arXiv:2304.03113v1 [stat.ML] 6 Apr 2023

通过因果结构学习实现高效 SAGE 估计

Christoph Luther*
维也纳大学 UniVie 博士学院 计算机科学

贡纳尔·克尼希¨*
慕尼黑大学 维也纳大学 慕尼黑机器学习中心 (MCML)

Moritz Grosse-Wentrup
维也纳大学 数据科学 @ 维也纳大学 维也纳认知科学中心

摘要

Shapley 加性全局重要性（SAGE）值是一种在理论上颇具吸引力的可解释性方法，能够公平地将全局重要性归因于模型的特征。然而，其精确计算需要计算特征集合数量呈指数增长的剩余性能贡献。这在计算上代价高昂，尤其是因为估计剩余贡献需要从条件分布中进行采样。因此，SAGE 近似算法只能将一部分特征集合纳入考虑。我们提出了 d-SAGE，这是一种能够加速SAGE 近似的方法。d-SAGE 的提出源于这一观察：特征与模型目标变量之间的条件独立关系（CIs）意味着剩余贡献为零，因此可以跳过相关计算。为了识别条件独立关系，我们利用因果结构学习（CSL）推断一个图，该图将数据中的（条件）独立关系编码为 d-分离。由于一次性图推断以及d-分离查询的开销，与剩余贡献评估的开销相比微不足道，因此这种方法在计算上更加高效。通过实验证明，d-SAGE 能够高效且准确地估计 SAGE 值。

1 引言

机器学习（ML）正日益应用于各个领域，从科学领域（Reichstein 等，2019；Schmidt 等，2019；Luan 和 Tsai，2021；Farrell 等，2018)到针对个人的高风险决策（Ragha- van 等人，2020；Zeng 等人，2017；Obermeyer 和 Mullainathan，2019）。尽管在预测性能方面取得了令人瞩目的成功（Senior 等人，2020；Bhatt 等人，2020），但ML 模型的复杂性使得评估其可信度或获取有关数据生成过程的知识变得困难。近年来，可解释机器学习的出现带来了大量能够提供模型和数据洞见的方法（Molnar，2020）。其中，基于博弈论中的 Shapley值（Shapley，1953）的可解释性方法因满足理想的公平性特征而广受欢迎（Strumbelj和 Kononenko，2014；Datta 等人，2016；Lundberg和 Lee，2017；Sundararajan 和 Najmi，2020；Covert 等人，2020）。

SAGE 值（Covert 等人，2020）将 Shapley 值应用于公平地将模型的预测性能归因于各个特征，从而深入揭示数据中的依赖关系。由于 SAGE 值与数据生成过程的性质相关联（Covert 等人，2020；Freiesleben 等人，2022），因此它们尤其适用于科学推断。SAGE 值的构建基础是所谓的 SAGE 值函数 $\nu(\mathbf{X}_S)$，用于衡量任意特征子集$\mathbf{X}_S$对性能的贡献。基于这些值函数，通过计算某个特征 X_j 在其余特征的所有可能子集$\mathbf{X}_S$ 中的平均额外贡献$\nu(\mathbf{X}_{S\cup j}) - \nu(\mathbf{X}_S)$，即可得到该特征的重要性值 ϕ。这一过程在计算上开销很大，原因在于联盟$\mathbf{X}_S$ 的数量会随特征数量呈指数级增长（Covert 等人，2020；Van den Broeck 等人，2022），且评估 ν 的成本很高，这是因为估算它需要进行条件抽样。在实践中，（Covert 等人，2020）通过仅针对随机抽样子集中的联盟计算相应的额外贡献，来应对联盟数量呈指数级增长的问题。[1]

*共同贡献。第26届国际人工智能与统计学会议（AISTATS）2023，西班牙瓦伦西亚。PMLR：第206卷。版权所有 2023，作者所有。

[1]此外，Covert 等人（2020）通过采用边际采样来避免在评估 ν 时进行条件采样。如果特征之间存在依赖关系，这会导致外推，并且无法将 SAGE 值与数据生成过程的性质联系起来（Chen 等人，2020）。在本文中，我们重点关注条件 SAGE 值的估计。

在这项工作中，我们建议利用数据中的依赖结构，加速通过（基于条件采样的）SAGE 值估计，并将这一方法命名为 d-SAGE。更具体地说，我们证明，如果所关注的变量在给定其余特征的相应子集以及模型的目标变量的条件下相互独立（即，如果 $X_j \perp Y | \mathbf{X}_S$，定理 1），则对于最优预测器，额外贡献 $\nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S)$为零。因此，如果我们知道数据中的条件独立关系（条件独立关系），就可以跳过相应的价值函数评估。由于依赖结构通常是未知的，且条件独立性检验代价高昂，我们借助因果结构学习（CSL）领域的研究，以贪心方式学习能够编码数据中依赖结构的图模型。

总体而言，该方法基于以下理由：SAGE近似的质量取决于对 ν 的评估次数，而每次评估都需要估计条件期望，因此计算成本高昂。[2] d- SAGE依赖于对因果图的一次性估计，在实践中，这可以通过多项式时间内运行的贪心搜索算法完成（Scutari etal.,2019b）。随后，估计得到的图可以通过线性时间的 d-分离查询来识别条件独立关系（Hagberget al.,2008;Darwiche, 2009）。反过来，每个发现的 d-分离都意味着可以省去一次对 $\nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S)$的昂贵评估。由于图学习只需执行一次，且 d-分离查询的效率很高，因此，只要有充分理由，就可以跳过 $\nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S)$的计算，从而显著缩短SAGE估计的运行时间。我们通过实证表明，节省的运行时间大致等于条件独立关系所占的比例。

1.1 贡献

我们提出 d-SAGE，这是首个利用数据中的依赖结构来提高 SAGE 估计效率的方法。具体而言，我们发现，数据中的条件独立关系意味着可以跳过相应的（计算开销较大的）剩余量评估，并建议利用贪心 CSL 来识别这些关系（第 4 节）。为选择合适的 CSL 算法，我们进行了基准测试；与以往工作不同，该测试评估了各算法在数据中高效识别条件独立关系的能力（第 5.1 节）。在十二个合成数据集上，我们通过实证表明，d-SAGE 与 Covert 等人（2020）提出的近似算法会收敛到相同的估计值，但 d-SAGE 的速度显著更快。我们发现，与剩余量评估的计算成本相比，学习因果结构所带来的计算开销可以忽略不计，因此总体运行时间的缩减幅度约等于数据中发现的条件独立关系所占的比例（第 5.2 节）。因此，d-SAGE 使得 SAGE 能够应用于更大型的模型，尤其是在稀疏场景下。

[2]计算开销取决于需要计算条件期望的数据类型。该领域先前的研究假设此操作具有多项式复杂度（Van den Broeck 等人，2022年）。

2 相关工作

尽管已有许多尝试旨在解决基于 Shapley 值的方法的复杂性，但大多数现有工作都致力于加速 SHAP（Lundberg 和 Lee，2017）估计（Jethani 等，2021；Covert 和 Lee，2021；Li 等，2020），或者仅适用于随机森林（B´enard 等，2022）。相比之下，我们的方法与模型无关，旨在改进 SAGE估计。此外，现有工作均未利用数据中的依赖结构来提高效率。因此，我们认为我们的工作与 Mitchell 等人（2022）提出的方法互为补充；后者建议谨慎选择排列。

近年来，因果关系领域的概念也被引入基于 Shapley 值的重要性度量中，以使其能够回答特定问题或改进模型解释。例如，Frye 等人（2020b）提出了非对称 *Shapley* 值，可以将所有变量的解释力沿因果链转移至根本原因（远端方法），也可以转移至直接原因（近端方法）。此外，Heskes 等人（2020）利用 Pearl 的 do-calculus 开发了因果 *Shapley* 值；Wang 等人（2021）则提出，应衡量因果图中边的重要性，而不是解释变量（即图中的节点）。与现有文献不同，我们希望借助因果推断研究提升特征归因的效率，同时保持 SAGE 值的原有原则不变。

不过，我们确实采用了 CSL。Scutari 等人（2019a）和 Constantinou 等人（2021）对大规模结构学习算法进行了基准测试。简而言之，两项研究均认为，基于贪心搜索算法的评分型结构学习优于基于约束的方法和混合方法。这些发现促使我们选择 CSL 算法进行 d- 分离推断。与现有研究不同，我们的基准测试不关注因果结构的恢复，而是关注检测数据中的条件独立关系。

3 背景

本节旨在帮助读者熟悉理解本文所需的基本概念。首先，我们介绍用于衡量全局特征重要性的 SAGE 值（第 3.1 节）。然后，我们回顾 SAGE 值为何难以估计，并介绍 SAGE 近似算法（第 3.2 节）。最后，我们解释 CSL，后文将使用它来加速 SAGE估计（第 3.3 节）。

3.1 Shapley加性全局重要性

Shapley 值最初由博弈论提出（Shapley，1953），现已广泛用于量化特征 ˇ 相关性（Strumbelj 和 Kononenko，2014；Datta 等人，2016；Lundberg 和 Lee，2017；Sundararajan 和 Najmi，2020；Covert 等人，2020）。在合作博弈研究中，它用于将博弈结果公平地归因于所有参与者。该原则可用于评估变量对于预测器f的重要性，其中预测性能是博弈结果，而变量则是参与者。Covert 等人（2020）利用 Shapley 值推导出一种衡量特征重要性的全局指标，即 SAGE 值。这里的“全局”是指评估样本中所有实例ˆ 的某一特征的重要性。对于使用输入 $x_1, ..., x_d$ 的任意模型 f，Covert 等人（2020）将第 j 个特征的 SAGE 值定义为：

$$\phi_j(\nu) = \frac{1}{d!} \sum_{\pi \in \Pi(d)} \big(\nu(\{X_i : \pi(i) \leq \pi(j)\}) - \nu(\{X_i : \pi(i) < \pi(j)\})\big) \tag{1}$$

其中， X_j 是与特征观测值 x_j对应的随机变量， $\Pi(d)$ 是索引 $\{1, ..., d\}$的所有排列的集合， π 是一个特定排列，而 $\pi(j)$是特征 j 在排列 π中的位置。为便于阅读，我们使用更一般的记号$\mathbf{X}_S$ 代替$\{X_i : \pi(i) < \pi(j)\}$ 作为值函数 ν 的输入，其中$\mathbf{X}_S$ 是任意特征集合， S 是其中所含特征的索引集合¯，即 $S \subseteq \{1, ..., d\}$ (S 是其补集)。 $\nu(\mathbf{X}_S)$定义为

$$\nu(\mathbf{X}_S) = \mathbb{E}_{\mathbf{X},Y}[\ell(\hat{f}_\emptyset(\mathbf{X}_\emptyset), Y)] - \mathbb{E}_{\mathbf{X},Y}[\ell(\hat{f}_S(\mathbf{X}_S), Y)],$$

ˆ 其中 $\ell(\cdot)$是任意适用的损失函数，且 $f_S(\mathbf{x}_S) =$ˆ $\mathbb{E}_{\mathbf{X}_{\bar{S}}|\mathbf{X}_S}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S]$。因此， $\nu(\mathbf{X}_S)$表示加入$\mathbf{X}_S$所带来的风险降低。由此，SAGE 值通过对比特征与任何已有联盟相比所带来的额外风险降低的平均值，来衡量该特征 j的重要性。

