arXiv:2304.03265v1 [cs.LG] 6 Apr 2023

具有任意噪声的加性模型中的因果发现与得分匹配

Francesco Montagna* FRANCESCO.MONTAGNA@EDU.UNIGE.IT
MaLGa-DIBRIS，热那亚大学
尼科莱塔·诺切蒂 NICOLETTA.NOCETI@UNIGE.IT
MaLGa-DIBRIS，热那亚大`学
Lorenzo Rosasco LROSASCO@MIT.EDU
MaLGa-DIBRIS，热那亚大`学
MIT，*CBMM* 意大利理工学院

Kun Zhang KUNZ1@CMU.EDU
卡内基梅隆大学 *MBZUAI*

Francesco Locatello LOCATELF@AMAZON.COM
AWS

编辑： Mihaela van der Schaar、Dominik Janzing 和 Cheng Zhang

摘要

因果发现方法本质上受到确保结构可识别性所需假设集合的限制。此外，为了简化推断任务，通常还会施加额外限制：对加性非线性模型采用高斯噪声假设便是其中一种，这一假设在许多因果发现方法中都很常见。本文分析了在该假设下进行推断的缺陷，研究了噪声项违反高斯性时边反转的风险。随后，我们提出了一种新方法，用于根据由具有通用噪声分布的加性非线性模型生成的数据，推断因果图中变量的拓扑排序。这一方法形成了 NoGAM（不仅限于高斯加性噪声模型），一种仅需最小假设集合且性能达到当前先进水平的因果发现算法，并通过合成数据进行了实验基准测试。

关键词： 因果发现；任意噪声分布；得分匹配

1. 引言

从观测数据中推断因果关系是《因果关系》研究的核心目标，因为这使得我们能够对系统干预进行形式化推理（Peters 等人(2017)，Pearl (2009年)），尤其是在干预成本高昂、不合伦理，甚至无法实施时。结构可识别性研究结果提出了仅凭联合分布中的纯观测数据能够推断因果图哪些部分的界限，并为完整识别数据背后的因果图所需的假设提供了形式化指导。传统的因果发现方法通常仅限于识别马尔可夫等价类（Glymour 等人 (2019)；PC、FCI（Spirtes 等人(2000)以及 GES（Chickering (2002）即属此类。近年来，人们提出了基于明确定义的结构因果模型（结构因果模型）的方法，以区分观测

* 部分工作是在图宾根亚马逊云科技实习期间完成的

© 2023 F. Montagna、N. Noceti、L. Rosasco、K. Zhang 和 F. Locatello。

数据，通过对结构因果模型的函数类施加额外假设来实现：Hoyer 等人（2009）和 Zhang and Hyv¨arinen（2009）表明，非线性加性噪声模型通常具有可识别性。SCORE（Rolland 等人（2022））和 CAM（B¨uhlmann 等人（2014））便是如此：在高斯扰动的假设下，它们能在推断过程中得到唯一且渐近一致的图。在非线性加性模型可识别的条件下，Peters 等人（2014）和 Mooij 等人（2009）展示了如何利用估计残差的独立性来推断因果效应，而不对噪声分布施加限制。这些方法受限于条件独立性检验的使用，而这类检验难以实施（Shah and Peters（2018））。与我们的工作更为接近的是，Bloebaum 等人（2018）通过比较回归误差来区分原因与结果，但仅适用于受限的双变量模型。其他方法，如 Lachapelle等人（2020）和 Zheng 等人（2018），则将问题表述为连续优化问题，从而得到唯一的有向无环图（DAG）。

一般而言，若要从观测数据中推断因果结构并获得理论保证，可识别性结果需要依赖一些假设。这种方法的不足之处在于，以假设形式提出的约束会缩小算法的适用范围。相比之下，我们希望有一些方法能够在不同条件下适用于广泛的问题，并且最好能对违反模型假设的情况表现出一定程度的鲁棒性。这一观点的优势在深度学习实践中显而易见：主流做法是应用能够解决目标任务的算法，而不论其基础假设是否遭到违反。本文旨在调和这两种理念，提供一种因果发现工具：从可识别性的角度出发，去除一些最突出的计算方法常常采用的不必要假设。为此，我们设计了一种算法，用于推断加性非线性模型背后的因果图，该模型具有一般形式的噪声项，并且不再采用常见的高斯分布假设。去除这些约束拓宽了有原则的因果发现方法的适用范围，也为重视理论保证、且工作于难以验证高斯噪声假设是否成立的关键场景中的实践者，提供了一种先进方法。

本文其余部分安排如下：第 2 节概述所研究的模型，并定义当前问题；第 3 节分析违反高斯噪声假设时推断反向边的风险；第 4 节介绍一种理论依据严谨的方法，通过迭代识别因果图的叶节点来确定其拓扑排序；具体而言，我们证明了得分函数（即对数似然的梯度）与残差估计量之间的重要关系；第 5 节定义 NoGAM[1]，，这是一种根据数据推断因果图的算法；第 6 节概述该方法相较于经典基准方法和最先进基准方法的实验表现。

2. 背景知识

模型定义 令$\mathbf{V} = \{1, \dots, d\}$ 为有向无环图 $\mathcal{G}$的顶点，令$\mathbf{X} \in \mathbb{R}^d$为根据结构因果模型（结构因果模型）生成的一组随机变量

$$X_i := f_i(\mathrm{PA}_i(\mathbf{X})) + N_i, \quad \forall\, i \in \mathbf{V}\,, \tag{1}$$

1. NoGAM 的代码作为 DoDiscover 库的一部分提供，网址为 https://www.pywhy.org/dodiscover/dev/index.html

其中，PA$_i$(X)是图 $\mathcal{G}$ 中节点 i 的父节点向量。模型 (1)被称为非线性加性噪声模型（ANM，Hoyer等人(2009年)）。我们假设因果机制 f_i 是连续可微的非线性函数，并满足因果最小性。噪声项$N_i \in \mathbb{R}$ 是具有密度 $p_i(N_i)$、均值 $\mu_i = 0$ 和方差 $\sigma_i^2 > 0$ 的连续随机变量。我们假设它们相互独立，使得其联合分布为 $p_{\mathbf{N}}(\mathbf{N}) = \prod_i p_i(N_i)$。在这些假设下，模型 (1)导出一个关于 $\mathcal{G}$满足马尔可夫性质的联合分布 $p_{\mathbf{X}}(\mathbf{X})$，从而它具有如下因子分解:

$$p_{\mathbf{X}}(\mathbf{X}) = \prod_i^d p_i(X_i \mid \mathrm{PA}_i(\mathbf{X}))\,, \tag{2}$$

其中，略微滥用记号，我们通过参数将边缘分布 $p_i(N_i)$与 $p_i(X_i \mid \mathrm{PA}_i(\mathbf{X}))$区分开。

在本文的其余部分，我们将使用 X_i 同时表示随机变量和相应的节点 $i \in \mathbf{V}$。

可识别性假设为了确保能够根据观测数据识别因果图，使得了解$\mathbf{X}$ 的联合分布足以区分底层因果图中的原因和结果，我们需要对函数机制 f_i 以及 ANM 的噪声项分布作出额外假设。下文将针对如Peters 等人（2014年）所示的二元图给出可识别性条件。这些条件可以无缝推广到多变量情形，相关讨论见附录A。

条件 *1*（*Peters* 等人（*2014*）的条件 *19*）给定一个双变量模型 $X_i := N_i$ *and* $X_j := f_j(X_i) + N_j$ *with* $\{i, j\} = \{1, 2\}$，其中 $\{i, j\} = \{1, 2\}$ 根据 *(1)* 生成，若三元组 (f_j, p_{N_i}, p_{N_j}) 不满足以下微分方程（对所有满足 $f_j'(x_i)g''(x_j - f_j(x_i)) \neq 0$ 的 x_i, x_j），则称该结构方程模型为可识别的双变量加性噪声模型:

$$k''' = k''\left(-\frac{g'''f'}{g''} + \frac{f''}{f'}\right) - 2g''f''f' + g'f''' + \frac{g'g'''f''f'}{g''} - \frac{g'(f'')^2}{f'}\,. \tag{3}$$

其中，$f := f_j$, $k := $对数 p_{N_i}, $g := $对数 p_{N_j}。为提高可读性，已省略 g、k和 f各自的参数 $x_j - f_j(x_i)$、x_i和 x_i。

Hoyer 等人 (2009年)首先证明，如果满足条件 1，那么与双变量加性噪声模型相关联的图可以从联合分布中识别出来 $p_{\mathbf{X}}$ (定理20Peters 等人 (2014年)。

直观而言，我们预期条件1 对一般的三元组(f_j, p_{N_i}, p_{N_j})成立。更形式化地说，这是因为，对于固定的一对(f_j, p_{N_j})，使条件 p_{N_i}1不成立的连续分布空间包含在一个三维空间中。由于连续分布空间是无限维的，因此我们可以说，条件1对大多数p_{N_i}选择成立。严谨的论述参见Peters 等人(2014年)。

在实践中，如果噪声项具有处处严格为正的密度 p_{N_i}, p_{N_j} 并且

$$f_j'(x_i)g''(x_j - f_j(x_i)) \neq 0 \tag{4}$$

除有限个点之外的所有(x_i, x_j) (Zhang 和 Hyv¨arinen (2009年Peters 等人(2014年）。此外，我们明确指出，条件 1 蕴含

$$\partial_{x_i}(k'(x_i) - f_j'(x_i)g'(x_j - f_j(x_i)) \neq 0\,, \tag{5}$$

对于所有满足 $f_j'(x_i)g''(x_j - f_j(x_i)) \neq 0$ 的 x_i, x_j。这可以直接通过 Peters 等人（2014）中定理20的证明（公式14）验证：若（5）的左侧为零，则对于满足 $f_j'(x_i)g''(x_j - f_j(x_i)) \neq 0$ 的 x_i, x_j，公式（3）总是成立，这与可识别性的条件相矛盾。

在本文余下部分中，我们假设结构因果模型（A）满足这些要求（以及附录中讨论的多变量情形下的对应要求 1）。

拓扑排序定义 设 $\mathcal{G} = (\mathbf{X}\ \mathcal{E})$ 是一个有向无环图。节点 $\mathbf{X}^{\pi} = X_{\pi_1}, \ldots\ldots, X_{\pi_d}$ 是相对于 $\mathcal{G}$ 的拓扑排序，如果每当有 $X_{\pi_i} \to X_{\pi_j} \in \mathcal{E}$，则 $i < j$ (Koller和 Friedman(2009年)。

问题定义给定来自联合分布的独立同分布样本 $p_{\mathbf{X}}$，我们希望推断生成这些数据的模型所对应的真实因果图 $\mathcal{G}$。解决这一问题的一种方法是将任务分为两个步骤：首先，我们为图中的顶点找到一个拓扑排序，然后对全连接图进行剪枝；该全连接图通过在排序中从每个节点向其所有后继节点绘制一条边而得到。在本文中，我们使用得分函数 $\nabla \log p_{\mathbf{X}}(\mathbf{X})$ 来提出一种一致的拓扑排序推断方法，该方法无需对噪声项的具体分布作出任何假设 N_i，如方程（1）中所示。

