Applied Intelligence (2022) 52:12957–12974
https://doi.org/10.1007/s10489-022-03161-8

Rethinking the framework constructed by counterfactual functional model

Chao Wang[1] · **Linfang Liu**[1] · **Shichao Sun**[2] · **Wei Wang**[1]

Accepted: 29 December 2021 / Published online: 17 February 2022

Abstract

The causal inference represented by counterfactual inference technology breathes new life into the current field of artificial intelligence. Although the fusion of causal inference and artificial intelligence has an excellent performance in many various applications, some theoretical justifications have not been well resolved. In this paper, we focus on two fundamental issues in causal inference: probabilistic evaluation of counterfactual queries and the assumptions used to evaluate causal effects. Both of these issues are closely related to counterfactual inference tasks. Among them, counterfactual queries focus on the outcome of the inference task, and the assumptions provide the preconditions for performing the inference task. Counterfactual queries are to consider the question of what kind of causality would arise if we artificially apply the conditions contrary to the facts. In general, to obtain a unique solution, the evaluation of counterfactual queries requires the assistance of a functional model. We analyze the limitations of the original functional model when evaluating a specific query and find that the model arrives at ambiguous conclusions when the unique probability solution is 0. In the task of estimating causal effects, the experiments are conducted under some strong assumptions, such as treatment-unit additivity. However, such assumptions are often insatiable in real-world tasks, and there is also a lack of scientific representation of the assumptions themselves. We propose a mild version of the treatment-unit additivity assumption coined as M-TUA based on the damped vibration equation in physics to alleviate this problem. M-TUA reduces the strength of the constraints in the original assumptions with reasonable formal expression.

Keywords Causal effect · Counterfactual approach · Functional model · Treatment-unit additivity assumption

1 Introduction

Counterfactual inference, as an indispensable method of causal inference, helps create human self-awareness and imbue life experiences with meaning, which is embodied when we modify a factual prior event and then evaluate the consequences of that change [1]. In the classic Rubin causal model (RCM), counterfactual results usually refer to unobserved potential outcomes [2]. A typical application representative is *counterfactual queries* (CQs) [3]. A counterfactual query is a question of what kind of causality would arise if we artificially adopt the conditions contrary to the facts. Formally, the evaluation of CQs can be expressed as "*If $\mathcal{C}$ happened, would $\mathcal{B}$ have occurred?*", where $\mathcal{C}$ is the *counterfactual antecedent*.

CQs embody our reflections on what already happening in the real world. For example, in Fig. 1,[1] data released by Johns Hopkins University (JHU) shows that as of August 19, 2021, EST, the cumulative number of confirmed cases of COVID-19 (coronavirus disease 2019) in the United States amounted to 37,155,209 cases and the cumulative number of deaths amounted to 624,253 cases. The data also shows that the current cumulative number of confirmed cases in

✉ Chao Wang
cwang17@fudan.edu.cn

Linfang Liu
liulf19@fudan.edu.cn

Shichao Sun
bruce.sun@connect.polyu.hk

Wei Wang
wangwei1@fudan.edu.cn

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China

[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

[1] https://coronavirus.jhu.edu/map.html

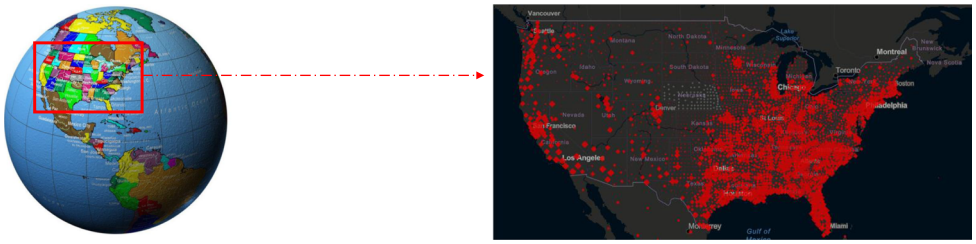

Fig. 1 COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at JHU

the United States accounts for about 17.75% of the more than 200 million confirmed cases worldwide; the cumulative number of deaths in the United States accounts for about 14.20% of the more than 4.3 million deaths worldwide. In face of the onslaught of the epidemic, one might ponder the following query: *If the U.S. has taken decisive measures, would the number of confirmed cases have been effectively controlled instead of spreading as wildly as it is now?*

Originally, most studies on counterfactual inferences (such as the query above) focus on the field of philosophy. Philosophers establish the form of a logical relationship constituting a logical world, which is consistent with the counterfactual antecedent and must be the *closest to the real world* (for the convenience of description, we call it the *closest world* approach) [4]. Further, Ginsberg [5] applies similar counterfactual logic to analyze problems of AI tasks, which relies on logic based on the *closest world* approach. However, the disadvantage of the *closest world* approach is that it lacks constraints on closeness measures.

Regarding the above issue, Balke and Pearl [3] are committed to explaining the *closest world* approach. Specifically, they suggest that turning a CQ into a probability problem, named, the *probabilistic evaluation of counterfactual queries* (PECQs). In other words, PECQs focus more on the probability of an event occurring in a specific CQ, rather than just outputting "True" or "False" (or "Yes" or "No", etc.) for this query. PECQs motivate us to deeply rethink counterfactual problems in many AI applications. For example, we know that COVID-19 has caused economic losses and increased unemployment in the United States [6]. An important reason is that the government has not dealt with the epidemic promptly.[2] Based on the facts that have already occurred, we may reflect on the following question, **CQ1**: *If the government issued effective policies in time to control the spread of COVID-19, would the unemployment rate in the United States still have raised?*

[2] https://www.nytimes.com/2020/05/28/business/unemployment-stock-market-coronavirus.html

Note that, in CQ1, there is a clear causal relationship, that is, COVID-19 has caused the unemployment rate in the United States to rise. Therefore, in response to CQ1, an essential task is to be able to evaluate the degree of belief in the counterfactual consequence (i.e., probability evaluation) after considering the facts that have already happened. In other words, it is equivalent to evaluating the probability of a potential (or counterfactual) outcome given the antecedent. Moreover, in **CQ1**, it is a fact that the COVID-19 sweeps the world and causes the unemployment rate in the United States to rise. Hence, we should focus on analyzing what is the probability that the unemployment rate in the United States will rise if there is no COVID-19? This is undoubtedly an influence on the government to make decisions. Therefore, evaluating counterfactual queries like these has far-reaching significance for practical application.

With the widespread application of causal inference in the field of AI [7, 8], the current popular method is to adopt the *functional model* (FM) [9] for inference. FM takes a CQ as an input and finally outputs the probability evaluation of the CQ by combining prior knowledge and internal inference mechanisms. The evaluation of CQs has benefited many research fields and tasks, such as the determination of person liable [10], marketing and economics [11], personalized policies [12], medical imaging analysis [13, 14], Bayesian network [7], high dimensional data analysis [15], abduction reasoning [16], the intervention of tabular data [8], epidemiology [17], natural language processing (NLP) [18, 19] and graph neural networks (GNN) [20, 21]. In particular, FM can provide powerful interpretability for machine learning model decisions [22–25], which is one of the most concerning issues in the Artificial Intelligence (AI) community today.

1.1 Motivation

Judea Pearl discusses the limitations of the current machine learning theory and points out that current machine learning models are difficult to be used as the basis for strong

AI [9]. An important reason is that the current machine learning approach is almost entirely in the form of statistics or "black box", which brings serious theoretical limitations to its performance [26]. For example, it is difficult for current smart devices to make counterfactual inferences. A large number of researchers are increasingly interested in combining counterfactual inference with AI [27, 28], such as explaining consumer behavior [29], the study of viral pathogenesis [30], and predicting the risk of flight delays [31]. In addition, counterfactual inference has shown advantages in improving the robustness of the model [32, 33] and optimizing text generation tasks [34] and classification tasks [35]. Although counterfactual inference has set off a new upsurge in the field of machine learning, a deeper understanding of the existing models and methods is notably lacking.

In our work, we focus on two basic aspects in the task of counterfactual inference. The first aspect focuses on the counterfactual framework and this aspect is related to the inference results of the model. The second aspect focuses on the preconditions for the counterfactual inference tasks. Specifically, the first aspect is based on a type of counterfactual approach (e.g., the functional model) in causal science. We analyze the credibility of some results obtained by using this counterfactual approach to evaluate CQs. Another aspect we are concerned about is the assumptions used in causal inference to estimate causal effects. Since causal effects depend on the potential results, however, we cannot observe all the potential outcomes of the experimental individual simultaneously (unobservable outcomes are usually called counterfactual outcomes). Therefore, some assumptions are often needed when estimating the causal effect. We pay attention to a commonly used strong assumption (i.e., the Treatment-Unit Additivity (TUA) assumption) and weaken it using some mathematical methods. Next, we specify the above two aspects to the following two issues (we use a real inference task (i.e., PECQs) as an example to explain the relationship between the two issues in Fig. 2.

1) *In the CQs tasks, although the output result of the FM is unique, this unique solution sometimes is ambiguous.* For example, in the task of evaluating the probability solution of CQs by FM, if the model predicts that the probability of a CQ is 0, the result may be ambiguous. In other words, although the probability value predicted by the model in this situation is 0, it is still possible that the event will happen. Intuitively, the existence of statistical uncertainty may cause ambiguity of the inference results. Dawid [36] proves that even if the statistical uncertainty can be eliminated, the inference may also produce ambiguity. Therefore, when ambiguity cannot be eliminated, we must consider what may cause ambiguity and how to avoid trouble caused by ambiguity.
2) *The assumptions used to estimate causal effects in the data are strong, which are often violated in real-world applications.* Some strong assumptions tend to constrain on individuals (e.g., individuals u in an experimental population $\mathcal{U}$) to obtain the ideal environment in an experiment. This neglects to obtain the equivalent form of the assumption directly from the *abstract level* (e.g., the experimental population $\mathcal{U}$, the dataset itself). In some practical applications of causal inference, a challenging task requires researchers to make causal inferences in the absence of data. For example, in RCM, the *causal effect* is described as $O_{t,u} - O_{c,u}$, where $O_{t,u}$ ($O_{c,u}$) is the result variable O displayed by subject (or individual) u under the **c**ontrol (c) group or **t**reatment (t) group. Unfortunately, we have no idea how to obtain $O_{t,u}$, and $O_{c,u}$ at the same time no matter how large the dataset is. This situation is also called the *fundamental problem of causal inference* (FPCI) [37].

Owing to the existence of FPCI, we can only apply additional assumptions on the data distribution to avoid it. Some typical assumptions are shown below:

- Stable Unit Treatment Value Assumption (SUTVA) [38], where each O of u is treated as an independent event;
- Assumption of Homogeneity (AOH) [39], which requires that for any individual u_i and $u_{j,j\neq i}$, and any intervention method t', $O_{t',u_i} = O_{t',u_{j,j\neq i}}$ always holds;
- Treatment-Unit Additivity (TUA) [36], some studies also call it *the assumption of constant effect* (AOCE). The TUA assumption constrains such an equivalence relationship that, for all individuals, the causal effect is the same for each individual under a defined intervention method.

$$\epsilon(u_1) = \epsilon(u_2) = \cdots = \epsilon(u_{|\mathcal{U}|}), \tag{1}$$

where $\epsilon(u_i)$ denotes the individual causal effect of $u_i \in \mathcal{U}$, and $|\mathcal{U}|$ is the cardinality of set $\mathcal{U}$. Apparently, AOH is stronger than TUA. Therefore, in the second aspect, we focus on TUA, aiming to obtain the milder TUA assumption.

