arXiv:2302.00860v3 [stat.ML] 9 Oct 2024

利用扩散模型对因果机制建模，以回答干预与反事实查询

Patrick Chao* PatrickBl¨obaum 宾夕法尼亚大学 亚马逊 pchao@wharton.upenn.edu bloebp@amazon.com
Sapan Patel ShivaPrasad Kasiviswanathan 亚马逊 亚马逊
sapanp@amazon.com kasivisw@gmail.com

摘要

我们研究在因果充分设定下回答观测、干预和反事实查询的问题，其中仅能获取观测数据和因果图。借助扩散模型领域的最新进展，我们提出基于扩散的因果模型（DCM）来学习因果机制，并生成唯一的潜在编码。这些编码使我们能够直接对干预进行采样，并通过溯因处理反事实。扩散模型非常适合这一任务，因为它们可以将每个节点编码为潜在表示，充当外生噪声的代理。我们的实证评估表明，在回答因果查询方面，该方法显著优于现有最先进的方法。此外，我们还给出了理论结果，为分析一般编码器-解码器模型中的反事实估计提供了一套方法，这套方法或可用于我们所提方法之外的其他场景。

摘要

我们研究在因果充分设定下如何回答观测、干预和反事实查询的问题，其中仅有观测数据和因果图可用。借助扩散模型的最新进展，我们提出基于扩散的因果模型（DCM）来学习因果机制，并生成唯一的潜在编码。这些编码使我们能够直接对干预情形进行采样，并通过溯因处理反事实。扩散模型非常适用于这一任务，因为它们可以将每个节点编码为潜在表示，作为外生噪声的代理。我们的实证评估表明，在回答因果查询方面，该方法显著优于现有最先进的方法。此外，我们还给出了理论结果，为分析一般编码器-解码器模型中的反事实估计提供了一套方法，这些结果也可能适用于我们提出的方法之外的场景。

1 引言

理解复杂问题中的因果关系，对于开展分析、得出结论和进行泛化预测至关重要。为此，我们需要因果模型和因果查询。结构因果模型（SCM）是描述变量间因果关系的生成模型，支持观测、干预和反事实查询（Pearl，2009a）。结构因果模型通过一组结构方程，规定如何根据一组具有先验分布的外生（未观测）随机变量生成一组内生（已观测）随机变量。

给定一个因果有向无环图，我们着重于使用扩散模型近似各个结构因果模型。作为应用，我们提出一个灵活的框架，用于回答所有三类因果查询（观测、干预和反事实）。对于反事实，我们采用 Pearl 的结构因果模型框架（Pearl，2009a），并尝试量化如下形式的个体层面陈述：已观测到一组 K 变量 $(X_1$、.....、$X_k)$ 的事实样本 $(x_1^{\mathrm{F}}$、.....、$x_K^{\mathrm{F}})$，那么如果某个集合 $X_{\mathcal{I}}$（其中包含 $I \subseteq [K]$）被设为某个 x_1^{F}，这些 K 变量的结果会是什么？

*在亚马逊实习期间完成的工作。

在本文中，我们始终假设因果充分性，即不存在隐藏混杂因素。请注意，因果充分性是仅凭观测数据回答因果查询的必要假设。

在结构因果模型框架中，可以通过学习未观测到的外生噪声及结构方程的代理来回答因果查询。这表明，将数据编码到潜在空间的（条件）生成模型可能是建模结构因果模型的一种理想选择，因为潜在表示可作为外生噪声的代理。在这些模型中，编码过程从观测值中提取潜在表示，而解码过程则从潜在表示生成样本，从而近似结构方程。

我们的贡献。在这项工作中，我们提出并分析了使用扩散模型建模结构因果模型的有效性。扩散模型（Sohl-Dickstein 等人，2015年；Ho 等人，2020年；Song 等人，2021年）近年来广受欢迎，这得益于其强大的表达能力以及在生成任务中的卓越表现（Saharia 等人，2022年；Ramesh 等人，2022年；Kong 等人，2021年）。我们的主要贡献在于确定如何将扩散模型应用于因果场景。我们的目标并非聚焦于某一种因果推断场景，而是通过这种建模提供一个灵活统一的框架，以回答广泛的因果查询。将扩散模型应用于因果场景并非易事，因为扩散模型通常用于学习标准高斯分布与数据分布之间的随机过程映射。我们的核心思路是将因果图中的每个节点建模为一个扩散模型，并按照拓扑序级联生成的样本，以回答因果查询。对于每个节点，相应的扩散模型以该节点及其父节点值作为输入，对潜在表示进行编码和解码。为了实现这一扩散模型，我们采用了近期提出的去噪扩散隐式模型（DDIM）（Song 等人，2021年）；该模型可被视为一种编码时不进行降维的确定性自编码器。我们利用 DDIM 的确定性正向和反向扩散过程，将扩散模型用作编码器-解码器模型。我们将所得模型称为基于扩散的因果模型（DCM），并表明该模型具备结构因果模型所需的特性。我们的主要贡献包括：

(1) [章节 3] 我们提出了基于扩散的因果模型（DCM），这是一类用于建模结构因果模型的新模型，为近似干预（do 算子）和反事实（溯因-行动-预测步骤）提供了灵活且实用的框架。我们提出了一种仅给定因果图和观测数据即可训练 DCM 的方法，并证明训练得到的模型能够从观测分布和干预分布中采样，并有助于回答反事实查询。

(2) [章节 4] 我们的理论分析考察了 DCM 生成的反事实估计的准确性，并证明在一些合理假设下，这些估计的误差可以得到界定。重要的是，我们的分析并不局限于扩散模型，也适用于其他编码器-解码器设置。据我们所知，这是首批误差界，能够解释使用编码器-解码器模型（如扩散模型）处理反事实查询时所观察到的性能提升。该结果的另一特点是，在增加一项假设后，也可推广至更具挑战性的多变量情形。

(3) [节 5] 我们在一系列合成数据集上评估了 DCM 在三种因果查询中的表现。这些数据集采用不同的结构方程类型生成。结果表明，DCM 的表现始终优于现有最先进方法（S´anchez-Martin 等人，2022年；Khemakhem 等，2021）。事实上，对于某些干预查询和反事实查询（例如涉及非加性噪声 模型的查询），DCM 的表现比这些现有方法高出一个数量级甚至更多。此外，我们还展示了 DCM 在一项基于功能磁共振成像数据开展的干预查询实验中的出色表现。

相关工作。 多年来，因果推断领域开发了多种方法，用于回答干预查询和／或反事实查询，包括非参数方法（Shalit 等人，2017；Alaa 与 Van Der Schaar，2017；Muandet 等人，2021年）以及概率建模方法（Ze˘cevi´c 等人，2021a）。与我们的方法更相关的是近期一系列研究，包括（Moraffah 等人，2020年；Pawlowski 等人，2020年；Kocaoglu 等人，2018年；Parafita 与 Vitri`a，2020年；Ze˘cevi´c 等人，2021b；Garrido 等人，2021年；Karimi 等人，2020年；S´anchez-Martin 等人，2022年；Khemakhem 等，2021；Sanchez 与 Tsaftaris，2022），这些研究展现了使用深度（条件）生成模型完成此任务的成效。

Karimi 等人 (2020年) 提出了一种回答干预查询的方法，即针对因果图所蕴含的马尔可夫分解中的每个条件分布拟合一个条件变分自编码器 。Pawlowski 等人 (2020年) 也运用变分推断和归一化流的思想，提出了反事实推断方案。

在 Khemakhem 等人（2021年）的研究中，作者提出了一种用于因果发现和因果查询的自回归归一化流，称为 *CAREFL*。即使像我们所要求的那样，仅知道因果排序而不了解完整的因果图，CAREFL 也适用。然而，正如 S´ánchez-Martin 等人（2022年）也指出的，当因果图

存在时，CAREFL 无法充分利用边的缺失，因为它会将因果图简化为因果排序（该排序可能并不唯一）。(Balgi 等人，2022年) 也探索了使用归一化流基于因果有向无环图回答因果查询的方法，但他们的方法对其反事实估计没有任何理论保证，而且实验评估相当有限。

S´anchez-Martin 等人（2022）提出了 *VACA*，该方法以变分图自编码器 的形式使用图神经网络（GNNs），从观测分布、干预分布和反事实分布中采样。VACA 可以通过 GNN 利用图的固有结构，但实证表现不佳（见第5节）。此外，使用 GNN 会导致不理想的设计限制，例如，编码器图神经网络不能包含隐藏层（S´anchez-Martin 等人，2022）。

Javaloy 等人（2023）最近的一项工作结合了使用自回归归一化流（如 CAREFL 中所用）和将整个因果图建模为一个模型（如 VACA 中所用）的思路，以减少图中可能出现的误差传播。作者进一步将 Khemakhem 等人（2021）的理论结果推广到仿射自回归归一化流之外，并通过提供一种更直接地应用 do 算子的方法，建立了与结构因果模型更清晰的联系。与我们的工作不同，Javaloy 等人（2023）专注于将整个因果图建模为一个模型。与单一模型相比，按节点分别建模在灵活性和计算效率方面具有多项优势，因为这种方式允许并行训练各个节点模型。例如，在一次实验中，我们观察到，我们的 DCM方法训练速度约为 CAREFL 的七倍、VACA 的九倍（见附录D.1）。

Sanchez 与 Tsaftaris（2022年）使用扩散模型进行反事实估计，重点研究图像类别导致图像的双变量图情形。作者训练扩散模型生成图像，并使用 Pearl 等人（2016年）提出的溯因-行动-预测程序，以及分类器引导（Dhariwal 与 Nichol，2021年）来生成反事实图像。然而，该方法仅适用于双变量模型，并且需要为中间扩散图像训练单独的分类器；对于更复杂的图像（例如 ImageNet（Deng 等人，2009年）），其性能也较差。我们的方法区别于 Sanchez 与 Tsaftaris（2022年）之处在于，它能够处理一般因果图（超越更简单的双节点设定），并且适用于连续变量（超越离散标签与图像这一更简单的情形）。我们的实验评估更加全面，还涵盖了 Sanchez 与 Tsaftaris（2022年）未涉及的干预查询。最后，在理论贡献方面，我们为潜变量模型和结构方程提供了严格条件，在这些条件下可以估计反事实。即使对于 Sanchez 与 Tsaftaris（2022年）所研究的双变量设定，此类理论理解此前也尚不存在。

最后，近期也有研究提出一种基于扩散模型的方法，用于加性噪声模型下的因果发现这一不同任务，其中扩散模型作为海森函数的近似（Sanchez 等人，2022年）。这引出了一个有趣的问题：能否调整扩散模型，以解决从因果发现到因果推断的端到端问题。

2 预备知识

符号约定。 为区分随机变量及其具体取值，前者用大写字母表示，后者用相应的小写字母表示。为区分因果图中的节点和扩散随机变量，我们用下标表示图节点。令 $[n] := \{1, \dots, n\}$。

结构因果模型。考虑一个有向无环图（DAG） $\mathcal{G}$，其节点按拓扑排序排列为 $\{1, \dots, K\}$，其中节点 i 由某个通用空间 $\mathcal{X}_i \subset \mathbb{R}^{d_i}$中的一个（随机）变量 X_i 表示。令 pa$_i$ 表示图 $\mathcal{G}$ 中节点 i 的父节点，并令 $X_{\mathrm{pa}_i} := \{X_j\}_{j \in \mathrm{pa}_i}$ 表示节点 i的父节点对应的变量。结构因果模型 $\mathcal{M}$ 描述观测/内生节点 i 与其因果父节点之间的关系。形式上，SCM $\mathcal{M} :=$ (F p(U))通过一组结构方程 F $:= (f_1, \dots, f_K)$ 确定如何由一组外生随机变量 U $:= \{U_1, \dots, U_K\}$ 及其先验分布 p(U)，生成一组 K 内生随机变量 X $:= \{X_1, \dots, X_K\}$；其中，对于 $i \in [K]$，有 $X_i := f_i(X_{\mathrm{pa}_i}, U_i)$。本文始终假设未观测随机变量联合独立（马尔可夫结构因果模型），且有向无环图 $\mathcal{G}$ 是由 $\mathcal{M}$诱导的图。每个 SCM $\mathcal{M}$ 都蕴含一个唯一的联合观测分布，并满足因果马尔可夫假设：p(X) $= \prod_{i=1}^{K} p(X_i \mid X_{\mathrm{pa}_i})$。

结构因果模型处理 Pearl 的因果层级（或“因果阶梯”），该层级由复杂度逐渐增加的三“层”因果查询组成（Pearl，2009a）：观察性（或关联性）、干预性和反事实。例如，干预查询可以表述为：“如果将变量 X_i 设为固定值 γ_i，对总体 **X** 会产生什么影响？”do 算子 do($X_i := \gamma_i$) 表示设定

变量 X_i 为 γ_i 的影响。请注意，我们提出的框架也支持更一般的干预集合，例如对多个变量进行干预，记作 $\text{do}(X_{\mathcal{I}} := \gamma)$（其中 $\mathcal{I} \subseteq [K]$, $X_{\mathcal{I}} := (X_i)_{i\in\mathcal{I}}$, $\gamma \in \mathbb{R}^{|\mathcal{I}|}$）。干预操作 $\text{do}(X_{\mathcal{I}} := \gamma)$ 会将原始联合分布转换为干预分布，记作 $p(\mathbf{X} \mid \text{do}(X_{\mathcal{I}} := \gamma))$。另一方面，反事实查询可以表述为："如果将 $X_{\mathcal{I}}$ 的值设为 γ，某个特定事实样本 $x^{\text{F}} := (x_1^{\text{F}}, \ldots, x_K^{\text{F}})$ 会得到什么结果？"反事实估计可以通过以下三步程序进行：1）溯因：估计外生噪声 U；2）行动：进行干预 $\text{do}(X_{\mathcal{I}} := \gamma)$；3）预测：使用溯因得到的噪声和干预值估计 x^{CF}。请注意，在步骤 1 中，我们进行确定性反事实推理，重点关注总体中单个个体的反事实。

扩散模型。给定来自分布 $X^0 \sim Q$ 的数据，扩散模型的目标是构造一个可高效采样、近似于 Q 的分布。去噪扩散概率模型（DDPM）（Sohl-Dickstein 等人，2015年；Ho 等人，2020年）通过引入前向加噪过程（在每个时间步添加各向同性高斯噪声）以及经过学习的反向去噪过程来实现这一目标。扩散模型的一种常见表示形式是一个固定的马尔可夫链，以方差 $\beta_1, \ldots, \beta_T \in (0, 1)$ 添加高斯噪声，生成潜变量 $X^1, \ldots, X^T$，

$$q(X^t \mid x^{t-1}) = \mathcal{N}(X^t; \sqrt{1-\beta_t}x^{t-1}, \beta_t I)$$ 和 $$q(X^t \mid x^0) = \mathcal{N}(X^t; \sqrt{\alpha_t}x^0, (1-\alpha_t)I)$$，其中：$\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$。这里，$T \in \mathbb{Z}^+$，而 $t \in \{0, \ldots, T\}$ 表示时间索引。

通过选择足够大的、收敛于 0 的 T 和 α_t，我们有 X^T 服从各向同性高斯分布。学习到的反向扩散过程尝试使用神经网络近似难以处理的 $q(X^{t-1} \mid x^t)$，并被定义为具有高斯转移的马尔可夫链，$p_\theta(X^{t-1} \mid x^t) = \mathcal{N}(X^{t-1}; \mu_\theta(x^t, t), \Sigma_\theta(x^t, t))$。网络不直接预测 μ_θ，而是可以预测来自 $x^t = \sqrt{\alpha_t}x^0 + \sqrt{1-\alpha_t}\varepsilon$的高斯噪声 ε。Ho 等人（2020年）发现，对 ε 而非 μ_θ进行建模、固定 Σ_θ，并使用以下重加权损失函数

$$\mathbb{E}_{\substack{t\sim\text{Unif}\{[T]\} \\ X^0\sim Q \\ \varepsilon\sim\mathcal{N}(0,I)}}[\|\varepsilon - \varepsilon_\theta(\sqrt{\alpha_t}X^0 + \sqrt{1-\alpha_t}\varepsilon, t)\|^2], \tag{1}$$