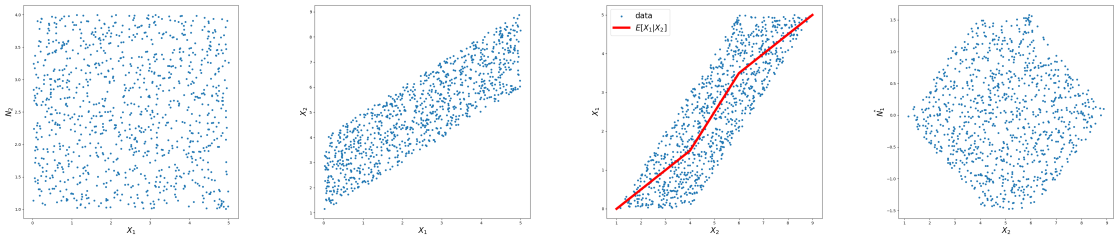
实际上，经典的和最先进的因果发现算法，如 SCORE（Rolland 等人（2022））、CAM（B¨uhlmann 等人（2014年））、GES（Chickering （2002））和 GraN-DAG（Lachapelle 等人（ 2020 年））都假设噪声项服从正态分布。在下一节中，我们将说明这一假设的局限性，以及当该假设不成立时如何错误地推断拓扑排序，从而导致估计出包含反向边的图。

3. 高斯噪声假设的局限性

已有研究表明，对于由非线性模型和加性噪声生成的数据，一般而言，两个变量之间的因果方向是可识别的，因为在反向中，无法找到独立残差（Hoyer 等人(2009年）；ZhangandHyv¨arinen (2009年））， 因此，反向回归模型（假设残差相互独立）给出的数据似然低于因果方向上的数据似然。可识别性结果（Zhang andHyv¨arinen (2009年）表明，如果因果模型中的噪声项服从高斯分布，而函数是非线性的，那么两个变量之间的因果方向是可识别的。也就是说，正确方向上的回归模型似然高于反向上的似然；或者等价地，因果方向上估计得到的噪声项（包括假设的原因变量）的总熵更小（ZhangandHyv¨arinen (2009年））。

然而，这一结果并不意味着，在假设噪声分布为高斯分布时，正确的因果方向总能使估计出的噪声项具有更高的似然或更低的总熵。

命题1给定一个真实值可识别的非线性加性因果模型，任何基于观测数据且错误地假设噪声项服从高斯分布的推断算法，都无法保证恢复底层图中边的正确方向。



(a) X_1 和 N_2 的散点图(b) X_1 和 X_2 的散点图(c) X_2 和 X_1 的散点图(d) X_2 和估计的 N_1的散点图以及回归曲线

图 1：用于说明图 $X_1 \to X_2$上高斯噪声假设局限性的示例。

在本文余下部分，我们将命题 1 视为成立，并通过一个示例证明我们的主张。我们从含加性均匀噪声的双变量模型中生成数据，并按照 Zhang 和 Hyv¨arinen (2009年) 提出的方法，通过比较残差的总熵来推断因果效应。我们表明，假设噪声项服从高斯分布会导致该方法失效，造成推断方向相反。有关该示例实验设计的更多细节，请参阅附录 E。

示例1 我们首先为线性加性噪声模型定义一个示例，使得可以找到当前回归问题的闭式解。然后，我们将该示例推广到非线性情形。

令 $X_1 \to X_2$ 满足 $X_2 := X_1 + N_2$，其中 X_1 和 N_2 均服从均匀分布，如图 *1a*所示。图 *1b* 展示了 X_1 和 X_2的散点图。图 *1c*给出了 X_2和 X_1的散点图，以及回归曲线；在此例中，该曲线是分段线性的。图 *1d*ˆ展示了 X_2 和 N_1的散点图，其中后者是反向回归方向上的估计回归噪声。

可以通过比较估计噪声项的总熵来找到正确的因果方向。令 $H(\cdot)$ 表示微分熵。可以计算得到

$$H(X_1) + H(N_2) = 2.708 < H(X_2) + H(\hat{N}_1) = 2.954$$

ˆ其中 $H(X_1)$ $H(X_2)$ 是利用边缘分布的知识计算得出的。 $H(N_1)$ 是ˆ根据数据估计得出的，而我们用 $H(N_2)$ 代替 $H(N_2)$ 的估计值，因为使用精确的熵可以使计算更加精确。令 $H_{\mathcal{G}}(\cdot)$ 表示高斯性假设下的微分熵。于是可以计算得出

$$H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) = 3.061 > H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(\hat{N}_1) = 3.043.$$

也就是说，在高斯性假设下，反向因果方向会得到更低的总熵（或等价地，更高的似然），因此会推断出错误的因果方向。
为什么会发生这种情况？请注意，对于方差固定的变量，高斯分布具有最高的微分熵。因此，很自然地， $H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) > H(X_1) + H(N_2)$ 和ˆ ˆ$H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(N_1) > H(X_2) + H(N_1)$。此外，完全有可能（在这个例子中显然ˆ如此），与原始的相互独立变量 X_1 和 N_2相比， X_2 和 N_1分别更接近高斯分布。也就是说，高斯性假设所引起的变化

估计噪声项的总熵在反向方向上更小。因此，在高斯分布假设下，与使用其真实分布的情况相反，反向方向可能会得到更低的估计噪声项总熵。

现在考虑非线性模型 $X_2 := X_1^{1+\delta} + N_2$，其中 $\delta > 0$，且噪声项 X_1 和 N_2均匀分布。显然，对于 $\delta = 0$，这等同于前面讨论过的线性示例：这里我们将 $\delta = 0.1$ 设为一个值，以在生成过程中引入弱非线性。同样，可以通过比较正确方向和反向方向下估计噪声项的总熵，来识别真实因果方向。可以计算得出，

$$H(X_1) + H(N_2) = 2.708 < H(X_2) + H(\hat{N}_1) = 2.926 \,.$$

同样，在高斯性假设下可得

$$H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) = 3.061 > H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(\hat{N}_1) = 3.001 \,,$$

表明命题 *1* 成立。

鉴于高斯假设的局限性，我们现在提出一种适用于具有一般噪声项的加性非线性模型的因果发现方法，从而保证推断出的因果排序（以及边的方向）与因果图一致。

4. 通过得分函数进行因果发现

本节将推导一种有理论依据的方法，通过得分函数识别叶节点 $s(\mathbf{X}) = \nabla \log p_{\mathbf{X}}(\mathbf{X})$，且无需对生成模型（1）中噪声随机变量的分布作任何假设$\mathbf{X}$。

4.1. 数据分布的得分函数

给定分布 $p_{\mathbf{X}}(\mathbf{X})$ 由模型（1）诱导，我们可以将得分函数向量定义为 $s(\mathbf{X}) = \nabla \log p_{\mathbf{X}}(\mathbf{X})$。利用公式（2）中联合分布的因子分解，我们可以写为：

$$\log p_{\mathbf{X}}(\mathbf{X}) = \sum_{i=1}^{d} \log p_i(X_i \mid \mathrm{PA}_i) \,. \tag{6}$$

使得 SCORE 的单个条目 $s_i(\mathbf{X})$等于：

$$s_i(\mathbf{X}) = \partial_{X_i} \log p_i(X_i \mid \mathrm{PA}_i) + \sum_{k \in \mathrm{CH}_i} \partial_{X_i} f_k(\mathrm{PA}_k) \partial_{X_k} \log p_k(X_k \mid \mathrm{PA}_k) \,. \tag{7}$$

在父节点条件化下，X_i 的边缘分布与 N_i 的分布相同，只是偏移了机制 $f_i(\mathrm{PA}_i)$ 的值，这意味着 $p_i(X_i \mid \mathrm{PA}_i)$可以替换为 $p_i(N_i = X_i - f_i(\mathrm{PA}_i) \mid \mathrm{PA}_i)$。因此，SCORE 可以改写为：

$$s_i(\mathbf{X}) = \partial_{N_i} \log p_i(N_i) - \sum_{k \in \mathrm{CH}_i} \partial_{X_i} f_k(\mathrm{PA}_k) \partial_{N_k} \log p_k(N_k) \,. \tag{8}$$

然后，对于每个 $i=1$、…、d，我们定义函数 $g_i(N_i) \coloneqq \log p_i(N_i)$，使得第 i 个评分项为

$$s_i(\mathbf{X}) = \partial_{N_i} g_i(N_i) - \sum_{k \in \mathrm{CH}_i} \partial_{X_i} f_k(\mathrm{PA}_k) \partial_{N_k} g_k(N_k) \,. \tag{9}$$

对于叶节点 X_l，得分方程（9）变为

$$s_l(\mathbf{X}) = \partial_{N_l} g_l(N_l) \,, \tag{10}$$

这表明，如果已知 g_l，就可以根据噪声 N_l预测叶节点 $s_l(\mathbf{X})$。在下一节中，基于这一思路，我们为每个变量 X_i定义一个回归问题，其中输入变量是其余所有项$\mathbf{X}^{\setminus\{i\}} := \mathbf{X} \setminus \{X_i\}$：随后我们证明，如果 X_i 预测目标是叶节点，那么该学习问题的残差是模型（1）对应结构方程中噪声项的一致估计量。

4.2. 残差估计

给定来自 $p_{\mathbf{X}}$ 的独立同分布样本 $X \in \mathbb{R}^{n \times d}$，对于每个 $i=1$，……，d，我们定义一个回归问题，用其余变量预测 X_i $\mathbf{X}^{\setminus\{i\}}$:

$$\begin{aligned} &\min_{q \in \mathcal{Q}} L(q), \;\; L(q) = \int_{\mathbb{R}^d} (q(\mathbf{X}_{\setminus\{i\}}) - X_i)^2 dp_{\mathbf{X}}(\mathbf{X}) \\ &\text{given } \mathcal{D} = \left\{ \left(\mathbf{X}^k_{\setminus\{i\}}, X^k_i \right) \right\}_{k=1}^n \end{aligned} \,, \tag{11}$$

其中，$\mathcal{Q}$ 是从输入空间到 $\mathbb{R}$ 的可测函数空间。对于输入空间中的所有 $\mathbf{X}^{\setminus\{i\}}$，$L$ 的极小化者 $q^* \in \mathcal{Q}$ 是条件期望 $\mathbf{E}[X_i \mid \mathbf{X}_{\setminus\{i\}}]$，根据均值的线性性质，这等价于:

$$q^*(\mathbf{X}_{\setminus\{i\}}) = \mathbf{E}[f_i(\mathrm{PA}_i) \mid \mathbf{X}_{\setminus\{i\}}] + \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,. \tag{12}$$

鉴于 $\mathrm{PA}_i \subset \mathbf{X}^{\setminus\{i\}}$，我们可以直接从求和的第一项中去掉期望算子，得到:

$$q^*(\mathbf{X}_{\setminus\{i\}}) = f_i(\mathrm{PA}_i) + \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,. \tag{13}$$

现在，我们将式（11）中学习问题的残差定义为响应与目标函数之间的差:

$$\begin{aligned} R_i &\coloneqq X_i - q^*(\mathbf{X}_{\setminus\{i\}}) \\ &= N_i - \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,, \end{aligned} \tag{14}$$

其中，第二个等式由公式（13）得出。注意到 $\mathbf{X}^{\setminus\{i\}} = \mathrm{DE}_i \cup \mathrm{ND}_i$，其中 DE_i 和 ND_i 分别表示节点 X_i的后代节点集和非后代节点集：于是，公式（14）可改写为

$$R_i = N_i - \mathbf{E}\left[N_i \mid \mathrm{DE}_i \cup \mathrm{ND}_i\right] \,. \tag{15}$$

对于一个叶节点 X_l 我们可以利用以下事实 $\mathrm{DE}_l = \emptyset$ 来简化上述残差表达式中的 $R_l = N_l - \mathbf{E}\left[N_l \mid \mathrm{ND}_l\right]$此外，我们可以使用 d-分离准则得出 N_l 与 ND_l无条件独立，如图 2所示：误差的期望