To address the two issues mentioned above, in this paper, our contributions are three-fold:

- We focus on a basic problem in the FM and primarily analyze the evaluation method of [3]. We find that FM sometimes produces ambiguous output results for some CQs, even if the final output result is unique. One of the important reasons is that FM needs to calculate the

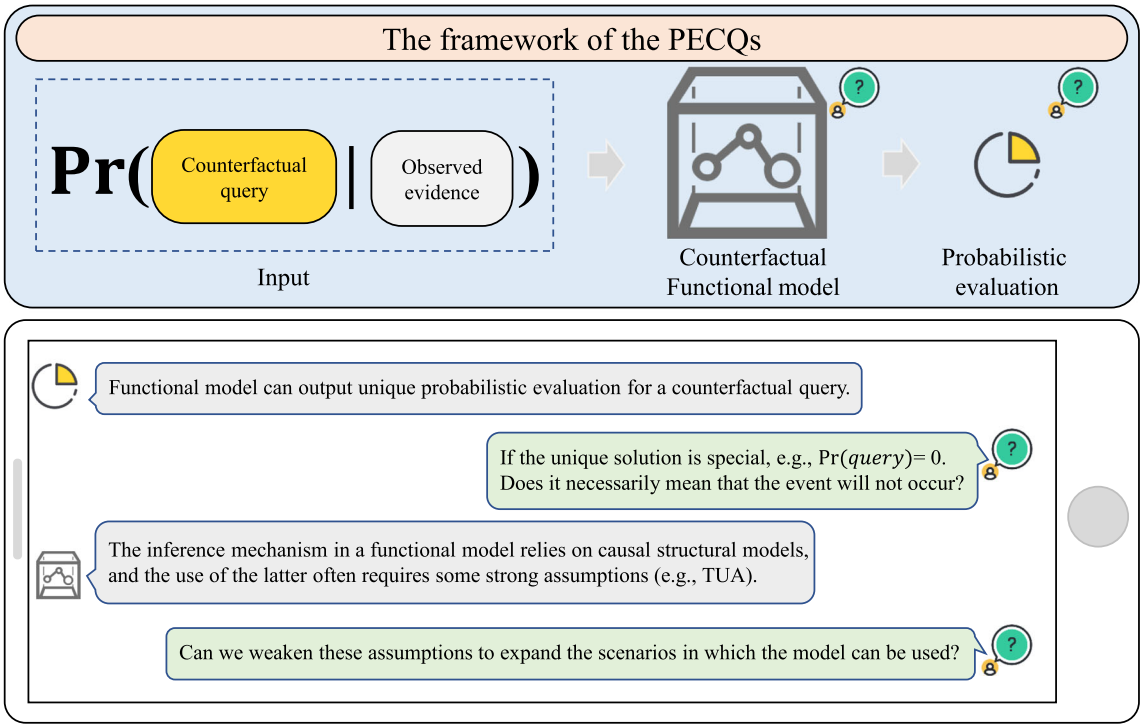

Fig. 2 The framework of the probabilistic evaluation of counterfactual queries: these two issues spread over the same inference task, and these two issues are independent of each other. However, for the same counterfactual inference task, the plausibility of the output affects, the user's confidence, and the strong assumptions premise determines the scope of the task

intersection between the two sets to get the final result when estimating the output probability. However, the intersection may be an empty set $\emptyset$, when estimating some special CQs.
- We provide a mild TUA assumption, called M-TUA, which incorporates the idea of the damped vibration equation.
- We prove theoretically that M-TUA can be applied to large datasets, and give a reasonable and rigorous mathematical description of this theory (see Theorem 1). Especially for some complex internal principles, we do not choose to use the "black box" method but hope to use M-TUA to try to reveal the complex internal relationship between certain parameters and assumptions and make some reasonable description and explanation.

1.2 Paper organization

The rest of this paper is organized as follows: In Section 2, we give the mathematical notation and their descriptions. In Section 3, we give a visualization of the FM inference mechanism and analyze the pitfalls of this inference mechanism based on concrete examples. In Section 4 and Section 5, we give a mild version of the TUA assumption (i.e., M-TUA), and theoretically prove the equivalent representation of the TUA assumption in the vector space and analyze the rationality and limitations of M-TUA. The comparison between TUA and M-TUA is in Section 6. Section 7 summarizes this paper.

2 Notation

In this section, the key mathematical notations and their descriptions are listed in Table 1.

3 Inference mechanism and result credibility analysis of FM

In this section, we first introduce the definition of PECQs [3], which is a probabilistic description of the counterfactual query. Second, we review the inference mechanism of FM in Fig. 3. Finally, we exhaustively analyze the inference mechanism in FM by some examples and find that when the probabilistic evaluation of a CQ is 0, the result causes unreliable guidance for decision-making.

3.1 Definition of PECQs

Definition 1 (Probabilistic Evaluation of Counterfactual Queries, PECQs [3]) The core idea of PECQs is to transform

Table 1 Key Notations and Descriptions

Notation	Description
$\emptyset$	the empty set
$\mathcal{R} = \{R_1, ..., R_n\}$	the set of variables R_i
$r_i/\hat{r}_i$	the value of R_i in the real/counterfactual world
$\{t, c\}$	t and c represent two different treatments (or intervention variables)
$\mathcal{U} = U_t \cup U_c$	a population with a huge number of units u_i
$U_t = \{u_1^t, ..., u_k^t\}$	the set of some units receiving treatment t
$U_c = \{u_1^c, ..., u_{k'}^c\}$	the set of other units receiving treatment c, i.e., $U_t \cap U_c = \emptyset$
$\mathbb{R}, \mathbb{Z}^+, \mathbb{C}$	the set of real numbers, positive integers, and complex numbers
$A^* \in \mathbb{C}$	the complex conjugate of $A \in \mathbb{C}$
$\lvert\mathcal{S}\rvert$	the cardinality of finite set $\mathcal{S}$, e.g., $\lvert\mathcal{R}\rvert = n$, $\lvert U_t\rvert = k$ and $\lvert U_c\rvert = k'$
$\{\cdot\}_n$	the finite set containing n elements, e.g., $\mathcal{R} = \{R_1, ..., R_n\} = \{R_i\}_n$
c_n	all unknown factors that may influence β in the inference mechanism of FM
$\Pr(c_n)$	the probability distribution of c_n in the inference mechanism of FM
L_{ao}	the Euclidean distance from point a to point o coordinate system
$\triangleq$	$p(x) \triangleq q(x)$ means that function $p(x)$ is equivalent to function $q(x)$

a CQ into a probabilistic evaluation problem, which can be formalized as:

$$\Pr(\hat{\beta}_1|\hat{\alpha}_1)_{|(\alpha_0,\beta_0)}, \tag{2}$$

where "$|(\alpha_0, \beta_0)$" represents the *evidence* (or observed data) we have observed in the real world, and the value of evidence be considered as a conditional probability (e.g., $(\alpha_0, \beta_0) \triangleq \Pr(\beta_0|\alpha_0) = p_0$). $\Pr(\hat{\beta}_1|\hat{\alpha}_1)$ is the counterfactual outcome that we need to infer based on evidence. The probabilistic evaluation of (2) can be obtained by the inference mechanism of FM [3] (i.e,. Figure 3).

Example 1 **CQ1** can be translated into (2) for evaluation. Specifically, for (α_0, β_0), we observe that there is an *ineffective policy* (i.e., α_0) that causes the *unemployment rate rise* (i.e., β_0); $\Pr(\hat{\beta}_1|\hat{\alpha}_1)$ indicates the probability of *unemployment rate falls* (i.e., $\hat{\beta}_1$) if we implement *effective policies* (i.e., $\hat{\alpha}_1$).

3.2 Inference mechanism of FM

The inference mechanism of FM is shown in Fig. 3. More detailed information on the inference mechanism of FM is

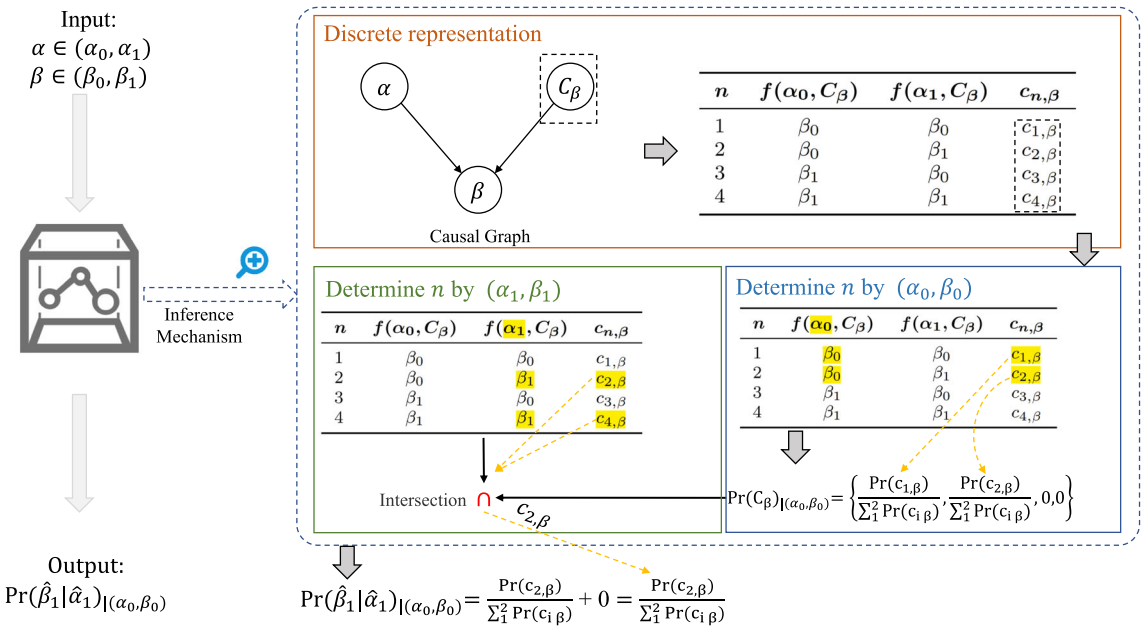

Fig. 3 The inference mechanism of FM when evaluating the **CQ1**

elaborated upon in [3], and we will not repeat it here in this section.

3.3 Analysis of the inference mechanism of FM

Although FM can output a unique solution for a CQ, however, we find that the results are not credible when the probability estimate of the FM output is 0. In other words, the output value of $\Pr(\cdot) = 0$ does not mean that the event will not occur. Next, we introduce some simple examples to reveal the untrustworthy guidance that this ambiguity may bring to the decision-making.

