

Model-Based Counterfactual Synthesizer for Interpretation

Fan Yang¹, Sahan Suresh Alva¹, Jiahao Chen², Xia Hu¹

¹Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA

²J.P. Morgan AI Research, New York, NY, USA

{nacoyang,sahanalva,xiahu}@tamu.edu

jiahao.chen@jpmchase.com

ABSTRACT

Counterfactuals, serving as one of the emerging type of model interpretations, have recently received attention from both researchers and practitioners. Counterfactual explanations formalize the exploration of “what-if” scenarios, and are an instance of example-based reasoning using a set of hypothetical data samples. Counterfactuals essentially show how the model decision alters with input perturbations. Existing methods for generating counterfactuals are mainly algorithm-based, which are time-inefficient and assume the same counterfactual universe for different queries. To address these limitations, we propose a **Model-based Counterfactual Synthesizer (MCS)** framework for interpreting machine learning models. We first analyze the model-based counterfactual process and construct a base synthesizer using a conditional generative adversarial net (CGAN). To better approximate the counterfactual universe for those rare queries, we novelly employ the umbrella sampling technique to conduct the MCS framework training. Besides, we also enhance the MCS framework by incorporating the causal dependence among attributes with model inductive bias, and validate its design correctness from the causality identification perspective. Experimental results on several datasets demonstrate the effectiveness as well as efficiency of our proposed MCS framework, and verify the advantages compared with other alternatives.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics; Neural networks; Generative and developmental approaches; Supervised learning by classification.**

KEYWORDS

Counterfactual sample; causal explanation; model interpretation

ACM Reference Format:

Fan Yang¹, Sahan Suresh Alva¹, Jiahao Chen², Xia Hu¹. 2021. Model-Based Counterfactual Synthesizer for Interpretation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*, August 14–18, 2021, Virtual Event, Singapore. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3447548.3467333>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8332-5/21/08...\$15.00

<https://doi.org/10.1145/3447548.3467333>

1 INTRODUCTION

Recently, machine learning (ML) models have been widely deployed in many real-world applications, and achieved a huge success in many domains. For high-stake scenarios, such as medical diagnosis [23] and policy making [3], explaining the behaviors of ML models is necessary for humans to be able to scrutinize the model outcomes before making a final decision. Various interpretation techniques [6] have thus been proposed to handle the ML explainability issue. Feature attribution [6], for example, is one commonly used interpretation technique in many ML tasks, including image captioning, machine translation and question answering, where important features that contribute most are highlighted as evidences for model predictions.

However, most ML interpretations like feature attribution typically lack the reasoning capability, since they are not discriminative in nature [5], which makes them limited in understanding *how*. To further investigate the decision boundaries of ML models, *counterfactual* explanation [39], emerging as a new form of interpretation, has gradually raised attentions in recent years. Counterfactuals are essentially a series of hypothetical samples, which are synthesized within a certain data distribution and can flip the model decisions to preferred outcomes. With valid counterfactuals, humans can understand how model predictions change with particular input perturbations, and further conduct reasoning under “what-if” circumstances. Given a loan rejection case for instance, attribution analysis simply indicates those important features for rejection (e.g., applicant income), while counterfactuals are able to show how the application could be approved with certain changes (e.g., increase the annual income from \$50,000 to \$80,000).

Several initial attempts on generating counterfactuals have been made in recent work. The first, and most straightforward, line of methodology is to utilize the adversarial perturbations to synthesize the hypothetical samples [29, 30], where the search process is conducted in the input data space. One significant drawback of this methodology is that out-of-distribution (OOD) samples cannot be avoided, which can largely limit the reasoning capability of generated counterfactuals. To cope with this issue, another line of methodology proposes to synthesize the counterfactuals in the latent code space with the aid of pre-trained generative models [16, 32, 42]. In this way, the generated counterfactuals are then well guaranteed to be within certain data distribution for interpretation purposes. Besides, considering the huge computational complexity for deriving high-dimensional samples (e.g., images), there is also a line of methodology using feature composition in data space to synthesize counterfactuals [11, 38]. The extra requirement for this methodology is a proper distractor instance (could be the query itself), which provides relevant feature resources for synthesis.

Despite existing efforts, effective generation of counterfactuals still remains challenging. First, common counterfactual frameworks are mostly algorithm-based ones, which makes them inefficient for sample generation, because each new query necessitates solving one specific optimization problem at one time. For cases with multiple inputs, algorithm-based frameworks could be extremely time-consuming. Second, the counterfactual universe is unknown in advance, and its high-quality approximation is preferred for better explanation. Existing frameworks mostly assume the same counterfactual universe for different queries, though this may not be consistent with the settings of real-world counterfactuals which are related to the inputs. Third, causal dependence among attributes needs be considered for counterfactual feasibility. Although there are a few work [18, 25] trying to incorporate causal constraints into counterfactuals, they simply achieve it by adding extra regularization terms, and do not have any theoretical guarantees on the generation process from a causality perspective.

To address the aforementioned challenges, we propose a **Model-based Counterfactual Synthesizer (MCS)** framework, which can faithfully capture the counterfactual universe and properly incorporate the attribute causal dependence. Specifically, by analyzing the counterfactual process, we are motivated to employ conditional generative adversarial net (CGAN) [27] as the base, and further build a model-based synthesizer by introducing relevant counterfactual objectives. To make MCS better approximate the potential counterfactual universe, we novelly apply the umbrella sampling [19] in synthesizer training, aiming to properly consider the influence of those rare events in data on counterfactual reasoning. Moreover, we also use model inductive bias to design the generator architecture in our proposed MCS framework, so as to incorporate the causal dependence of attributes into the generated samples, which is further validated from the causality identification perspective. Our main contributions are summarized as follows:

- Design a model-based counterfactual explanation framework (i.e., MCS) based on CGAN, whose goal is to help humans better understand the decision boundaries of deployed ML models;
- Apply the umbrella sampling technique in MCS training, which significantly enhances the synthesizer in capturing the influence of those rare events in data for counterfactual explanation;
- Use the concept of model inductive bias to design the generator architecture in the proposed MCS, and further validate the design correctness through a causality identification process;
- Demonstrate the advantages of our proposed MCS on different datasets, and compare the performance with other alternatives.

2 PRELIMINARIES

In this section, we briefly introduce some involved concepts, as well as some basics of the employed techniques.

Counterfactual Explanation. This is one particular ML interpretation technique developed from example-based reasoning [35], where hypothetical data samples are provided to promote the understandings of model boundaries. As a specific example, consider a classification model $f_\theta : \mathbb{R}^d \rightarrow \{-1, 1\}$, with -1 and 1 respectively denoting the undesired and desired outputs. The counterfactual

explanation problem can be generally formulated as:

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \sim C} l(\mathbf{x}, \mathbf{q}_0) \quad \text{s.t. } f_\theta(\mathbf{q}_0) = -1, f_\theta(\mathbf{x}^*) = 1, \quad (1)$$

where $\mathbf{q}_0$ represents the input query, and $\mathbf{x}^*$ is the derived counterfactual sample. Here, C indicates the counterfactual universe of the observed data space $\mathbb{R}^d$, and $l : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_0^+$ denotes a distance measure in the input space. From Eq. 1, we can see that counterfactuals are essentially data samples within some distributions, which can flip the model decisions as desired, while keeping similar to the query input. Conceptually, Eq. 1 can be solved either in an algorithm-based way, or a model-based way. Algorithm-based methods typically employ different optimization strategies to solve Eq. 1 for each query $\mathbf{q}_0$, while model-based ones try to approximate the particular C given relevant constraints and further conduct sampling. In this paper, we mainly explore the counterfactual explanation problem using the model-based methods.

