediation problems, whether the treatment is imposed on specific individuals is irrelevant to the characteristics represented by covariates [10,12]. Therefore, the treatment variable T is typically considered independent of $\boldsymbol{X}$ in many mediation problems, and as such, it is treated as a separate component without coupling with $\boldsymbol{X}$ in the structure of G_M. On the other hand, if T appears in a linear term in the mediation model, it is convenient to calculate the direct and indirect treatment effects and compare our network model with the traditional methods, by examining the coefficient associated with T. Consequently, the network structure related to T is specifically designed to be linear in line with standard mediation analysis practices.

The linear structures in GAMN. For G_Y, with binary treatment $T \in \{t_0^\star, t_1^\star\}$, only the samples from $(Y, \boldsymbol{X}, M(t_0^\star), t_0^\star)$ and $(Y, X, M(t_1^\star), t_1^\star)$ can be observed. $(Y, \boldsymbol{X}, M(t_1^\star), t_0^\star)$ and $(Y, \boldsymbol{X}, M(t_0^\star), t_1^\star)$ are the counterfactuals to be estimated based on the observations. For prediction, the neural network G_Y can also be considered as a continuous function, essentially making predictions based on its continuations. It is noted that the distances between the observations and counterfactuals (i.e., $\|(M(t_1^\star), t_0^\star) - (M(t_0^\star), t_0^\star)\|$, $\|(M(t_1^\star), t_0^\star) - (M(t_1^\star), t_1^\star)\|$, $\|(M(t_0^\star), t_1^\star) - (M(t_0^\star), t_0^\star)\|$ and $\|(M(t_0^\star), t_1^\star) - (M(t_0^\star), t_0^\star)\|$) could be quite large, which means the observations used to train G_Y are distinct from the counterfactuals in terms of component (M, T). The continuations of G_Y along the directions of (M, T) can be difficult and inaccurate. With a complex model structure (deep network structure associated with (M, T)), G_Y probably makes very biased predictions for inferring the counterfactuals. To address this, we adhere to Occam's razor principle in learning theory, favoring simplicity in the architecture related to M and T. Therefore, we position Z_Y and $\boldsymbol{X}$ in the deeper layers of G_Y, while M and T are designed to be in the shallower layers and participate in linear components of the network. This arrangement enhances the generalization ability of our model. Additionally, the deep structure associated with $\boldsymbol{X}$ and $\boldsymbol{Z}_Y$ naturally accommodates nonlinear covariate effects. Consequently, these architectural choices in GAMN contribute to more accurate counterfactual estimations and improved mediation analysis.

Model interpretability. Following the network structures shown in Fig. 2, the formulations of G_M and G_Y given by (7) and (9) can be further specified as

$$\hat{M} = G_M(\boldsymbol{Z}_M, T, \boldsymbol{X}) = aT + f_1(\boldsymbol{X}, \boldsymbol{Z}_M), \tag{40}$$

$$\hat{Y} = G_Y(\boldsymbol{Z}_Y, T, \boldsymbol{X}, M) = bM + c'T + f_2(\boldsymbol{X}, \boldsymbol{Z}_Y), \tag{41}$$

where a, b and c' are the parameters of the linear components. f_1 and f_2 represent two unknown functions corresponding to the nonlinear components of G_M and G_Y. On the one hand, by (4), the direct effects and indirect effects estimated in the counterfactual framework can be also given by $\triangle_{T\to Y} = (t_1 - t_0)c'$ and $\triangle_{T\to M\to Y} = (t_1 - t_0)ab$. Thus, the estimations of the linear parameters are crucial for mediation analysis. On the other hand, it is noted that $\partial\hat{M}/\partial T = a$, $\partial\hat{Y}/\partial T = b$ and $\partial\hat{Y}/\partial M = c'$, the linear parameters also directly quantify the marginal effects with respect to the treatment and mediator variables, which explicitly demonstrate the impact of changes in the treatment on the outcome and mediator variables. Therefore, our partially linear network designment can not only bring better generalization and more accurate counterfactual predictions, but also provide a certain level of model interpretability in the context of CMA. In Section 4, we will thoroughly discuss and provide proof regarding the convergence of our proposed model, specifically focusing on the convergence of the estimations for these linear parameters.

4. Theoretical view

We investigate the weak convergence of GAMN to conditional distribution and the convergence of the estimators of treatment and mediation effects. For simplicity, we consider the uniform GAMN function to represent modeling for G_Y and G_M hereafter. Consider $(X, Y, T) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{T}$, where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{T}$ are the value domains of X, Y and T, respectively. In the mediator block, Y can be replaced by M for the mediator. Likewise, we can replace T with (T, M) in the outcome block to have the original modeling. With (Y_i, T_i, X_i) $(i = 1, 2, \ldots, n)$, we have

$$\begin{aligned}(\hat{G}, \hat{D}) &= \min_G \max_D \mathcal{L}_n(G, D) \\ &\equiv \min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}[\log D(Y_i, T_i, X_i)] \\ &+ \frac{1}{n}\sum_{i=1}^{n}[\log(1 - D(G(Z_i, T_i, X_i), T_i, X_i))],\end{aligned} \tag{42}$$

where $\mathcal{D}$ and $\mathcal{G}$ are function spaces spanned by ReLU-activated FNN for D and G, respectively. Inspired by recent advances [49], we first provide a different insight to regard CGANs (8) and (10), which are conditional on the continuous and categorical variables X instead of finite labels, as the generative learning targeting the joint distribution. $p_{X,Y,T}$ and $p_{X,G(Z,T,X),T}$ denote the densities of the joint distributions of (X, Y, T) and $(X, G(Z, T, X), T)$. By Lemma 2.2 in [49], suppose that Z is independent of X and T. Then $p_{G(Z,T,X)} \doteq p_{Y|X,T}$ if and only if $p_{X,G(Z,X,T),T} \doteq p_{X,Y,T}$, where $\doteq$ represents that the two density functions are the same. The lemma facilitates investigating the optimization to minimize the discrepancy between $p_{X,G(Z,T,X),T}$ and $p_{X,Y,T}$, instead of directly working on conditional distribution $p_{Y|X,T}$. Let $\mathbb{D}_{JS}$ and $\mathbb{D}_{KL}$ be JS and Kullback–Leibler (KL) divergence, respectively. According to [17,23], we have

$$\begin{aligned}&\mathbb{D}_{JS}(p_{X,Y,T} \parallel p_{X,G(Z,T,X),T}) \\ &= \sup_D \left\{ \mathbb{E}_{W_1\sim p_{X,Y,T}} \log D(W_1) + \mathbb{E}_{W_2\sim p_{X,G(Z,T,X),T}} \log(1 - D(W_2)) \right\} + \log 4 \\ &= \sup_D \left\{ \mathbb{E}_{X,Y,T\sim p_{X,Y,T}} \log D(X, Y, T) + \mathbb{E}_{X,T,Z\sim p_{X,G(Z,T,X),T}} \log(1 - D(X, G(Z, X, T), T)) \right\} \\ &\equiv \sup_D \mathcal{L}(G, D).\end{aligned} \tag{43}$$

Therefore, the optimal G and D are defined as $(G^*, D^*) = \arg\min_G \arg\max_D \mathcal{L}(G, D)$, and their empirical counterparts are defined in (42). In our network structure, T only appears in the last layer and servers as a linear part for Y. Hence, $G(Z, X, T)$ can be further specified as

$$G(Z, X, T) = G_1(Z, X) + \beta T, \tag{44}$$

where $G_1(Z, X)$ is a fully ReLU-activated FNN and β is the coefficient of the treatment variable T. First, we show the convergence of $(X, G(Z, X, T), T)$ in distribution, and then apply the continuous mapping (44) to obtain the desired result. Using Scheff Lemma [see50, Chap 8.2], let $\mathbb{D}_{TV}$ be the total variation defined as

$$\mathbb{D}_{TV}(p_{X,Y,T}, p_{X,G(Z,T,X),T}) = \frac{1}{2}\|p_{X,Y,T} - p_{X,G(Z,T,X),T}\|_1, \tag{45}$$

where $\|\cdot\|_1$ is the L_1 norm. Further, we can bound the total variation by its' JS divergence, i.e.,

$$\begin{aligned}&\mathbb{D}_{JS}(p_{X,Y,T} \parallel p_{X,G(Z,T,X),T}) \\ &= \frac{1}{2}\left[\mathbb{D}_{KL}\left(p_{X,Y,T} \middle\| \frac{p_{X,Y,T} + p_{X,G(Z,T,X),T}}{2} \right) \right. \\ &+ \left. \mathbb{D}_{KL}\left(p_{X,G(Z,T,X),T} \middle\| \frac{p_{X,Y,T} + p_{X,G(Z,T,X),T}}{2} \right) \right] \\ &\geq \frac{1}{4}\left[\mathbb{D}_{TV}^2\left(p_{X,Y,T}, \frac{p_{X,Y,T} + p_{X,G(Z,T,X),T}}{2} \right) \right. \\ &+ \left. \mathbb{D}_{TV}^2\left(p_{X,G(Z,T,X),T}, \frac{p_{X,Y,T} + p_{X,G(Z,T,X),T}}{2} \right) \right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{8}\left\| \frac{p_{X,Y,T} - p_{X,G(Z,T,X),T}}{2} \right\|_1^2 \\
&= \frac{1}{8}\mathbb{D}_{TV}^2(p_{X,Y,T}, p_{X,G(Z,T,X),T}),
\end{aligned} \tag{46}
$$

where the inequality in the third line follows Pinsker's inequalities [51]. Therefore, we have paved the path to show a practical solution that $(\hat{G}, X, T)$ converge in distribution to (X, Y, T) by the technique of excess risk bound:

$$
\begin{aligned}
&\mathbb{D}_{TV}^2(p_{X,Y,T}, P_{X,\hat{G}(Z,T,X),T}) \lesssim \mathbb{D}_{JS}(p_{X,Y,T} \parallel p_{X,\hat{G}(Z,T,X),T}) \\
&= \sup_{D} \mathcal{L}(\hat{G}, D) - \sup_{D} \mathcal{L}(G^*, D) \\
&= \underbrace{\sup_{D\in\mathcal{D}} \mathcal{L}(\hat{G}, D) - \sup_{D} \mathcal{L}(\hat{G}, D)}_{\Delta_1} + \underbrace{\sup_{D\in\mathcal{D}} \mathcal{L}(\hat{G}, D) - \sup_{D\in\mathcal{D}} \mathcal{L}_n(\hat{G}, D)}_{\Delta_{21}} \\
&+ \underbrace{\sup_{D\in\mathcal{D}} \mathcal{L}_n(\hat{G}, D) - \sup_{D\in\mathcal{D}} \mathcal{L}_n(\bar{G}, D)}_{\Delta_{41}} + \underbrace{\sup_{D\in\mathcal{D}} \mathcal{L}_n(\bar{G}, D) - \sup_{D\in\mathcal{D}} \mathcal{L}(\bar{G}, D)}_{\Delta_{22}} \\
&+ \underbrace{\sup_{D\in\mathcal{D}} \mathcal{L}(\bar{G}, D) - \sup_{D} \mathcal{L}(\bar{G}, D)}_{\Delta_{42}} + \underbrace{\sup_{D} \mathcal{L}(\bar{G}, D) - \sup_{D} \mathcal{L}(G^*, D)}_{\Delta_3} \\
&\leq \Delta_1 + \Delta_3 + (\Delta_{21} + \Delta_{22}) \\
&\leq \Delta_1 + \Delta_3 + \underbrace{\sup_{D\in\mathcal{D}} \sup_{G\in\mathcal{G}} |\mathcal{L}(G, D) - \mathcal{L}_n(G, D)|}_{\Delta_2},
\end{aligned} \tag{47}
$$

where $\bar{G}$ is any element that belongs to $\mathcal{G}$. Δ_1 and Δ_3 represent the approximation errors of $\mathcal{D}$ and $\mathcal{G}$ for their optimal counterparts, respectively. Δ_2 is the supremum of estimation error for $\mathcal{L}_n$ to $\mathcal{L}$, and thus can be further controlled by the empirical process theorem. Like [49], the following mild assumptions are made for the target generators to derive the asymptotic results.

A.1 Target conditional generator G^* is continuous and its l_∞ norm is upper bounded.
A.2 For the optimal discriminator, $\frac{p_{X,Y,T}}{p_{X,G,Y}+p_{X,Y,T}}$ is lower and upper bounded in the support.
A.3 $\partial t_0 \mathbb{E}(Y|X, T = t_0)$ exists and is finite (for $T \in \{t_1, t_0\}$, $\partial t_0 = t_1 - t_0$).

For the FNN used for generators in $\mathcal{G}$ and discriminators in $\mathcal{D}$, we consider the following assumptions to show the asymptotic result as the network size grows with the sample size. Suppose a network has depth $\mathcal{H}$, width $\mathcal{W}$, and the whole size of the network is $\mathcal{S}$.

B.1 The l_∞ norm generator G within its support is upper bounded by constant B.
B.2 As sample size n goes to infinity, $\mathcal{H}\mathcal{W} \to 0$ and $\frac{B\mathcal{S}\mathcal{H}\log(\mathcal{S})\log n}{n} \to 0$.

Additionally, we consider the approximation error of specific structures in linear part βT in $\mathcal{G}$. Assuming that the true data-generating process conforms to a linear relationship between Y and T while controlling for both random error and covariate X. One can express this relationship as $Y = \beta^* T + f(X, \epsilon)$, where f represents an unknown function. Then $\partial t_0 \mathbb{E}(Y|X, T = t_0) = \partial t_0 [t_0 \beta^* + \mathbb{E}(f(X, \epsilon)|X)] = \beta^*$. The following theorem formally demonstrates weak convergence of the joint distributions of (X, T, Y) and $(X, T, G(Z, T, X))$, and the desired outcome is obtained for capturing the exact treatment effect β^*.

Theorem 1. *Under the assumptions A.1-A.3, B.1 and B.2, we have*

$$
\mathbb{E}_{X,T,Y,Z}\, \mathbb{D}_{TV}(p_{X,Y,T}, p_{X,\hat{G}(Z,T,X),T}) \xrightarrow{p} 0 \quad \textit{and} \quad P_{X,\hat{G}(Z,T,X),T} \xrightarrow{d} P_{X,Y,T}.
$$

Hence, we have further

$$
\begin{aligned}
\hat{\beta} &= \frac{1}{N_g}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_g}\left[\hat{G}(Z_j, X_i, T = t_1) - \hat{G}(Z_j, X_i, T = t_0)\right] \\
&\xrightarrow{p} \partial t_0 \mathbb{E}(Y|X, T = t_0).
\end{aligned}
$$

The above consistency results hold for our proposed network.

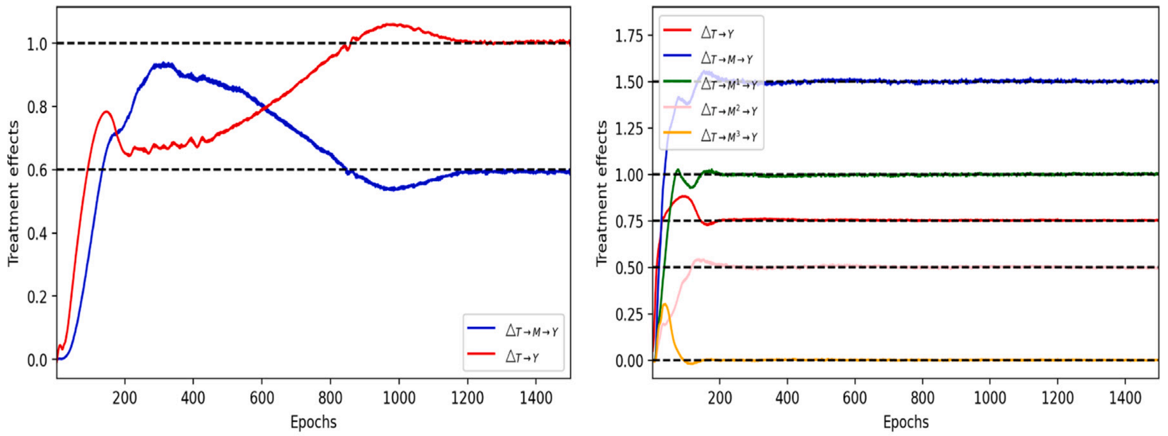
Proof. See Appendix. ◆

Remark on the convergence of parameters. In GAMN, the min–max optimization objective typically revolves around minimizing the conditional JS divergence between the conditional distribution of actual data and the conditional distribution of generated data [17,23]. The conditional JS divergence serves as a measure of the discrepancy between these two conditional distributions. The discriminator computes the JS divergence between the current generated distribution and the distribution of actual data. Subsequently, the generator produces data to establish a new generated distribution, thereby reducing JS divergence between the generated and actual distributions. Through iterative updates using Adam algorithm, we can achieve the optimal solution to the min–max optimization problem, effectively guiding the generator to learn the underlying conditional distribution of actual data correctly. It is noteworthy that the unknown parameters in GAMN implicitly define the conditional distribution function of the generated data. Hence, when the generated distribution converges to the objective distribution, the unknown network parameters are also updated and converge to their target values. Theorem 1 further provides the corresponding theoretical results regarding the convergence of the parameters.

Remark on our contributions. This paper's main contributions lie in three aspects. First, this study approaches the benchmark CMA problem from an unprecedented generative machine learning perspective, reinterpreting the CMA problem as an image-to-image problem, and also introduces novel GAN-based mediation models with adaptive architectures. Despite the similarities between the CMA problem and image-to-image translation problem, the widely used GANs in image processing cannot be straightforwardly applied to mediation analysis. Mediation models are commonly applied in causal inference and treatment effect evaluation in fields such as medicine and social sciences, demanding high levels of interpretability and accuracy in counterfactual estimation. The network architecture of conventional CGANs cannot meet the requirements of these CMA problems. In our GAMN, we specifically design the linear structure for the treatment variable T and the mediator variable M. Compared with the conventional CGANs, this designment not only provides more accurate counterfactual estimation but also enhances model interpretability, which is particularly beneficial for analyzing the mediation problems we are concerned with. Traditional mediation models make strict assumptions about model structure and random terms, which can also be relaxed in our framework. This allows us to explore the inherent complex patterns in the data in a more flexible manner and develop more accurate estimation and hypothesis testing methods.

Second, we prove the convergence of our method, with particular attention to the coefficients within the linear components of our GAMN. These convergence results offer a robust theoretical underpinning for our proposed GAN-based approach. Additionally, we conduct comprehensive empirical studies to demonstrate the effectiveness of our method.

Third, we demonstrate that GAN approach can be more effective for the development of novel mediation and causal inference models. Through the demonstration of reformulating the mediation model, this study broadens insights into the construction of CMA models from a generative learning perspective. Thus, the methodology presented in this study can be potentially extended to tackle a wide range of crucial



(a)

(b)

Fig. 5. (a) Displays the convergence curves of the estimations for system (48). The red solid line represents the estimations of the direct effect, while the blue solid line corresponds to the estimations of the indirect effect for each epoch. Notably, these estimations converge toward their target values, with the direct effect converging to 1 and the indirect effect converging to 0.6, achieved approximately after 1200 epochs. (b) displays the convergence curves of the estimations for system (49). The red solid line corresponds to the estimations of the direct effect, while the blue solid line represents the estimations of the total indirect effect for each epoch. Additionally, the green line, pink line, and yellow line represent the estimations of the indirect effects through M^1, M^2, and M^3, respectively. These estimations converge toward their respective target values, achieving convergence approximately after 300 epochs. Specifically, the target values are set as follows: 0.75 for the direct effect, 1.5 for the total indirect effect, and 1, 0.5, and 0 for the indirect effects through M^1, M^2, and M^3, respectively.

mediation problems, including high-dimensional mediation challenges and mediation scenarios with censored outcomes.