Example 2 **CQ2** [36]: *Patient $\mathcal{P}$ has a headache. Will it help if $\mathcal{P}$ takes aspirin?* The information we observe is that the current *patient has a headache* (denoted as β_0) and is *not taking aspirin* (denoted as α_0). Therefore, $\Pr(\hat{\beta}_1|\hat{\alpha}_1)_{|(\alpha_0,\beta_0)}$ is equivalent to the probability evaluation of **CQ2** (the query of this form like **CQ2** can also be called the *effects of causes* [36]). However, consider a situation (denoted as the variant of **CQ2**, which is abbreviated as **V-CQ2**) where *the patient still does not take aspirin. What is the probability of the headache disappearing?* It is equivalent to evaluating $\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)}$. If we still choose to use FM to estimate this query, we first determine the value of $n_{(\alpha_0,\beta_0)} \in \{1, 2\}$ ($n_{(\alpha_0,\beta_0)}$ refers to the value of n, which is determined according to (α_0, β_0)), and then we determine the new value of $n_{(\hat{\alpha}_1,\hat{\beta}_1)} \in \{3, 4\}$ ($n_{(\hat{\alpha}_1,\hat{\beta}_1)}$ refers to the value of n, which is determined according to $(\hat{\alpha}_1, \hat{\beta}_1)$). Finally, the evaluation of $\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)}$ is the sum of $\Pr(c_{3,\beta})_{|(\alpha_0,\beta_0)}$ and $\Pr(c_{4,\beta})_{|(\alpha_0,\beta_0)}$, i.e.,

$$\begin{aligned}\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)} &= \Pr(c_{3,\beta})_{|(\alpha_0,\beta_0)} + \Pr(c_{4,\beta})_{|(\alpha_0,\beta_0)} \\ &= 0 + 0 = 0 \end{aligned} \tag{3}$$

Why is the evaluation of V-CQ2 equal to 0, and what does this mean? 1) When using FM to estimate the results of **CQ1** and **V-CQ2**, a key step is to calculate the intersection of $N_{(\alpha_0,\beta_0)}$ and $N_{(\hat{\alpha}_1,\hat{\beta}_1)}$, where $N_{(\alpha_0,\beta_0)} = \{n_{(\alpha_0,\beta_0)}\}$ refers to the set of $n_{(\alpha_0,\beta_0)}$, which is determined by the observed evidence in the real world, and $N_{(\hat{\alpha}_1,\hat{\beta}_1)} = \{n_{(\hat{\alpha}_1,\hat{\beta}_1)}\}$ refers to the set of $n_{(\hat{\alpha}_1,\hat{\beta}_1)}$, which is updated in the counterfactual world. For example, $N_{(\alpha_0,\beta_0)} = \{1, 2\}$ and $N_{(\hat{\alpha}_1,\hat{\beta}_1)} = \{2, 4\}$ in **CQ1** can be derived from Fig. 3. Therefore, the probabilistic evaluation of **CQ1** is uniquely determined by $N_{(\alpha_0,\beta_0)} \cap N_{(\hat{\alpha}_1,\hat{\beta}_1)} = n = 2$. Hence, the probabilistic evaluation of **CQ1** is $\Pr(c_{2,\beta})_{|(\alpha_0,\beta_0)}$.

However, unlike **CQ1**, the $N_{(\hat{\alpha}_1,\hat{\beta}_1)}$ in **V-CQ2** is $\{3, 4\}$, which causes the probability evaluation of **V-CQ2** to be 0 (i.e., (3), because $N_{(\alpha_0,\beta_0)} \cap N_{(\hat{\alpha}_1,\hat{\beta}_1)} = \emptyset$). This probability estimate is not completely credible. The reason is that we cannot be sure whether the output results are derived from real predictive inferences or the processing of some special counterfactual queries (e.g., **V-CQ2**) by the inference mechanism. Therefore, when the probabilistic evaluation of a CQ is 0, the decision based on this result is not credible, that is, the result is ambiguous.

2) In addition, in **V-CQ2**, α_0 does not constitute a counterfactual condition, it still belongs to the assumptions in the real world, in this case, the $\hat{\beta}_1$ is also known evidence in the real world, i.e., $\hat{\beta}_1 = \beta_1$. Hence, we have

$$\begin{aligned}\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)} &= \Pr(\beta_1|\alpha_0)_{|(\alpha_0,\beta_0)} \\ &= 1 - \Pr(\beta_0|\alpha_0) = 1 - p_0, \end{aligned} \tag{4}$$

which contradicts with the result of (3). This shows that α_0 does not constitute an intervention that affects the outcome of the counterfactual world. Therefore, the estimated value of (3) obtained by FM violates the counterfactual consistency rule [40].

The impact of ambiguity in inference results on decision-making We discuss the impact of the unique solution on decision-making by two examples as follows:

Example 3 In predicting the probability value of 0.8 or 0.9 for an earthquake to occur at a certain location, there is little difference in decision-making for this probability. However, when the probability of an earthquake is estimated to be 0 and unique, it is essential for us to verify its rationality, because this may directly lead to the need for the corresponding deployment. In other words, how confident are we to ensure that there will be no earthquake based on the prediction of FM? Therefore, the fact that there exist queries that cannot be answered using FM does not mean that the evaluation of these queries is meaningless.

Example 4 **CQ3**:*The murderer assassinated President Kennedy, if the assassination had failed, would Kennedy still be alive?* Formally, if the shot hits the target (α_0) with a high probability (p_0) that the hit target will die (β_0), then we estimate $\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)}$ =? We will eventually get $\Pr(\hat{\beta}_1|\alpha_0)_{|(\alpha_0,\beta_0)} = 0$ using FM (the prediction process is similar to predicting **V-CQ2**). Obviously, if the assassination failed (that is, the shot was successfully fired but did not cause the target to die) and Kennedy is still alive, this situation may affect the assassin's further decisions and deployment. For Kennedy's team, this may affect the deployment of security measures for similar activities. Therefore, when the estimated result of a CQ is 0, the result cannot provide credible and sufficient opinions for decision-making.

A straightforward solution Through the above series of analyses, it is not difficult to find that when the probability of a CQ is evaluated as 0, for this situation, further

verification and analysis are indispensable. Because the inference mechanism of FM itself will inevitably introduce ambiguity for the evaluation result of $\Pr(\cdot) = 0$. Since the evaluation of the FM determines the final output solution through the intersection between two sets, there is a certain probability that the intersection is an empty set.

A straightforward solution is that if an empty set appears in the estimation process, we need to stop using the FM for estimation because the above analysis shows that we cannot define the empty set as $\Pr(\cdot) = 0$. Therefore, when this happens, we should estimate the output probability in the real world instead of the counterfactual world to avoid the appearance of ambiguous results. In this case, $\Pr(\cdot) = 0$ plays a role in prompting a replacement prediction strategy. Therefore, to comply with the counterfactual consistency rule, we must use the prior probability (4) (i.e.,$1 - p_0$) to replace $\Pr(\cdot) = 0$.

4 The mild treatment-unit additivity assumption

For the second reflection in Fig. 2, in this section, we analyze the TUA assumption, which is often used as a strong prerequisite for estimating causal effects in data. We first review the potential outcome framework (Section 4.1), individual causal effect (Section 4.2), the definition of TUA (Section 4.3), and provide an equivalent description of TUA utilizing vectorization (Section 4.4). Second, based on the idea of the Damped Vibration Equation (DVE) [41], we propose a mild TUA assumption (called M-TUA) (Section 4.5). M-TUA not only weakens the original assumption but also has good mathematical properties and interpretability .

Our main conclusion in this section is presented based on two lemmas, and the specific proof process is mainly divided into the following two steps. First, we describe the relationship between TUA and ICE in the counterfactual approach, and we explore the equivalence of ICE and **r**esidual **c**ausal **e**ffect (RCE) in the TUA assumption (i.e., Lemma 1). Second, we innovatively introduce the definitions of *positive effects* and *negative effects*, and on this basis, we obtain the equivalent form of TUA in vector space by Lemma 2.

4.1 Potential outcome framework

According to the viewpoint of Rubin [42], there is an *intervention* in the causal inference, which means that there is no cause and effect without intervention, and one intervention state corresponds to a potential outcome. When the intervention state is realized, we can only observe the potential outcomes in the realization state, that is, we cannot observe the potential outcomes (i.e, counterfactual outcomes) in the counterfactual world (e.g., O_{c,u_2} in Example 6). This situation where all potential outcomes of units cannot be observed simultaneously is also called FPCI we mentioned earlier. Formally, for binary intervention variables, let $d \in \{t = 1, c = 0\}$, the observation outcome O_{d,u_i} and the potential outcome Y_o can be expressed by the following formula:

$$Y_o = d \cdot O_{1,u_i} + (1-d) \cdot O_{0,u_i} = \begin{cases} Y_o = O_{1,u_i}, \text{ if } d = 1 \\ Y_o = O_{0,u_i}, \text{ if } d = 0. \end{cases} \tag{5}$$

Where $O_{d,u_i} \in \{O_{t,u_i}, O_{c,u_i}\}$ represents the potential outcome of treatment $d \in \{t, c\}$ on unit $u_i \in \mathcal{U}$.

For a more intuitive description, we focus on the following 2-dimensional Gaussian distribution model

$$G(O_{t,u_i}, O_{c,u_i}) \sim \mathcal{N}(\mu_t, \mu_c, \sigma_t, \sigma_c, \rho). \tag{6}$$

Specifically, we introduce the following example [36] and use it as a basic background for subsequent analysis.

Example 5 Given the pair (O_{t,u_i}, O_{c,u_i}), O_{t,u_i}, and O_{c,u_i} are **i**ndependent and **i**dentically **d**istributed (i.i.d.), each with the 2-dimensional Gaussian distribution with means (μ_t, μ_c), $\sigma_c = \sigma_t = \sigma_o$ (for simplicity of calculation, we assume that the distribution has a common variance σ_o), and the correlation $\rho \in (0, 1)$. Furthermore, we use the mixed model to describe the specific structure, i.e.,

$$\begin{aligned} & O_{d,u_i} \triangleq \mu_d + \tau_{u_i} + \lambda_{d,u_i} \\ & \text{s.t.} \begin{cases} \mu_d = \mu_t \text{ or } \mu_c \\ \tau_{u_i} \sim \mathcal{N}(0, \sigma_\tau) = \mathcal{N}(0, \rho\sigma_o) \\ \lambda_{d,u_i} \sim \mathcal{N}(0, \sigma_\lambda) = \mathcal{N}(0, (1-\rho)\sigma_o) \end{cases}, \end{aligned} \tag{7}$$

where μ_d indicates the *treatment effects* applicable to all units. τ_{u_i} represents the effect on unit $u_i \in \mathcal{U}$, called *unit effects*, and this effect applies to all units, i,e., $\tau_{u_i} = \tau_{u_{j,j\neq i}}$. λ_{d,u_i} stands for the effect between treatment and unit, called *unit-treatment interaction*. This internal mechanism reveals the change from one treatment to another for unit u_i. τ_{u_i} and λ_{d,u_i} are independent random variables.

4.2 Individual causal effect

Dawid [36] adopts the model of (7) to analyze the pros and cons of the counterfactual based on the idea of decision-making and mentions an assumption that is often used in the counterfactual analysis, which is called TUA (Definition 2). As the TUA assumption has strong constraints on data, it will lead to a reduction in the practicability and scope of use of TUA. *Hence, in this paper, another goal of a study is to design a mild TUA assumption that constrains the dataset itself or the experimental population as a whole, rather than a strong constraint on each individual, as in*

the traditional TUA assumption. In the rest of this section, we try to optimize TUA to make it have a broader scope of application in the context of large data.

Specifically, we first analyze the individual and average causal effect based on (7). In an experimental study, the ***individual causal effect*** (ICE) is the basic object (or a basic measure). It describes the differences in various potential outcomes of a given unit $u_i \in \mathcal{U}$ under all possible treatments $d \in \{t, c\}$. Generally, for one unit $u_i \in \mathcal{U}$, the ICE can be represented as

$$\epsilon(u_i) \triangleq O_{t,u_i} - O_{c,u_i}. \tag{8}$$

For different tasks, the ICE can also have other forms of description, such as $\epsilon(u_i) = \log\left(O_{t,u_i}/O_{c,u_i}\right)$. Therefore, from a broader perspective, the subtraction in the definition of ICE may not necessarily be a subtraction in $\mathbb{R}$. Note that no matter which form is used, only one potential outcome can be observed [43]. Researchers usually do not pay attention to ICE directly, but focus on the average value of the causal effect of all units, that is, ACE, also known as ***average treatment effect*** (ATE). ACE can be expressed by the following formula,

$$\epsilon_{\text{ACE}}(u_i) \triangleq \mathbb{E}(\epsilon(u_i)) = \mathbb{E}\left(O_{t,u_i} - O_{c,u_i}\right). \tag{9}$$

Apparently, in (7), $\epsilon_{\text{ACE}}(u_i) = \mu_t - \mu_c$.

