arXiv:2601.04673v1 [cs.LG] 8 Jan 2026

估计 有限数据下高斯线性结构因果模型中的因果效应估计

Aurghya Maiti * 1 **Prateek Jain** * 1

摘要

从观测数据中估计因果效应仍是因果推断中的一项基础挑战，尤其是在存在潜在混杂因素时。本文聚焦于高斯线性结构因果模型（GL-SCM）中的因果效应估计。这类模型因具有良好的解析处理性而得到广泛应用。然而，由于过参数化，有限数据下的GL-SCM 参数估计往往难以实现。为此，我们提出集中式高斯线性结构因果模型（CGL-SCM）这一类别。它是一个简化但表达能力强的子类，其中外生变量服从标准化分布。我们证明，就从观测分布中识别因果效应而言，CGL-SCM 的表达能力与 GL-SCM 相同；此外，我们还提出一种新颖的、基于 EM 的估计算法，能够学习 CGL-SCM 参数，并利用有限的观测样本估计可识别的因果效应。我们通过合成数据和基准因果图上的实验验证了理论分析，结果表明，学习得到的模型能够准确恢复因果分布。

1. 引言

从观测数据中识别因果效应，是因果推断中的一项基础性挑战，这在很大程度上源于普遍存在的未观测混杂问题。过去几十年来，人们在非参数框架下应对这一挑战方面取得了重大进展，最终形成了 Pearl 的 do 演算等稳健的理论框架，以及用于识别的系统化算法方法（Pearl，1995；Tian& Pearl，2002；Shpitser & Pearl，2007；Huang & Valtorta，2006；Bareinboim & Pearl，2012；2016；Correa& Bareinboim，2017；Leeet al.，2019；Correaetal.，2021；Lee etal.，2020；Correa & Bareinboim，2024）。具体而言，在非参数领域，对识别干预分布（L_2 查询）和反事实分布（L_3 查询）的问题已得到广泛研究，并在很大程度上得以解决。

相比之下，计量经济学和统计学领域传统上是在参数假设下研究因果推断，重点关注结构化模型中因果效应的可识别性，尤其是线性系统（Brito & Pearl，2002；Tian，2005；Chen et al.，2017；Kumor et al.，2019；2020；Shimizu，2014）。然而，这些研究中有许多假设可以获取无限数据、对分布参数有部分了解，或满足马尔可夫假设；后者排除了潜在混杂，在实际场景中可能过于严格。另一方面，近期一些研究致力于利用有限数据估计因果效应，但主要关注非参数模型（Jung et al.，2021a；b）。

本研究旨在弥合这两种观点之间的鸿沟，解决仅使用有限观测数据估计高斯线性结构因果模型（GL-SCM）中可识别因果效应的问题。GL-SCM 因其可解释性和分析便利性而得到广泛应用，但通常具有较高的参数复杂度——尤其是在同时建模观测到的混杂和潜在混杂时——这使得仅凭有限样本进行准确的参数估计变得不可行。

为此，我们引入了集中式高斯线性结构因果模型（CGL-SCM）这一类模型。它是 GL-SCM 的一个简化但具有表达能力的子类，其中所有外生变量均标准化为均值为零、方差为一。这样的集中化处理显著减少了自由参数的数量，并实现了从有限数据中进行可处理的学习。我们的主要贡献有两点：

1. 我们证明，就从观测分布中识别因果效应而言，CGL-SCM 保留了与 GL-SCM 相同的表达能力（定理2.4）。

2. 其次，我们开发了一种可扩展的期望最大化（EM）算法，用于从有限数据中估计 CGL-SCM 的参数（算法 1和算法 2），即使存在潜变量也是如此；随后可利用该算法识别因果效应。

*共同贡献 [1]哥伦比亚大学计算机科学系，纽约州纽约市 10025。联系方式：Au- rghya Maiti <aurghya@cs.columbia.edu>、Prateek Jain <pra- teek.jain@columbia.edu>。

