arXiv:2303.04209v2 [cs.LG] 5 Jul 2023

因果依赖图

Joshua R. Loftus
伦敦政治经济学院统计系，伦敦，英格兰，英国 j.r.loftus@lse.ac.uk

Lucius E. J. Bynum
纽约大学数据科学中心，纽约，纽约州，美国
lucius@nyu.edu

Sakina Hansen
伦敦政治经济学院统计系，伦敦，英格兰，英国
s.a.hansen1@lse.ac.uk

摘要

解释人工智能或机器学习模型正变得日益重要。要明智地使用这类数据驱动系统，我们必须了解它们如何与世界互动，包括它们在因果上如何依赖数据输入。在本文中，我们提出因果依赖图（CDP），用于可视化一个变量——结果——如何随另一个变量——预测变量——的变化而变化，以及其他预测变量随之产生的任何因果变化。至关重要的是，CDP 不同于那些通过固定其他预测变量或假设它们相互独立来进行分析的标准方法。CDP 使用辅助因果模型，因为因果结论需要因果假设。通过模拟和真实数据实验，我们展示了 CDP 可以与因果学习或敏感性分析方法以模块化方式结合。由于人们常常从因果角度思考输入输出依赖关系，CDP 可以成为可解释人工智能或可解释机器学习工具箱中的强大工具，并有助于科学机器学习和算法公平性等应用。

1 引言

本文提出因果依赖图（CDP），用于可视化预测变量与结果变量之间的因果关系。这个想法具有普遍性，但我们的研究动机是解释或阐释人工智能或机器学习模型 [6, 13, 14, 27]。为简单起见，我们考虑监督学习，即使用一组特征来预测结果，例如回归或分类。我们还聚焦于与模型无关或“黑箱”解释情境，在这种情境下，解释者可以查询模型，但无法访问其内部结构。此情境下的解释方法更广泛适用于分布式研究，但其功能也仅限于观察模型如何响应输入变化。我们的一般方法还可扩展至这一初始应用之外。

一次聚焦于一个输入变量的可视化和简单解释，是帮助人们理解的有力工具。两种流行的可视化方法——Friedman 的部分依赖图（PDP）[10] 和 Goldstein 等人的个体条件期望（ICE）图 [11],——展示了模型输出如何依赖于某个输入变量。然而，正如解释线性回归模型系数时一样，一次只聚焦于一个预测变量所揭示的关系可能具有误导性。改变一个输入变量时，我们必须决定其他输入变量采用哪些取值。部分依赖图将其他预测变量视为与所绘制的变量相互独立，因此，如果预测变量彼此独立且模型具有加性，它们就能正确捕捉模型对各变量的依赖关系 [15]。一般而言，在解释黑箱对某个输入变量的依赖时，我们处理模型其他输入变量的方式，可能会破坏或保留预测变量之间现有的统计或因果依赖关系。由此，我们提出如下问题：

问题陈述。 如果现实世界中的预测变量之间存在因果关系，但我们的可视化、解释或解释方法没有尊重这些关系，那么由此得出的模型

预印本。审稿中。

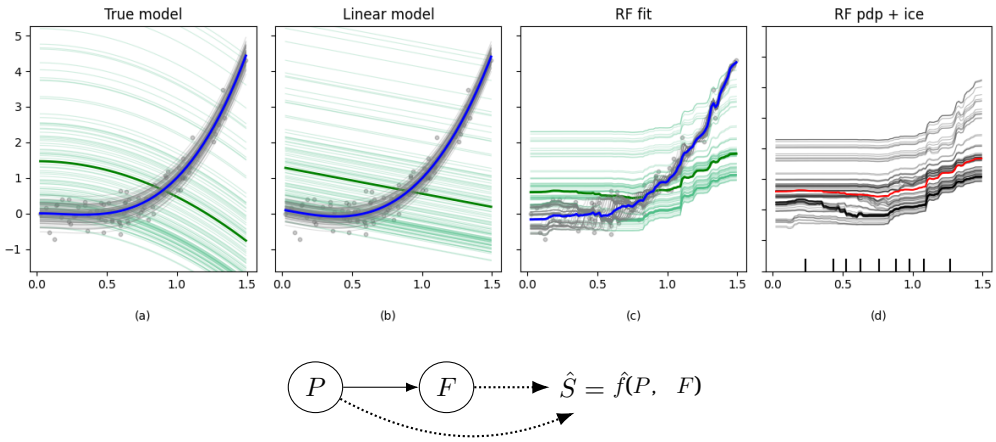


图 1：示例说明。示例的因果依赖图（上方）及其解释性因果模型的结构图。在上排中，面板（a）展示了 ECM 的关系。总依赖以蓝色表示，自然直接依赖以绿色表示。各个点的反事实曲线以细而浅的线条表示，聚合结果则以粗而深的线条表示。面板（b-c）分别展示了针对线性模型和随机森林模型的模型解释图。面板（d）展示了标准的部分依赖图（红色）及个体条件期望曲线，以便与我们的 NDDP（绿色）进行比较。

解释对于现实世界的用途而言可能无关紧要或具有误导性 [28, 41]。例如，这类解释可能导致人们在监管或调整算法系统时作出错误决策，依据模型预测分配资源不当，造成人类反馈与强化学习系统之间出现脱节，或引发其他形式的伤害。在科学机器学习领域，解释可用于提出假设以供后续研究，因此有缺陷的解释可能会引导我们提出虚假的假设。基于这些原因，我们可能会关注模型解释的因果有效性。

我们的方法。 概括而言，我们使用辅助性的解释性因果模型（ECM）来解读或解释给定的机器学习模型。对于希望进行解释的每个输入预测变量，我们使用 ECM 来确定操纵该预测变量时其他输入如何变化，而不是将它们视为相互独立或固定不变。我们将所得图称为因果依赖图，即 CDP。

因果模型可以根据所需的解释进行设计，利用先验领域知识加以指定，和/或从数据中潜在地学习并估计得到。

动机示例。 在介绍完整细节之前，我们先用一个简单示例说明这一思想。考虑一个关于父母收入 P、学校经费 F 和毕业生平均起薪

$$S \qquad P \sim \mathcal{U}[0, 1.5] \quad F = 2P^3 + \mathcal{N}\ 0\ 0.2^2 \qquad S = \text{的中介模}$$

$F - P^2 + \mathcal{N}\ 0\ 0.2^2$
型，其结构方程为；(，)；以及(，)。相应的有向无环图（DAG）显示在图 1 底行，顶行左侧面板绘制了数据，其余面板则展示了用于预测 $\hat{}_S = \hat{}f(P, F)$ 的监督模型的可视化解释。
从这幅图中可以看出几个重要结论。首先，蓝色曲线和绿色曲线之间的差异表明直接（或部分）依赖与总依赖之间可能存在定性差异；当我们考虑因果干预可能如何改变结果时，这一点影响重大。其次，对模型的解释可能与其背后的因果关系存在定性差异。例如，即使是像面板 (c) 中的随机森林这样灵活的模型，也显示出 S 对 P 的直接依赖呈递增趋势，而真实模型中 S 对 P 的直接依赖却呈递减趋势。再如，面板 (b) 表明，线性模型对某个预测变量的总依赖可能是非线性的，因为中介变量 F 对 P 的依赖是非线性的。最后，我们看到以因果模型生成解释图的框架，涵盖了一些现有模型解释图作为特例，例如 *ICE* 和部分依赖图。我们将在后文再次讨论这一点。

2 方法论

We 在第 2.1–2.2 节介绍背景符号和定义，并在第 2.4–2 节介绍我们的方法。 6.