Limitations of the counterfactual approach focused on ICE We utilize the above Example 5 for our analysis. Specifically, according to (7) and (8), we have that,

$$\begin{aligned}\epsilon(u_i) = O_{t,u_i} - O_{c,u_i} &= (\mu_t - \mu_c) + (\lambda_{t,u_i} - \lambda_{d,u_i}) \\ &= \epsilon_{\text{ACE}}(u_i) + \Delta(\lambda_{u_i}),\end{aligned} \tag{10}$$

where $\Delta(\lambda_{u_i}) \triangleq \lambda_{t,u_i} - \lambda_{d,u_i}$ is called ***residual causal effect*** (RCE) [36]. It is easy to verify that $\Delta(\lambda_{u_i}) \sim \mathcal{N}(0, 2(1-\rho)\sigma_o)$. Thus, according to (7)-(9), we could obtain the distribution of ICE as follows:

$$\epsilon(u_i) \sim \mathcal{N}(\epsilon_{\text{ACE}}(u_i), 2(1-\rho)\sigma_o). \tag{11}$$

However, in (11), $2(1-\rho)\sigma_o$ cannot be inferred from observed data and has nothing to do with the size of the data. Because even if the marginal distributions of O_{t,u_i} and O_{c,u_i} are known, the joint distribution of random variables $G(O_{t,u_i}, O_{t,u_i})$ cannot be determined, and the marginal distribution of the Gaussian distribution does not depend on the parameter σ_o.

Moreover, according to (7), we have

$$\begin{cases} 2(1-\rho)\sigma_o = 2\sigma_\lambda \in (0, 2\sigma_o), & \text{if } \rho \in (0, 1) \\ 2(1-\rho)\sigma_o = 2\sigma_\lambda \in (2\sigma_o, 4\sigma_o), & \text{if } \rho \in (-1, 0) \end{cases}. \tag{12}$$

(12) indicates that different values of ρ determine different variances of the distribution of $\epsilon(u_i)$. We can only get a range of σ_λ, and a different ρ will lead to a different σ_λ, which will cause a variety of uncertain results for reasoning.

For example, we can use (11) to estimate the ICE of the new unit u_{new}. Because inferring $\epsilon(u_{new})$ is equivalent to inferring $\epsilon_{\text{ACE}}(u_i)$ and $2(1-\rho)\sigma_o$ under (11). Unfortunately, we cannot accurately determine the value of $2(1-\rho)\sigma_o$.

Example 6 (Calculation of causal effect parameters (i.e., ICE, ACE) in the ideal case). In Table 2, we construct a simple example to demonstrate the calculation of the causal effect parameters, such as ICE, ACE. Suppose a population contains four subjects, labeled as u_1, u_2, u_3 and u_4, respectively. For each u_i, the potential outcomes in both intervention states are known (in reality only one potential outcome can be observed). Where individuals 1 and 2 are in the intervention group (i.e., the set of some units receiving treatment t) and individuals 3 and 4 are in the control group (the set of some units receiving treatment c).

According to Table 2, we can obtain:

$$\begin{aligned}\epsilon_{\text{ACE}}(u_i) &= \mathbb{E}(O_{t,u_i} - O_{t,u_i}) \\ &= \frac{1}{4}(30 + 0 + 10 + 0) = 10.\end{aligned} \tag{13}$$

Meanwhile, based on the information in Table 2, we can further obtain information on two other causal effect parameters, one is *average treatment effect for the treated* (ATT) and the other is *average treatment effect for the control* (ATC). Where,

$$\epsilon_{\text{ATT}}(u_i) = \mathbb{E}(O_{t,u_i} - O_{t,u_i} | d = t) = \frac{1}{2}(30+0) = 15. \tag{14}$$

and

$$\epsilon_{\text{ATC}}(u_i) = \mathbb{E}(O_{t,u_i} - O_{t,u_i} | d = c) = \frac{1}{2}(10+0) = 5. \tag{15}$$

Unfortunately, in the real world, the **boldface numbers** (e.g., O_{c,u_2}, O_{t,u_3}) in Table 2 are not observable to us. The reason is that the treatment received by subject u_2 is $d = t$, we can not observe the potential outcome of u_2 receiving treatment $d = c$ at the same time. Therefore, in the real world, the calculation and estimation of the causal effect parameters require additional constraints (e.g., Treatment-unit additivity assumption (Definition 2) to be imposed on the data.

Table 2 Causal effect parameters

Subject	O_{t,u_i}	O_{c,u_i}	O_{t,u_i}	d	$O_{t,u_i} - O_{c,u_i}$
u_1	30	**0**	30	t	30
u_2	10	**10**	10	t	0
u_3	**10**	0	0	c	10
u_4	**10**	10	10	c	0

4.3 Treatment-unit additivity

In summary, the POF focuses on the inference of causal effects but does not explain the mechanism of influence between variables [44]. A computational bottleneck is the prediction of parameter ρ through the marginal distribution. Therefore, in the task of using the causal model for inference, additional constraints (e.g., Example 7) are usually required to ensure that the inference result is obtained under this constraint.

Example 7 Under the TUA, $\epsilon(u_{new}) = \epsilon_{\text{ACE}}(u_i)$ implies that $\rho = 1$.

Definition 2 (Treatment-Unit Additivity (TUA) [36]). The TUA assumption is to deal with the non-uniformity of data through a strong processing method. Specifically, TUA requirements that $\epsilon(u_i)$ in $\epsilon(u_i) \triangleq O_{t,u_i} - O_{c,u_i}$ has the same effect on all units in $\mathcal{U}$, e.g., $\epsilon(u_1) = \epsilon(u_2) = \cdots = \epsilon(u_{|\mathcal{U}|}) = \epsilon_{\text{ACE}}(u_i)$.

TUA can be equivalently regarded as the ***a**ssumption **o**f **c**onstant **e**ffect* (AOCE). For example, we can set $\epsilon(u_i) = \epsilon(u_{j,j\neq i})$= *a specific constant* (e.g., $\epsilon_{\text{ACE}}(u_i)$. Generally speaking, AOCE uses the average effect in the sample to estimate the causal effect. Next, we will give a simple example to demonstrate the relation between TUA and ACE and the application of TUA.

Example 8 Considering a fundamental problem of causal inference, let u_1 be a patient. We want to know whether certain medication has a therapeutic effect on u_1. Suppose that the data about patient u_1 is shown in Table 3.

According to Table 3, we only know that $O_{t,u_1} = 13$. Due to the existence of FPCI, we cannot simultaneously observe the effects of u_1 taking the medication and not taking the medication. Therefore, we rely on adding additional constraints (i.e., TUA) to estimate the value of $O_{c,u}$.

Suppose we also have additional data (as shown in Table 4), we can then use TUA assumption to infer the values of O_{c,u_i} and $O_{t,u_i} - O_{c,u_i}$ $(i = 1, 2, 3, 4, 5)$. For example, according to

$$\epsilon(u_i) = O_{t,u_i} - O_{c,u_i} = \epsilon_{\text{ACE}}(u_i) = -1, \tag{16}$$

we can obtain the following complete prediction data (see Table 5).

Table 3 The data of u_1, where O_{c,u_1} and $O_{t,u_1} - O_{c,u_1}$ are unknown

Subject	O_{t,u_i}	O_{c,u_i}	$O_{t,u_i} - O_{c,u_i}$
u_1	13	?	?

Table 4 Additional information about all u_i

Subject	O_{t,u_i}	O_{c,u_i}	$O_{t,u_i} - O_{c,u_i}$
u_1	13	?	?
u_2	?	12.5	?
u_3	10	?	?
u_4	?	13	?
u_5	?	12	?
mean	11.5	12.5	−1

4.4 Equivalent form of TUA

TUA assumes that the causal effect $\epsilon(u_i)$ has the same effect on all units in $\mathcal{U}$, e.g., $\epsilon(u_i) = \epsilon_{\text{ACE}}(u_i), i \in [1, ..., |\mathcal{U}|]$. Unfortunately, as a commonly used prerequisite, TUA is a strong assumption, which cannot be tested on observable data and lacks a more transparent explanation in the real world [36]. This leads to some interesting questions worth exploring, such as:

- *For applications of TUA, how to obtain a mild version of the TUA assumption to make the TUA more broadly applicable?*
- *For interpretability of TUA, based on the TUA assumption (or a mild TUA), how to establish a formal expression to describe the impact of the main factors inside the data on estimating ICE?*

To address these issues, next, we first provide an equivalent form of the TUA assumption under the 2-dimensional Gaussian distribution (i.e., Lemma 1).

Lemma 1 *If the data follows a Gaussian distribution as Example 5, then the TUA assumption has the following equivalent form, i.e.,*

$$\begin{aligned} &\underbrace{\epsilon(u_1) = \epsilon(u_2) = \cdots = \epsilon(u_q)}_{TUA} \\ &\Longrightarrow \lim_{q \to q'} \sum_{q} \left(\Delta(\lambda_{d,u_i}) - \Delta(\lambda_{d,u_{j,j\neq i}})\right) = 0. \end{aligned} \tag{17}$$

Table 5 Assignment mechanism based on TUA assumption with $\epsilon_{\text{ACE}}(u_i) = -1$

Subject	O_{t,u_i}	O_{c,u_i}	$O_{t,u_i} - O_{c,u_i}$
u_1	**13**	**14**	**−1**
u_2	11.5	12.5	−1
u_3	10	11	−1
u_4	12	13	−1
u_5	11	12	−1
mean	11.5	12.5	−1

Where $u_i, u_{j,j\neq i} \in \mathcal{U}$, $i, j \in [1, ...q]$, $q = |\mathcal{U}|$, q' *is a sufficiently large positive integer* ($q \ll q'$). $\Delta(\lambda_{d,u_i}) = \lambda_{t,u_i} - \lambda_{c,u_i}$.

Proof Given two units u_i and $u_{j,j\neq i}$, according to (7) and (8), we have that

$$\begin{aligned}\epsilon(u_i) - \epsilon(u_{j,j\neq i}) &= \left(\lambda_{t,u_i} - \lambda_{c,u_i}\right) - \left(\lambda_{t,u_{j,j\neq i}} - \lambda_{c,u_{j,j\neq i}}\right) \\ &\triangleq \Delta(\lambda_{d,u_i}) - \Delta(\lambda_{d,u_{j,j\neq i}}). \end{aligned} \tag{18}$$

Hence, a reasonable idea based on (18) is that we can shift our attention from the constraint on λ_{d,u_i} to constraint on RCE $\Delta(\lambda_{d,u_i})$. Note that the predicted average value of $O_{t,u_i} - O_{c,u_i}$ (denoted as $\overline{O_{t,u_i}} - \overline{O_{c,u_i}}$) will be closer to $\mathbb{E}\left(O_{t,u_i} - O_{c,u_i}\right)$ if the size of the data is large enough. Therefore, $\epsilon_{\text{ACE}}(u_i)$ can be identified, from a large experiment, as $\overline{O_{t,u_i}} - \overline{O_{c,u_i}}$. This means that the impact of $\Delta(\lambda_{d,u_i})$ on the data may be related to the size of the data.