Generative Modeling with CGAN. Generative adversarial net is a novel way to train generative models, which typically has a generator G and a discriminator D [9]. The training objective of G is to capture the data distribution, while the objective of D is to estimate the probability that a sample comes from the data rather than G . CGAN is a natural extension of this framework, where G and D are both conditioned on some additional information $\mathbf{a}$ (e.g., labels or attributes). The min-max game between G and D conducted in CGAN training can be expressed as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_x} \log D(\mathbf{x}|\mathbf{a}) + \mathbb{E}_{\mathbf{z} \sim \mathcal{P}_z} \log(1 - D(G(\mathbf{z}|\mathbf{a}))), \quad (2)$$

where $\mathcal{P}_x$ indicates the data distribution over $\mathbf{x}$, $\mathcal{P}_z$ denotes a prior noise distribution, and V represents a value function of the two players in the min-max game. With a well-trained CGAN, we can effectively capture the conditional distribution given certain constraints or regularizations. In this paper, we use the CGAN framework to approximate the potential counterfactual universe.

Model Inductive Bias. When multiple decisions are equally good, model inductive bias enables the learning process to prioritize some decisions over the others [28], which is independent of the data observed. Model inductive bias can be incorporated in different ways. In early days, connectionist models commonly indicated their inductive bias through relevant regularization terms [26]. For conventional Bayesian models, inductive bias is typically expressed through the prior distribution [12], either from its selection or parameterization. In other contexts, inductive bias can also be encoded with model architectures [1, 43], where the structure itself indicates the data-generating assumption or the potential decision space. In this paper, we make use of the inductive bias specifically instilled by generator architecture to properly consider the causal dependence among attributes for generated samples.

3 MODEL-BASED COUNTERFACTUAL SYNTHESIZER FRAMEWORK

In this section, we first analyze the counterfactual universe given a deployed ML model, and then formulate the problem of model-based counterfactual explanation. Further, we introduce the proposed synthesizer design based on the CGAN framework.

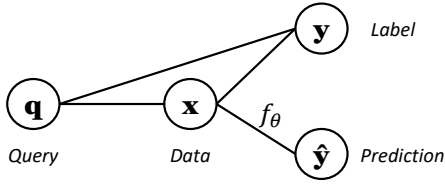


Figure 1: The general graphical model for counterfactual generation process given a deployed ML model.

3.1 Model-Based Counterfactual Explanation

To effectively design a counterfactual synthesizer, it is crucial to make clear the counterfactual universe we are focusing on. Given a deployed ML model f_θ , we can characterize the whole universe with the graphical model illustrated in Fig. 1. In general, x and y represent the data and label variables respectively, while $\hat{y}$ denotes the decision variable output from f_θ . The query variable q is introduced to incorporate relevant constraints for counterfactual reasoning, which brings about the fact that hypothetical samples (i.e., x and y) are typically generated under the influence from q . According to the graphical model in Fig. 1, we can further factorize the joint distribution of the whole counterfactual universe as follows:

$$\mathcal{P}_{q,x,y,\hat{y}} = \mathcal{P}_q \cdot \mathcal{P}_{x|q} \cdot \mathcal{P}_{\hat{y}|x} \cdot \mathcal{P}_{y|x,q} = \mathcal{P}_q \cdot \mathcal{P}_{\hat{y}|x} \cdot \mathcal{P}_{x,y|q}. \quad (3)$$

Within Eq. 3, $\mathcal{P}_q$ is typically known as the prior, and $\mathcal{P}_{\hat{y}|x}$ is considered as fixed since the model f_θ is pre-deployed. Thus, the key to capturing the counterfactual universe lies in the proper approximation of $\mathcal{P}_{x,y|q}$ (i.e., the joint distribution of x and y conditioned on q), which reflects the latent sample generation process with certain query. Thus, to achieve the model-based counterfactual analysis, we need to investigate the hypothetical sample generation under particular query conditions. With this insight, we now formally define the problem of model-based counterfactual explanation below.

DEFINITION 1. A **model-based counterfactual** is a data point sampled from a perturbed **hypothetical distribution**, which statistically satisfies the counterfactual requirements (indicated by Eq. 1). Given a specific query q_0 , counterfactual x^* can be obtained through sampling $x^* \sim C_{x|q_0}$, where $C_{x|q_0}$ is a hypothetical distribution marginalized from $C_{x,y|q_0}$. In general, $C_{x,y|q_0}$ can be derived by

$$C_{x,y|q_0} = \arg \min_{\mathcal{P}_{x,y|q_0}} L^{cf}(\mathcal{P}_{x,y|q_0}), \quad (4)$$

where q_0 follows the prior $\mathcal{P}_q$, and L^{cf} indicates a counterfactual loss.

By definition, a model-based counterfactual does not focus on the instance optimization for each individual query. Instead, it tries to capture the latent sample generation process with particular query conditions, within the whole counterfactual universe. One significant merit brought by model-based explanations is that it largely enhances the efficiency for counterfactual generation, since we only need to obtain the certain hypothetical distribution once for all potential queries following the prior $\mathcal{P}_q$. Nevertheless, modeling such latent generation processes is nontrivial, and we need some specific designs to make it effectively work.

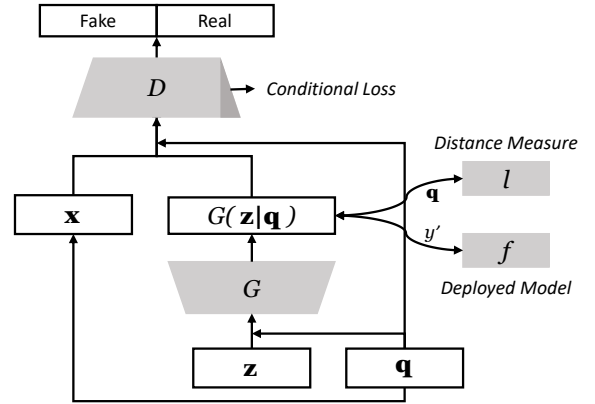


Figure 2: The overall framework design for the proposed counterfactual synthesizer through conditional modeling.

3.2 Conditional Synthesizer Design

Designing a model-based synthesizer typically involves how to build an effective generative model for counterfactuals. According to the previous analysis, we know that the key lies in the approximation of hypothetical distribution $C_{x,y|q}$, which can be formulated as a conditional modeling problem. To this end, we propose a conditional generative framework based on CGAN in this paper, specifically designing for counterfactual explanation. The overall architecture is illustrated by Fig. 2.

In the designed framework, D represents the discriminator module, and G denotes the generator module. Similar as CGAN, D and G are jointly trained as adversaries to each other, aiming to achieve a min-max game. The major difference with CGAN comes from how we prepare the conditional vectors for the framework training. Here, instead of simply employing label information, we use the query as conditions for counterfactual generation. Besides, to guarantee the quality of generated counterfactual samples, we further incorporate a distance measure l and the deployed model f to regularize the training of G . Throughout this process, we aim to obtain the counterfactuals which are similar to the query q and have the preferred output decision y' . Specifically, the training objective of this counterfactual min-max game can be indicated as:

$$\min_G \max_D \mathbb{E}_{x \sim \mathcal{P}_x} \log D(x|q) + \mathbb{E}_{z \sim \mathcal{P}_z} \log [1 - D(G(z|q))] + L^{cf}(G(z|q)), \quad (5)$$

where z represents the noise vector following a distribution $\mathcal{P}_z$. Within Eq. 5, the counterfactual loss term L^{cf} can be further expressed as follows:

$$L^{cf}(G(z|q)) = L^{ce}(f(G(z|q)), y') + l(G(z|q), q), \quad (6)$$

in which L^{ce} indicates the cross-entropy loss between the predictions from f and the preferred decision y' , regarding to the generated samples. Overall, L^{cf} is expected to be minimized for counterfactual reasoning purposes, and it can be treated as an additional regularization term appended with the conventional CGAN training. By utilizing this conditional design, we can then employ the well-trained generator G to synthesize a series of hypothetical distributions parametrized by query q , and further achieve the model-based counterfactual generation through sampling.