2.1 解释监督学习模型

为便于具体说明，我们考虑这样一种监督学习设定：包含一组预测特征 $\mathbf{X}$ 和结果变量 Y。我们希望解释或阐释由函数 f表示的预测模型，该模型通常通过经验风险最小化（ERM）进行估计或学习

$$\hat{f} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell\left(h(\mathbf{x}_i), y_i\right)$$

使用某个损失函数 ℓ、预先指定的函数类 $\mathcal{H}$，以及独立同分布的训练样本 $\{(y_i,\ \mathbf{x}_i)\colon\ i = 1,$ ……，$n\}$，其特征向量为 $\mathbf{x}_i^T \in \mathbb{R}^p$。在 2.5 节中，我们将聚焦于一项称为中介分析的特定解释任务；再次强调，这一任务既便于具体说明，也是一个应用广泛的例子。在该节中，我们将预测变量划分为若干子集，使得 X和 M 都表示预测变量，而 M 表示中介变量。

模型无关解释。有时，出于实际原因，解释方法无法获知 f的内部数学结构。在这种模型无关的设定下，我们可以通过提供合成输入，根据输入输出依赖关系生成解释

$$\tilde{\mathbf{X}} \mapsto \hat{f}\left(\tilde{\mathbf{X}}\right),$$

记录相关预测，然后以某种方式对其进行汇总。我们将以这种方式生成的模型无关解释记作 $\mathcal{E}(f; \mathbf{X})$，或简写为 $\mathcal{E}(f)$。但请注意，所有此类解释（包括我们提出的方法）都取决于 f和合成输入。为了便于解释和创建低维可视化，我们通常一次只针对一个特征。我们将针对特征 j 的解释记作 $\mathcal{E}_j(f)$；在这种情况下，合成输入为

$$\tilde{\mathbf{x}}_j := (x_1, \ldots, \tilde{x}_j, \ldots, x_p) \mapsto \hat{f}\left(\tilde{\mathbf{x}}_j\right).$$

如果特征 j 是数值型的，通常会沿其取值范围内的网格变化。在大多数现有的模型无关解释方法中，其他特征的值会固定为用于生成解释的数据集中的观测值。请注意，该解释数据集可能与 f的训练数据并不相同。为强调这一点，我们将用于生成解释的数据记作 $\{(y_i',\ \mathbf{x}_i')\colon\ i = 1,$ ……，$m\}$，其中 y_i'可能不提供，也可能因解释类型而无需提供。当解释特征为类别型时，可以使用条形图。

定义 2.1（个体条件期望（ICE）图）。 对于解释数据集 $x_i', i = 1, \ldots, m$ 中的每个观测值，我们通过绘制映射得到一条单独的曲线

$$\tilde{x}_j \mapsto \hat{f}\left(\tilde{\mathbf{x}}_{ij}'\right), \text{ where } \tilde{\mathbf{x}}_{ij}' := (x_{i1}', \ldots, \tilde{x}_j, \ldots, x_{ip}').$$

特征 j 的个体条件期望图会显示所有 m 这些曲线。

定义 2.2（部分依赖图（PDP））。 特征 j 的部分依赖图可以通过计算网格中每个点处解释数据的经验均值，从特征 x_j 的个体条件期望图得到，即 $\tilde{x}_j \mapsto \frac{1}{m}\sum_{i=1}^{m} \hat{f}\left(\tilde{\mathbf{x}}_{ij}'\right)$.

还有各种其他模型无关解释方法，例如参见 [27]。但由于我们当前提出的是一种可视化方法，因此这些是用于比较的主要替代方法。

单变量解释的根本问题。要创建一种模型无关解释，用于说明^模型对单个特征的依赖性，例如以 x_j 为横轴、 f 为纵轴的图，我们必须决定在合成变化 $\tilde{x}_j$时如何处理其他特征。几乎所有现有解释方法都采用与PDP和个体条件期望图相同的方法：它们将其他特征固定在辅助解释数据集中的取值上。例如，如果其他特征在因果地依赖于 x_j ，这种做法可能并不现实。如果特征相互构成，例如人口、国内生产总值和人均国内生产总值，那么这种做法在数学上可能没有定义，或者无法作出任何解释。

2.2 结构因果模型

以下符号约定和定义受到 [5, 32, 33]的影响。令$\mathbf{U}$ 为一组外生噪声变量，$\mathbf{V}$ 为一组 $p = |\mathbf{V}|$ 可观测变量，并令$\mathbf{G}$ 为一组函数，使得对于每个 $j \in 1, \ldots, p$，都有 $V_j = g_j(\mathbf{PA}_j, U_j)$，其中$\mathbf{PA}_j \subseteq \mathbf{V}$ 和 $U_j \subseteq \mathbf{U}$ 分别是变量 V_j的可观测父变量和外生父变量。令有向无环图（DAG）$\mathcal{G}$ 的顶点为各变量；对于每个 $V_j \in \mathbf{V}$ 及其在$\mathbf{PA}_j$ 和U_j中的每个父变量，都添加一条从父变量指向 V_j的有向边。该图可以通过直观展示$\mathbf{G}$中每个函数的输入变量，帮助进行解释。

定义 2.3（结构因果模型（SCM））。（概率）结构因果模型 $\mathcal{M}$ 是一个元组 $\langle \mathbf{U}, \mathbf{V}, \mathbf{G}\ P_{\mathbf{U}} \rangle$，其中 $P_{\mathbf{U}}$ 是外生变量的联合分布。该分布和函数 $\mathbf{G}$ 决定了 $\mathcal{M}$中所有变量的联合分布 $P^{\mathcal{M}}$。最后，该模型中的因果关系通过额外假设来表示，这些假设使得 $\mathcal{M}$ 能够按照下文的定义对干预和/或反事实进行建模。

定义 2.4（干预）。对于结构因果模型 $\mathcal{M}$，干预 I 会产生一个修改后的结构因果模型，记为 $\mathcal{M}^{\mathrm{do}(I)}$，它可能具有不同的结构方程 $\mathbf{G}^I$。相应地，一些变量的父节点集可能发生变化，因此有向无环图表示 $\mathcal{G}^{\mathrm{do}(I)}$ 也可能随之改变。我们将新的干预分布记为 $P^{\mathcal{M};\mathrm{do}(I)}$。一类简单的干预是对单个变量进行干预，例如：

$$I = \mathrm{do}\left(V_j := \tilde{g}(\tilde{\mathbf{PA}}_j, \tilde{U}_j)\right),$$

这会改变 $\mathcal{G}$ 中 V_j 及其所有有向路径上的变量的生成方式。更简单的子类是原子干预，即将一个变量 V_j 设为某个常数值 v，记作 $I_{j,\ v} := \mathrm{do}(V_j = v)$。请注意，在这种情况下，图中 V_j 没有父节点 $\mathcal{G}^{\mathrm{do}(I)}$；干预本身的来源位于模型所描述的世界之外。

干预有助于对数据生成过程（DGP）的变化进行建模，例如，通过实验控制某个特定变量，观察其取值如何改变其他变量；又如，通过政策变更来改变或消除现有的因果关系。除了作为数据生成过程生成新的观测值外，结构因果模型（SCM）还可用于对已确定观测值的反事实值进行建模。反事实分布是在特定数据集上定义的干预分布，其中包含该数据中部分观测值所提供的信息或约束，下面将对此进行说明。这里我们稍微滥用一下记号，用粗体表示数据集，例如，$\mathbf{V}$ 表示先前生成的数据集中所有观测对应的观测变量。

定义 2.5（反事实）。对于变量 V_j，已知其父变量 $\mathbf{PA}_j = v$ 的观测值时，我们可以固定 v 中的部分或全部变量，并改变 $U_j := u$，再将这些值传入 $g_j(v, u)$（如果我们还进行了干预，改变了 $\mathbf{PA}_j$ 中任意变量的取值，则传入 $\tilde{g}_j(\tilde{v}, u)$）。反事实 $V_j(\tilde{v}, u)$ 是指 V_j 在其观测到的父变量和/或外生父变量取不同值（$\tilde{v}, u$）时本会取到的值。给定干预 I，要定义反事实分布 $P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\mathrm{do}(I)}$，我们使用后验分布或条件分布（取决于所采用的概率模型方法）$P_{\mathbf{U}|\mathbf{V}=\mathbf{v}}$，在修改后的结构因果模型 $\mathcal{M}^{\mathrm{do}(I)}$ 中，为先前生成的观测计算任意变量的反事实值时，用它来描述关于 $\mathbf{U}$ 的不确定性。