Given a group $\mathcal{U}_q = \{u_1^d, u_2^d, ..., u_q^d\}$ containing q units, where $u_{j,j\neq i}^d$ means the unit $u_{j,j\neq i}$ will receive treatment d. We can assign "treatment" through Randomized Controlled Trials (RCT) and collect all potential outcomes, i.e., $\mathcal{O}_t = \{O_{t,u_{j,j\neq i}}\}_k$ and $\mathcal{O}_c = \{O_{c,u_{j,j\neq i}}\}_{q-k}$.

Suppose that q is a large positive integer and naturally let $\mathbb{E}\left(O_{t,u_i} - O_{c,u_i}\right) = \overline{O_{t,u_i}} - \overline{O_{c,u_i}}$, we have that

$$\begin{aligned}&\epsilon_{\text{ACE}}(u_i) = \left(\frac{1}{k}\sum_{j=1}^{k} O_{t,u_{j,j\neq i}} - \frac{1}{q-k}\sum_{j=k+1}^{q} O_{c,u_{j,j\neq i}}\right) = \hat{\epsilon} \\ &\text{s.t. } O_{t,u_{j,j\neq i}} \sim \mathcal{N}(\mu_t, \sigma_o),\ O_{c,u_{j,j\neq i}} \sim \mathcal{N}(\mu_c, \sigma_o). \end{aligned} \tag{19}$$

Where $\frac{1}{k}\sum_{j=1}^{k} O_{t,u_{j,j\neq i}}$ represents the average of the responses of k units receiving treatment t, and $\frac{1}{q-k}\sum_{j=k+1}^{q} O_{c,u_{j,j\neq i}}$ is the average of the responses of $q-k$ units receiving treatment c. q, k, and $q-k$ are both large numbers. Therefore, $\epsilon_{\text{ACE}}(u_i) = \hat{\epsilon}$ is estimable and close to the true value.

Next, we employ the TUA constraint on (18), which is equivalent to the setting $\epsilon(u_i) - \epsilon(u_{j,j\neq i}) = 0$. According to (18), it is unnecessary for us to constrain every $\lambda_{d,u}$ to a fixed value if q is large enough (e.g., $q \longrightarrow q'$). The alternative solution is that we consider the difference between two $\Delta(\lambda_{d,u})$, and formally characterize $\Delta(\lambda_{d,u_i}) - \Delta(\lambda_{d,u_{j,j\neq i}})$ so that it gradually approaches 0 when q is a large number. Therefore, in the case of the considered RCE, we obtain the equivalent form of the TUA assumption,

$$\lim_{q\to q'} \sum_{q} \left(\Delta(\lambda_{d,u_i}) - \Delta(\lambda_{d,u_{j,j\neq i}})\right) = 0, \tag{20}$$

which proves the lemma. □

4.5 The properties of $\Delta(\lambda_{d,u_i})$ in 2-dimensional vector space

Further, we will analyze the properties of TUA in 2-dimensional vector space. Through the above analysis, it is not difficult to find that both the TUA and the equivalent form given by Lemma 5 are only numerical constraints (e.g., $\epsilon(u_i) = \epsilon(u_{j,j\neq i})$, $\Delta(\lambda_{d,u_1}) - \Delta(\lambda_{d,u_2})$). In other words, neither the TUA assumption itself nor Lemma 5 reflects their internal influence on the data. To explore the internal influence of TUA on the data, our core idea is to transform the original TUA assumption of constraints on values (i.e., scalars) into constraints on vectors. Specifically, we analyze the TUA assumption by vectorizing λ_{d,u_i} (i.e., Lemma 2) and introducing a definition of the *positive and negative effects* of λ_{d,u_i} (i.e., Definition 3) on the data.

Lemma 2 *For any* λ_{d,u_i}, *let* $\Delta_+(\lambda_{d,u_i})$ *denote the positive effect of* λ_{d,u_i} *on the data, and* $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ *denote the negative effect of* $\lambda_{d,u_{j,j\neq i}}$ *on the data. Then the TUA assumption has the following equivalent form in the vector space, i.e.,*

$$\begin{aligned}&\underbrace{\epsilon(u_1) = \epsilon(u_2) = \cdots = \epsilon(u_q)}_{TUA} \\ &\Longrightarrow \lim_{q\to q'} \left(\sum_{q^+} \Delta_+(\lambda_{d,u_i}) - \sum_{q^-} \Delta_-(\lambda_{d,u_{j,j\neq i}})\right) = 0, \end{aligned} \tag{21}$$

where $q^+ + q^- = q$.

Before proving Lemma 2, we need to introduce the definition of the vectorization of λ_{d,u_i}, positive effects, and negative effects.

Definition 3 (**The vectorization of** λ_{d,u_i}.) Let $\lambda_{d,u_i} = L_{ao}$ represent the distance from a certain point a to the point o in the coordinate system (e.g., in Fig. 4a, L_{ao} represents λ_{d,u_i} and L_{bo} represents $\lambda_{d,u_{j,j\neq i}}$). The vectorization of λ_{d,u_i} refers to assigning the characteristics of a vector to λ_{d,u_i} to describe the possible positive or negative effect of λ_{d,u_i} on the data. As shown in Fig. 4b, for each λ_{d,u_i},

$$\Delta(\lambda_{d,u_i}) \triangleq \begin{cases} \text{positive effect}, & \text{if } \Delta(\lambda_{d,u_i}) \text{ is in the first, second quadrants.} \\ \text{negative effect}, & \text{if } \Delta(\lambda_{d,u_i}) \text{ is in the third, fourth quadrants.} \end{cases} \tag{22}$$

As shown in Fig. 4-(*c*) and (*d*), for $\sum \lambda_{d,u_i}$,

$$\sum \Delta(\lambda_{d,u_i}) \triangleq \begin{cases} \text{positive effect}, & \text{if } \sum \Delta(\lambda_{d,u_i}) \text{ are in the first, second quadrants.} \\ \text{negative effect}, & \text{if } \sum \Delta(\lambda_{d,u_i}) \text{ are in the third, fourth quadrants.} \end{cases} \tag{23}$$

There is a one-to-one correspondence between positive effects and negative effects. In other words, if a positive effect "+" exists, there must be a negative effect "-" corresponding to it.

Rationality analysis According to Definition 3, we transform the original TUA assumption of constraints on the scalars into constraints on vectors. For example, some individuals insist on eating *nuts* in actual life because nuts are good for their health (i.e., positive effect), but some people are allergic to nuts, and eating them will bring pains and even life-threatening effects (i.e., negative effect). Therefore, we argue that it is necessary to consider the positive or negative effects of $\lambda_{d,u}$. Definition 3 provides an intuitive representation of positive/negative effect in the vector space, and according to the definition, next, we give a proof of Lemma 2 as follows.

Proof For ease of understanding, we will combine Fig. 4 for the proof. Considering the representation of λ_{d,u_i} in a 2-dimensional plane. As shown in Fig. 4a, we first represent λ_{d,u_i} as the Euclidean distance in the plane, i.e.,

$$L_{ao} = \Delta(\lambda_{d,u_i}), \text{ and } L_{bo} = \Delta(\lambda_{d,u_{j,j\neq i}}). \tag{24}$$

According to Lemma 1, $\epsilon(u_i) = \epsilon(u_{j,j\neq i})$ can be regarded as $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$. Then, we can use $L_{ao} = L_{bo}$ to equivalently describe $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$.

Second, we consider the representation of the TUA in 2-dimensional vector space. According to Definition 3, we can vectorize $\lambda_{d,u}$. The meaning of vectorization is to give each $\Delta(\lambda_{d,u_i})$ a measure, which aims to describe the positive or negative effects of $\Delta(\lambda_{d,u_i})$ on the data. In order to maintain consistency with the original TUA assumption, we assume that $|\Delta(\lambda_{d,u_i})| = |\Delta(\lambda_{d,u_{j,j\neq i}})|$. For instance, as shown in Fig. 4b, let $|\Delta_+(\lambda_{d,u_i})|$ ($|\Delta_-(\lambda_{d,u_{j,j\neq i}})|$) denote the positive (negative) effect of $\Delta(\lambda_{d,u_i})$ on the data, although $|\Delta(\lambda_{d,u_i})| = |\Delta(\lambda_{d,u_{j,j\neq i}})|$, $\Delta(\lambda_{d,u_i}) \neq \Delta(\lambda_{d,u_{j,j\neq i}})$.

Third, we consider extending $\Delta(\lambda_{d,u_i})$ to the entire dataset. Since the background of our research is in the context of large datasets, we implied a condition here, that is, in the entire data, the positive effects $\sum \Delta_+(\lambda_{d,u_i})$ and negative effects $\sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$ on data generation are basically the same. Furthermore, since $|\Delta(\lambda_{d,u_i})| = |\Delta(\lambda_{d,u_{j,j\neq i}})|$, we can visualize the entire data as a circle in a 2-dimensional plane, where $|\Delta(\lambda_{d,u_i})| = |\Delta(\lambda_{d,u_{j,j\neq i}})| = r$.

Intuitively, under the TUA constraint, $\sum \Delta_+(\lambda_{d,u_i}) = \sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$ always holds. However, $\sum \Delta_+(\lambda_{d,u_i}) = \sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$ does not necessarily have to be under a strong constraint of $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$ to hold. In other words, in Fig. 4-(*b*), it is sufficient that the area of *red* is the same as the area of *blue*. Therefore, we can relax the restriction on $\Delta(\lambda_{d,u_i})$ by only assuming $\sum \Delta_+(\lambda_{d,u_i}) = \sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$ without $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$. In summary, we obtain the following conclusion based on TUA, i.e.,

$$\lim_{q\to q'} \left(\sum_{q^+} \Delta_+(\lambda_{d,u_i}) - \sum_{q^-} \Delta_-(\lambda_{d,u_{j,j\neq i}}) \right) = 0, \tag{25}$$

which proves the lemma. □

Rationality analysis The traditional TUA strongly constrains all λ_{d,u_i} (or $\epsilon(u_i)$) to be the same for $u_i \in \mathcal{U}$, which undoubtedly ignores the effect of λ_{d,u_i} on the data and the estimated ICE. However, ignoring this effect by applying TUA does not mean that the effect of λ_{d,u_i} on the data does not exist. Therefore, we did not directly ignore this potential impact but pioneered to represent it by introducing the vectorization method (i.e., positive and negative effects in Definition 3). In addition, Lemma 2 relaxes the constraint on the data to the level of the entire dataset $\mathcal{U}$ rather than imposing a strong constraint on each unit u_i. Therefore, Lemma 2 can be considered as an equivalent form of TUA at the abstract level.

5 The convergence of $\Delta_+(\lambda_{d,u})$ and $\Delta_-(\lambda_{d,u})$

Through the above analysis, we provide the equivalent form of the TUA, which is based on 2-dimensional Gaussian distribution and a large dataset. By performing vectorization operations on $\Delta\lambda_{d,u_i}, u_i \in \mathcal{U}$, we introduce the definition of positive and negative effects, respectively, aiming to study the effect of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ on the data under the premise $\Delta(\lambda_{d,u_i}) \neq \Delta(\lambda_{d,u_{j,j\neq i}})$. Although we assume that the effects of $\sum \Delta_+(\lambda_{d,u_i})$ and $\sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$ are equal in a large dataset, we hope that $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ will have less and less impact on the data as q approaches q'. This concern is necessary because if the sample size is not large enough, the positive and negative effects may not cancel each other out. For example, the positive effects may be greater than the negative effects or vice versa. Quantifying $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ requires rigorous and rational mathematical expressions. Therefore, a natural question is: *how to describe the convergence of* $\Delta_+(\lambda_{d,u_i})$ *and* $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ *when* q *approaches* q'*?* We will give the answers to the above questions in Theorem 1.