SAGE 值尤其具有吸引力，因为它们满足六项理想的公平性公理，使其区别于其他特征重要性度量：效率、虚设特征性质、对称性、单调性、线性性[3] 以及对单调变换的不变性。尽管 SAGE 值拥有充分的数学基础，并满足上述期望，但它们存在一个重大缺点：需要评估指数数量的余量，这在计算上是不可行的。在实践中，只会评估所有可能联盟中的一个子集（参见第 3.2 节）。

[3]为简洁起见，我们采用附录 D 中所述 SAGE 性质的 Shapley值性质名称。

要估计 SAGE 值，需要获取特征条件分布；更具体地说，我们需要从 $P(\mathbf{X}_{\bar{S}}|\mathbf{X}_S)$ 中采样，以估计边际化后的预测值 $f_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{X}_{\bar{S}}|\mathbf{X}_S}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S]$。遗憾的是，在实践中，条件采样器可能并不容易获得。Covert 等人（2020）建议通过改为从 $P(\mathbf{X}_{\bar{S}})$ 中采样来规避这一问题（边际采样）。尽管边际采样易于实现（且计算效率高），但它可能生成不真实的数据点（$\mathbf{x}_S$，$\mathbf{x}_{\bar{S}}$），因此，基于边际采样的 SAGE 值并不适用于推断数据生成过程，也不适用于理解模型在观测分布中的行为（Frye 等人，2020a；Chen 等人，2020；Aas 等人，2021；Molnar 等人，2022）。因此，我们聚焦于条件 SAGE，并在条件分布事先未知时对其进行估计。

估计条件分布时可以采用多种技术：对于分类变量，条件分布的估计可归结为使用交叉熵损失的标准监督学习。对于线性高斯数据，可以根据协方差矩阵解析地估计条件分布（Page Jr，1984）。对于具有非线性关系的连续情形，已有多种方法（Bishop，1994；Bashtannyk 和 Hyndman，2001；Sohn 等人，2015；Trippe 和 Turner，2018；Winkler 等人，2019；Hothorn 和 Zeileis，2021）。对于混合数据，可以采用序贯设计（Blesch 等人，2022年）。

3.2 SAGE 的难处理性与近似算法

对于基于 Shapley 值的可解释性方法 SHAP，其难处理性已得到证明（Van den Broeck 等人，2022年）。要进行精确计算，必须评估剩余特征所有可能子集的超额贡献。可能子集的数量会随着特征数量的增加而呈指数级增长。

精确的 SAGE估计同样受到联盟数量呈指数增长的影响。为解决这一问题，Covert 等人(2020) 提出了一种近似算法，该算法不考虑所有可能的联盟。更具体地说，作者提议从特征索引中反复采样排列 π。然后，对于当前排列中的每个元素，从第一个元素开始，他们依次计算 $\Delta_{j|S} := \nu(\mathbf{X}^{S \cup j}) - \nu(\mathbf{X}_S)$，其中集合$S$ 包含位于感兴趣特征 j 在 π中的所有特征。通过对不同排列 π中的所有 $\Delta_{j|S}$ 值取平均，我们得到 $\phi_j(\nu)$的估计重要性，其中X_j。该近似算法是无偏的，并且估计量的方差以 $O(\frac{1}{n})$的速度降低(Covert等人, 2020)。然而，考虑到估计 ν所需的风险评估，基于条件采样的过程在计算上仍然要求很高。

3.3 因果结构学习

本节介绍用于估计表示 d-分离的图的 CSL。 d-分离是条件独立性在底层分布中的图形等价形式。在两个标准假设下，这两个概念确实等价：(1) 满足马尔可夫性质；(2) 分布相对于图是忠实的。当变量 X_j 在图 $\mathcal{G}$ 中给定 $\mathbf{X}_S$ 时，与 Y d- 分离，我们记作 $X_j \perp_{\mathcal{G}} Y|\mathbf{X}_S$。由于我们仅使用图来读取 d-分离，因此不在此全面介绍，而请读者参阅 Darwiche（ 2009）和 Pearl（2009）。此处只需说明，我们所说的是有向无环图（DAG），其节点表示底层分布中的随机变量，边则反映数据中的直接依赖关系。边的方向还被解释为因果关系。现在我们简要总结如何从数据中推断这类图。

一般来说，方法可分为基于约束的方法和基于评分的方法。前者使用从数据中推断出的条件独立关系，作为边的连接位置约束。后者会探索给定变量上所有可能 DAG 的空间，并为访问过的每个图赋予评分。算法的输出是评分最高的图。由于一组变量或节点上 DAG 的空间大小会随着该集合基数的增加而超指数增长，基于评分的方法通常依赖贪婪搜索技术。此外，混合方法会结合条件独立关系约束和图评分来评估候选图。

在本研究中，我们聚焦于在近期基准测试中表现最佳的贪婪结构学习（Scutari et al., 2019a; Constantinou et al., 2021）。具体而言，我们采用基于爬山法（HC）和 TABU 搜索的结构推断方法（Russell and Norvig, 2009; Scutari et al., 2019b）。关键在于，这两种算法都使用贝叶斯信息准则（Schwarz, 1978），该准则满足两个重要性质：一致性 和局部一致性[4] （G´amez et al., 2011; Chickering, 2003）。G´amez et al. (2011) 表明，对于大小为 n 的数据集和独立同分布 数据，如果 $n \to \infty$且评分函数满足一致性 和局部一致性，那么 HC 的输出图就是底层分布的一个最小I-映射。根据最小 I-映射的定义，图中由 d-分离表示的条件独立关系集合是分布中条件独立关系的子集。因此，尽管数据的底层分布中可能存在未由 d-分离表示的独立关系，但不会出现与独立关系不对应的 d-分离。请注意，HC 会为输出图引入 DAG 结构，但对数据的假设仅仅是其为一个独立同分布 样本。不过，该证明依赖于忠实性假设。然而，对于线性模型，研究表明，如果模型参数是从正密度中随机抽取的，则忠实性被违反的概率为零（参见 Peters et al. (2017), Spirteset al. (2000)）。虽然 TABU 尚无类似的理论结果，但它是 HC 的扩展，在实践中表现出类似的行为（参见第 5.1 节）。

[4]贝叶斯狄利克雷等价均匀（BDeu）评分也满足这些性质，是一种有效的替代方案。

4 用于高效 SAGE 估计的因果结构学习

SAGE 估计在计算上颇具挑战。要进行精确计算，必须计算感兴趣特征 j 相对于其余特征的每一种可能联盟 $\mathbf{X}_S$ 的剩余贡献。剩余贡献的定义见第 3.2 节。

$$\Delta_{j|S} = \nu(\mathbf{X}_{S\cup j}) - \nu(\mathbf{X}_S) \tag{2}$$

可能的联盟数量会随着特征数量呈指数增长，使得在高维设置下进行精确计算变得不可行。因此，通过随机抽样联盟，直到估计值收敛，来估计SAGE值（第3.2节）。然而，由于评估 $\Delta_{j|S}$ 需要从条件分布中进行抽样，因此即使进行一次评估也是一项重大的计算挑战。因此，在实践中，近似质量受到能够计算的剩余贡献数量的限制。

我们提出 d-SAGE，这种方法能够识别并跳过不必要的冗余评估，从而提高近似质量。该方法基于以下观察：如果给定 X_S， X_j 与 Y 条件独立，那么 $\Delta_{j|S}$ 的取值为零：

定理 1. 对于 ℓ 为交叉熵损失或均方误差、 f^* 为相应的最优预测器且 ν_{ℓ,f^*}为对应的 *SAGE* 值函数的情况，有以下结论：

$$X_j \perp Y|\boldsymbol{X}_S \Rightarrow \nu_{\ell,f^*}(\boldsymbol{X}_{S\cup j}) - \nu_{\ell,f^*}(\boldsymbol{X}_S) = 0.$$

证明（略，完整证明见附录 *A*）：*Covert* 等人（*2020*）表明，对于交叉熵损失函数及其相应的最优模型——贝叶斯分类器，式 *2* 等于 X_j 和 Y 在给定 $\boldsymbol{X}_S$ 条件下的互信息，即 $I(X_j;\ Y|\boldsymbol{X}_S)$。对于以均方误差（*MSE*）为损失函数的最优回归模型，也存在类似结果。在这种情况下，额外贡献被证明等于 $\mathbb{E}_{\boldsymbol{X}_S}[Var(\mathbb{E}[Y|\boldsymbol{X}_S,\ X_j]|\boldsymbol{X}_S)]$（*Covert* 等人，*2020*）。对于这两个表达式，很容易看出，当 X_j 在给定 $\boldsymbol{X}_S$ 的条件下与 Y 条件独立时，即当 $X_j \perp Y \mid \boldsymbol{X}_S$ 时，它们的值均为零。

作为定理1的推论，了解数据中的依赖结构可以加快SAGE估计流程：如果 $X_j \perp Y \mid \mathbf{X}_S$，则可以跳过对 $\nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S)$的评估。

算法 1： 基于采样的近似 d-SAGE

数据 $\{\mathbf{x}_i, y_i\}_{i=}^{n}$ $_1$ 使用$\mathbf{x}_i \in \mathbb{R}^d$、模型 $\hat{}f$、损失函数 ℓ，排列数量 n_π

从**DAG** $\mathcal{G}$ 数据中推断 $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ 结合结构所选的学习算法。

对于 i 属于 $\{1, ..., n_\pi\}$ **执行**

  采样一个排列 π

  $S = \emptyset$

  对于 j in $\{1, ..., d\}$ **执行**

    如果 $X_{\pi_j} \not\perp_{\mathcal{G}} Y|\boldsymbol{X}_S$ **然后**

      从 $p(\mathbf{x}_{\bar{S}}|\mathbf{x}_S)$ 中采样 $\mathbf{x}_{\bar{S}}$

      采样 $\mathbf{x}_{\overline{S\cup\pi_j}}$ 从 $p(\mathbf{x}_{\overline{S\cup\pi_j}}|\mathbf{x}_S)$，其中 π_j 是 π的第 j 个元素

      $\hat{\Delta}_{j|S} = \ell(\hat{f}(\mathbf{x}_S, \mathbf{x}_{\bar{S}})) - \ell(\hat{f}(\mathbf{x}_{S\cup\pi_j}, \mathbf{x}_{\overline{S\cup\pi_j}}))$

    else

      $\hat{\Delta}_{j|S} = 0$

    end

    $S = S \cup \pi_j$

  end

end

return $\hat{\phi}_j = \frac{1}{n_\pi}\sum_{i=1}^{n_\pi} \hat{\Delta}_{j|S}$ 对于 $j = 1, ..., d$

请注意，为便于阅读，我们省略了 $\hat{}\Delta_{j|S}$ 的索引。

为识别数据中的条件独立关系，我们建议利用最初为学习数据中的因果结构而开发的贪心过程。因果结构学习算法能够在多项式时间内估计因果图（Scutari等，2019b）。鉴于马尔可夫性质和忠实性均得到满足，可以使用线性时间的 d-分离查询从图中读取数据中的（条件）独立关系（Hagberg等，2008；Darwiche，2009）。我们的理由是，学习因果图的一次性工作，以及额外的线性时间 d-分离查询，与计算剩余贡献的计算开销相比微不足道。[5]

总而言之，d-SAGE估计对原始 SAGE 近似算法做出了两项关键改动。首先，在所有随机变量、特征和目标变量上拟合一个图 $\mathcal{G}$。其次，如果排列 π 中当前特征 X_j 在给定集合 $\mathbf{X}_S = \{X_i$ 的条件下与目标变量 d-分离，则跳过对 $\Delta_{j|S}$ 的估计：$\pi(i) < \pi(j)\}$。算法 1 中以蓝色突出显示了这些改动。