2.3 单变量因果解释

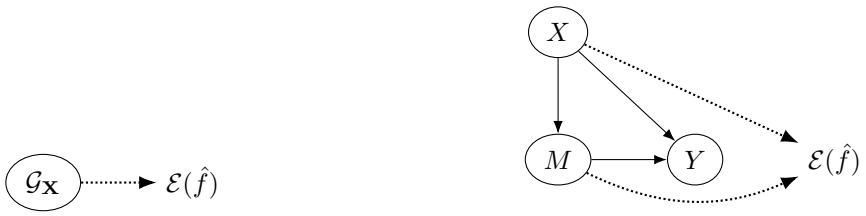
针对单变量解释中凸显的根本问题，我们提出的解决方案是使用辅助 ECM $\mathcal{M}_j$，并让该因果模型决定其他特征如何随 x_j 变化。我们将这些解释记作 $\mathcal{E}_j(f;\ \mathcal{M}_j)$；若上下文明确，则记作 $\mathcal{E}_j(f)$。在确定性或无噪声情形下，假设我们知道函数 g_{kj}，使得 $x_k = g_{kj}(x_j)$，其中 g_{jj} 为恒等映射。在这种情况下，模型 $\mathcal{M}_j$ 告诉我们

$$x_j \mapsto g(x_j) =: (g_{1j}(x_j), \ldots, x_j, \ldots, g_{pj}(x_j))$$

$\hat{}$ 是 $\mathbb{R}^p$ 中一条由 x_j 参数化的曲线，我们使用以下方式生成解释 $\mathcal{E}_j(f)$

$$\tilde{x}_j \mapsto \hat{f}\left(g(\tilde{x}_j)\right).$$

接下来，我们将给出更多定义，以便把这一策略扩展到更一般的非确定性因果模型。我们提出，使用辅助 ECM $\mathcal{M}_{\mathbf{X}}$（它刻画了预测变量之间的因果关系），生成监督学习模型 $\hat{f}$（可能是黑箱）的各种因果解释。借助 $\mathcal{G}_{\mathbf{X}}$ 对应的有向无环图，我们在图2中以图形方式表示这一结构，其中从子图 $\mathcal{G}_{\mathbf{X}}$ 指向解释 $\mathcal{E}(f)$ 的 $\hat{}$ 箭头以虚线表示，以区别于特征之间的箭头。不同的解释对应于在 $\mathcal{M}_{\mathbf{X}}$ 中执行的各种因果操作，例如干预。



$\hat{}$ 图2：使用预测变量 $\mathcal{G}_{\mathbf{X}}$ 的结构因果模型生成对预测模型 f 的解释 $\mathcal{E}(f)$。在中介示例 (b) 中，预测变量 X 直接导致 Y，也通过中介变量 M 导致 Y。实线箭头表示数据生成过程中的因果关系，虚线箭头表示模型解释对预测变量的形式依赖。

备注 2.6。 请注意，如果所需的因果解释使用反事实，我们很可能会从辅助解释数据集中获取观测值。不过，由于结构因果模型能够生成数据，我们也可以用它来生成初始观测值，并在计算解释所需的反事实时再次使用这些值。也就是说，我们可以从 $P^{\mathcal{M}}$ 生成辅助解释数据集 $\mathbf{V}$，然后通过干预 I，从 $P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\mathrm{do}(I)}$ 生成反事实。

结构因果模型的通用性和灵活性使我们能够提出许多不同的解释性问题；由于结构因果模型可以生成合成数据，我们可以用它们计算许多不同类型的解释。在接下来的两节中，我们将重点介绍如何使用图作为解释，并描述几种典型类型的解释性问题。

2.4 因果依赖图

这项工作的一个重要动机，是定义 PDP 和个体条件期望图的各种因果类比形式；现在我们已经可以着手定义了。对于以下定义，我们假设存在预测变量 $\hat{}\mathbf{X} \in \Omega_{\mathbf{X}}$、感兴趣的结果 $Y \in \Omega_Y$，以及一个黑箱函数 $f(x) : \Omega_{\mathbf{X}} \to \Omega_Y$，其输出为 $\hat{}$，也可记作 Y。结构因果模型 $\mathcal{M}_{\mathbf{X}}$ 要么是预先假设的，要么是从数据中学习得到的。重要的是，$\mathcal{M}_{\mathbf{X}}$ 仅指定预测变量 $\mathbf{X}$ 之间的因果关系，无须涉及结果 Y。

个体反事实和期望效应。 我们用简写 $\hat{} f(P^{\mathcal{M}})$ 表示将来自分布 $P^{\mathcal{M}}$ 的数据作为输入传入 $\hat{}$ 黑箱函数 f；其中，$\hat{} f$ 以分布 $P^{\mathcal{M}}$ 作为参数。对于每种图，我们都会展示一组个体反事实曲线 $\hat{} f(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\mathrm{do}(I)})$，以及它们在解释数据集上的经验平均值

$$\hat{\mathbb{E}}\left[\hat{f}(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\mathrm{do}(I)})\right]. \tag{1}$$

对于给定命名效应的每种因果解释类型，我们将个体反事实命名效应曲线定义为解释数据集中每个个体对应的反事实曲线 $\hat{} f(P^{\mathcal{M}})$ 的集合；将命名效应函数定义为这些曲线的期望；将命名依赖图定义为展示所有这些曲线的图（对于效应函数，则展示其经验估计）。

在定义不同类型的因果依赖图之前，我们先介绍一种有用的抽象。生成因果解释需要基于结构因果模型执行溯因、行动和预测；该模型经过扩展，纳入了我们希望解释的预测。

定义 2.7（解释性因果模型（ECM））。 ECM $\mathcal{M}'$ 通过将预测结果 Y 作为一个额外变量，并将 f 作为其结构方程，扩展了原始结构因果模型 $\mathcal{M}_{\mathbf{X}}$ˆ ˆ。然后，例如，我们可以使用 ECM 计算 ˆ $f(P^{\mathcal{M}_{\mathbf{X}}}|_{\mathbf{V}=\mathbf{v};\text{do}(I)})$，即 $P^{\mathcal{M}'}|_{\mathbf{V}=\mathbf{v};\text{do}(I)\hat{Y}}$。我们在算法 1 中介绍 ECM 的构造方法。

我们使用算法 2 至算法 5 中的这一流程来计算下文介绍的各种效应。首先介绍一种或许最直观且最重要的因果效应。

定义 2.8（总依赖图（TDP））。 对于干预 I，个体反事实总效应（ICTE）曲线

$$\mathsf{TE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I)}) \tag{2}$$

展示了解释数据集中每个个体观测的干预 I 对黑盒输出的总效应。对解释数据集中的这些效应取经验平均值，可得到总效应函数（TEF）的（蒙特卡洛）估计；展示 ICTE 和 TEF 的图即为总依赖图（TDP）。我们按照算法 2 计算 TDP。

备注 2.9。 在余下的定义中，我们仅给出个体反事实曲线的符号表示，其他对象则不再明确定义。

ˆ 我们通常希望分解 X 对 Y（或 Y）的总效应，了解其中有多少可归因于变量之间不同的可能路径。为此，除了上文介绍的总效应之外，还需要理解若干因果量，现将其定义如下。

定义 2.10（部分控制依赖图（PCDP））。 考虑对有向无环图 $\mathcal{G}_{\mathbf{X}}$ 中某些变量子集施加干预 I，以及原子干预 C，该干预将未受 I 干预的其他变量子集固定不变。个体反事实部分控制效应曲线

$$\mathsf{PCE}(I, C) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I,C)}) \tag{3}$$