在经验上效果良好。我们在训练中也采用该损失函数。

Song等人（2021）证明，可以将预训练的标准去噪扩散概率模型（DDPM）推广到非马尔可夫过程。具体而言，可以使用预训练的 DDPM 模型，根据噪声 X^T获得确定性样本，这被称为去噪扩散隐式模型（DDIM），其具有反向隐式扩散过程

$$X^{t-1} := \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}X^t - \varepsilon_\theta(X^t, t)\left(\sqrt{\alpha_{t-1}(1-\alpha_t)/\alpha_t} - \sqrt{1-\alpha_{t-1}}\right). \tag{2}$$

注意，X^t 这里是确定性的。我们还使用 Song等人（2021）提出的前向隐式扩散过程，该过程通过将 DDIM 过程公式 2 改写为常微分方程（ODE），并考虑正向的欧拉方法近似推导得出

$$X^{t+1} := \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}X^t + \varepsilon_\theta(X^t, t)\left(\sqrt{1-\alpha_{t+1}} - \sqrt{\alpha_{t+1}(1-\alpha_t)/\alpha_t}\right). \tag{3}$$

我们在本文中采用 DDIM框架，具体而言，公式 3 和 2 分别定义编码（正向）和解码（反向）过程。请注意，这确保了编码与解码的确定性。这种构造为每个观测生成唯一的潜变量以及唯一的解码结果，并确保对于重复的反事实查询，我们获得相同的输出。

3 扩散因果模型：基于扩散的因果模型

在本节中，我们介绍用于对结构因果模型进行建模并回答因果查询的 DCM 方法。DCM 方法属于一类通用技术，这类技术尝试使用编码器-解码器对来建模结构因果模型。考虑一个数据生成过程 $X = f(X_{pa}, U)$。目标是构建一个编码函数 g和一个解码函数 h。编码函数 g尝试表示 U中的信息：对于一对 (X, X_{pa})，$Z := g(X, X_{pa})$ 是潜变量。解码器将输入 Z 和 X_{pa} 作为输入，尝试重构 X：

$\hat{}X = h(Z, X_{pa})$，其中在完美重构的情况下，$\hat{}X = X$。解码函数 h 模拟真实的结构方程 f，但不必与之完全相等。例如，对于可逆函数 r，存在无穷多个编码，使得对所有 U，均满足$Z = r(U)$。

我们首先介绍 DCM 的构建和训练过程，然后说明如何利用该模型回答各种因果查询。我们从一些符号约定开始。

- 定义 Z_i^t 为正向隐式扩散过程（公式3）中扩散步骤 t 的第 i个内生节点值，并令 $Z_i := Z_i^T$。
- 定义 $\hat{} X_i^t$ 为反向隐式扩散过程（公式 2）中扩散步骤 t 的第 i个内生节点值，并令 $\hat{} X_i := \hat{}X_i^0$。

训练 DCM。 我们为每个节点训练一个扩散模型，并将去噪后的父节点值作为输入。父节点值可以被视为模型的额外协变量，此时可以选择使用无分类器引导来纳入这些协变量（Ho & Salimans，2021年）。从经验上看，我们发现，直接拼接协变量比使用无分类器引导能够取得更好的性能。

我们采用 Ho 等人（2020年）提出的扩散模型 ε_θ 参数化，将节点 i 的扩散模型表示为 $\varepsilon_\theta^i(X, X_{\mathrm{pa}_i}, t)$。算法1中给出的完整训练过程仅在通常训练过程的基础上进行了少量修改，增加了将父节点作为协变量，以及为每个节点训练一个扩散模型。由于为不同内生节点生成数据所学习的生成模型不会相互影响训练，因此这些模型可以并行训练。对于图中的每个节点，我们都可以并行训练一个模型，因为每个扩散模型只需要当前节点值和父节点值。最终的 DCM模型就是这些 K 训练好的扩散模型 ε_θ^1、.....、ε_θ^K的组合。

算法1DCM 训练

输入： 分布 Q，尺度因子 $\{\alpha_t\}_{t=1}^T$，因果有向无环图 $\mathcal{G}$，其中节点 i 表示为 X_i

1: **当**未收敛**时**2: 采样 $X^0 \sim Q$3: **对于** $i = 1, \ldots, K$，**执行**4: $t \sim \mathrm{Unif}[\{1, \ldots, T\}]$5: $\varepsilon \sim \mathcal{N}(0, I_{d_i})\{d_i$是 $X_i\}$6: 更新节点 i 的扩散模型参数 ε_θ^i，方法是最小化以下损失：$\|\varepsilon - \varepsilon_\theta^i(\sqrt{\alpha_t} X_i^0 + \sqrt{1-\alpha_t}\varepsilon, X_{\mathrm{pa}_i}^0, t)\|_2^2$ (基于公式 1)7: **结束 for**8: **结束 while**

训练过程中，我们使用所有变量（$X_1, \ldots, X_K$），因为先验上我们不对可能的因果查询作任何假设，即允许所有可能的目标变量、干预变量等。但是，如果我们只关注某个预定义的查询集合，那么可以相应地简化该图。例如，如果我们只关注某个节点相对于对其前驱节点进行干预的反事实估计，则可以将其简化为一个仅包含目标节点、干预节点和后门调整集（例如，干预节点的祖先节点）的子图。这样便只需学习一个扩散模型。

我们提出的 DCM方法的一项主要优势是能够推广到更大的图。由于每个扩散模型仅将父节点作为输入，对每个节点进行建模只取决于该节点的入度（即因果父节点的数量）。虽然扩散模型的数量会随非根节点的数量而增加，但从参数规模来看，每个模型通常都较小，并且可以并行训练。此外，我们可以将所提出的 DCM方法应用于任何适用扩散模型的场景：连续变量、高维场景、类别数据、图像等。

DCM 编码与解码步骤。 对于 DCM，编码（分别为解码）过程与 DDIM 编码（分别为解码）过程相同，只是将父节点值作为额外协变量纳入其中。请注意，给定模型 ϵ_θ，DDIM 是一个确定性过程，如公式 2 和 3 所述。我们聚焦于一个节点 $i \in [K]$（对每个节点 i重复相同的过程）。编码过程将 X_i 及其父节点值 X_{pa_i} 作为输入，并将其映射为一个潜变量 Z_i。解码过程将 Z_i 和 X_{pa_i} 作为输入，以构建 X_i（ X_i的近似值）。形式上，使用公式 3 中的前向隐式扩散过程，给定一个样本 X_i，我们编码得到一个唯一的潜变量

$Z_i := Z_i^T$，使用递归公式

$$Z_i^{t+1} := \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} Z_i^t + \varepsilon_\theta^i(Z_i^t, X_{\mathrm{pa}_i}, t) \left(\sqrt{1-\alpha_{t+1}} - \sqrt{\frac{\alpha_{t+1}(1-\alpha_t)}{\alpha_t}} \right), \forall t = 0, .., T-1, \tag{4}$$

其中 $Z_i^0 := X_i$。潜变量 Z_i 充当外生噪声 U_i的代理。使用公式 2 中 DDIM 的反向隐式扩散过程，给定一个潜向量 Z_i，我们得到一个确定性解码 $\hat{} X_i := \hat{} X_i^0$，其递归公式为

$$\hat{X}_i^{t-1} := \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \hat{X}_i^t - \varepsilon_\theta^i(\hat{X}_i^t, X_{\mathrm{pa}_i}, t) \left(\sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}} - \sqrt{1-\alpha_{t-1}} \right), \text{ for all } t = T, \ldots, 1, \tag{5}$$

其中 $\hat{} X^T := Z_i$。下文中，我们用 $\mathsf{Enc}_i(X_i, X_{\mathrm{pa}_i})$ 和 $\mathsf{Dec}_i(Z_i, X_{\mathrm{pa}_i})$ 分别表示式 4 和式 5 中定义的节点 i 的编码函数和解码函数。详细伪代码见算法 2 和算法 3。

算法 2 $\mathsf{Enc}_i(X_i, X_{\mathrm{pa}_i})$

Input: X_i, X_{pa_i}

$Z_i^0 \leftarrow X_i$
for $t = 0, \ldots, T-1$ **do**
  $Z_i^{t+1} \leftarrow \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} Z_i^t + \varepsilon_\theta^i(Z_i^t, X_{\mathrm{pa}_i}, t) \left(\sqrt{1-\alpha_{t+1}} - \sqrt{\frac{\alpha_{t+1}(1-\alpha_t)}{\alpha_t}} \right)$
end for
Return $Z_i := Z_i^T$

算法 3 $\mathsf{Dec}_i(Z_i, X_{\mathrm{pa}_i})$

Input: Z_i, X_{pa_i}

$\hat{X}^T \leftarrow Z_i$
for $t = T, \ldots, 1$ **do**
  $\hat{X}_i^{t-1} \leftarrow \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \hat{X}_i^t - \varepsilon_\theta^i(\hat{X}_i^t, X_{\mathrm{pa}_i}, t) \left(\sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}} - \sqrt{1-\alpha_{t-1}} \right)$
end for
Return $\hat{X}_i := \hat{X}_i^0$

使用训练好的 DCM 回答因果查询。现在我们介绍如何使用训练好的 DCM模型来近似回答因果查询。回答观测查询和干预查询分别需要从观测分布和干预分布中采样。对于反事实查询，查询是在个体层面进行的，此时结构赋值会发生变化，但外生噪声与观测数据中的相同。

(a) 为观测查询/干预查询生成样本。可以按如下方式生成来自 DCM模型的样本，以近似干预分布 $p(\mathbf{X} \mid \mathrm{do}(X_\mathcal{I} := \gamma))$。对于干预节点 $i \in \mathcal{I}$，其干预值为 γ_i，采样值始终是干预值，因此我们生成 $X_i := \gamma_i$。对于非干预节点 i，假设根据归纳法，我们已得到生成的父节点值 X_{pa_i}。要生成 X_i，我们首先采样潜向量 $Z_i \sim \mathcal{N}(0, I_{d_i})$，其中 d_i 是 X_i的维度。然后将 Z_i 作为节点 i的噪声，计算 $\hat{} X_i := \mathsf{Dec}_i(Z_i, \hat{} X_{\mathrm{pa}_i})$，作为节点 i生成的样本值。随后将该值 $\hat{} X_i$ 用作节点 i的子节点的父节点值。通过设置 $\mathcal{I} = \emptyset$，可以生成来自 DCM模型的样本，以近似观测分布 $p(\mathbf{X})$。伪代码见算法 4。

（b）反事实查询。考虑事实观测 $x^\mathrm{F} := (x_1^\mathrm{F}, \ldots\ldots, x_K^\mathrm{F})$，以及对一组节点 $\mathcal{I}$ 施加值为 γ的干预。我们使用 DCM模型构造反事实估计 $\hat{} x^\mathrm{CF}$，方法如下。反事实估计仅在干预节点或干预节点的后代节点上与事实值不同。与干预查询类似，对于每个干预节点 $i \in \mathcal{I}$，$\hat{} x_i^\mathrm{CF} := \gamma_i$。对于每个非干预节点 i，若其为任一干预节点的后代节点，则根据归纳假设，我们已生成反事实估计 $\hat{} x_{\mathrm{pa}_i}^\mathrm{CF}$。要获得 $\hat{} x_i^\mathrm{CF}$，

算法 4观测/干预采样

输入： 干预集 $\mathcal{I}$ 及其对应的值 γ （$I = \emptyset$ 用于观测采样）

对于 $i = 1, \ldots, K$ **执行** {按拓扑序排列}
  $Z_i \sim \mathcal{N}(0, \ I_{d_i})$
  如果 $i \in \mathcal{I}$ **那么**
    $\hat{X}_i \leftarrow \gamma_i$
  else
    $\hat{X}_i \leftarrow \text{Dec}_i(Z_i, \ \hat{X}_{\text{pa}_i})$
  结束 if
结束 for
返回 $\hat{X} := (\hat{X}_1, \ldots, \hat{X}_K)$

算法 5反事实估计

输入： 干预集 $\mathcal{I}$，取值为 γ，事实样本 $x^{\text{F}} := (x_1^{\text{F}}, \ \ldots\ldots, \ x_K^{\text{F}})$

对于 $i = 1, \ldots, K$ **执行** {按拓扑序}
  如果 $i \in \mathcal{I}$ **则**
    $\hat{x}_i^{\text{CF}} \leftarrow \gamma_i$
  否则，如果 i 不是 $\mathcal{I}$ 中任何干预节点的后代节点**则**
    $\hat{x}_i^{\text{CF}} \leftarrow x_i^{\text{F}}$
  else
    $z_i^{\text{F}} \leftarrow \text{Enc}_i(x_i^{\text{F}}, \ x_{\text{pa}_i}^{\text{F}})$ {溯因步骤}
    $\hat{x}_i^{\text{CF}} \leftarrow \text{Dec}_i(z_i^{\text{F}}, \hat{x}_{\text{pa}_i})$ {行动和预测步骤}
  结束 if
结束 for
返回 $\hat{x}^{\text{CF}} := (\hat{x}_1^{\text{CF}}, \ \ldots\ldots, \ \hat{x}_K^{\text{CF}})$

我们首先将估计的事实噪声定义为 $\hat{z}_i^{\text{F}} := \text{Enc}_i(x_i^{\text{F}}, \ x_{\text{pa}_i}^{\text{F}})$。然后，我们将 $\hat{z}_i^{\text{F}}$用作节点 i的噪声，通过解码生成反事实估计：$\hat{x}_i^{\text{CF}} := \text{Dec}_i(\hat{z}_i^{\text{F}}, \ \hat{x}_{\text{pa}_i}^{\text{CF}})$。伪代码见算法 5。

请注意，使用 x^{F}时，我们假设完全可观测性，[1] 因为算法 5 会为每个节点生成反事实估计。然而，当对 $X_{\mathcal{I}}$ 进行干预，且唯一感兴趣的量是某个 $X^\star$上的反事实时，只需要来自 $\{X_i$ 的事实样本：X_i 位于从 $X_{\mathcal{I}} \to X^\star\}$ 出发的一条路径上（Saha & Garain, 2022）。在实践中，还可以通过填补缺失数据进一步放宽这一要求，但这超出了本文的研究范围。

4 反事实误差界定

现在，我们将确立足以界定反事实估计误差的条件。事实上，本节中的结果不仅适用于扩散模型，也适用于满足某些性质的更一般的条件潜变量模型。此外，在额外假设下，这一结果还可扩展到更具挑战性的高维情形。本节的所有证明汇总于附录 A。

我们重点学习给定其父节点 X_{pa_i} 时内生节点 X_i 的单个条件潜变量模型，因为针对不同内生节点学习的模型不会相互影响。由于节点 i 的选择无关紧要，下面省略下标 i，并将感兴趣的节点称为 X，其因果父节点称为 X_{pa}，对应的外生变量称为 U，结构方程称为 $X := f(X_{\text{pa}}, \ U)$。设编码函数为 $g\colon \mathcal{X} \times \mathcal{X}_{\text{pa}} \to \mathcal{Z}$，解码函数为 $h\colon \mathcal{Z} \times \mathcal{X}_{\text{pa}} \to \mathcal{X}$，其中 $\mathcal{Z}$ 是潜在空间。在 DCM 场景中，函数 g 和 h 分别对应 Enc 和 Dec 函数。

众所周知，即使满足因果充分性，若不对函数关系作出假设，某些反事实查询也无法从观测数据中识别（Pearl, 2009b）。因此，近期研究致力于探究在什么条件下能够获得可识别性结果（Lu等,

[1] 这是文献中的一个常见假设，所有相关工作也都采用了这一假设，例如（S´anchez-Martin 等人，2022年；Khemakhem 等，2021；Pawlowski 等人，2020年）。

2020年；Nasr-Esfahany & Kiciman, 2023；Nasr-Esfahany等人，2023）。我们定理1中的假设2确保真实的反事实结果具有可识别性，参见例如（Lu等，2020年，定理1）或（Nasr-Esfahany & Kiciman, 2023，定理5）。在学习得到的结构因果模型这一背景下，要判断给定的反事实查询能否以足够的准确度得到回答，还需要对学习得到的结构因果模型作出假设，例如本例中的编码器和解码器。

我们的第一个结果给出了关于潜变量编码函数和结构方程的充分条件，在这些条件下，我们可以将潜在外生变量恢复到一个（可能是非线性的）可逆函数的程度。我们从一维外生噪声 U 和变量 $X \in \mathcal{X} \subset \mathbb{R}$ 开始。在第4.1节中，我们为更高维的情形给出了类似的定理；在定理 2 中，$X \in \mathbb{R}^m$ 对于 $m \geq 3$，并对 f和 g的雅可比矩阵作出了更强的假设。

定理1. 假设对于 $X \in \mathcal{X} \subset \mathbb{R}$ 和外生噪声 $U \sim$ 均匀分布$[0,1]$，X 满足如下结构方程：$X := f(X_{\mathrm{pa}}, U)$，其中 $X_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$是节点 X的父节点，且 $U \perp\!\!\!\perp X_{\mathrm{pa}}$。考虑一个编码器-解码器模型，其编码函数为 $g: \mathcal{X} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{Z}$，解码函数为 $h: \mathcal{Z} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{X}$，$Z := g(X, X_{\mathrm{pa}})$，$\hat{X} := h(Z, X_{\mathrm{pa}})$。假设满足以下条件：

1. 编码独立于父节点值，$g(X, X_{\mathrm{pa}}) \perp\!\!\!\perp X_{\mathrm{pa}}$.

