arXiv:2304.03382v1 [cs.LG] 6 Apr 2023

基于分数匹配的可扩展因果发现

Francesco Montagna* FRANCESCO.MONTAGNA@EDU.UNIGE.IT
热那亚大学 *MaLGa-DIBRIS*

Nicoletta Noceti NICOLETTA.NOCETI@UNIGE.IT
热那亚大学 *MaLGa-DIBRIS*

Lorenzo Rosasco LROSASCO@MIT.EDU
热那亚大学 *MaLGa-DIBRIS*、
MIT CBMM、意大利技术研究院

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

Francesco Locatello LOCATELF@AMAZON.COM
AWS

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

摘要

本文展示了如何从观测到的非线性加性高斯噪声模型中对数似然的二阶导数发现完整因果图。通过利用可扩展的机器学习方法近似分数函数 ∇对数 $p(\mathbf{X})$，我们拓展了 Rolland等人 (2022) 的研究。该研究只能从分数中恢复拓扑序，并且需要一个计算开销很大的剪枝步骤，以移除排序所推断出的伪边。我们的分析得出了 DAS（Discovery At Scale 的缩写），这是一种实用算法，可将剪枝的复杂度降低与图规模成比例的倍数。实践中，DAS 在速度快一个数量级以上的同时，实现了与当前最先进方法相当的准确率。总的来说，我们的方法实现了有理论依据且可扩展的因果发现，显著降低了计算门槛。**关键词：** 高维因果发现；分数匹配；可扩展性

1. 引言

从观测数据中进行因果发现，是影响几乎所有科学领域的核心问题，例如生物学、遗传学、经济学和机器学习（Sachs 等人 (2005年)；Koller 和 Friedman (2009年)；Pearl (2009年)；Peters 等人(2017年)）。给定一个因果模型，无需获取成本高昂、难以实施或在伦理上不宜收集的干预数据，便可以预测干预对系统变量的结果。另一方面，从观测数据中推断因果关系通常不可识别，并且需要额外的假设。在传统因果研究中，从观测值中发现因果关系的算法可分为三类（Glymour 等人 (2019年)；Sch¨olkopf 等人 (2021年)）。基于约束的方法，如 PC（Spirtes 和 Glymour (1991)）、FCI 和 SGS（Spirtes 等人 (2000年)），会检验变量之间的条件独立性，并在忠实性假设下搜索满足这些条件的图结构

* 部分工作是在亚马逊云科技蒂宾根实习期间完成的

© 2023 F. MONTAGNA、N. NOCETI、L. ROSASCO、K. ZHANG 和 F. LOCATELLO。

。通常，它们不会输出唯一的图，而是输出一个等价类。这类方法的主要瓶颈在于，条件独立性检验众所周知地难以进行（Shah 和 Peters (2020年）。基于分数的方法会定义合适的分数函数，并搜索与数据拟合度最高的图。贪心方法（如 GES（Chickering (2002））） 用于在这个庞大的搜索空间中进行搜索。该空间的规模会随着节点数量的增加而超指数级增长（因此限制了可扩展性）。最后，受限模型类假设（例如非线性关系和加性高斯噪声）可以识别观测值背后的有向无环图（DAG）（Peters 等人 (2017年）；B¨uhlmann 等人 (2014）；Lachapelle 等人 (2020年）；Zheng 等人 (2018年）；Shimizu 等人(2006年）。

影响因果图发现的一个主要挑战是，执行 DAG约束会使每次迭代的成本随变量数量呈立方增长，从而使优化成为计算瓶颈。降低计算需求的一种方法是将因果发现任务解耦为两个步骤：首先找到一个拓扑排序，使得一个节点只能作为该排序中其后继节点的父节点，从而确保无环约束。然后，通过剪枝步骤，在推断出的排序所允许的边中选择正确的子集，移除图中的所有伪连接。在这种情况下，Rolland等人Rolland等人(2022) 最近提出了 SCORE 算法，这是提升可扩展性的一项工作：首先，他们高效地估计分数函数 $\nabla \log p(\mathbf{X})$，然后从分数函数的雅可比矩阵中恢复拓扑排序，最后采用 CAM 中提出的方法对全连接 DAG 进行剪枝（B¨uhlmann等人(2014）） 。剪枝步骤是 SCORE 的瓶颈，在包含 95% 个节点的图上占运行时间的 50 ，并且其成本随节点数量呈立方增长。

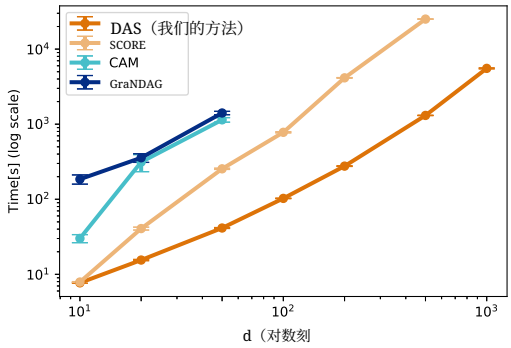


图 1：不同方法在稠密图（ER4 数据）上的执行时间与 d 节点数的关系。

在本研究中，我们表明，对数似然的二阶导数能够恢复因果有向无环图中的骨架和边的方向。从理论上讲，这意味着我们可以省去 SCORE 中代价高昂的剪枝步骤（Rolland等人(2022）），因为关于因果结构的所有信息都已包含在得分函数的雅可比矩阵中。尽管我们的分析得出了一种实用的算法，但我们发现，先识别少数候选边并保留最后一个简单的剪枝步骤更为有益。这样效率高得多，因为大多数边已经被检测出来，只需纠正得分函数有限样本近似所导致的错误，从而将复杂度降低至与图中节点数成反比。这个优势在实际中显而易见——

tice，将 SCORE 的运行时间缩短了一个数量级以上，如图1所示。我们的贡献总结如下：

- 我们演示了如何根据数据分布的 SCORE从理论上恢复完整的因果有向无环图。这扩展了先前的工作，该工作表明可以从 SCORE 中恢复拓扑序（Rolland 等人 (2022）。

• 我们介绍 DAS[1]（Discovery At Scale 的缩写），这是一种用于高效、可扩展因果发现的算法。在降低因果发现技术应用于大量变量时的计算门槛的同时，我们还提供了清晰的示例，并指出了算法可能失效的情形。我们希望这些内容能帮助有意在自身数据上使用 DAS 的实践者，以及从事可扩展因果发现研究的人员。

• 我们的分析得出了一种实用方法，可在最后的剪枝步骤中筛选候选边。尽管我们的方法准确度略低于（Rolland等人 (2022)），但在图规模方面，运行时间缩短了一个数量级。我们通过最多包含一千个节点的合成图，展示了速度提升效果。

2. 背景知识

我们从结构因果模型的形式化定义入手，介绍分析所需的背景知识。

2.1. 结构因果模型

形式化变量之间因果关系的一种方法是使用加性结构因果模型（SCM）。考虑有向无环图中的一组$\mathbf{X} = \{X_i\}_{i=1}^d$可观测顶点。我们假设图的结构可以用以下函数关系表示

$$X_i = f_i(\mathrm{pa}_i(\mathbf{X})) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i), \quad \forall i = 1, \ldots, d \ , \tag{1}$$

其中 $\mathrm{pa}_i(\mathbf{X})$ 表示有向网络中 X_i 的父节点集合。我们假设 $X_i \in \mathbb{R}$、可加且独立抽取的高斯噪声元素 ϵ_i，以及 f_i 均二阶连续可微，且在每个分量上都是非线性的。

递归应用 (1) 可以推导出联合概率分布 $p(X_1, X_2, \ldots, X_d)$。由于此概率定义在有向无环图的顶点上，因此有如下因子分解（Pearl (2009年)；Lauritzen (1996)）：

$$p(\mathbf{X}) = \prod_{i=1}^{d} p(X_i | \, \mathrm{pa}_i(\mathbf{X})) \ . \tag{2}$$