5. Experiments

In this section, numerical examples on artificial and realistic datasets are presented to demonstrate the effectiveness of our method (GAMN-S and GAMN-M). All the datasets and codes for this section shall be released on Github.

5.1. GAMN-S with simulated data

Consider the following data generation process

$$\begin{cases} M_i = 1 + 0.2\left(X_i + 1\right)^2 + 1.5T_i + \varepsilon_{M,i}, \\ Y_i = 2.5 - 0.1X_i^2 + \sin(1.5X_i) + 0.4M_i + T_i + \varepsilon_{Y,i}. \end{cases} \tag{48}$$

For $\forall i$, the treatment variable $T_i \sim \mathcal{B}(0.5)$ and pre-treatment covariate $X_i \sim \mathcal{N}(0,1)$, where $\mathcal{B}(0.5)$ is the Bernoulli distribution with a success probability of 0.5. The random error term is set as

$$\varepsilon_{M,i} = \vartheta_{M,0}^{(1-T_i)}\vartheta_{M,1}^{T_i} \quad \text{and} \quad \varepsilon_{Y,i} = \vartheta_{Y,0}^{(1-T_i)}\vartheta_{Y,1}^{T_i},$$

where $\vartheta_{M,0} \sim \mathcal{N}(0, 0.25^2)$, $\vartheta_{M,1} \sim \mathcal{U}(-0.5, 0.5)$, $\vartheta_{Y,0} \sim N(0, 0.025X_i^2)$ and $\vartheta_{Y,1} \sim \mathcal{N}(0, 0.1\cos(X_i)^2)$, $\mathcal{U}(-0.5, 0.5)$ is the uniform distribution on $[-0.5, 0.5]$. Thus, (48) is a nonlinear system and perturbed by non-Gaussian and highly heterogeneous noises. Obviously, the objective direct effect is 1, and the objective indirect effect is 0.6 (0.4×1.5).

Following Fig. 2, the GAMN-S for (48) is designed as follows. In G_M, T is the linear component of the network. The structure associated with Z_M and X is designed as a 4-layer FNN. The second and third layers of the FNN contain 32 and 128 neurons, respectively. In G_Y, M and T formulate the linear component of the network. The structure associated with Z_Y and X are designed as a 4-layer network, the second and third layers of which also contain 32 and 128 neurons, respectively. Both Z_M and Z_Y are drawn from $\mathcal{N}((0,1)^2)$. For the output layers of G_M and G_Y, the identity function is adopted to generate a continuous outcome. The architecture of discriminators D_M and D_Y are the same except the input variables. D_Y takes X, T, M and Y for its input layer, while D_M takes X, T and M as input variables. The second and third layers of both networks contain 128 and 32 neurons, respectively. For the output layers of D_M and D_Y, sigmoid functions are adopted in order to identify real or generated sample. In G_M, D_M, G_Y and D_Y, all the hidden layers are activated by Leaky ReLU function with slope coefficient 0.2.

A total of 5000 samples are generated and used for the modeling. When applying CGANs to generate images in the conventional tasks, it is not necessary to separate the samples into training set and testing set. However, the objective of GAMN is to estimate the counterfactuals. To clearly demonstrate our method from a machine learning perspective, 80% samples are randomly selected as training set and the remaining 20% samples are selected as testing set. Our GAMN-S is first trained based on the training set and then used to estimate the treatment effects on testing set. To obtain the optimal performance, Adam algorithm with learning rate $\eta = 0.0001$ is utilized, and the GAMN-S is trained for 1500 epochs for convergence.

Fig. 5(a) shows the training process and convergence curves of estimating direct effect $\triangle_{T\to Y}$ given by (18) and indirect effect $\triangle_{T\to M\to Y}$ given by (22) on the testing set with $N_g = 1$ (for reducing computational load). The objective values are presented by black dotted lines. The estimates obtained by our model converge to the objective values after about 1200 epochs, which illustrates the effectiveness of our method. The direct effects $\triangle_{T\to Y}(j)s$ and indirect effects $\triangle_{T\to M\to Y}(j)s$ $(j = 1, 2, \ldots, N_g)$ are calculated on the whole dataset with $N_g = 1000$. The corresponding empirical distributions and 95% confidence intervals (CI) can be calculated according to the method developed in Section 3.2.1. $\triangle_{T\to Y}$ and $\triangle_{T\to M\to Y}$ are also calculated accordingly. We remark that since the samples are randomly and sequentially added into the training of our network, the estimations slightly fluctuate around the objective ones, which is inevitable in the learning process. Therefore, to improve the accuracy and stability, all the estimations of the last 100 epochs are averaged to formulate our final results. Three benchmark traditional methods (OLS-based method (OLS) [6,52], SEM with nonparametric bootstrap (SEM) [53], and Bayesian Monte Carlo method (Bayesian) [26]) and three benchmark GAN-based methods (WGAN, WGAN-GP and LSGAN) are used for comparison. It is worth noting that the network structures associated with WGAN, WGAN-GP, and LSGAN are configured to be identical to GAMN-S. To further demonstrate our method, we conduct experiments with different sample sizes (2000, 5000, and 10000) using GAMN-S. All the results are reported in Table 2.

Table 2
The estimated treatment effects by GAMN-S and other benchmark methods for system (48).

Methods	$\triangle_{T\to Y}$	CI($\triangle_{T\to Y}$)	$\triangle_{T\to M\to Y}$	CI($\triangle_{T\to M\to Y}$)
True value	1	–	0.6	–
GAMN-S	0.9887	[0.9886, 0.9888]	0.5991	[0.5989, 0.5994]
OLS	1.3062	[1.2439, 1.3685]	0.2942	[0.2387, 0.3498]
SEM	1.3062	[1.1789, 1.4245]	0.2943	[0.1824, 0.4239]
Bayesian	1.3050	[1.1823, 1.4270]	0.2955	[0.1714, 0.4190]
WGAN	0.9792	[0.9789, 0.9794]	0.5975	[0.5972, 0.5980]
WGAN-GP	0.9874	[0.9874, 0.9875]	0.5918	[0.5917, 0.5919]
LSGAN	0.9978	[0.9976, 0.9980]	0.5811	[0.5807, 0.5814]
GAMN-S(2000)	0.9840	[0.9837, 0.9843]	0.5969	[0.5964, 0.5974]
GAMN-S(10000)	0.9986	[0.9984, 0.9987]	0.5879	[0.5876, 0.5881]
GAMN-S(25000)	0.9816	[0.9815, 0.9817]	0.5994	[0.5992, 0.5996]

Note: The results presented in this table are based on $N_g = 1000$ repeated trials, the estimates and CIs are obtained through multiple averages and calculating percentiles of them.

It can be seen that both the direct and indirect effects are effectively estimated by GAMN-S and other GAN-based methods, showcasing superior performance compared to the traditional methods. As the sample size increases, the estimation accuracy improves. Moreover, in comparison to the traditional methods, GAN-based methods yield significantly improved results.

5.2. GAMN-M with simulated data

Consider the following data generation process

$$\begin{cases} M_i^1 = 0.5 + 0.2X_{1,i} + 0.8X_{2,i} + 0.5X_{1,i}X_{2,i} + 0.2X_{3,i} + 0.3X_{4,i} \\ \quad + 0.1X_{5,i} + 0.4X_{6,i} + 2T_i + \varepsilon_{M^1,i}, \\ M_i^2 = 0.5 + 0.4X_{1,i} + 0.2\left(X_{1,i}+1\right)^2 + 0.1\exp(X_{2,i}) + 0.2X_{5,i} + 0.2X_{6,i} \\ \quad + 0.3X_{7,i} + 0.3X_{8,i} + 0.5T_i + \varepsilon_{M^2,i}, \\ M_i^3 = -0.5 + 0.3X_{1,i} + 0.2X_{2,i} + 0.1X_{1,i}^3 + 0.6X_{1,i}X_{2,i} + 0.3X_{7,i} + 0.1X_{8,i} \\ \quad + 0.1X_{9,i} + 0.3X_{10,i} + T_i + \varepsilon_{M^3,i}, \\ Y_i = 2.25 - 2\sqrt{X_{1,i}+5} + \sin(1.5X_{2,i}) + 0.5M_i^1 + M_i^2 + 0.25X_{3,i} + 0.3X_{5,i} \\ \quad + 0.2X_{8,i} + 0.2X_{10,i} + 0.75T_i + \varepsilon_{Y,i}. \end{cases} \tag{49}$$

For $\forall i$, the treatment variable $T_i \sim \mathcal{B}(0.5)$, the pre-treatment covariates $X_{1,i}, X_{2,i}, \ldots, X_{10,i}$ are independently drawn from $\mathcal{N}(0,1)$. The random error terms are set as

$$\varepsilon_{M^1,i} = \vartheta_{M^1,0}^{(1-T_i)}\vartheta_{M^1,1}^{T_i}, \quad \varepsilon_{M^2,i} = \vartheta_{M^2,0}^{(1-T_i)}\vartheta_{M^2,1}^{T_i},$$
$$\varepsilon_{M^3,i} = \vartheta_{M^3,0}^{(1-T_i)}\vartheta_{M^3,1}^{T_i}, \quad \varepsilon_{Y,i} = \vartheta_{Y,0}^{(1-T_i)}\vartheta_{Y,1}^{T_i},$$

where $\vartheta_{Y,0} \sim \mathcal{N}(0, 0.01(X_{1,i}^2 + X_{2,i}^2))$, $\vartheta_{Y,0} \sim \mathcal{N}(0, 0.1\cos(5X_{1,i}X_{2,i})^2)$, $\vartheta_{M^1,0} \sim \mathcal{N}(0, 0.25^2)$, $\vartheta_{M^1,1} \sim \mathcal{U}(-0.5, 0.5)$, $\vartheta_{M^2,0} \sim 0.2t(10)$, $\vartheta_{M^2,1} \sim \mathcal{N}(0, 0.25\cos(X_{1,i})^2)$, $\vartheta_{M^3,0} \sim \mathcal{U}(-0.5, 0.5)$, $\vartheta_{M^3,1} \sim \mathcal{N}(0, 0.1X_{2,i}^2)$, and $t(10)$ represents t-distribution with a degree of freedom 10. Obviously, the objective direct effect is 0.75. The objective indirect effect through M^1 is 1 (2×0.5), through M^2 is 0.5 (0.5×1). Since M^3 is not involved in (49), the corresponding indirect effect is 0.