2. 结构方程 f 对 U 可微且严格递增，对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{p}}$
a。*3.* 编码 g 对 X 可逆且可微，对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$。

那么，对于一个可逆函数 $\tilde{q}$，有 $g(X, X_{\mathrm{pa}}) = \tilde{q}(U)$。

定理1所依据假设的讨论。

(1) 关于编码与父节点值相互独立的假设1看似严格，但实际上往往成立。例如，在加性噪声设定中，若 $f(X_{\mathrm{pa}}, U) := f'(X_{\mathrm{pa}}) + U$，其中 X_{pa}和 U相互独立，且拟合模型为 $\hat{f} \equiv f'$，那么编码器 $g(X, X_{\mathrm{pa}}) = f(X_{\mathrm{pa}}, U) - \hat{f}(X_{\mathrm{pa}}) = U$，根据定义，$U$ 与 X_{pa}相互独立。[2] 在双射结构因果模型的反事实可识别性方面，其他相关结果也采用了相同假设，例如（Nasr-Esfahany等人，2023，定理5.3）以及（Nasr-Esfahany & Kiciman，2023）中的定理5证明。我们通过考察父节点与编码值之间的依赖关系，进行实证检验以进一步确认这一假设。实验结果表明，扩散因果模型始终无法拒绝独立性原假设。这意味着实践中可以找到相互独立的编码。我们在附录B中提供这些实验的详细信息。[3]。

(2) 在加性噪声模型下，假设2始终成立，其中 $f(X_{\mathrm{pa}}, U) := f'(X_{\mathrm{pa}}) + U$。在满足 Zhang & Hyvarinen, 2012 所述可识别性的标准要求时，后非线性模型也满足该假设。[4] 异方差噪声模型也满足假设 2（Strobl & Lasko, 2023）。[5] 同样，近期有关反事实可识别性的结果，例如（Nasr-Esfahany & Kiciman, 2023, 定理5）和（Lu等，2020, 定理1），也采用了相同的假设。

通过假设 U严格递增，我们无需区分诸如 $f(X_{\mathrm{pa}}, U) := X_{\mathrm{pa}} + U$与$f(X_{\mathrm{pa}}, U) := X_{\mathrm{pa}} - U$ 这样的情况；对于对称分布而言，这些情况无法区分 U（另请参见下文的备注）。例如，对于任意固定值 X_{pa}，$f(X_{\mathrm{pa}}, U) := X_{\mathrm{pa}} - U$ 并非关于 U递增，因此这类结构方程会在定理1中自动排除。

(3) 我们可以对均匀噪声 U 进行变换，以得到其他设定，例如加性高斯噪声。对于累积分布函数 F可逆的连续随机变量 U'，以及结构方程 $f(\cdot, F(\cdot))$，我们有$U' \stackrel{d}{=} F^{-1}(U)$，相应的结果同样成立。

现在，我们讨论定理1对反事实结果估计的一些推论。

---

[2]一般而言，如果我们能够用某个 $\hat{f}$对 f作出良好近似，那么编码 $g(X, X_{\mathrm{pa}}) = X - \hat{f}(X_{\mathrm{pa}})$ 将接近 U，Hoyer等（2008）也指出了这一点。[3]为了促进独立性，也可以修改原始扩散模型的训练目标，加入希尔伯特–施密特独立性准则（HSIC）（Gretton 等人，2007）正则化项。我们的实验并未显示采用这一修改后目标有明显益处，因此将进一步研究留待未来工作。[4]Zhang & Hyvarinen（2012）（见公式2）将后非线性模型定义为 $f(X_{\mathrm{pa}}, U) := f_2(f_1(X_{\mathrm{pa}}) + U)$，其中 X_{pa}和 U相互独立，函数 f_1 为非恒定函数，且 f_2 可逆。对 f_2 的可逆性假设意味着 f关于 U严格递增（或递减）。此外，Zhang & Hyvarinen（2012）为保证可识别性，对 f_2 作出了可微性假设（见假设A1），这意味着 f 关于 U可微。[5]异方差噪声模型通常定义为 $f(X_{\mathrm{pa}}, U) := f'(X_{pa}) + g(X_{pa})$，其中假设函数 g 严格为正（见 Strobl & Lasko（2023）中的定义1），这使其与定理1中的假设2兼容。

1. 完美估计。 现在利用定理1考察一种条件：在该条件下，编码器-解码器模型产生的反事实估计与真实反事实结果相符。[6]。这里的思路是，如果编码与解码步骤没有丢失任何信息，即 $h(g(X, X_{\mathrm{pa}})\ X_{\mathrm{pa}}) = X$，并且假设定理1成立（$g(X, X_{\mathrm{pa}}) = \tilde{q}(U)$），那么我们有 $h(\tilde{q}(U)\ X_{\mathrm{pa}}) = X = f(X_{\mathrm{pa}}, \tilde{q}^{-1}(U))$。这意味着在溯因步骤中，编码器-解码器模型能够恢复 $\tilde{q}(U)$；而在预测步骤中，它先对恢复出的外生变量应用 $\tilde{q}$ 的逆变换，然后 f。因此，反事实估计等于真实反事实结果。我们在推论1中对此作形式化表述。

推论*1*。假设定理*1*的条件成立。此外，假设编码器-解码器模型对 (g, h) 满足： $h(g(X, X_{\mathrm{pa}})\ X_{\mathrm{pa}}) = X$。考虑一个事实样本对 $x^{\mathrm{F}} := (x, x_{\mathrm{pa}})$，其中 $x := f(x_{\mathrm{pa}}, u)$，以及干预 $do(X_{\mathrm{pa}} := \gamma)$。那么，由 $h(g(x, x_{\mathrm{pa}})\ \gamma)$ 给出的反事实估计与真实反事实结果 $x^{\mathrm{CF}} := f(\gamma, u)$ 相匹配。

与近期相关工作的比较。 Nasr-Esfahany & Kiciman (2023) 和 Nasr-Esfahany等人 (2023) 的近期研究探讨了如何利用学习得到的结构因果模型估计反事实结果。具体而言，Nasr-Esfahany & Kiciman (2023, 定理5) 考虑了这样一种情形：结构因果模型 $X := f(X_{\mathrm{pa}}, U)$ 是通过一种双射（深度条件生成）模型 $\hat{} f(X_{\mathrm{pa}}, \hat{} U)$ 学得的。Nasr-Esfahany等人 (2023, 定理5.3) 考虑了一个密切相关的问题，即学习真实双射结构因果模型。由于与我们的设定不同，他们没有明确考虑编码器-解码器模型，因此我们这些结果所依据的条件与他们的条件无法直接比较。我们的结果为编码器 g 和解码器 h恢复正确反事实结果提供了精确条件，并且可将这些结果扩展到在较宽松假设下获得反事实估计误差界；此前尚无研究探讨这一问题。

（Lu等，2020）也在强化学习的另一种情境下确立了反事实可识别性。他们的结果依赖于关于状态-动作对的不同假设。此外，我们的证明技巧与Lu等人（2020）截然不同；他们依赖于分析条件分位数的技巧，而本文采用的是代数技巧。

2. 估计误差。定理1的另一个推论是，它可以根据编码器-解码器模型的重构误差来界定反事实误差。非正式地说，以下推论表明，如果重构$h(g(X,\ X_{\mathrm{pa}})\ X_{\mathrm{pa}})$与 X（在某个指标 $d(\cdot,\ \cdot)$下衡量）“足够接近”，那么此类编码器-解码器模型便能提供“良好”的反事实估计。据我们所知，这是首个确立此类编码器-解码器模型的反事实误差与重构误差之间界限的结果。

推论 *2.* 令 $\gamma \geq 0$。假设定理*1*的条件成立。此外，假设编码器-解码器模型对 (g, h) 在某个指标 d（例如 $\|\cdot\|_2$）下的重构误差小于 τ： $d(h(g(X, X_{\mathrm{pa}})\ X_{\mathrm{pa}})\ X) \leq \tau$。考虑事实样本对 $x^{\mathrm{F}} := (x, x_{\mathrm{pa}})$，其中 $x := f(x_{\mathrm{pa}}, u)$，以及干预 $do(X_{\mathrm{pa}} := \gamma)$。那么，真实反事实 $x^{\mathrm{CF}} := f(\gamma, u)$ 与由 $h(g(x, x_{\mathrm{pa}})\ \gamma)$ 给出的反事实估计之间的误差至多为τ。换言之， $d(h(g(x, x_{\mathrm{pa}})\ \gamma)\ x^{\mathrm{CF}}) \leq \tau$。

上述结果表明，重构误差可以作为反事实误差的估计。虽然 τ 的真实值未知，但我们可以通过计算数据集上的重构误差，得到一个合理的上界。

4.1 将定理1扩展到高维情形

本节将定理1扩展到高维情形，并利用该扩展给出反事实可识别性和估计误差的结果。由于现在处理的是向量值函数，我们使用雅可比矩阵。令 $Jf_{x_{\mathrm{pa}}}$ 表示在 $X_{\mathrm{pa}} = x_{\mathrm{pa}}$ 处对 $f(X_{\mathrm{pa}},\ U)$ 关于 U 求雅可比矩阵后得到的雅可比矩阵。令 $Jg_{x_{\mathrm{pa}}}$ 表示在 $X_{\mathrm{pa}} = x_{\mathrm{pa}}$ 处对$g(X,\ X_{\mathrm{pa}})$ 关于 X 求雅可比矩阵后得到的雅可比矩阵。

定理 2. 假设对于 $X \in \mathcal{X} \subset \mathbb{R}^m$ 和连续外生噪声 $U \sim$ 均匀分布$[0,1]^m$，其中 $m \geq 3$，且 X 满足结构方程

$$X = f(X_{\mathrm{pa}}, U) \tag{6}$$

[6]请注意，反事实结果的可识别性并不要求结构因果模型具有可识别性。

其中 $X_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$ 是节点 X 的父节点，且 $U \perp\!\!\!\perp X_{\mathrm{pa}}$。考虑一个编码器-解码器模型，其编码函数为 $g: \mathcal{X} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{Z}$，解码函数为 $h: \mathcal{Z} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{X}$，

$$Z := g(X, X_{\mathrm{pa}}), \quad \hat{X} := h(Z, X_{\mathrm{pa}}). \tag{7}$$

假设满足以下条件：

1. 编码独立于父节点值，$g(X, X_{\mathrm{pa}}) \perp\!\!\!\perp X_{\mathrm{pa}}$ *2.* 结构方程 f 关于 U 可逆且可微，并且对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$，$Jf_{x_{\mathrm{pa}}}$ 均为正定。*3.* 编码 g 关于 X 可逆且可微，并且对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$，$Jg_{x_{\mathrm{pa}}}$ 均为正定。*4.* 编码 $q_{x_{\mathrm{pa}}}(U) := g(f(x_{\mathrm{pa}}, U)\ x_{\mathrm{pa}})$ 满足 $Jq_{x_{\mathrm{pa}}}|_{q^{-1}_{x_{\mathrm{pa}}}(z)} = c(x_{\mathrm{pa}})A$，对所有 $z \in \mathcal{Z}$ 和 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$ 均成立，其中 c 是标量函数，A 是正交矩阵。于是，$g(f(X_{\mathrm{pa}}, U)\ X_{\mathrm{pa}}) = \tilde{q}(U)$，其中存在一个可逆函数 $\tilde{q}$。

定理 2 中假设1、2和3 背后的解释与定理1中的相应假设类似。不过，假设 4 是技术性的，我们将在下文解释其必要性。在推论3和4（附录 A）中，我们将推论1和2重新表述为适用于这一高维情形。其证明与推论1和2的证明完全相同，唯一的变化是这些证明中定理1的作用现在由定理 2 取代。

关于 Nasr-Esfahany & Kiciman (2023) 的否定性结果。Nasr-Esfahany & Kiciman (2023) 在多维外生噪声条件下提出了一项一般性的反事实不可识别性结果。Nasr-Esfahany & Kiciman (2023) 的构造考虑了两个在分布上无法区分的结构方程 f，f'。形式上，设 $R \in \mathbb{R}^{m \times m}$ 为旋转矩阵，$U \in \mathbb{R}^m$ 为标准（各向同性）高斯随机向量。定义如下：

$$f'(X_{\mathrm{pa}}, U) = \begin{cases} f(X_{\mathrm{pa}}, U) & \text{for } X_{\mathrm{pa}} \in A \\ f(X_{\mathrm{pa}}, R \cdot U) & \text{for } X_{\mathrm{pa}} \in B \end{cases}, \tag{8}$$

其中，定义域 $\mathcal{X}_{\mathrm{pa}}$ 被划分为互不相交的 A 和 B。现在，对于证据位于 A 且干预位于 B（或反过来）的反事实查询，f 和 f' 会产生不同的反事实结果。

在定理 2 中，我们通过假设能够构造一种由假设 4 刻画的“特殊”编码，避开了这一不可识别性结果。具体而言，考虑在特定父变量取值 x_{pa} 下，将外生噪声 U 映射为编码 $q_{x_{\mathrm{pa}}}$。该假设指出，编码的雅可比矩阵等于 $c(x_{\mathrm{pa}})A$，其中 c 是标量函数，A 是正交矩阵。然而，必须承认，这一假设限制性很强且难以验证，更不用说仅凭观测数据在实践中构造这种编码也颇具挑战。我们的初衷是希望这些初步想法能成为应对这一不可识别性结果的起点，并期待后续结果进一步完善和拓展这些想法。

5 实验评估

在本节中，我们在合成数据和真实世界数据上评估 DCM 回答因果查询的经验性能。附录 D.2 中给出了额外的半合成实验评估。对于半合成实验，我们使用*Sachs* 数据集（Sachs 等人，2005年）。

扩散模型实现与训练。 在 DCM 中实现 ε_θ 模型时，我们使用一个包含三个隐藏层的简单全连接神经网络，每层大小为 $[128, 256, 256]$，并采用 SiLU 激活函数（Elfwing 等人，2018年）。我们使用 Adam 训练模型，学习率为 1e-4，批量大小为 64，训练 500 个训练轮数。由于根节点没有父节点，唯一形式的反事实推理是直接对根节点进行干预，这可以轻易完成。因此，对于根节点，我们不训练扩散模型，而是从训练数据的经验分布中进行采样。有关扩散模型参数的更多细节，请参见附录 C.1。

对比方法。 为确保比较公平，我们的评估重点关注同时支持干预估计和反事实估计的方法。基于这一标准，我们主要将 DCM 与近期提出的两种先进方案 VACA（S´anchez-Martin et al., 2022）和 CAREFL（Khemakhemet al.,2021），以及一种通用

回归模型进行比较。该模型假设采用加性噪声模型，我们将其称为 ANM。[7] 对于 VACA 和 CAREFL，我们使用各自作者提供的代码。ANM 方法会在多种模型中进行模型选择，包括线性回归和梯度提升回归器；我们使用广受欢迎的因果推断软件包 *DoWhy* 中的实现（Sharma et al., 2019;Bl¨obaum et al., 2022）。关于 ANM 如何回答因果查询的更多详情见附录 C.2； VACA 、 CAREFL 和 ANM 的实现细节见附录 C.1。