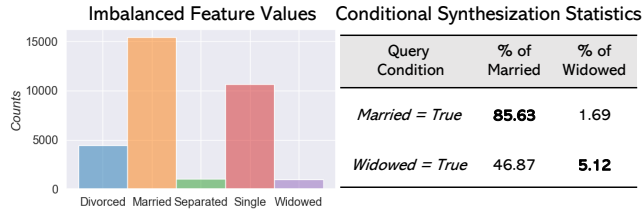


Figure 3: Case results for base synthesizer on conditional performance of the ‘marital-status’ attribute in *Adult* data.

4 ENHANCEMENT FOR COUNTERFACTUALS

With the previous design, we now consider two practical enhancements for model-based counterfactuals. First, we propose a novel training scheme for counterfactual synthesizer based on the umbrella sampling technique. Second, we utilize the model inductive bias to consider the causal dependence among attributes.

4.1 Effective Synthesizer Training

4.1.1 Query imbalance during training. To effectively train the designed counterfactual synthesizer shown in Fig. 2, we need to let G well capture the conditions indicated by q . However, this may not be as straightforward as in conventional CGAN, since $\mathcal{P}_q$ is typically imbalanced among different attribute values. Such query imbalance results in the fact that the hypothetical distributions conditioned on those rare values cannot be effectively approximated, due to the limited number of instances. To illustrate the point, we show some case results of the base synthesizer on *Adult* dataset¹ in Fig. 3. Here, we assume that q and x share a same prior distribution (i.e., $\mathcal{P}_q = \mathcal{P}_x$), because queries are usually collected from similar data sources in most real-world scenarios. According to the results, we note that the imbalanced values of ‘marital-status’ attribute² lead to significantly different conditional performance on synthesis. When conditioned on those majority values (e.g., Married), the synthesizer can reasonably capture the corresponding hypothetical distribution. In contrast, the synthesizer fails when conditioned on minority values (e.g., Widowed), and its conditional performance is bad. Thus, to better train the designed counterfactual synthesizer, we need to prepare a proper set of training queries, which contains sufficient samples with attribute values in the tails of prior $\mathcal{P}_q$.

4.1.2 Umbrella sampling for rare instances. To properly reflect the influence of rare values, we need some enhanced sampling strategies for training, instead of the simple random way. An intuitive way is to relatively increase the probability mass for rare values. In work [41], the authors used the frequency logarithm to curve the probability mass, aiming to make the sampling process have higher chances in obtaining those “tail” values. However, such mass curves may not be well suited for our training scenario, because it distorts the prior $\mathcal{P}_q$ and further leads to an unfaithful hypothetical distribution $C_{x,y|q}$. Considering this, we novelly apply the umbrella sampling technique here to enhance the synthesizer training, which was originally used in computational physics and chemistry for

¹<http://archive.ics.uci.edu/ml/datasets/Adult>

²We merge Married-civ-spouse, Married-spouse-absent, Married-AF-spouse in ‘marital-status’ attribute all to value Married for simplicity.

	age	workclass	education		hours_per_week	Prediction
Query	22.0	Private	HS-Grad		45.0	<=50K
CF 1	--	Self-Employ	Doctorate		--	>50K
CF 2	--	--	Some-College		53.0	>50K

*(CF $\rightarrow$ Counterfactual; -- $\rightarrow$ No Changes)

Figure 4: Case study on counterfactual samples (selected) generated by DiCE algorithm in *Adult* data.

molecular simulation [19]. Umbrella sampling recasts the whole sampling process into several unique samplings over umbrella-biased distributions in a weighted-sum manner, where the added artificial umbrellas are expected to cover the full value domain with overlaps. By calculating the weight of each biased distribution, we can then reconstruct the original distribution and conduct evaluations with the umbrella samples obtained. Specifically for counterfactual synthesizer training, we can thus guarantee the sufficient number of queries with balanced values by sampling under different umbrella biases. In particular, the corresponding weight of each biased distribution for query preparation can be calculated as below.

THEOREM 1. Consider the sampling process $q \sim \mathcal{P}_q$ for counterfactual synthesizer training. Let $w = [w_1, \dots, w_N]$ denote the weight vector for N umbrella-biased distributions [19], where w_i indicates the normalized weight of the i -th biased distribution $\mathcal{P}_q^i$, and $u = [u_1, \dots, u_N]$ denote the profile of the added artificial umbrellas. Then, the optimal w can be derived by solving the equation $wM = w$, where $M = M(w)$ represents the overlap matrix defined as:

$$M_{ij} = \left\langle \frac{u_j/w_i}{\sum_{k=1}^N u_k/w_k} \right\rangle_{\mathcal{P}_q^i}. \quad (7)$$

Here, the operation $\langle \cdot \rangle_{\mathcal{P}}$ indicates the average over distribution $\mathcal{P}$. The proof for deriving w is shown in Appendix A.

4.2 Causal Dependent Generation

4.2.1 Causality for generated counterfactuals. To make the generated counterfactuals have better feasibility on reasoning, it is also preferred to consider the causality among different input attributes. We here use another case study on the *Adult* data, shown in Fig. 4, to illustrate the point. In particular, we employ an existing counterfactual explanation method, DiCE [30], to generate related counterfactual samples. The results show in Fig. 4 that DiCE is not able to reflect the causality between the attribute ‘education’ and ‘age’, since it suggests to improve the education level without changing the age for flipping model decisions. In real-world settings, such counterfactuals are usually considered infeasible, because input attributes cannot be altered independently in counterfactuals. Specifically in this example, education level typically improves with age, so feasible counterfactuals should indicate such causality for human reasoning. We now enhance our proposed MCS by showing how to incorporate such causal dependencies among attributes.

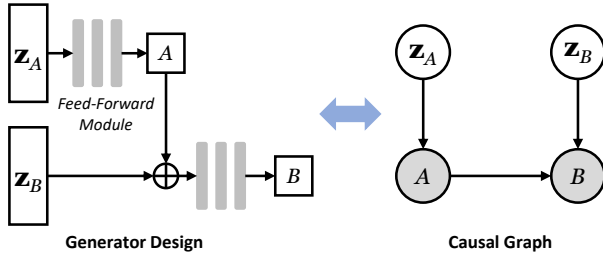


Figure 5: Example design of G for incorporating the causation: $A \rightarrow B$, as well as the represented causal graph.

4.2.2 Model inductive bias for causal dependence. In contrast with the algorithm-based counterfactuals, model-based counterfactuals provide new possibilities to incorporate domain-specific causal dependence for explanation, instead of simply adding extra regularization or constraints for generation [25]. In particular, we propose to utilize the inductive bias of G to encode relevant causal dependence, where the architecture of G is intentionally designed to mimic the structural equations [13] of corresponding causation. Fig. 5 shows an example of designing G for the causal relationship $A \rightarrow B$, where A indicates the cause and B denotes the effect. Essentially, by purposely manipulating the generator architecture, we encode the causal structure as inductive biases in G , so as to achieve the causal dependent generation for counterfactuals. We now state a theorem proving the correctness of this approach from the perspective of causality identification, where the corresponding considered causal dependence is shown to be existed from the generated counterfactual samples.