Following Fig. 4, the GAMN-M for (49) is designed as follows. For all the G_M^ps ($p = 1,2,3$) in mediation block, their network structures associated with Z_M^p and X are the same and designed as 4-layer FNNs. In all the G_M^ps, T is the linear component of the network. The second and third hidden layers of G_M^ps contain 32 and 128 neurons, respectively. In G_Y, $\boldsymbol{M}$ and T formulate the linear components of the network. The structure associated with Z_Y and X is a 4-layer network, the second and third layers of which also contain 32 and 128 neurons, respectively. $\boldsymbol{Z}_M^p$s and $\boldsymbol{Z}_Y$ are drawn from $\mathcal{N}((0,1)^2)$. For the output layers of G_M^p and G_Y, the identity function is adopted to generate a continuous outcome. The architecture of discriminators $D_M^p (p = 1,2,3)$ and D_Y are the same except the input variables. For $\forall p$, D_M^p takes X, T and M^p as input variables, while D_Y takes X, T, Y and three mediators (M^1, M^2, M^3) for its input layer. The second and third layers of D_M^p and D_Y contain 128 and 32 neurons, respectively. For the output layers of D_M^ps and D_Y, sigmoid functions are adopted in order to identify real or generated samples. All the hidden layers of G_M^p, D_M^p, G_Y and D_Y are formulated using Leaky ReLU function with slope coefficient 0.2. A total of 25000 samples are generated and used for the modeling. Following Section 5.1, 20000 samples are randomly selected for training and the rest 5000 samples are for testing. Our GAMN-M is first trained based on the training set and then used to estimate the treatment effects on testing set. To obtain the optimal performance, Adam with learning rate $\eta = 0.0001$ is utilized, and the training is conducted for 1500 epochs.

Fig. 5(b) shows the learning process and convergence curves of estimating direct effect $\triangle_{T\to Y}$ given by (30), total indirect effect $\triangle_{T\to \boldsymbol{M}\to Y}$ given by (34), and indirect effects through different paths ($\triangle_{T\to M^1\to Y}$, $\triangle_{T\to M^2\to Y}$ and $\triangle_{T\to M^3\to Y}$) given by (38) on the testing set with $N_g = 1$. The objective values are presented by black dotted lines. Although multiple mediators are involved, our model can achieve convergence and accurate estimations after about 300 epochs, demonstrating the effectiveness of our method for this example. The direct effects $\triangle_{T\to Y}(j)$s, indirect effects via the pth mediator $\triangle_{T\to M^p\to Y}(j)$s and total indirect effects $\triangle_{T\to \boldsymbol{M}\to Y}(j)$s ($j = 1,2,\ldots,N_g$) are calculated on the whole dataset with $N_g = 1000$. Then, $\triangle_{T\to Y}$, $\triangle_{T\to M^p\to Y}$ ($p = 1,2,3$), $\triangle_{T\to \boldsymbol{M}\to Y}$ and the corresponding 95% CI are calculated following the method in Section 3.2.2. To improve the accuracy and stability, the estimations of the last 100 epochs are averaged to formulate our final results.

Similar to Section 5.1, the benchmark traditional and GAN-based methods are used for comparison. We conduct experiments with different sample sizes (2000, 5000, and 10000) using GAMN-M. All the numerical results obtained by different methods are reported in Table 3, which demonstrates that the proposed method and other GAN-based methods significantly outperforms the traditional methods in this example.

5.3. GAMN-S with realistic data: China education panel survey

The China Education Panel Survey (CEPS) is a large-scale, nationally representative and longitudinal survey, which is conducted by the National Survey Research Center (NSRC) at Renmin University of China. Documenting educational processes and transitions by which students progress through various educational stages, the CEPS aims to explain the linkages between individuals' educational outcomes and multiple contexts of families, school processes, communities and social structure and further study the effects of educational outcomes. The baseline survey of CEPS is completed in the 2013–2014 academic year. A stratified, multistage sampling design with probability proportional to size is used to randomly select a school-based, nationally representative sample of 19487 students in 438 classrooms of 112 schools in 28 county-level units in mainland China (http://ceps.ruc.edu.cn/).

In this study, the interested outcome (Y) is the total score of the sampled students on core subjects (Chinese, mathematics, and English).

Table 3
The estimated treatment effects by GAMN-M and other benchmark methods for system (49).

Methods	$\triangle_{T\to Y}$ CI($\triangle_{T\to Y}$)	$\triangle_{T\to \boldsymbol{M}\to Y}$ CI($\triangle_{T\to \boldsymbol{M}\to Y}$)	$\triangle_{T\to M^1\to Y}$ CI($\triangle_{T\to M^1\to Y}$)	$\triangle_{T\to M^2\to Y}$ CI($\triangle_{T\to M^2\to Y}$)	$\triangle_{T\to M^3\to Y}$ CI($\triangle_{T\to M^3\to Y}$)
True value	0.75	1.5	1.0	0.5	0
GAMN-M	0.7448 [0.7447, 0.7449]	1.4841 [1.4837, 1.4844]	0.9916 [0.9914, 0.9918]	0.4922 [0.4919, 0.4924]	0.0002 [0.0001, 0.0003]
OLS	0.9139 [0.8394, 0.9915]	1.3167 [1.2377, 1.4282]	0.8747 [0.7925, 0.9679]	0.4317 [0.3922, 0.4879]	0.0133 [−0.0283, 0.0575]
SEM	0.9187 [0.8555, 0.9820]	1.3080 [1.2471, 1.3689]	0.8720 [0.8205, 0.9234]	0.4228 [0.3971, 0.4485]	0.0132 [−0.0068, 0.0333]
Bayesian	0.9178 [0.8538, 0.9851]	1.3087 [1.2411, 1.3730]	0.8717 [0.7942, 0.9475]	0.4229 [0.3983, 0.4496]	0.0136 [−0.0172, 0.0427]
WGAN	0.7438 [0.7437, 0.7439]	1.4872 [1.4868, 1.4876]	0.9822 [0.9819, 0.9824]	0.5019 [0.5014, 0.5022]	0.0032 [0.0031, 0.0033]
WGAN-GP	0.7380 [0.7380, 0.7381]	1.4966 [1.4964, 1.4969]	0.9918 [0.9917, 0.9919]	0.5047 [0.5044, 0.5049]	0.0001 [0.0001, 0.0001]
LSGAN	0.7486 [0.7486, 0.7486]	1.4904 [1.4901, 1.4907]	0.9921 [0.9919, 0.9923]	0.4976 [0.4973, 0.4978]	0.0007 [0.0007, 0.0007]
GAMN-M(2000)	0.8007 [0.8006, 0.8008]	1.4102 [1.4092, 1.4109]	0.9485 [0.9483, 0.9486]	0.4732 [0.4722, 0.4740]	−0.0115 [−0.0116, −0.0114]
GAMN-M(5000)	0.7512 [0.7512, 0.7512]	1.4610 [1.4605, 1.4616]	0.9827 [0.9826, 0.9829]	0.4827 [0.4821, 0.4832]	−0.0044 [−0.0044, −0.0044]
GAMN-M(10000)	0.7457 [0.7457, 0.7457]	1.4928 [1.4923, 1.4932]	0.9893 [0.9890, 0.9896]	0.5047 [0.5043, 0.5050]	−0.0012 [−0.0012, −0.0012]

Note: The results presented in this table are based on $N_g = 1000$ repeated trials, the estimates and CIs are obtained through multiple averages and calculating percentiles of them.

Table 4
Descriptions of variables in the CEPS dataset.

Factor	Variable	Description
Academic performance	Y	Total score of Chinese, math, and English
Cognitive ability	M	Standardized cognitive ability test scores
Parental involvement	T	=1 if greater than average, =0 otherwise
Student's gender	X_1	=1 if male, =0 otherwise
Grade	X_2	=1 if grade 9, =0 grade 7
Ethnic nationality	X_3	=1 if Han nationality, =0 otherwise
Location of Hukou	X_4	=1 if located in suburban district, =0 otherwise
Nearsightedness	X_5	=1 if short-sighted, =0 otherwise
Only-children	X_6	=1 if the only child of family, =0 otherwise
Boarding status	X_7	=1 if live in school at night, =0 otherwise
Mother's education	X_8	=1 if complete senior high school, =0 otherwise
Father's education	X_9	=1 if complete senior high school, =0 otherwise
Financial condition	X_{10}	=1 if moderate or above moderate income, =0 otherwise
Head teacher's gender	X_{11}	=1 if male, =0 otherwise
Teaching experience	X_{12}	Years of teaching experience