加性噪声模型（1）已知在温和的假设下（Hoyer 等人（2009年）；Zhang 和 Hyv¨arinen （2009年）），这意味着我们可以从根据关于$\mathbf{X}$的联合分布生成的观测数据中唯一地恢复因果图。

问题定义 通常，模型中的 f_i （1）形式是未知的，式（2）中的概率也同样未知，而我们只能获取联合分布的一组观测值。给定这些观测值，任务是识别 SCM 所对应图的因果结构。这个问题称为因果发现。一种解决方案是利用数据估计 $\mathbf{X}$ 中变量的拓扑排序，然后根据该排序在 DAG 中选择边。在我们的方法中，我们选择满足由 SCORE 的雅可比矩阵推导出的约束的边。

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

2.2. 图的拓扑排序

给定一组节点 X $= \{X_i\}_{i=1}^d$，令 $\mathcal{G} =$ (X $\mathcal{E}$) 为一个有向无环图。相对于 $\mathcal{G}$的拓扑序是节点的一个排列 $\mathbf{X}^\pi$，使得 $\mathcal{G}$ 中的边（i, j）意味着 X_i 在 $\mathbf{X}^\pi$ 中出现在 X_j 之前，记作 $X_j \succ X_i$。直观地说，在该排序中，一个节点只能是其后继节点的父节点。根据这一定义，对于给定的有向无环图，允许存在多个拓扑排序。另一方面，给定的排列 $\mathbf{X}^\pi$ 与完全连接图 $\mathcal{G}^\pi$ 之间存在一一对应关系；对于每个 i、$j = 1$、.....、d，只要 $X_j \succ X_i$ 位于 $\mathbf{X}^\pi$ 中，我们就在其中绘制边（i, j）。如果 $\mathbf{X}^\pi$ 等于 $\mathbf{X}^*$，即真实有向无环图 $\mathcal{G}$ 的一个正确排列，那么 $\mathcal{G}^\pi =$ (X $\mathcal{E}^\pi$) 是 $\mathcal{G}$ 的超图，这意味着其边集合满足 $\mathcal{E} \subseteq \mathcal{E}^\pi$。下一节将概述 CAM边选择过程，该过程能够在给定拓扑排序 $\mathbf{X}^\pi$ 及其对应的稠密集合 $\mathcal{E}^\pi$的情况下，正确识别 $\mathcal{E}$。

2.3. 初步邻居搜索（PNS）和 CAM剪枝

现在我们简要介绍CAM 的两步剪枝方法（B¨uhlmann 等, （2014）），即初步邻居搜索（PNS），随后进行*CAM*剪枝，用于移除虚假边，即 $\mathcal{G}^\pi$由排列$\mathbf{X}^\pi$诱导的稠密图中的边。

PNS PNS 是一种针对非线性加性高斯噪声模型的邻居选择方法，遵循Meinshausen和 B¨uhlmann（2006年）针对线性高斯情形提出的图中的变量选择思想。具体而言，给定一个顺序 $\mathbf{X}^\pi$，通过对每个 $j = 1, \ldots, d$ 拟合一个以 X_j 为因变量、以所有其他变量 $\{X_i$ 为自变量的加性模型来执行变量选择： $X_j \succ X_i$ in$\mathbf{X}^\pi\}$，并选择 K 个最重要的预测变量作为 X_j 的可能父节点。这将 X_j 的可能入边数量限制为任意固定值 K，从而降低边搜索过程的总体复杂度以及 CAM剪枝步骤的工作量。作者使用加性模型拟合的提升方法实现了PNS（B¨uhlmann 和 Hothorn（2007）；B¨uhlmann和 Yu（2003））。因此，PNS 的总体复杂度为 $\mathcal{O}(dt\mathbf{r}(n, d))$，因为对于每个节点 1,. . . , d，提升算法都会拟合 t 个复杂度为 $\mathbf{r}(n, d)$ 的模型，其中 n 为数据集中的样本数量。复杂度项$\mathbf{r}(n, d)$取决于所选的加性模型拟合技术，使用迭代加权最小二乘法（Minka (2003)）时为 $\mathcal{O}(nd^2)$。

CAM剪枝 在初步邻居搜索之后，通过对协变量进行显著性检验来执行最后的剪枝步骤，以移除多余的边，从而减少假阳性的数量。在这种情况下，计算复杂度可以忽略不计，因为它受 PNS 中确定的固定参数 K 限制。

我们的工作旨在用一种新方法取代 PNS，使 CAM 剪枝能够应用于高维图。具体而言，我们将边搜索过程的渐近复杂度从 $\mathcal{O}(nd^3)$降低到 $\mathcal{O}(d^2)$。

3. 从分数中推导因果结构

对于正在分析的因果发现问题，我们考虑一个可观测的$\mathbf{X} \in \mathbb{R}^d$，其条目$X_i$ 是根据模型（1）生成的图的顶点。在下一节中，我们将展示，分数函数原则上足以解决这一任务。首先，我们阐述 SCORE 算法背后的思想，该算法可以估计分数函数的雅可比矩阵，并找到该

感兴趣的变量。然后，我们推导出关于得分函数雅可比矩阵的其他约束，从而能够识别因果图中的边。

3.1. SCORE 概述

Rolland等人 (2022) 提出了一种高效估计分数函数 $s(\mathbf{X})$ 及其雅可比矩阵的方法，该方法利用了 Stein 恒等式（Stein (1972)）。此外，他们提出了一种方法，通过检查分数函数雅可比矩阵的对角元素，识别根据（1）生成的因果图中的叶节点。

为了推导叶节点识别步骤，我们首先需要求出 $s(\mathbf{X}) = \nabla \log p(\mathbf{X})$的闭式表达。从方程（2）出发，我们有：

$$\begin{aligned} \log p(\mathbf{X}) &= \sum_{i=1}^{d} \log p(X_i | \operatorname{pa}_i(\mathbf{X})) = \\ &= -\frac{1}{2} \sum_{i=1}^{d} \left(\frac{X_i - f_i(\operatorname{pa}_i(\mathbf{X}))}{\sigma_i^2} \right)^2 - \frac{1}{2} \sum_{i=1}^{d} \log(2\pi\sigma_i)^2 \ . \end{aligned} \tag{3}$$

j 阶条目 $\nabla \log p(\mathbf{X})$ 因此为

$$s_j(\mathbf{X}) = -\frac{X_j - f_j(\operatorname{pa}_j(\mathbf{X}))}{\sigma_j^2} + \sum_{i \in \operatorname{ch}_j(\mathbf{X})} \frac{\partial f_i}{\partial x_j}(\operatorname{pa}_i(\mathbf{X})) \frac{X_i - f_i(\operatorname{pa}_i(\mathbf{X}))}{\sigma_i^2} \ , \tag{4}$$

其中，$ch_j(\mathbf{X})$ 表示节点的子节点集合 X_j。现在，设 X_j 为一个叶节点：其子节点集合 $ch_j(\mathbf{X}) = \emptyset$为空。根据方程（4），我们注意到分数函数可简化如下：

$$s_j(\mathbf{X}) = -\frac{X_j - f_j(\operatorname{pa}_j(\mathbf{X}))}{\sigma_j^2}. \tag{5}$$

此外，很容易验证 $\frac{\partial s_j(\mathbf{X})}{\partial x_j} = -\frac{1}{\sigma_j^2}$，使得与叶节点相关的分数函数的雅可比矩阵的对角元素为常数。基于这一关系，Rolland等人（2022）的引理 1 提供了在因果图中识别叶节点的形式化判据。给定一个节点 X_i，有：

$$\frac{\partial s_i(\mathbf{X})}{\partial x_i} = c \iff X_i \text{ is a leaf}, \ \forall i = 1, \ldots, d, \tag{6}$$