THEOREM 2. Let $\mathcal{D} = (\mathcal{A}, \mathcal{E})$ represent a causal graph with vertex set $\mathcal{A} = \{A_1, A_2, \dots, A_S\}$ being attributes used for counterfactual generation, and edge set $\mathcal{E}$ consisting of directed edges from causes to effects. Then, the generator G can be designed with S feed-forward modules $G = [F_1, F_2, \dots, F_S]$, mimicking the corresponding structural equations of $\mathcal{D}$, such that A_s can be generated by

$$A_s = F_s \left(\{A_p\}_{p \in \text{Pa}_s^{\mathcal{D}}}, \mathbf{z}_{A_s} \right), \quad \forall s \in [1, \dots, S], \quad (8)$$

where $\text{Pa}_s^{\mathcal{D}}$ denotes the set of parent attributes (i.e., causes) for A_s in graph $\mathcal{D}$. As a result, the counterfactual samples generated by G are further guaranteed to have the incorporated causality, which can be identified from the observational perspective.

The relevant discussion & proof are shown in Appendix B.

5 IMPLEMENTATION

In this section, we briefly introduce the employed practical techniques when implementing the proposed MCS framework. We also show the overall pipeline of MCS for counterfactual generation.

Data Representation. When implementing MCS, we employ different modeling techniques for different types of data attributes. We use Gaussian mixture models [2] to encode continuous attributes, and normalize the values according to the selected mixture mode. We represent discrete attributes directly using one-hot encoding. Furthermore, one specific data instance $\mathbf{r}_{ins}$ can be then

Algorithm 1: Building Pipeline of MCS for Interpretation

```
# Setup Phase
- Prepare  $f$  for explanation, and select  $l$  for measurement;
- Design the generator  $G$  with domain  $\mathcal{D}$  based on Theo. 2;
# Training Phase
- Data Modeling with Eq. 9;
- for training batch  $k$  do
  Utilize umbrella sampling to prepare a set of queries  $\mathbf{q}_k$ ;
  Weigh batch  $k$  with  $\mathbf{w}$  over  $\mathbf{q}_k$  based on Theo. 1;
  Update  $G$  and  $D$  in the min-max game of Eq. 5;
# Interpretation Phase
- Feed the user query  $\mathbf{q}$  to  $G$  for counterfactual generation.
```

represented as

$$\mathbf{r}_{ins} = \mathbf{c}_1^{ins} \oplus \dots \oplus \mathbf{c}_{N_c}^{ins} \oplus \mathbf{d}_1^{ins} \oplus \dots \oplus \mathbf{d}_{N_d}^{ins}, \quad (9)$$

where N_c and N_d indicate the number of continuous and discrete attributes in the data respectively. Also, we represent the query $\mathbf{q}$ by value masking, thus giving humans control over the semantics of the counterfactuals as appropriate for any particular use case.

Umbrella Sampling for Discrete Attributes. Conventional umbrella sampling technique only applies to continuous distributions. For discrete attributes, we use the Gumbel-Softmax [15] method to relax the categorical distribution into a continuous one. This reparameterization trick is proved to be faithful and effective in many cases with appropriate temperature τ , which controls the trade-off of the distribution relaxation. When $\tau \rightarrow 0$, the relaxed distribution becomes into the original discrete one. When $\tau \rightarrow \infty$, it gradually converges into a uniform distribution. In our proposed MCS framework, we conduct the implementation with a fixed temperature as $\tau = 0.5$.

Overall MCS Pipeline. To clearly show the procedures of building MCS for counterfactual explanation, we give an overview of the pipeline, illustrated by Algorithm 1. Compared with the existing algorithm-based counterfactuals, the proposed MCS is much more efficient during the interpretation phase, since it avoids the iterative perturbation step regarding to each input $\mathbf{q}$. In exchange, MCS pushes the computational complexity to the setup and training phase, which largely depends on the data scale as well as the value space we focus on.

6 EXPERIMENTS

In this section, we empirically evaluate the proposed MCS on both synthetic and real-world datasets from several different aspects, and aim to answer the following key research questions:

- How effective and efficient is MCS in generating counterfactual samples, compared with existing algorithm-based methods?
- How well does MCS model the original observational distribution as well as the conditional counterfactual distribution, with the umbrella sampling technique?
- Can we identify the incorporated domain-specific causal dependence from the counterfactual samples generated by MCS?

Table 1: Data statistics in experiments.

Dataset	#Row	#Col	Attribute Type
Syn_Moons	500	3	Continuous
Syn_Circles	500	3	Continuous
Adult	48,842	9	Continuous & Categorical
Home_Credit	344,971	39	Continuous & Categorical

6.1 Evaluation on Counterfactual Generation

In this part, we evaluate the effectiveness and efficiency of MCS, comparing with existing algorithm-based counterfactual methods.

6.1.1 Experimental settings. We consider two synthetic datasets and two real-world datasets for counterfactual generation evaluation. Specifically, the statistics of the datasets are shown in Tab. 1.

- **Synthetic³:** We synthesize two datasets for classification purpose, i.e., “Syn_Moons” and “Syn_Circles”, with different separation boundaries. To facilitate visualization, the synthetic data only contains two continuous attributes. These two synthetic datasets are mainly used to evaluate the MCS effectiveness.
- **Adult:** This is a real-world benchmark dataset for income prediction, where each instance is labelled as “>50K” or “≤50K”. In the experiments, the counterfactuals on this data aim to help understand how to flip model decisions from “≤50K” to “>50K”. To facilitate our task, we only consider a subset of the attributes.
- **Home_Credit⁴:** This is a larger real-world dataset for client risk assessment, where the goal is to predict clients’ repayment abilities of given loans. The counterfactuals here are to help reason how to make improvements for risky clients to become non-risky. We drop some unimportant attributes in experiments.

As for the deployed classifier, we prepare three different f for counterfactual generation evaluation as below. Those classifiers are all trained with 80% of the data, and tested with the rest 20%.

- **RBF SVM:** This is a pre-trained support vector machine (SVM) with the RBF kernel, where a squared l_2 penalty is applied. Related SVM hyperparameter γ is set to 2, and C is set to 1.
- **Random Forest (RF):** This is a pre-trained tree-based RF classifier with 10 estimators. The maximum depth of each tree is set as 5. We use the Gini impurity as our splitting criterion.
- **Neural Net (MLP):** This is a pre-trained ReLU neural classifier with multi-layer perceptron (MLP). Here, we have 1 hidden layer and 100 hidden units. The relevant l_2 regularization coefficient is set to 1, and the maximum iteration number is set as 1000.

Furthermore, we select four recent algorithm-based counterfactual explanation methods as our baselines for comparison. These methods are all set up with their default settings.

- **DiCE** [30]: This method generates diverse counterfactual samples by providing feature-perturbed versions of the query, where the perturbations are derived by iterative optimization.
- **C-CHVAE** [32]: This method utilizes a pre-trained variational auto-encoder (VAE) to transform the data space into a latent embedding space, and then perturbs the latent representation of the query for counterfactual generation.

³<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets>

⁴<https://www.kaggle.com/c/home-credit-default-risk/data>

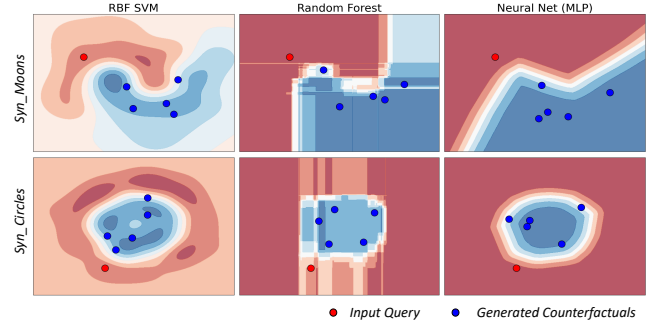


Figure 6: Visualization of the generated counterfactual samples by the proposed MCS on two synthetic datasets.

- **CADEX** [29]: This method employs the gradient-based scheme to perturb the query for flipping outcomes, which is an application of adversarial attack methods for counterfactual generation.
- **CLEAR** [40]: This method uses the concept of b -perturbation to construct potential counterfactuals through local regression, where the corresponding fidelity error is minimized iteratively.