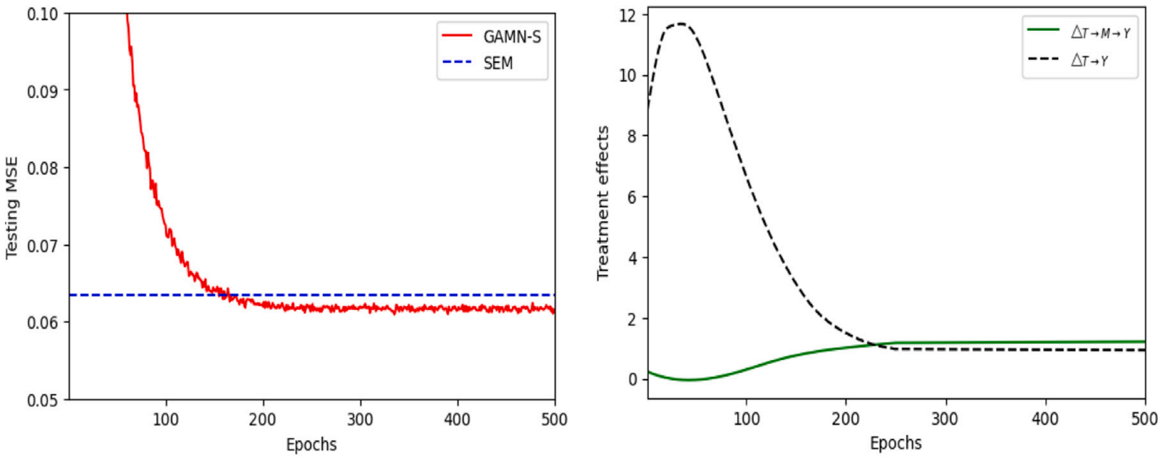
The treatment variable (T) is home-based parental involvement and is measured from three perspectives. The first one is "parent tutoring and supervision", which includes whether the parents have checked children's homework this semester ($Q1$), and whether parents have tutored their children in the last week before exams ($Q2$). The second one is "proactively talking with children", which includes whether the parents discuss school life topics with their children, such as things happening at school ($Q3$), children's relationship between friends ($Q4$), teachers ($Q5$), worries and troubles ($Q6$), and children's mood ($Q7$). The third one is "time spending with children", which includes whether the parents do accompanying activities with children more than once a month, such as having dinner ($Q8$), reading books ($Q9$), watching TV ($Q10$), playing sports ($Q11$), visiting museums or zoos ($Q12$), and going out to watch movies/shows/sports games ($Q13$). For each parent, the number of answering 'yes' in $Q1$ to $Q13$ are counted and used to quantify the treatment T. $T = 1$, if the number of 'yes' is larger than 6. Otherwise, $T = 0$. In this setting, $T = 1$ implies the home-based involvement of parents is greater than average.

The mediator M here is cognitive ability, which is measured by standardized test scores for students' logical thinking and problem-solving skills. The pre-treatment covariates include students' demographic characteristics, family financial conditions, and head teacher's characteristics. The covariates are summarized in Table 4.

In this experiment, we aim at modeling the direct effect of parental involvement (T) on student academic performance (Y) and the mediating role of the cognitive ability (M) playing in this system. Following Fig. 2, the GAMN-S for this example is designed as follows. The nonlinear parts of G_M (G_Y) are designed as 4-layer FNNs with the same structure. The input variables are set as Z_M (Z_Y), and X_i ($i = 1, 2, \ldots, 12$). $\boldsymbol{Z}_M \sim \mathcal{N}((0,1)^2)$ and $\boldsymbol{Z}_Y \sim \mathcal{N}(((0,1)^2))$. T is the linear component of G_M. T and M formulate the linear component of G_Y. For the hidden layers, the second and third layers contain 32 and 128 neurons, respectively. For the discriminators, both D_M and D_Y are also designed as 4-layer FNNs with the same structure. D_M takes X, T and M as input variables, and D_Y takes X, T, M and Y for its input layer. The second and third hidden layers contain 64 and 32 neurons, respectively. All of the hidden layers in G_M, D_M, G_Y and D_Y are set to be activated by Leaky ReLU function with slope coefficient 0.2.

After removing the observations with missing values, 16,394 samples are selected for our analysis. A total of 13,116 (80% of the selected dataset) samples are randomly chosen for training, and the remaining 3278 samples are for testing. Adam with learning rate 0.00001 is used as the optimizer for the training. The GAMN-S is trained for 500 epochs. OLS, SEM and Bayesian method are also used for comparison.

Theocratically, all the traditional methods and our GAMN-S can be available for modeling. However, since the true model and the



(a) (b)

Fig. 6. (a) Displays the convergence curve of the testing MSE by GAMN-S. The red solid line illustrates the testing MSE by GAMN-S for each epoch, while the blue dotted line represents the testing MSE obtained through SEM. The testing MSE by GAMN-S converges after approximately 250 epochs and, moreover, remains lower than the testing MSE achieved by SEM. (b) displays the convergence curves of the estimations of direct effect and indirect effects. The black dotted line represents the estimation of the direct effect, and the green solid line represents the estimation of the direct effect for each epoch. The estimations reach their convergence after approximately 250 epochs, and the convergence points provide the final estimated values for both the direct and indirect effects.

underlying true treatment effects are unknown for realistic data, it is quite difficult to appraise the results obtained by different methods. Noticing that all the traditional methods and our GAMN can provide the filtered factuals (observations) and out-of-sample predictions for M_i and Y_i, we propose to first appraise these methods by comparing their predictions on the testing set. Suppose that $G_M(\cdot;\hat{\theta}_{G_M(l)})$ and $G_Y(\cdot;\hat{\theta}_{G_Y(l)})$ are the generators trained at lth epoch using the training set. $\hat{\theta}_{G_M(l)}$ and $\hat{\theta}_{G_Y(l)}$ are corresponding network parameters. For the ith sample in the testing set, we have

$$\hat{M}_i(l) = G_M(\boldsymbol{Z}_{M_i}(l), T_i, \boldsymbol{X}_i; \hat{\theta}_{G_M(l)}), \quad \hat{Y}_i(l) = G_Y(\boldsymbol{Z}_{Y_i}(l), T_i, \boldsymbol{X}_i, \hat{M}_i(l); \hat{\theta}_{G_Y(l)}), \tag{50}$$

where $\hat{M}_i(l)$ and $\hat{Y}_i(l)$ are the predictions of M_i and Y_i for the lth epoch, respectively. To save the computational cost, N_g is set as 1. Then, the mean square error (MSE) on the testing set for the lth epoch is defined as

$$MSE(l) = \frac{1}{3278}\sum_{i=1}^{3278}(\hat{Y}_i(l) - Y_i)^2, \tag{51}$$

where Y_is correspond to the samples in the testing set. Accordingly, the MSEs by the traditional methods can also be trivially defined and calculated.

Fig. 6(a) shows the prediction performance in terms of the testing MSEs given by (51). Considering that the prediction results using the three traditional models are similar, we only compare GAMN-S with the SEM model. Our method produces higher prediction accuracy than the SEM model, implying that the proposed GAMN can approximate the underlying reality more efficiently from a machine learning perspective. Thus, the treatment effects inferred by our GAMN are more convincing compared with the traditional methods. Fig. 6(b) presents the treatment effects calculated on the testing set, which shows that the estimations achieve convergence after about 300 epochs. The direct effects $\triangle_{T\to Y}(j)s$ and indirect effects $\triangle_{T\to M\to Y}(j)s$ $(j = 1, 2, \ldots, N_g)$ calculated on the whole dataset with $N_g = 1000$. The corresponding empirical distributions and 95% CI can be calculated according to the method developed in Section 3.2.1. $\triangle_{T\to Y}$ and $\triangle_{T\to M\to Y}$ are also calculated accordingly. Similar to Section 5.1, to improve the accuracy and stability, the results of the last 100 epochs are averaged to formulate our final estimations.

Table 5
Estimated treatment effects by GAMN-S and other benchmark methods for CEPS dataset.

Methods	$\triangle_{T\to Y}$	CI($\triangle_{T\to Y}$)	$\triangle_{T\to M\to Y}$	CI($\triangle_{T\to M\to Y}$)
GAMN-S	0.9358	[0.9289, 0.9426]	1.1978	[1.1909, 1.2051]
OLS	1.4036	[0.6647, 2.1426]	1.5252	[1.2235, 1.8268]
SEM	1.404	[0.660, 2.09]	1.525	[1.241, 1.83]
Bayesian	1.394	[0.679, 2.14]	1.517	[1.200, 1.83]
WGAN	0.9602	[0.8974, 1.0191]	1.2775	[1.1966, 1.3636]
WGAN-GP	1.0417	[1.0389, 1.0448]	1.2394	[1.2363, 1.2425]
LSGAN	0.9395	[0.8982, 0.9908]	1.1751	[1.1171, 1.2289]

Note: The results presented in this table are based on $N_g = 1000$ repeated trials, the estimates and CIs are obtained through multiple averages and calculating percentiles of them.

All the numerical results obtained by our method and the other benchmark methods are reported in Table 5. Our estimation results differ from those of the other methods and probably offer a more reliable mediation analysis compared with the traditional methods given its smaller MSEs and shorter length of CIs.

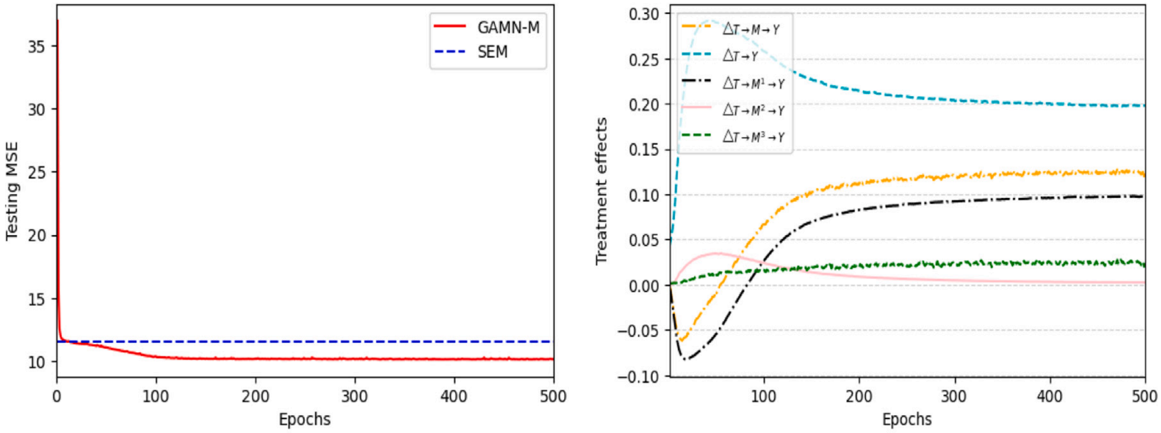
5.4. GAMN-m with realistic data: China health and nutrition survey

The China Health and Nutrition Survey (CHNS), an international collaborative project between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute for Nutrition and Health at the Chinese Center for Disease Control and Prevention, is a large-scale longitudinal study (https://www.cpc.unc.edu/projects/china). The survey was started in 1989 and conducted ten times until 2015. CHNS was designed to investigate how social and economic transformations in Chinese society affect the health and nutritional status of the population.