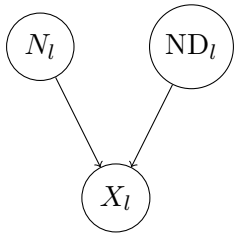


图 2：考虑一个叶节点 X_l，其根据（1）中定义的结构因果模型生成：根据 d-分离，噪声项 N_l 与非后代节点集合 ND_l无条件独立，因为这两个节点之间的唯一路径包含一个碰撞点，即 X_l。

叶节点的项随后为 $\mathbf{E}\left[N_l \mid \mathrm{ND}_l\right] = \mathbf{E}\left[N_l\right]$。最后，在模型（1）的噪声为零均值的假设下，我们得出方程（15）的残差为

$$R_l = N_l \,. \tag{16}$$

利用这一等价关系，我们可以将公式（10）中叶节点的 SCORE 重写为 R_l 的函数，使得评分项 s_l 满足：

$$\boxed{s_l(\mathbf{X}) = \partial_{N_l} g_l(R_l)} \,. \tag{17}$$

将方程（14）中的残差代入方程（9），我们可以推导出与非叶节点 s_i 相关的评分项 X_i：

$$s_i(\mathbf{X}) = \partial_{N_i} g_i(R_i + \mathbf{E}\left[N_i \mid \mathbf{X}_{\setminus\{i\}}\right]) - \sum_{k \in \mathrm{CH}_i} \partial_{X_i} f_k(\mathrm{PA}_k) \partial_{N_k} g_k(R_k + \mathbf{E}\left[N_k \mid \mathbf{X}_{\setminus\{k\}}\right])) \,. \tag{18}$$

讨论 对于叶节点 X_l，公式（17）表明，相应的评分项是单个变量的函数，即残差 R_l：这表明，我们可以期望从这样的残差中一致地估计评分 s_l（$\mathbf{X}$）。另一方面，如果考虑非叶节点 X_i，其评分 s_i（$\mathbf{X}$）的形式更为复杂（即取决于更多的变量），如公式（18）所示：直观地看，我们可以看到 R_i作为预测变量，并不足以找到 s_i（$\mathbf{X}$）的一致估计量。在下一节中，我们希望将这些直觉形式化，从而能够通过考察由相应残差对评分项进行预测时的误差，推导出一种具有理论依据的识别叶节点的方法。

4.3. 识别叶节点

考虑一个叶节点 X_l：给定一组独立同分布观测值 $\left\{(R_l^k,\ s_l(\mathbf{X}^k))\right\}_{k=1}^n$，我们希望以残差作为输入，找到一个得分估计量。与（11）类似，我们定义一个回归问题

$$\min_{h \in \mathcal{H}} L(h), \quad L(h) = \int_{\mathbb{R} \times \mathbb{R}^d} (h(R_l) - s_l(\mathbf{X}))^2 dp(R_l, \mathbf{X}) \,, \tag{19}$$

其中 $\mathcal{H}$ 是从输入空间到输出空间的可测函数集合。对于每个 $R_l \in \mathbb{R}$，使期望风险最小化的目标函数为

$$\begin{aligned} h^*(R_l) &:= \mathbf{E}\left[s_l(\mathbf{X}) \mid R_l\right] \\ &= \mathbf{E}\left[\partial_{N_l} g_l(R_l) \mid R_l\right] \,. \end{aligned} \tag{20}$$

显然，由于我们是在 R_l的条件下进行计算，因此可以去掉期望算子，得到：

$$h^*(R_l) = \partial_{N_l} g_l(R_l),\ \forall R_l \in \mathbb{R}\,. \tag{21}$$

然后，给定一个样本（$R_l, s_l(\mathbf{X})$），预测 $h^*(R_l)$ 与真实值 $s_l(\mathbf{X}) = g_l(R_l)$ 之间的差值为

$$\boxed{h^*(R_l) - s_l(\mathbf{X}) = 0}\,. \tag{22}$$

类似地，可以为非叶节点 X_i 定义式（19）的回归问题：由此得到的回归函数为 $h^*(R_i) = \mathbf{E}\left[s_i(\mathbf{X}) \mid R_i\right]$，即式（18）的条件期望。现在，我们可以利用这些结果定义叶节点识别准则。

引理 1 设 $\mathbf{X}$ 是根据模型生成的随机向量 (1)，并令 $X_i \in \mathbf{X}$。则

$$\mathbf{E}\left[\left(h^*(R_i) - s_i(\mathbf{X})\right)^2\right] = 0 \Leftrightarrow X_i \text{ is a leaf.}$$

讨论 根据公式（22），我们表明可以找到一个一致估计量 h^*，该估计量能够精确预测与叶节点相关的得分函数 X_l，前提是我们观测到残差 R_l。一般而言，对于节点 X_i（其不是叶节点），情况并非如此。通过考虑在所有输入和输出实现上的预测平方误差均值，引理 1将这一认识形式化。该引理的证明见附录 B。

基于本节的结果，我们现在介绍一种用于因果发现的算法，该算法适用于噪声项分布一般的非线性加性模型。随后，我们将在合成数据上将其性能与若干现有方法进行比较。

5. 方法

在第4.3节中，我们展示了如何在样本数量一致增长的情况下，识别由模型（1）生成的观测值所对应的因果图中的叶节点：其思想是，给定一组观测值，对于每个节点 $i = 1$、.....、d，我们可以根据相应的残差 R_i预测得分 s_i(X)，并将泛化误差最小的条目所对应的节点 l 选为叶节点。识别出一个叶节点后，将其从图中移除，并迭代地重复该过程，直至源节点，从而推断出渐近一致的拓扑排序X^{π}。实际上，给定有限的观测值集合 $X \in \mathbb{R}^{n\times d}$，我们首先使用Stein梯度估计器（Li 和 Turner（2017），详见附录D）估计数据的 ∇log p(X)得分函数，其输出为向量ˆs(X)。然后，我们通过核岭回归估计残差R $=$(R_1、.....、 R_d)，求解公式（11）中定义的问题。接下来，对于每个变量 $i = 1$、.....、 d，我们定义向量估计器ˆ ˆ~s(R)，根据 R_i预测ˆs_i(X)：为避免过拟合，我们通过K折交叉验证训练K个不同的模型（即仅使用部分观测值进行训练），每个模型都在训练期间未见过的相应测试集上进行预测。最后，我们计算预测值 ~s(R)与SCORE输出所提供的真实值ˆs(X)之间的均方误差（MSE）：选择MSE向量中argmin所对应的节点作为叶节点。重复这一过程，使得每次迭代都识别出一个叶节点，并将其加入拓扑排序估计中。算法1的框中给出了所述算法实现的更多细节。

给定算法1所估计的排序，我们使用一种剪枝方法，即CAM的剪枝过程（*CAM-pruning*，附录C），从由该排序确定的全连接图中移除多余的边。

算法1： NoGAM 因果发现 输入：数据矩阵 $X \in \mathbb{R}^{n \times d}$ $X^\pi \leftarrow [\,]$ $nodes \leftarrow [1, \ldots, d]$ **对于** $i = 1, \ldots, d$ **执行** $\hat{\mathbf{s}} \leftarrow$ SCORE(X) $\hat{\mathbf{R}} \leftarrow \{\hat{X}_j\}_{j=1}$ 根据 $\mathbf{X}_{\setminus\{i\}}$ $\tilde{\mathbf{s}} \leftarrow \{\tilde{s}_j\}_{j=1}$ 根据 $\hat{R}_i$ $MSE \leftarrow$ 平均值 $[\hat{\mathbf{s}} - \tilde{\mathbf{s}}]^2$ $l_{index} \leftarrow$ argmin MSE $l \leftarrow nodes[l_{index}]$ $X^\pi \leftarrow [l, X^\pi]$ 从 X 中移除第 l_{index} 列；从 $nodes$ 中移除 l

end

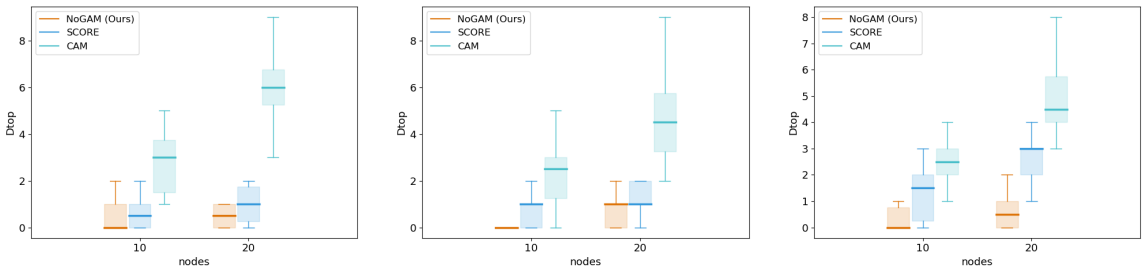
$X^\pi \leftarrow reverse(X^\pi)$（第一个节点是源节点，最后一个节点是叶节点）$\hat{\mathcal{G}} \leftarrow$ CAM-剪枝(X^π)（CAM-剪枝：CAM 算法的剪枝方法）**返回** $\hat{\mathcal{G}}$

6. 实验

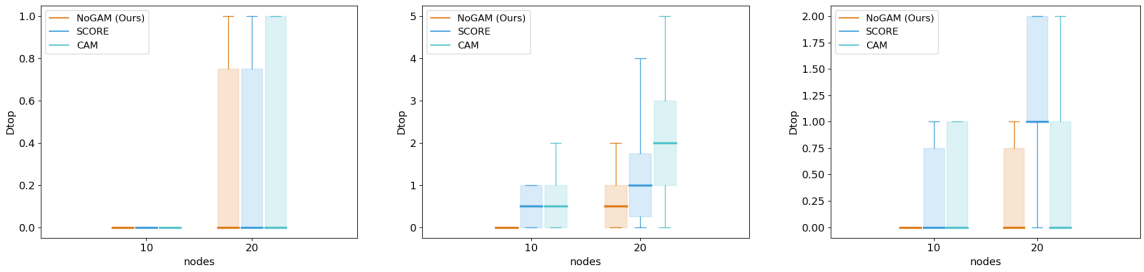
本节通过实证研究 NoGAM（算法 1）的性能。实验分析基于使用 Erdős–Rényi (1960)（ER）模型合成生成的数据进行。仿照 Rolland 等人 (2022)、Lachapelle 等人 (2020) 和 Zhu 等人 (2020年) 的设置，我们通过使用具有单位带宽径向基函数核的高斯过程进行采样来生成机制 f_i。当节点数 d 等于 10 和 20，且期望边数等于 d（ER1）和 $4d$（ER4）时，重复进行实验，以模拟在更稀疏和更稠密的图上的推断。数据集大小为 $N = 1000$。我们在噪声项服从以下分布之一的数据集上测试性能：贝塔、指数分布、伽马、贡贝尔分布、拉普拉斯和正态分布。通过将 NoGAM 的性能与基于高斯性假设运行的最先进方法进行比较，我们能够提供关于算法对分布变化的鲁棒性的实证证据。

$\hat{}$ $\hat{}$残差$\mathbf{R}$ 及得分函数各项 $\tilde{}\mathbf{s}(\mathbf{R})$ 通过核岭回归估计，使用`scikit-learn` (Pedregosa 等人(2011)) 对算法的实现，超参数为 $\alpha = 0.01$，$\gamma = 0.1$：这些值通过最小化估计残差上的泛化误差进行调优，且不使用因果图真实结构上的性能。对于 CAM 剪枝步骤，截断阈值设为 0.001。

用于评估推断图质量的指标包括结构汉明距离（SHD），即假阳性、假阴性和反向边的总和；以及结构干预距离（SID，Peters 和 B¨uhlmann (2015)），它统计由推断图导致的干预分布计算错误的数量。我们还单独评估拓扑排序估计的质量，方法如下：给定一个排序 $\hat{}\pi$ 以及真实值邻接矩阵