5.1 合成数据实验

为生成定量结果，我们使用合成实验，因为我们知道精确的结构方程，因此能够获得真实的观测分布、干预分布和反事实分布。我们提供了两组合成实验：一组是本文给出的两个较大图，另一组是附录D.3中给出的四个较小图。这两个较大图包括一种阶梯图结构（见图9）和一种随机生成的图结构。两个图均由10个三维节点组成，而随机图则是随机采样得到的有向无环图（更多详情见附录 C.3）。由于每个扩散模型仅将父节点作为输入，因此对每个节点建模只取决于该节点的入度（即因果父节点的数量）。

按照 S´anchez-Martin 等人（2022年）的做法，对于观测分布和干预分布，我们报告真实分布与估计分布之间的最大均值差异（最大均值差异）（Gretton 等人，2012年）。对于反事实估计，我们报告真实反事实值与估计反事实值之间的均方误差（均方误差）。同样按照 S´anchez-Martin 等人（2022年）的做法，我们考虑两大类结构方程：

1. 加性噪声模型（NLIN）： $f_i(X_{\mathrm{pa}_i},\ U_i) = f'(X_{\mathrm{pa}_i}) + U_i$。特别是，我们关注其中 f_i为非线性的情况。

2. 非加性噪声模型（NADD）： $f_i(X_{\mathrm{pa}_i},\ U_i)$ 是 X_{pa_i} 和 U_i 的任意函数。

为防止超参数过拟合，我们会为每次初始化随机生成这些结构方程。每个结构方程由一个神经网络构成，该网络包含一个由 16 个单元组成的隐藏层，并采用 SiLU 激活函数（Elfwing 等人，2018年）；其随机权重从 $[-1, 1]$ 中采样。

ˆ 每次模拟都会生成 $n = 5000$ 个训练数据样本。设 M 为拟合得到的因果模型， $M^\star$ 为真实因果模型；二者均能够生成观测样本和干预样本，并回答反事实查询。我们针对每种图与结构方程类型组合进行 20 次不同的初始化评估，并报告均值。关于观测、干预和反事实评估框架的更多细节，请参见附录 C.4。

合成实验结果。表1展示了所有评估模型在观测查询、干预查询和反事实查询上的表现，这些结果是对模型和训练数据分别进行20次初始化后取平均所得，每行中的最小值以粗体标出。为便于阅读，数值均乘以100。我们还在图1（附录D）中提供了性能箱线图。

我们发现，DCM和ANM是最具竞争力的方法，在观测查询和干预查询上的表现相近。如果ANM是设定正确的模型，那么在回归模型对数据拟合良好的前提下，ANM的编码应接近真实编码。非线性设置下的结果印证了这一点：ANM表现良好，但难以胜过DCM，这或许是因为使用经典模型拟合神经网络的复杂性较高。需要注意的是，由于观测查询和干预查询本身比反事实查询更简单，因此相较于其他基线方法，我们观察到的性能提升较小是合理的。

与 VACA 和 CAREFL 相比，我们提出的 DCM 方法表现更优，优势往往可达一个数量级。DCM 优于 CAREFL，可能是因为 DCM 使用了因果图，而 CAREFL 仅依赖于非唯一的因果排序。对于 VACA，其图神经网络编码器-解码器的表达能力有限，可能是其表现较差的原因，尤其是在处理本文所考虑的多变量、多维复杂结构方程时；VACA 的作者也承认了这一不足（S´anchez-Martin 等，2022年）。此外，VACA 执行近似推断，例如，在执行 do($X_1 = 2$) 反事实查询时，由于重构不完善，预测的 X_1 反事实值并不恰好为 2。这种设计选择可能导致下游误差不断累积，或许可以解释性能差异。为避免对此特性造成惩罚，所有指标均使用干预节点的下游节点计算。为评估计算效率，我们在单一实验设置下比较了 DCM、VACA 和 CAREFL（见附录D.1）。结果表明，DCM

[7]尽管我们已尽最大努力，但由于难以将 Javaloy 等人提出的最新方法（2023年）中的代码适配到我们的设置中，我们未能与之进行恰当的比较。

SCM		指标	DCM ($\times 10^{-2}$)	ANM ($\times 10^{-2}$)	VACA ($\times 10^{-2}$)	CAREFL ($\times 10^{-2}$)
Ladder	NLIN	观测 MMD	**0.44±0.14**	0.63±0.21	2.82±0.83	13.41±1.14
		积分最大均值差异	**1.63±0.20**	1.80±0.17	4.48±0.78	15.01±1.23
		反事实均方误差	**3.42±1.67**	10.65±2.48	41.03±19.00	17.46±6.04
	NADD	观测 MMD	**0.32±0.11**	0.40±0.16	3.22±1.05	14.60±1.34
		积分最大均值差异	**1.54±0.17**	1.57±0.15	5.13±1.16	16.87±1.85
		反事实均方误差	**4.28±2.39**	10.71±5.47	27.42±12.34	22.26±13.75
Random	NLIN	观测 MMD	**0.28±0.12**	0.47±0.15	1.82±0.73	12.11±1.21
		积分最大均值差异	**1.45±0.07**	1.88±0.22	3.52±1.03	14.15±2.34
		反事实均方误差	**9.51±13.12**	23.68±28.49	82.10±78.49	52.57±82.03
	NADD	观测 MMD	**0.19±0.05**	0.31±0.14	2.09±0.60	12.63±1.10
		积分最大均值差异	**1.42±0.25**	1.73±0.44	4.24±1.40	14.65±1.76
		反事实均方误差	**20.13±57.52**	44.76±86.02	124.82±275.09	54.29±83.68

表1：在非线性和非加性设定下，对阶梯型和随机结构因果模型进行观测、干预和反事实查询所得结果的均值±标准差；模型和训练数据采用20次随机初始化。为便于查看，数值均乘以100。

算法	绝对误差中位数	平均绝对误差
DCM	**0.5981 ± 9.5e-3**	**0.5779 ± 1.9e-3**
CAREFL	0.5983 ± 3.0e-2	0.6004 ± 2.2e-2
ANM	0.6746 ± 4.8e-8	0.6498 ± 1.4e-6
线性结构因果模型	0.6045 ± 0	0.6042 ± 0

表2：在功能性磁共振成像数据集上进行干预预测的表现，形式为中位数/平均绝对误差±标准差，取 10 个随机种子的均值。由于实现上的困难，我们未纳入 VACA。CAREFL、ANM 和线性结构因果模型的结果与 (Khemakhem 等，2021) 报告的结果一致。需要指出的是，线性结构因果模型的标准差为零，因为岭回归模型不会随随机种子而变化。

在计算效率上显著高于 VACA 和 CAREFL，训练耗时少了 7 倍。最后，DCM 的标准差相较于其他模型较小，表明其表现较为稳定，也体现了我们所提方法的稳健性。

5.2 真实数据实验 I

我们使用 (Khemakhem 等，2021) 的实验设置，在 (Thompson 等，2020) 的电刺激干预功能磁共振成像数据上评估模型，从而检验 DCM 在真实世界干预数据上的表现。该功能磁共振成像数据包含 14 名药物难治性癫痫患者的样本，记录了扣带回 (CG) 和赫施尔回 (HG) 的时间序列。假定的底层因果结构为二元图 CG $\rightarrow$ HG。我们的干预真实值数据包含一个经过干预的 CG 值和一个观测到的 HG 样本。有关该数据集的更详细讨论，请参阅 (Thompson 等，2020; Khemakhem 等，2021)。

在表2中，我们注意到性能差异比合成数据结果中的差异更小。我们认为这是由两个原因造成的。首先，数据似乎本质上接近线性，这一点可以从标准岭回归模型（线性结构因果模型）表现相对相似看出。其次，我们只有一个真实的干预

值，而不是来自干预分布的多个样本。因此，我们只能基于这一个值计算绝对误差，而无法评估真实干预分布与预测干预分布之间的最大均值差异（MMD）。具体而言，在上表中，我们针对 14 位患者中的每一位，计算模型预测值与干预样本之间的绝对误差，并报告其均值/中位数。只有一个干预值可能会引入大量不可约误差，从而人为地抬高误差值。有关误差膨胀的更多详情，请参见附录 C.5。

6 总结

我们证明，扩散模型，尤其是 DDIM 形式（支持唯一的编码与解码），为近似干预（do 算子）和反事实（溯因-行动-预测）步骤提供了灵活且实用的框架。我们提出的 DCM 方法不受有向无环图结构的限制。实验结果表明，在各种结构方程和图类别下，DCM 在三种因果场景——观测、干预和反事实查询——中的表现均优于竞争方法。

尽管本文未探讨这一点，所提出的 DCM 方法也可以自然地扩展到任何适用扩散模型的场景，例如类别数据、图像等。对于更高维的空间，我们认为扩散因果模型应能良好扩展，因为扩散模型通常用于高维图像场景，并达到最先进水平。此外，我们还可以借鉴许多优化和实现技巧，因为扩散模型是一个非常活跃的研究领域。

所提出的方法确实存在一些局限性。例如，与前文提到的所有相关方法一样，DCM 不考虑未观测混杂。理论分析需要满足一些假设，而其中并非所有假设都易于检验。不过，我们的实践结果表明，即使违反了理论保证所需的部分假设，DCM 仍能取得有竞争力的实证表现。

参考文献

Ahmed M Alaa 和 Mihaela Van Der Schaar。利用多任务高斯过程对个体化治疗效应进行贝叶斯推断。神经信息处理系统进展，30，2017年。

Sourabh Balgi、Jose M Pena 和 Adel Daoud。利用因果图归一化流对社会科学中的个性化公共政策进行分析。载于人工智能促进协会会议论文集，第36卷，第11810–11818页，2022年。

David E Blair。反演理论与保角映射，第9卷。美国数学学会，2000年。

Patrick Bl¨obaum、Peter G¨otz、Kailash Budhathoki、Atalanti A. Mastakouri 和 Dominik Janzing。DoWhy-GCM：DoWhy 的扩展，用于图形因果模型中的因果推断。*arXiv* 预印本 *arXiv:2206.06821*，2022年。

Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层次化图像数据库。载于*2009*年 *IEEE* 计算机视觉与模式识别会议，第248–255页，2009年。doi：10. 1109/CVPR.2009.5206848。

Prafulla Dhariwal 和 Alexander Nichol。《扩散模型在图像合成方面胜过生成对抗网络》。载于 M. Ranzato、A. Beygelz- imer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，第 34 卷，第 8780–8794 页。Curran Associates, Inc.，2021年。网址：https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf。

Conor Durkan、Artur Bekasov、Iain Murray 和 George Papamakarios。《神经样条流》。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett（编），神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf。

Stefan Elfwing、Eiji Uchibe 和 Kenji Doya。用于强化学习中神经网络函数逼近的 Sigmoid加权线性单元。《神经网络》，107：3–11，2018年。ISSN 0893-6080。doi: https: //doi.org/10.1016/j.neunet.2017.12.012。URL https://www.sciencedirect.com/science/article/ pii/S0893608017302976。深度强化学习特刊。

Sergio Garrido、Stanislav Borysov、Jeppe Rich 和 Francisco Pereira。使用神经自回归密度估计器估计因果效应。因果推断期刊，9(1)：211–218，2021年。Arthur Gretton、Kenji Fukumizu、Choon Teo、Le Song、Bernhard Sch¨olkopf 和 Alex Smola。核独立性统计检验。载于 J. Platt、D. Koller、Y. Singer 和 S. Roweis（编），神经信息处理系统进展，第20卷。Curran Associates, Inc.，2007年。URL https://proceedings.neurips.cc/paper/2007/file/d5cfead94f5350c12c322b5b664544c1-Paper.pdf。Arthur Gretton、Karsten M. Borgwardt、Malte J. Rasch、Bernhard Sch¨olkopf 和 Alexander Smola。核双样本检验。机器学习研究期刊，13(25)：723–773，2012年。URL http://jmlr.org/ papers/v13/gretton12a.html。Jennifer L Hill。用于因果推断的贝叶斯非参数建模。计算与图形统计学期刊，20(1)：217–240，2011年。J onathan Ho 和 Tim Salimans。无分类器扩散引导。载于*NeurIPS 2021* 深度生成模型与下游应用研讨会，2021年。URL https://openreview.net/forum?id=qw8AKxfYbI。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第33卷，第6840–6851页。Curran Associates, Inc.，2020年。URL https://proceedings.neurips.cc/paper/2020/ file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf。Patrik Hoyer、Dominik Janzing、Joris M Mooij、Jonas Peters 和 Bernhard Sch¨olkopf。基于加性噪声模型的非线性因果发现。神经信息处理系统进展，21，2008年。Adri´an Javaloy、Pablo S´anchez-Mart´ın 和 Isabel Valera。因果归一化流：从理论到实践，2023年。Amir-Hossein Karimi、Julius Von K¨ugelgen、Bernhard Sch¨olkopf 和 Isabel Valera。在不完备因果知识下的算法补救：一种概率方法。神经信息处理系统进展，33：265–277，2020年。Ilyes Khemakhem、Ricardo Monti、Robert Leech 和 Aapo Hyvarinen。因果自回归流。载于 Arindam Banerjee 和 Kenji Fukumizu（编），第*24*届人工智能与统计国际会议论文集，机器学习研究论文集第130卷，第3520–3528页。PMLR，2021年4月13–15日。URL https://proceedings.mlr.press/v130/khemakhem21a.html。Murat Kocaoglu、Christopher Snyder、Alexandros G Dimakis 和 Sriram Vishwanath。因果生成对抗网络：通过对抗训练学习因果隐式生成模型。载于国际学习表征会议，2018年。Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro。DiffWave：一种多功能音频合成扩散模型。载于国际学习表征会议，2021年。URL https://openreview. net/forum?id=a-xFK8Ymz5J。Robert J LaLonde。使用实验数据评估培训项目的计量经济学评估。美国经济评论，第604–620页，1986年。Chaochao Lu、Biwei Huang、Ke Wang、Jos´e Miguel Hern´andez-Lobato、Kun Zhang 和 Bernhard Sch¨olkopf。通过基于反事实的数据增强实现样本高效强化学习。*arXiv* 预印本 *arXiv:2012.09092*，2020年。Raha Moraffah、Bahman Moraffah、Mansooreh Karami、Adrienne Raglin 和 Huan Liu。CAN：一种用于学习观测分布和干预分布的因果对抗网络。*arXiv* 预印本 *arXiv:2008.11376*，2020年。Krikamol Muandet、Motonobu Kanagawa、Sorawit Saengkyongam 和 Sanparith Marukatat。反事实均值嵌入。*J. Mach. Learn. Res.*，22：162–1，2021年。Arash Nasr-Esfahany 和 Emre Kiciman。学习得到的结构因果模型的反事实（非）可识别性。*arXiv* 预印本 *arXiv:2301.09031*，2023年。

Arash Nasr-Esfahany、Mohammad Alizadeh 和 Devavrat Shah。双射因果模型的反事实可识别性。*arXiv* 预印本 *arXiv:2302.02228*，2023年。Alexander Quinn Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第8162–8171页。PMLR，2021年7月18–24日。网址：https: // proceedings.mlr.press/v139/nichol21a.html。´Alvaro Parafita 和 Jordi Vitri`a。利用深度因果图进行因果推断。*arXiv* 预印本 *arXiv:2006.08380*，2020年。Nick Pawlowski、Daniel Coelho de Castro 和 Ben Glocker。用于可处理反事实推断的深度结构因果模型。神经信息处理系统进展，33：857–869，2020年。J. Pearl、M. Glymour 和 N.P. Jewell。《统计学中的因果推断：入门》。Wiley，2016年。ISBN 9781119186847。网址：https://books.google.com/books?id=L3G-CgAAQBAJ。Judea Pearl。统计学中的因果推断：概述。*Statistics Surveys*，3(none)：96–146，2009a。doi：10.1214/ 09-SS057。网址：https://doi.org/10.1214/09-SS057。

Judea Pearl。《统计学中的因果推断：概述》，2009年b。

Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。《基于 CLIP 潜变量的分层文本条件图像生成》，2022年。网址：https://arxiv.org/abs/2204.06125。

Karen Sachs、Omar Perez、Dana Pe'er、Douglas A. Lauffenburger 和 Garry P. Nolan。《从多参数单细胞数据中推导因果蛋白质信号网络》。《科学》，308(5721):523–529，2005年。doi：10.1126/science. 1105809。网址：https://www.science.org/doi/abs/10.1126/science.1105809。