As urbanization brings dietary changes and increases the risk of obesity through people's intake, this example focuses on how urbanization affects people's health conditions by evaluating the mediating role of primary nutrients. The interested dependent variable (Y) is the body mass index (BMI), which is commonly used to measure the obesity of a person. The treatment (T) is urbanization. $T = 1$ if the individual lived in urban area during the survey year, while $T = 0$ for those in rural areas. Three main nutrients in food (in grams): carbohydrate (M^1), fat (M^2) and protein (M^3), are considered mediators, and the corresponding data are collected from the participants' 24-h recall of

Table 6
Descriptions variables in the CHNS dataset.

Factor	Variable	Description
BMI	Y	ratio of weight (kg) to the square of height (m^2)
Carbohydrate	M_1	Daily intake of carbohydrate (grams)
Fat	M_2	Daily intake of fat (grams)
Protein	M_3	Daily intake of protein (grams)
Urbanization	T	=1 if live in urban, =0 otherwise
Gender	X_1	=1 if male, =0 otherwise
Marital status	X_2	=1 if married, =0 otherwise
Age	X_3	age of the individual
Education	X_4	=1 if possess upper middle school degree, =0 otherwise
Smoking	X_5	=1 if smokes cigarettes, =0 otherwise
Alcohol	X_6	=1 if drinks alcohol, =0 otherwise
Chronic disease	X_7	=1 if diagnosed with chronic disease, =0 otherwise
	X_8	=1 if in 1991,1993 or 1997, =0 otherwise
Year	X_9	=1 if in 2000,2004,2006 or 2009, =0 otherwise
	X_{10}	=1 if in 2011, =0 otherwise



(a) (b)

Fig. 7. (a) Displays the convergence curve of the testing MSE by GAMN-M. The red solid line illustrates the testing MSE by GAMN-M for each epoch, while the blue dotted line represents the testing MSE obtained through SEM. The testing MSE by GAMN-M converges after approximately 200 epochs and, and remains lower than the testing MSE achieved by SEM. (b) displays the convergence curves of the estimations of direct effect and indirect effects. The blue line represents the estimation of the direct effect, and the yellow line represents the estimation of the total direct effect for each epoch. Additionally, the black line, yellow line, and green line represent the estimations of the indirect effects through M^1, M^2, and M^3, respectively. The estimations reach their convergence after approximately 200 epochs, and the convergence points provide the final estimated values for both the direct and indirect effects.

food consumption in the past three days. The pre-treatment covariates include individuals' demographic characteristics. The covariates are summarized in Table 6.

GAMN-M is utilized for this experiment. For each G_M^p $(p = 1, 2, 3)$ in the mediation block, the structure associated with Z_M^p $(p = 1, 2, 3)$ and X is set as a 4-layer FNN. T formulates the linear component. The second and third layers of G_M^p contain 32 neurons and 128 neurons, respectively. In G_Y, M and T formulate the linear component. The structure associated with Z_Y and X is designed as the same as G_M^p. $\boldsymbol{Z}_M^p$s and $\boldsymbol{Z}_Y$ are drawn from $N((0, 1)^2)$. In order to generate continuous outcome, identity functions are adopted to the output layers of G_M^ps and G_Y. D_M^ps and D_Y are 4-layer FNNs. D_M^ps take X, T and M^ps as input variables, while D_Y takes X, T, Y and three mediators (M^1, M^2, M^3) for its input layer. The second and third layers of D_M^ps and D_Y contain 128 and 32 neurons, respectively. The output layers of D_M^k and D_Y are activated by sigmoid functions. Leaky ReLU with slope coefficient 0.2 is used to activate all the hidden layers of the networks.

A total of 102,575 observations were collected from year 1991 to 2011. After removing the samples with missing values, 80,341 samples are selected for modeling, among which 64,272 (80% of the dataset) samples are randomly chosen for training, and the remaining 16,069 samples are for testing. Adam algorithm with learning rate 0.0001 is utilized for training. The GAMN-M model is trained for 500 epochs.

Fig. 7(a) presents the testing MSEs by different methods, showing that better predicting results are obtained by our method. As discussed in Section 5.3, it implies that our method can approximate the underlying reality more efficiently and offer more reliable mediation analysis. Fig. 7(b) presents the estimations of treatment effects at each epoch. The estimations achieve convergence after about 400 epochs, which illustrates the feasibility of our GAMN. The generated direct effects and indirect effects are calculated on the whole dataset with $N_g = 1000$. The empirical distributions and 95% CI of the treatment effects can be calculated accordingly. The results of the last 100 epochs are averaged to formulate our final estimations.

All the numerical results obtained by our method and the other benchmark methods are reported in Table 7. Our results suggest a more significant total mediation effect and a positive mediation effect through carbohydrate, which is more conformable to the reality of Chinese society [54].

5.5. Further discussions on the implementation of our method

Implementation challenges. While our method has demonstrated its effectiveness in these examples, the implementation of our method comes with certain challenges. Due to the involvement of deep neural networks, GAMN requires a larger amount of data for training compared to the traditional models. When dealing with real-world problems

Table 7
The estimated treatment effects by GAMN-M and other benchmark methods for CHNS dataset.

Methods	$\triangle_{T\to Y}$	$\triangle_{T\to \boldsymbol{M}\to Y}$	$\triangle_{T\to \boldsymbol{M}_1\to Y}$	$\triangle_{T\to \boldsymbol{M}_2\to Y}$	$\triangle_{T\to \boldsymbol{M}_3\to Y}$
	CI($\triangle_{T\to Y}$)	CI($\triangle_{T\to \boldsymbol{M}\to Y}$)	CI($\triangle_{T\to \boldsymbol{M}_1\to Y}$)	CI($\triangle_{T\to \boldsymbol{M}_2\to Y}$)	CI($\triangle_{T\to \boldsymbol{M}_3\to Y}$)
GAMN-M	0.1964	0.1227	0.0961	0.0027	0.0239
	[0.1963, 0.1965]	[0.1225, 0.1229]	[0.0960, 0.0962]	[0.0026, 0.0028]	[0.0238, 0.0240]
OLS	0.223	−0.026	−0.156	0.087	0.039
	[0.172, 0.272]	[−0.065, 0.010]	[−0.211, −0.092]	[0.061, 0.110]	[0.031, 0.046]
SEM	0.1999	0.0315	−0.0201	0.0257	0.0259
	[0.1483, 0.2515]	[0.0179, 0.0451]	[−0.0312, −0.0090]	[0.0197, 0.0316]	[0.0208, 0.0310]
Bayesian	0.2007	0.0314	−0.0202	0.0258	0.0259
	[0.1450, 0.2520]	[0.0155, 0.0479]	[−0.0340, −0.0056]	[0.0194, 0.0324]	[0.0209, 0.0314]
WGAN	0.2043	0.1213	0.0932	0.0024	0.0256
	[0.2007,0.2080]	[0.1185,0.1244]	[0.0924,0.0941]	[0.0023,0.0025]	[0.0245,0.0267]
WGAN-GP	0.2034	0.1346	0.1031	0.0033	0.0282
	[0.1998, 0.2065]	[0.1343, 0.1349]	[0.1029, 0.1033]	[0.0031, 0.0035]	[0.0280, 0.0284]
LSGAN	0.1825	0.1299	0.1032	0.0039	0.0228
	[0.1819,0.1833]	[0.1254,0.1348]	[0.0997,0.1106]	[0.0034,0.0043]	[0.0219,0.0231]

Note: The results presented in this table are based on $N_g = 1000$ repeated trials, the estimates and CIs are obtained through multiple averages and calculating percentiles of them.

where the dataset is limited in size, the model's estimation accuracy may be compromised, and it might even fail to converge. This is particularly relevant in specific topics, such as medical outcome evaluation [55], impact of behavior interventions on mental health [56]. In these topics, mediation analysis is commonly needed, typically involving survey data with limited sample sizes. This limitation adds challenges to the implementation of GAMN. Additionally, even with an ample dataset, the training process of GAN-based models is more time consuming and computationally intensive compared to the traditional models. Fine-tuning the hyperparameters is also necessary to achieve the best learning performance.

Computational complexity. The computational complexity of GANs depends on several factors, including the network architecture, dataset size, and data dimensionality, as well as the intricacy of data patterns [57]. GANs are commonly developed for image processing tasks where both the input and output of the network are images. In such cases, the dimensions of the network's input and output layers tend to be very high. Image data itself is usually characterized by complex patterns, especially in tasks like image-to-image translation, which require deep neural network architectures for effective handling. These factors contribute to the complexity of GAN architectures, involving a large number of parameters, and consequently demanding a substantial volume of training samples. In contrast, CMA is primarily applied in fields such as social sciences, psychology, and medicine. These domains typically involve data with underlying patterns that are less complex and datasets that are much smaller compared to image-related problems. In GAMN, both the input and output of the network consist of low-dimensional variables, especially with the generator's output being one-dimensional. The crucial partially linear design in GAMN further reduce the network complexity. Therefore, our GAMN, designed specifically for mediation problems, exhibits significantly lower dimensions in terms of input and output variables, network depth, and complexity when compared to conventional GANs designed for image tasks. With similar optimization objectives and training algorithms (Adam), the computational complexity of GAMN is substantially lower than that of GANs used for image-related tasks. The training of GANs also benefits from highly developed computing frameworks such as PyTorch, and our proposed GAN-based approach is computationally feasible for CMA.