(a) 贝塔分布 (b) 指数分布 (c) 伽马分布



(d) 正态分布 (e) 贡贝尔分布 (f) 拉普拉斯分布

图3：在稀疏图 ER1 上运行 10 次所得的箱线图，展示了拓扑排序偏差 D_{top} （越低越好）。GES 算法未显示，因为它不需要显式估计拓扑排序。总体而言，NoGAM 的表现明显优于 CAM 和 SCORE。

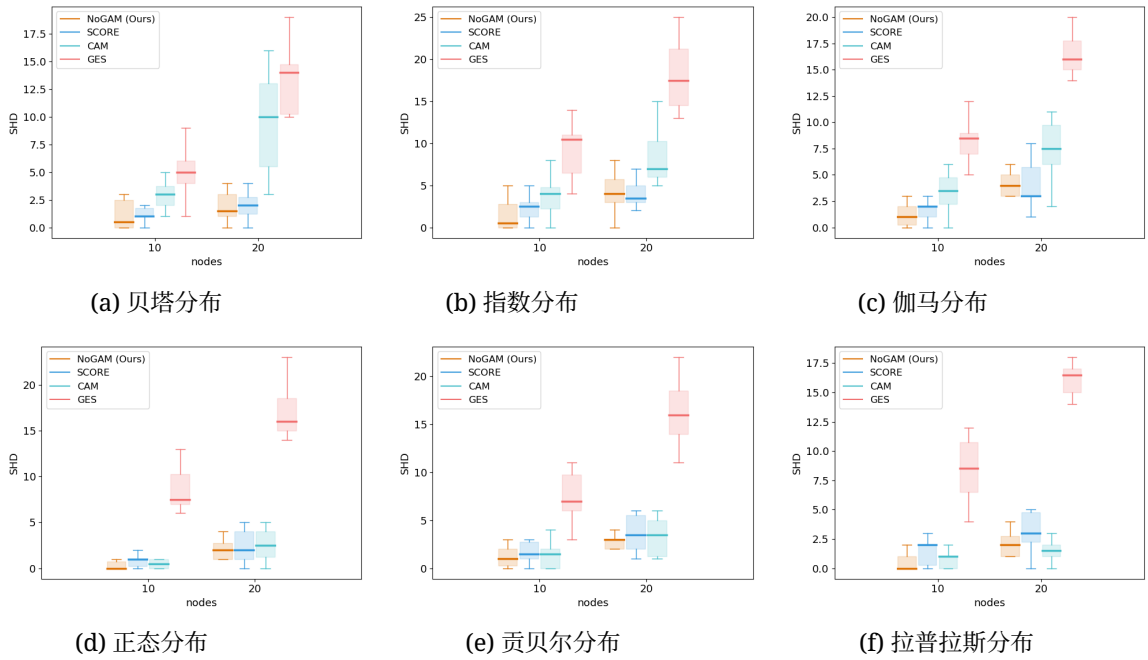
矩阵 A，我们使用 SCORE 中定义的拓扑排序偏差（Rolland 等人 (2022)）：

$$D_{top}(\hat{\pi}, A) = \sum_{i=1}^{d} \sum_{j:\hat{\pi}_i \succ \hat{\pi}_j} A_{ij} \,, \tag{23}$$

其中 $\hat{\pi}_i \succ \hat{\pi}_j$ 表示节点 i 在排序中紧接着 j，而 $A_{ij} = 1$ 当 $X_i \in \text{PA}_j(\mathbf{X})$。换言之，这表示由于拓扑排序的选择而无法恢复的边的总数。如果相对于 A 而言，$\hat{\pi}$ 正确，那么 $D_{top}(\hat{\pi}, A) = 0$。

我们将算法的实验表现与 SCORE（Rolland 等人（2022年）、CAM（B¨uhlmann 等人（2014年））和 GES（Chickering （2002））进行比较。这些因果发现方法均假设噪声服从高斯分布。我们不纳入 PC 和 FCI，因为它们通常表现差得多（B¨uhlmann 等人 （2014年）、Lachapelle 等人 （2020年）。

图3 展示了在稀疏情形（ER1）下，NoGAM 的拓扑排序估计在所有非高斯噪声分布下均系统性地优于 D_{top} CAM 和 SCORE 得到的结果。只有在噪声服从高斯分布的数据集上，准确率差距才会缩小，这与我们的预期一致。如附录中的图5 所示，在稠密情形（ER4）下也观察到了类似的性能差距，且有利于我们的方法。总体而言，与其他方法相比，SCORE 和 NoGAM 的 SHD（图4 和6）明显更优。请注意，这两种算法仅在拓扑排序推断步骤上有所不同，而它们采用的剪枝方法均与 CAM 相同。ER 图的实验结果完整概述见附录G。



(a) 贝塔分布 (b) 指数分布 (c) 伽马分布

(d) 正态分布 (e) 贡贝尔分布 (f) 拉普拉斯分布

图4：展示稀疏图 ER1 上结构汉明距离性能的 10 次运行箱线图。总体而言，NoGAM 优于所有其他方法，而 SCORE 的性能相当。

我们在附录中对实验进行了大幅扩展。我们在 Sachs 真实数据集（Sachs 等人（2005），附录 H.1）和无标度合成图（Barabasi 和 Albert （1999年），附录H.2）上测试了 NoGAM 的性能。在附录H.3 中，我们分析了在将假设空间限制为线性函数时，NoGAM 在公式（11）和（20）所述回归问题上的表现。此外，在附录F 中，我们将其与Mooij 等人（2009年）以及Peters 等人（2014年）的方法进行比较；这些方法依靠残差的独立性，在不限制噪声项分布的情况下发现非线性 ANM 中的因果效应。

关于 SCORE 鲁棒性的讨论 尽管 SCORE 算法的表现始终逊于我们的方法（图3和5），但它在不同噪声项分布下表现出显著的鲁棒性。这里我们想简要讨论一下，在我们看来，这一现象背后的原因。Rolland等人（2022）的引理 1 提出，可从得分函数 s(X) 的雅可比矩阵中识别一个叶节点。具体而言，计算雅可比矩阵对角元素的方差 Var $\left[\partial_{X_i} s_i(\mathbf{X})\right], \forall i = 1, \ldots, d$，可以证明以下结论：在高斯假设下，叶节点对应零方差，即 Var $\left[\partial_{X_l} s_l(\mathbf{X})\right] = 0 \iff \mathrm{CH}_l(\mathbf{X}) = \emptyset$，当且仅当 X_l 是叶节点。实践中，由于估计存在统计误差，该表达式通常不会恰好为零。为此，SCORE 算法通过迭代选择满足 $l = \operatorname{argmin}_{i \in \{1,\ldots,d\}} \mathrm{Var}\left[\partial_{X_i} s_i(\mathbf{X})\right]$ 的叶节点。我们认为，这一启发式方法是该算法在高斯假设之外仍具有经验鲁棒性的关键。如果噪声项不服从正态分布，那么得分函数等价于公式（7）。显然，$\partial_{X_i} s_i(\mathbf{X})$ 的方差与求和项中的子节点数量成正比。具体而言，得分函数雅可比矩阵对角元素的总方差等于公式（7）中两个右侧项偏导数的方差之和，再加上它们的协方差：

	方法	排序时间 [s]	总时间 [s]
d=10	NoGAM	6.98 ± 0.57	10.56 ± 0.84
	SCORE	4.28 ± 0.28	7.72 ± 01.07
	CAM	24.12 ± 1.58	27.98 ± 2.61
	GES	–	0.43 ± 0.11
d=20	NoGAM	25.75 ± 1.48	53.98 ± 2.01
	SCORE	9.51 ± 1.12	37.3 ± 4.21
	CAM	501.92 ± 11.92	529.61 ± 18.41
	GES	–	5.35 ± 1.67

表1：实验执行时间。在固定方法、节点数量和噪声项分布的 ER1 数据上运行 10 次，并据此计算经验均值和偏差。

如果协方差恰好为负且其绝对值足够大，则与非叶节点相关的方差可能小于相对于所有其他节点的方差，从而导致推断出的拓扑排序出现错误。然而，一般来说，这种情况似乎并不常见：尽管叶节点方差为零的保证不再成立，但在实践中，基于 argmin 运算符的叶节点选择准则仍可被认为是有效的。在附录 D 中，我们扩展了关于Rolland 等人 (2022) 的讨论，重点关注我们的方法与其方法之间的差异。

6.1. 算法复杂度

现在，我们对 NoGAM 拓扑排序方法的算法复杂度进行分析。记 n 为样本数量，记 d 为节点数量。根据算法 1的实现，for 循环的每次迭代都需要针对每个 $\mathcal{O}(n^3)$ˆ残差求解一个 d 回归问题，该问题的 R_i 估计成本为。对 ˜**s** 由ˆ**R**进行估计时，也可进行同样的分析。这给出了总体的 $\mathcal{O}(d^2n^3)$ 复杂度。与 NoGAM 类似，SCORE 通过迭代方式每次识别一个叶节点：每次迭代都需要对数据的 $n \times d$维矩阵求逆，因此拓扑排序推断时间按 $\mathcal{O}(dn^3)$进行伸缩。表1给出了实验执行时间的概览。

7. 结论

在加性非线性模型中，如果高斯噪声项这一假设不成立，因果发现算法可能会推断出边方向反转的图。本文证明了这一局限性，并针对这一问题提出了 NoGAM。基于Rolland 等人(2022)提出的得分匹配与因果发现之间的联系，我们的算法提出了一种新颖且一致的拓扑排序推断方法，且不对噪声项的分布作任何假设。通过系统性实验，我们证明，在噪声项服从任意分布时，我们的方法在几乎所有合成数据集上的表现均优于传统因果发现算法和最先进的因果发现算法。

致谢

本研究得到美国空军科学研究办公室资助，资助编号：FA8655-20-1-7035。FM 由国家研究与创新行动计划 *2014-2020*年资助。

参考文献

Albert-Laszlo Barabasi 和 Reka Albert。随机网络中尺度律的涌现。《科学》，286(5439):509–512，1999年。doi: 10.1126/science.286.5439.509。网址：http://www.sciencemag.org/cgi/content/abstract/286/5439/509。

Patrick Bloebaum、Dominik Janzing、Takashi Washio、Shohei Shimizu 和 Bernhard Schoelkopf。通过比较回归误差进行因果效应推断。在 Amos Storkey 和 Fernando Perez- Cruz 编，第二十一届人工智能与统计国际会议论文集，第 84of 卷，机器学习研究论文集，第900–909页。PMLR，09–11 Apr 2018年。网址 https://proceedings.mlr.press/v84/bloebaum18a.html。

Peter B¨uhlmann, Jonas Peters, and Jan Ernest. CAM：因果加性模型、高维序搜索和惩罚回归。《统计年鉴》，42(6)，2014年12月。URL https://doi.org/10.1214%2F14-aos1260。

David Chickering。使用贪心搜索进行最优结构识别。机器学习研究期刊，3：507–554，2002年01月。

Paul Erdos 和 Alfred Renyi。《随机图的演化》。匈牙利科学院数学研究所出版物，5:17–61，1960 年。

Clark Glymour、Kun Zhang 和 Peter Spirtes。基于图模型的因果发现方法综述。《遗传学前沿》，10，2019 年。ISSN 1664-8021。doi：10.3389/fgene。2019.00524。URLhttps://www.frontiersin.org/articles/10.3389/fgene。2019.00524。

Jackson Gorham 和 Lester Mackey。使用核衡量样本质量。载于 Doina Precup 和 Yee Whye Teh（编），第*34*届国际机器学习会议，第70卷，机器学习研究论文集，第1292–1301页。PMLR，2017年8月6日至11日。网址：https://proceedings.mlr.press/v70/gorham17a.html。