[5]一般来说，条件采样的复杂性取决于对数据生成过程的假设。在他们针对SHAP的可处理性分析中，Van den Broeck 等人（2022年）假设，以多项式复杂度计算形如 $\mathbb{E}_{\mathbf{X}_{\bar{S}}|\mathbf{X}_S}[\hat{}f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S]$ 的条件期望。

5 实验

本节分为三个部分。前两部分中，我们在具有已知真实值的合成数据上评估所提出的方法：由于我们在估计图中使用 d-分离查询来进行 d-SAGE 近似，因此首先评估学习到的结构中的 d-分离相对于数据中真实条件独立关系的准确性（第 5.1 节）。随后，我们将 d-SAGE 与常规 SAGE 值近似进行比较（第 5.2 节）。第三部分展示该方法在一项现实应用中的实用性（第 5.3 节）。[6]

5.1 因果结构学习基准测试

现有的结构学习基准测试通过评估算法恢复真实因果结构的能力来进行比较（Constantinou 等，2021；Scutari 等，2019a）。然而，对于 d-SAGE，我们只关注学习依赖结构。因此，我们评估估计图中以 d-分离表示数据中的条件独立关系的效果。

5.1.1 设置

我们评估贪心搜索算法爬山算法（HC）和 TABU（Scutari 等人，2019b；Russell 和 Norvig，2009）。我们根据这些方法在近期 CSL 基准测试中的优异表现选用了它们（Constantinou 等人，2021；Scutari 等人，2019a）。作为性能指标，我们采用 F1 分数来评估针对随机抽样目标变量 Y 的 d-分离检测能力，并计算相应的错误发现率。具体而言，对于每个形式为 $X_j \perp_{\mathcal{G}} Y|\mathbf{X}_S$ 的潜在 d-分离，我们检查其在真实图和估计图中的状态是否一致。为应对高维图（DAG$_{sm}$、DAG$_m$ 和 DAG$_l$）中呈指数增长的大量 d-分离，我们随机抽取一个感兴趣的节点 X_j 和一个条件集 $\mathbf{X}_S$，重复一百万次，而不是遍历所有潜在的 d-分离语句。对于这两种算法，我们都使用了其在 R 中的实现 *bnlearn*（Scutari，2010）。[7] 我们考虑了十二种具有已知真实情况的不同合成数据设置：

DAG$_s$、DAG$_{sm}$、DAG$_m$ 和 DAG$_l$ 我们采样了节点数各异（$s = 10$、$sm = 20$、$m = 50$和 $l = 100$）、密度各不相同的三类合成图（平均邻接度为 2、3 和 4）。基于这些图，我们从对应的线性高斯数据模型中采样数据，其中边的绝对值

[6]所有代码均已公开：https://github.com/gcskoenig/csl-experiments/tree/ camera-ready。[7] 所有图学习实验均在搭载 Intel Core i7- 8700K 的台式电脑上运行。

权重范围为 0.5 到 2。我们将方差标准化为（近似）1，以避免其随拓扑排序而增加，并抵消基准测试中可能存在的偏见（Reisach 等人，2021）。采样本身使用了在 R 中实现的 *pcalg* 软件包（Kalisch 等人，2012；R 核心团队，2022年）。

5.1.2 结果

首先，我们观察到，TABU 虽然耗时大约是两倍，但其性能要么与爬山算法相当，要么优于爬山算法（参见图 1、图2 和附录 C）。因此，本节仅讨论 TABU搜索的结果，我们也采用 TABU搜索进行 d-SAGE估计。图 1 展示了所有十二个图中，图学习的运行时间如何随样本量变化，以及用于 d-分离推断的相应 F1分数。关键结论是，对于最稀疏的图（平均邻接度为2），当 $n \geq 10{,}000$ 时，F1 分数大于 0.88。然而，对于规模更大的图，性能会略有下降，这是预料之中的。只有在最稠密的图设置（平均邻接度为 4）下，并且节点数为 50 和 100 时，才需要更大的样本量，即 $n \geq 100{,}000$，才能以合理的比率推断 d-分离。正如我们将在第5.2 节中看到的，在d-SAGE估计的背景下，图学习的运行时间可以忽略不计。

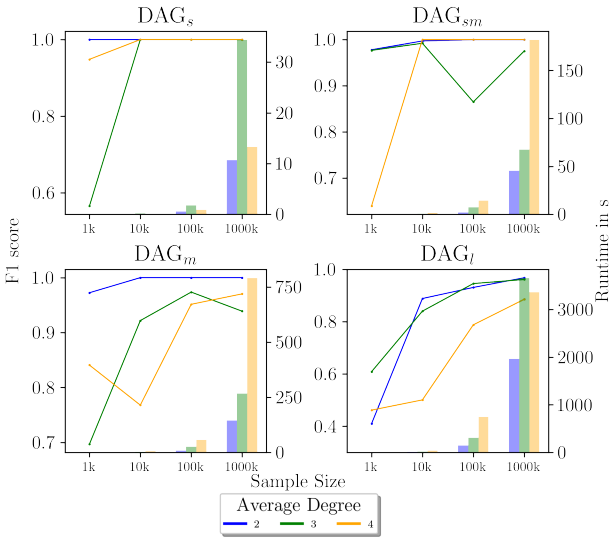


图 1：使用 TABU 搜索时，d-分离的 F1 分数（折线，左侧纵轴）和图学习的运行时间（柱形，右侧纵轴）随样本量的变化。

我们注意到，最低 F1 分数并没有明确的阈值，无法据此确定 SAGE估计何时会受益于因果结构学习，因为不同错误类型会产生不同后果。错误推断出的 d-分离可能会导致有偏估计，而未检测到的 d-分离则只会降低 CSL 在

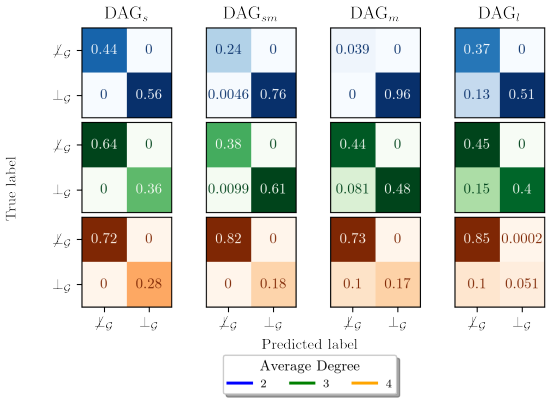


图 2：基于 TABU 搜索，针对全部十二个图、样本量为 $n = 10,\ 000$ 时，真实与预测的 d-连接（$\not\perp\!\!\!\perp_{\mathcal{G}}$）及 d-分离（$\perp\!\!\!\perp_{\mathcal{G}}$）的混淆矩阵。

缩短运行时间方面带来的益处。值得注意的是，我们在图 2 中的模拟结果几乎没有假发现（即底层分布中不存在条件独立关系，但却推断出一个 d-分离），但存在一些假阴性实例，导致跳过的 $\Delta_{j|S}$ 评估次数少于应有的次数。这一结果与第 3.3 节中的推理一致。因此，使用 CSL 是一种保守的条件独立推断方法。请注意，对于基准测试中使用的数据，真实图是已知的，且马尔可夫性质和忠实性均成立，因此 d-分离确实与统计独立性相对应。

5.2 评估 d-SAGE 的效率与准确性

在第 5.1 节的基准测试研究中，我们强调了结构学习能够高效且保守地估计 d-分离，并将其视为条件独立关系的等价形式。现在，我们将评估 d-SAGE 的效率和准确性。

5.2.1 设置

为了评估 d-SAGE 在实践中的表现，我们在十二个数据集上分别拟合线性模型（LM）和随机森林（RF）（使用 *scikit-learn* 实现，采用默认设置（Pedregosa 等，2011））。两者均使用均方误差（MSE）作为损失函数。因此，线性模型（LM）属于理论论证所要求的最优模型类别。随机森林模型则用于对高性能但非最优模型进行合理性检查（模型性能参见附录 D）。为进行公平比较，我们基于完全相同的特征排序比较 d-SAGE 和 SAGE。这也使我们能够将跳过的 $\Delta_{j|S}$ 评估（其值设为零）与估计值进行比较；后者应当非常接近零。总体而言，对于每种设置（图 + 模型），我们分别对 SAGE 和 d-SAGE 值进行了五次估计。

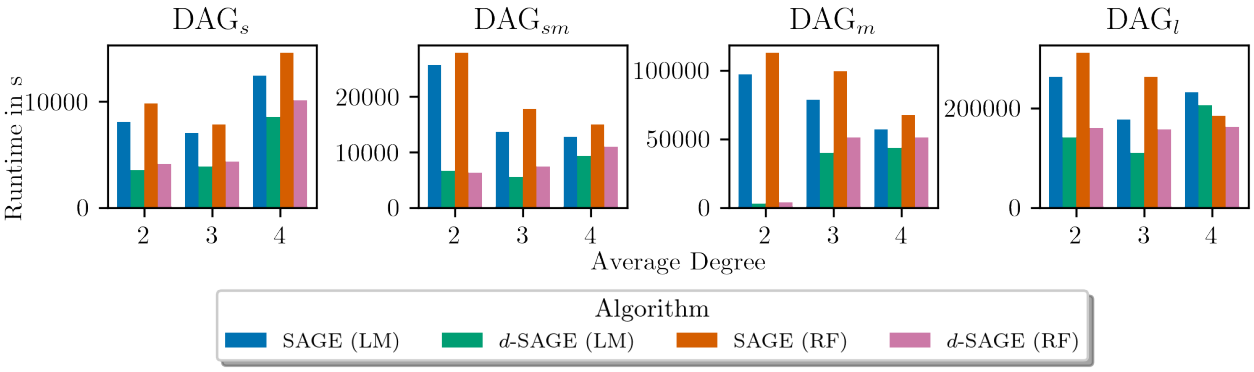
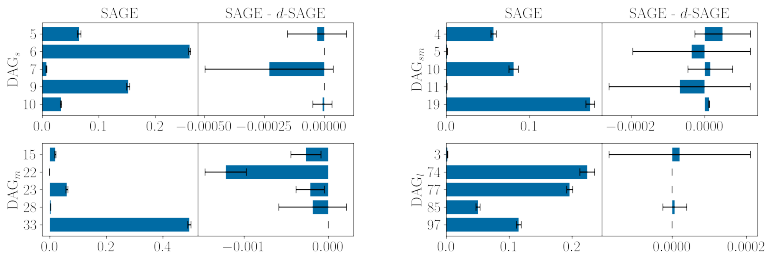
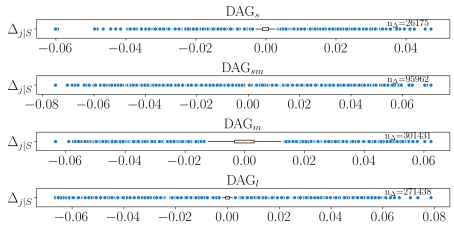