其中 $c \in \mathbb{R}$ 是一个常数标量值。该关系直接意味着

$$\operatorname{Var}\left[\frac{\partial s_i(\mathbf{X})}{\partial x_i} \right] = 0 \iff X_i \text{ is a leaf}, \ \forall i = 1, \ldots, d. \tag{7}$$

为了找到完整的拓扑排序，Rolland等人（2022）的 SCORE 算法ˆ设计如下：首先，它估计分数函数的雅可比矩阵 J（s（$\mathbf{X}$）），并通过（7）利用该矩阵识别图中的叶节点。然后，它从图中移除该叶节点，并为其在排序向量中分配一个位置。通过反复执行这两个步骤，直到源节点为止，$\mathbf{X}$中的所有变量最终都会在因果排序中被分配一个位置。

在以下部分中，我们将展示，在已知拓扑排序的情况下，可以推导出关于分数函数的雅可比矩阵的非对角元素的附加约束，从而识别图中的有向边。

3.2. 推导边选择的约束

我们观察到，对于叶节点 l，$X_l \in \mathbf{X}$，(4) 对 X_j 在 $j \neq l$ 下的偏导数为：

$$\frac{\partial s_l(\mathbf{X})}{\partial X_j} = \begin{cases} \frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X})) \neq 0 & \text{if } X_j \in \mathrm{pa}_l(\mathbf{X}) \\ 0 & \text{else} \end{cases} . \tag{8}$$

值得注意的是，$\frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))$ 对于某些$\mathrm{pa}_l(\mathbf{X})$ 的取值仍可能趋于零，$X_j \in \mathrm{pa}_l(\mathbf{X})$即使在受限区间上考察时仍为非线性；例如，当函数达到最大值或最小值时，就可能出现这种情况：根据关于 f_l 的假设，这些事件发生的概率为零，因此 $\frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X})) \neq 0$ 几乎必然成立几乎必然。我们证明，稍作改造8式中的条件即可导出一项判据，用于识别给定叶节点的父节点。该结果是对Rolland等人(2022)

引理 1（改编自 Rolland等人 (2022)） 设 p 是随机变量 $\mathbf{X} \in \mathbb{R}^d$ 通过非线性加性高斯噪声模型 (1)定义的概率密度函数。另设$s(\mathbf{X}) = \nabla \log p(\mathbf{X})$ 为相应的分数函数。不失一般性，假设存在拓扑排序 $\mathbf{X}^\pi = (X_1, \ldots, X_d)$。那么给定一个叶节点 *l:</style}*

$$\mathbf{E}\left[\left|\frac{\partial s_l(\mathbf{X})}{\partial X_j}\right|\right] \neq 0 \iff X_j \in \mathrm{pa}_l(\mathbf{X}),\ \forall j \in \{1, \ldots, l-1\} . \tag{9}$$

为完整起见，附录中给出了证明A。

引理 1 与 Rolland 等人（2022）的差异 Rolland 等人（2022）提出的公式要求 Var $\left[\frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))\right] \neq 0 \Leftrightarrow X_j \in \mathrm{pa}_l(\mathbf{X})$，其中 X_l 是叶节点。我们通过一个简单的双变量情形来说明其中的问题，设图为 $X_1 \longrightarrow X_2$：如果父节点 X_1 的方差为零，他们的选择条件就会失效，从而预测出一个包含 X_1 且 X_2 相互独立的图。虽然每个节点的方差都大于零这一假设可以排除这种情形，但在实践中仍可能出现问题。给定有限样本 $X \in \mathbb{R}^{n \times d}$ 及其拓扑排序 $\mathbf{X}^\pi$，如果叶节点 X_l 的父节点在样本中的方差较小，我们仍可能把在 $\frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))$ 中观察到的波动误认为有限样本估计导致的统计误差，从而舍弃一条实际存在的边。与拓扑排序相比，估计边需要更高的统计准确度，因为在前一种情况下，我们不知道一个节点有多少个父节点；而在后一种情况下，我们知道至少始终存在一个叶节点。因此，我们对 Rolland 等人（2022）的引理进行了改写，使其在理论上等价，但在实践中更为稳健。我们在实现引理 1 时，采用分数函数的雅可比矩阵元素 $\frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))$ 的绝对值的样本均值：这种估计量可能面临相同的问题，但表现出比样本方差更好的稳健性（得益于绝对值），且估计低阶矩会带来较低的误差（估计方差需要先估计均值，因此均值估计量中的任何统计误差都会影响方差估计量），所以这是更可取的选择。

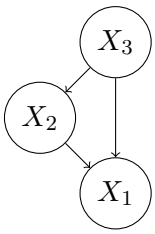
在实际中，只有在提供排序 1 的情况下，我们才能利用引理重建整个图$\mathbf{X}^\pi$。为了解其原因，考虑最后一项 X_l 在 $\mathbf{X}^\pi$中：根据拓扑排序的定义，X_l 是一个叶节点。然后我们可以应用引理 1 对 $s_l(\mathbf{X})$ 关于所有节点 $\{X_j$ 求偏导数：$X_j \prec X_l$ 在 $\mathbf{X}^\pi\}$ 并将满足所需约束的节点识别为父节点。此时，我们移除

X_l 从排序$\mathbf{X}^\pi$ 开始，并在顶点为$\mathbf{X} \setminus X_l$的剪枝后图上重复该过程。通过对排序中的每个节点从最后一个节点到源节点迭代这些步骤，我们可以识别出精确的图。

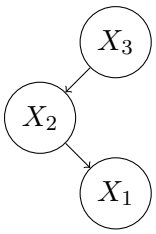
例 1 为阐明这些想法，我们讨论一个简单的三变量示例，说明引理中的结果如何*1*实现边的发现。

设 $\boldsymbol{X} = (X_1, X_2, X_3)$，其中 $X_i \in \mathbb{R}$ 根据模型（*1*）生成。假设给定拓扑排序 $\boldsymbol{X}^\pi = (X_3, X_2, X_1)$。目标是恢复真实的因果图，其中 X_3 为源节点，$pa_1(X) = \{X_2\}$ 和 $pa_2(X) = \{X_3\}$。

图 $\mathcal{G}^\pi$ 由$\mathbf{X}^\pi$接纳



(b) $\mathbf{X}$的正确因果图



我们继续分析 X_1 和 X_2 的分数函数的雅可比矩阵，同时从排序$\mathbf{X}^\pi$ 中很容易看出，X_3 没有父节点。

- X_1：根据(5)，其得分仅为 $s_1(\mathbf{X}) = -^{X_1 - f_1(X_2)}\sigma_1^2$。其非对角偏导数为

$$\frac{\partial s_1}{\partial X_2}(\mathbf{X}) = \frac{1}{\sigma_1^2}\frac{\partial f_1}{\partial X_2}(X_2) \ ,$$
$$\frac{\partial s_1}{\partial X_3}(\mathbf{X}) = 0 \ .$$

应用引理 *1* 后，我们可以正确地得出结论，X_1 的父节点集合为 $\mathrm{pa}_1(\mathbf{X}) = \{X_2\}$。

- X_2：为了也将引理*1* 应用于 X_2，我们需要从图中删除 X_1，使得 X_2 是一个叶节点。给定变量集合$\mathbf{X} = \mathbf{X} \setminus \{X_1\}$，我们得到图 *3*。然后 $s_2(\tilde{\mathbf{X}}) = -^{X_2 - f_2(X_3)}\sigma_2^2$，且偏导数为

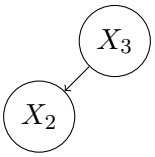
$$\frac{\partial s_2}{\partial X_3}(\tilde{\mathbf{X}}) = \frac{1}{\sigma_2^2}\frac{\partial f_2(X_3)}{\partial X_3} \ .$$

By Lemma 1 we see that $\mathbf{E}\left[\left|\frac{1}{\sigma_2^2}\frac{\partial f_2(X_3)}{\partial X_3}\right|\right] \neq 0$, *such that* X_3 *is a parent of* X_2.