Patrik O 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》。载于《神经信息处理系统进展》，第 689–696 页，2009年。

Daphne Koller 和 Nir Friedman。《概率图模型：原理与技术》——自适应计算与机器学习。麻省理工学院出版社，2009年。ISBN 0262013193。S´ebastien Lachapelle、Philippe Brouillard、Tristan Deleu 和 Simon Lacoste-Julien。基于梯度的神经有向无环图学习。载于国际学习表征会议，2020年。网址：https://openreview.net/forum?id=rklbKA4YDS。

Yingzhen Li 和 Richard Turner。隐式模型的梯度估计量。2017年5月。

Qiang Liu、Jason D. Lee 和 Michael Jordan。《用于拟合优度检验的核化斯坦差异》。载于第 *33* 届国际机器学习会议论文集——第 *48* 卷，ICML'16，第 276–284 页。JMLR.org，2016年。

Giampiero Marra 和Simon Wood。《广义加性模型的实用变量选择》。计算统计与数据分析，55:2372–2387，072011。doi：10.1016/j.csda.2011.02.004。Joris Mooij、Dominik Janzing、Jonas Peters和 BernhardSch¨olkopf。《通过依赖最小化进行回归及其在因果推断中的应用》。第 94页，06 2009年。doi：10.1145/1553374. 1553470。JudeaPearl。《因果关系》。剑桥大学出版社，英国剑桥，第2 版，2009年。ISBN 978-0-521-89560-6。doi：10.1017/CBO9780511803161。F. Pedregosa、G.Varoquaux、A.Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Pretten- hofer、R.Weiss、V.Dubourg、J. Vanderplas、A.Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E.Duchesnay。Scikit-learn：Python 中的机器学习。《机器学习研究杂志》，12:2825–2830，2011。JonasPeters 和PeterB¨uhlmann。《用于评估因果图的结构干预距离》。《神经计算》，27(3):771–799，032015。ISSN 0899-7667。doi：10.1162/NECO a 00708。URLhttps://doi.org/10.1162/NECO_a_00708。Jonas Peters、Joris M. Mooij、Dominik Janzing和 Bernhard Sch¨olkopf。《连续加性噪声模型的因果发现》。《机器学习研究杂志》，15(1):2009–2053，2014年 1 月。ISSN 1532- 4435。Jonas Peters、DominikJanzing 和 Bernhard Schlkopf。《因果推断要素：基础与学习算法》。麻省理工学院出版社，2017。ISBN 0262037319。Paul Rolland、Volkan Cevher、Matth¨aus Kleindessner、Chris Russel、Bernhard Sch¨olkopf、DominikJanzing和FrancescoLocatello。得分匹配实现非线性加性噪声模型的因果发现。载于国际机器学习会议（ICML），2022。URL https: //arxiv.org/abs/2203.04413。Karen Sachs、Omar Perez、Dana Pe'er、Douglas A.Lauffenburger和Garry P. Nolan。源自多参数单细胞数据的因果蛋白质信号网络。《科学》，308(5721): 523–529，2005。URLhttps://www.science.org/doi/abs/10.1126/science. 1105809。RajenShah和Jonas Peters。条件独立性检验的难度与广义协方差度量。《统计年鉴》，48，04 2018年。doi：10.1214/19-AOS1857。P. Spirtes、C. Glymour 和 R.Scheines。《因果、预测与搜索》。麻省理工学院出版社，第 2版，2000。Charles M.Stein。相关随机变量之和的分布的正态近似误差界。1972。K.Zhang 和A.Hyv¨arinen。加性扰动下的因果发现：信息论视角。载于 2009年欧洲机器学习会议与数据库知识发现原理与实践（ECML PKDD）会议论文集，布莱德，斯洛文尼亚，2009年。

KunZhang和AapoHyv¨arinen。关于后非线性因果模型的可识别性。收录于第二十五届人工智能不确定性大会，UAI'09，第647–655页，阿灵顿，弗吉尼亚州，美国，2009年。AUAI出版社。ISBN 9780974903958。

Xun Zheng、Bryon Aragam、PradeepKRavikumar 和 Eric PXing。《没有眼泪的有向无环图：用于结构学习的连续优化》。载于 S.Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，《神经信息处理系统进展》，第31卷。Curran Associates, Inc.，2018年。网址：
https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf。

Shengyu Zhu、Ignavier Ng 和 Zhitang Chen。《通过强化学习进行因果发现》。载于国际学习表征会议，2020年。网址：https://openreview.net/forum?id=S1g2skStPB。

附录 A. 多变量 ANM 的可识别性

Peters 等人（2014）表明，双变量模型的条件 1 足以证明多变量情形下的可识别性。直观地说，给定形如 $X_j := f_j(\mathrm{PA}j) + N_j$ 的结构方程 d，如模型 (1) 所示，只需固定 f_j 的所有自变量（一个父节点 X_i 和噪声变量 N_j 除外），即可重现双变量加性噪声模型。更正式地说，Peters 等人（2014）的定义 27 将受限加性噪声模型定义为这样一种结构因果模型：对于所有节点 $j \in \{1,\dots,d\}$、$i \in \mathrm{PA}j$，以及所有满足 $\mathrm{PA}_j \setminus \{i\} \subseteq S \subseteq \mathrm{ND}_j \setminus \{i,j\}$ 的集合 $S \subseteq \{1,\dots,d\}$，都存在一个 $\mathbf{X}_S$ 的取值，其联合密度为 $p_{\mathbf{X}_S}(\mathbf{x}_S) > 0$，使得三元组

$$(f_j(\mathrm{pa}_j \setminus \{i\}, X_i), p_{X_i|\mathbf{X}_s}(X_i \mid \mathbf{x}_S), p_{N_j}(N_j))$$

满足条件 1。为方便起见，我们滥用符号 PA_i，同时表示因果图中的节点及其对应的随机变量。此外，我们用 $\mathrm{pa}j$ 表示随机变量向量 $\mathrm{PA}j$的观测值。一般而言，我们用大写字母表示随机变量，用小写字母表示其取值。

Peters 等人的定理 28 （2014年）证明，若某个分布由一个受限加性噪声模型生成，且该模型满足因果最小性（即因果机制 f_j 对其任一参数均非恒定），那么因果图可从观测数据中识别。我们假设模型（1）是根据 Peters 等人受限 *ANM*的定义 27 构造的，（2014年），从而确保因果图的可识别性。

附录 B. 引理 1的证明

(i) X_i 叶节点 $\Rightarrow \mathbf{E}\left[(h^*(R_i) - s_i(\mathbf{X}))^2\right] = 0$：由方程（22）可知此结论成立。

(ii) $\mathbf{E}\left[(h^*(R_i) - s_i(\mathbf{X}))^2\right] = 0 \Rightarrow X_i$ 叶节点：零期望可以改写为

$$\int_{\mathbb{R}\times\mathbb{R}^d} (h^*(R_i) - s_i(\mathbf{X}))^2 \, dp(R_i, \mathbf{X}) = 0\,.$$

由于积分是对一个正函数进行的，因此显然，当且仅当 $h^*(R_i) = s_i(\mathbf{X})$ 以概率为 1 成立时，等式才等于零，此时 $\mathbb{P}_{\mathbf{X}|R_i}(s_i(\mathbf{X}) = h^*(R_i) \mid R_i) = 1, \forall\, R_i \in \mathbb{R}$。由此可知，条件方差满足 $\mathrm{Var}_{\mathbf{X}}[s_i(\mathbf{X}) \mid R_i] = 0$ 对所有 $R_i \in \mathbb{R}$ 均成立，并且给定已观测的 R_i 后，$s_i(\mathbf{X})$ 几乎必然为常数。我们将通过引入额外符号来明确说明这一事实。定义 $c_{R_i} := h^*(R_i)$，这样便可清楚地看出，当 R_i 固定时，c_{R_i} 是一个常数。

现在，我们将证明 X_i 在模型（1）的二元情形中必须是一个叶节点。随后，我们将把这些论证推广到一般的 n 变量情形。设 $(X_i,\ X_j)$ 是二元图 $\mathcal{G}$的节点。反证法设 X_i 在该图中不是叶节点。根据方程（18），SCORE 的 i阶项可以写成

$$\begin{aligned} s_i(X_i, X_j) &= \partial_{N_i} g_i(R_i + \mathbf{E}[N_i|X_j]) - \partial_{X_i} f_j(X_i)\partial_{N_j} g_j(X_j - f_j(X_i)) = \\ &=: g_i'(R_i + \mathbf{E}[N_i|X_j]) - f_j'(X_i) g_j'(X_j - f_j(X_i))\,. \end{aligned} \tag{24}$$

在 R_i 条件下，评分项几乎必然为常数，即 $s_i(X_i, X_j) \mid R_i = c_{R_i}$，概率为 1。

根据生成 ANM的可识别性假设，$f_j'(x_i)\, g_j''(x_j^* - f_j(x_i)) \neq 0$ 除有限个 $(x_i, x_j) \in \mathbb{R}^2$ （公式（4））。定义 $\mathcal{X}$ 为满足该条件的不可数对 (x_i, x_j) 的集合，即

$$\mathcal{X} \coloneqq \{(x_i, x_j) \in \mathbb{R}^2 \mid f_j'(x_i) g_j''(x_j - f_j(x_i)) \neq 0\} \subset \mathbb{R}^2 .$$

我们的下一个目标是证明，与 $\mathrm{Var}_{\mathbf{X}}[s_i(\mathbf{X}) \mid R_i] = 0$（假设期望为零的直接推论）这一事实相矛盾的是，存在某个R_i，使得 $s_i(x_i, x_j) \neq s_i(x_i^*, x_j^*) \mid R_i$，其中 (x_i, x_j)、(x_i^*, x_j^*) 是 $\mathbf{X} \mid R_i$ 的支持集中的不同点对。根据公式 (14)的定义，我们有 $R_i = N_i - \mathbf{E}[N_i \mid X_j]$；根据噪声项具有严格正密度的可识别性假设，我们知道，对于每个处于支持集中的 n_i，$p_{N_i}(n_i) > 0$。然后，很明显，$\mathbf{X}$ 的支持集及其变换 $s_i(\mathbf{X})$ 的支持集并不受 R_i 的观测限制：事实上，对于每个观测 $R_i = r_i$，对于任意 $X_j = x_j, \exists\, n_i$，只要 $p_{N_i}(n_i) > 0$，就允许 $r_i = n_i - \mathbf{E}\,[N_i \mid x_j]$。

因此，我们知道存在一个不可数的点集 $(x_i, x_j) \in \mathcal{X}$，满足$s_i(x_i, x_j) = c_{R_i}$和 $f_j'(x_i) g_j''(x_j - f_j(x_i)) \neq 0$，条件是 R_i。根据公式 (5)，我们还知道 $f_j'(x_i) g_j''(x_j - f_j(x_i)) \neq 0$ 意味着 $g_i'(x_i) - f_j'(x_i) g_j'(x_j - f_j(x_i))$ 在 x_i 上对所有点对 $(x_i, x_j) \in \mathcal{X}$的偏导数均非零；也就是说，它在 x_i上绝不会是常数。鉴于 $g_i'(x_i) - f_j'(x_i) g_j'(x_j - f_j(x_i))$ 恰好是第 i 个评分项的解析表达式，因此，对于任意(x_i, x_j)、$(x_i^*, x_j^*) \in \mathcal{X}$以及 $x_i \neq x_i^*, x_j \neq x_j^*$，都有 $s_i(x_i, x_j) \neq s_i(x_i^*, x_j^*) \mid R_i$。因此，在 R_i条件下，我们有 $\mathrm{Var}_{\mathbf{X}}[s_i(\mathbf{X}) \mid R_i] > 0$，这与假设矛盾。