图 3：基于 $n = 10,000$，对全部十二个图以及线性模型（语言模型）和随机森林（随机森林）进行一次 SAGE 或 d-SAGE 运行的总估计运行时间。



(a) SAGE 值以及五个最大值中 SAGE 与 d-SAGE 之间的差异。



(b) 箱线图显示 $\Delta_{j|S}$ 的分布
被跳过的多余评估。

图4：基于平均度为2的有向无环图，对最优模型（语言模型）中 d-SAGE近似质量的结果。基于五个（d-）SAGE估计值。

评估所用的合成数据集与第 5.1 节中的相同。

5.2.2 结果

我们发现，d-SAGE 确实如预期那样加快了 SAGE 的近似计算。具体而言，对于两类模型，在所有图上，估计运行时间[8]的降幅大致等于条件独立关系在模型目标变量中的占比（参见附录 C）（参见图 3）。此外，d-SAGE 在实现加速的同时，并未扭曲估计结果。请注意，图 3 中未包含图学习的运行时间，因为其耗时介于 0.06 秒（平均度为 2 的 DAG$_s$）和 39.86 秒（平均度为 4 的 DAG$_l$ ）之间，因此在此情境下可以忽略不计。

[8]完整的 SAGE 估计在多台不同的机器上进行。为公平评估运行时间，我们采用了在同一 CPU（Intel Core i7-8700K 台式机 CPU）上完成的估计值：两种方法均使用相同的 100 次排列进行，这些排列对 SAGE 和 d-SAGE 相同；运行时间再乘以因子 $\frac{n_\pi}{100}$，其中 n_π 是一次 SAGE 运行收敛所需的排列次数。有关收敛行为，请参见附录 E。

线性模型 图4 (a)展示了平均度为二的四个图中绝对值最大的五个 SAGE 值，以及 SAGE 与 d-SAGE 估计值之间相应的差异。总体而言，这些差异比原始 SAGE 值小约三个数量级，即通常低于百分之一。即使是 DAG$_s$ 中变量 7 的差异最为显著，也仅约为 SAGE 值（约 0.007）的 2.7%。在其余用于识别重要特征的 SAGE 值中，我们未发现其他显著差异；这些特征的 SAGE 值绝对值最大。SAGE 值判定为不重要的特征，d-SAGE 也同样将其检测为不重要。值得注意的是，如果给定所有（采样得到的）联盟后，感兴趣的特征与目标变量条件独立，那么某些 d-SAGE 估计值会等于零。这里我们认为，观测数据的 SAGE 值会被我们向零偏移；对于真正独立的特征而言，这更接近（或等于）在采用最优预测器和无限数据时得到的“忠实于数据”的估计值。

图4 (b) 展示了每个 $\Delta_{j|S}$ 值，这些值源自被检测为条件独立的特征

因此，在 d-SAGE 近似中将其设为零。可以清楚地看到，大多数值都非常接近零，这一点也体现在箱体较窄上，凸显了我们方法的实用性。

随机森林 为了检验结果的敏感性，我们使用一种性能较高但并非最优的 RF 回归器（而非最优 LM）重复了完全相同的研究。虽然运行时间的节省与使用 LM 时相同，但 d-SAGE 值相对于原始估计值的偏差略为明显（参见附录 D）。结果表明，我们的方法对于接近最优的模型同样适用。

5.3 真实世界应用

为了展示 d-SAGE 在实践中的实用性，我们将该方法应用于 UCI ML 仓库中的药物消费数据（Dua 和 Graff，2017）。我们使用逻辑回归，根据一个样本量为 $n = 1885$ 的数据集中的十二个解释变量，预测目标变量“尼古丁消费量”。使用 TABU 搜索算法拟合图，耗时 0.035 秒。对五次不同运行进行 SAGE估计，耗时约 12 小时 14 分钟[9]。为得出 d-SAGE 值，我们没有重新运行基于 d-SAGE 的估计，而是直接将输出中与拟合图中 d-分离相对应的相应 $\Delta_{j|S}$（其中包含所有此类 $\Delta_{j|S}$）替换为零。我们发现，此类 $\Delta_{j|S}$ 值约占 38%，可以跳过这些值，从而实现几乎同等幅度的相对加速。

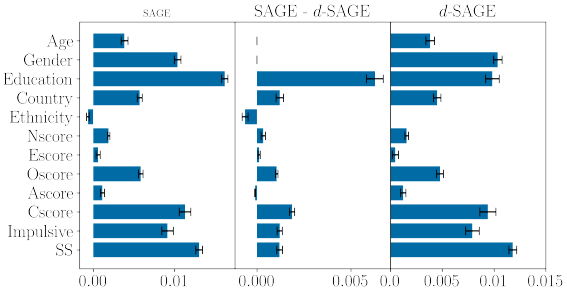


图 5：药物消费数据的 SAGE 值、SAGE 与 d-SAGE 以及 d-SAGE 值之间的差异。基于五个（d-） SAGE 估计值。

图 5 显示，d-SAGE 值大多与原始SAGE 估计值一致。在重要变量中，只有‘教育’的 d-SAGE 值明显不同，因为与SAGE 估计值相比，它减少了约三分之一。然而，它仍被赋予了相对较高的重要性。d-SAGE 在实际应用中的有效性将在后文进一步

[9]所有计算均在 Intel Core i7-8700K 台式机 CPU 上运行。

如图 6 所示，$\Delta_{j|S}$ 值集中在零附近，凸显了这一点。

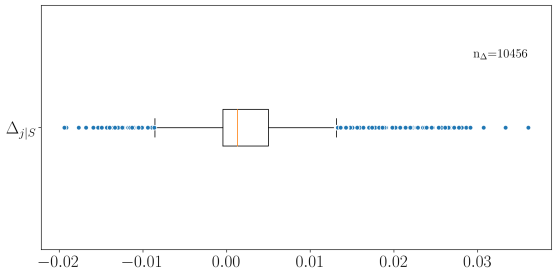


图6：箱线图显示了跳过的剩余评估中 $\Delta_{j|S}$ 的分布。基于五个（d-）SAGE 估计值。

6 讨论

模型最优性 条件 SAGE 值对于科学推断尤其具有吸引力，即用于了解数据（Chen 等人，2020；Covert等人，2020）。因此，通常需要准确的预测器（Molnar 等人，2022年）。然而，对于 d-SAGE而言，这一要求的重要性更高，因为如果模型最优性的假设不成立，跳过对非零剩余贡献的评估可能会进一步导致解释产生偏见（定理 1）。

CSL的假设CSL 依赖于因果充分性、马尔可夫性质和忠实性（Peters 等，2017）。这些假设确保所有相关变量都被观测到，并确保数据中的条件独立关系与真实因果图中的 d-分离相一致（我们假设该图为 DAG）。我们推测，违反这些假设对我们的方法并非至关重要，因为学习真实因果图并不是我们的目标。相反，我们只关注学习一个图，以编码观测分布中存在的（条件）独立关系（无论这些关系源自哪种因果机制）。由 HC学习得到的有向无环图是底层分布的最小I-映射，这使其适用于条件独立推断，但无法保证得到正确的图，也无法保证发现的条件独立关系数量正确。

然而，实践者在应用 d-SAGE 之前应仔细评估相关假设。例如，若存在潜在混杂因素或循环赋值，可以考虑其他概念，例如 m-分离和 σ-分离（参见 Bongers 等人，2021）。此外，建议进行合理性检查，以确认被跳过的剩余贡献是否确实被评估为零。

基于评分的 CSL 的使用由于效率较高，分析仅限于使用基于评分的 CSL。爬山算法

尤其具有吸引力，因为正如第3.3节所解释的那样，它能够推断出底层分布的最小I-映射，而TABU在经验上表现良好。然而，条件独立推断并不局限于这些技术。正如上文所述，可以选择某种算法并考虑所采用的假设来进行图学习。此外，我们方法背后的理由是用CSL替代条件独立检验。例如，偏相关检验的效率明显低于 d-分离查询（参见附录F），因此需要更多的条件独立关系才能实现SAGE的加速。

7 结论

我们提出了 d-SAGE，这是一种利用数据中的依赖结构来加速 SAGE 估计的方法。更具体地说，我们发现，数据中的条件独立性意味着相应的剩余贡献可以直接评估为零。我们修改了基于排序的 SAGE 近似算法：首先使用因果结构学习算法学习数据中的依赖结构，然后在图中编码了条件独立关系时跳过剩余贡献的评估。学习到的图中的错误可能会减慢收敛速度（如果未发现条件独立关系），也可能使结果偏向零（如果出现假发现）。然而，在我们的实验中，几乎没有假发现，因此，对于给定任意联盟后并非条件独立的特征，其估计值基本收敛到与原始 SAGE 近似算法相同的值。此外，因果结构学习算法能够发现大多数条件独立关系，因此我们观察到了显著的性能提升。因此，在计算预算固定的情况下，d-SAGE 的效率提升可以实现比 Covert 等人（2020）提出的近似算法更准确的 SAGE 值估计。未来，将 d-SAGE 与 Mitchell 等人（2022）的排列抽样相结合，并评估结果能否推广到其他基于 Shapley 的可解释性方法（例如 SHAP），将是一个有趣的研究方向。

致谢

该项目得到德国联邦教育与研究部（BMBF）的支持。所呈现的计算结- 果部分借助维也纳科学集群（VSC）完成。感谢 Bernd Bischl 就基准测试设计提供建议。

参考文献

Aas, K.、Jullum, M. 和 Løland, A.（2021）。解释特征存在依赖关系时的个体预测：更准确地近似 Shapley 值。《人工智能》，298：103502。

Bashtannyk, D. M. 和 Hyndman, R. J.（2001）。核条件密度估计的带宽选择。《计算统计与数据分析》，36(3):279–298。B´enard, C., Biau, G., Da Veiga, S. 和 Scornet, E.（2022年）。Shaff：通过随机森林快速且一致地估计 Shapley 效应。载于《国际人工智能与统计会议论文集》，第 5563–5582 页。PMLR。Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. 和 Eckers- ley, P.（2020）。部署中的可解释机器学习。载于《2020年公平、问责与透明度会议论文集》，第 648–657 页。

C.M.Bishop(1994). 混合密度网络。技术报告，阿斯顿大学。

Blesch, K.、Watson, D. S. 和 Wright, M. N.（2022年）。混合数据的条件特征重要性。arXiv 预印本 arXiv:2210.03047。

Bongers, S.、Forr´e, P.、Peters, J. 和 Mooij, J. M.（2021）。含有环和潜变量的结构因果模型基础。《统计年刊》，49(5)：2885 – 2915。Chen, H.、Janizek, J. D.、Lundberg, S. 和 Lee, S.-I.（2020）。忠于模型，还是忠于数据？arXiv 预印本 arXiv:2006.16234。Chickering, D. M.（2003）。通过贪心搜索进行最优结构识别。《机器学习研究杂志》，3(null)：507–554。Constantinou, A. C.、Liu, Y.、Chobtham, K.、Guo, Z. 和 Kitson, N. K.（2021）。含噪数据下贝叶斯网络结构学习算法的大规模实证验证。《近似推理国际期刊》，131：151–188。

Covert, I. 和 Lee, S.-I.（2021）。改进 KernelSHAP：使用线性回归进行实用的 Shapley 值估计。载于《国际人工智能与统计会议》，第 3457–3465 页。PMLR。

Covert, I.、Lundberg, S. M. 和 Lee, S.-I.（2020）。通过加性重要性度量理解全局特征贡献。《神经信息处理系统进展》，33：17212–17223。

Darwiche, A.（2009）。《贝叶斯网络建模与推理》。剑桥大学出版社。

Datta, A.、Sen, S. 和 Zick, Y.（2016）。通过定量输入影响实现算法透明性：理论及其在学习系统中的实验。载于《2016 年 IEEE 安全与隐私研讨会（SP）》，第 598–617 页。IEEE。