Saptarshi Saha 和 Utpal Garain。《用于回答反事实查询的噪声溯因：实践展望》。机器学习研究汇刊，2022年。

Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S. Sara Mahdavi、Rapha Gontijo Lopes、Tim Salimans、Jonathan Ho、David J Fleet 和 Mohammad Norouzi。具备深度语言理解的照片级文本生成图像扩散模型，2022年。网址：https://arxiv.org/abs/2205.11487。

Pedro Sanchez 和 Sotirios A. Tsaftaris。用于反事实估计的扩散因果模型。在*CLeaR*，2022年。

Pedro Sanchez、Xiao Liu、Alison Q O'Neil 和 Sotirios A Tsaftaris。通过拓扑排序实现因果发现的扩散模型。在第十一届国际学习表征会议，2022年。

Uri Shalit、Fredrik D Johansson 和 David Sontag。个体处理效应估计：泛化界与算法。在国际机器学习会议，第3076–3085页。PMLR，2017年。

Amit Sharma、Emre Kiciman 等。DoWhy：用于因果推断的 Python 软件包。https://github.com/microsoft/dowhy，2019年。

Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。使用非平衡热力学进行深度无监督学习。载于 Francis Bach 和 David Blei（编），第*32*届国际机器学习会议论文集，第37卷，机器学习研究论文集，第2256–2265页，法国里尔，2015年7月7–9日。PMLR。网址：https://proceedings.mlr.press/v37/sohl-dickstein15. html。

Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=St1giarCHLP。

Eric V Strobl 和 Thomas A Lasko。使用异方差噪声模型识别患者特异性根本原因。《计算科学杂志》，72:102099，2023年。

Pablo S´anchez-Martin、Miriam Rateike 和 Isabel Valera。Vaca：为因果查询设计变分图自编码器。人工智能促进协会会议论文集，36(7):8159–8168，2022年6月。doi: 10.1609/ aaai.v36i7.20789。URL https://ojs.aaai.org/index.php/AAAI/article/view/20789。

W. H. Thompson、R. Nair、H. Oya、O. Esteban、J. M. Shine、C. I. Petkov、R. A. Poldrack、M. Howard 和 R. Adolphs。关于人脑同步颅内刺激与功能性磁共振成像的数据资源。*Scientific Data*，7(1):258，2020年8月。ISSN 2052-4463。doi：10.1038/s41597-020-00595-y。URL：https://doi.org/10. 1038/s41597-020-00595-y。

Matej Ze˘cevi´c、Devendra Dhami、Athresh Karanam、Sriraam Natarajan 和 Kristian Kersting。干预式和网络：使用可处理概率模型进行因果推断。神经信息处理系统进展，34:15019–15031，2021a。

Matej Ze˘cevi´c、Devendra Singh Dhami、Petar Veli˘ckovi´c 和 Kristian Kersting。《将图神经网络与结构因果模型联系起来》。*arXiv* 预印本 *arXiv:2109.04173*，2021b。

Kun Zhang 和 Aapo Hyvarinen。《后非线性因果模型的可识别性》。*arXiv* 预印本 *arXiv:1205.2599*，2012年。

第4节缺失的细节

符号约定。对于两个集合 $\mathcal{X}, \mathcal{Y}$，一个映射 $f\colon \mathcal{X} \mapsto \mathcal{Y}$，以及一个集合 $S \subset \mathcal{X}$，我们定义 $f(S) = \{f(x)\colon x \in S\}$。对于 $x \in \mathcal{X}$，我们定义 $x + S = \{x + x'\colon x' \in \mathcal{X}\}$。对于随机变量 X，将 $p_X(x)$ 定义为其在 x 处的概率密度函数（PDF）。我们用 p.d. 表示正定矩阵，并用 $Jf|_x$ 表示在 x 处计算的 f 的雅可比矩阵。对于有两个输入的函数 $f(\cdot,\cdot)$，我们定义 $f_x(Y) := f(x, Y)$ 和 $f_y(X) := f(X, y)$。

引理1。对于 $\mathcal{U}$、$\mathcal{Z} \subset \mathbb{R}$，考虑一族可逆函数 $q_{x_{\mathrm{pa}}}\colon \mathcal{U} \to \mathcal{Z}$，其中 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$，则 $\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z)) = c(z)$ 对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$ 成立，当且仅当 $q_{x_{\mathrm{pa}}}$ 可以表示为

$$q_{x_{\mathrm{pa}}}(u) = q(u + r(x_{\mathrm{pa}}))$$

对于某个函数 r 和可逆的 q。

证明。首先，对于反向，设 $q_{x_{\mathrm{pa}}}(u) = q(u + r(x_{\mathrm{pa}}))$。那么

$$\frac{dq_{x_{\mathrm{pa}}}}{du}(u) = \frac{dq}{du}(u + r(x_{\mathrm{pa}})).$$

现在代入 $u = q^{-1}_{x_{\mathrm{pa}}}(z) = q^{-1}(z) - r(x_{\mathrm{pa}})$，

$$\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z)) = \frac{dq}{du}(q^{-1}(z) - r(x_{\mathrm{pa}}) + r(x_{\mathrm{pa}})) = \frac{dq}{du}(q^{-1}(z)) = c(z).$$

因此，$\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z))$ 是 z 的函数。

对于正向，设 $\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z)) = c(z)$。定义 $s_{x_{\mathrm{pa}}}\colon \mathcal{Z} \to \mathcal{U}$ 为 $q_{x_{\mathrm{pa}}}$ 的逆函数。根据反函数定理以及假设。

$$\frac{ds_{x_{\mathrm{pa}}}}{dz}(z) = \frac{dq^{-1}_{x_{\mathrm{pa}}}}{dz}(z) = \frac{1}{\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z))} = \frac{1}{c(z)}$$

对所有 x_{pa} 均成立。由于对于所有 x_{pa}，$s_{x_{\mathrm{pa}}}$ 的导数都相等，根据中值定理，所有 $s_{x_{\mathrm{pa}}}$ 彼此之间都只相差加性平移。不失一般性，我们可以考虑任意一个固定的 $x_{\mathrm{pa}_0} \in \mathcal{X}_{\mathrm{pa}}$，并将 $s_{x_{\mathrm{pa}}}$ 重新参数化为

$$s_{x_{\mathrm{pa}}}(z) = s_{x_{\mathrm{pa}_0}}(z) - r(x_{\mathrm{pa}}).$$

令 $u = s_{x_{\mathrm{pa}}}(z)$。那么我们有

$$s_{x_{\mathrm{pa}_0}}(z) = u + r(x_{\mathrm{pa}})$$
$$q_{x_{\mathrm{pa}}}(u) = z = q_{x_{\mathrm{pa}_0}}(u + r(x_{\mathrm{pa}})),$$

通过选择$q = q_{x_{\mathrm{pa}_0}}$，我们便得到了所需的表示。 □

定理1。 假设对于 $X \in \mathcal{X} \subset \mathbb{R}$ 和外生噪声 $U \sim \mathrm{Unif}[0,1]$， X 满足结构方程：$X := f(X_{\mathrm{pa}}, U)$，其中 $X_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$是节点 X和 $U \perp\!\!\!\perp X_{\mathrm{pa}}$的父节点。考虑一个编码器-解码器模型，其编码函数为 g：$\mathcal{X} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{Z}$，解码函数为 h： $\mathcal{Z} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{X}$， $Z := g(X, X_{\mathrm{pa}})$，$\hat{X} := h(Z, X_{\mathrm{pa}})$。假设满足以下条件：

1. 编码独立于父节点值， $g(X, X_{\mathrm{pa}}) \perp\!\!\!\perp X_{\mathrm{pa}}$.
2. 结　　构方程 f 可微，并且对于所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$，相对于 U 严格递增　　.
3. 编码 g 可逆，并且对于所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$，相对于 X 可微。

于是，对于可逆函数 $\tilde{q}$， $g(X, X_{\mathrm{pa}}) = \tilde{q}(U)$。

证明。 首先，我们证明 $g(X, X_{\mathrm{pa}}) = g(f(X_{\mathrm{pa}}, U) X_{\mathrm{pa}})$ 仅是 U 的函数。

由于连续性和可逆性意味着严格单调，不失一般性，假设 g 是严格递增函数（若不是，我们可以用 $-g$ 替换 g，并使用 $h(-Z, X_{\mathrm{pa}})$）。根据函数复合的性质，$q_{x_{\mathrm{pa}}}(U) := g(f(x_{\mathrm{pa}}, U) x_{\mathrm{pa}})$ 关于 U 也可微且严格递增。此外，由于其严格单调，它也是可逆的。

根据编码 Z 独立于 X_{pa} 这一假设，

$$Z = q_{x_{\mathrm{pa}}}(U) \perp\!\!\!\perp X_{\mathrm{pa}}. \tag{9}$$

因此， Z 的条件分布不依赖于 X_{pa}。利用 $U \perp\!\!\!\perp X_{\mathrm{pa}}$ 这一假设，对于 Z 支撑集中的所有$x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$ 和 z，根据密度变换公式，

$$p_Z(z) = \frac{p_U(q^{-1}_{x_{\mathrm{pa}}}(z))}{\left|\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z))\right|} = \frac{\mathbb{1}\{q^{-1}_{x_{\mathrm{pa}}}(z) \in [0,1]\}}{\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z))} = c_1(z). \tag{10}$$

分子可由噪声服从均匀分布这一事实推出。项 $\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z))$ 非负，因为 $q_{x_{\mathrm{pa}}}$ 是递增的。此外，由于 $p_Z(z) > 0$，式 10 中的分子始终等于 1，因此分母不得依赖于 X_{pa}，

$$\frac{dq_{x_{\mathrm{pa}}}}{du}(q^{-1}_{x_{\mathrm{pa}}}(z)) = c_2(z)$$

对于某个函数 c_2。由引理1（将 a 替换为 x_{pa}），我们可以表示为

$$q_{x_{\mathrm{pa}}}(u) = q(u + r(x_{\mathrm{pa}})) \tag{11}$$

对于一个可逆函数 q。

接下来，由于 $Z \perp\!\!\!\perp X_{\mathrm{pa}}$， Z 的支撑集不依赖于 X_{pa}；等价地，对于所有 x_1， $x_2 \in \mathcal{X}_{\mathrm{pa}}$， q_{x_1} 和 q_{x_2} 的值域相等，

$$q_{x_1}([0,1]) = q_{x_2}([0,1]). \tag{12}$$

应用式 11 以及 q 的可逆性，

$$q([0,1] + r(x_1)) = q([0,1] + r(x_2))$$
$$[0,1] + r(x_1) = [0,1] + r(x_2)$$
$$[r(x_1), r(x_1) + 1] = [r(x_2), r(x_2) + 1]$$

由于这对所有 x_1， $x_2 \in \mathcal{X}_{\mathrm{pa}}$ 都成立，因此 $r(x_{\mathrm{pa}})$ 是一个常函数，或者 $r(x_{\mathrm{pa}}) \equiv r$。因此，令 $\tilde{q}$ 为 $\tilde{q}(u) = q(u + r) = q_{x_{\mathrm{pa}}}(u)$，它对于所有 x_{pa} 都只是 U 的函数。对于所有 x_{pa}，

$$g(f(x_{\mathrm{pa}}, U), x_{\mathrm{pa}})) = q_{x_{\mathrm{pa}}}(U) = \tilde{q}(U) \implies g(f(X_{\mathrm{pa}}, U), X_{\mathrm{pa}}) = \tilde{q}(U), \tag{13}$$

对于可逆函数 $\tilde{q}$。证明至此完成。 □

推论*1*。假设定理*1*的条件成立。此外，假设编码器-解码器模型对 (g, h) 满足：$h(g(X, X_{\mathrm{pa}}) X_{\mathrm{pa}}) = X$。考虑事实样本对 $x^{\mathrm{F}} := (x, x_{\mathrm{pa}})$，其中 $x := f(x_{\mathrm{pa}}, u)$，以及干预 $\mathrm{do}(X_{\mathrm{pa}} := \gamma)$。那么，由 $h(g(x, x_{\mathrm{pa}}) \gamma)$ 给出的反事实估计与真实反事实结果 $x^{\mathrm{CF}} := f(\gamma, u)$ 相匹配。

证明。对于干预 $\mathrm{do}(X_{\mathrm{pa}} := \gamma)$，真实的反事实结果为 $x^{\mathrm{CF}} := f(\gamma, u)$。根据假设，$h(g(x^{\mathrm{CF}}, x_{\mathrm{pa}}) x_{\mathrm{pa}}) = x^{\mathrm{CF}}$。现在，由于公式13对所有 X_{pa} 和 U 均成立，因此它也适用于事实样本和反事实样本。于是，

$$g(x, x_{\mathrm{pa}}) = g(f(x_{\mathrm{pa}}, u), x_{\mathrm{pa}}) = \tilde{q}(u) = g(f(\gamma, u), \gamma) = g(x^{\mathrm{CF}}, \gamma).$$

因此，编码器-解码器模型给出的反事实估计

$$h(g(x, x_{\mathrm{pa}}), \gamma) = h(g(x^{\mathrm{CF}}, \gamma), \gamma) = x^{\mathrm{CF}}.$$

证明至此完成。 □

推论 *2*。设 $\gamma \geq 0$。假设定理*1*的条件成立。此外，假设在某个指标 d（例如 $\|\cdot\|_2$）下，编码器-解码器模型对 (g, h) 的重构误差小于 τ：$d(h(g(X, X_{\mathrm{pa}}) X_{\mathrm{pa}}) X) \leq \tau$。考虑事实样本对 $x^{\mathrm{F}} := (x, x_{\mathrm{pa}})$，其中 $x := f(x_{\mathrm{pa}}, u)$，以及干预 $\mathit{do}(X_{\mathrm{pa}} := \gamma)$。那么，真实反事实结果 $x^{\mathrm{CF}} := f(\gamma, u)$ 与 $h(g(x, x_{\mathrm{pa}}) \gamma)$ 给出的反事实估计之间的误差至多为τ。换言之，$d(h(g(x, x_{\mathrm{pa}}) \gamma) x^{\mathrm{CF}}) \leq \tau$。

证明。对于干预 $\mathrm{do}(X_{\mathrm{pa}} := \gamma)$，反事实结果为 $x^{\mathrm{CF}} := f(\gamma, u)$。由于公式13对所有 X_{pa} 和 U均成立，因此它对事实样本和反事实样本也成立。于是我们有，

$$g(x, x_{\mathrm{pa}}) = g(f(x_{\mathrm{pa}}, u), x_{\mathrm{pa}}) = \tilde{q}(u) = g(f(\gamma, u), \gamma) = g(x^{\mathrm{CF}}, \gamma). \tag{14}$$

根据关于编码器-解码器重构误差的假设以及公式14，

$$d(h(g(x^{\mathrm{CF}}, \gamma), \gamma), x^{\mathrm{CF}}) \leq \tau \tag{15}$$

$$d(h(g(x, x_{\mathrm{pa}}), \gamma), x^{\mathrm{CF}}) \leq \tau. \tag{16}$$

□

现在我们讨论一个引理，它将引理1扩展到了高维情形。

引理 2。对于 $\mathcal{U}, \mathcal{Z} \subset \mathbb{R}^m$，考虑一族可逆函数 $q_{x_{\mathrm{pa}}}$：$\mathcal{U} \to \mathcal{Z}$，其中 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$，若对所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$ 均有 $Jq_{x_{\mathrm{pa}}}(q_{x_{\mathrm{pa}}}^{-1}(z)) = cA$，则 $q_{x_{\mathrm{pa}}}$ 可表示为

$$q_{x_{\mathrm{pa}}}(u) = q(u + r(x_{\mathrm{pa}}))$$

对于某个函数 r 和可逆的 q。

证明。假设 $Jq_{x_{\mathrm{pa}}}(q_{x_{\mathrm{pa}}}^{-1}(z)) = cA$。根据反函数定理，

$$Jq_{x_{\mathrm{pa}}}^{-1}|_z = \left(Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)}\right)^{-1}. \tag{17}$$

定义 $s_{x_{\mathrm{pa}}}$：$\mathcal{Z} \to \mathcal{U}$ 是 $q_{x_{\mathrm{pa}}}$ 的逆。根据假设

$$\det Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = \det Jq_{x_{\mathrm{pa}}}^{-1}|_z = \det Js_{x_{\mathrm{pa}}}|_z = cA$$