5.1 The descriptive equation of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$

In classical physics, *damping* refers to the characteristic that the amplitude of vibration gradually decreases in any oscillating system, which may be caused by external influences or the system itself [45]. We introduce the above ideas into the study of the descriptive equation of $\Delta_+(\lambda_{d,u_i})$

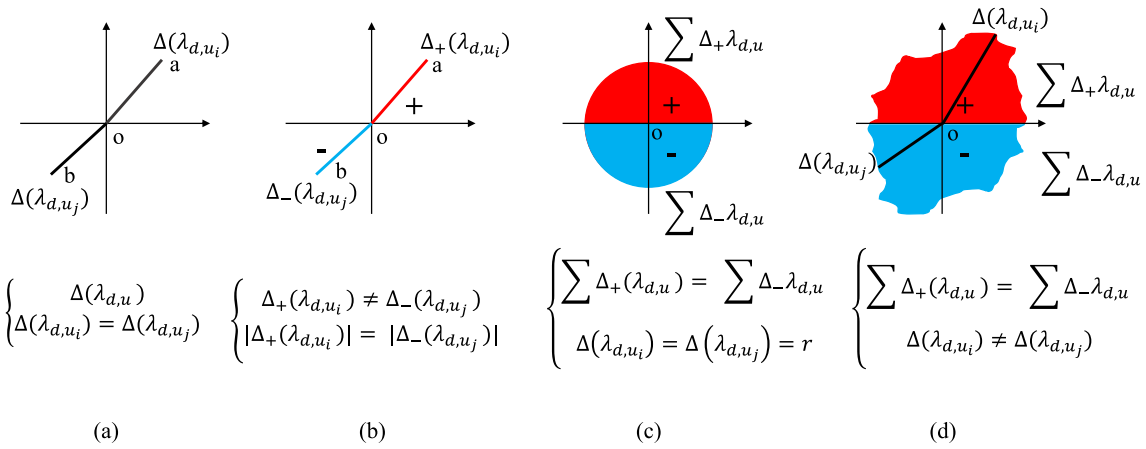

Fig. 4 Figures $(a) - (d)$ describe the equivalent representation of the TUA in the vector space by vectorizing λ_{d,u_i}. **(a)** is the geometric description of the traditional TUA assumption in the coordinate system. According to Lemma 1, $\epsilon(u_i) = \epsilon(u_{j,j\neq i})$ can be regarded as $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$. Hence, in the 2-dimensional plane, we can use Euclidean distance $L_{ao} = L_{bo}$ to describe $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$; **(b)** describes the vectorization of $\Delta(\lambda_{d,u_i})$. According to the definitions of *positive* (red), *negative* (blue) effects and the TUA assumption, we have $|\Delta_+(\lambda_{d,u_i})| = |\Delta_-(\lambda_{d,u_{j,j\neq i}})|$; **(c)** describes the vectorization of $\sum \Delta(\lambda_{d,u_i})$. It should be noted that the positive and negative effects of λ_{d,u_i} on the data are almost equal when the number of samples is large enough. Since $|\Delta_+(\lambda_{d,u_i})| = |\Delta_-(\lambda_{d,u_{j,j\neq i}})|$, all after vectorization of $\Delta(\lambda_{d,u_i})$ can form a circle in a 2-dimensional plane; **(d)** reflects the expansion of TUA assumption in the vector space. It can be regarded as a visualization of the TUA assumption at an abstract level (that is, constraints are applied to the dataset $\mathcal{U}$ rather than to each u_i). In other words, it is no longer necessary that $\Delta(\lambda_{d,u_i}) = \Delta(\lambda_{d,u_{j,j\neq i}})$

and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$. In this section, we provide a description equation about $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$, which satisfies that when q approaches q', $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ converge strictly to 0 (see Theorem 1).

Theorem 1 *For λ_{d,u_i}, if there are positive effect $\Delta_+(\lambda_{d,u_i})$ and negative effect $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ of $\Delta(\lambda_{d,u_i})$ on the data, $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ satisfy (or approximately satisfy) the following equation,*

$$\begin{aligned} S(\Delta_+(\lambda_{d,u_i}), q) &\triangleq A_+ \mathrm{e}^{-\eta_+ \cdot q} \cos(n \cdot \eta_+ \cdot q) \\ S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q) &\triangleq A_- \mathrm{e}^{-\eta_- \cdot q} \cos(n \cdot \eta_- \cdot q) \end{aligned}, \tag{26}$$

where $n \in \mathbb{Z}^+$, and $\eta_+ > 0$, $\eta_- > 0$ are adjustment parameters. $\mathrm{e}^{-\eta_+ \cdot q}$ and $\mathrm{e}^{-\eta_- \cdot q}$ are attenuation parameters. A_+ and A_- are the initial values of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$, respectively. Then $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ will gradually converge to 0 as q approaches q'.

Proof Let's analyze the first term of (26), i.e.,

$$\begin{aligned} S_1(\Delta_+(\lambda_{d,u_i}), q) &\triangleq A_+ \mathrm{e}^{-\eta_+ \cdot q} \\ S_2(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q) &\triangleq A_- \mathrm{e}^{-\eta_- \cdot q} \end{aligned}, \tag{27}$$

where A_+ and A_- are the initial values of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$, respectively. Because of $\eta_+ > 0$, $\eta_- > 0$, the two terms $\mathrm{e}^{-\eta_+ \cdot q}$ and $\mathrm{e}^{-\eta_- \cdot q}$ in the equation decay with the data size q.

Unfortunately, if the equation only uses (27) to describe the exponential decay trend of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$, it cannot reflect the potential impact of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ on the data. In other words, $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ do not necessarily follow a strictly monotonically decreasing function for convergence (see Fig. 5). Therefore, we need to consider the *volatility* effect of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ on the data.

Consider that the influence of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ on the data may be volatile. Therefore, we add the term "$\cos(n \cdot \eta_+ \cdot q)$" to (27) to describe the volatility effect of $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ on the data. We can rewrite (27) as follows:

$$\begin{aligned} S(\Delta_+(\lambda_{d,u_i}), q) &\triangleq A_+ \mathrm{e}^{-\eta_+ \cdot q} \cdot \cos(n \cdot \eta_+ \cdot q) \\ S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q) &\triangleq A_- \mathrm{e}^{-\eta_- \cdot q} \cdot \cos(n \cdot \eta_- \cdot q), \end{aligned} \tag{28}$$

where n and $\eta_+ > 0$, $\eta_- > 0$ are *adjustment parameters*, $\mathrm{e}^{-\eta_+ \cdot q}$ and $\mathrm{e}^{-\eta_- \cdot q}$ are *attenuation parameters*. Not only does the $\cos(n \cdot \eta_{+/-} \cdot q)$ function ensure that $A_+ \mathrm{e}^{-\eta_{+/-} \cdot q}$ decays exponentially, but also it ensures that (26) decays.

According to Fig. 5, we can intuitively understand the meaning of parameter A_+ and parameter η_+ in (27). The parameter A_+ determines the initial maximum value of the positive effect. The parameter η_+ determines the convergence speed of the function $S_1(\Delta_+(\lambda_{d,u_i}), q)$. Although $S_1(\Delta_+(\lambda_{d,u_i}), q)$ can describe that the positive effect converges to 0 quickly as the number of samples increases, it ignores the volatility of positive effects. The proof for $S_2(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ is similar.

Similarly, according to Fig. 6, we can intuitively understand the meaning of parameter A_+ and parameter η_+ in (26). The parameter A_+ determines the initial maximum value of the positive effect, the parameter η_+ determines the convergence speed of the function $S(\Delta_+(\lambda_{d,u_i}), q)$, and

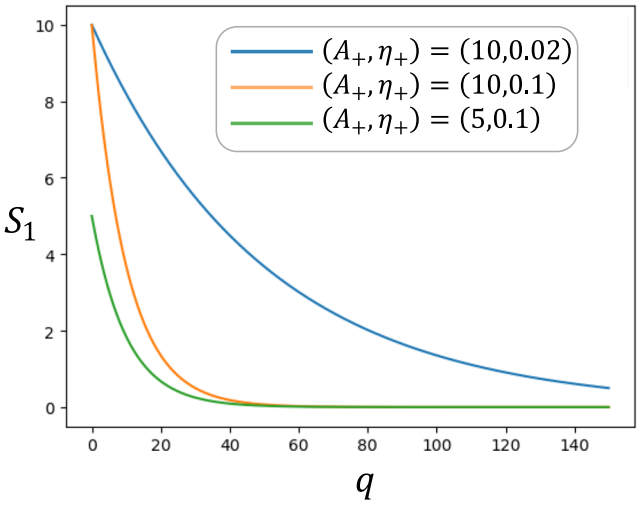

Fig. 5 A visualization of the influence of parameter (A_+, η_+) on equation $S_1(\Delta_+(\lambda_{d,u_i}), q)$. The situation of $S_2(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ is similar to the description of $S_1(\Delta_+(\lambda_{d,u_i}), q)$

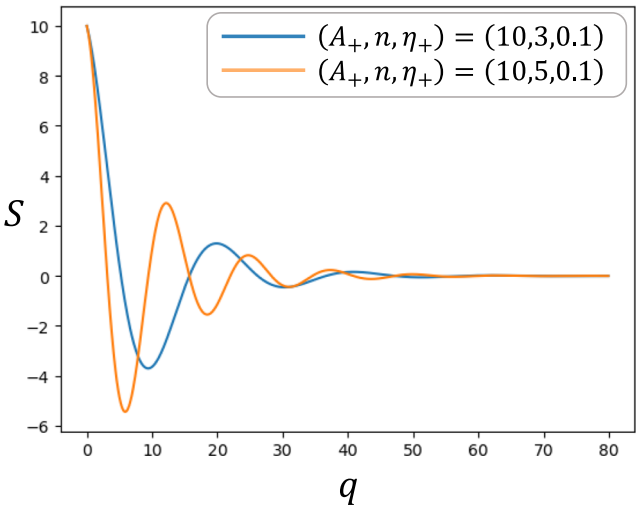

Fig. 6 A visualization of the influence of parameters (A_+, n, η_+) and $\cos(n \cdot \eta_+ \cdot q)$ on equation $S(\Delta_+(\lambda_{d,u_i}), q)$. The situation of $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ is similar to the description of $S(\Delta_+(\lambda_{d,u_i}), q)$

the $\cos(n \cdot \eta_+ \cdot q)$ reflects the volatility of the positive and negative effect. The purpose of introducing $\cos(n \cdot \eta_+ \cdot q)$ is to reflect the conversion between the positive effect and the negative effect as much as possible. Regarding the form of conversion, it can either be a positive effect that becomes a negative effect or vice versa. However, no matter how it is converted, it will eventually converge to 0 strictly under the $A_+ e^{-\eta_+ \cdot q}$. The proof for $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ is similar. □

5.2 The rationality analysis of equations $S(\Delta_+(\lambda_{d,u_i}), q)$ and $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$

The rationality analysis of equations $S(\Delta_+(\lambda_{d,u_i}), q)$ and $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ mainly includes two aspects:

- One is about the analysis of the visualization results of $S(\Delta_+(\lambda_{d,u_i}), q)$ and $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$.
- The other is the interpretability of $S(\Delta_+(\lambda_{d,u_i}), q)$ and $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$.