Dua, D. 和 Graff, C.（2017）。UCI 机器学习资料库。

Farrell, S., Calafiura, P., Mudigonda, M., Anderson, D., Vlimant, J.-R., Zheng, S., Bendavid, J., Spiropulu, M.,

Cerati, G., Gray, L., 等（2018）。用于轨迹重建的新型深度学习方法。arXiv 预印本 arXiv:1810.06111。Freiesleben, T., K¨onig, G., Molnar, C. 和 Tejero-Cantero, A.（2022）。利用可解释机器学习进行科学推断：分析模型以了解现实世界现象。arXiv 预印本 arXiv:2206.05487。Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M. 和 Feige, I.（2020a）。数据流形上的 Shapley 可解释性。Frye, C., Rowat, C. 和 Feige, I.（2020b）。非对称 Shapley 值：将因果知识纳入模型无关的可解释性。载于第 34 届神经信息处理系统进展国际会议论文集，NIPS'20，美国纽约州雷德胡克。Curran Associates Inc. G´amez, J., Mateo, J. L. 和 Puerta, J.（2011）。通过爬山法学习贝叶斯网络：基于逐步限制邻域的高效方法。《数据挖掘与知识发现》，22(1):106–148。Hagberg, A., Swart, P. 和 S Chult, D.（2008）。使用 NetworkX 探索网络结构、动态特性和功能。技术报告，洛斯阿拉莫斯国家实验室（LANL），美国新墨西哥州洛斯阿拉莫斯。Heskes, T., Sijben, E., Bucur, I. G. 和 Claassen, T.（2020）。因果 Shapley 值：利用因果知识解释复杂模型的个体预测。神经信息处理系统进展，33:4778–4789。Hothorn, T. 和 Zeileis, A.（2021）。使用变换森林进行预测分布建模。《计算与图形统计学杂志》，30(4):1181–1196。Jethani, N., Sudarshan, M., Covert, I. C., Lee, S.-I. 和 Ranganath, R.（2021）。FastSHAP：实时估计 Shapley 值。载于国际学习表征会议。Kalisch, M., M¨achler, M., Colombo, D., Maathuis, M. H. 和 B¨uhlmann, P.（2012）。使用 R 软件包 pcalg，基于图模型进行因果推断。《统计软件杂志》，47(11):1–26。Li, X., Zhou, Y., Dvornek, N. C., Gu, Y., Ventola, P. 和 Duncan, J. S.（2020）。在不确定性条件下高效估计特征重要性的 Shapley 解释。载于国际医学图像计算与计算机辅助干预会议，页码 792–801。Springer。Luan, H. 和 Tsai, C.-C.（2021）。机器学习方法在精准教育中的应用综述。《教育技术与社会》，24(1):250–266。

Lundberg, S. M. 和 Lee, S.-I.（2017）。一种解释模型预测的统一方法。神经信息处理系统进展，30。Mitchell, R., Cooper, J., Frank, E. 和 Holmes, G.（2022）。用于 SHAP 值估计的排列采样。机器学习研究杂志，23(43):1–46。Molnar, C.（2020）。可解释机器学习。Lulu. com。Molnar, C., K¨onig, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M. 和 Bischl, B.（2022）。机器学习模型的模型无关解释方法常见陷阱。在拓展深度模型和分类器之外的可解释人工智能国际研讨会论文集，第39–68页。Springer。Obermeyer, Z. 和 Mullainathan, S.（2019）。剖析一种为7000万人提供医疗决策指导的算法中的种族偏见。在公平、问责与透明度会议论文集，第89– 89页。Page Jr, T. J.（1984）。多元统计：一种向量空间方法。JMR,《营销研究杂志》（1986年前），21(000002):236。Pearl, J.（2009）。因果关系。剑桥大学出版社，第2版。Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cour- napeau, D., Brucher, M., Perrot 和 Duchesnay, E.（2011）。Scikit-learn：Python 中的机器学习。机器学习研究杂志，12:2825–2830。Peters, J., Janzing, D. 和 Sch¨olkopf, B.（2017）。因果推断要素：基础与学习算法。麻省理工学院出版社，美国马萨诸塞州剑桥。R 核心团队（2022年）。R：统计计算语言与环境。R 统计计算基金会，奥地利维也纳。Raghavan, M., Barocas, S., Kleinberg, J. 和 Levy, K.（2020）。缓解算法招聘中的偏见：评估主张与实践。在2020年公平、问责与透明度会议论文集，FAT* '20，第469–481页，美国纽约州纽约市。计算机协会。Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. 等（2019）。数据驱动的地球系统科学中的深度学习与过程理解。《自然》，566(7743):195–204。Reisach, A. G., Seiler, C. 和 Weichwald, S.（2021）。当心模拟 DAG！因果发现基准测试可能很容易被操纵。Russell, S. 和 Norvig, P.（2009）。《人工智能：一种现代方法》。Prentice Hall，第3版。

Schmidt, J.、Marques,M. R.、Botti, S. 和 Marques, M. A.（2019）。固态材料科学中机器学习的最新进展与应用。《npj计算材料》，5(1)：1–36。Schwarz, G.（1978）。估计模型的维度。《统计学年刊》，6(2)：461–464。Scutari, M.（2010）。使用 bnlearn R 软件包学习贝叶斯网络。《统计软件杂志》，35(3)：1–22。Scutari, M.、Graafland, C. E.和Guti´errez, J.M.（2019a）。谁能更好地学习贝叶斯网络结构：结构学习算法的准确性与速度。《近似推理国际期刊》，115：235–253。Scutari,M.、Vitolo, C. 和Tucker, A.（2019b）。通过贪心搜索从大数据中学习贝叶斯网络：计算复杂度与高效实现。《统计与计算》，29(5)：1095–1108。Senior, A. W.、Evans, R.、Jumper, J.、Kirkpatrick,J.、Sifre,L.、Green, T.、Qin,C.、ˇZ´ıdek, A.、Nelson, A.W.、Bridg-land, A. 等（2020）。使用来自深度学习的势函数改进蛋白质结构预测。《自然》，577(7792)：706–710。Shapley,L. S.（1953）。随机博弈。《美国国家科学院院刊》，39(10)：1095–1100。Sohn, K.、Lee, H. 和 Yan, X.（2015）。使用深度条件生成模型学习结构化输出表示。神经信息处理系统进展，28。Spirtes, P.、Glymour,C.和 Scheines,R.（2000）。因果关系、预测与搜索。麻省理工学院出版社，第2版。ˇStrumbelj, E. 和 Kononenko, I.（2014）。使用特征贡献解释预测模型和个体预测。《知识与信息系统》，41(3)：647–665。Sundararajan, M. 和 Najmi, A.（2020）。用于模型解释的多种 Shapley值。在《国际机器学习会议》中，第9269–9278页。PMLR。Trippe,B.L. 和 Turner, R. E.（2018）。使用贝叶斯归一化流进行条件密度估计。arXiv预印本arXiv:1802.04908。Vallat, R.（2018）。Pingouin：Python 中的统计学。《开源软件杂志》，3(31)：1026。Van den Broeck, G.、Lykov,A.、Schleich,M. 和 Suciu, D.（2022）。关于 SHAP 解释的可处理性。《人工智能研究杂志》，74：851–886。Wang,J.、Wiens, J. 和 Lundberg,S.（2021）。Shapley流：一种用于解释模型预测的基于图的方法。在《国际人工智能与统计学会议》中，第721–729页。PMLR。

Winkler, C., Worrall, D., Hoogeboom, E., 和 Welling, M. (2019). 使用条件归一化流学习似然。arXiv 预印本 arXiv:1912.00042。Zeng, J., Ustun, B., 和 Rudin, C. (2017). 用于再犯预测的可解释分类模型。英国皇家统计学会会刊：A辑（社会统计学），180(3):689–722。

定理 1 的证明

定理1。对于 ℓ 为交叉熵损失或均方误差、 f^* 为相应的最优预测器且$\nu_{\ell,\ f^*}$ 为相应的 *SAGE* 价值函数，有

$$X_j \perp Y|\boldsymbol{X}_S \Rightarrow \nu_{\ell,f^*}(\boldsymbol{X}_{S\cup j}) - \nu_{\ell,f^*}(\boldsymbol{X}_S) = 0.$$

证明。均方误差：Covert等人(2020)表明，对于 ℓ 为均方误差且 f^* 为相应的最优预测器，有：

$$\nu(\mathbf{X}_{S\cup j}) - \nu(\mathbf{X}_S) = \mathbb{E}[\text{Var}(Y|\mathbf{X}_S)] - \mathbb{E}[\text{Var}(Y|\mathbf{X}_{S\cup j})]$$

在条件独立性 $Y \perp X_j|\mathbf{X}_S$ 下可得

$$\begin{aligned}\mathbb{E}[\text{Var}(Y|\mathbf{X}_{S\cup j})] &= \mathbb{E}[\mathbb{E}[\text{Var}(Y|\mathbf{X}_{S\cup j})|\mathbf{X}_S]] \\ &= \mathbb{E}[\text{Var}(Y|\mathbf{X}_S)]\end{aligned}$$

因此 $Y \perp X_j|\mathbf{X}_S \Rightarrow \nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S) = 0$。

交叉熵： Covert 等人（2020）表明，给定交叉熵作为损失函数以及相应的损失最优预测器 f^*则有：

$$\nu(\mathbf{X}_{S\cup j}) - \nu(\mathbf{X}_S) = I(Y; X_j|\mathbf{X}_S)$$

互信息 $I(Y; X_j|\mathbf{X}_S)$ 当且仅当 $Y \perp X_j|\mathbf{X}_S$ 时为零。因此， $\nu(\mathbf{X}^{S\cup j}) - \nu(\mathbf{X}_S) = 0 \Leftrightarrow X_j \perp Y|\mathbf{X}_S$。

□

B SAGE 值的性质

如第 3 节所述，SAGE 值满足某些公平性性质，这些性质是从适用于 Shapley 值的性质推导而来的（Covert 等，2020）。虽然这些性质并未明确以 Shapley 值的性质（效率性、虚拟特征性质、对称性、单调性、线性性）命名，但为简便起见，我们将这些术语用于 SAGE 的性质：

1. 效率性：$\sum_{j=1}^{d} \phi_j(\nu) = \nu(\mathbf{X})$，其中$\mathbf{X}$ 是所有特征的集合。

ˆ 2. 虚拟特征性质： $\phi_j(\nu) = 0$当 $X_j \perp f(\mathbf{X})|\mathbf{X}_S$对所有 $S \subseteq \{1, ..., d\} \setminus j$均成立时。

3. 对称性： $\nu(\mathbf{X}^{S\cup j}) = \nu(\mathbf{X}_{S\cup i})$ 对于两个具有确定性关系的变量 X_j 和 X_i。

ˆ ˆ 4. 单调性：对于两个目标变量 Y、 Y' 及对应的模型 f、 f'：如果对于所有 $S \subseteq \{1, ..., d\} \setminus j$， $\nu_{\hat{}f}(\mathbf{X}^{S\cup j}) - \nu_{\hat{}f}(\mathbf{X}_S) \geq \nu_{\hat{}f'}(\mathbf{X}^{S\cup j}) - \nu_{\hat{}f'}(\mathbf{X}_S)$，则 $\phi_j(\nu_{\hat{}f}) \geq \phi_j(\nu_{\hat{}f'})$。

5.根据 线性性： $\phi_j(\nu) = \mathbb{E}_{\mathbf{X},\ Y}[\phi_j(\nu_{\hat{}f,\ x,\ y})]$, 其中 $\phi_j(\nu_{\hat{f},\ x,\ y})$ 是博弈 $\nu_{\hat{f},\ x,\ y}(\mathbf{X}_S) = \hat{}\hat{}\ell(f_\emptyset(\mathbf{X}_\emptyset),\ y) - \ell(f_S(\mathbf{X}_S),\ y)$ 的 Shapley 值