对所有 x_{pa} 以及常数 c 和正交矩阵 A均成立。由于 $s_{x_{\mathrm{pa}}}$ 的雅可比矩阵是一个缩放后的正交矩阵，因此 $s_{x_{\mathrm{pa}}}$ 是保角函数。因此，根据刘维尔定理，$s_{x_{\mathrm{pa}}}$ 是一个莫比乌斯函数（Blair，2000），这意味着

$$s_{x_{\mathrm{pa}}}(z) = b_{x_{\mathrm{pa}}} + \alpha_{x_{\mathrm{pa}}} A_{x_{\mathrm{pa}}}(z - a_{x_{\mathrm{pa}}})/\|z - a_{x_{\mathrm{pa}}}\|^{\varepsilon}, \tag{18}$$

其中 $A_{x_{\mathrm{pa}}}$ 是一个正交矩阵， $\varepsilon \in \{0, 2\}$、 $a_{x_{\mathrm{pa}}} \in \mathbb{R}^m$和 $\alpha_{x_{\mathrm{pa}}} \in \mathbb{R}$。根据假设，$s_{x_{\mathrm{pa}}}$ 的雅可比矩阵等于 cA

$$Js_{x_{\mathrm{pa}}}\big|_z = \frac{\alpha_{x_{\mathrm{pa}}} A_{x_{\mathrm{pa}}}}{\|z - a_{x_{\mathrm{pa}}}\|^{\varepsilon}} \left(I - \varepsilon \frac{(z - a_{x_{\mathrm{pa}}})(z - a_{x_{\mathrm{pa}}})^T}{\|z - a_{x_{\mathrm{pa}}}\|^2} \right) = cA.$$

这对变量 α、 a 和 ε 施加了约束。选择 z，使得 $z - a_{x_{\mathrm{pa}}} = kv$ 对于单位向量 v 成立，并乘以 $A_{x_{\mathrm{pa}}}^{-1}$,

$$\begin{aligned} cAA_{x_{\mathrm{pa}}}^{-1} &= \frac{\alpha_{x_{\mathrm{pa}}}}{\|kv\|^{\varepsilon}} \left(I - \varepsilon \frac{k^2 vv^T}{k^2 \|v\|^2} \right) \\ I &= \varepsilon vv^T + \left(\frac{ck^{\varepsilon}}{\alpha_{x_{\mathrm{pa}}}} \right) AA_{x_{\mathrm{pa}}}^{-1}. \end{aligned}$$

如果 $\varepsilon = 2$，选择 k 的不同取值会导致 z 取不同的值，从而使右侧的取值发生变化，而右侧应为常数单位矩阵。因此，我们必须有 $\varepsilon = 0$。这也意味着 $\alpha_{x_{\mathrm{pa}}} = c$ 和 $A = A_{x_{\mathrm{pa}}}$。由此可得进一步的参数化

$$s_{x_{\mathrm{pa}}}(z) = b_{x_{\mathrm{pa}}} - cA(z - a_{x_{\mathrm{pa}}}) = b'(x_{\mathrm{pa}}) + cAz$$

其中 $b'(x_{\mathrm{pa}}) = b_{x_{\mathrm{pa}}} - cAa_{x_{\mathrm{pa}}}$。

不失一般性，我们可以考虑任意一个固定的 $x_{\mathrm{pa}_0} \in \mathcal{X}_{\mathrm{pa}}$,

$$s_{x_{\mathrm{pa}_0}}(z) - s_{x_{\mathrm{pa}}}(z) = r(x_{\mathrm{pa}}) \coloneqq b'(x_{\mathrm{pa}_0}) - b'(x_{\mathrm{pa}}).$$

令 $u = s_{x_{\mathrm{pa}}}(z)$。则有

$$\begin{aligned} s_{x_{\mathrm{pa}_0}}(z) &= u + r(x_{\mathrm{pa}}) \\ q_{x_{\mathrm{pa}}}(u) = z &= q_{x_{\mathrm{pa}_0}}(u + r(x_{\mathrm{pa}})), \end{aligned}$$

选择 $q = q_{x_{\mathrm{pa}_0}}$ 即可得到所需的表示。 □

定理 2. 假设对于 $X \in \mathcal{X} \subset \mathbb{R}^m$ 和连续外生噪声 $U \sim \mathrm{Unif}[0, 1]^m$，对于 $m \geq 3$，且 X 满足结构方程

$$X = f(X_{\mathrm{pa}}, U) \tag{6}$$

其中 $X_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}} \subset \mathbb{R}^d$ 是节点 X 的父节点，且 $U \perp\!\!\!\perp X_{\mathrm{pa}}$。考虑一个编码器-解码器模型，其编码函数为 $g\colon \mathcal{X} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{Z}$，解码函数为 $h\colon \mathcal{Z} \times \mathcal{X}_{\mathrm{pa}} \to \mathcal{X}$,

$$Z \coloneqq g(X, X_{\mathrm{pa}}), \quad \hat{X} \coloneqq h(Z, X_{\mathrm{pa}}). \tag{7}$$

假设满足以下条件：

1. 编码独立于父节点值， $g(X, X_{\mathrm{pa}}) \perp\!\!\!\perp X_{\mathrm{pa}}$.

2. 结构方程 f 关于 U 可逆且可微，并且对于所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$， $Jf_{x_{\mathrm{pa}}}$ 均为正定。

3. 编码器 编码 g 相对于 X可逆且可微，并且对于所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$， $Jg_{x_{\mathrm{pa}}}$ 均为正定。 .

4. 编码 $q_{x_{\mathrm{pa}}}(U) \coloneqq g(f(x_{\mathrm{pa}}, U)\ x_{\mathrm{pa}})$满足 $Jq_{x_{\mathrm{pa}}}\big|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = c(x_{\mathrm{pa}})A$，对所有 $z \in \mathcal{Z}$和 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$均成立，其中 c 是一个标量函数， A 是一个正交矩阵。

那么，对于可逆函数 $\tilde{q}$， $g(f(X_{\mathrm{pa}}, U)\ X_{\mathrm{pa}}) = \tilde{q}(U)$。

证明。 我们可以证明， $g(X, X_{\mathrm{pa}}) = g(f(X_{\mathrm{pa}}, U)\ X_{\mathrm{pa}})$仅是 U的函数。

根据函数复合的性质，$q_{x_{\mathrm{pa}}}(U) \coloneqq g(f(x_{\mathrm{pa}}, U)\ x_{\mathrm{pa}})$也可逆且可微。由于$Jf_{x_{\mathrm{pa}}}$ 和 $Jg_{x_{\mathrm{pa}}}$ 均为正定，且 $Jq_{x_{\mathrm{pa}}} = Jf_{x_{\mathrm{pa}}} Jg_{x_{\mathrm{pa}}}$, 因此对于所有 $x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$， $Jq_{x_{\mathrm{pa}}}$ 也为正定。

根据假设，编码 Z 与 X_{pa}相互独立,

$$Z = q_{X_{\mathrm{pa}}}(U) \perp\!\!\!\perp X_{\mathrm{pa}}. \tag{19}$$

因此， Z 的条件分布不依赖于 X_{pa}。利用假设 $U \perp\!\!\!\perp X_{\mathrm{pa}}$，对于 Z支撑集中的所有$x_{\mathrm{pa}} \in \mathcal{X}_{\mathrm{pa}}$ 和 z，根据密度变换公式，

$$p_Z(z) = \frac{p_U(q_{x_{\mathrm{pa}}}^{-1}(z))}{\left|\det Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)}\right|} = \frac{2^{-m}\mathbb{1}\{q_{x_{\mathrm{pa}}}^{-1}(z) \in [0,1]^m\}}{\det Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)}} = c_1(z). \tag{20}$$

分子可由噪声服从均匀分布这一事实推出。雅可比矩阵项$Jq_{x_{\mathrm{pa}}}(q_{x_{\mathrm{pa}}}^{-1}(z))$的行列式非负，因为 $Jq_{x_{\mathrm{pa}}}$ 是正定的。此外，由于 $p_Z(z) > 0$，式 20 中的分子始终等于 2^{-m}，而分母不能依赖于 X_{pa}，

$$\det Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = c_2(z) \tag{21}$$

对于某个函数 c_2。根据我们的假设，对于所有 z，$Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = c(x_{\mathrm{pa}})A$，其中 A 是一个正交矩阵。将其应用于式 21，

$$\det Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = \det c(x_{\mathrm{pa}})A = c(x_{\mathrm{pa}}) = c_2(z),$$

这意味着 $c(x_{\mathrm{pa}}) \equiv c$ 是一个常函数，或者$Jq_{x_{\mathrm{pa}}}|_{q_{x_{\mathrm{pa}}}^{-1}(z)} = cA$。根据引理 2，我们可以将$q_{x_{\mathrm{pa}}}(u)$表示为

$$q_{x_{\mathrm{pa}}}(u) = q(u + r(x_{\mathrm{pa}})) \tag{22}$$

对于一个可逆函数 q。

接下来，由于 $Z \perp\!\!\!\perp X_{\mathrm{pa}}$，$Z$ 的支撑集不依赖于 X_{pa}；等价地，对于所有 $x_1, x_2 \in \mathcal{X}_{\mathrm{pa}}$，$q_{x_{\mathrm{pa}_1}}$ 和$q_{x_{\mathrm{pa}_2}}$ 的值域相等，

$$q_{x_1}([0,1]^m) = q_{x_2}([0,1]^m). \tag{23}$$

应用式 22 以及 q的可逆性，

$$\begin{aligned} q([0,1]^m + r(x_1)) &= q([0,1]^m + r(x_2)) \\ [0,1]^m + r(x_1) &= [0,1]^m + r(x_2) \\ [r(x_1), r(x_1)+1]^m &= [r(x_2), r(x_2)+1]^m. \end{aligned}$$

由于这对所有 $x_1, x_2 \in \mathcal{X}_{\mathrm{pa}}$ 都成立，因此我们有 $r(x)$ 是常数，或者 $r(x) \equiv r$。因此，令 $\tilde{q}$ 为 $\tilde{q}(u) = q(u + r) = q_{x_{\mathrm{pa}}}(u)$，它对所有 x_{pa} 都仅是 U 的函数。对于所有 x_{pa}，

$$g(f(x_{\mathrm{pa}}, U, x_{\mathrm{pa}})) = q_{x_{\mathrm{pa}}}(U) = \tilde{q}(U) \implies g(f(X_{\mathrm{pa}}, U), X_{\mathrm{pa}}) = \tilde{q}(U). \tag{24}$$

证明至此完成。 □

推论 3。假设定理 *2* 的条件成立。此外，假设编码器-解码器模型对 (g, h)满足

$$h(g(X, X_{\mathrm{pa}}), X_{\mathrm{pa}}) = X. \tag{25}$$

考虑事实样本对 (x, x_{pa})，其中 $x := f(x_{\mathrm{pa}}, u)$，以及干预 $\mathrm{do}(X_{\mathrm{pa}} := \gamma)$。那么，由$h(g(x, x_{\mathrm{pa}})\ \gamma)$ 给出的结果与真实反事实结果 $x^{\mathrm{CF}} := f(\gamma, u)$相匹配。

推论 4。令 $\gamma \geq 0$。假设定理 *2* 的条件成立。此外，假设编码器-解码器模型对(g, h) 在某个指标 d（例如 $\|\cdot\|_2$）下的重构误差小于 τ，

$$d(h(g(X, X_{\mathrm{pa}}), X_{\mathrm{pa}}), X) \leq \tau. \tag{26}$$

考虑一个事实样本对（x，x_{pa}），其中 $x := f(x_{\mathrm{pa}}, u)$，以及一个干预 do（$X_{\mathrm{pa}} := \gamma$）。那么，真实反事实 $x^{\mathrm{CF}} := f(\gamma, u)$ 与反事实估计 $h(g(x, x_{\mathrm{pa}})\ \gamma)$ 之间的误差至多为 τ，即 $d(h(g(x, x_{\mathrm{pa}})\ \gamma)\ x^{\mathrm{CF}}) \leq \tau$。

B 检验父节点与编码之间的独立性

我们通过实证评估编码与父节点值之间的依赖关系。我们考虑一个二元非线性结构因果模型 $X_1 \to X_2$，其中 $X_2 = f(X_1, U_2) = X_1^2 + U_2$，并且 X_1 和 U_2 独立地从标准正态分布中采样。我们评估 X_1 与 X_2 的编码之间的 HSIC。我们在 $n = 5000$ 个样本上拟合模型，并使用来自同一分布的 1000 个测试样本评估 HSIC 分数。我们使用基于核的独立性检验计算 p 值，并将我们的性能与 ANM 进行比较；在这一设置下，ANM 是一个正确设定的模型（Gretton 等人，2007）。我们重复该试验 100 次。在真实独立的情况下，我们预期 p 值服从均匀分布。在下表中，我们展示了 100 次试验所得 p 值的一些汇总统计量，最后一行表示真实均匀 p 值下的期望值（零假设成立时即会出现这种情况）。

	Mean	标准差	10% 分位数	90% 分位数	Min	Max
DCM	0.196	0.207	0.004	0.515	6e-6	0.947
ANM	0.419	0.255	0.092	0.774	3e-5	0.894
真实均匀分布（零假设）	0.500	0.288	0.100	0.900	1e-2	0.990

表 3：p 值描述表。

我们给出了设定正确的 ANM 方法所得到的 p 值，它们接近均匀分布，表明编码可以接近独立。尽管我们的 DCM 方法产生的 p 值并非完全均匀，但这些编码并不会一致地拒绝独立性原假设。这些结果表明，在经验上可以获得独立于父变量的编码。我们还注意到，在这种情况下，ANM 的设定是正确的，而 DCM 尽管通用得多，表现仍具有相当的竞争力。

C 缺失的实验细节

本节补充第5节中缺失的细节。

C.1 模型超参数

在所有评估实验中，我们将模型超参数保持不变。对于 DCM，我们使用 $T = 100$ 个总时间步，并采用线性回归 β_t 调度，在 1e-4 和 0.1 之间进行插值；或者对于 $t \in [T]$，使用 $\beta_t = \left(0.1 - 10^{-4}\right)\frac{t-1}{T-1} + 10^{-4}$。为子纳入父节点的值和时间步 t，我们只需将父节点值和 t/T 拼接起来，作为 ε_θ 模型的输入。我们发现，采用常用的余弦调度（Nichol & Dhariwal，2021年）会导致经验表现较差，使用时间 t的位置编码也会如此。我们认为，位置编码导致性能下降，是因为问题的维度较低，位置编码器的维度会主导其他输入的维度。

我们还评估了使用无分类器引导（CFG）（Ho & Salimans，2021年）来增强对父节点值的依赖，但发现这也会降低性能。我们通过定理1给出了一个合理的解释。根据定理1，我们希望编码 $g(Y, X)$ 与 X 独立；然而，使用 CFG 编码 $(1 + w)g(Y, X) - wg(Y, 0)$ 只会增强 $g(Y, X)$ 对 X 的依赖，这与我们的目标背道而驰。

对于 VACA，我们采用默认实现[8], , 训练 500 个训练轮数，学习率为 0.005，编码器和解码器的隐藏维度大小分别为 $[8, 8]$ 和 $[16]$，潜向量维度为 4，父节点丢弃率为 0.2。

对于 CAREFL，我们同样使用默认实现[9] 和神经样条自回归流（Durkan 等人，2019年），训练 500 轮，学习率为 0.005，使用四个流和十个隐藏单元。

[8]https://github.com/psanch21/VACA[9]
https://github.com/piomonti/carefl/

对于 ANM，我们同样使用默认实现来选择回归模型。给定一组拟合后的回归模型，ANM 会选择在数据划分上平均均方根误差最低的模型。ANM 考虑以下回归模型：线性回归、岭回归、LASSO、弹性网络、随机森林、直方图梯度提升、支持向量机、极端随机树、k近邻和 AdaBoost。

C.2 加性噪声模型（ANM）的详细信息

对于给定的节点 X_i 及其父节点 X_{pa_i}，考虑拟合一个回归模型 f_i，其中 $f_i(X_{\mathrm{pa}_i}) \approx X_i$。利用该回归模型和训练数据集，便足以生成来自观测分布和干预分布的样本，并计算反事实。