The function of $\cos(n \cdot \eta_{+/-} \cdot q)$ To simplify the presentation, we only analyze positive effects in this subsection. The analysis of negative effects is similar. As shown in Fig. 5, $S_1(\Delta_+(\lambda_{d,u_i}), q)$ only reflects the nature of exponential decay as q increases. Although $S_1(\Delta_+(\lambda_{d,u_i}), q)$ also can eventually converge to 0, $S_1(\Delta_+(\lambda_{d,u_i}), q)$ does not reflect its potential impact on data, because $S_1(\Delta_+(\lambda_{d,u_i}), q)$ directly describes the positive effect as a strict monotonic decreasing function. However, a representation based on strict monotonic decrement ignores the description of its internal complexities. The effect of $\Delta_+(\lambda_{d,u_i})$ on data may be volatile (the situation may also be more complex). Therefore, in order to describe the volatility of $\Delta_+(\lambda_{d,u_i})$, we introduce the $\cos(\cdot)$ function. Apparently, $S(\Delta_+(\lambda_{d,u_i}), q)$ presents a trend of exponential decay with volatility. Finally, as q increases, $S(\Delta_+(\lambda_{d,u_i}), q)$ will strictly converge to zero.

Attenuation parameters $e^{(-\eta_{+/-}) \cdot q}$ The purpose of introducing the attenuation parameter $e^{-\eta_{+/-} \cdot q}$ is to ensure that the positive effect and the negative effect can exhibit exponential decay characteristics as q increases. Although we improve TUA by vectorization, we hope that $S(\Delta_+(\lambda_{d,u_i}), q)$ and $S(\Delta_-(\lambda_{d,u_{j,j\neq i}}), q)$ will have minimal impact on the overall data. Therefore, even while acknowledging the existence of positive and negative effects, we hope that $\Delta_+(\lambda_{d,u_i})$ and $\Delta_-(\lambda_{d,u_{j,j\neq i}})$ can decay as quickly as possible in an exponential decay manner.

In fact, according to Lemma 1, Lemma 2, and Theorem 1, we provide a milder TUA assumption (referred to as M-TUA for short) through vectorization operations. In particular, (26) provides a formal description of positive effects and negative effects, which makes M-TUA interpretable. In summary, the above conclusion provides a mild form of TUA at the abstract level and an explicit (but not unique) mathematical description.

6 Comparison of TUA and M-TUA

In this section, we compare the traditional TUA and M-TUA to illustrate the similarities and differences between each other.

- $\epsilon(u_i)$ and $\Delta(\lambda_{d,u_i})$. TUA assumes that the value of ICE is the same for all $u_i \in \mathcal{U}$ ($|\mathcal{U}| = q$), e.g., $\epsilon(u_i) = \epsilon_{\text{ACE}}(u_i)$, where $i \in [1, ..., q]$. M-TUA transfers

the above problem to the constraint of $\Delta(\lambda_{d,u_i})$ by vectorization operation, that is,

$$\text{TUA} \Longrightarrow \begin{cases} \lim\limits_{q \to q'} \sum\limits_{q} (\Delta(\lambda_{d,u_i}) - \Delta(\lambda_{d,u_{j,j\neq i}})) = 0, \\ \lim\limits_{q \to q'} \left(\sum\limits_{q^+} \Delta_+(\lambda_{d,u_i}) - \sum\limits_{q^-} \Delta_-(\lambda_{d,u_{j,j\neq i}}) \right) = 0, \end{cases} \tag{29}$$

where $q^+ + q^- = q$.

- *Vector* $\Delta_{+/-}(\lambda_{d,u_{i/j}})$ *and Scalar* $\Delta(\lambda_{d,u_i})$. M-TUA provides a vector description of positive and negative effects for $\Delta(\lambda_{d,u_i})$ (i.e., $\Delta_{+/-}(\lambda_{d,u_{i/j}})$), aiming to distinguish M-TUA from traditional TUA. The vectorization operation allows for differences between individuals to exist, that is, $\Delta(\lambda_{d,u_i}) \neq \Delta(\lambda_{d,u_{j,j\neq i}})$ is allowed under the premise of $\sum \Delta_+(\lambda_{d,u}) = \sum \Delta_-(\lambda_{d,u_{j,j\neq i}})$. Therefore, M-TUA achieves the weakening of TUA.
- *Variance*. For a randomized experiment, the assumption of TUA implies that the variance is constant for all treatments. Constant variance is not a necessary condition for MTUA, MTUA should be used in data with small variance to constrain the dispersion of the population.

For the intuitiveness of description, we use a simple example to further illustrate how M-TUA weakens the TUA assumption.

Example 9 (Difference between data generated by TUA and M-TUA) TUA is different from M-TUA in a number of respects. A simple goal in this example is to compare the differences in the data under different assumptions via estimating the unobserved potential outcomes from Table 6.

- Similar to Example 8, in Table 7, we construct a set of data (including 10 subjects $u_i, i \in [1, 2, ..., 10]$) that meets the TUA assumption, where

$$\begin{aligned} \epsilon_{\text{ACE}}(u_i) &= \mathbb{E}(O_{t,u_i} - O_{c,u_i}) \\ &= \frac{1}{10} \sum_{i=1}^{10} (O_{t,u_i} - O_{c,u_i}) = 1. \end{aligned} \tag{30}$$

- Tables 8 and 9 are constructed based on M-TUA assumption.

As can be seen from Table 7, we know that the data only follows two situations, i.e., $O_{c,u_i} < O_{t,u_i}$ (i.e., $\epsilon_{\text{ACE}}(u_i) > 0$), or $O_{c,u_i} > O_{t,u_i}$ (i.e., $\epsilon_{\text{ACE}}(u_i) < 0$). *However, this strong assumption is often violated in the real world, which forces all subjects to have the same* $\epsilon(u_i)$. M-TUA alleviates this scenario and makes it more in line with the complex situations in real data (note that the values of $\Delta(\lambda_{d,u_i})$ in Tables 8 and 9 are not unique).

Table 6 Observation data with $\epsilon_{\text{ACE}}(u_i) = 1$

Subject	O_{t,u_i}	O_{c,u_i}	$O_{t,u_i} - O_{c,u_i}$
u_1	13	?	?
u_2	?	9.5	?
u_3	?	8	?
u_4	?	10	?
u_5	11	?	?
u_6	15	?	?
u_7	?	9.5	?
u_8	9	?	?
u_9	?	10	?
u_{10}	?	9	?
mean	?	?	1

As shown in Tables 8 and 9, it is not difficult to see that based on the assumption of M-TUA (i.e., $\sum_{i=1}^{10} \Delta(\lambda_{d,u_i}) = (0.2+0.2+0.2+0.2+0.1+0.1)-(0.2+0.2+0.3+0.3) = 0$), the data can be more in line with the assignment mechanism on the condition that the ACE value remains unchanged, thereby avoiding either $O_{c,u_i} < O_{t,u_i}$ (i.e., $\epsilon_{\text{ACE}}(u_i) > 0$), or $O_{c,u_i} > O_{t,u_i}$ (i.e., $\epsilon_{\text{ACE}}(u_i) < 0$). For example, according to (10), we have that

$$\epsilon(u_i) = \epsilon_{\text{ACE}}(u_i) + \Delta(\lambda_{d,u_i}), i \in [1, 2, ..., 10]. \tag{31}$$

Further, we obtain that,

$$\epsilon_{\text{ACE}}(u_i) = \frac{1}{10} \cdot \sum_{i=1}^{10} \epsilon(u_i) = \frac{1}{10} \cdot \sum_{i=1}^{10} \left(\epsilon_{\text{ACE}}(u_i) + \Delta(\lambda_{d,u_i}) \right). \tag{32}$$

Since $\epsilon_{\text{ACE}}(u_i) = \epsilon_{\text{ACE}}(u_i)$, in (32), only $\sum_{i=1}^{10} \Delta(\lambda_{d,u_i}) = 0$ needs to be satisfied. There are countless equations that satisfy $\sum_{i=1}^{10} \Delta(\lambda_{d,u_i}) = 0$.

Table 7 Assignment mechanism based on TUA assumption with $\epsilon_{\text{ACE}}(u_i) = 1$

Subject	O_{t,u_i}	O_{c,u_i}	$O_{t,u_i} - O_{c,u_i}$
u_1	13	12	1
u_2	11.5	10.5	1
u_3	10	9	1
u_4	12	11	1
u_5	11	10	1
u_6	15	14	1
u_7	13	12	1
u_8	9	8	1
u_9	8.5	7.5	1
u_{10}	12	11	1
mean	11.5	10.5	1

Table 8 Assignment mechanism based on M-TUA assumption with $\epsilon_{\text{ACE}}(u_i) = 1$

Subject	O_{t,u_i}	O_{c,u_i}	$\epsilon(u) = O_{t,u_i} - O_{c,u_i}$	$\Delta(\lambda_{d,u_i}) = \lambda_{t,u_i} - \lambda_{c,u_i}$
u_1	13	11.8	1.2	0.2
u_2	10.7	9.5	1.2	0.2
u_3	9.2	8	1.2	0.2
u_4	11.2	10	1.2	0.2
u_5	11	9.9	1.1	0.1
u_6	15	13.9	1.1	0.1
u_7	10.3	9.5	0.8	−0.2
u_8	9	8.2	0.8	−0.2
u_9	10.7	10	0.7	−0.3
u_{10}	9.7	9	0.7	−0.3
mean	10.98	9.98	1	0

Example 9 shows that the data constructed based on the M-TUA assumption allows for differences between various u_i's (e.g., $\epsilon(u_7, u_8, u_9, u_{10}) < 0$, $\epsilon(u_{1,2,3,4,5}) > 0$ and $\epsilon(u_6) = 0$), while ensuring that $\epsilon_{\text{ACE}}(u_i)$ is constant (e.g., $\epsilon_{\text{ACE}}(u_i) = 1$), which is more in line with the diversity of experimental samples in real tasks.

However, note that it is not sufficient to simply require that $\sum_{i=1}^{10} \Delta(\lambda_{d,u_i}) = 0$ holds, which does not guarantee that the data keeps good dispersion with this constraint. Therefore, an indispensable measure is to introduce variance as a metric to constrain the data so that the data constructed based on M-TUA maintains good dispersion. The reason is that the population is larger and the variance is less, the ACE would be closer to the true ACE regardless of the specific units randomly assigned to treatments. As mentioned above, for a randomized experiment, the TUA implies that the variance is constant for all treatments, which means that a necessary condition for TUA is that the variance is constant, while M-TUA only requires a small value of variance (e.g., the variance of $\Delta(\lambda_{d,u_i})$ in Table 8 is less than 0.5, and the variance of $\Delta(\lambda_{d,u_i})$ in Table 9 is close to 1).

Limitations Although M-TUA has realized the weakening of TUA to a certain extent and expanded the use scope of the original TUA, M-TUA itself is based on some assumptions and finally achieves the equivalence with TUA in the case of large samples, i.e., $q \to q'$. Therefore, M-TUA still has the following limitations.

– *Dimensionality limitation of vector space*. We take the 2-dimensional Gaussian distribution as an example. Based on Example 5, we analyze the equivalent form of TUA in 2-dimensional vector space. The vectorization operation in 2-dimensional space can easily be extended to 3-dimensional space. However, the equivalent form of the TUA for data in high-dimensional space has not been rigorously established.