6. Conclusion

GANs have achieved remarkable success in image tasks. However, developing a generative learning approach for mediation analysis remains an unexplored but promising area. This paper proposes two novel GAN-based mediation models by reinterpreting the CMA problem from a generative learning perspective. Our carefully designed network architecture and novel encoding scheme for complex noise enable our models to provide more accurate counterfactual predictions than existing benchmark methods when dealing with complex data patterns. Furthermore, by efficiently inferring individual direct/indirect treatment effects from the estimated counterfactuals, our models lead to more promising CMA results. Numerical examples are presented. In the two examples with artificial data, our approach yielded highly precise estimation results. In contrast, the traditional methods exhibit estimation errors exceeding 30% for both direct and indirect effects in the single-mediator case, and nearly 20% in the multi-mediator case. In the benchmark real-world data examples, our method outperforms the traditional methods by 15% in out-of-sample predictions. This suggests that the treatment effects deduced through GAMN are more compelling compared to the traditional methods from a machine learning perspective. The encouraging quantitative results further illustrate the effectiveness and efficiency of our method. Our study also represents a significant step toward the development of effective approaches for deploying generative learning methods in mediation problems.

Despite the efficiency, the proposed mediation method still has several limitations. First, the architecture of our GAMN is developed for single-level mediation problems. The architecture of GAMN is not well-suited for handling multi-layer mediation problems with complex causal pathways. To address the multi-layer mediation issues, the network structure must be redesigned. Second, high-dimensional mediation problems have been a focal point of research in both the statistical and machine learning communities. The central challenge in these problems revolves around variable selection and dimension reduction with respect to the sparsity of the mediator variables. However, our optimization schemes and model structure are not available for high-dimensional settings and cannot be trivially extended to address the high-dimensional problems. Third, a significant advantage of our approach lies in employing deep neural networks to model the distribution of noise, allowing for coupling between random terms and covariates. However, when the underlying ture relationships between variables and the characteristics of random terms are relatively straightforward within the data, our method may face a risk of overfitting. Consequently, in such cases, our approach might not consistently outperform traditional methods. Therefore, our future research will be dedicated to developing GAN-based models for addressing high-dimensional mediation problems and multi-layer mediation problems, leveraging the GAN-based approach to conduct empirical studies and address important practical challenges.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jiaming Zhang: Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Yiqi Lin:** Writing –

original draft, Investigation, Formal analysis. **Xinyuan Song:** Writing – review & editing, Validation, Resources, Funding acquisition. **Hanwen Ning:** Writing – original draft, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by Humanities and Social Science Fund of Ministry of Education, China under project 23YJC910011, National Social Science Foundation of China under Project 19BTJ025, Innovation and Talent Base for Digital Technology and Finance (B21038), General Research Fund 14303622 from the Research Grant Council of the Hong Kong Special Administrative Region, Fundamental Research Funds for the Central Universities (Zhongnan University of Economics and Law), China under Project 2722022BY020, 2722022BQ044, and 2722023EJ002.

Appendix

By the argument in the main text, we already have

$$\mathbb{D}_{TV}(p_{X,Y,T}, p_{X,\hat{G}(Z,T,X),T}) \lesssim \Delta_1 + \Delta_2 + \Delta_3, \tag{52}$$

where $\Delta_1 = \sup_D \mathcal{L}(\hat{G}, D) - \sup_{D \in D} \mathcal{L}(\hat{G}, D)$, $\Delta_2 = \sup_{D\in D} \sup_{G\in\mathcal{G}} |\mathcal{L}(G, D) - \mathcal{L}_n(G, D)|$, and $\Delta_3 = \sup_D \mathcal{L}(\bar{G}, D) - \sup_D \mathcal{L}(G^*, D)$. First, the standard empirical process argument in [49] (Lemma B.2) provides the convergence rate of estimation error Δ_2:

$$\Delta_2 \lesssim \mathcal{O}(n^{-\frac{2}{2+\bar{d}+m}} + n^{-\frac{2}{2+\bar{d}+q}}) = \mathcal{O}(n^{-\frac{2}{3+\bar{d}}}), \tag{53}$$

where $\bar{d}$ is the sum of dimension of X and T, m is the dimension of outsource noise Z and $q = 1$ is the dimension of Y. The term simplicity tells the fact that the higher dimension we use in Y, X, T, Z, the slower convergence rate of estimation error we face. A low dimensional structure in X, T, Y is preferred. Followed by [49] (Lemma B.3), as n goes to infinity, we have

$$\mathbb{E}_{X,Y,Z}\, \Delta_1 \to 0. \tag{54}$$

Then, we turn to control the approximation error Δ_3 in $\mathcal{G}$. Note that $\bar{G}$ is any element in $\mathcal{G}$. For the fixed $\bar{G}$, using the optimality in D [23], we have the following result:

$$\tilde{D}(\varsigma) = \arg\max_D \mathcal{L}(\bar{G}, D) = \frac{p_{X,Y,T}(\varsigma)}{p_{X,\bar{G},Y}(\varsigma) + p_{X,Y,T}(\varsigma)}, \tag{55}$$

where $\varsigma \in \mathcal{X} \times \mathcal{Y} \times \mathcal{T}$. Then we are able to derive

$$\begin{aligned}\Delta_3 &= \sup_D \mathcal{L}(\bar{G}, D) - \sup_D \mathcal{L}(G^*, D) \\ &= \mathcal{L}(\bar{G}, \frac{p_{X,Y,T}}{p_{X,\bar{G},Y} + p_{X,Y,T}}) - \mathcal{L}(G^*, \frac{p_{X,Y,T}}{p_{X,G^*,Y} + p_{X,Y,T}}) \\ &\equiv \tilde{L}(\bar{G}) - \tilde{L}(G^*). \end{aligned} \tag{56}$$

Since $\mathcal{L}$ is continuous and $\tilde{L}$ is a composite function of $\mathcal{L}$ and optimal discriminator, it turns out $\tilde{L}$ maintains the continuity. Note (56) holds for any $\bar{G} \in \mathcal{G}$, that $\Delta_3 \to 0$ for some $\bar{G} \in \mathcal{G}$ suffices to show existence of $\bar{G} \in \mathcal{G}$ can closed to G^* with any given tolerance. However, our neural architecture is not a standard fully connected network, hence the standard approximation result is not applicable. Consider the following decomposition,

$$\begin{aligned}\inf_{\bar{G}\in\mathcal{G}} \|\bar{G} - G^*\| &\le \inf_{\bar{G}\in\mathcal{G}} \Big\{ \|\bar{G} - G_{FNN}\| + \|G_{FNN} - G^*\| \Big\} \\ &\le \inf_{\bar{G}\in\mathcal{G}} \|\bar{G} - G_{FNN}\| + \mathcal{O}\Big(\frac{\log n}{n^{2+d+m}}\Big), \end{aligned} \tag{57}$$

where G_{FNN} is a fully connected ReLU network with a specific setup considered in [49, Lemma B.1], and its convergence rate is provided in the second inequality above.

In turn, we denote $G_{FNN} = G_1(Z, X) + G_2(T) + G_3(Z, T, X)$, where G_1 and G_2 are the parts of G_{FNN} that only contain information from (Z, X) and T, respectively, and G_3 only contains its remaining interaction information between (Z, X) and T. Therefore, we consider a specific $\bar{G} \in \mathcal{G} = G_1(Z, X) + \beta T$, remaining a linear part βT to be optimized. Then, we have

$$\epsilon_{lienar} \equiv \inf_{\bar{G}\in\mathcal{G}} \|\bar{G} - G_{FNN}\| \le \inf_{\beta} \|\beta T - G_1(T) - G_3(Z, T, X)\|, \tag{58}$$

as a measure of linear projection error. Hence, only using fully connected ReLU network, $\Delta_3 = o(1)$, but when $\mathbb{E}_{Z,T,X}\epsilon_{linear} = o(1)$, by the continuity of $\tilde{L}$, we still have Δ_3 with practical solution in $\mathcal{G}$ to varnish. Combined with the aforementioned three varnished results, we have

$$\mathbb{E}_{X,T,Y,Z}\, \mathbb{D}_{TV}(p_{X,Y,T}, p_{X,\hat{G}(Z,T,X),T}) \lesssim \mathbb{E}_{X,T,Y,Z}\,(\Delta_1 + \Delta_2 + \Delta_3) \to 0. \tag{59}$$

It further implies the weak convergence of the joint distribution, $(X, G, T) \xrightarrow{d} (X, Y, T)$. By the Portmanteau Theorem and the weak law of large number, it follows

$$\begin{aligned}\hat{\beta} &= \frac{1}{N_g} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_g} \Big[\hat{G}(Z_j, X_i, T = t_1) - \hat{G}(Z_j, X_i, T = t_0)\Big] \\ &\xrightarrow{p} \partial t_0 \mathbb{E}(Y|X, T = t_0) \end{aligned} \tag{60}$$