那么 X_i 必定是一个叶节点，这就证明了引理在二元情形下的结论。

现在，我们考虑式（1）中的多元受限 ANM。同样，通过反证法，假设X_i 是一个非叶节点。由于模型可识别，我们知道，对于每个节点 $q \in \{1, \ldots, d\}$，$u \in \mathrm{PA}_q$，以及所有满足 $S \subseteq \{1, \ldots, d\}$ 且 $\mathrm{PA}_q \setminus \{i\} \subseteq S \subseteq \mathrm{ND}_q \setminus \{u, q\}$的集合，都存在一个 $\mathbf{x}_S$，其联合密度为 $p_{\mathbf{x}_S}(\mathbf{x}_S) > 0$ 使得三元组

$$(f_q(\mathrm{pa}_q \setminus \{u\}, X_u), p_{X_u \mid \mathbf{x}_s}(X_u \mid \mathbf{X}_S), p_{N_q}(N_q))$$

满足条件 1。令 i_c 为节点 i 的一个子节点，且 $i_c \notin \mathrm{PA}_k$ 对每个 $k \in \mathrm{CH}_i$均成立。由于因果图满足无环约束，这样的节点 i_c 总是存在。令 $S = \mathrm{ND}_{i_c} \setminus \{i, i_c\}$：在给定随机变量集合 $\mathbf{X}_S = \mathbf{x}_S$ 的条件下，节点 i 的评分项为

$$\begin{aligned} s_i(\mathbf{X}) = {} & g_i'(R_i + \mathbf{E}[N_i \mid \mathrm{ch}_{i \setminus \{X_{i_c}\}}, X_{i_c}]) + \\ & - \partial_{X_i} f_{i_c}(\mathrm{pa}_{i \setminus \{X_i\}}, X_i) g_{i_c}'(X_{i_c} - f_{i_c}(\mathrm{pa}_{i_c \setminus \{X_i\}}, X_i)) + \\ & - \sum_{k \in \mathrm{CH}_{i \setminus \{i_c\}}} \partial_{X_i} f_k(\mathrm{pa}_{k \setminus \{X_i\}}, X_i) \partial_{N_k} g_k(n_k = x_k - f_k(\mathrm{pa}_{k \setminus \{X_i\}}, X_i)) \end{aligned} \tag{25}$$

如果我们固定 $X_i = x_i^*$，使得对于所有不可数的对 (x_i^*, x_{i_c})，满足 $g_i'(N_i) - \partial_{X_i} f_{i_c}(\mathrm{pa}^{i \setminus \{X_i\}}, x_i^*) g_{i_c}'(X_{i_c} - f_{i_c}(\mathrm{pa}^{i \setminus \{X_i\}}, x_i^*))$ 不为常数，那么我们可以观察到：方程 (25) 的前两项与二元情形中的方程 (24) 完全类似，已证明在观测 R_i 时方差不为零；而对其余子节点的求和在给定 $\mathbf{x}_S$ 和 x_i^* 的条件下并不对方差作出贡献。因此，我们知道 $\mathrm{Var}_{\mathbf{X}}[s_i(\mathbf{X}) \mid R_i] > 0$，这与假设相矛盾。我们得出结论：$X_i$ 必须是一个叶节点。

附录C. 使用CAM剪枝有向无环图

一旦通过拓扑排序施加有向无环图约束，我们的 NoGAM 算法就会使用 *CAM-pruning* 过程（B¨uhlmann 等人（2014年））。事实上，与符合该拓扑排序的全连接图相比，较小的有向无环图通常能够对干预分布进行统计效率更高的估计。其思想是，在假设 f_i 中的非线性机制具有加性结构（1）的情况下，可以对潜在父节点进行回归，并使用加性假设检验（Marra 和 Wood（2011）来判断一条边是否存在。更多详细信息，请参阅B¨uhlmann 等人 （2014年）的原论文。

附录D. 与SCORE的比较

本节概述我们的工作与Rolland 等人(2022）的主要区别，并简要介绍 Stein 梯度估计器；它是 SCORE 和 NoGAM 算法实现中共同的核心组成部分。Rolland 等人(2022）提出了一种通过检查得分函数雅可比矩阵的对角元素来识别叶节点的方法：具体而言，他们通过估计对角项的方差来识别其中的非线性，而这只有在假设噪声项服从高斯分布时才有可能。此外，为了保证得分函数雅可比矩阵中的非线性，因果机制 f_i 必须对各自的每个自变量均为非线性：这是一个很强的假设；若该假设不成立，SCORE 就会推断出错误的拓扑排序。为了更清楚地说明这一点，我们明确考察得分函数雅可比矩阵对角项的解析形式：

$$\partial_{X_i} s_i(\mathbf{X}) = \partial^2_{X_i} \log p_i(X_i \mid \mathrm{PA}_i) + \sum_{k \in \mathrm{CH}_i} \partial_{X_i}(\partial_{X_i} f_k(\mathrm{PA}_k) \partial_{X_k} \log p_k(X_k \mid \mathrm{PA}_k)) \,.$$

通过将 $p_k(X_k \mid \mathrm{PA}_k)$ 正态分布明确写出，可以很容易看出，当 f_k 关于 X_i 是线性函数时，求和中对其子节点求和的各项会如何消失（这是因为$f_k(\mathrm{PA}_k)$的二阶偏导数为零）。相比之下，我们的论文基于最小化从方程（14）残差预测得分分量时的泛化误差，提出了在对噪声和函数机制的假设限制较少的情况下识别因果图中叶节点的理论。

从实践角度来看，SCORE 和 NoGAM 方法都依赖于利用基于得分匹配的 Stein 梯度估计器高效近似 $\nabla_{\mathbf{X}} \log p(\mathbf{X})$ （Li和Turner (2017）的具体做法如下文所述。

Stein 梯度估计器该估计器基于 Stein 恒等式。该恒等式最初是针对高斯随机变量提出的（Stein (1972），随后推广到了一般情形（GorhamandMackey (2017）；Liu 等人 (2016年）。对于任意满足以下条件的测试函数$\mathbf{h}: \mathbb{R}^d \to \mathbb{R}^{d'}$，若 $\lim_{\mathbf{x}\to\infty} \mathbf{h}(\mathbf{x}))p(\mathbf{x}) = 0$，则下列恒等式成立：

$$\mathbf{E}_p \left[\mathbf{h}(\mathbf{x}) \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x})\right] = 0.$$

我们注意到，所关注的量 $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ 出现在该表达式中。由于其中的期望积分难以求解，Li和Turner(2017)提议利用蒙特卡洛方法对期望进行近似，并随后证明，可以推导出一个在大样本极限下保持一致性的得分估计量。如需更多细节，请参阅原始论文。

附录 E：关于示例的讨论1

示例 1 通过实验说明，当真实值生成模型中的高斯噪声项假设不成立时，在该假设下进行推断的不足之处。接下来，我们将更详细地介绍该示例的实验设计，以及因果效应方向的推断方法。对于线性和非线性情形，我们生成2000 个真实值噪声项的样本 X_1 和 N_2，它们分别在$[0,5]$ 和 $[1,4]$ 区间内均匀分布（见图 1a）。ZHANG 和 Hyv¨arinen (2009年) 的定理 1 证明，当拟合加性模型（1）且其因果结构由有向无环图表示时，扰动项的总熵，即 $\sum_i^d H(N_i)$，在 $H(X_1,\dots,X_d)$的最小值处达到最小（这对应于负对数似然的最小值）。因此，在示例 1的二元情形中，我们可以选择估计噪声项总熵达到最小值的方向作为正确的因果方向。

附录 F：基于估计残差独立性的因果发现

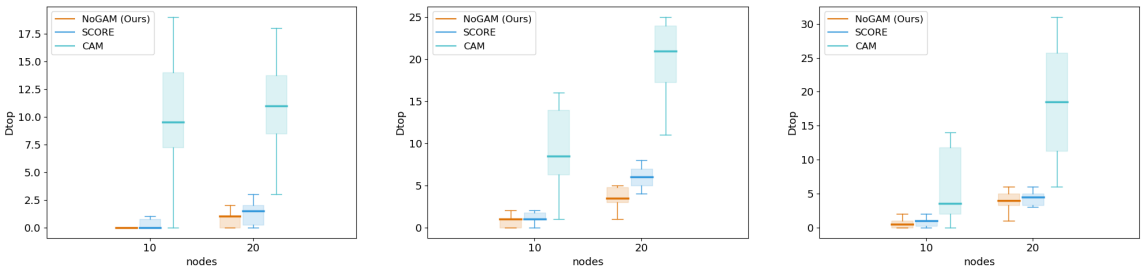
Mooij 等人 (2009年) 和 Peters 等人 (2014年) 提出了一种与我们的算法 1相当的因果发现方法，该方法基于可识别的加性噪声模型（ANM），且不对噪声项的分布作任何限制。在实践中，他们通过迭代方式估计因果结构，进行类似于我们公式（11）的回归，并检验推断出的残差R_i 与变量 $\mathbf{X}_{\setminus\{i\}}$之间的独立性。这种方法的主要瓶颈在于使用独立性检验；独立性检验不仅难以实施（Shah 和 Peters (2018年)），而且难以扩展到高维图和大型数据集。此外，除非假设存在预言机独立性检验，否则该方法无法保证总体情形下推断出的图具有一致性。表 2 比较了 NoGAM 与 Peters 等人 提出的 RESIT 方法 (2014年) 的实证表现。我们复现了 Peters 等人 的实验设置 (2014年)（第 5.1.2 节）：从高斯过程采样非线性机制，带宽设为 1，噪声项相互独立且服从正态分布，方差则从均匀分布中选取。实验在样本数为 $n \in \{100, 500\}$ 、变量数为 $d \in \{4, 15\}$且图为稀疏结构的设置下进行。可以看到，在所有实验配置下，NoGAM 的表现都优于 RESIT。

表 2：NoGAM 与 RESIT 的实验性能对比。RESIT 的结果取自Peters 等人（2014年），表 3 和表 4。

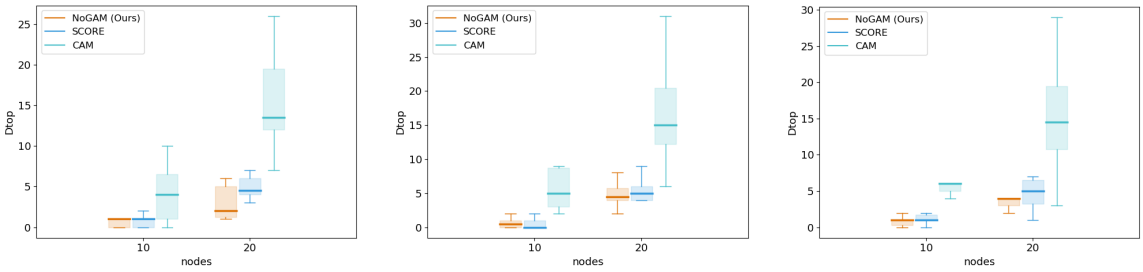
	方法	结构汉明距离 (n =100)	结构汉明距离 (n =500)
d=4	**NoGAM**	$\mathbf{1.6 \pm 1.1}$	$\mathbf{0.4 \pm 0.4}$
	RESIT	1.7 ± 1.3	0.8 ± 0.9
d=15	**NoGAM**	$\mathbf{11.7 \pm 2.4}$	$\mathbf{7.6 \pm 4.1}$
	RESIT	15.4 ± 5.7	10.1 ± 5.7

附录 G。在 ER 图上的实验

从表3 到 8我们提供了第6节所述实验的完整概览。每个指标均取10 次运行的平均值，并记录经验均值和偏差。图5展示了在稠密设置（ER4）中，NoGAM 推断拓扑排序的方法。