该算法恢复了精确的结构。

讨论 鉴于先前关于马尔可夫网络的研究成果，基于分数函数识别因果结构的这些发现并不完全令人意外（Spantini 等人 (2018年)；Morrison 等人 (2017年)）。给定一组随机变量 $\mathbf{X} = (X_1, X_2, \ldots\ldots, X_d)$，其联合密度为 $p(\mathbf{X})$，变量$\mathbf{X}$ 之间条件独立关系的信息可以

˜图3：$\mathbf{X} = $的因果图（$X_2$，$X_3$）。



嵌入一个简单的无向马尔可夫网络 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中边 (i, j) 编码了随机变量对之间某种概率相互作用 X_i，X_j。特别是，Spantini 等人 (2018年证明了如何通过读取随机变量对的条件独立性来构造马尔可夫图，如下所示：

$$X_j \perp\!\!\!\perp X_i \,|\, \mathbf{X}_{\mathcal{V}\setminus\{i,j\}} \iff \partial_{ij} \log p(\mathbf{X}) = 0 \ , \tag{10}$$

其中，$\partial_{ij}(\cdot)$ 表示 ij阶混合偏导数，$\partial_{ij} \log p(\mathbf{X})$ 是得分函数雅可比矩阵的一个元素。通过在每一对不满足公式（10）的节点之间添加边，我们得到一个无向图，它编码了变量$\mathbf{X}$之间所有且仅有的条件独立关系。

本文的公式（8）在略有不同的设定下发现了相同的约束：我们不再针对每个节点逐一评估 $\partial_{ij} \log p(\mathbf{X})$ 与其他所有节点的关系，而是采用迭代方法：首先识别一个叶节点 X_l，然后仅针对拓扑排序中排在它之前的节点检验其混合偏导数。找到一条边时，我们便知道它的方向，因为已知 X_l 是叶节点，这打破了关系中的对称性。此外，引理1 确保能正确识别有向$\mathcal{V}$-结构，例如 $i \to j \leftarrow k$；而在条件独立图中，这些结构会被道德化，并添加一条连边(i, k)，从而能够恢复所有边及其方向。

接下来，我们基于引理推导一种因果发现算法1，并展示它在性能方面如何保持与其他最先进方法相当，同时在节点数量方面展现出更好的可扩展性。

3.3. DAS：一种可扩展的因果发现算法

我们希望利用引理1 对分数函数施加的约束，推导出一种用于因果发现的算法；据我们所知，与其他任何技术相比，该算法速度更快，并且随着节点数量增加表现出更好的扩展性。给定一组根据 ˆ 模型（1）生成的 $X \in \mathbb{R}^{n\times d}$ 个 n 观测值，我们首先通过 SCORE算法估计拓扑排序 $\mathbf{X}^{\pi}$ 。然后，我们从最后一个节点开始，遍历该排序中的每个节点 X_l；我们知道最后一个节点是叶节点，ˆ 并按如下方式识别 X_l 的入边：首先，我们通过SCORE算法估计分数函数的雅可比矩阵 J；现在，对于每个 n 数据点，我们都有一个分数函数雅可比矩阵的 n 个估计值。根据引理 1，我们考察 n 雅可比矩阵的第 l行的绝对值，并寻找具有非零均值的条目：这可以通过统计假设检验实现，其思想是检验样本均值是否不同于零。在实践中，我们按照以下启发式方法执行假设检验：计算第 l行中各条目绝对值的平均值，并将与最大平均值对应的 $K+1$ 节点选为 X_l的潜在父节点。该参数 K与我们在 PNS算法中发现的参数相同，它限制了输入剪枝算法的节点最大数量。在这些 $K+1$ 条目中，我们

找到参考节点 X_{ref}，其 $\left|\hat{J}_{l,ref}\right|$ 的样本平均值最接近零：此时，对于其余的 K 个节点，我们执行 Welch t 检验，原假设为总体均值相等 H_0： $\mathbf{E}[|J_{l,j}|] = \mathbf{E}[|J_{l,ref}|]$，备择假设为 H_1： $\mathbf{E}[|J_{l,j}|] > \mathbf{E}[|J_{l,ref}|]$，其中 X_j 是 X_l的潜在父节点。如果我们以 p 值 ≤ 0.01拒绝原假设，则将 X_j 添加到邻接矩阵中表示的推断图里 X_l 的父节点中。

最终， X_l 列将从数据矩阵 X中移除，并对排序中找到的另一个叶节点重复此过程。当 $\mathbf{X}^\pi$ 中的每个节点都经过考察后，我们通过*CAM*剪枝（第2.3节）对生成的邻接矩阵进行剪枝，目的仅仅是减少假阳性的数量。

DAS 的实现细节在算法1 .

3.4. 算法复杂度

考虑一个输入矩阵 $n \times d$，其中 n 表示样本数， d 表示节点数，则 DAS 的总体复杂度为 $\mathcal{O}(dn^3 + d^2)$。事实上，使用 SCORE 估计拓扑序需要求逆一个 $n \times n$矩阵，共进行 d次，每次迭代都用于识别一个叶节点：因此， $\mathcal{O}(dn^3)$ 构成了相应的开销。此外，边搜索步骤需要遍历排序中的 d 个元素，每次从一个大小为 K 的列表中选出最大的 $\leq d$ 个条目（见算法1），产生 $\mathcal{O}(d^2)$的开销。

另一方面，SCORE 可以说是用于因果发现的最具可扩展性的最先进算法，它使用 PNS 和 CAM剪枝来选择边，同时采用与 DAS 相同的排序。SCORE 执行过程中的瓶颈是初步的邻居搜索步骤，其复杂度已在第 2.3 节中详细研究，复杂度为 $\mathcal{O}(dtr(n,d))$，其中 $\mathrm{r}(n,d)$ 表示拟合广义加性模型所需的运算次数（ $\mathcal{O}(nd^2)$ (Minka (2003))，使用迭代加权最小二乘法）。因此，我们使用分数函数来选择候选边，大幅缩短了执行时间，使高维因果发现的规模扩展能力提升了 $\mathcal{O}(d)$。

线性模型的不可识别性 接下来，我们进一步展示算法的一致性，说明在假设线性 f_j 模型（1）中 $\forall j = 1, \ldots, d$时，它无法识别因果图。事实上，已有研究证明，根据线性加性高斯噪声模型生成的观测数据，无法恢复潜在的因果结构（Peters 等人 (2017)； Comon (1994)，除非作出额外假设（Peters 和 B¨uhlmann(2013)）。在仅有两个变量的最简单情形下， X, Y，它们在因果图中相连，不可识别性意味着无法确定边的方向，也就是说，无法判断X 是 Y的原因还是结果。

鉴于变量之间的拓扑排序本身消除了有向无环图中边方向的歧义，我们必须证明，在集合 $\mathbf{X} \in \mathbb{R}^d$ 上，我们的拓扑排序方法在线性设置下会失败。SCORE 算法通过寻找分数函数的雅可比矩阵对角线上方差为零的项来识别叶节点，如公式（7）所规定。在定义于（1）中但函数为线性的 SCM 情况下 f_j, 我们发现 $\frac{\partial s_j}{\partial X_j}(\mathbf{X}) = -\frac{1}{\sigma_j^2} +$ 对于图中的每个节点都是常数 j。因此很容易看出，

$$\mathrm{Var}\left[\frac{\partial s_j}{\partial X_j}(\mathbf{X})\right] = 0, \forall j = 1, \ldots, d \ . \tag{11}$$