观测/干预样本。样本按照拓扑序构造。对于干预节点，采样值始终为干预值。SCM 中未受干预的根节点从训练集的经验分布中采样。通过归纳采样父节点值 X_{pa_i}，并从经验残差分布中采样 $\hat{U}$，即可生成 X_i 的新样本，输出 $\hat{X}_i := \hat{f}_i(X_{\mathrm{pa}_i}) + \hat{U}$。

反事实估计。对于事实观测 x^{F} 以及对节点 $\mathcal{I}$ 施加值为 γ 的干预，反事实估计与事实估计仅在所有受到干预或位于干预节点下游的节点上有所不同。我们按照拓扑序进行。对于每个干预节点 i，$\hat{x}_i^{\mathrm{CF}} := \gamma_i$。对于每个下游存在干预节点的未干预节点 i，定义 $\hat{u}_i^{\mathrm{F}} := x_i^{\mathrm{F}} - \hat{f}_i(x_{\mathrm{pa}_i}^{\mathrm{F}})$，即事实样本的残差和估计噪声。令 $\hat{x}_{\mathrm{pa}_i}^{\mathrm{CF}}$ 为 X_i 的父节点的反事实估计。则 $\hat{x}_i^{\mathrm{CF}} := \hat{f}_i(\hat{x}_{\mathrm{pa}_i}^{\mathrm{CF}}) + \hat{u}_i^{\mathrm{F}}$。

因此，对于反事实查询，如果真实的函数方程 f_i 是加性噪声模型，那么如果 $f_i \approx f_i$，回归模型的反事实误差就会很低。事实上，如果 $f_i \equiv f_i$，那么回归模型的反事实表现将达到完美。

C.3 随机图生成的详细信息

该随机图由十个节点组成。我们通过生成一个随机上三角邻接矩阵来对该图进行随机采样，其中上三角部分的每个元素以 30% 的概率取值为 1。随后，我们检查该图是否由单个连通分量构成（如果不是，则重新采样）。图 4 中给出了一个图示示例。

C.4 合成数据实验的查询评估框架

观测评估。我们从拟合的观测分布和真实的观测分布中各生成 1000 个样本，并报告二者之间的最大均值差异。由于 DCM 和 ANM 对根节点使用经验分布，因此我们仅计算非根节点之间的最大均值差异。

干预评估。我们考虑对单个节点进行干预。对于干预节点 i，我们选择 20 个干预值 $\gamma_1, \ldots, \gamma_{20}$，在节点 i 边缘分布的 10% 和 90% 分位数之间进行线性插值，以表示真实的干预。随后，对于每个干预 $\mathrm{do}(X_i := \gamma_j)$，我们从拟合模型和真实因果模型中分别生成 100 个值，X 和 $X^\star$ 分别表示来自拟合模型和真实模型的样本。由于干预只会影响节点 i的后代节点，我们对 X 和 $X^\star$ 取子集，仅保留节点 i的后代节点，并在 X 和 $X^\star$ 上计算最大均值差异（MMD），得到距离 $\delta_{i,\,j}$，用于衡量特定节点和干预值对应的干预分布之间的差异。最后，我们报告所有 20 个干预值和所有干预节点上的平均 MMD。对于梯形图，我们选择对 X_2 和 X_3 进行干预，因为它们是最远的上游节点，能够体现干预的最大难度。对于随机图，我们随机选择三个非汇节点进行干预。干预评估框架的正式描述见算法 6。

反事实评估。与干预评估类似，我们考虑对单个节点进行干预。对于节点 i，我们选择 20 个干预值 γ_1，...，γ_{20}，在节点 i 的边缘分布的 10% 分位数和 90% 分位数之间进行线性插值，以表示现实的干预。然后，对于每个干预 $\mathrm{do}(X_i := \gamma_j)$，我们生成 100 个未干预的事实样本 x^{F}，并分别查询估计的反事实值和真实反事实值 $\hat{x}^{\mathrm{CF}}$ 和 x^{CF}。与之前类似，$\hat{x}^{\mathrm{CF}}$ 和 x^{CF} 仅在节点 i 的后代节点上有所不同，因此我们只考虑节点 i 的后代节点子集。我们计算均方误差 $\delta_{i,\,j}$，由于反事实估计和真实值都是点值，因此得出特定节点和干预值对应的误差。最后，我们报告 the

算法 6干预查询的评估

1: **对于**每个干预节点 i **执行**2: γ_1，...， γ_{20} 对节点 i 的 10% 至 90% 分位数进行线性插值3: **对于**上述生成的每个干预 γ_j **执行**4: 执行 do($X_i := \gamma_j$) ^ ^ 5: 从 M 生成 X 个样本，并从 $M^\star$ 生成 $X^\star$ 个样本，这些样本均来自 i 的后代节点 ^ ^ 6: $\delta_{i,\ j} \leftarrow$ 最大均值差异(X， $X^\star$) 7: **结束循环**8: **结束循环**9: 输出所有δ_i的均值$_{,\ j}$

算法 7反事实查询的评估

1: **对**每个干预节点 i **执行**2: $\gamma_1, \ldots, \gamma_{20}$ 对节点 i的 10% 至 90% 分位数进行线性插值3: **对**上面生成的每个干预 γ_j **执行**4: 生成 100 个事实样本 $x_1^{\mathrm{F}}, \ldots, x_{100}^{\mathrm{F}}$5: 执行干预 do($X_i := \gamma_j$) 6: 使用 $\{x^{\mathrm{F}}\}_{k=1}^{100}$，计算 i所有后代节点的反事实估计 $\{\hat{} x_k^{\mathrm{CF}}\}_{k=1}^{100}$ 和真实反事实 $\{x_k^{\mathrm{CF}}\}_{k=1}^{100}$ 7: $\delta_{i,j} \leftarrow$ 均方误差($\{\hat{} x_k^{\mathrm{CF}}\}_{k=1}^{100}, \{x_k^{\mathrm{CF}}\}_{k=1}^{100}$) 8: **结束循环**9: **结束循环**10: 输出所有$\delta_{i,j}$的均值

对所有 20 个干预值和所有干预节点的均方误差取均值。我们使用与上文干预评估中相同的干预节点。反事实评估框架的正式描述见算法 7（附录 C）。

C.5 功能性磁共振成像实验中误差膨胀的解释

在我们的功能性磁共振成像实验中，我们针对单个干预样本计算绝对误差，并取其绝对值。为了直观理解为何我们不能指望误差接近于零，以及为何这些误差彼此相对更接近，请看下面这个玩具问题。

假设 X_1，.....， X_n， Y_1，.....， Y_n 独立同分布$\sim$ $\mathcal{N}(\theta,\ 1)$，并观测到数据 X_1，.....， X_n。考虑以下两个统计问题：

^ 1. 学习一个分布 D，使得从 D 中抽取的样本与样本 Y_1，.....， Y_n 之间的最大均值差异较小。

2. 学习一个在平方误差意义下能够良好估计 Y_1 的估计器。

^ 对于第一个统计问题，如果 D 是一个合理的估计器，我们应当预期更多数据会带来更低的最大均值差异，例如误差可能以 $1/n$ 的速率衰减。我们应当预期最大均值差异值接近于零，而且其表现的大小可以直接解释。

在第二个统计问题中，在平方误差准则下，该问题等价于

$$\min_c \mathbb{E}_{Y_1 \sim \mathcal{N}(\theta,1)} (Y_1 - c)^2 = \min_c \theta^2 + 1 - 2c\theta + c^2.$$

最优估计量是均值 θ，其平方误差为 1。当然，在实践中我们并不知道 θ；如果使用 X_1，.....， X_n的样本均值，误差的量级将为 $1 + 1/n$。虽然我们仍然可以比较各种估计量，例如样本均值、样本中位数和深度神经网络，但所有损失都会增加 1，导致性能差异看起来微乎其微。

我们的合成实验旨在直接估计干预分布，并计算真实分布与模型分布的样本之间的最大均值差异，这表明它们属于第一类统计问题。fMRI 真实数据实验则旨在从一个分布中估计单个干预值，这表明它们属于第二类统计问题。

统计问题。我们不应期待这些结果非常小，所有指标值都应加上一个内在的不可约误差。

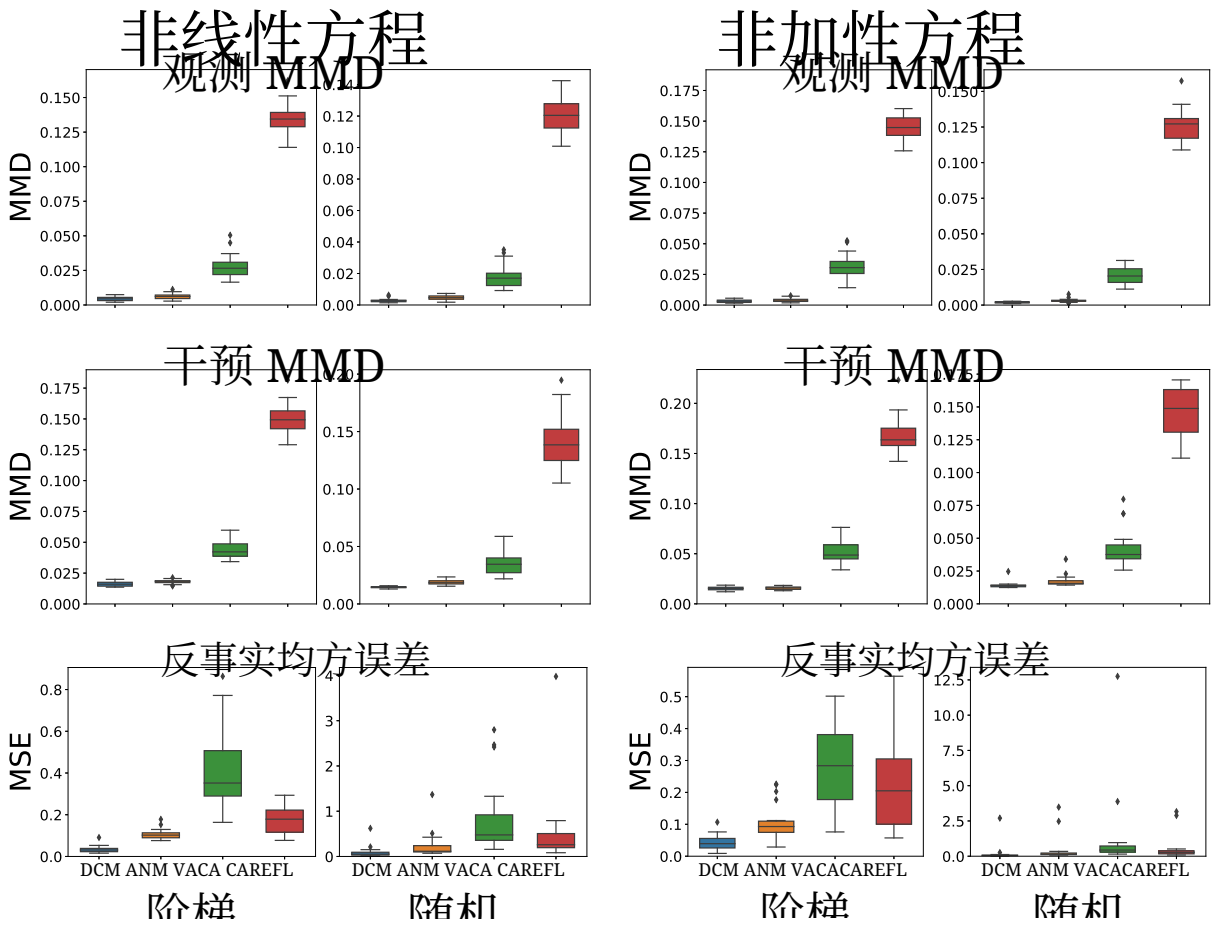

图1：左：非线性设置（NLIN），右：非加性设置（NADD）。箱线图展示了在模型和训练数据的 20 次随机初始化下，对阶梯和随机结构因果模型进行的观察、干预和反事实查询结果。

D 其他实验

D.1 运行时间实验

我们报告了使用默认实现和参数，在梯形图上以一个随机种子进行训练所需的时间（分钟）。为确保公平比较，所有实验均在 CPU 上评估。请注意，ANM 速度最快，因为它使用标准

	DCM	ANM	VACA	CAREFL
训练时间（分钟）	15.3	4.3	142.8	110.5

表 4：单次运行的训练时间表。

回归模型，而我们提出的 DCM 方法比 CAREFL 和 VACA 快约 7–9 倍。所有方法的生成（推理）时间都在 1 秒左右。对于 VACA 和 CAREFL，我们使用各自作者提供的实现。

D.2 半合成实验

为了进一步评估 DCM 的有效性，我们基于 Sachs 数据集（Sachs 等人，2005年）开展了一项半合成实验。我们使用了包含 11 个节点的真实世界图[10]。该图呈现了人类 T 细胞内错综复杂的信号通路网络。图中的 11 个节点分别对应其研究中考察的磷酸化蛋白质或磷脂。

在我们的实验中，我们以半合成的方式对数据进行采样。对于根节点，我们从经验边缘分布中采样。对于非根节点，由于真实结构方程未知，我们采用随机神经网络作为结构方程，这与第5节中的做法相同。实验结果见表5。可以看到，我们的DCM方法带来的性能提升印证了此前的发现。

			DCM	ANM	VACA	CAREFL
SCM		指标	($\times 10^{-2}$)	($\times 10^{-2}$)	($\times 10^{-2}$)	($\times 10^{-2}$)
Sachs	NLIN	观测 MMD	0.31±0.15	**0.21±0.04**	0.53±0.24	7.30±0.95
		积分最大均值差异	**1.25±0.11**	1.37±0.12	2.03±0.36	5.77±0.85
		反事实均方误差	**0.72±0.19**	4.72±1.71	17.71±9.23	9.59±2.52
	NADD	观测 MMD	**0.18±0.07**	0.18±0.06	0.39±0.24	6.10±1.14
		积分最大均值差异	**1.42±0.26**	1.86±0.51	2.21±0.98	5.29±2.07
		反事实均方误差	**1.99±2.49**	4.68±5.77	7.30±11.57	8.30±8.34

表5：在非线性（NLIN）和非加性（NADD）设置下，对 Sachs 结构因果模型进行观测、干预和反事实查询所得结果的均值±标准差。模型和训练数据采用 10 次随机初始化。为便于呈现，数值均乘以 100。

D.3 额外合成实验

本节展示了更多实验 证据，说明 DCM 在处理各种类型的因果查询方面表现优异。我们还考察了四种较小的图结构，分别称为链、三角形、*Diamond*和*Y*形（见图9）。

具体方程列于表6。这些函数方程经过选择，以平衡协变量与噪声的信噪比，从而体现现实场景。此外，这些结构方程是在进行超参数选择之后选定的；这意味着我们没有根据模型表现来调整 DCM 的参数，也没有据此调整结构方程。

考虑一个取值为 X 、父节点为 X_{pa} 且外生噪声为 U 的节点，其中 $X_{\mathrm{pa}} \perp\!\!\!\perp U$，其对应的函数方程为 f，使得

$$X = f(X_{\mathrm{pa}}, U).$$

进一步假设 加性噪声模型， $X := f_1(X_{\mathrm{pa}}) + f_2(U)$。在这种加性设定下，由于 $X_{\mathrm{pa}} \perp\!\!\!\perp U$，我们有 e

$$\mathrm{Var}\,[X] = \mathrm{Var}\,[f_1(X_{\mathrm{pa}}] + \mathrm{Var}\,[f_2(U)].$$

我们选择 f_1 和 f_2，使得

$$0.05 \leq \frac{\mathrm{Var}\,[f_2(U)]}{\mathrm{Var}\,[f_1(X_{\mathrm{pa}}]} \leq 0.5,$$

这表示噪声相对于父节点的影响之比大致相当，或小一个数量级。

对于非加性情形，我们使用全方差定律分解方差，

$$\mathrm{Var}\,[X] = \mathbb{E}\,\mathrm{Var}\,[f(X_{\mathrm{pa}}, U) \mid U] + \mathrm{Var}\,\mathbb{E}[f(X_{\mathrm{pa}}, U) \mid U].$$