ˆ 表示干预 I 对黑盒输出 Y 的影响，同时将其他变量（通过干预）设为常数值。我们使用算法3计算 PCDP。

定义 2.11（自然直接依赖图（NDDP））。 考虑原子干预 I 以及相应的干预 J：该干预会作用于所有受干预 I 影响的节点的子节点，并将这些子节点设为它们在数据集 $\mathbf{x}$ 中的观测值。例如，如果 $I = \text{do}(A = a, B = b)$，那么干预 J 会将变量 A 和 B 的所有子节点设为它们在$\mathbf{x}$中的观测值。接着，我们将个体反事实自然直接效应曲线定义为

$$\mathsf{NDE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I,J)}). \tag{4}$$

ˆ 该量表示干预 I 对黑盒输出 Y 的影响，同时将所有未受干预的变量设为其“自然”值，也就是解释数据集中的干预前取值。算法 4 展示了如何计算 NDDP。

通过构造 NDDP，并将其与定义 2.2 对比，我们可以看出它等价于 PDP，这也证实了我们在图1中观察到的现象。

命题2.12. 部分依赖图展示自然直接效应。使用解释数据 $D \sim \mathcal{M}_{\mathbf{X}}$生成图时，个体条件期望图的曲线与个体反事实自然依赖曲线完全相同，因此部分依赖图和 *NDDP* 图也完全相同。

定义 2.13（自然间接依赖图（NIDP））。 考虑原子干预 I 以及相应的干预 K：该干预从有向无环图 $\mathcal{G}_{\mathbf{X}}$ 中移除由干预 I 所作用的任意节点发出的所有边，并将这些节点设置为解释数据集中的观测值。例如，如果 $I = \text{do}(A = a,\ B = b)$，则干预 K 会移除从 A 和 B 发出的所有边，并将 A 和 B 设置为其原始观测值。随后，我们定义个体反事实自然间接效应曲线

$$\mathsf{NIE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}^{\text{do}(I)}|\mathbf{X}=\mathbf{x};\text{do}(K)}). \tag{5}$$

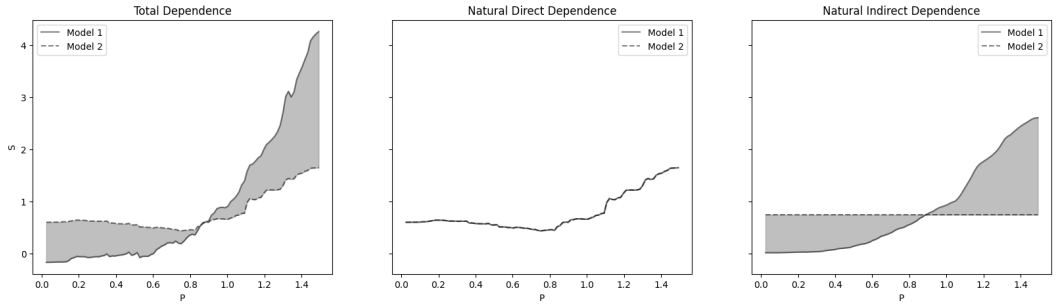


图3：薪资示例中的 TDP、NDDP 和 NIDP，纳入了因果不确定性，现在展示由两个相互竞争的候选因果模型所诱导的一系列可能的效应函数——其中一个包含中介，另一个不包含中介。

注意，干预 I 在干预 K之前实施。该量表示干预 I 对黑盒输出 Y的效应，且该效应仅归因于通向 Y的任何间接路径。我们根据算法 5计算 NIDP。该函数两个值之间的差异可用于将自然间接效应表示为一种特殊情形。

补充材料包含计算上述所有图的算法的完整说明。

2.5 中介分析

许多应用都涉及一种我们称之为中介三角结构的因果结构，图2b展示了一个示例。在中介分析中，我们通常希望分解 X 对 Y的总效应中，有多少可归因于通过 M 的路径，以及有多少属于直接效应。上述定义使我们能够将这一情境下经常研究的感兴趣量可视化。例如，PCDP 两个值之差可用于将控制直接效应表示为一种特殊情形——具体而言，在中介三角结构中，干预 I、C 分别定义为 $I = \text{do}(X = x)$ 和 $C = \text{do}(M = m)$。虽然中介分析为 CDP 提供了动机并有助于建立直观理解，但我们强调，第 2.4 节中的定义可用于任意结构因果模型。其他更复杂的示例请参见第 3 节。

2.6 将不确定性纳入因果依赖

将有关真实因果模型的不确定性纳入因果依赖图，有多种方法。现在，我们来探讨对因果依赖图的一种自然的初步扩展，它展示了由一组辅助 *ECM*所产生的效应函数的可能范围。ECM 集合可以预先指定，也可以是因果结构学习算法输出的有向无环图的马尔可夫等价类。回到第 1 节中的启发性示例，我们可能会质疑父母收入 P 是否会影响学校经费 F，并转而考虑一个不含中介的结构因果模型： $P \to S \leftarrow F$。图 3 展示了针对 TDP、NDDP 和 NIDP，效应函数在不含间接效应的 ECM 与第 1 节原始 ECM 之间插值所形成的可能范围。这里，我们假设两个模型共有的边具有相同的结构方程。这些图展示了当我们不确定另一个预测变量如何依赖于*P*时，预测可能在多大范围内取决于一个预测变量 P。补充材料中，我们展示了一个真实数据示例，其中使用了由 PC算法发现的候选经验因果模型。

在这个示例中，我们展示了将多个因果模型纳入因果依赖图，如何让我们直接可视化对潜在因果模型的不确定性，以及这种不确定性对我们预期预测变量 X 会对黑盒输出 Y产生的影响ˆ。更广泛地说，这一过程让我们能够描述在多种条件下， X 将如何影响 Y或 Yˆ，从而实现灵活的敏感性分析。

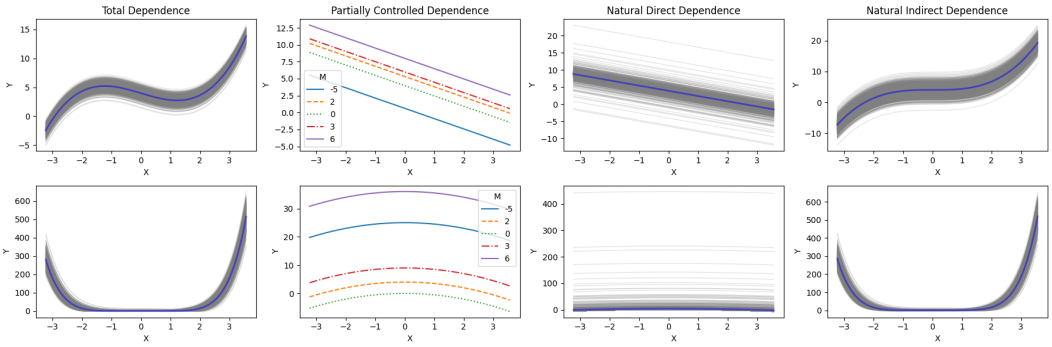


图4：第 3 节模拟示例的因果依赖图，分别展示了「正确」黑箱模型（下排）和「不正确」黑箱模型（上排）。

3 实验

现在，我们将展示本方法能够生成的几种因果解释。

模拟。 考虑由以下数据生成过程支配的非线性中介示例 P.

$$X \sim \mathcal{N}(0,1), \quad M = \frac{1}{2}X^3 + \mathcal{N}(0,1), \quad Y = M^2 - \frac{1}{2}X^2 + \mathcal{N}(0,1).$$

我们使用此数据生成过程拟合两个不同的黑箱模型：一个假设函数形式正确（即上述数据生成过程所示的 Y 关系），另一个「不正确」的模型则通过线性回归预测Y。在实验中，我们使用 Python DoWhy软件包 [3, 40] 采样反事实数据，以构建图。图4展示了以黑箱模型的训练数据作为解释数据时，各模型对应的因果依赖图。从图4中可以得出几点见解。首先，所有因果依赖图都对黑箱模型的函数形式假设是否符合真实值数据生成过程敏感。其次，我们可能想要研究的结果 Y上的不同效应，会在不同的图中直观呈现。例如，不正确模型的 TDP 显示出对 X 的依赖关系，看起来呈立方关系，而真实关系则是二次且六次关系。通过观察 NIDP 和 NDDP，我们可以看出，对于不正确的模型，这种立方关系是由通过中介变量 M产生的间接效应所致。正确模型与不正确模型的 PCDP 直接显示了二次项——即对 Y的控制直接效应——在黑箱模型中是否存在。