由于式（11）中的方差对于每个节点 j 都会消失，而不仅仅是对于叶节点消失，因此方程（7）的判据不再成立。这意味着用于该问题的拓扑排序方法失效。

算法 1： DAS

输入：数据矩阵 $X \in \mathbb{R}^{n \times d}$, $K \in \mathbb{R}$

$X^{\pi} \leftarrow$ SCORE(X)（$X^{\pi}[d]$ 叶节点）$X^{\pi} \leftarrow reverse$（$X^{\pi}$）（$X^{\pi}[d]$源节点）

$A \leftarrow d \times d$ 全零邻接矩阵

对于 X_l 在 X^{π} **执行**$\hat{J} \leftarrow \left[\left|\frac{\partial s_l}{\partial X_j}\right|\right]_{X_j \prec X_l \text{ 在 } X^{\pi}}$（从 SCORE 估计）$K \leftarrow \min(K,$ length($\hat{J}$))$topK = \overline{topK}$（平均值($\hat{J}$)）（topK(·)：返回输入中 K 个最大值的索引）

$j_{ref} = topK[\text{argmin } J[topK]]$**对于**$k$ 在 $topK$ **执行**$p =$ 检验的 p 值 $H_0 : \mu_{J_k} = \mu_{J_{ref}}, H_1 : \mu_{J_k} > \mu_{J_{ref}}$（$\mu$ = 均值）**如果** $p < 0.01$**则**$A[k, l] = 1$**结束 结束**从 X中移除第 l 列

结束$\hat{\mathcal{G}} \leftarrow$CAM剪枝($A$) **返回** $\mathcal{G}$

与我们的论断一致的线性情形。

在下一节中，我们将研究 DAS 的算法复杂度，并强调其相较于 SCORE 的更高效率。

4. 实验

现在我们总结DAS方法与几种用于因果发现的最先进算法进行比较的实验结果。我们报告了采用PNS和CAM-剪枝步骤的SCORE-排序结果（表中简称为SCORE）（Rolland等人(2022)）、CAM [2]（B¨uhlmann等人（2014））和GraNDAG（Lachapelle等人（2020年））。其他算法（如PC和FCI）因其表现差得多而被省略（B¨uhlmann等人（2014）；Lachapelle等人（2020年））。对于最多200 个节点，我们在一台配备16GB内存和8个Intel(R) Core(TM) i5-8265U CPU处理器（1.60GHz）的机器上进行了实验。对于500 个或更多节点，我们使用了一台配备256GB内存和64个AMDEPYC 7301 16-Core Processor处理器（2.20GHz）的机器。

使用的指标包括精确率、召回率、结构汉明距离（SHD）——其计算方式为假阳性、假阴性和方向错误的边之和——以及结构干预距离（SID）（Peters 和 B¨uhlmann (2015年)），后者衡量由推断图导致的错误干预分布数量。

我们重点关注 在合成数据上进行实验，以展示 DAS 的可扩展性特性，并不断增—

2. CAM 指原论文中提出的拓扑排序和剪枝步骤。PNS 仅针对 $d \geq 50$ 应用。

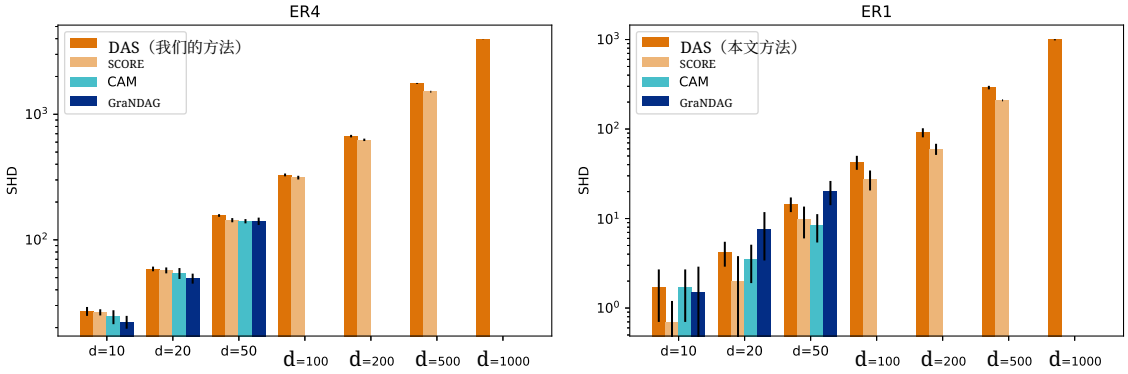


图 4：不同方法在稠密（左）和稀疏（右）图上的 SHD 与 d 节点数量的关系。当 d 的值较高时，部分方法未显示，因为运行这些方法耗时过长。样本数量为 $n = 1000$。

节点数量。为了从方程（1）的非线性加性高斯噪声模型中采样数据，我们模拟 Rolland等人 (2022) 的实验设置。因果图使用 Erd¨os-Renyi 模型（Erdos 和 Renyi (1960）生成。我们进行实验时固定节点数量 d，并通过将边的期望数量设为 d（ER1）或 $4d$（ER4）来设定图的稀疏度。对于 $d > 200$，我们省略 SID 指标，因为计算速度太慢。如果表格中没有出现某种方法的结果，这意味着我们无法在合理的时间内完成这些运行。我们将实验重复 10 次，并报告各项指标的经验均值和标准差。样本数量固定为 $n = 1000$，并设定 $K = 20$（与 SCORE 和 CAM 实验设置中采用的值相同）。

从表 1 可以看出，在较稠密的图（ER4）上，对于最多 50 个节点的图，我们的方法与其他三种方法的性能相近，同时速度显著更快，尤其是相较于 GraNDAG 和 CAM。随着 d 增加，与 SCORE 的差距逐渐缩小，直到200 个节点时，我们的算法在 SID 上表现更佳。在 $d \geq 500$ 的情况下，想要在个人电脑上于有限时间内运行 SCORE 几乎是不可能的，而 DAS 是唯一合理的选择。
如表 2所示，在较稀疏的图（ER1）上运行推断时，不同方法的性能同样相当。这些结果可直接从图 4中看出：随着节点数量增加，每种算法的性能都以相近幅度下降，且各柱形对应的 SHD 值相近。尽管如此，图 1 清楚地显示，DAS 能以显著更短的时间达到这些指标，支持了我们的方案在速度和可扩展性方面效率更高的主张。

在附录 B中，我们提供了更多实证结果，重点关注 DAS 对假设检验阈值（均值大于零）的变化的稳定性。此外，我们大幅扩展了在无标度合成图（Barabasi 和 Albert(1999）上测试 DAS 的实验（附录 C），以及在 Sachs 真实数据（Sachs 等人 (2005年）和从 SynTReN 生成器采样的半合成数据（Van den Bulcke 等人(2006年））上的实验，详见附录 D）。