机器学习国际会议 42nd 干预模型扩展研讨会，加拿大温哥华，2025 年。版权所有 2025，归作者所有。

2. 中心化高斯线性结构因果模型

本节首先研究高斯线性结构因果模型这一类模型，然后介绍一种新的中心化高斯线性结构因果模型类别；对于从观测数据中识别效应而言，两者具有相同的表达能力。在高斯线性结构因果模型中，外生变量独立地从正态分布中采样，而内生变量是其父变量的线性组合。接下来，我们将对此类模型给出形式化定义。

定义 2.1（高斯线性结构因果模型）。 GL-SCM 由元组 $\langle(\boldsymbol{U'},\ \boldsymbol{\varepsilon'})、\mathbf{X}\ \mathbb{P}、\mathbf{F_X}\rangle$ 定义，其中：(i) $\mathbf{U}'\ \sim\mathcal{N}(\boldsymbol{\mu_{U'}},\ \boldsymbol{\Sigma^2})$ 是外生混杂变量的集合（$\boldsymbol{\Sigma^2}$ 是对角矩阵）；(ii) $\boldsymbol{\varepsilon'} \sim \mathcal{N}(\boldsymbol{\mu_{\varepsilon'}},\ \boldsymbol{\Psi^2})$ 是外生非混杂变量的集合；(iii) 内生变量 X_i 定义为

$$X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki} U'_k + \mu'_i + \varepsilon'_i$$

其中，$Pa^o(X_i)$ 是 X_i 的观测父节点，$Pa^u(X_i)$ 是 X_i 的未观测混杂父节点。μ'_i 是 X_i 的偏差。

估计此类模型中的精确参数可能很困难，甚至通常是不可能的，因为与观测数据所能提供的信息相比，参数数量可能过多。在本节接下来的部分中，我们将介绍中心化线性结构因果模型，其中外生变量仅从标准分布$\mathcal{N}(0,\ 1)$中采样，误差则从零均值正态分布中采样。与 GL-SCM 相比，此类模型的参数数量可能少得多，而且在许多情况下，这些参数可以仅根据数据进行估计。

定义 2.2（集中式高斯线性结构因果模型）。 CGL-SCM 由元组 $\langle(\boldsymbol{U},\ \boldsymbol{\varepsilon}),\ \mathbf{X}\ \mathbb{P},\ \mathbf{F_X}\rangle$ 定义，其中 (i) $\mathbf{U}$：外生混杂变量集合，且 $\mathbf{U}\ \sim\ \mathcal{N}(\mathbf{0},\ \mathbf{I})$；(ii) $\boldsymbol{\varepsilon}$：外生非混杂变量集合，且$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\Psi^2})$；(iii) 内生变量 X_i 由下式给出

$$X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U_k \in Pa^u(X_i)} \alpha_{ki} U_k + \mu_i + \varepsilon_i$$

其中，$Pa^o(X_i)$ 是 X_i 的观测父节点，而$Pa^u(X_i)$ 是 X_i 的未观测混杂父节点。μ_i 是 X_i的偏差。

现在一个自然的问题是：从 GL-SCM 到 CGL-SCM，我们会损失多少？是否存在某个模型可以用 GL-SCM 表示，却无法用 CGL-SCM 建模？是否存在某种效应，在图模型假设和观测数据下可以在 GL-SCM 中识别，却无法在 CGL-SCM 中识别？下述定理表明，就观测分布而言，GL-SCM 和 CGL-SCM 具有相同的表达能力。

定理 2.3 (CGL-SCM 的表达能力)。 对于 *GL-SCM* 中的任意模型M'，都存在一个因果图相同的 *CGL-SCM* 模型 M，使得 $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$。

CGL-SCM 的一个优点是，当 GL-SCM 因参数数量较多而无法估计时，仍可根据数据进行估计。接下来，我们陈述上述讨论得出的最重要观察结论。假设查询 Q 在 GL-SCM M' 的因果图中可根据观测数据识别。根据可识别性的定义，这意味着任何具有相同因果图和 $P(\mathbf{X})$ 的结构因果模型都会对 Q 得出相同结果。因此，在具有相同 $P(\mathbf{X})$ 的 CGL-SCM 中估计 Q，即可得到 GL-SCM 中 Q 的估计值，尽管无法从现有数据中获知后者的确切参数。我们正式陈述如下定理：