[10]https://www.bnl earn.com/bnrepository/discrete-small.htm l#sachs

SCM		非线性情形	非加性情形
Chain	$f_2(U_2, X_1)$	$\exp(X_1/2) + U_2/4$	$1/((U_2 + X_1)^2 + 0.5)$
	$f_3(U_3, X_2)$	$(X_2 - 5)^3/15 + U_3$	$\sqrt{X_2 + \lvert U_3\rvert}/(0.1 + X_2)$
Triangle	$f_2(U_2, X_1)$	$2X_1^2 + U_2$	$X_1/((U_2 + X_1)^2 + 1) + U_2/4$
	$f_3(U_3, X_1, X_2)$	$20/(1 + \exp(-X_2^2 + X_1)) + U_3$	$(\lvert U_3\rvert + 0.3)(-X_1 + X_2/2 + \lvert U_3\rvert/5)^2$
Diamond	$f_2(U_2, X_1)$	$X_1^2 + U_2/2$	$\sqrt{\lvert X_1\rvert}(\lvert U_2\rvert + 0.1)/2 + \lvert X_1\rvert + U_2/5$
	$f_3(U_3, X_1, X_2)$	$X_2^2 - 2/(1 + \exp(-X_1)) + U_3/2$	$1/(1 + (\lvert U_3\rvert + 0.5)\exp(-X_2 + X_1))$
	$f_4(U_4, X_2, X_3)$	$X_3/(\lvert X_2 + 2\rvert + X_3 + 0.5) + U_4/10$	$(X_3 + X_2 + U_4/4 - 7)^2 - 20$
Y	$f_3(U_3, X_1, X_2)$	$4/(1 + \exp(-X_1 - X_2)) - X_2^2 + U_3/2$	$(X_1 - 2X_2 - 2)(\lvert U_3\rvert + 0.2)$
	$f_4(U_4, X_3)$	$20/(1 + \exp(X_3^2/2 - X_3)) + U_4$	$(\cos(X_3) + U_4/2)^2$

表6：定义 NLIN 和 NADD 情况下数据生成过程的方程。

类似地，我们选择函数方程 f，使得 f 满足

$$0.05 \leq \frac{\operatorname{Var}\mathbb{E}[f(X_{\mathrm{pa}}, U) \mid U]}{\mathbb{E}\operatorname{Var}[f(X_{\mathrm{pa}}, U) \mid U]} \leq 0.5$$

对于所有图，U_i 均独立同分布$\sim \mathcal{N}(0, 1)$，并且我们选择 f，使得当 X_i 是根节点时，$X_i = U_i$，即 f 是恒等函数。最后，我们对每个节点 X_i 进行归一化，使得 $\mathrm{Var}(X_i) \approx 1$。为清晰起见，我们省略公式中的所有归一化项，并省略下文根节点的函数方程。

结果。对于这 4 种图结构（链、三角形、*Diamond* 和 *Y*形），表 7 展示了所有评估模型在观测、干预和反事实查询上的表现；这些结果是对模型和训练数据分别进行 10 次独立初始化后取平均所得，每行中的最低值以粗体标出。为便于查看，数值均乘以 100。图 2 展示了同一组实验的箱线图。

此处的结果与表 1 中较大图结构下观察到的结果相似。在 12 种非线性设置中，DCM 有 7 种设置的误差最低，其余 5 种则是正确指定的 ANM 误差最低。此外，与其他竞争方法相比，DCM 和 ANM 的标准差通常都较低。对于非加性设定，DCM 在全部 12 个因果查询中均表现出最低值。

E RealData ExperimentII

为了评估我们的DCM方法，我们评估其计算个体处理效应（ITE）的性能。个体处理效应定义为 $D_i := Y_i(1) - Y_i(0)$，其中 $Y_i(0) \in R$ 是单位 i 在被分配到对照组时的潜在结果，而 $Y_i(1)$ 是被分配到处理组时的潜在结果。请注意，我们无法

SCM		指标	DCM ($\times 10^{-2}$)	ANM ($\times 10^{-2}$)	VACA ($\times 10^{-2}$)	CAREFL ($\times 10^{-2}$)
Chain	NLIN	观测 MMD	0.27±0.11	**0.19±0.27**	1.63±0.42	4.25±1.12
		积分最大均值差异	1.71±0.27	**1.70±0.47**	10.10±1.81	8.63±0.52
		反事实均方误差	**0.33±0.16**	2.43±2.49	25.99±6.47	19.62±4.01
	NADD	观测 MMD	**0.22±0.16**	1.51±0.43	1.53±0.41	3.48±1.06
		积分最大均值差异	**2.85±0.45**	7.34±0.62	10.42±2.77	11.02±1.32
		反事实均方误差	**75.84±1.65**	88.56±2.63	98.82±4.16	105.80±11.04
Triangle	NLIN	观测 MMD	0.16±0.11	**0.12±0.07**	3.12±0.93	4.64±1.03
		积分最大均值差异	**1.50±0.30**	3.28±0.79	18.43±1.72	7.08±0.82
		反事实均方误差	**1.12±0.26**	9.80±1.70	178.69±16.45	41.85±19.58
	NADD	观测 MMD	**0.25±0.12**	0.51±0.08	2.42±0.48	5.12±1.10
		积分最大均值差异	**2.81±0.21**	5.54±0.60	11.09±0.85	4.17±0.44
		反事实均方误差	**26.28±6.68**	97.25±16.45	173.67±16.28	121.99±31.34
Y	NLIN	观测 MMD	**0.11±0.05**	0.14±0.08	2.29±0.69	6.82±0.85
		积分最大均值差异	**1.23±0.08**	1.40±0.14	9.50±0.96	14.97±1.29
		反事实均方误差	**0.28±0.25**	1.22±0.27	28.79±4.02	27.85±5.33
	NADD	观测 MMD	**0.21±0.15**	1.00±0.19	1.51±0.44	3.39±0.58
		积分最大均值差异	**1.54±0.23**	5.62±0.31	5.37±0.72	6.51±0.40
		反事实均方误差	**33.45±2.06**	47.55±2.56	60.67±3.24	52.41±6.69
Diamond	NLIN	观测 MMD	0.22±0.10	**0.13±0.07**	2.77±0.45	8.28±1.29
		积分最大均值差异	3.21±0.62	**2.56±0.31**	25.30±1.39	18.23±3.01
		反事实均方误差	**14.74±6.09**	32.02±37.74	138.65±12.33	607.62±241.21
	NADD	观测 MMD	**0.25±0.18**	0.28±0.09	2.36±0.44	5.50±0.81
		积分最大均值差异	**1.88±0.23**	4.40±0.52	12.54±0.89	16.34±1.72
		反事实均方误差	**1.36±0.14**	8.58±0.77	57.40±4.23	24.61±7.05

表 7：在非线性和非加性设置下，对四种不同结构因果模型的观测、干预和反事实查询结果取均值±标准差。模型和训练数据均采用 10 次随机初始化。为便于阅读，数值均乘以 100。

观测到任何单位的 D_i i，因为对于训练数据集中的每个单位，我们只能观测到其在对照条件下或处理条件下的结果，但绝不可能同时观测到两者。

对于非模拟数据，由于反事实事件并未发生，因此无法得知反事实真实值。然而，对所有个体的个体处理效应取平均值，即 $\mathbb{E}[Y(1) - Y(0)]$, ，可以估计平均处理效应（ATE），而我们有真实世界数据集，其真实值已知。因此，为了评估我们的方法，我们首先通过估计每个样本对应的反事实结果，来估计各样本的个体处理效应。随后，我们对这些估计值取平均，得到平均处理效应，再将其与平均处理效应的真实值进行比较。由于我们的技术旨在构建个体层面反事实，而非直接估计平均处理效应，因此此处的目标不是与其他平均处理效应估计方法进行比较，而是展示通过反事实计算出的个体处理效应具有较高的准确性。

以下实验基于常用数据集。每项实验均重复进行 10 次，对每个问题使用相同的模型超参数，且不针对具体问题进行微调。对于每个数据集，我们使用

数据集	算法	平均处理效应（根据个体处理效应计算）	相对绝对误差（%）
IHDP	DCM	4.013	0.199%
IHDP	ANM	3.957	1.592%
Lalonde	DCM	1579.59	3.672%
Lalonde	ANM	1516.20	7.538%

表8：通过计算个体层面反事实进行平均处理效应估计。IHDP和Lalonde数据集提供的真实平均处理效应分别为4.021和1639.80。

文献中的真实图结构，并为每个非根节点指定一个 DCM模型。对于根节点，我们可以直接使用经验分布，无需拟合特定模型。

婴儿健康与发展计划数据集。 该数据集旨在预测专业托儿服务对婴儿认知测试分数的影响（Hill，2011）。这里的目标是考察，与未接受处理的婴儿相比，接受处理（专业托儿服务）是否能够提高婴儿的认知能力。这里的平均处理效应是do（专业托儿服务 = 1）下的预期认知测试分数与do（专业托儿服务 = 0）下的预期认知测试分数之差。与该数据集相关的真实平均处理效应（认知测试分数）为4.021。

Lalonde数据集。 该数据集包含不同的人口统计变量（例如性别、年龄、教育程度等），旨在考察培训项目是否能提高收入（LaLonde，1986）。此处的平均处理效应是 do(training program = 1) 下的预期收入与 do(training program = 0) 下的预期收入之差。与该数据集相关的真实平均处理效应（针对收入）为 1639。80。

结果。 结果汇总于表8。如上所述，本文的重点是计算个体层面反事实。通过 DCM 方法计算的平均处理效应估计值与真实平均处理效应相符，这可以视为个体层面反事实计算准确的迹象。作为基准，我们还展示了用加性噪声模型（ANM）方法替代 DCM 方法来建模结构因果模型的结果。

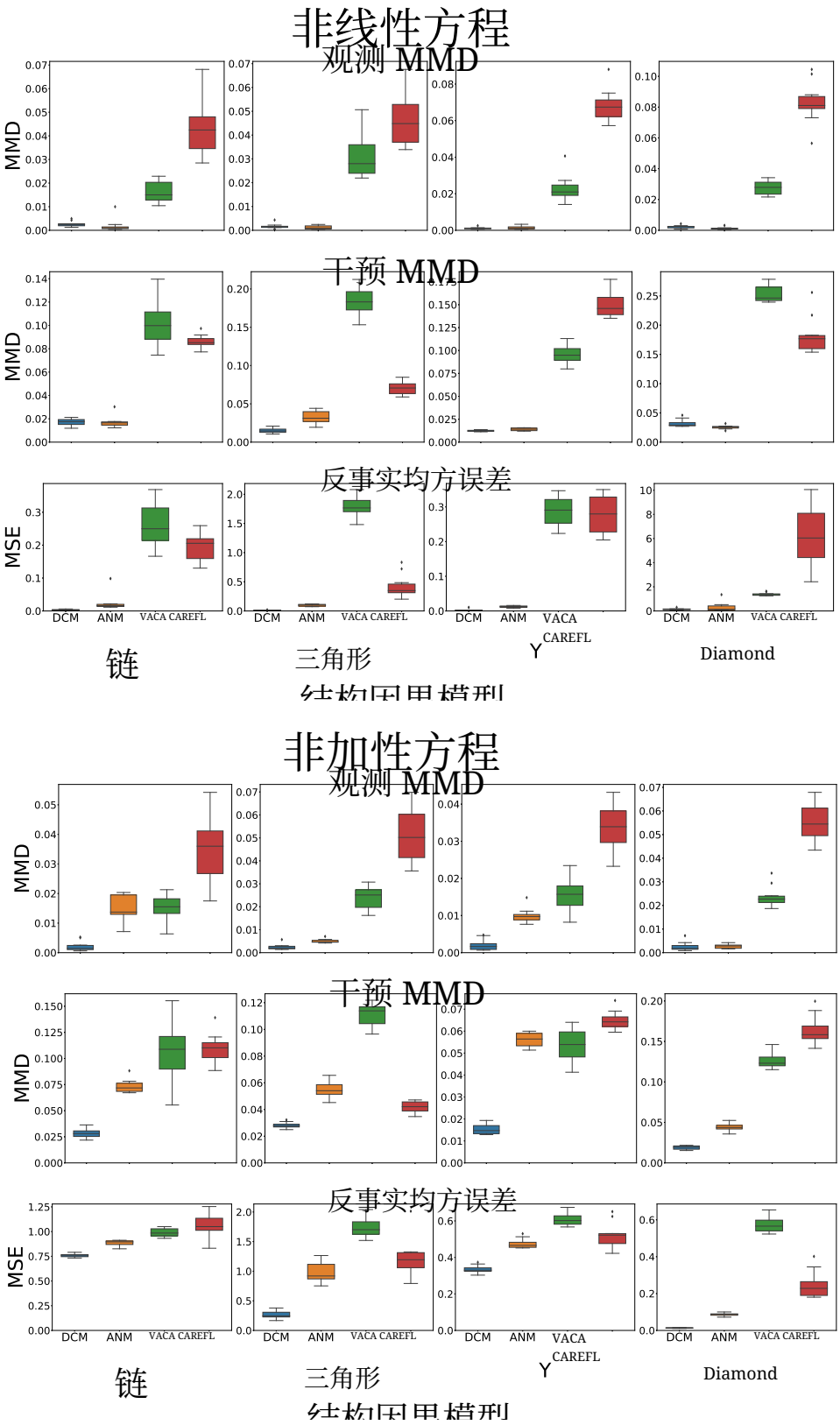

图 2：上：非线性设置（NLIN）；下：非加性设置（NADD）。四种不同结构因果模型的观测、干预和反事实查询的箱线图，基于模型和训练数据的 10 次随机初始化。

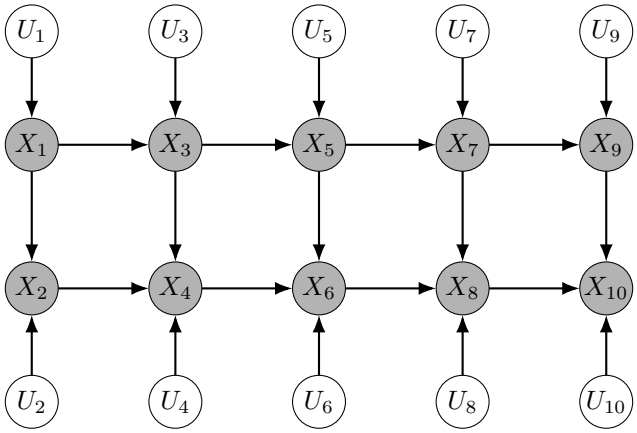

图3：第5节中使用的梯形图。

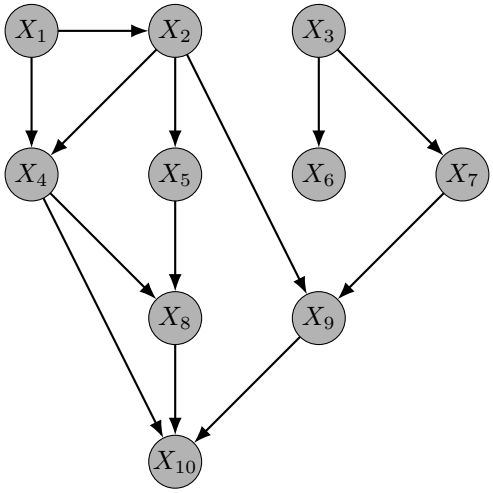

图4：第5节中使用的随机图示例，为清晰起见省略了外生噪声节点。

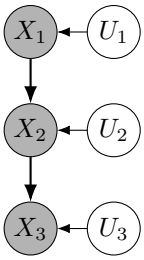

图5：链图。

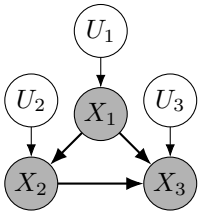

图6：三角图。

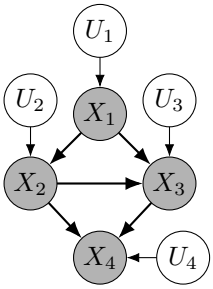

图7：菱形图。

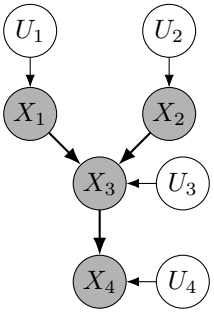

图8：Y形图。

图9：附录D.3中实验所用的因果图。