表1：ER4数据上的实验。对于CAM和GraNDAG，我们报告了Rolland等人(2022)中给出的结果。对于较高的 d，部分方法因运行耗时过长而未能纳入。

	方法	SHD	SID	精确率	Rec.	时间 [s]
d=10	DAS（我们的方法）	27.0 ± 2.2	43.6 ± 5.8	1.00 ± 0.00	0.33 ± 0.02	$\mathbf{7.7 \pm 0.1}$
	SCORE	26.5 ± 1.5	42.3 ± 2.9	0.99 ± 0.00	0.33 ± 0.02	7.9 ± 0.1
	CAM	24.4 ± 3.1	45.2 ± 10.2	–	–	30.1 ± 3.7
	GraNDAG	$\mathbf{22.2 \pm 2.6}$	$\mathbf{42.0 \pm 6.2}$	–	–	185 ± 26
d=20	DAS（我们的方法）	56.4 ± 2.5	213 ± 28	0.99 ± 0.00	0.27 ± 0.04	$\mathbf{16.1 \pm 0.3}$
	SCORE	57.17 ± 3.1	229 ± 23	0.99 ± 0.01	0.30 ± 0.04	40.7 ± 1.8
	CAM	54.2 ± 5.4	202 ± 29	–	–	313 ± 80
	GraNDAG	$\mathbf{49.3 \pm 4.5}$	$\mathbf{211 \pm 37}$	–	–	357 ± 47
d=50	DAS（我们的方法）	156 ± 4	1460 ± 67	0.96 ± 0.02	0.24 ± 0.03	$\mathbf{48.3 \pm 1.1}$
	SCORE	144 ± 6	1346 ± 57	0.97 ± 0.01	0.30 ± 0.03	245 ± 5
	CAM	$\mathbf{141 \pm 6}$	$\mathbf{1337 \pm 94}$	–	–	1143 ± 79
	GraNDAG	$\mathbf{141 \pm 10}$	1432 ± 110	–	–	1410 ± 73
d=100	DAS（我们的方法）	335 ± 5	6695 ± 224	0.91 ± 0.03	0.21 ± 0.04	$\mathbf{114 \pm 2}$
	SCORE	$\mathbf{313 \pm 11}$	$\mathbf{5965 \pm 273}$	0.91 ± 0.03	0.27 ± 0.06	779 ± 13
d=200	DAS（我们的方法）	691 ± 13	$\mathbf{25147 \pm 835}$	0.89 ± 0.06	0.21 ± 0.05	$\mathbf{279 \pm 4}$
	SCORE	$\mathbf{626 \pm 14}$	25707 ± 891	0.88 ± 0.04	0.30 ± 0.05	4142 ± 35
d=500	DAS（我们的方法）	1761 ± 15	–	0.80 ± 0.04	0.19 ± 0.03	$\mathbf{1308 \pm 7}$
	SCORE[3]	**1642**	–	0.82	0.27	25307
d=1000	DAS（我们的方法）	$\mathbf{3951 \pm 9}$	–	0.76 ± 0.05	0.08 ± 0.00	$\mathbf{5539 \pm 81}$

[2] 对于 $d = 500$和 SCORE 方法，由于无法在合理时间内重复实验，因此未显示标准差。表中的数值对应单次运行。

5. 结论

在数据服从非线性加性高斯噪声模型的假设下，我们展示了如何根据分数函数的雅可比矩阵从理论上恢复精确的因果图。我们的发现扩展了Rolland等人（2022）的工作，即利用分数从数据中学习变量的拓扑排序。此外，我们还展示了可以根据分数发现并确定所有边的方向。基于我们的分析，我们设计了一种算法，其效率提高了 $\mathcal{O}(d)$，与Rolland等人（2022）相比，在保持准确率相当的同时，实际运行速度得到了显著提升。

虽然我们已经显著提升了速度，但还可以通过在子图甚至多个图之间分摊分数计算（L¨owe 等人（2020年））或使用更具可扩展性的估计量来进一步改进。然而，未来工作的主要障碍（也是我们工作的最大局限）是缺乏公开、经过精心整理且规模庞大的因果发现基准。事实上，我们的可扩展性实验仅限于在合成数据上进行推断。没有真实数据，就很难在受控的合成环境之外评估算法的性能。由于庞大且完整标注的因果图可能永远无法获得，我们确定了未来工作的三个重要方向。首先，将我们的方法扩展到模型假设之外的输入变量，尤其是

表2：ER1数据上的实验。对于CAM和GraNDAG，我们报告了Rolland等人（2022）中给出的结果。当 d 的值较高时，部分方法因运行耗时过长而未能列出。

	方法	SHD	SID	精确率	Rec.	时间 [秒]
d=10	DAS（我们的方法）	1.0 ± 0.8	$\mathbf{3.3 \pm 3.7}$	0.97 ± 0.01	0.84 ± 0.05	$\mathbf{7.8 \pm 0.1}$
	SCORE	$\mathbf{0.7 \pm 0.5}$	4.5 ± 4.3	0.98 ± 0.01	0.98 ± 0.01	8.0 ± 0.2
	CAM	1.7 ± 1.0	6.4 ± 4.2	–	–	30.1 ± 3.7
	GraNDAG	1.5 ± 1.4	6.5 ± 7.2	–	–	185 ± 26
d=20	DAS（我们的方法）	3.2 ± 1.4	18.1 ± 10.1	0.98 ± 0.02	0.85 ± 0.03	$\mathbf{16.7 \pm 0.4}$
	SCORE	$\mathbf{2.0 \pm 1.8}$	$\mathbf{8.3 \pm 9.9}$	0.99 ± 0.01	0.91 ± 0.03	36.4 ± 1.8
	CAM	3.5 ± 1.6	14.3 ± 9.8	–	–	313 ± 80
	GraNDAG	7.6 ± 4.2	31.6 ± 22.7	–	–	357 ± 47
d=50	DAS（我们的方法）	17.3 ± 3.4	99.6 ± 42.1	0.96 ± 0.04	0.77 ± 0.04	$\mathbf{49.2 \pm 2.1}$
	SCORE	9.8 ± 3.8	69.6 ± 41.3	0.98 ± 0.01	0.87 ± 0.03	251 ± 7
	CAM	$\mathbf{8.3 \pm 2.9}$	$\mathbf{53.7 \pm 31.9}$	–	–	1143 ± 79
	GraNDAG	20.2 ± 6.1	135 ± 456	–	–	1410 ± 73
d=100	DAS（我们的方法）	46.6 ± 9.1	327 ± 47	0.92 ± 0.06	0.68 ± 0.04	$\mathbf{111 \pm 6}$
	SCORE	$\mathbf{27.5 \pm 6.9}$	$\mathbf{288 \pm 115}$	0.97 ± 0.02	0.83 ± 0.05	776 ± 12
d=200	DAS（我们的方法）	110 ± 6.6	899 ± 191	0.88 ± 0.07	0.68 ± 0.06	$\mathbf{282 \pm 4}$
	SCORE	$\mathbf{59.9 \pm 8.5}$	$\mathbf{495 \pm 161}$	0.95 ± 0.03	0.85 ± 0.07	4237 ± 22
d=500	DAS（我们的方法）	291 ± 13	–	0.78 ± 0.07	0.65 ± 0.05	$\mathbf{1329 \pm 7}$
	SCORE[4]	**209**	–	0.8	0.85	25115
d=1000	DAS（我们的方法）	$\mathbf{994 \pm 15}$	–	0.59 ± 0.02	0.09 ± 0.00	$\mathbf{5544 \pm 73}$

[3] 对于 $d = 500$和 SCORE 方法，由于无法在合理的时间内重复实验，因此未显示标准差。表中的数值指的是单次运行结果。

尤其是考虑来自干预且存在潜在混杂因素的数据。第二，采用部分标注的因果图和外部收集的干预分布的评估方案。第三，合并许多较小且经过标注、变量相似但互不重叠的数据集（Mejia 等人（2022））。

致谢

我们感谢 Volkan Cevher 所作的宝贵讨论。本研究得到 AFOSR 资助，资助编号为 FA8655-20-1-7035。FM 得到国家研究与创新行动计划 *2014-2020*年的支持。

参考文献

Albert-LaszloBarabasi 和 Reka Albert。随机网络中尺度的涌现。《科学》，286(5439):509–512，1999。doi: 10.1126/science.286.5439.509。