定理 2.4（GL-SCM 中的识别）。 设 M' 和 M分别为 *GL-SCM* 和 *CGL-SCM*，二者具有相同的因果图 G和 $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$。若查询 Q在因果图 G中可识别，则

$$P^{M'}(Q) = P^M(Q) \tag{1}$$

所有证明和推导均见附录。

3. 使用有限数据估计 CGL-SCM

对于我们的问题，假设因果图已知，并且给定来自观测（L_1）分布的有限样本。本节介绍一种仅使用有限观测数据来估计 CGL-SCM 参数的方法。需要注意的是，在这种情况下，参数估计问题可能没有唯一解。然而，我们的目标不一定是恢复真实的数据生成模型，而是找到一个能诱导相同 L_1 分布的 CGL-SCM。估计出这样的模型后，就可以使用该估计出的结构因果模型计算任何可识别的反事实或干预查询。根据定理 2.4，这些查询的结果将与生成数据的原始模型所得结果一致。

为了估计 CGL-SCM 的参数，我们采用期望最大化（EM）算法。这是一种成熟的方法，可在存在潜变量时进行最大似然估计。然而，要在我们的设定中应用 EM，我们必须先将 CGL-SCM 改写为与经典线性高斯框架相符的向量化形式。

3.1. 向量化形式的 CGL-SCM

给定一个 CGL-SCM M，我们首先定义其因果结构的矩阵表示。令$\mathbf{X}$ 表示内生变量向量，并令 G 表示$\mathbf{X}$上的已知因果图。构造一个 $|\mathbf{X}| \times |\mathbf{X}|$ 矩阵 T，其中每个元素

$(i,\ j)$ 定义为:

$$t_{ij} = \begin{cases} \alpha_{ij} & \text{if causal edge } X_i \to X_j \text{ exists in } G \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

接下来，定义矩阵 B，用以刻画图中一条变量到另一条变量的所有有向路径上的总影响。令 d 为 G 中最长路径的长度。矩阵 B 的计算方式如下:

$$B = I + \sum_{i=1}^{d} T^i \tag{3}$$

该求和包含单位矩阵，以计入自我影响，并将其权重定义为 1。因此， B 的第 $(i,\ j)$ 个元素表示图中所有有向路径上从 X_i 到 X_j 的总（加性）影响。

除 B 外，还定义一个 $|\mathbf{U}| \times |\mathbf{X}|$ 矩阵 C，其中每个元素 $(i,\ j)$ 表示外生变量 U_i 对内生变量 X_j 的直接效应。

由此，CGL-SCM 中的内生变量 $\mathbf{X}$ 可仅用 $\mathbf{U}$ 和 $\boldsymbol{\varepsilon}$ 表示，其向量化形式如下:

$$\mathbf{X} = B^T\boldsymbol{\mu} + B^T C^T \boldsymbol{U} + B^T\boldsymbol{\varepsilon} \tag{4}$$

3.2. 算法

在 E 步中，我们计算$\mathbf{U}^i \ \ |\mathbf{x}^i$的分布； $\boldsymbol{\mu}, B,$ $C \sim \mathcal{N}\big(\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}, \Sigma_{\mathbf{U}^i|\mathbf{x}^i}\big)$ 其中,

$$\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i} = CB((CB)^T CB + B^T B)^{-1}(\mathbf{x}^i - B^T\boldsymbol{\mu})$$

$$\Sigma_{\mathbf{U}^i|\mathbf{x}^i} = \mathbf{I} - CB((CB)^T CB + B^T B)^{-1} B^T C^T$$

在 M 步中，我们希望最大化

$$\begin{aligned} max_{B,C,\mu}\ & -n\log|B^T B| \\ & -\sum_{i=1}^{N} \mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\Big[\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^T C^T\mathbf{U}^i\big)^T \\ & \quad \big(B^T B\big)^{-1}\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^T C^T\mathbf{U}^i\big)\Big] \end{aligned}$$