6.1.2 Counterfactual generation effectiveness. To demonstrate the effectiveness of MCS, we visualize the generated counterfactuals on two synthetic datasets, illustrated by Fig. 6. In this set of experiments, we select a fixed query q_0 for each synthetic dataset with negative model decision (i.e., predicted as ‘-1’), and randomly show 5 samples generated by MCS. Fig. 6 shows that all generated samples successfully flip the query prediction from the negative ‘-1’ to the positive ‘1’, across different classifiers f . Thus, it is noted that MCS is capable of synthesizing valid counterfactuals for prediction reasoning. To further evaluate the counterfactuals generated by MCS, we employ the *average Euclidean distance* as the metric, indicated by Eq. 10, to reflect how close of the generated samples regarding to the input query:

$$\delta^{\text{avg}} = \sum_{e \in S^f} \text{Euclid_Dist}(q_0, e) / |S^f|, \quad (10)$$

where S^f denotes the set of the generated counterfactuals. In our experiments, we set $|S^f| = 20$, and compare δ^{avg} among different counterfactual methods under different f over different datasets. Fig. 7 shows that counterfactuals generated by MCS are generally farther to the queries, compared with those generated by baselines (except C-CHVAE). This observation suggests that algorithm-based methods are good at finding those “nearest” counterfactuals through iterative perturbations, while MCS mainly focus on the counterfactual distribution approximation instead of sample searching. For C-CHVAE, it is designed to conduct perturbations in the latent space, and hence its generated counterfactuals are also observed to be farther than those from DiCE, CADEX and CLEAR in the data space. C-CHVAE can be treated as a data density approximator, thus the corresponding δ^{avg} are almost the same over different f within the same dataset. Overall, we know that MCS can effectively generate valid counterfactuals for reasoning, but its synthesized samples may not be as close to the query as those generated from existing algorithm-based methods (i.e., DiCE, CADEX, CLEAR), which becomes more significant with the increase of the data scale.

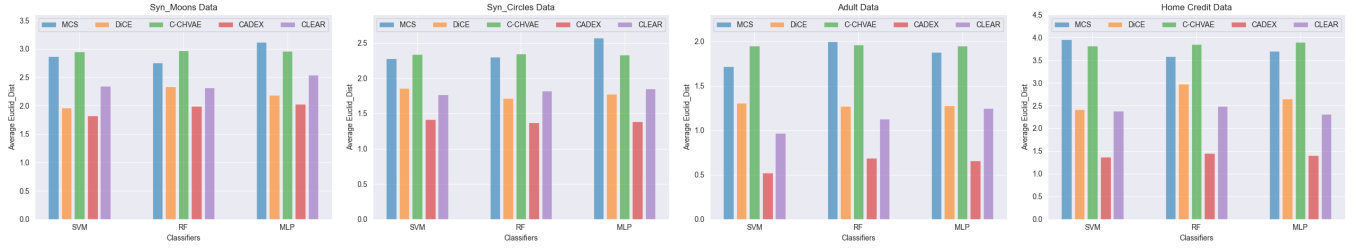


Figure 7: The average Euclidean distance comparison between the query and the generated counterfactuals.



Figure 8: The time efficiency comparison with MLP.

6.1.3 Counterfactual generation efficiency. To evaluate the efficiency of different counterfactual explanation methods, we compare the time cost of sample generation process regarding to multiple queries. In this set of experiments, we only consider the *Adult* and *Home_Credit* dataset for illustration, with a MLP classifier f , and test the counterfactual generation under 10, 20 and 30 queries. The relevant results are reported by averaging 5 runs over different sets of input. Fig. 8 shows that MCS consumes significantly less time for counterfactual generation, and such merit over algorithm-based methods is more remarkable with a larger scale of input. Besides, we can also observe that MCS is much more stable (i.e., lower standard deviation) among different queries, and its unit time cost for each query almost keeps identical despite the specific inputs.

6.2 Evaluation on Distribution Modeling

In this part, we evaluate the modeling performance of MCS aided by umbrella sampling technique. Overall, we aim to demonstrate the faithfulness of the generated counterfactuals.

6.2.1 Experimental settings. We only consider the two real-world datasets (i.e., *Adult* and *Home_Credit* data) for this part of evaluation. For the artificial umbrellas used during MCS training, we respectively set $N = 8$ and $N = 32$ for training query sampling in the *Adult* and *Home_Credit* dataset. Detailed hyper-parameters, as well as the related influence studies, of the umbrella sampling process are introduced in Appendix C.

6.2.2 Observational distribution modeling. To evaluate the MCS modeling on observational data distribution, we use the metric *model compatibility* in [31] focusing on learning efficacy. The intuition behind this metric is that, models trained on data with similar

Table 2: Average F-score of testing on learning efficacy.

Dataset	$\mathcal{F}_o$		$\mathcal{F}_t$		$\mathcal{F}_p$	
	RF	MLP	RF	MLP	RF	MLP
Adult	0.616	0.435	0.593	0.429	0.586	0.431
Home_Credit	0.602	0.386	0.421	0.307	0.418	0.306

distributions should have similar test performances. Furthermore, to conduct such evaluation, we prepare three different sets of classifiers (i.e. $\mathcal{F}_o$, $\mathcal{F}_t$ and $\mathcal{F}_p$) for test comparison, which are respectively trained on the original data, table-GAN [31] synthesized data and MCS synthesized data. Here, we employ table-GAN as a baseline synthesizer, since it has been proven to be an effective way to create hypothetical samples which follow a particular observational distribution. As for MCS, we regard the label y as an additional attribute for synthesis in this part, and remove L^c_f in Eq. 5 during training. The final synthesized data is then generated by G in an unconstrained manner with $\mathbf{q}$ set to *None* (all zero). In experiments, we consider two types of classifiers (RF & MLP), and report the average F-score over five rounds. The results in Tab. 2 show that classifiers in $\mathcal{F}_p$ have competitive test performances with those in $\mathcal{F}_t$, and thus demonstrates that MCS can reasonably model the observational distribution with its synthesized data samples.

6.2.3 Counterfactual distribution modeling. Evaluating MCS in modeling counterfactual distribution is essentially to assess the conditional generation performance under particular human priors. When users request hypothetical samples for reasoning, a good synthesizer should have a reasonable conditional performance in generating those in-need samples which are consistent with human priors. In the experiments herein, we select the ‘marital-status’ and ‘housing-type’ attributes in *Adult* and *Home_Credit* data for illustration, considering assumptive priors on rare values for testing. In particular, we evaluate with priors on ‘marital-status=Widowed’ and ‘housing-type=Office-apt’. Originally, Widowed takes around 3%, and Office-apt takes around 1%. The derived counterfactual distributions from MCS of ‘marital-status’ and ‘housing-type’ are illustrated by Fig. 9. Here, *MCS-base* indicates the synthesizer trained directly with Eq. 5, *MCS-LF* represents the synthesizer trained with logarithm frequency curve [41], and *MCS-US* denotes the synthesizer trained with the umbrella sampling technique. Based on the results, it is noted that, aided with umbrella sampling, our proposed MCS can effectively approximate the hypothetical distribution presumed by human priors, and generate in-need counterfactuals to facilitate the reasoning process.

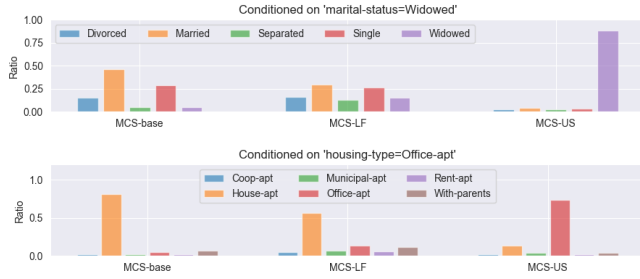


Figure 9: The counterfactual distributions derived by MCS.