结合因果发现的真实数据。威斯康星乳腺癌（原始）数据集 [44] 是一个公开可用的数据集，常用于测试医学数据上的算法。该数据集包含 9 个有序变量，表示乳腺肿块中细胞的属性：细胞团厚度、细胞大小均匀性、细胞形状均匀性、边缘黏附、单层上皮细胞大小、裸核、温和染色质、正常核仁和有丝分裂。结果变量是乳腺肿瘤的类别，即良性或恶性。

为简单起见，我们在较小的预测变量子集上使用一种因果结构学习算法，具体来说是 J ulia CausalInference 中实现的 PC算法 [42]，以学习该数据集的有向无环图。图 5 展示了所得的有向无环图，以及用于对类别变量进行分类的随机森林模型的因果依赖图。

这表明，因果依赖图可以与结构学习算法等其他因果方法结合使用。PC算法的输出显示，细胞大小和细胞形状之间存在一条无向边，因此我们在补充材料中研究了对与该输出一致的不同图结构的敏感性。

4 应用、扩展、相关工作

协变量偏移下的解释：模型训练完成后，通常会用于预测可能不符合训练数据生成过程的数据。如果我们掌握偏移后的协变量分布中的因果关系，

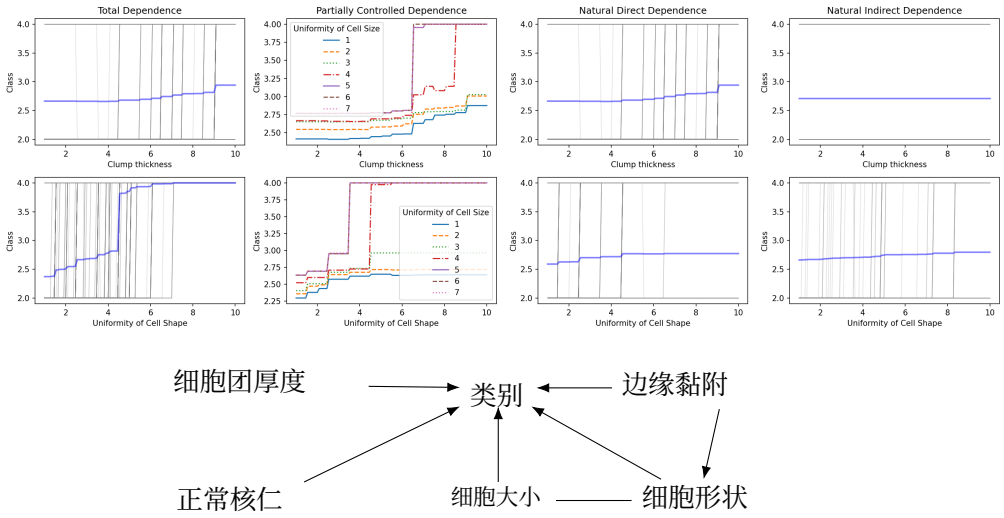


图 5：乳腺癌数据示例。随机森林分类器的因果依赖图（CDPs），以及预测变量细胞团厚度（第一行）和细胞形状均匀性（第二行）。由 PC算法学习得到的 ECM 结构图 $\mathcal{G}_B$（最后一行）。结果类别为二元变量：2 表示良性，4 表示恶性。

就可以利用这些信息来选择 ECM，而因果依赖图则可以直观展示模型在协变量偏移下的表现。

因果半监督学习：已知预测变量之间的因果结构，并拥有一个用于预测结果的监督学习模型后，我们的方法可用于尝试推断从预测变量到结果的因果关系，其思路与 [39] 和 [49]类似。

公平性或其他期望属性的审计：先前研究已将因果方法论应用于公平性[4, 8, 20–22, 25, 26, 29, 36, 46, 48], 补救 [19, 34, 45], 以及其他期望属性。因果依赖图（CDPs）可用于探查黑箱是否具备此类属性。

科学理论发展：大型且复杂的机器学习模型可能需要在海量数据集上进行拟合，而这些数据集的底层结构大多未知。在这种情况下，相对简单的 ECM 可用于提出关于多个预测变量之间关系的假设，并绘制各种因果依赖关系，从而生成新的假设，或检验既有假设是否合理。

不确定性与敏感性分析：未来的工作可以开发用于可视化不确定性的方法。例如，基于保形预测的敏感性分析 [7, 16, 24, 47]。如果多个候选经验因果模型之间存在不确定性，第 2.6 节中的方法或 [36] 中类似的方法可以生成复合因果依赖图。

相关工作：近年来，受补救 [18]启发的研究采用了对比解释或反事实解释[43]。其中一些研究虽然使用了“反事实”一词，但并未以因果关系为基础；另一些则着重研究因果依赖 [38]。Blöbaum 和 Shimizu [2] 通过找出总效应最大的预测变量来生成因果解释，这种方法在假设线性的情况下最为适用。Zhao 和 Hastie [49] 研究了部分依赖图的因果解释。该论文旨在利用这类图对底层数据生成过程进行因果推断，而非解释黑箱；研究表明，当数据生成过程满足后门准则 [30]时，部分依赖图会呈现预测变量的总效应（TE）。Cox Jr [9] 观察到部分依赖图与自然直接效应（NDE）之间存在关联；据我们所知，我们在命题2.12中首次正式确立了这种等价关系。Lazzari 等人 [23] 提出，在计算部分依赖图时对观测值加权，这或许可以用于混杂因素调整。我们的统一框架展示了如何将原本并非出于因果动机或论证而提出的各种模型解释图与因果解释联系起来。除了部分依赖图和 M图分别与自然直接效应和总效应之间的偶然联系外，我们不知道此前是否有研究为适用于我们框架的其他类型因果解释提供了解释或解释性图。

5 讨论

局限性。 因果建模通常存在一些局限性，我们不再赘述，但可参见例如 [12]。模型无关解释方法也存在局限性 [1, 27], ，其中有几点我们将特别指出，因为我们认为使用本方法时应牢记这些重要事项。

黑箱与真实数据生成过程之间的不匹配：如果预测模型无法拟合数据生成过程，那么当我们的解释目标是了解数据生成过程时，实际上任何模型解释也都会失败 [49]。在模型无关的场景中，这一问题尤其令人担忧，因为我们无法进行模型诊断，也无法探查黑箱的内部机制。

解释性结构因果模型与真实/目标数据生成过程之间的不匹配：如果真实数据生成过程在重要方面不同于 ECM，因果依赖图可能会产生误导。然而，标准部分依赖图及类似的解释方法同样需要辅助解释数据，而这些数据也可能不同于目标数据生成过程。因此，这并不是我们的方法所特有的额外局限，影响并不大。

ECM 的可用性：如果我们将模型视为工具，并不关心数据生成过程是否存在一个“真实”的因果模型，那么在生成解释时，我们仍然需要选择要使用哪些工具。从这个意义上说，完全掌握一个结构因果模型可能是一个很强的假设。然而，在第 2.6 节和第 4 节中，我们讨论了改进这一点的一些方法。总的来说，如果需要或必须提供因果解释，那么我们无法避免作出因果假设。