6. SAGE 值在对输入应用可逆映射后保持不变，例如，对于原始输入数据及其对数值，SAGE 值相同。

C 图基准

本节详细介绍第 5 节中使用的图、图学习算法和图基准。此外，我们还展示了通过 HC 算法对 CSL 得出的结果。

C.1 图概述

表 1 概述了第5 节中使用的全部十二个图、随机抽取的目标变量、目标变量的邻接度，以及相对于目标变量的 d-分离所占比例。这进一步揭示了图稀疏性、目标变量的邻接度与 d-分离所占比例之间的关系。后者可以被视为SAGE近似潜在的相对运行时间下降幅度。

表 1：第 5 节中使用的全部十二个图、随机抽取的目标变量、目标变量的邻接度，以及相对于目标变量的 d-分离所占比例。

图（平均度）	目标变量度	目标变量占比	$\perp_{\mathcal{G}}$
$\text{DAG}_s(2)$	8	2	0.556
$\text{DAG}_s(3)$	1	2	0.357
$\text{DAG}_s(4)$	1	4	0.283
$\text{DAG}_{sm}(2)$	17	1	0.765
$\text{DAG}_{sm}(3)$	2	1	0.623
$\text{DAG}_{sm}(4)$	16	4	0.185
$\text{DAG}_m(2)$	4	1	0.961
$\text{DAG}_m(3)$	32	5	0.556
$\text{DAG}_m(4)$	2	3	0.274
$\text{DAG}_l(2)$	4	3	0.632
$\text{DAG}_l(3)$	66	3	0.552
$\text{DAG}_l(4)$	66	7	0.151

表 2 展示了依托 R（R 核心团队，2022年）的 *bnlearn* 软件包（Scutari，2010年）进行 CSL 时所使用的超参数设置。

表 2：用于图学习的超参数

算法	超参数
HC	最大迭代次数 ∞，最大入度： ∞；评分：贝叶斯信息准则
TABU	列表大小：10；最大迭代次数 ∞，最大入度： ∞；评分：贝叶斯信息准则

C.2 用于 d-分离推断的蒙特卡洛抽样

在算法 2 中，我们详细说明了如何在估计图中推断真阳性、假阳性、真阴性和假阴性d-分离的数量，尤其是如何处理较大图中潜在条件集数量呈指数级增长的问题。

C.3 结果 - 爬山算法

图 7 展示了爬山算法的图学习基准测试结果，并与第 5 节的结果进行了对比。由于爬山算法的表现从未优于 TABU，且在部分实验中表现更差，因此我们选择后者用于 d-SAGE。

算法 2： 用于 d-分离推断的蒙特卡洛采样

输入： 节点集 $\{X_1, X_2, ...X_d, Y\}$ 上目标节点 Y 的真实图 $\mathcal{G}^*$ 和估计图 $\mathcal{G}$；MC 样本数 n_{mc}

输出： $\mathcal{G}$ 中推断出的 d-分离的真阳性、真阴性、假阳性和假阴性：TP、TN、FP、FN

集合 $TP = TN = FP = FN = 0$

对于 $m = 1, ..., n_{mc}$ **do**

从 $\{X_1, X_2, ...X_d\}$ 中随机抽取一个节点 X_j

随机抽取大小为 n_s 的条件集 $\mathbf{X}_S$，来自离散概率分布 $P(n_s = i) = \frac{\binom{d-1}{i}}{2^{d-1}}, i \in \{0, ..., d-1\}$

随机抽取元素 $X_i, i = 1, ...n_s$, 从 $\{X_1, X_2, ...X_d\} \setminus X_j$ 不放回抽样并设置 $\mathbf{X}_S = \{X_i\}_{i=1,...,n_s}$

如果 $X_j \perp_{\mathcal{G}^*} Y | \mathbf{X}_S$ **那么**

  如果 $X_j \perp_{\mathcal{G}} Y | \mathbf{X}_S$ **那么**

    真阳性 = 真阳性

  else

    假阴性 = 假阴性 $^{+1}$

  end

else

  如果 $X_j \not\perp_{\mathcal{G}} Y | \mathbf{X}_S$ **那么**

    真阴性 = 真阴性

  else

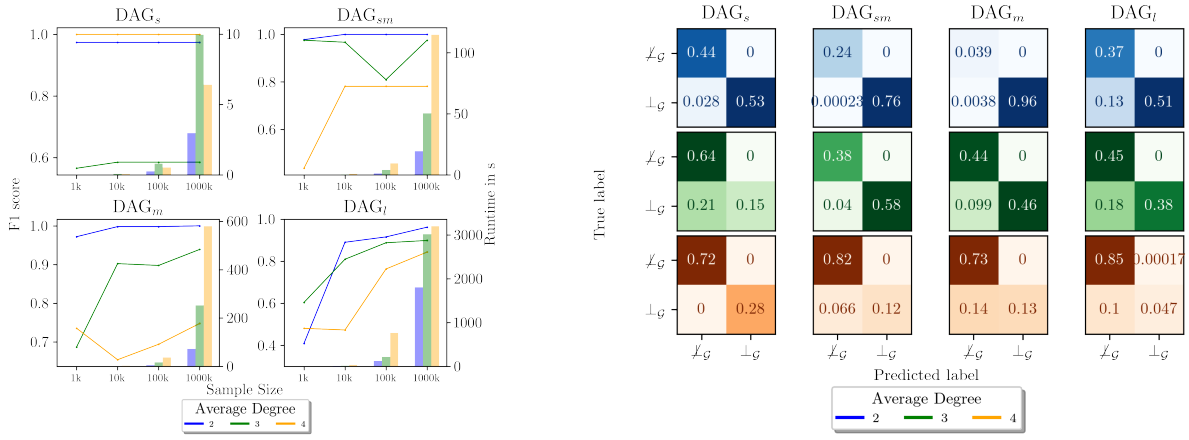
    假阳性 = 假阳性 $^{+1}$

  end

end

end

返回： 真阳性, TN, FP, FN



(a) 使用爬山算法时，不同样本量下 d-分离的 F1 分数（折线，左侧 y 轴）和图学习的运行时间（柱形，右侧 y 轴）。(b) 基于爬山算法、针对全部十二个图且 $n = 10{,}000$ 时，真实与预测的 d-连接（$\not\perp_{\mathcal{G}}$）和 d-分离（$\perp_{\mathcal{G}}$）的混淆矩阵。

图 7：HC 算法在图学习基准测试中的结果。

D SAGE - 实验

在本节中，我们简要说明实验设置，随后呈现缺失的结果。对于我们的分析，我们为每个数据集拟合了两个模型，即语言模型和随机森林，所使用的目标变量与图中d分离分析时随机采样的目标变量相同。我们使用 $n_{train} = 8000$ 进行模型拟合，并使用 $n_{test} = 2000$ 进行模型评估（与用于图拟合和 SAGE 推断的相同 $n = 10000$ 数据点）。随后，我们使用数据估计 SAGE 和d-SAGE 五次，也就是说，对于每个图和模型，我们获得了五个（d-）SAGE 的近似值，并据此给出误差界限。 $\Delta_{j|S}$ 图基于对上述每次运行跳过的评估。

在表3中，我们提供了模型的性能指标，并在附录D.1中展示了与第5节未展示的实验相关的图表。请注意，表3凸显了在所有设置下，就MSE和 R^2而言，RF的表现始终略逊于最优语言模型。

表3：Li 的详细信息 近语言模型（语言模型）和随机森林（随机森林）；基于100个树估计器的随机森林 .

数据（平均度）$\mathbf{n}_{train}$;$\mathbf{n}_{test}$		均方误差$_{LM}$	$\mathbf{R}^2_{LM}$	均方误差$_{RF}$	$\mathbf{R}^2_{RF}$
DAG$_s$(2)	8000; 2000	0.541	0.495	0.572	0.466
DAG$_{sm}$(2)	8000; 2000	0.035	0.963	0.038	0.960
DAG$_m$(2)	8000; 2000	0.474	0.522	0.498	0.498
DAG$_l$(2)	8000; 2000	0.070	0.930	0.103	0.897
DAG$_s$(3)	8000; 2000	0.382	0.616	0.480	0.517
DAG$_{sm}$(3)	8000; 2000	0.072	0.926	0.078	0.921
DAG$_m$(3)	8000; 2000	0.089	0.914	0.174	0.832
DAG$_l$(3)	8000; 2000	0.065	0.938	0.082	0.922
DAG$_s$(4)	8000; 2000	0.101	0.902	0.161	0.843
DAG$_{sm}$(4)	8000; 2000	0.075	0.925	0.086	0.914
DAG$_m$(4)	8000; 2000	0.163	0.840	0.194	0.810
DAG$_l$(4)	8000; 2000	0.004	0.996	0.059	0.943

D.1 结果 - SAGE 和 d-SAGE

本节针对所有缺失设置以及两种模型——语言模型和随机森林——提供与第5节相同的结果，同时还提供第5节所呈现设置的前十五个值。总体而言，我们可以在不同设置下确认我们的发现。

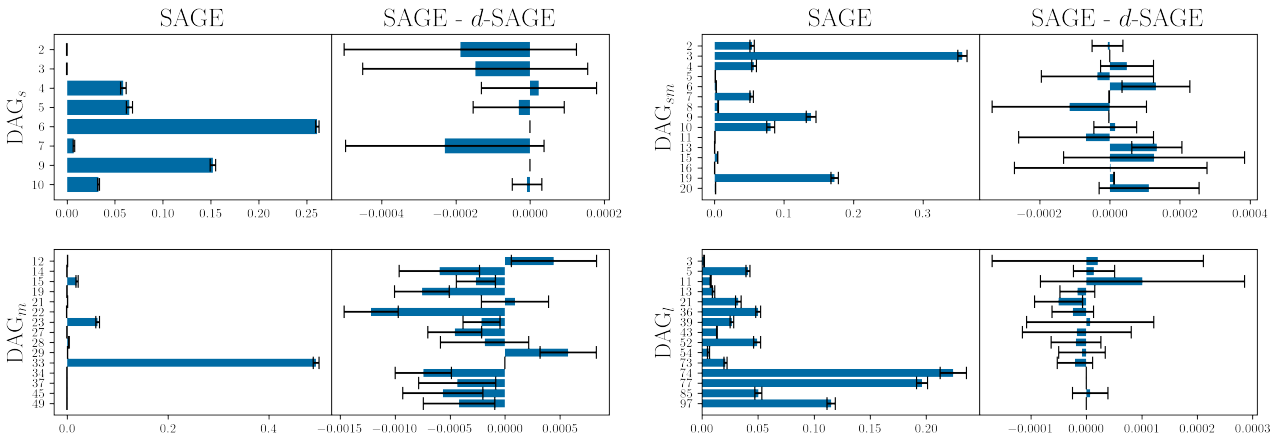
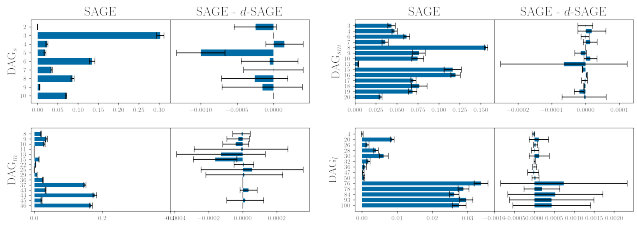
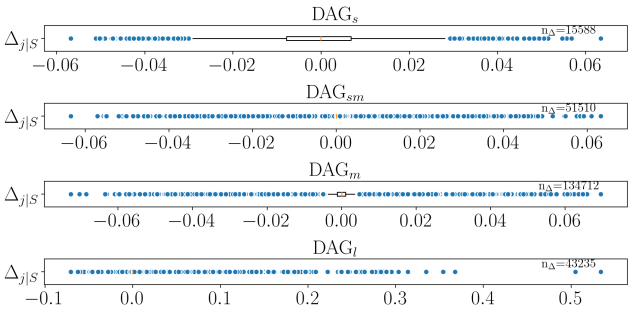