URLhttp://www.sciencemag.org/cgi/content/abstract/286/5439/509。

Peter B¨uhlmann 和 Torsten Hothorn。提升算法：正则化、预测与模型拟合。《统计科学》，22(4)，2007年11月。URL https://doi.org/10.1214% 2F07-sts242。Peter B¨uhlmann 和 B. Yu。使用 l 2 损失进行提升。《美国统计学会杂志》- J AMER STATIST ASSN，98：324–339，2003年06月。Peter B¨uhlmann、Jonas Peters 和 Jan Ernest。CAM：因果加性模型、高维顺序搜索与惩罚回归。《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。《机器学习研究杂志》，3：507–554，2002年01月。Pierre Comon。独立成分分析，一个新概念? 《信号处理》，36(3)：287–314，1994。ISSN 0165-1684。URL https://www.sciencedirect.com/science/ article/pii/0165168494900299。高阶统计。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。URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524。Patrik O Hoyer、Dominik Janzing、Joris M Mooij、Jonas Peters 和 Bernhard Sch¨olkopf。利用加性噪声模型进行非线性因果发现。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。URL https://openreview.net/forum?id=rklbKA4YDS。Steffen L. Lauritzen。《图模型》。牛津大学出版社，1996年。ISBN 0-19-852219-3。Sindy L¨owe、David Madras、Richard Zemel 和 Max Welling。摊销因果发现：学习从时间序列数据中推断因果图。arXiv 预印本 arXiv:2006.10833，2020年。Nicolai Meinshausen 和 Peter B¨uhlmann。高维图与基于套索的变量选择。《统计年鉴》，34(3)，2006年6月。URL https://doi.org/10.1214%2F009053606000000281。Sergio H Garrido Mejia、Elke Kirschbaum 和 Dominik Janzing。通过使用 maxent 合并数据集获取因果信息。载于《国际人工智能与统计会议》，第581–603页。PMLR，2022。Thomas P Minka。逻辑回归数值优化器比较。未发表稿，页1–18，2003年。

Rebecca Morrison、RicardoBaptista 和YoussefMarzouk。超越正态性：在非高斯设定中学习稀疏概率图模型。载于 I.Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett编，神经信息处理系统进展，第 30卷。Curran Asso- ciates, Inc.，2017年。URL
h
t
t
p
s://
pr
oceedings.neurips.cc/paper/2017/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf。JudeaPearl。因果性：模型、推理与推断。剑桥大学出版社，美国，第 2 版，2009年。ISBN 052189560X。J.Peters和 P.B¨uhlmann。具有相等误差方差的高斯结构方程模型的可识别性。Biometrika，101(1):219–228，2013年11月。URL https://doi.org/10.1093% 2Fbiomet% 2Fast043。JonasPeters和 Peter B¨uhlmann。用于评估因果图的结构干预距离。NeuralComputation，27(3):771–799，2015年03 月。ISSN 0899-7667。doi: 10.1162/NECO a 00708。URLhttps://doi.org/10.1162/NECO_a_00708。Jonas Peters、Dominik Janzing和 BernhardSch¨olkopf。因果推断基础：基础与学习算法。自适应计算与机器学习。麻省理工学院出版社，马萨诸塞州剑桥，2017年。ISBN 978-0-262-03731-0。URL https://mitpress.mit.edu/books/elements-causal-inference。Paul Rolland、Volkan Cevher、Matth¨aus Kleindessner、Chris Russel、Bernhard Sch¨olkopf、DominikJanzing和FrancescoLocatello。分数匹配支持非线性加性噪声模型的因果发现。载于（即将出版）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年。URLhttps://www.science.org/doi/abs/10.1126/science.1105809。Bernhard Sch¨olkopf、Francesco Locatello、StefanBauer、Nan Rosemary Ke、Nal Kalchbrenner、Anirudh Goyal和 Yoshua Bengio。迈向因果表征学习。IEEE会刊，109(5):612–634，2021年。RajenD.Shah和 Jonas Peters。条件独立性检验的难点与广义协方差测度。《统计年鉴》，48(3)，2020年 6 月。Shohei Shimizu、Patrik O. Hoyer、Aapo Hyvarinen 和Antti Kerminen。用于因果发现的线性非高斯无环模型。机器学习研究杂志，7(72):2003–2030，2006年。URL http://jmlr.org/papers/v7/shimizu06a.html。Alessio Spantini、Daniele Bigoni 和 Youssef Marzouk。通过低维耦合进行推断。J. Mach.Learn. Res.，19(1):2639–2709，2018年1 月。ISSN 1532-4435。P. Spirtes、C. Glymour 和 R.Scheines。因果、预测与搜索。MIT press，第 2 版，2000年。

PeterSpirtes 和 ClarkGlymour。用于快速恢复稀疏因果图的算法。《社会科学计算机评论》，9(1):62–72，1991。URL https://doi.org/10.1177/089443939100900106。

Charles M. Stein。《依赖随机变量之和的分布正态近似误差界》。1972年。

Tim Van den Bulcke、Koenraad Leemput、Bart Naudts、Piet Remortel、Hongwu Ma、Alain Verschore、Bart De Moor 和 Kathleen Marchal。SynTReN：用于设计和分析结构学习算法的合成基因表达数据生成器。*BMC* 生物信息学，7:43，2006年02月。doi：10.1186/1471-2105-7-43。

K. ZHANG 和 A. Hyv¨arinen。基于加性噪声的无环因果关系发现：信息论视角。发表于欧洲机器学习与数据库知识发现原理及实践会议（*ECML PKDD*）*2009*年，斯洛文尼亚布莱德，2009年。

Xun Zheng、Bryon Aragam、PradeepKRavikumar 和 Eric P Xing。有向无环图无泪：用于结构学习的连续优化。在 S.Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编辑的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。

附录 A. 引理证明1

在本节中，我们为引理 1 提供证明，以完整起见。
证明 对于叶节点 l，公式（4）的 SCORE 变为 $s_l(\mathbf{X}) = -\frac{X_l - f_l(\mathrm{pa}_l(\mathbf{X}))}{\sigma_l^2}$。我们计算偏导数

$$\frac{\partial s_l(\mathbf{X})}{\partial X_j} = \frac{1}{\sigma_l^2}\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X})) \tag{12}$$

并注意到：

(i) $\mathbf{E}\left[\left|\frac{\partial s_l(\mathbf{X})}{\partial X_j}\right|\right] \neq 0 \Rightarrow X_j \in \mathrm{pa}_l(\mathbf{X})$。通过反证法，考虑 $X_j \notin \mathrm{pa}_l(\mathbf{X})$：由于 $f_l(\mathrm{pa}_l(\mathbf{X}))$ 关于 X_j 为常数，因此根据导数的定义，对每个 $\mathbf{X} \in \mathbb{R}^d$，都有 $\frac{\partial f_l(\mathrm{pa}_l(\mathbf{X}))}{\partial X_j} = 0$。于是，$\mathbf{E}\left[\left|\frac{\partial s_l(\mathbf{X})}{\partial X_j}\right|\right] = 0$，这与假设矛盾。

(ii) $X_j \in \mathrm{pa}_l(\mathbf{X}) \Rightarrow \mathbf{E}\left[\left|\frac{\partial s_l(\mathbf{X})}{\partial X_j}\right|\right] \neq 0$：我们从公式（8）中观察到，$\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X})) \neq 0$ 几乎必然，因此 $\left|\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))\right| > 0$ 几乎必然。由于 $\left|\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))\right|$ 取零的概率等于零，因此期望 $\mathbf{E}\left[\left|\frac{\partial s_l(\mathbf{X})}{\partial X_j}\right|\right]$ 等价于积分 $\int_{\mathcal{X}^+}\left|\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))\right| dP(\mathbf{X})$，其中 $\mathcal{X}^+ \subseteq \mathbb{R}^d$ 是 $\left|\frac{\partial f_l}{\partial X_j}(\mathrm{pa}_l(\mathbf{X}))\right|$ 严格为正的值的子集。由于严格正函数的积分本身也是严格为正的，因此 $\mathbf{E}\left[\left|\frac{\partial s_l}{\partial X_j}(\mathbf{X})\right|\right] > 0$。