6.3 Evaluation on Causal Dependence

In this part, we evaluate the causal dependence over the samples from MCS, and test the causation reflected by model inductive bias.

6.3.1 Experimental settings. We only use *Adult* and *Home_Credit* data for causal dependence evaluation. When designing the generator G , we consider the following cause-effect pairs for synthesis:

$$\begin{aligned} \text{Adult} : \text{education} &\rightarrow \text{age}; \\ \text{Home_Credit} : \text{income_type} &\rightarrow \text{income_total}. \end{aligned} \quad (11)$$

To validate such pairwise causality in synthesized data, we employ two different methods that are commonly used as below:

- **ANM** [14]: It is a popular approach for pairwise causality identification, which bases on the data fitness to the additive noise model on one direction and the rejection on the other direction.
- **CDS** [7]: It measures the variance of marginals after conditioning on bins, which indicates statistical features of joint distribution.

6.3.2 Pairwise causation identification. To conduct the evaluation on pairwise causal dependence, we utilize a simple causation score τ^c to indicate the strength of particular cause-effect pairs. Specifically for $A \rightarrow B$, the causation score can be calculated as follows:

$$\tau_{A \rightarrow B}^c = \tau_{B \rightarrow A}^f - \tau_{A \rightarrow B}^f, \quad (12)$$

where τ^f denotes the data fitness score for a given direction. For different methods, τ^f has different statistical meanings. In our case, τ^f indicates an independence test⁵ score for ANM, and represents the standard deviation of fitness for CDS. Thus, we know that, the larger the τ^f is, the causation on the given direction is less likely to happen from the observational perspective. In practice [17], the causation $A \rightarrow B$ is considered to exist when $\tau_{A \rightarrow B}^c \geq 1$, and $B \rightarrow A$ when $\tau_{A \rightarrow B}^c \leq -1$. When $-1 < \tau_{A \rightarrow B}^c < 1$ holds, it usually indicates that there is no obvious causal dependence identified based on statistical analysis. In experiments, we test the causation pairs in Eq. 11, and show the relevant results in Fig. 10. Here, we compare τ^c between the original and synthesized data with both ANM and CDS under different data scale. We observe a stronger causal dependence for the considered pairs in the synthesized data, which demonstrates the effectiveness of MCS on causal dependent generation. With such advantages, humans can easily obtain feasible counterfactuals by properly incorporating relevant domain-specific knowledge into MCS for interpretation.

⁵We test the independence using the Hilbert-Schmidt independence criterion (HSIC).

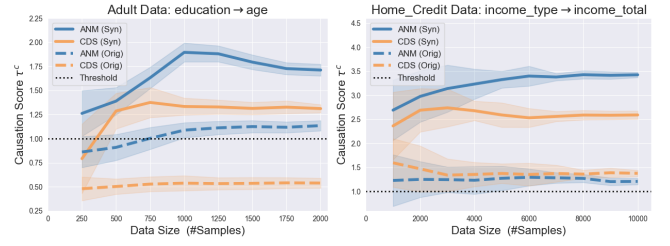


Figure 10: The scores of considered pairwise causation.

7 RELATED WORK

Counterfactuals are one of many interpretation techniques for ML models. In general, according to the format of *explanation carrier* (i.e., how explanations are delivered to humans), related interpretation methods can be categorized as follows.

The first category of interpretation methods uses instance features as indicators to demonstrate which part of the input contributes most to the model predictions. A representative work is *LIME* [33], which uses linear models to approximate the local decision boundary and derive feature importance by perturbing the input sample. Some other methods, utilizing input perturbations for feature importance calculation, can also be found in *Anchors* [34] and *SHAP* [24]. Besides, employing model gradient information for feature attribution is another common methodology under this category. Related examples can be found in *GradCAM* [36] and *Integrated Gradients* [37].

The second category uses abstracted concepts as high-level features to indicate the prediction attribution process. One of the earliest work in this category is *TCAV* [21], which trains linear concept classifiers to derive concept representations, and measures the concept importance based on sensitivity analysis. Besides, the authors in [45] decompose the model prediction semantically according to the projection onto concept vectors, and quantify the contributions over a large concept corpus. Inspired by *SHAP*, a corresponding attribution method with human concepts, named *ConceptSHAP*, is proposed in [44] to quantify the concept contributions with game theories. Some other follow-up work based on *TCAV* can also be found in [8, 10].

The third category uses data samples to deliver relevant explanations to humans. One line of research is to select out prototype or criticism data samples in training set to interpret model behaviors [4, 20]. Similarly in [22], influential training samples to particular model predictions are selected, with the aid of influence functions in measuring sample importance. Furthermore, beyond the real samples in training set, synthesized hypothetical one (e.g., counterfactual sample) is yet another way to interpret predictions for model reasoning. Several representative work along this direction can be found in [16, 30, 32]. Our work generally lies in this category of methods for interpreting ML model behaviors.

8 CONCLUSIONS

In this paper, a general interpretation framework, named MCS, has been proposed to synthesize model-based counterfactuals for prediction reasoning. By analyzing the focusing counterfactual universe, we first formally defined the problem of model-based

counterfactual explanation, and then employed the CGAN structure to train our proposed MCS in an end-to-end manner. To better capture the hypothetical distributions, we novelly applied the umbrella sampling technique to enhance the synthesizer training. Furthermore, we also showed a promising way to incorporate the attribute causal dependence into MCS with model inductive bias, aiming to achieve better feasibility for the derived counterfactuals. Experimental results on both synthetic and real-world data validated several advantages of MCS over other alternatives. Future work extensions may include the model-based counterfactual explorations under more challenging contexts, such as involving high-dimensional data space, time-series sequential nature, or some ethical concerns.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. This work is in part supported by NSF IIS-1900990, and NSF IIS-1939716. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies. This work was partially done when F. Yang was an intern at J. P. Morgan AI Research.