我们应用 EM 算法来估计 $B,\ C$ 和 $\boldsymbol{\mu}$。请注意，直接应用 EM 算法可能会得到一个 $B,\ C$，而它不满足图约束。为此，我们定义掩码 B_m 和 C_m。令 T_m 为一个矩阵，其中元素 $(i,\ j)$ 在存在边 $X_i \to X_j$时为 1。于是， B_m 定义为

$$A = \sum_{i=1}^{d} T_m^i, \qquad b_{ij}^m = \begin{cases} 1 & \text{if } a_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

算法1 CGL-Go

初始化： 将 $B, C, \boldsymbol{\mu}$ 和 B_m, C_m 初始化为定义为见 3.2
$B = I + B \cdot B_m,\quad C = C \cdot C_m$
重复
  E步（期望步）：

$$\mu_{U^i|X^i} = CB((CB)^T CB + B^T B)^{-1}(\mathbf{X}^i - \mathbf{B}^T\boldsymbol{\mu})$$

$$\Sigma_{U^i|X^i} = I - CB((CB)^T CB + B^T B)^{-1}(CB)^T$$

  M步（最大化步）：
最大化给定的 $l(\mathbf{B}、\mathbf{C}\ \boldsymbol{\mu})$

$$\begin{aligned} & -n\log|\mathbf{B}^T\mathbf{B}| \\ & -\sum_{i=1}^{N}\mathbb{E}_{U^i|x^i}[(\mathbf{X}^i - (B^T\boldsymbol{\mu} + (CB)^T\mathbf{U}^i))^T \\ & \quad (B^T B)^{-1}(\mathbf{X}^i - (B^T\boldsymbol{\mu} + (CB)^T\mathbf{U}^i))] \end{aligned}$$

更新$\mathbf{B}$:
  重复
    对于 每个训练样本$\mathbf{X}^i$ **执行**
      计算梯度： $\nabla_B l(B,\ C,\ \boldsymbol{\mu})$
      更新参数： $B \leftarrow B + \eta\nabla_B l(B,\ C,\ \boldsymbol{\mu}) \cdot B_m$
    结束循环
  直到收敛
  更新$\mathbf{C}$:
  重复
    对于每个训练样本 $\mathbf{X}^i$ **执行**
      计算梯度： $\nabla_C l(B, C, \boldsymbol{\mu})$
      更新参数： $C \leftarrow C + \eta\nabla_C l(B, C, \boldsymbol{\mu}) \cdot C_m$
    结束循环
直到收敛
  更新 $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\Big((B^T)^{-1}\mathbf{X}^i - C^T\boldsymbol{\mu}_{U^i|X^i}\Big)$$

直到收敛

其中 a_{ij} 和 b_{ij}^m 分别是 $(i,\ j)$ 中 A 和 B_m 的元素，且 d 是图中的最长路径长度。在 C_m中，元素 $(i,\ j)$ 在存在边 $U_i \to U_j$ 时为 1，否则为 0。在最大化过程中，我们没有找到 $B,\ C$ 的闭式解，因此在此步骤中使用梯度上升来求得这些矩阵。在使用梯度上升进行优化时，我们对梯度进行掩码处理，以维护图约束。最终算法见算法 1。推导过程见附录。