结论。 本文提出了因果依赖图，这是一种利用给定的（可能是学习得到的）解释性因果模型生成各种具有因果解释的图的方法。这使我们能够运用结构因果模型的强大语言，提出并回答各种具有因果意义的问题。我们的框架将此前彼此分散、并非出于因果动机而设计的解释工具（如部分依赖图）统一起来，并揭示了一些此前文献中尚未探讨过的新型因果解释。未来在这一方向上的进一步工作可以探索其他相对简单的典型因果结构，以用于实际应用；也可以与其他类型的模型结合，例如通过应用因果表征学习，将方法扩展到非表格数据。将解释方法与 Pearl因果阶梯联系起来 [31],大多数以往的可解释机器学习和可解释人工智能方法——如部分依赖图——关注的是关联，因此局限于阶梯的第一层。借助因果依赖图，我们得以沿阶梯向上，创建旨在改变世界的模型解释。尽管可解释性是提出因果依赖图的最初动机，但我们相信，绘制此类因果关系在其他场景中也会很有用。

参考文献

[1] T Altmann、J Bodensteiner、C Dankers、T Dassen、N Fritz、S Gruber 等。可解释机器学习方法的局限性。2020年。[2] Patrick Blöbaum 和 Shohei Shimizu。特征对预测的干预效应估计。载于*2017*年 *IEEE* 第*27*届国际信号处理机器学习研讨会（*MLSP*），第1–6页。IEEE，2017年。[3] Patrick Blöbaum、Peter Götz、Kailash Budhathoki、Atalanti A. Mastakouri 和 Dominik Janzing。Dowhy-gcm：将 Dowhy 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。[4] Lucius Bynum、Joshua Loftus 和 Julia Stoyanovich。通过细分干预减少不平等。载于算法、机制与优化中的公平与可及性，第1–13页。计算机协会，美国纽约州纽约，2021年10月。ISBN 978-1-4503- 8553-4。URL https://doi.org/10.1145/3465416.3483286。[5] Lucius Bynum、Joshua Loftus 和 Julia Stoyanovich。面向未来的反事实。载于*AAAI*人工智能会议论文集，2023年。[6] Diogo V. Carvalho、Eduardo M. Pereira 和 Jaime S. Cardoso。机器学习可解释性：方法与指标综述。*Electronics*，8(8):832，2019年8月。ISSN 2079-9292。

doi: 10.3390/ electronics8080832. 网址 htt ps://www.m dpi.com/2079 -9292/8/8/832。编号：8 出版商：多学科数字出版研究所。[7] Victor Chernozhukov、Kaspar Wüthrich 和 Yinchu Zhu。《用于反事实与合成控制的精确稳健保形推断方法》。《美国统计学会杂志》，116(536):1849–1864，2021年。[8] Silvia Chiappa。路径特定反事实公平。在*AAAI*人工智能会议论文集中，第33卷，第7801–7808页，2019年。[9] Louis Anthony Cox Jr.。在观察性数据中评估因果关系的 Bradford Hill 标准现代化。《毒理学评论》，48(8):682–712，2018年。[10] Jerome H Friedman。贪心函数逼近：梯度提升机。《统计年刊》，第1189–1232页，2001年。[11] Alex Goldstein、Adam Kapelner、Justin Bleich 和 Emil Pitkin。窥探黑箱内部：通过个体条件期望图可视化统计学习。《计算与图形统计学杂志》，24(1):44–65，2015年。[12] Sander Greenland 和 Mohammad Ali Mansournia。个体因果模型、因果图和可忽略性假设的局限性：随机混杂与设计不忠实性的例证。《欧洲流行病学杂志》，30:1101–1110，2015年。[13] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Franco Turini、Fosca Giannotti 和 Dino Pedreschi。黑箱模型解释方法综述。《*ACM*计算调查》，51(5):93:1–93:42，2018年8月。ISSN 0360-0300。doi: 10.1145/3236009。URL https://doi.org/10.1145/3236009。[14] David Gunning、Eric Vorm、Jennifer Yunyan Wang 和 Matt Turek。美国国防高级研究计划局的可解释人工智能（XAI）计划：回顾。《应用人工智能快报》，2(4):e61，2021年。ISSN 2689-5595。doi: 10.1002/ail2.61。URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61。_电子预印本：https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61。[15] Trevor Hastie 和 Robert Tibshirani。广义加性模型。《统计科学》，1 (3):297 – 310，1986年。doi: 10.1214/ss/1177013604。URL https://doi.org/10.1214/ss/ 1177013604。[16] Ying Jin、Zhimei Ren 和 Emmanuel J Candès。个体治疗效应的敏感性分析：一种稳健保形推断方法。《美国国家科学院院刊》，120(6):e2214889120，2023年。[17] Diviyan Kalainathan、Olivier Goudet 和 Ritik Dutta。因果发现工具箱：在 Python 中揭示因果关系。《机器学习研究杂志》，21:37:1–37:5，2019年。[18] Amir-Hossein Karimi、Gilles Barthe、Bernhard Schölkopf 和 Isabel Valera。算法补救综述：定义、形式化、解决方案与展望。*arXiv:2010.04050 [cs,stat]*, 2021年3月。URL http://arxiv.org/abs/2010.04050。arXiv：2010.04050。[19] Amir-Hossein Karimi、Julius von Kügelgen、Bernhard Schölkopf 和 Isabel Valera。迈向因果算法补救。载于 Andreas Holzinger 、 Randy Goebel 、 Ruth Fong 、 Taesup Moon 、 Klaus-Robert Müller 和 Wojciech Samek 编，《xxAI——超越可解释人工智能：与 ICML 2020 联合举办的国际研讨会，2020年7月18日，奥地利维也纳，修订与扩展论文集》*xxAI——超越可解释人工智能：与 ICML 2020 联合举办的国际研讨会，2020年7月18日，奥地利维也纳，修订与扩展论文集*，《计算机科学讲义》，第139–166页。施普林格国际出版，查姆，2022年。ISBN 978-3-031-04083-2。doi: 10.1007/978-3-031-04083-2_8。URL https://doi.org/10.1007/978-3-031-04083-2_8。[20] Niki Kilbertus、Mateo Rojas Carulla、Giambattista Parascandolo、Moritz Hardt、Dominik Janzing 和 Bernhard Schölkopf。通过因果推理避免歧视。《神经信息处理系统进展》，第30卷，2017年。[21] Matt Kusner、Chris Russell、Joshua Loftus 和 Ricardo Silva。制定减少歧视性影响的决策。载于第*36*届国际机器学习会议论文集，第3591–3600页。机器学习研究论文集，2019年5月。URL https://proceedings.mlr.press/ v97/kusner19a.html。ISSN：2640-3498。

[22] Matt J Kusner、Joshua Loftus、Chris Russell 和 Ricardo Silva。反事实公平性。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html。[23] Matilde Lazzari、Jose M Alvarez 和 Salvatore Ruggieri。预测并解释员工离职意向。《国际数据科学与分析杂志》，14(3):279–292，2022年。

[24] Lihua Lei 和 Emmanuel J Candès。《英国皇家统计学会会刊B辑》发表的《反事实与个体治疗效应的保形推断》，83(5):911–938，2021年。

[25] Joshua R. Loftus、Chris Russell、Matt J. Kusner 和 Ricardo Silva。《算法公平性的因果推理》。*arXiv:1805.05859 [cs]*, 2018年5月。URL http://arxiv.org/abs/ 1805.05859。arXiv: 1805.05859。

[26] Karima Makhlouf、Sami Zhioua 和 Catuscia Palamidessi。关于基于因果的机器学习公平性概念的综述。*arXiv*预印本 *arXiv:2010.09553*，2020年。

[27] Christoph Molnar。《可解释机器学习：黑箱模型解释指南》。2022年。网址 `https://christophm.github.io/interpretable-ml-book/`。

[28] Raha Moraffah、Mansooreh Karami、Ruocheng Guo、Adrienne Raglin 和 Huan Liu。机器学习问题的因果可解释性：方法与评估。*ACM SIGKDD* 探索通讯，22(1):18–33，2020年。

[29] Razieh Nabi 和 Ilya Shpitser。关于结果的公平推断。载于*AAAI*人工智能会议论文集，第32卷，2018年。

[30] Judea Pearl。 [专家系统中的贝叶斯分析]：评论：图模型、因果关系与干预。《统计科学》，8(3):266–269，1993年。ISSN 08834237。URL http: //www.jstor.org/stable/2245965。