Disclaimer. This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

- [1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261* (2018).
- [2] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. Springer.
- [3] Tim Brennan and William L Oliver. 2013. Emergence of machine learning techniques in criminology: implications of complexity in our data and in research questions. *Criminology & Pub. Pol'y* 12 (2013), 551.
- [4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*. 8928–8939.
- [5] Amit Dhurandhar, Pin-Yu Chen, et al. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*. 592–603.
- [6] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
- [7] José AR Fonollosa. 2019. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*. Springer, 339–347.
- [8] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129* (2019).
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, et al. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.
- [10] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *arXiv:1907.07165* (2019).
- [11] Yash Goyal, Ziyang Wu, Jan Ernst, et al. 2019. Counterfactual Visual Explanations. In *International Conference on Machine Learning (ICML)*. 2376–2384.
- [12] Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences* 14, 8 (2010), 357–364.
- [13] Ned Hall. 2007. Structural equations and causation. *Philosophical Studies* 132, 1 (2007), 109–136.
- [14] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, et al. 2008. Nonlinear causal discovery with additive noise models. *NeurIPS* 21 (2008), 689–696.
- [15] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv:1611.01144* (2016).
- [16] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, et al. 2018. xGEMs: Generating explainers to explain black-box models. *arXiv:1806.08867* (2018).
- [17] Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. 2020. Causal Discovery Toolbox: Uncovering causal relationships in Python. *JMLR* 21, 37 (2020), 1–5.
- [18] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv:2010.04050* (2020).
- [19] Johannes Kästner. 2011. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1, 6 (2011), 932–942.
- [20] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NeurIPS*. 2280–2288.
- [21] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, et al. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*. 2668–2677.
- [22] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *ICML*. 1885–1894.
- [23] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, et al. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.
- [24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NeurIPS*. 4765–4774.
- [25] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv:1912.03277* (2019).
- [26] JL McClelland. 1992. The interaction of nature and nurture in development: A parallel distributed processing perspective (Parallel Distributed Processing and Cognitive Neuroscience PDP. CNS. 92.6).
- [27] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv:1411.1784* (2014).
- [28] Tom M Mitchell. 1980. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research.
- [29] Jonathan Moore, Nils Hammerla, and Chris Watkins. 2019. Explaining deep learning models with constrained adversarial examples. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 43–56.
- [30] Ramaravind K Muthilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings on Fairness, Accountability, and Transparency (FAccT)*. 607–617.
- [31] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. *arXiv:1806.03384* (2018).
- [32] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020*. 3126–3132.
- [33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD conference on knowledge discovery and data mining*. 1135–1144.
- [34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *32nd AAAI on Artificial Intelligence*.
- [35] Edwina L Rissland. 1991. Example-based reasoning. *Informal reasoning in education* (1991), 187–208.
- [36] Ramprasaath R Selvaraju, Michael Cogswell, et al. 2017. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*. 618–626.
- [37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.
- [38] Tom Vermeire and David Martens. 2020. Explainable Image Classification with Evidence Counterfactual. *arXiv:2004.07511* (2020).
- [39] Sandra Wachter et al. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
- [40] Adam White and Artur d'Avila Garcez. 2019. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020* (2019).
- [41] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. In *NeurIPS*. 7335–7345.
- [42] Fan Yang, Ninghao Liu, Mengnan Du, et al. 2021. Generative Counterfactuals for Neural Networks via Attribute-Informed Perturbation. *arXiv:2101.06930* (2021).
- [43] Fan Yang, Ninghao Liu, Mengnan Du, Kaixiong Zhou, Shuiwang Ji, and Xia Hu. 2020. Deep Neural Networks with Knowledge Instillation. In *SDM*. 370–378.
- [44] Chih-Kuan Yeh, Been Kim, et al. 2020. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. *NeurIPS* (2020).
- [45] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *ECCV*. 119–134.

A PROOF OF THEOREM 1

Through the umbrella sampling process, we aim to reconstruct the target distribution $\mathcal{P}_q$ with N biased distributions $\mathcal{P}_q^i$ by a weighted sum manner, where $i = 1, \dots, N$. In the following, we demonstrate how to derive the corresponding weight vector $\mathbf{w}$ with the aid of the pre-defined overlap matrix $\mathbf{M}$ in Eq. 7.

Without loss of generality, we here assume $\mathbf{q}$ only contains continuous attributes. By adding the umbrella profile u_i , the biased distribution $\mathcal{P}_q^i$ can then be indicated by:

$$\mathcal{P}_q^i = u_i \cdot \mathcal{P}_q / w_i, \quad (13)$$

where w_i denotes the normalized weight for $\mathcal{P}_q^i$ we aim to obtain. Since $\int \mathcal{P}_q^i d\mathbf{q} = 1$ always holds for each biased distribution, we can further represent w_i as follows:

$$w_i = \int u_i \mathcal{P}_q d\mathbf{q} = \langle u_i \rangle_{\mathcal{P}_q}. \quad (14)$$

To evaluate the generator G over $\mathcal{P}_q$, we have

$$\begin{aligned} \langle G \rangle_{\mathcal{P}_q} &= \int G \mathcal{P}_q d\mathbf{q} = \int G \frac{\sum_{i=1}^N u_i / w_i}{\sum_{j=1}^N u_j / w_j} \mathcal{P}_q d\mathbf{q} \\ &= \sum_{i=1}^N \int \frac{G}{\sum_{j=1}^N u_j / w_j} \frac{u_i}{w_i} \mathcal{P}_q d\mathbf{q} \\ &= \sum_{i=1}^N \int \frac{G}{\sum_{j=1}^N u_j / w_j} \mathcal{P}_q^i d\mathbf{q} = \sum_{i=1}^N \left\langle \frac{G}{\sum_{j=1}^N u_j / w_j} \right\rangle_{\mathcal{P}_q^i}. \end{aligned} \quad (15)$$

Thus, we know that the original evaluation of G over $\mathcal{P}_q$ can be possibly conducted over the sum of N biased distributions with proper weights. Here, the distribution sum operation can be simply achieved by the direct sampling over each $\mathcal{P}_q^i$, where $i = 1, \dots, N$.

To finally obtain the weights, we utilize the overlap matrix $\mathbf{M}$ defined in Eq. 7. Within $\mathbf{M}$, it is noted that M_{ij} equals to 0 if there is no overlap between $\mathcal{P}_q^i$ and u_j . Further, we can derive the product of the weight vector $\mathbf{w}$ and the j -th column of $\mathbf{M}$ as below, based on Eq. 14 and Eq. 15.

$$\sum_{i=1}^N w_i \cdot M_{ij} = \sum_{i=1}^N \left\langle \frac{u_j}{\sum_{k=1}^N u_k / w_k} \right\rangle_{\mathcal{P}_q^i} = \langle u_j \rangle_{\mathcal{P}_q} = w_j. \quad (16)$$

Now, considering all the columns of $\mathbf{M}$, we can then obtain the weight vector $\mathbf{w}$ by solving the equation $\mathbf{wM} = \mathbf{w}$. This finalizes the proof of the Theorem 1.

B PROOF OF THEOREM 2

For simplicity without loss of generality, we here only consider the case with causal dependence between two attributes, which is illustrated by Fig. 5. According to the causal graph of $A \rightarrow B$, we can express the corresponding structural equations as below:

$$\begin{cases} B = F_B(A, Z_B) \\ A \perp\!\!\!\perp Z_B, A \sim \mathcal{P}_A, Z_B \sim \mathcal{P}_{Z_B} \end{cases}, \quad (17)$$

where Z_B generally indicates the disturbance term of the structural equation for B . However, the generated samples from Eq. 17 may not necessarily be identified with the causality $A \rightarrow B$ purely from the observational perspective, since there could exist a symmetric

case for $B \rightarrow A$ which derives the same joint distribution of A and B . Take the following two set of structural equations for example:

$$\begin{cases} B = \alpha A + Z_B, A \perp\!\!\!\perp Z_B \\ A \sim \mathcal{N}(\mu_A, \sigma_A^2) \\ Z_B \sim \mathcal{N}(\mu_{Z_B}, \sigma_{Z_B}^2) \end{cases} \quad \begin{cases} A = \beta B + Z_A, B \perp\!\!\!\perp Z_A \\ B \sim \mathcal{N}(\mu_B, \sigma_B^2) \\ Z_A \sim \mathcal{N}(\mu_{Z_A}, \sigma_{Z_A}^2) \end{cases}, \quad (18)$$

where α, β are constants, and μ, σ respectively denotes the mean and standard deviation of the normal distribution $\mathcal{N}$. Now, assuming we have the following relationships satisfied:

$$\begin{cases} \beta = \alpha \sigma_A^2 / (\alpha^2 \sigma_A^2 + \sigma_{Z_B}^2) \\ \mu_B = \alpha \mu_A + \mu_{Z_B}, \sigma_B^2 = \alpha^2 \sigma_A^2 + \sigma_{Z_B}^2 \\ \mu_{Z_A} = (1 - \alpha\beta) \mu_A - \beta \mu_{Z_B}, \sigma_{Z_A}^2 = (1 - \alpha\beta)^2 \sigma_A^2 + \beta^2 \sigma_{Z_B}^2 \end{cases},$$

the two set of structural equations in Eq. 18 then derives the exact same joint distribution for A and B , from which we cannot identify the related causal dependence with the generated samples.