图8：SAGE 值和 SAGE 与 d-SAGE 之间的差异，针对十五个（均对应 DAG$_s$）最大的最优值用于有向无环图的模型，具有平均年龄度为二。基于五个（d-）SAGE 估计值。

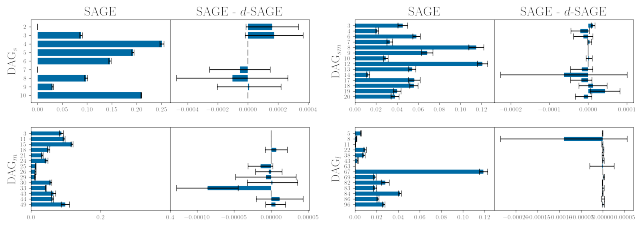


(a) SAGE 值以及 SAGE 与 d-SAGE 之间的差异，用于最优模型中最大的十五个值（全部对应于 DAG_s）。

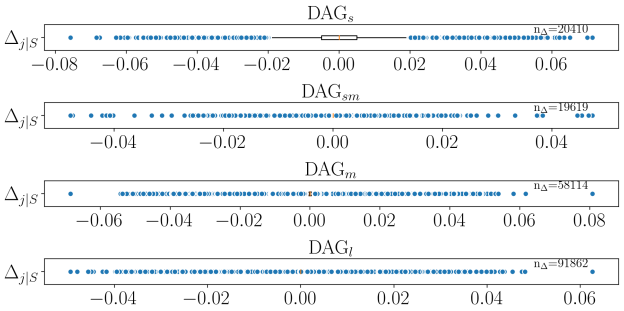


(b) 箱线图展示了被跳过的 sur- 的 $\Delta_{j|S}$ 分布。以及评估。

图 9：基于平均度为三的每个 DAG 和语言模型，对 d-SAGE 的估计质量结果。

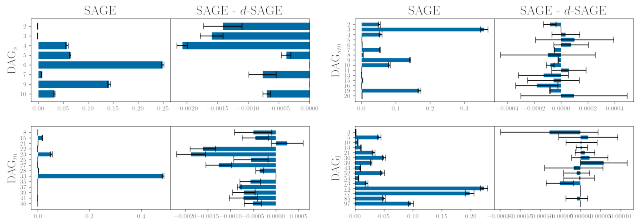


(a) SAGE 值以及 SAGE 与 d-SAGE 之间的差异
最优模型的十五个最大值（全部针对 DAG_s）

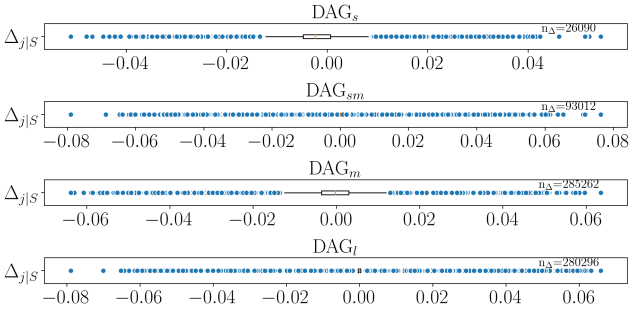


(b) 箱线图展示了被跳过的 sur- 的 $\Delta_{j|S}$ 分布
以及评估。

图10：基于平均度为四的每个DAG和语言模型，对 d-SAGE的估计质量的结果。基于五个（d-）SAGE估计值。

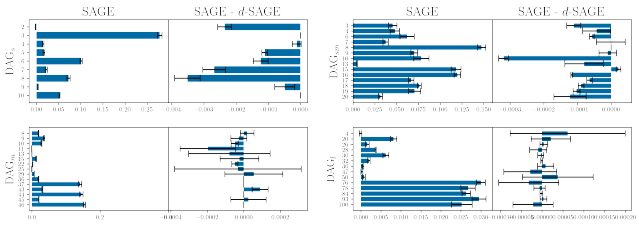


(a) 最优模型中十五个（全部针对DAG_s）最大值的SAGE 值，以及 SAGE 与 d-SAGE 之间的差异。

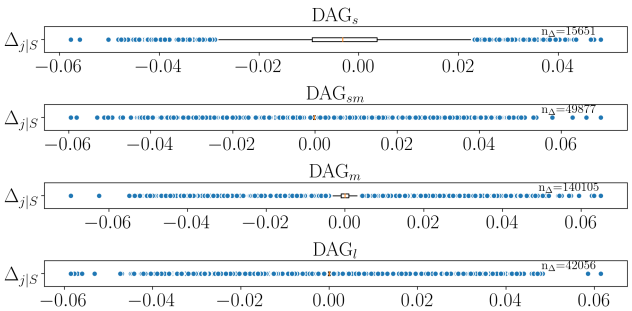


b）箱线图展示了被跳过的 sur 的 $\Delta_{j|S}$ 分布。
以及评估。

图 11：基于每个平均度为 2 的有向无环图和随机森林，对 d-SAGE 的估计质量进行的结果。基于五个（d-）SAGE 估计值。

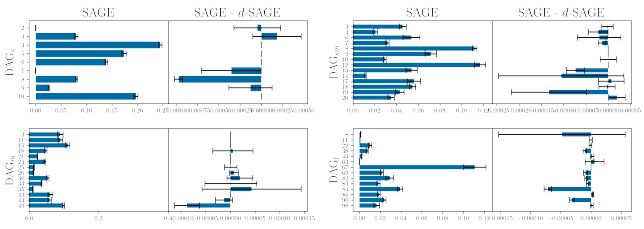


(a) SAGE 值以及 SAGE 与 d-SAGE 之间的差异，用于该最优模型中最大的十五个值（全部针对 DAG_s）。

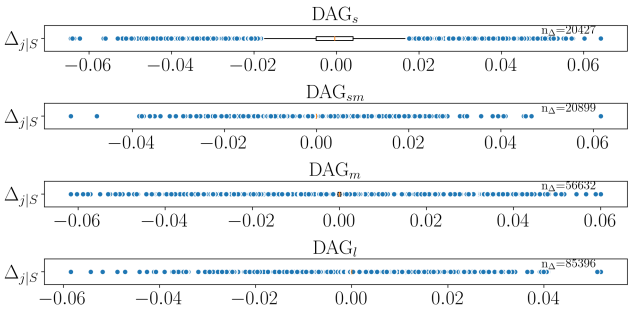


(b) 箱线图展示了跳过的 sur- 的 $\Delta_{j|S}$ 分布以及评估。

图 12：基于平均度为三的每个 DAG 和随机森林，对 d-SAGE 的估计质量的结果。基于五个（d-）SAGE 估计值。



(a) SAGE 值以及 SAGE 与 d-SAGE 之间的差异，对于 DAG_s 中最优模型的十五个最大值。



(b) 箱线图显示跳过的 sur- 中 $\Delta_{j|S}$ 的分布以及评估。

图 13：基于每个平均度为四的DAG和随机森林的 d-SAGE 估计质量结果。基于五个（d-）SAGE 估计值。

E 收敛图

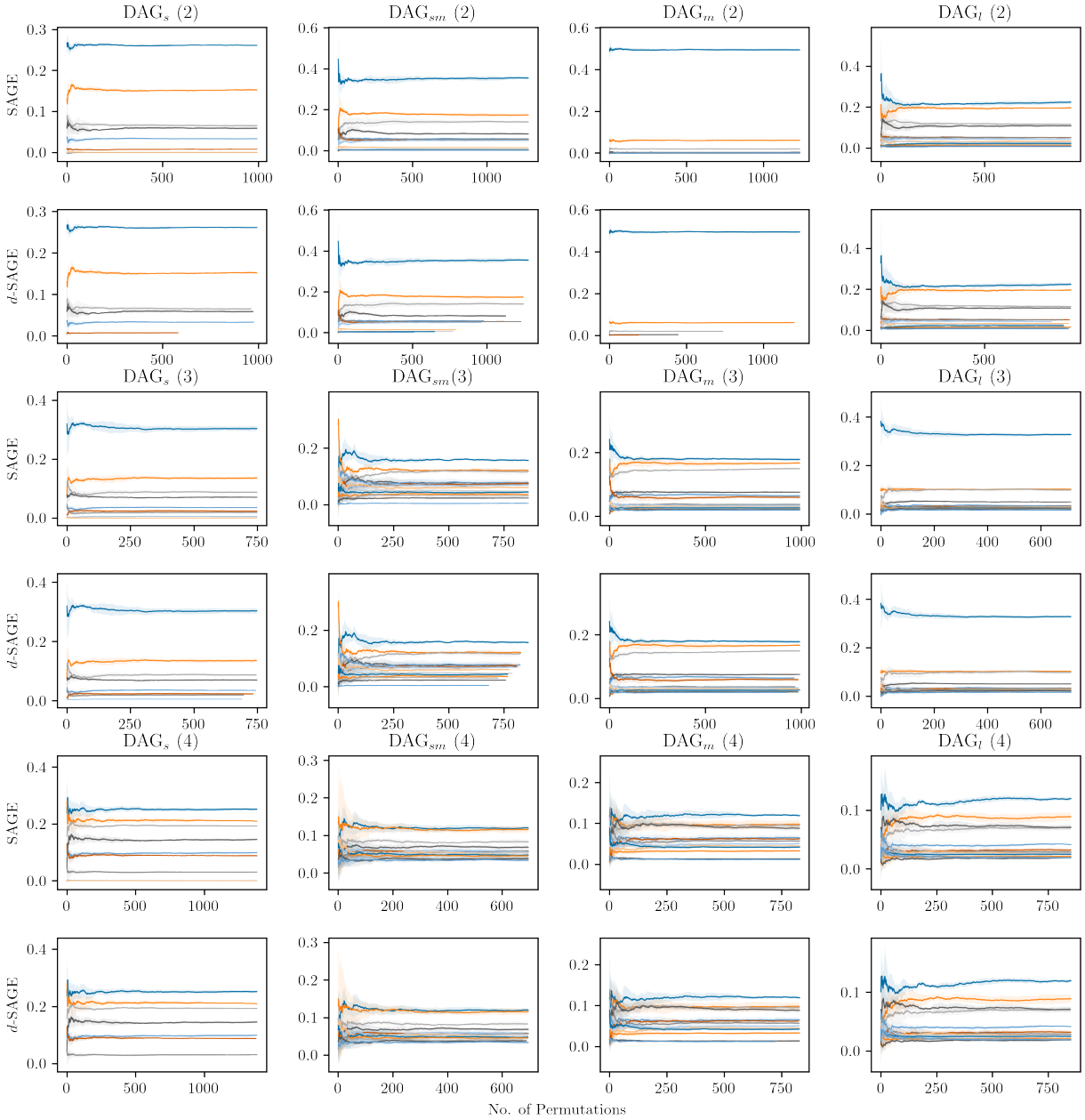


图 14：每个 DAG（平均邻接度）的最优模型（语言模型）中最大的十五个 SAGE 和 d-SAGE值的收敛。对于给定的图，每种颜色代表 SAGE和 d-SAGE 图中相同的特征（如果该特征同时存在于两者中）。基于五个（d-）SAGE 估计值。为便于阅读，省略图例。