■

附录 B. DAS 对假设检验阈值的稳定性

在第 4 节的实验中，我们将 α 非零均值假设检验的阈值设为0.01。标准 α 值为0.1 或更低。在缺乏关于数据的具体信息时，大多数基于条件独立性检验的因果发现方法都采用默认阈值0.05 （例如，参见知名库 `DoDiscover` 和 `causal-learn` 中 FCI 和 PC 的实现）：作为比较，我们在表 3 中给出了 DAS 在 α阈值设为0.05时的实验结果。我们进行了10 次使用不同随机种子的实验，并报告经验均值和标准差。实验表明，$\alpha = 0.05$ 不会显著影响 DAS 的性能（与平均 SHD 的距离在 $\alpha = 0.01$ 误差条范围内）。

附录 C. SF实验

本节展示了无标度图上的实验结果，包括较稀疏的图（表 4）和较稠密的图（表 5）。

表 3：对 ER1 和 ER4 数据进行 DAS 实验，采用不同的 α 截断阈值进行假设检验。

节点	阈值 SHD（ER1）		SHD (ER4)
$d = 10$	$\alpha = 0.01$	1.0 ± 0.8	27.0 ± 2.2
	$\alpha = 0.05$	0.9 ± 0.9	26.7 ± 2.0
$d = 20$	$\alpha = 0.01$	3.2 ± 1.4	56.4 ± 2.5
	$\alpha = 0.05$	3.1 ± 1.6	58.2 ± 2.5
$d = 50$	$\alpha = 0.01$	17.3 ± 3.4	156 ± 4
	$\alpha = 0.05$	17.1 ± 3.7	155 ± 5
$d = 100$	$\alpha = 0.01$	46.6 ± 9.1	355 ± 5
	$\alpha = 0.05$	47.6 ± 5.3	339 ± 6
$d = 200$	$\alpha = 0.01$	110 ± 6	691 ± 13
	$\alpha = 0.05$	114 ± 8	698 ± 9

表 4：SF1 数据上的实验。对于 SCORE、CAM 和 GraNDAG，我们报告了Rolland等人（2022）中的结果。

	方法	SHD	SID	精确率	Rec.	时间 [s]
d=10	DAS（我们的方法）	0.4 ± 0.7	2.2 ± 3.4	0.99 ± 0.04	0.84 ± 0.15	7.6 ± 0.1
	SCORE	0.3 ± 0.6	2.7 ± 5.8	–	–	–
	CAM	0.4 ± 0.5	2.8 ± 3.6	–	–	–
	GraNDAG	1.4 ± 1.0	12.5 ± 9.7	–	–	–
d=20	DAS（我们的方法）	1.9 ± 1.6	19.1 ± 7.4	0.99 ± 0.02	0.84 ± 0.11	16.6 ± 0.4
	SCORE	0.9 ± 0.9	13.8 ± 12.6	–	–	–
	CAM	0.9 ± 0.9	12.9 ± 14.0	–	–	–
	GraNDAG	3.2 ± 1.9	25.5 ± 15.6	–	–	–
d=50[5]	DAS（我们的方法）	10.9 ± 4.9	225.3 ± 72.1	0.96 ± 0.03	0.74 ± 0.08	53.1 ± 0.9
	SCORE	4.6 ± 2.4	132.6 ± 75.8	–	–	–
	CAM	3.6 ± 1.9	115.4 ± 72.6	–	–	–
	GraNDAG	9.2 ± 3.3	281.8 ± 129.8	–	–	–
d=100	DAS（我们的方法）	39.4 ± 8.0	214 ± 42	0.94 ± 0.02	0.51 ± 0.07	102 ± 5
d=200	DAS（我们的方法）	112 ± 11	626 ± 83	0.97 ± 0.02	0.33 ± 0.03	279 ± 6
d=500	DAS（我们的方法）	271 ± 15	–	0.93 ± 0.09	0.68 ± 0.01	1315 ± 9
d=1000	DAS（我们的方法）	910 ± 12	–	0.59 ± 0.02	0.09 ± 0.00	5442 ± 61

[4] 对于 $d > 50$，实验仅针对 DAS 执行。

附录 D. 真实与半合成数据

除了在合成图上进行实验，我们还测试了 DAS 在 Sachs 真实数据（Sachs 等人(2005年）上的实证表现。这是一项常用于生物数据因果发现的基准数据，包含 17 条边和 853 个观测值；此外，我们还测试了从基因 SynTReN 生成器中采样得到的半合成数据。

表5：SF4数据上的实验。对于 SCORE、CAM 和 GraNDAG，我们报告的结果见Rolland等人（2022）。

	方法	SHD	SID	精确率	Rec.	时间 [s]
d=10	DAS（我们的方法）	10.1 ± 2.39	35.7 ± 9.1	0.99 ± 0.01	0.75 ± 0.06	7.8 ± 0.1
	SCORE	4.6 ± 1.7	21.5 ± 9.6	–	–	–
	CAM	9.6 ± 2.0	40.4 ± 11.4	–	–	–
	GraNDAG	4.7 ± 1.8	23.0 ± 7.3	–	–	–
d=20	DAS（我们的方法）	30.4 ± 5.78	248.1 ± 20.0	0.98 ± 0.02	0.59 ± 0.06	17.1 ± 0.7
	SCORE	17.5 ± 3.5	179.2 ± 23.8	–	–	–
	CAM	26.4 ± 3.9	253.7 ± 28.8	–	–	–
	GraNDAG	14.7 ± 4.0	168.0 ± 39.2	–	–	–
d=50[6]	DAS（我们的方法）	115.5 ± 10.8	703.1 ± 87.5	0.97 ± 0.01	0.45 ± 0.06	51.3 ± 1.2
	SCORE	68.3 ± 3.6	1724 ± 109	–	–	–
	CAM	85.3 ± 4.2	1935 ± 99	–	–	–
	GraNDAG	63.8 ± 9.7	1677 ± 118	–	–	–
d=100	DAS（我们的方法）	295.3 ± 10.8	3212 ± 145	0.97 ± 0.01	0.25 ± 0.02	108 ± 2
d=200	DAS（我们的方法）	674.2 ± 16.6	21314 ± 891	0.95 ± 0.02	0.11 ± 0.01	283 ± 3
d=500	DAS（我们的方法）	1890 ± 7.1	–	0.97 ± 0.03	0.02 ± 0.01	1212 ± 31
d=1000	DAS（我们的方法）	3715 ± 21.3	–	0.92 ± 0.04	0.03 ± 0.01	5016 ± 53

[5] 仅对 DAS 进行 $d > 50$ 实验。

表达式记录（Van den Bulcke 等人（2006年））。在表 6 中，我们看到 DAS 在 Sachs 上的表现与 SCORE 相当；而在 SynTReN 实验中，其表现与 SCORE、CAM 和 GraNDAG 相当（表 7）。

表 6：Sachs 数据集上的实验。

方法	SHD	SID
DAS（我们的方法）	12	45
SCORE	12	45
CAM	12	55
GraNDAG	13	47

表 7：SynTReN 数据上的实验（20 个节点，500 个样本）。经验均值和误差是基于不同随机种子的 10 次运行计算得出的。

方法	SHD	SID
DAS（我们的方法）	38.0 ± 4.1	188.5 ± 71.2
SCORE	36.8 ± 4.7	193.4 ± 60.2
CAM	40.5 ± 6.8	152.3 ± 48.0
GraNDAG	34.0 ± 8.5	161.7 ± 53.4