[31] Judea Pearl 和 Dana Mackenzie。《为什么之书：因果关系的新科学》。Basic Books，2018年。

[32] Judea Pearl 等。《模型、推理与推断》。英国剑桥：剑桥大学出版社，19(2)，2000年。

[33] Jonas Peters、Dominik Janzing 和 Bernhard Schölkopf。《因果推断要素：基础与学习算法》。麻省理工学院出版社，2017年。

[34] Rafael Poyiadzi、Kacper Sokol、Raul Santos-Rodriguez、Tijl De Bie 和 Peter Flach。FACE：可行且可采取行动的反事实解释。载于*AAAI/ACM*人工智能、伦理与社会会议论文集，第344–350页。计算机协会，纽约，纽约州，美国，2020年2月。ISBN 978-1-4503-7110-0。URL https://doi.org/10. 1145/3375627.3375850。

[35] Joseph Ramsey 和 Bryan Andrews。借助干预知识，Fask 从 Sachs模型中恢复边。*arXiv*, abs/1805.03108，2018年。

[36]Chris Russell、Matt J Kusner、Joshua Loftus 和 Ricardo Silva。《当世界碰撞：整合公平性中的不同反事实假设》。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。URL https://proceedings.neurips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html。[37] Karen Sachs、Omar Perez、Dana Pe'er、Douglas A. Lauffenburger 和 Garry P. Nolan。基于多参数单细胞数据构建的因果蛋白质信号网络。《科学》，308(5721)：523–529，2005年。doi：10.1126/science.1105809。URL https://www.science.org/doi/abs/10.1126/science.1105809。[38] Numair Sani、Daniel Malinsky 和 Ilya Shpitser。通过因果学习解释黑箱预测算法的行为。*arXiv*预印本 *arXiv:2006.02482*，2020年。

[39] B Schölkopf、D Janzing、J Peters、E Sgouritsa、K Zhang 和 J Mooij。论因果学习与反因果学习。载于第*29*届国际机器学习会议（*ICML 2012*），第1255–1262页。国际机器学习学会，2012年。[40] Amit Sharma 和 Emre Kiciman。Dowhy：用于因果推断的端到端库。*arXiv*预印本 *arXiv:2011.04216*，2020年。[41] Donghee Shin。可解释性与因果可解释性对感知、信任和接受度的影响：对可解释人工智能的启示。《国际人机研究杂志》，146：102551，2021年。[42] Peter Spirtes、Clark N Glymour、Richard Scheines 和 David Heckerman。因果、预测与搜索。麻省理工学院出版社，2000年。[43] Ilia Stepin、Jose M. Alonso、Alejandro Catala 和 Martín Pereira-Fariña。可解释人工智能的对比解释与反事实解释生成方法综述。*IEEE Access*，9：11974–12001，2021年。ISSN 2169-3536。doi：10.1109/ACCESS.2021.3051315。会议名称：IEEE Access。[44] W Nick Street、William H Wolberg 和 Olvi L Mangasarian。乳腺肿瘤诊断的核特征提取。载于生物医学图像处理与生物医学可视化，第1905卷，第861–870页。SPIE，1993年。[45] Berk Ustun、Alexander Spangher 和 Yang Liu。线性分类中的可行补救。载于公平、问责与透明度会议论文集，FAT* ’19，第10–19页，美国纽约州纽约，2019年1月。计算机协会。ISBN 978-1-4503-6125-5。doi：10.1145/3287560.3287566。URL https://doi.org/10.1145/ 3287560.3287566。[46] Ke Yang、Joshua R. Loftus 和 Julia Stoyanovich。因果交叉性与公平排序。载于Katrina Ligett 和 Swati Gupta 编，《第二届负责任计算基础研讨会（*FORC 2021*）》，第192卷莱布尼茨国际信息学会议录（*LIPIcs*），第7:1–7:20页，德国达格施图尔，2021年。达格施图尔城堡—莱布尼茨信息学中心。ISBN 978-3-95977-187-0。doi：10.4230/LIPIcs.FORC.2021.7。URL https: // drops.dagstuhl.de/opus/volltexte/2021/13875。ISSN：1868-8969。[47] Mingzhang Yin、Claudia Shi、Yixin Wang 和 David M Blei。个体治疗效应的保形敏感性分析。《美国统计学会杂志》，第1–14页，2022年。[48] Junzhe Zhang 和 Elias Bareinboim。决策中的公平性——因果解释公式。载于*AAAI*人工智能会议论文集，第32卷，2018年。[49] Qingyuan Zhao 和 Trevor Hastie。黑箱模型的因果解释。《商业与经济统计学杂志》，39(1)：272–281，2021年。doi：10.1080/07350015.2019.1624293。URL https://doi.org/10.1080/07350015.2019.1624293。

补充材料：因果依赖图

6 因果依赖图算法

本节将介绍用于计算正文中定义的各类因果依赖图的算法。

算法 1解释性因果模型（ECM）

输入：$\mathcal{M}_{\mathbf{X}}$（结构因果模型）$\hat{f}$（黑箱预测器），$\mathbf{S} \subseteq \mathbf{X}$（黑箱使用的协变量）

复制 $\mathcal{M}'$ 结构因果模型 $\mathcal{M}_{\mathbf{X}}$
添加节点 $\hat{Y}$到结构因果模型 $\mathcal{M}'$的因果图 $\mathcal{G}$
对 x 在 **S 中执行**
  添加边：从 $\mathcal{G}$ x 到 $\hat{Y}$
结束循环
设置节点的结构方程 $\hat{Y}$ to $\hat{f}$
设置外生变量 $U_{\hat{Y}} \leftarrow 0$
返回 $\mathcal{M}'$

算法 2总依赖图（TDP）

输入：$\mathcal{M}_{\mathbf{X}}$（结构因果模型）$\hat{f}$（黑箱预测器），D（解释数据集），X_s（感兴趣的协变量）

通过算法 1 获取 ECM $\mathcal{M}$
获取 X_S 的可能取值并将其设为 X
将 N 设为 D中的观测值数量
初始化 $N \times |X|$ 估计矩阵 $\hat{Y}$
对 x中的 X **执行**
  定义干预 $I = \text{do}(X_S = x)$
  抽取由 $P^{\mathcal{M}|D;\text{do}(I)}$蕴含的反事实数据集 D_c
  Set $\hat{Y}[:, x] \leftarrow D_c[:, y]$ 对应节点的索引 y $\hat{Y}$
结束循环
绘制 N 线条（$X, \hat{Y}[i,:]$） ▷ (个体反事实)
绘制平均值（X，$\sum_i \hat{Y}[i,:]/N$) ▷ (因果依赖)

算法3部分控制依赖图（PCDP）

输入：$\mathcal{M}_{\mathbf{X}}$(结构因果模型), $\hat{f}$ (黑箱预测器), D (解释数据集), X_s (感兴趣的协变量), C (对 $\mathcal{M}_{\mathbf{X}}$中的其他变量进行控制的干预)

通过算法 1 获取 ECM $\mathcal{M}$
获取 X_S 可能的取值，并设为 X
将 N 设为 D 中观测值的数量
初始化 $N \times |X|$ 估计矩阵 $\hat{Y}$
对于 x中 X，**执行**
  定义干预 $I = \text{do}(X_S = x,\ C)$
  由 $P^{\mathcal{M}|D;\text{do}(I)}$ 推导出的反事实数据集示例 D_c
  Set $\hat{Y}[:, x] \leftarrow D_c[:, y]$ 对于对应节点的索引 y $\hat{Y}$
结束循环
绘制 N 线条（$X, \hat{Y}[i,:]$） ▷ (个体反事实)
绘制平均值（$X, \sum_i \hat{Y}[i,:]/N$） ▷ （因果依赖）

算法 4自然直接依赖图（NDDP）
输入：$\mathcal{M}_{\mathbf{X}}$（结构因果模型）$\hat{f}$（黑箱预测器），D（解释数据集），X_s（感兴趣的协变量）