矩阵 C 直接给出每个外生混杂因子到内生变量的边权重

算法2CGL-Edge

1: **对于**每个节点 X_i，在**X**中**执行**
2: 确定内生变量的拓扑序 τ_i，
变量 $\bar{X}_i$ 通过从 X_i 出发的直接外向边连接
使用因果图 G
3: **对于**每个 X_k，在 $\bar{X}_i$ 按拓扑序 τ_i **执行**
4: 计算边权重 t_{ik}

$$t_{ik} = b_{ik} - \sum_{\tau_i(X_j)<\tau_i(X_k)} t_{ij} b_{jk}$$

5: **结束循环**
6: **结束循环**

中**U** 到内生变量 **X**的边权重。但矩阵 B 仅给出每对内生变量之间的聚合权重。为了识别矩阵 T，我们使用算法 2。

4. 实验

我们在由两个合成高斯线性结构因果模型生成的合成数据集上评估了我们的算法。这些因果图被称为前门图（图 1）和餐巾图（图 2），在因果推断文献中得到了广泛研究。对于每个图，我们使用其对应的 CGL-SCM 生成了 10,000 个内生变量样本，并应用算法 1和2，从观测数据中估计结构因果模型参数。随后，我们将得到的因果分布与原始模型的因果分布进行了比较。完整的参数矩阵见附录；具有代表性的比较结果如下表 1和 2所示。

4.1. 前门图

我们使用了以下 CGL-SCM $M = \langle (\boldsymbol{U},\ \varepsilon),$ 其中 $\mathbf{X}\ \mathbb{P},\ \mathbf{F_X}\rangle$诱导出如图 1 所示的前门图，其中 $\mu = \begin{bmatrix}0.3 & 0.1 & 0.2\end{bmatrix}^T$。

表 *1*。前门图的原始与估计的因果分布比较

	原始值	估计值
$P(X_3\|do(X_2=1))$	$\mathcal{N}$(1。1，1。09)	$\mathcal{N}(1.1018, 1.069)$
$P(X_3\|do(X_1=1))$	$\mathcal{N}(0.74, 1.9)$	$\mathcal{N}(0.7391, 1.881)$

4.2. 餐巾图

我们使用了以下 CGL-SCM $M = \langle(\boldsymbol{U},\ \varepsilon),\ \mathbf{X}\ \mathbb{P},\ \mathbf{F_X}\rangle$，生成了如图 2 所示的餐巾图，其中 $\mu = \begin{bmatrix}0.8 & -0.9 & 0.01 & -0.5\end{bmatrix}^T$。
正如这两个例子所示，估计得到的因果分布与原始分布非常接近。

表 *2*。餐巾图的原始与估计的因果分布

	原始	估计值
$P(X_4\|do(X_3=1))$	$\mathcal{N}(0.3, 1.16)$	$\mathcal{N}(0.3051, 1.1692)$
$P(X_4\|do(X_1=1))$	$\mathcal{N}(-1.068,\ 2。3248)$	$\mathcal{N}(-0.9721,\ 2。3274)$

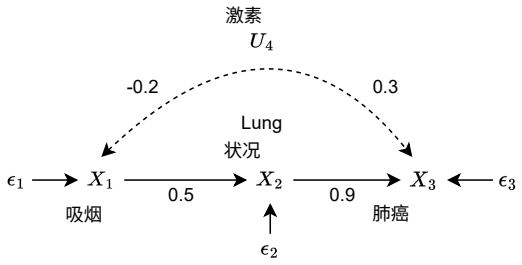

图 *1.* 带有边权重的前门因果图

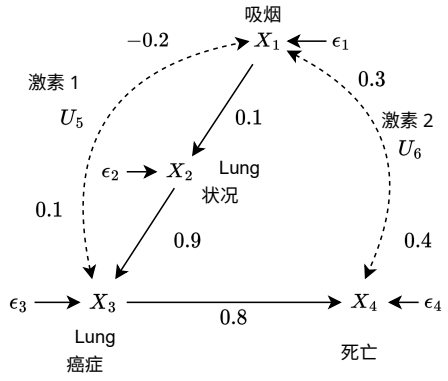

图*2.* 餐巾因果图及其边权重

5. 结论

在这项工作中，我们提出了中心化高斯线性结构因果模型，作为高斯结构因果模型的一个易于处理的子类，同时保留了从观测数据中识别因果效应的能力。我们证明了它们在表达能力和可识别性方面与 GL-SCM 等价，并提出了一种基于 EM 的学习算法来估计参数，进而在已知参数后估计因果效应。通过在前门结构和餐巾结构等典型图上的实验，我们表明，该方法能够从有限样本中准确恢复干预分布。未来的工作可以探索更一般的分布，或为基于有限数据的不可识别查询寻找界限。