Thus, to guarantee the causality we incorporate is able to be identified, we need to break the symmetry of potential structural equations. One promising way for our case is to utilize the properties of additive noise model (ANM) [14], which is proved to be able to generate the samples with related causal dependence identifiable. To prove a set of structural equations follow ANM, there are two key requirements for verification: (1) the transformation function F is non-linear; (2) the influence of the disturbance term can be isolated out of F in an additive manner. In the following, we verify our design in Theorem 2 is consistent with the ANM formulation.

- **Non-linear Transformation:** In our design, the transformation F is implemented with feed-forward modules, which consists of several dense layers for computation. As a result, the implemented transformation is thus a non-linear mapping. Specifically for Eq. 17, F_B is non-linear in our scenario.
- **Additive Disturbance:** For our case, the structural equation in Eq. 17 can be expressed as: $B = F_B(A \oplus Z_B)$, where $\oplus$ represents the direct sum, indicating the concatenation of A and Z_B before feeding into F_B . Since $A \perp\!\!\!\perp Z_B$, we know that there exists relevant transformations F'_B and F''_B , such that

$$B = F_B(A \oplus Z_B) = F'_B(A) + F''_B(Z_B), \quad (19)$$

based on the homomorphism property of direct sum. In Eq. 19, $F''_B(Z_B)$ can be further treated as a transformed disturbance. Thus, within our scenario, we see that Eq. 17 can isolate the influence of disturbance Z_B as an additive term.

Overall, we show that our design of G in Theorem 2 essentially builds a series of ANMs for counterfactual generation. With the identifiable causal dependence guaranteed by ANM, we finalize the proof of Theorem 2.

C DETAILS & ANALYSIS OF UMBRELLA SAMPLING SETTINGS

We utilize the umbrella sampling to obtain the training queries with rare values, aiming to effectively train the proposed MCS and let it well capture the potential hypothetical distributions. During implementation, we divide the particular sampling space into N different windows, where i -th window is appended with the umbrella profile u_i . In experiments, our biased samplers are implemented with the

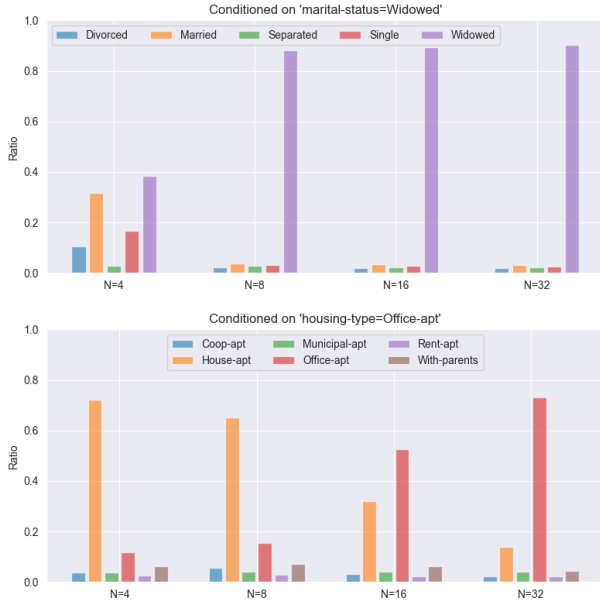


Figure 11: The conditional modeling performance of MCS with umbrella sampling on different number of windows.

Ensemble Sampler⁶, and N is set as 8, 32 respectively for the *Adult* and *Home_Credit* dataset. Besides, we employ 8 walkers for each umbrella-biased sampler, and run 1000 steps for each walker. The samplers will stop when the maximum Gelman-Rubin estimate falls below the threshold $\zeta = 0.01$.

Typically, when N is large enough, those rare values would have roughly the equal opportunities to be sampled compared with major ones, which basically balances the values appeared for MCS training. However, it is usually not acceptable with an overlarge N , because N also affects the training efficiency significantly. Increasing N will directly lead to the rise of computational complexity, referring Theo. 1, and will further result in the fact that each training batch consumes more time for updating. Thus, to achieve a reasonable trade-off on training between the effectiveness and efficiency, we need to select an appropriate number of windows N for umbrella sampling. Here, we attach some additional experimental results, considering the influences of N , shown in Fig. 11 and Fig. 12. According to the comparison, it is noted that MCS trained with more umbrella samplers shows a better modeling performance on hypothetical distributions. Meanwhile, the increase of N brings about heavier time cost for training, which becomes more significant with larger data scale and more training epochs. Therefore, properly choosing N for specific data should be an important pre-step before deploying MCS for interpretation in practice.

D SYNTHESIZER TRAINING WITH DIFFERENT DEPLOYED CLASSIFIERS

We show some additional results about the MCS training regarding to different deployed classifiers f (i.e., SVM, RF, MLP), and briefly discuss how f can potentially affect the training process.

⁶<https://emcee.readthedocs.io/en/stable/user/sampler/#emcee.EnsembleSampler>

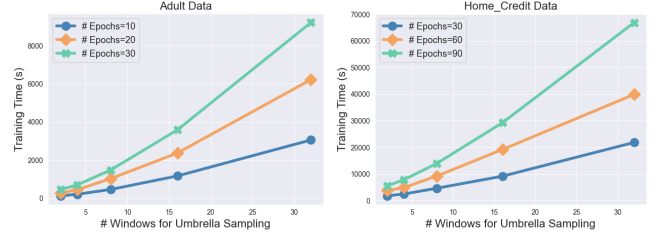


Figure 12: The time consumption of MCS training.

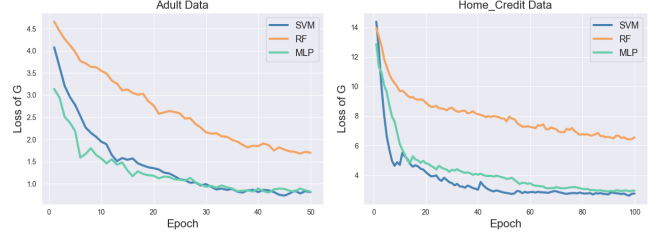


Figure 13: The training loss of generator G with different f .

In experiments, the dimension of the latent space (i.e., z) is set as 128, and D, G are designed with two-layer feed-forward modules whose intermediate dimensions are set as 256. Besides, the training batch size is fixed as 500, and the learning rate is 2×10^{-4} . The empirical results on the loss of generator G during training, with respect to different f , are illustrated in Fig. 13. Based on the training loss curve, we note that generator G converges faster when trained with deployed SVM and MLP, which typically takes less epochs to reach certain loss value compared with that under the RF case. This observation mainly results from the L^{ce} term in Eq. 6, which involves particular f to calculate the counterfactual loss. Thus, when decision boundaries of the deployed f are not smooth, the corresponding L^{ce} term may not be effectively minimized for training objectives, so that it would largely increase the difficulties of the optimizer in updating G . In our experiments, the deployed SVM and MLP classifier seem to have smoother boundaries than RF (intuitively validated by Fig. 6), and related training objectives are better optimized within certain epochs. From this perspective, it sheds light on the relationship between the counterfactual synthesizer and the deployed classifier, where smoother f would boost the MCS training towards better effectiveness.

E DATASET PREPROCESSING

In the conducted experiments of this paper, we preprocess the two real-world datasets employed (i.e., *Adult* and *Home_Credit*) for simplicity. The processing actions include the following aspects:

- *Remove the rows with missing values.* This action is mainly for the *Home_Credit* data, which contains many instances with *None*;
- *Delete some columns (attributes).* We ignore some columns which may not be that significant for prediction. For example, the 'fml-wgt' attribute in *Adult* data is removed in our experiments.
- *Merge some values.* We merge those values together which share overlap semantics. For example, in *Adult* data, value *Assoc-voc*, *Assoc-acdm* are merged as *Assoc*, and value 11th, ..., 1st-4th are all merged as *School*, within the attribute 'education'.