The proof is completed. ◆

References

[1] K.J. Preacher, Advances in mediation analysis: A survey and synthesis of new developments, Ann. Rev. Psychol. 66 (2015) 825–852.
[2] T.J. VanderWeele, Mediation analysis: A practitioner's guide, Ann. Rev. Public Health 37 (2016) 17–32.
[3] T.Q. Nguyen, I. Schmid, E.A. Stuart, Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn, Psychol. Methods 26 (2) (2021) 255.
[4] B. Shi, C. Choirat, B.A. Coull, T.J. VanderWeele, L. Valeri, CMAverse: A suite of functions for reproducible causal mediation analyses, Epidemiology 32 (5) (2021) e20–e22.
[5] V. Celli, Causal mediation analysis in economics: Objectives, assumptions, models, J. Econ. Surv. 36 (1) (2022) 214–234.
[6] M.J. Valente, J.J. Rijnhart, H.L. Smyth, F.B. Muniz, D.P. MacKinnon, Causal mediation programs in R, M plus, SAS, SPSS, and stata, Struct. Equ. Model.: Multidisc. J. 27 (6) (2020) 975–984.
[7] R.B. Kline, Principles and Practice of Structural Equation Modeling, Guilford publications, 2023.
[8] C. Zheng, X.-H. Zhou, Causal mediation analysis in the multilevel intervention and multicomponent mediator case, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (3) (2015) 581–615.
[9] M. Miočević, A tutorial in Bayesian mediation analysis with latent variables, Methodology 15 (4) (2019) 137–146.
[10] X. Zhou, X. Song, Mediation analysis for mixture Cox proportional hazards cure models, Stat. Methods Med. Res. 30 (6) (2021) 1554–1572.
[11] Y. Song, X. Zhou, J. Kang, M.T. Aung, M. Zhang, W. Zhao, B.L. Needham, S.L. Kardia, Y. Liu, J.D. Meeker, et al., Bayesian sparse mediation analysis with targeted penalization of natural indirect effects, J. R. Stat. Soc. Ser. C. Appl. Stat. 70 (5) (2021) 1391–1412.
[12] R. Sun, X. Zhou, X. Song, Bayesian causal mediation analysis with latent mediators and survival outcome, Struct. Equ. Model.: Multidisc. J. 28 (5) (2021) 778–790.
[13] J.J. Rijnhart, S.J. Lamp, M.J. Valente, D.P. MacKinnon, J.W. Twisk, M.W. Heymans, Mediation analysis methods used in observational research: A scoping review and recommendations, BMC Med. Res. Methodol. 21 (1) (2021) 1–17.
[14] J. Zhang, Z. Li, X. Song, H. Ning, Deep tobit networks: A novel machine learning approach to microeconometrics, Neural Netw. 144 (2021) 279–296.

[15] T.-H. Tsai, Y.-L. Chen, S.S.-F. Gau, Relationships between autistic traits, insufficient sleep, and real-world executive functions in children: A mediation analysis of a national epidemiological survey, Psychol. Med. 51 (4) (2021) 579–586.

[16] K. Kang, D. Pan, X. Song, A joint model for multivariate longitudinal and survival data to discover the conversion to Alzheimer's disease, Stat. Med. 41 (2) (2022) 356–373.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[18] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, B. Guo, Styleswin: Transformer-based GAN for high-resolution image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11304–11314.

[19] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[20] J. Pan, J. Dong, Y. Liu, J. Zhang, J. Ren, J. Tang, Y.-W. Tai, M.-H. Yang, Physics-based generative adversarial models for image restoration and beyond, IEEE Trans. Pattern Anal. Mach. Intell. 43 (7) (2020) 2449–2462.

[21] S. Palsson, E. Agustsson, R. Timofte, L. Van Gool, Generative adversarial style transfer networks for face aging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2084–2092.

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[23] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[24] K. Wang, Maximum likelihood analysis of linear mediation models with treatment–mediator interaction, psychometrika 84 (3) (2019) 719–748.

[25] D. Gunzler, T. Chen, P. Wu, H. Zhang, Introduction to mediation analysis with structural equation modeling, Shanghai Arch. Psychiat. 25 (6) (2013) 390–394.

[26] B.B. Yimer, M. Lunt, M. Beasley, G.J. Macfarlane, J. McBeth, BayesGmed: An R-package for Bayesian causal mediation analysis, Plos one 18 (6) (2023) e0287037.

[27] S. Jackman, Bayesian Analysis for the Social Sciences, John Wiley & Sons, 2009.

[28] K. Dong, X. Dong, X. Ren, Can expanding natural gas infrastructure mitigate CO2 emissions? analysis of heterogeneous and mediation effects for China, Energy Econ. 90 (2020) 104830.

[29] Y. Dou, F. Chen, Z. Kong, K. Dong, Re-estimating the trade openness–carbon emissions nexus: A global analysis considering nonlinear, mediation, and heterogeneous effects, Appl. Econ. (2023) 1–16.

[30] Q. Zhang, High-dimensional mediation analysis with applications to causal gene identification, Stat. Biosci. (2021) 1–20.

[31] J. Pearl, Direct and indirect effects, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 373–392.

[32] Y.-F. Yung, M. Lamm, W. Zhang, et al., Causal mediation analysis with the CAUSALMED procedure, in: Proceedings of the SAS Global Forum 2018 Conference, SAS Institute Inc, Cary, NC, 2018, pp. 1991–2018.

[33] Y.-T. Huang, H.-I. Yang, Causal mediation analysis of survival outcome with multiple mediators, Epidemiol. (Cambridge, Mass.) 28 (3) (2017) 370–378.

[34] Y. Li, Q. Wang, J. Zhang, L. Hu, W. Ouyang, The theoretical research of generative adversarial networks: An overview, Neurocomputing 435 (2021) 26–41.

[35] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems. Vol. 29, 2016.

[36] Y. Liu, Q. Li, Q. Deng, Z. Sun, M.-H. Yang, Gan-based facial attribute manipulation, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[37] S. Jain, G. Seth, A. Paruthi, U. Soni, G. Kumar, Synthetic data augmentation for surface defect detection and classification using deep learning, J. Intell. Manuf. (2022) 1–14.

[38] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[39] E. Batziou, K. Ioannidis, I. Patras, S. Vrochidis, I. Kompatsiaris, Artistic neural style transfer using CycleGAN and FABEMD by adaptive information selection, Pattern Recognit. Lett. 165 (2023) 55–62.

[40] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 994–1003.

[41] M. Zhu, S. Gong, Z. Qian, L. Zhang, A brief review on cycle generative adversarial networks, in: The 7th IIAE International Conference on Intelligent Systems and Image Processing, ICISIP, 2019, pp. 235–242.

[42] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems. Vol. 30, 2017.

[44] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.

[45] G.-J. Qi, Loss-sensitive generative adversarial networks on lipschitz densities, Int. J. Comput. Vis. 128 (5) (2020) 1118–1140.

[46] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: Algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3313–3332.

[47] A. Jabbar, X. Li, B. Omar, A survey on generative adversarial networks: Variants, applications, and training, ACM Comput. Surv. 54 (8) (2021) 1–49.

[48] S. Frolov, T. Hinz, F. Raue, J. Hees, A. Dengel, Adversarial text-to-image synthesis: A review, Neural Netw. 144 (2021) 187–209.

[49] X. Zhou, Y. Jiao, J. Liu, J. Huang, A deep generative approach to conditional sampling, J. Amer. Statist. Assoc. (2022) 1–12.

[50] S. Resnick, A Probability Path, Springer, 2019.

[51] I. Sason, S. Verdú, f-Divergence inequalities, IEEE Trans. Inform. Theory 62 (11) (2016) 5973–6006.

[52] A. Sidhu, P. Bhalla, S. Zafar, Mediating effect and review of its statistical measures, Empir. Econ. Lett. 20 (2021) 29–40.

[53] S.Ö. Özdil, Ö. Kutlu, Investigation of the mediator variable effect using BK, Sobel and bootstrap methods (mathematical literacy case), Int. J. Progress. Educ. 15 (2) (2019) 30–43.

[54] Y. Ren, B. Castro Campos, Y. Peng, T. Glauben, Nutrition transition with accelerating urbanization? Empirical evidence from rural China, Nutrients 13 (3) (2021) 921.

[55] H. Zamanian, M. Amini-Tehrani, Z. Jalali, M. Daryaafzoon, S. Ala, S. Tabrizian, S. Foroozanfar, Perceived social support, coping strategies, anxiety and depression among women with breast cancer: Evaluation of a mediation model, Eur. J. Oncol. Nurs. 50 (2021) 101892.

[56] J.R. Seeley, L.B. Sheeber, E.G. Feil, C. Leve, B. Davis, E. Sorensen, S. Allan, Mediation analyses of internet-facilitated cognitive behavioral intervention for maternal depression, Cogn. Behav. Therapy 48 (4) (2019) 337–352.

[57] Z. Wang, Q. She, T.E. Ward, Generative adversarial networks in computer vision: A survey and taxonomy, ACM Comput. Surv. 54 (2) (2021) 1–38.

Jiaming Zhang received the B.S degree and Ph.D. degree in statistics from Zhongnan University of Economics and Law in 2016 and 2021, respectively. She is currently a lecturer in Zhongnan University of Economics and Law. Her current research interests include machine learning, deep learning and econometrics.

Yiqi Lin received the B.S degree in Statistics from the Department of Mathematics at the Southern University of Science and Technology, and the Ph.D degree in Statistics at The Chinese University of Hong Kong, in 2019 and 2023. His main research interests encompass High-Dimensional Statistics, Causal Inference, Machine Learning, and Bayesian Statistics.

Xinyuan Song received the M.S degree in statistics from Sun Yat-sen University, Guangzhou, China, and the Ph.D. degree in statistics from the Chinese University of Hong Kong, Hong Kong, China, in 1989 and 2000, respectively. She is currently a Chair and Professor in Department of Statistics, the Chinese University of Hong Kong. She is also an associate editor of Psychometrika, associate editor of Structural Equation Modeling: A Multidisciplinary Journal, associate editor of Electronic Journal of Statistics, associate editor of Computational Statistics & Data Analysis, associate editor of Statistics and Its Interface, associate editor of Journal of the Korean Statistical Society, associate editor of Statistical Theory and Related Fields, associate editor of Computational and Mathematical Methods in Medicine. Her research interests include Bayesian method, latent variable models, statistical computing, statistical diagnostics, and survival analysis.

Hanwen Ning received the B.S degree in applied mathematics and the Ph.D. degree in mathematical statistics from the Huazhong University of Science and Technology, Wuhan, China, in 2005 and 2010, respectively. He was a research Fellow at University of Glasgow, Glasgow, UK, and senior academic visitor at Monash University, Melbourne, Australia. He is currently a Professor in Zhongnan University of Economics and Law. His research interests include machine learning, nonlinear system identification, financial data analysis and analysis of PDE.