6. 影响声明

本文旨在推动机器学习领域的发展。我们的工作可能会对社会产生许多影响，但我们认为没有哪一项需要在此特别指出。

参考文献

Bareinboim，E. 和 Pearl，J. 通过替代实验进行因果推断：z-可识别性。载于《第二十八届人工智能不确定性会议论文集》，第 113–120 页，2012 年。

Bareinboim，E. 和 Pearl，J. 因果推断与数据融合问题。载于 Shiffrin，R. M.（编），《美国国家科学院院刊》，第 113 卷，第 7345–7352 页。美国国家科学院，2016 年。

Brito，C. 和 Pearl，J. 线性模型中因果效应可识别性的图形判据。载于《全国人工智能会议论文集》，第 2 卷，第 533–540 页，2002 年。网址：https://cdn.aaai.org/AAAI/2002/AAAI02-080.pdf。

Chen, B.、Kumor, D. 和 Bareinboim, E. 使用辅助变量在线性结构方程模型中进行识别与模型检验。载于 Precup, D. 和 Teh, Y. W.（编），第*34*届国际机器学习会议论文集，机器学习研究论文集第70卷机器学习研究论文集，第1156–1164页，澳大利亚悉尼国际会议中心，2017年8月。PMLR。

Correa, J. 和 Bareinboim, E. 混杂和选择偏差下通过调整识别因果效应。载于第*31*届*AAAI*人工智能会议论文集，第3740–3746页，美国加利福尼亚州旧金山，2017年2月。AAAI出版社。

Correa, J. 和 Bareinboim, E. 反事实图模型：约束与推断。技术报告 R-115，哥伦比亚大学因果人工智能实验室，2024年8月。

Correa, J., Lee, S., 和 Bareinboim, E. 从任意替代实验中识别嵌套反事实。《神经信息处理系统进展》，34: 6856–6867，2021。Huang, Y. 和 Valtorta, M. Pearl干预演算是完备的。载于第二十二届人工智能不确定性会议论文集，UAI'06，第 217–224 页，阿灵顿，弗吉尼亚州，美国，2006。AUAI Press。ISBN 0974903922。

Jung, Y., Tian, J. 和 Bareinboim, E. 通过双重机器学习估计可识别的因果效应。载于《*AAAI* 人工智能会议论文集》，第 35 卷，第 12113–12122 页，2021a。

Jung, Y.、Tian, J. 和 Bareinboim, E. 通过双重机器学习估计马尔可夫等价类中可识别的因果效应。载于国际机器学习会议，第 5168–5179 页。PMLR，2021b。

Kumor, D.、Chen, B. 和 Bareinboim, E. 在线性结构因果模型中利用工具变量割集进行高效识别。载于《神经信息处理系统进展》，第 32 卷，2019。网址：https://proceedings.neurips.cc/paper/2019/file/2a845d4d23b883acb632fefd814e175f-Paper.pdf.

Kumor, D.、Cinelli, C. 和 Bareinboim, E. 线性结构因果模型中利用辅助割集进行高效识别。载于国际机器学习会议，第 5501–5510 页。PMLR，2020 年。

Lee, S.、Correa, J. D. 和 Bareinboim, E. 任意替代实验下的一般可识别性。载于第 *35* 届人工智能不确定性会议论文集，2019 年。

Lee, S.、Correa, J. D. 和 Bareinboim, E. 任意替代实验下的一般可识别性。载于人工智能不确定性，第 389–398 页。PMLR，2020 年。

Pearl, J. 面向实证研究的因果图。*Biometrika*，82(4)：669–688，1995 年。

Shimizu, S. Lingam：用于估计因果结构的非高斯方法。*Behaviormetrika*，41(1)：65–98，2014 年。

Shpitser, I. 和 Pearl, J. 哪些反事实可以检验。载于第二十三届人工智能不确定性会议论文集，UAI'07，第 352–359 页，美国弗吉尼亚州阿灵顿，2007 年。AUAI Press。ISBN 0974903930。