Table 9 Assignment mechanism based on M-TUA assumption with $\epsilon_{\text{ACE}}(u_i) = 1$

Subject	O_{t,u_i}	O_{c,u_i}	$\epsilon(u) = O_{t,u_i} - O_{c,u_i}$	$\Delta(\lambda_{d,u_i}) = \lambda_{t,u_i} - \lambda_{c,u_i}$
u_1	13	10.98	2.02	1.02
u_2	11.53	9.5	2.03	1.03
u_3	10.04	8	2.04	1.04
u_4	12.04	10	2.04	1.04
u_5	11	9	2	1.00
u_6	15	15	0	−1.00
u_7	9.48	9.5	−0.02	−1.02
u_8	9	9.03	−0.03	−1.03
u_9	9.96	10	−0.04	−1.04
u_{10}	8.96	9	−0.04	−1.04
mean	11.001	10.001	1	0

– $\sum \Delta_{+/-}(\lambda_{d,u_{i/j\neq i}})$. As shown in Fig. 4d, M-TUA implies a premise that

$$\lim_{q\to q'}\left(\sum_{q^+}\Delta_+(\lambda_{d,u_i})-\sum_{q^-}\Delta_-(\lambda_{d,u_j})\right)=0, \tag{33}$$

where $q^+ + q^- = q$. It requires a large enough sample size to ensure that the equation

$$\sum_{q^+}\Delta_+(\lambda_{d,u_i})=\sum_{q^-}\Delta_-(\lambda_{d,u_j}) \tag{34}$$

holds with a high probability. Because the effects of any $\Delta(\lambda_{d,u_i})$ may be positive or negative (this is similar to the classical coin toss experiment, when the number of experiments is sufficient, the numbers of positive and negative coin occurrences are basically equal).

– *Decay rate*. The $e^{-\eta_{+/-}\cdot q}$ in Theorem 1 ensures that (26) will eventually converge to 0 with exponential decay. Of course, the purpose of choosing exponential decay is to make $\Delta_+(\lambda_{d,u_i})$ or $\Delta_-(\lambda_{d,u_-})$ converge quickly so that as the amount of sample data increases, the impact of $\Delta_+(\lambda_{d,u_i})$ or $\Delta_-(\lambda_{d,u_j})$ on the data will be minimal (or as small as possible) and eventually reach a negligible level.
– *Ignorability*. Since M-TUA is a constraint imposes on the task of making causal inferences in the POF, *ignorability*(i.e., $(O_{t,u_i}, O_{c,u_i}) \perp d$)) still needs to hold. In addition, we argue that estimating the variance of the data is still necessary (e.g, Example 9). Because if the population is larger and the variance is less, the ACE would be closer to the true ACE regardless of the specific units randomly assigned to treatment.

Interpretability Since the TUA cannot be tested and verified on the observed data, this will lead to limitations in the use of many models (e.g., the model of (7)) [36]. Therefore, it is necessary to obtain a milder and interpretable assumption. In general, M-TUA offers several advantages in terms of interpretability as follows:

– Based on the idea of DVE, we establish the relationship between TUA and RCE and try to provide some reasonable explanations for $\lambda_{d,u}$.
– Through vectorization operations, we endow $\lambda_{d,u}$ with the ability to describe positive and negative effects on data, and theoretically prove the rationality of M-TUA under the large dataset.
– M-TUA not only weakens the strength of the original TUA assumption but also provides a geometric description of the TUA.
– In particular, the M-TUA has an explicit mathematical expression that represents the meaning of the original TUA assumption at an abstract level through a set of interpretable parameters.

7 Conclusion and future work

In this paper, we first use an example to illustrate the underlying problems of using the functional model to estimate the probability solution of counterfactual queries. We analyze the inference mechanism of the functional model and point out that there are ambiguous conclusions when the unique output probability solution is 0 under the functional model. In other words, when the probability solution obtained by the functional model is 0, it does not mean that the estimated event will not occur. Secondly, for the TUA assumption commonly used in counterfactual models, we provide an equivalent description form of the TUA in the low-dimensional space. We weaken the TUA assumption by vectorizing the original TUA and finally obtain a milder TUA assumption, i.e., M-TUA. In addition, we also give theoretical proof and exhaustive analysis of the rationality and limitations of M-TUA.

As pointed out earlier, in M-TUA, the constraints on the unit are related to the dataset and RCE, instead of mandatory constraints for each unit. We argue this is very necessary, especially in the case of big data. Mild version assumption (not just M-TUA) can be viewed as an abstraction from the micro world to the macro world [46]. An intuitive example is that if we want to measure the water temperature of a swimming pool, it is impossible for us to measure every drop of water in the swimming pool. However, we do not think that the conclusion of this paper is the final form of the M-TUA. Therefore, we will focus on the following points in our future work.

Practicality Causal science has shown vigorous vitality in the field of AI and public health [47]. However, a large number of tasks can only be carried out under the premise of satisfying strong assumptions. The use of some assumptions is also not differentiated according to the different tasks. Therefore, including M-TUA, whether the version for different AI task scenarios can be further developed is a topic worthy of our further consideration.

Challenges posed by high-dimensional data . As a theoretical exploration of weakening TUA, M-TUA presents the equivalent form of TUA in vector space through vectorization and gives it a certain degree of interpretability. However, with the explosion of data, AI practitioners are confronted with data that are very large in both volume and dimensionality. Although our theorem shows that M-TUA is applicable in the case of big data, high-dimensional data

brings new challenges. Therefore, how to develop assumptions based on M-TUA with theoretical guarantees and applicable to high-dimensional data is also the focus of our future work.

Acknowledgements This work was supported by the National Key R&D Program of China under Grant 2018YFB1403200.

References

1. Heintzelman SJ, Christopher J, Trent J, King LA (2013) Counterfactual thinking about one's birth enhances well-being judgments. J Posit Psychol 8(1):44–49
2. Morgan SL, Winship C (2015) Counterfactuals and causal inference. Cambridge University Press
3. Balke A, Pearl J (1994) Probabilistic evaluation of counterfactual queries. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence, pp 230–237
4. Lewis D (1976) Probabilities of conditionals and conditional probabilities. In: Ifs. Springer, pp 129–147
5. Ginsberg ML (1986) Counterfactuals. Artif Intell 30(1):35–79
6. Kong E, Prinz D (2020) Disentangling policy effects using proxy data: Which shutdown policies affected unemployment during the covid-19 pandemic? J Public Econ 189:104257
7. Luo G, Zhao B, Du S (2019) Causal inference and bayesian network structure learning from nominal data. Appl Intell 49(1):253–264
8. Liu Y, Yu J, Xu L, Wang L, Yang J (2021) Sissos: intervention of tabular data and its applications. Appl Intell:1–15
9. Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic Books
10. Venzke I (2018) What if? counterfactual (hi) stories of international law. Asian J Int Law 8(2):403–431
11. Pesaran MH, Smith RP (2016) Counterfactual analysis in macroeconometrics: An empirical investigation into the effects of quantitative easing. Res Econ 70(2):262–280
12. Atan O, Zame WR, Feng Q, van der Schaar M (2019) Constructing effective personalized policies using counterfactual inference from biased data sets with many features. Mach Learn 108(6):945–970
13. Major D, Lenis D, Wimmer M, Sluiter G, Berg A, Bühler K (2020) Interpreting medical image classifiers by optimization based counterfactual impact analysis. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp 1096–1100
14. Castro DC, Walker I, Glocker B (2020) Causality matters in medical imaging. Nat Commun 11(1):1–10
15. Hao Z, Zhang H, Cai R, Wen W, Li Z (2015) Causal discovery on high dimensional data. Appl Intell 42(3):594–607
16. Qin L, Shwartz V, West P, Bhagavatula C, Hwang JD, Le Bras R, Bosselut A, Choi Y (2020) Backpropagation-based decoding for unsupervised counterfactual and abductive reasoning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 794–805
17. Nguyen T-L, Collins GS, Landais P, Le Manach Y (2020) Counterfactual clinical prediction models could help to infer individualised treatment effects in randomised controlled trials–an illustration with the international stroke trial. J Clin Epidemiol
18. Niu Y, Tang K, Zhang H, Lu Z, Hua X-S, Wen J-R (2021) Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12700–12710
19. Abbasnejad E, Teney D, Parvaneh A, Shi J, Hengel A (2020) Counterfactual vision and language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10044–10054
20. Bajaj M, Chu L, Xue ZY, Pei J, Wang L, Lam PC-H, Zhang Y (2021) Robust counterfactual explanations on graph neural networks. Adv Neural Inf Process Syst 34
21. Holzinger A, Malle B, Saranti A, Pfeifer B (2021) Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. Inf Fusion 71:28–37
22. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv JL Tech 31:841
23. Hendricks LA, Hu R, Darrell T, Akata Z (2018) Generating counterfactual explanations with natural language. arXiv:1806.09809
24. Ustun B, Spangher A, Liu Y (2019) Actionable recourse in linear classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp 10–19
25. Barocas S, Selbst AD, Raghavan M (2020) The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 80–89
26. Pearl J (2018) Theoretical impediments to machine learning with seven sparks from the causal revolution. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp 3–3
27. Marx A, Vreeken J (2019) Telling cause from effect by local and global regression. Knowl Inf Syst 60(3):1277–1305
28. Bertossi L (2021) Specifying and computing causes for query answers in databases via database repairs and repair-programs. Knowl Inf Syst 63(1):199–231
29. Hair Jr J F, Sarstedt M (2021) Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. J Mark Theory Pract:1–13
30. Zucker J, Paneri K, Mohammad-Taheri S, Bhargava S, Kolambkar P, Bakker C, Teuton J, Hoyt CT, Oxford K, Ness R et al (2021) Leveraging structured biological knowledge for counterfactual inference: A case study of viral pathogenesis. IEEE Trans Big Data 7(1):25–37
31. Truong D (2021) Using causal machine learning for predicting the risk of flight delays in air transportation. J Air Transport Manag 91:101993
32. Kumar V, Choudhary A, Cho E (2020) Data augmentation using pre-trained transformer models. arXiv:2003.02245
33. Wu X, Lv S, Zang L, Han J, Hu S (2019) Conditional bert contextual augmentation. In: International Conference on Computational Science. Springer, pp 84–95
34. Qin L, Bosselut A, Holtzman A, Bhagavatula C, Clark E, Choi Y (2019) Counterfactual story reasoning and generation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 5043–5053
35. Qian C, Feng F, Wen L, Ma C, Xie P (2021) Counterfactual inference for text classification debiasing. ACL-IJCNLP
36. Dawid AP (2000) Causal inference without counterfactuals. J Amer Stat Assoc 95(450):407–424
37. Holland PW (1986) Statistics and causal inference. J Amer Stat Assoc 81(396):945–960
38. Rubin DB (1980) Randomization analysis of experimental data: The fisher randomization test comment. J Amer Stat Assoc 75(371):591–593

39. Pearl J (2009) Causality. Cambridge university press
40. Pearl J, Glymour M, Jewell NP (2016) Causal inference in statistics: A primer. Wiley
41. Humar J (2012) Dynamics of structures. CRC press
42. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688
43. Imbens GW, Rubin DB (1997) Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Stat:305–327
44. Heckman JJ (2010) Building bridges between structural and program evaluation approaches to evaluating policy. J Econ Literature 48(2):356–98
45. Géradin M, Rixen DJ (2014) Mechanical vibrations: theory and application to structural dynamics. Wiley
46. Beckers S, Eberhardt F, Halpern JY (2020) Approximate causal abstractions. In: Uncertainty in Artificial Intelligence. PMLR, pp 606–615
47. Mohimont L, Chemchem A, Alin F, Krajecki M, Steffenel LA (2021) Convolutional neural networks and temporal cnns for covid-19 forecasting in france. Appl Intell:1–26