通过算法 1 获取 ECM $\mathcal{M}$
获取 X_S 的可能取值并设为 X
将 N 设为 D 中观测值的数量
初始化 $N \times |X|$ 估计矩阵 $\hat{Y}$
获取 X_S 中除以下节点之外的所有子节点 $\hat{Y}$，并存储在$\mathbf{C}$中
复制 $\mathcal{M}'$ 结构因果模型 $\mathcal{M}$
对 x in **C 执行**
    删除从 $\mathcal{M}'$指向 x 的所有入边
结束循环
对于 x属于 X **执行**
    对于 i 属于 N **执行**
        获取个体 i 的$\mathbf{C}$ 中所有变量的观测值，并存储在 $\mathbf{c}$中 $_i$
        定义干预 $I = \text{do}(X_S = x, \mathbf{C} = \mathbf{c}_i)$
        采样由 d_c 个体 i 蕴含的反事实观测 $P^{\mathcal{M}'|D[i];\text{do}(I)}$
        Set $\hat{Y}[i, x] \leftarrow d_c[y]$ 对于对应于节点的索引 y $\hat{Y}$
    结束循环
结束循环
绘制 N 条线（X，$\hat{Y}[i,:]$） ▷（个体反事实）
绘制平均值（X，$\sum_i \hat{Y}[i,:]/N$） ▷（因果依赖）

算法 5 自然间接依赖图（NIDP）ˆ 输入：$\mathcal{M}_{\mathbf{X}}$（结构因果模型），f（黑箱预测器），D（解释数据集），X_s（感兴趣的协变量）通过算法 1 获取 ECM $\mathcal{M}$获取 X_S 的可能取值，并将其设为 X将 N 设为 D中的观测值数量初始化 $N \times |X|$ 估计矩阵 ˆ$_Y$ˆ 获取 X_S 在 $\mathcal{M}$中的所有子节点（排除 Y），并将其存储在 C 中复制结构因果模型 $\mathcal{M}$，得到 $\mathcal{M}'$ 对于 C 中的 x，执行：移除 x 的所有入边，这些入边来自 $\mathcal{M}'$结束循环定义干预 $I = \text{do}(X_S = x)$对于 X 中的 x，执行：对于 N 中的 i，执行：采样由 $P^{\mathcal{M}|D[i];\text{do}(I)}$蕴含的个体 i 的反事实观测值 d_c 从观测值 d_c 中获取 C 中所有变量的反事实值，并将其存储在 $\mathbf{c}_i$中定义干预 $J = \text{do}(X_S = x, \mathbf{C} = \mathbf{c}_i)$采样由 $P^{\mathcal{M}'|D[i];\text{do}(J)}$蕴含的个体 i 的反事实观测值 d'_c 将索引 y 处对应于节点$_Y$的值设为 ˆ $Y[i, x] \leftarrow d'_c[y]$ 结束循环结束循环绘制 N线（X, ˆ$Y[i,:]$）▷（个体反事实）ˆ 绘制平均值（X, $\sum_i Y[i,:]/N$）▷（因果依赖）

7 使用因果发现分析真实数据

本节将探讨与主文中有向无环图 $\mathcal{G}_B$ 中的不确定边相一致的图结构，该图基于威斯康星州乳腺癌数据集。

图 S1 展示了一个学习得到的加性噪声模型（ANM）的 TDP、NDDP 和 NIDP。该模型有三种与 $\mathcal{G}_B$ 一致的不同结构：（1）一条连接细胞形状 $\rightarrow$ 细胞大小的边；（2）一条连接细胞大小 $\rightarrow$ 细胞形状的边；（3）细胞大小与

细胞形状。该图表明，关于细胞形状影响肿瘤类别这一结论，确实对我们如何假设不确定边十分敏感，尤其是在总依赖的情况下。

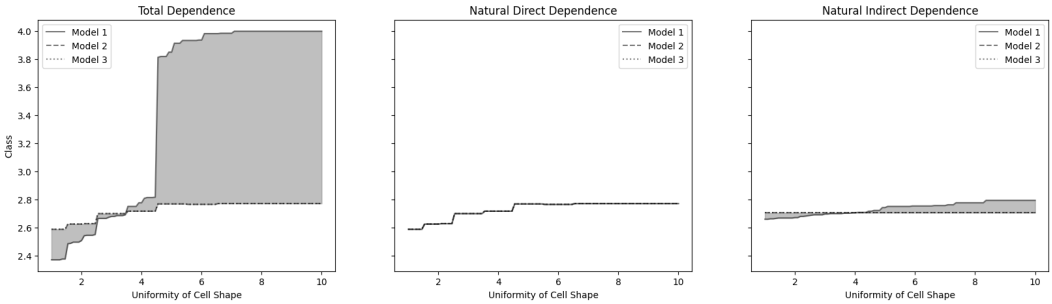


图 S1：在 PC 算法找到的三种可能有向无环图下，威斯康星乳腺癌数据集的总依赖图、自然直接依赖图和自然间接依赖图：(1) $\mathcal{G}_B$，其中有一条从细胞形状指向细胞大小的边 $\rightarrow$；(2) $\mathcal{G}_B$，其中有一条从细胞大小指向细胞形状的边 $\rightarrow$；(3)$\mathcal{G}_B$，其中细胞大小与细胞形状之间没有边。

8 具有领域专业知识的真实数据

作为一个利用领域专业知识构建底层因果模型的示例，图 S2 展示了 Sachs 等人 [37] 数据集的有向无环图、TDP 和 NDDP。该数据集的数据和真实有向无环图[1]已在因果发现工具箱 [17]中公开。尽管这里关注的重点并非问题本身的生物学，但图 S2 揭示了重要的含义。根据该模型，TDP 显示两个特征之间的关系与 NDDP 所显示的关系有着显著差异。请记住，NDDP 捕捉到的关系与部分依赖图或个体条件期望图所呈现的关系相同。如果我们不考虑图中其他变量会如何随着 PKA 的变化而改变，那么所发现的趋势实际上会完全相反。

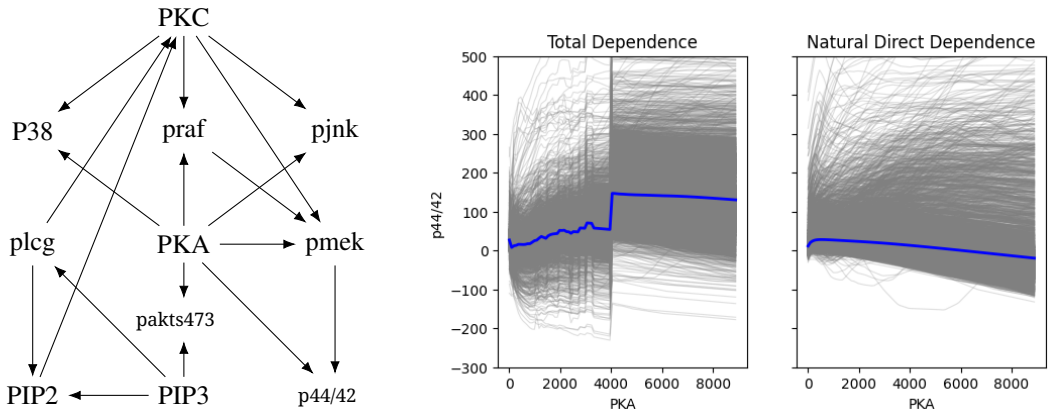


图 S2：Sachs 等人的数据集对应的有向无环图 $\mathcal{G}_S$，以及 PKA 对 p44/42 影响的相应总依赖图和自然直接依赖图 [37]。

[1]遵循 Ramsey 和 Andrews [35] 中的讨论，以及 Ramsey 和 Andrews 的图 5 中 Sachs 等人 [37] 数据集后续公布的真实有向无环图，我们选择 PIP3 $\rightarrow$ PIP2 这条边来消除一个可能形成的环，其余部分则保持发布的有向无环图不变。