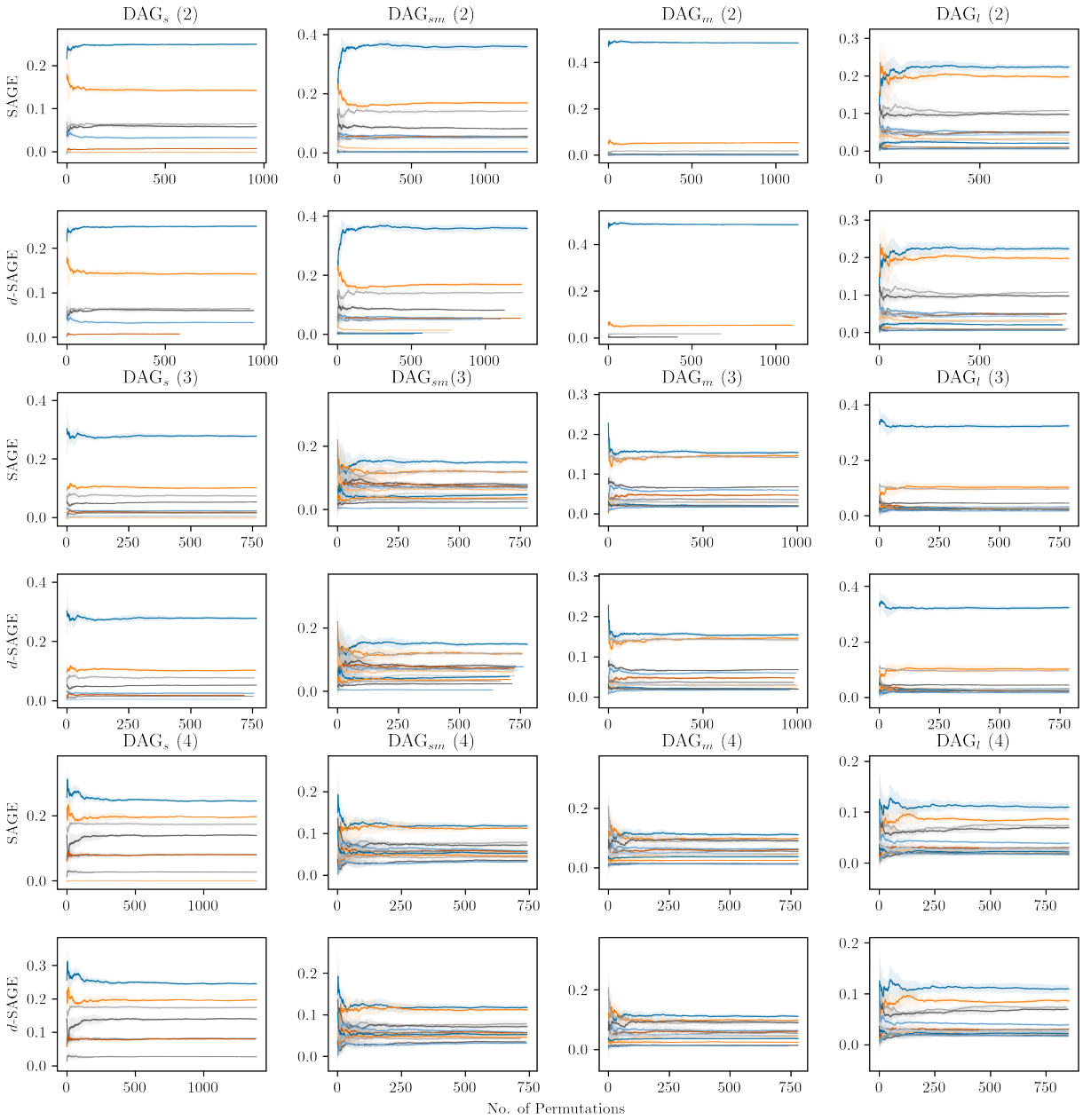


图 15：随机森林模型（RF）在每个 DAG（平均邻接度）下，排名前十五的 SAGE 和 d-SAGE 值的收敛情况。每种颜色代表给定图中 SAGE 和 d-SAGE 图里相同的特征（若两者均存在）。基于五个（d-）SAGE 估计值。为便于阅读，省略图例。

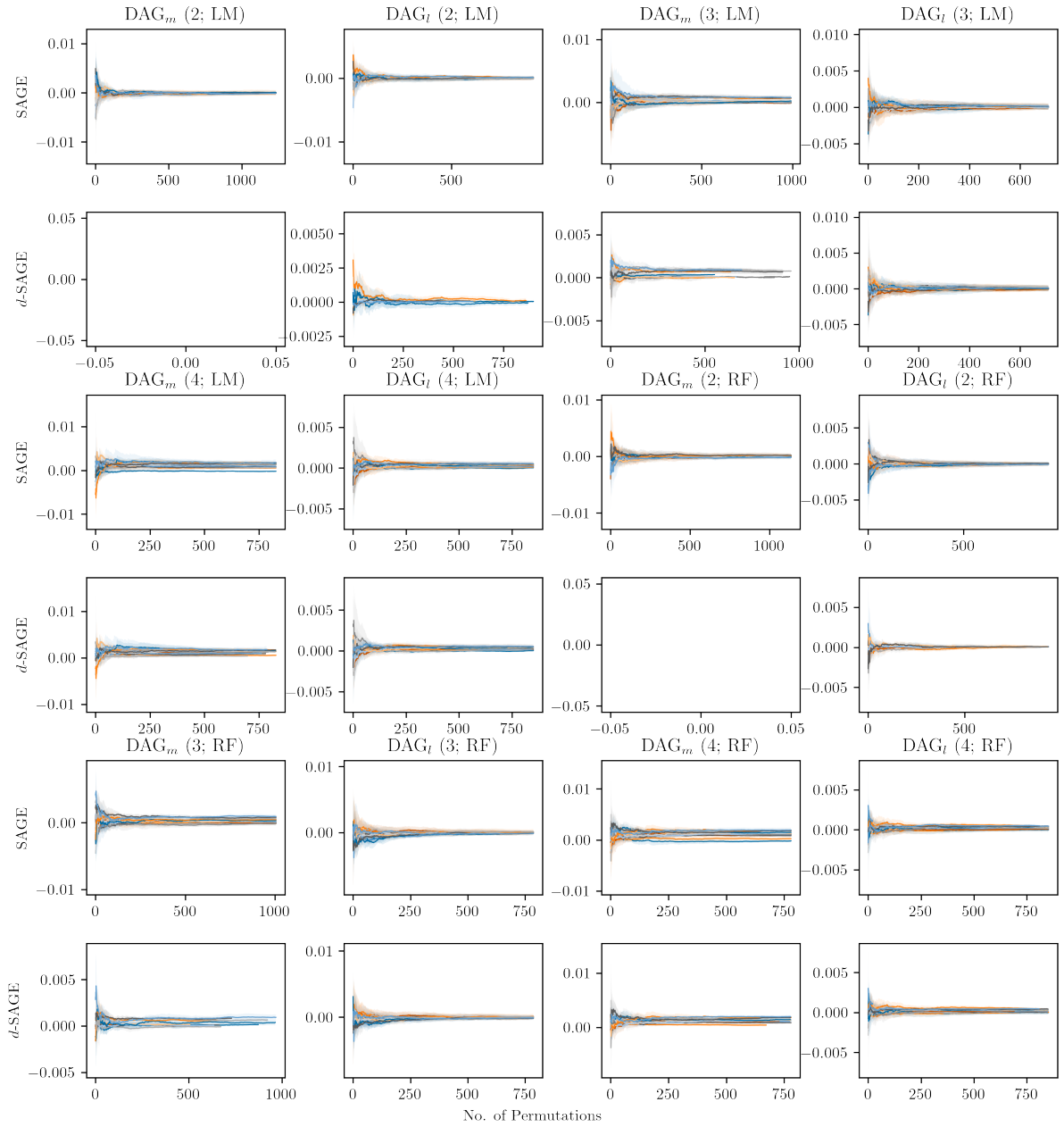


图16：最优模型（语言模型）和随机森林模型（随机森林）中，DAG$_m$ 和 DAG$_l$ （平均邻接度）的底部十五个 SAGE 和 d-SAGE 值的收敛情况。对于给定的图（如果同时存在于两者中），每种颜色代表SAGE和 d-SAGE 图中的同一特征。基于五个（d-）SAGE 估计值。为便于阅读，省略图例。

E.1 SAGE 值的收敛

近似算法的设计要求所有值都收敛后才停止。因此，有些值已经收敛，但仍会继续计算。然而， d-SAGE 的优势取决于条件独立关系所占的比例，而非收敛所需的排列数量，因此即使排列数量较少，也能获得相近的加速效果。收敛图中缺失的线对应于条件独立特征（在每个采样联盟给定的情况下），这凸显了（d-）SAGE 进行事后特征选择的能力。图 14 中对 DAG$_m$(2) 的 SAGE 与 d-SAGE 进行比较，展示了收敛速度更快的 d-SAGE 值示例；其中较小的值（略高于零）在 d-SAGE 中收敛更快。

F 部分相关检验与 d-分离查询

为了凸显 CSL 相较于统计独立性检验的优势，我们将 NetworkX Python 软件包（Hagberg 等，2008）中用于线性时间 d-分离查询（在 TABU 推断出的图中）的运行时间，与 Pingouin 软件包（Vallat，2018）中针对线性高斯数据的部分相关检验运行时间进行了比较。结果基于 100 次排列；每个图的特征数量（分别为 9、19、49 和 99）乘以 100，即为单独检验的次数。表 4 清楚地表明，与 d-分离查询（+ 图学习）相比，部分相关检验通常更准确，但运行时间要长得多。

表 4：基于 $n = 10{,}000$ 和 100 次排列的部分相关检验与 d-分离查询，如 (d-)SAGE 评估中所示；基于 TABU 的图学习；ACC = 准确率。

DATA	时间（d-分离）	时间（TABU）	时间（条件独立关系）	准确率（d-分离）	准确率（条件独立关系）
DAG$_s$ (2)	0.13 秒	0.06秒	46.82s	1.000	1.000
DAG$_{sm}$ (2)	0.39s	0.22秒	166.75 秒	0.996	0.999
DAG$_m$ (2)	1.95s	1.11s	1058.80s	1.000	1.000
DAG$_l$ (2)	15.48s	12.02 秒	4344.81秒	0.863	0.934
DAG$_s$ (3)	0.13 秒	0.18秒	47.15s	1.0	1.0
DAG$_{sm}$ (3)	0.41秒	0.71 秒	166.28s	0.996	0.992
DAG$_m$ (3)	2.12s	2.51s	1089.00s	0.908	0.983
DAG$_l$ (3)	16.85s	18.22s	4299.47秒	0.857	0.941
DAG$_s$ (4)	0.14 秒	0.09s	47.18s	1.0	0.998
DAG$_{sm}$ (4)	0.42秒	1.37s	163.48s	1.0	0.988
DAG$_m$ (4)	2.33s	5.65秒	1093.50秒	0.845	0.940
DAG$_l$ (4)	20.74s	39.86s	4312.16s	0.902	0.916