(a) 贝塔分布 (b) 指数分布 (c) 伽马分布



(d) 正态分布 (e) 贡贝尔分布 (f) 拉普拉斯分布

图5：在稠密图（ER4）上，10次运行的拓扑排序偏差 D_{top} 箱线图。GES 算法未出现，因为它不需要对拓扑排序进行显式估计。从图示中我们可以看到，NoGAM 通常优于 CAM 和 SCORE。

优于 CAM 和 SCORE，与稀疏设置（ER1）中观察到的情况类似，如图 3所示。在图 6 中，我们可以看到，SCORE 和 NoGAM 的整体结构汉明距离相当。

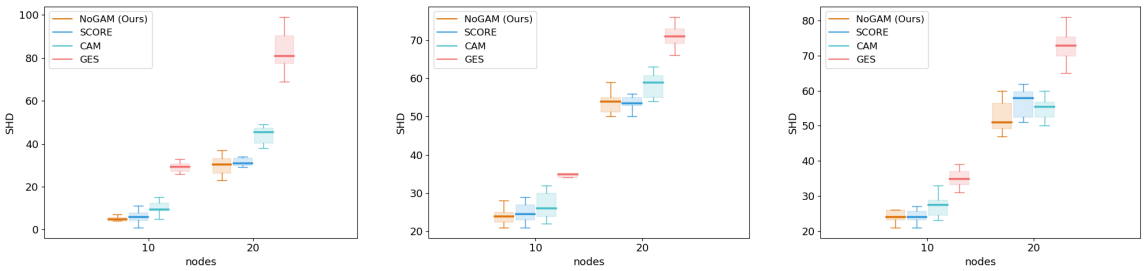
表 3：贝塔噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.1 ± 1.5	4.2 ± 5.4	0.6 ± 0.8	5.1 ± 1.3	11.4 ± 7.1	0.1 ± 0.3
	SCORE	1.6 ± 1.0	4.9 ± 3.5	0.6 ± 0.7	6.1 ± 2.0	9.6 ± 5.4	0.4 ± 0.7
	CAM	3.4 ± 1.9	13.8 ± 8.6	2.2 ± 1.4	14.4 ± 5.1	51.0 ± 14.3	7.7 ± 5.1
	GES	6.5 ± 2.0	22.0 ± 7.6	–	27.0 ± 5.2	72.0 ± 8.7	–
d=20	NoGAM	2.0 ± 1.6	8.0 ± 9.0	0.5 ± 0.5	30.0 ± 4.2	121.5 ± 22.6	0.9 ± 0.9
	SCORE	2.0 ± 1.5	8.0 ± 5.2	1.0 ± 0.3	33.1 ± 6.8	111.9 ± 29.1	1.4 ± 1.0
	CAM	9.5 ± 4.8	45.3 ± 16.0	5.7 ± 2.3	45.1 ± 7.2	220.8 ± 29.1	12.2 ± 7.0
	GES	12.9 ± 4.7	52.0 ± 14.4	–	78.5 ± 9.2	290.6 ± 25.9	–

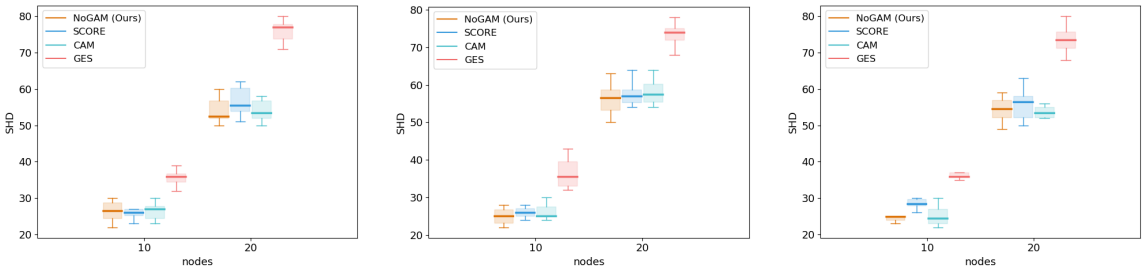
附录 H. 其他实验

H.1. Sachs 数据上的实验

我们在 Sachs 数据集（Sachs 等人（2005））上测试 NoGAM。该数据集由真实世界的生物学数据构成，是常用的因果发现基准。在表9的结果中，我们可以看到，相较于其他方法，NoGAM 仍保持着最先进的性能。



(a) 贝塔分布 (b) 指数分布 (c) 伽马分布



(d) 正态分布 (e) 贡贝尔分布 (f) 拉普拉斯分布

图 6：展示在稠密图（ER4）上结构汉明距离性能的 10 次运行箱线图。

表 4：指数分布噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.2 ± 1.4	4.3 ± 5.2	0.3 ± 0.6	24.1 ± 2.0	43.3 ± 5.2	0.8 ± 0.8
	SCORE	2.3 ± 1.2	4.1 ± 4.2	0.8 ± 0.9	25.1 ± 3.0	40.9 ± 6.6	1.1 ± 0.6
	CAM	3.4 ± 1.4	14.8 ± 5.8	2.4 ± 1.0	24.6 ± 2.0	54.1 ± 11.9	4.4 ± 2.1
	GES	9.0 ± 3.0	33.1 ± 12.2	–	33.9 ± 3.7	80.6 ± 5.8	–
d=20	NoGAM	2.5 ± 1.4	11.3 ± 9.1	0.7 ± 0.6	54.2 ± 4.5	218.1 ± 28.7	3.6 ± 1.6
	SCORE	4.2 ± 2.8	19.7 ± 13.6	1.4 ± 1.4	54.1 ± 4.0	215.1 ± 40.4	6.1 ± 1.3
	CAM	8.2 ± 1.8	46.4 ± 24.4	5.5 ± 2.1	54.8 ± 4.3	258.4 ± 57.4	18.8 ± 7.5
	GES	19.8 ± 3.5	87.8 ± 24.4	–	72.9 ± 5.4	336.2 ± 24.0	–

H.2. 无标度图上的实验

从表 10 至 15，我们给出了无标度（SF）图的更多实验结果（Barabasi 和 Albert （1999年）。数据集由 1000 个样本生成，节点数分别为 10 和 20，每个节点的平均边数分别为1 （稀疏图）和 4 （稠密图）。

H.3. 使用线性回归的 NoGAM

本节讨论我们的方法论可能存在的稳健性问题。具体而言，根据公式（11）和（20），我们的算法在估计残差及其得分函数时，需要在所有可测函数构成的空间上进行最小化。众所周知，在实践中，由于计算限制，我们需要进一步限制假设

表 5：伽马噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.2 ± 1.2	3.9 ± 5.4	0.3 ± 0.5	24.2 ± 1.8	41.4 ± 5.3	0.7 ± 0.8
	SCORE	1.6 ± 1.5	4.8 ± 6.3	1.4 ± 0.9	24.2 ± 0.9	41.6 ± 2.5	1.0 ± 1.1
	CAM	3.3 ± 1.6	17.3 ± 8.2	2.6 ± 1.7	26.5 ± 3.8	59.3 ± 15.1	8.0 ± 4.5
	GES	9.8 ± 2.2	34.6 ± 13.2	–	35.9 ± 3.6	80.8 ± 6.5	–
d=20	NoGAM	3.9 ± 1.7	13.5 ± 10.1	0.7 ± 0.8	52.7 ± 2.8	203.1 ± 24.9	3.7 ± 1.6
	SCORE	4.0 ± 1.8	14.8 ± 8.2	2.5 ± 1.3	56.5 ± 2.3	214.7 ± 21.7	4.7 ± 1.7
	CAM	8.3 ± 3.0	48.9 ± 23.3	5.1 ± 2.5	54.0 ± 3.5	264.5 ± 17.6	16.2 ± 4.6
	GES	18.8 ± 2.5	89.0 ± 19.4	–	73.9 ± 3.2	332.8 ± 19.2	–

表 6：高斯噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.4 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	26.4 ± 2.5	43.9 ± 4.8	0.8 ± 0.9
	SCORE	0.9 ± 0.7	3.2 ± 2.4	0.1 ± 0.3	26.1 ± 3.0	43.7 ± 8.7	0.7 ± 0.6
	CAM	0.6 ± 0.7	1.3 ± 1.6	0.0 ± 0.0	26.9 ± 2.2	47.4 ± 5.2	5.1 ± 3.4
	GES	8.3 ± 1.8	31.9 ± 9.5	–	36.2 ± 2.4	85.3 ± 3.8	–
d=20	NoGAM	2.0 ± 1.0	10.1 ± 5.7	0.3 ± 0.5	54.0 ± 3.0	195.0 ± 17.0	3.1 ± 2.0
	SCORE	2.3 ± 1.3	11.6 ± 6.9	0.4 ± 0.7	56.5 ± 2.9	197.8 ± 22.3	5.0 ± 1.3
	CAM	2.1 ± 1.8	9.8 ± 9.9	0.8 ± 1.0	57.1 ± 1.9	204.1 ± 34.2	21.4 ± 8.4
	GES	17.1 ± 3.7	70.0 ± 22.1	–	71.9 ± 3.8	340.1 ± 20.9	–

表 7：Gumbel噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.3 ± 1.4	3.3 ± 4.2	0.2 ± 0.4	24.9 ± 2.7	41.4 ± 4.8	0.6 ± 0.7
	SCORE	1.6 ± 1.4	6.3 ± 5.2	0.5 ± 0.5	26.1 ± 2.1	42.6 ± 5.7	0.5 ± 0.7
	CAM	1.4 ± 1.3	7.9 ± 8.8	0.8 ± 1.1	26.5 ± 1.9	55.9 ± 12.4	7.1 ± 4.3
	GES	9.7 ± 1.2	41.0 ± 11.1	–	38.2 ± 1.3	86.8 ± 2.5	–
d=20	NoGAM	2.7 ± 2.4	11.3 ± 6.8	0.6 ± 0.7	56.1 ± 3.9	207.9 ± 20.4	4.7 ± 2.4
	SCORE	3.5 ± 1.6	15.9 ± 17.4	1.3 ± 1.3	56.9 ± 2.2	205.6 ± 25.0	5.4 ± 1.6
	CAM	4.3 ± 1.4	29.7 ± 12.3	2.3 ± 1.4	57.0 ± 4.3	225.5 ± 24.7	17.6 ± 6.3
	GES	21.1 ± 1.9	95.4 ± 28.5	–	78.2 ± 2.4	348.5 ± 15.9	–

空间，即用于回归问题求解的函数类别所构成的空间，而这可能导致不可约误差。尽管如此，我们发现，使用均方误差的 argmin 迭代识别叶节点（如算法 1所提出）使 NoGAM 能够稳健地应对受限假设空间所引入的误差。直观来说，要在某次迭代中错误地识别出叶节点，每个叶节点得分的预测不可约误差都必须大于一般非叶节点的总预测误差，而后者也会因假设空间受限而增大。然而，如果某些变量具有目标

表 8：拉普拉斯噪声 ER 图

		ER1（稀疏）			ER4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.6 ± 0.7	2.2 ± 2.9	0.0 ± 0.3	24.5 ± 2.0	42.9 ± 6.5	0.8 ± 0.9
	SCORE	1.4 ± 1.0	3.0 ± 3.0	0.3 ± 0.4	27.9 ± 2.3	46.5 ± 6.1	1.1 ± 0.5
	CAM	1.1 ± 1.0	4.6 ± 4.9	0.4 ± 0.5	24.7 ± 1.7	43.8 ± 4.3	6.0 ± 3.1
	GES	10.0 ± 1.7	37.5 ± 10.1	–	35.6 ± 2.4	85.0 ± 2.3	–
d=20	NoGAM	2.0 ± 2.2	9.1 ± 6.4	0.4 ± 0.7	54.5 ± 3.4	204.2 ± 20.1	3.9 ± 1.5
	SCORE	3.1 ± 1.3	16.1 ± 6.6	1.3 ± 0.6	56.1 ± 1.9	209.6 ± 21.4	4.6 ± 2.0
	CAM	3.4 ± 1.2	11.1 ± 9.1	0.5 ± 0.7	54.8 ± 2.6	200.4 ± 16.1	18.5 ± 6.5
	GES	19.7 ± 2.4	90.4 ± 32.0	–	73.4 ± 3.1	329.4 ± 13.9	–

表 9：Sachs 数据集上的实验结果（11 个变量、17 条边、853 个观测变量）。

方法	D_{top}	结构汉明距离 结构干预距离	
NoGAM	8	12	45
SCORE	8	12	45
CAM	7	12	55
GES	–	17	62

表10：贝塔噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	2.6 ± 2.1	13.7 ± 15.9	1.7 ± 1.6	7.3 ± 2.3	12.9 ± 9.1	0.6 ± 0.7
	SCORE	2.3 ± 1.4	4.5 ± 4.0	0.7 ± 0.8	6.2 ± 1.4	16.0 ± 8.3	0.3 ± 0.5
	CAM	4.5 ± 1.2	29.9 ± 6.7	3.3 ± 0.8	14.3 ± 4.0	47.7 ± 18.6	6.4 ± 4.5
	GES	9.8 ± 2.9	56.3 ± 17.4	–	27.6 ± 5.8	73.7 ± 5.4	–
d=20	NoGAM	2.5 ± 1.7	24.8 ± 22.0	0.9 ± 1.0	31.1 ± 6.9	60.3 ± 20.0	0.8 ± 0.6
	SCORE	2.1 ± 1.4	5.8 ± 3.9	0.1 ± 0.3	31.6 ± 5.8	108.4 ± 18.7	1.2 ± 1.2
	CAM	6.3 ± 2.8	84.1 ± 40.8	4.6 ± 2.1	37.7 ± 7.4	178.9 ± 44.4	6.8 ± 4.2
	GES	21.3 ± 6.9	157.8 ± 48.5	–	93.3 ± 13.8	303.2 ± 13.1	–

函数比其他函数更接近假设空间，这可能会导致不同变量的不可约误差差异很大。如果函数空间有偏所引入的此类误差恰好在叶节点处更大，就可能导致排序出错，因为它可能会与估计残差混淆。尽管如此，我们认为这种情况通常并非如此：为了通过实验验证我们的主张，我们在 ER4 合成数据上运行 NoGAM 拓扑排序推断，将核岭回归器替换为 `scikit-learn` (Pedregosa 等人 (2011)) 中的线性模型套索回归，并将其同时用于残差估计和基于残差的分数预测。在表 16 中，我们看到，尽管生成模型的机制高度非线性，NoGAM 并未受到将搜索空间限制为线性函数的影响（如 Lasso 回归算法所致）：比较线性回归和非线性回归的结果后，我们发现它们几乎总是接近，且在误差条范围内具有可比性。

表11：指数分布噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.5 ± 1.1	7.7 ± 6.7	0.9 ± 0.8	8.3 ± 2.1	31.6 ± 6.0	0.6 ± 0.8
	SCORE	0.7 ± 0.8	2.4 ± 3.6	0.3 ± 0.6	8.0 ± 2.6	39.4 ± 6.1	0.4 ± 0.5
	CAM	3.6 ± 2.0	26.6 ± 12.6	2.8 ± 1.3	11.7 ± 3.3	54.0 ± 9.1	4.8 ± 4.0
	GES	11.7 ± 2.6	59.0 ± 10.7	–	26.2 ± 3.0	81.0 ± 5.9	–
d=20	NoGAM	5.3 ± 2.5	61.7 ± 28.5	3.6 ± 1.7	24.5 ± 3.9	233.3 ± 24.4	4.8 ± 2.4
	SCORE	3.4 ± 1.7	14.8 ± 7.0	1.3 ± 0.9	24.7 ± 3.7	209.4 ± 29.3	4.2 ± 1.2
	CAM	7.7 ± 2.6	114.0 ± 37.8	5.8 ± 1.8	29.2 ± 3.2	271.5 ± 38.4	10.1 ± 5.2
	GES	26.8 ± 3.5	203.5 ± 48.5	–	59.0 ± 4.1	350.1 ± 10.6	–

表12：伽马噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.7 ± 1.2	6.0 ± 12.0	0.6 ± 1.2	7.5 ± 2.6	36.4 ± 12.4	0.8 ± 1.2
	SCORE	0.7 ± 0.9	3.6 ± 4.7	0.5 ± 0.7	7.2 ± 2.5	41.3 ± 7.9	0.6 ± 1.0
	CAM	2.8 ± 2.4	19.3 ± 16.7	2.1 ± 1.8	10.4 ± 3.2	48.8 ± 14.8	4.2 ± 3.3
	GES	10.5 ± 2.5	58.5 ± 14.7	–	25.6 ± 2.3	80.4 ± 3.7	–
d=20	NoGAM	6.1 ± 2.3	71.5 ± 40.0	4.2 ± 1.9	26.6 ± 5.0	237.8 ± 21.4	4.3 ± 2.3
	SCORE	3.1 ± 2.1	13.2 ± 10.2	0.8 ± 0.6	27.1 ± 4.1	204.4 ± 23.3	3.7 ± 1.7
	CAM	8.6 ± 2.7	100.0 ± 38.3	5.6 ± 1.6	29.0 ± 6.6	270.7 ± 41.7	8.6 ± 4.3
	GES	27.1 ± 3.1	187.3 ± 31.1	–	59.8 ± 5.1	348.6 ± 11.1	–

表13：高斯噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.3 ± 0.5	2.0 ± 4.0	0.2 ± 0.4	6.4 ± 1.9	34.1 ± 5.3	0.0 ± 0.0
	SCORE	0.3 ± 0.6	2.7 ± 5.8	0.1 ± 0.3	7.6 ± 2.9	31.6 ± 9.4	1.1 ± 0.7
	CAM	0.2 ± 0.4	1.5 ± 3.0	0.0 ± 0.0	9.8 ± 2.3	39.7 ± 9.9	1.0 ± 0.9
	GES	10.6 ± 3.0	49.7 ± 15.2	–	24.1 ± 3.3	80.7 ± 3.5	–
d=20	NoGAM	1.2 ± 0.9	14.2 ± 16.3	0.7 ± 0.6	15.8 ± 4.6	224.6 ± 22.4	2.3 ± 1.6
	SCORE	0.9 ± 0.9	13.8 ± 12.6	0.7 ± 0.8	17.5 ± 3.5	179.2 ± 23.8	4.9 ± 3.0
	CAM	0.3 ± 0.5	1.9 ± 5.7	0.0 ± 0.0	24.8 ± 3.3	240.7 ± 29.8	3.1 ± 2.4
	GES	28.1 ± 7.6	212.6 ± 51.0	–	58.5 ± 3.6	360.5 ± 2.7	0.0 ± 0.0

表14：Gumbel噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.9 ± 0.9	6.0 ± 5.2	0.6 ± 0.7	7.2 ± 2.3	28.1 ± 7.5	0.2 ± 0.4
	SCORE	1.2 ± 1.3	5.0 ± 5.1	0.7 ± 0.6	7.6 ± 3.2	44.0 ± 9.3	0.5 ± 0.5
	CAM	0.4 ± 0.5	4.0 ± 4.9	0.4 ± 0.5	8.0 ± 3.0	37.5 ± 13.4	1.5 ± 2.3
	GES	10.5 ± 2.6	55.1 ± 13.0	–	24.0 ± 3.4	80.3 ± 6.1	–
d=20	NoGAM	2.1 ± 1.4	26.5 ± 20.0	1.5 ± 0.9	25.3 ± 3.7	237.3 ± 21.0	2.7 ± 2.0
	SCORE	3.2 ± 1.9	20.0 ± 13.6	1.1 ± 1.4	28.1 ± 4.1	212.3 ± 19.7	5.6 ± 2.4
	CAM	1.5 ± 1.6	13.7 ± 16.6	0.8 ± 0.8	26.0 ± 4.5	247.0 ± 31.6	4.7 ± 2.2
	GES	28.0 ± 8.7	195.9 ± 34.7	–	58.9 ± 4.8	353.3 ± 12.7	–

表15：拉普拉斯噪声无标度图

		SF1（稀疏）			SF4（稠密）		
	方法	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.1 ± 0.3	0.9 ± 2.7	0.1 ± 0.3	8.0 ± 2.1	30.4 ± 8.3	0.5 ± 0.9
	SCORE	0.5 ± 1.2	1.5 ± 3.4	0.0 ± 0.0	8.2 ± 1.7	43.6 ± 6.5	0.5 ± 0.7
	CAM	0.4 ± 0.5	1.9 ± 3.8	0.1 ± 0.3	7.0 ± 2.2	34.0 ± 12.1	1.0 ± 1.3
	GES	12.8 ± 4.7	60.4 ± 19.7	–	26.1 ± 1.9	82.4 ± 2.8	–
d=20	NoGAM	1.7 ± 0.9	12.6 ± 8.6	0.7 ± 0.6	25.9 ± 3.3	244.3 ± 22.9	2.5 ± 2.3
	SCORE	2.6 ± 1.1	16.7 ± 15.7	1.1 ± 1.0	27.4 ± 3.0	207.6 ± 18.4	4.8 ± 2.2
	CAM	1.1 ± 0.9	3.9 ± 7.8	0.1 ± 0.3	23.7 ± 3.9	227.2 ± 15.1	2.5 ± 1.8
	GES	29.0 ± 6.7	207.3 ± 34.7	–	58.0 ± 6.6	353.4 ± 7.7	–

表16：使用核岭回归和套索回归方法估计残差，以及根据残差估计得分函数时，NoGAM 的实验性能。（粗体标示性能最佳的回归方法。）

噪声	节点	回归方法	D_{top}
Beta	$d = 10$	**核岭回归**	0.1 ± 0.3
	$d = 10$	套索回归	0.2 ± 0.4
	$d = 20$	**核岭回归**	0.9 ± 0.9
	$d = 20$	套索回归	4.6 ± 2.9
伽马	$d = 10$	核岭回归	0.7 ± 0.8
	$d = 10$	**套索回归**	0.6 ± 0.7
	$d = 20$	**核岭回归**	3.7 ± 1.6
	$d = 20$	套索回归	4.0 ± 1.9
高斯	$d = 10$	核岭回归	0.8 ± 0.9
	$d = 10$	**套索回归**	0.6 ± 0.7
	$d = 20$	**核岭回归**	3.1 ± 2.0
	$d = 20$	套索回归	3.7 ± 1.7

噪声	节点	方法	D_{top}
贡贝尔分布	$d = 10$	**核岭回归**	0.6 ± 0.7
	$d = 10$	套索回归	0.8 ± 1.0
	$d = 20$	核岭回归	4.7 ± 2.4
	$d = 20$	**套索回归**	3.8 ± 2.2
指数分布	$d = 10$	核岭回归	0.8 ± 0.8
	$d = 10$	**套索回归**	0.6 ± 0.7
	$d = 20$	核岭回归	3.6 ± 1.6
	$d = 20$	**套索回归**	3.5 ± 1.7
拉普拉斯	$d = 10$	核岭回归	0.8 ± 0.7
	$d = 10$	**套索回归**	0.7 ± 0.6
	$d = 20$	核岭回归	3.9 ± 1.5
	$d = 20$	**套索回归**	3.1 ± 1.5