Tian, J. 在线性模型中识别直接因果效应。载于美国人工智能全国会议论文集，第 20 卷，第 346–353 页，2005 年。网址 https://cdn.aaai.org/AAAI/2005/AAAI05-055.pdf。

Tian、J. 和 Pearl、J. 因果效应的一般识别条件。载于第十八届全国人工智能会议，第 567–573 页，美国，2002 年。美国人工智能协会。ISBN 0262511290。

A. 背景

介绍并阐明干预和可识别性的概念至关重要——它们是因果关系领域中的关键方面，也是本文不可或缺的组成部分。

定义 A.1（干预）。给定结构因果模型 M 中两个互不相交的变量集合 $\mathbf{Z}$、$\mathbf{Y} \subseteq \mathbf{X}$，$\mathbf{Z}$ 对$\mathbf{Y}$的因果效应记为 $P(\mathbf{y}|do(\mathbf{z}))$，它是一个从 $\mathbf{Z}$ 到$\mathbf{Y}$的概率分布空间的函数。对于 $\mathbf{Z}$ 的每个取值 $\mathbf{z}$，$P(\mathbf{y}|do(\mathbf{z}))$ 给出通过从模型中删除所有与 $\mathbf{Z}$ 中变量对应的方程，并在剩余方程中代入$\mathbf{Z} = \mathbf{z}$后，$\mathbf{Y} = \mathbf{y}$ 的概率。

定义 A.2（可识别性）。给定观测变量集合 $\mathbf{X}$，其中$\mathbf{Z}$、$\mathbf{Y} \subseteq \mathbf{X}$，若$\mathbf{Z}$ 对$\mathbf{Y}$ 的因果效应可由因果图 G 识别，则称该效应具有可识别性；也就是说，量 $P(\mathbf{y} \mid do(\mathbf{z}))$ 能够根据观测变量的正概率唯一计算得出。换言之，对于每一对诱导出 G 的模型 M_1 和 M_2，

$$P^{M_1}(\mathbf{y} \mid do(\mathbf{z})) = P^{M_2}(\mathbf{y} \mid do(\mathbf{z}))$$

当 $P^{M_1}(\mathbf{x}) = P^{M_2}(\mathbf{x}) > 0$ 时。

B. 证明

B.1. 定理 2.3 的证明

设有一个 GL-SCM $M' = \langle (\boldsymbol{U}, \boldsymbol{\varepsilon}'), \mathbf{X}\ \mathbb{P}, \mathbf{F}_{\mathbf{X}} \rangle$。内生变量 X_i 则为

$$X_i = \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki} U'_k + \mu'_i + \varepsilon'_i$$

其中，$Pa^o(X_i)$ 是 X_i 的观测父节点，$Pa^u(X_i)$ 是 X_i 的未观测混杂父节点。μ'_i 是 X_i 的偏差。X_i 可重写为:

$$\begin{aligned} X_i &= \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki}(\mu_{U'_k} + \sigma^2_{U'_k} U_k) + \mu'_i + \mu_{\varepsilon'_i} + \varepsilon_i \\ &= \sum_{X_j \in Pa^o(X_i)} \alpha_{ji} X_j + \sum_{U_k \in Pa^u(X_i)} \alpha_{ki} U_k + \mu_i + \varepsilon_i \end{aligned}$$

其中,

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}^2), \quad \alpha_{ki} = \alpha'_{ki}\sigma^2_{U'_k}, \quad \mu_i = \left(\sum_{U'_k \in Pa^u(X_i)} \alpha'_{ki}\mu_{U'_k} \right) + \mu'_i + \mu_{\varepsilon'_i}$$

这正是 CGL-SCM 的定义。因此，对于每个 GL-SCM M'，都存在一个 CGL-SCM M，使得 $P^{M'}(\mathbf{X}) = P^M(\mathbf{X})$。

B.2. 定理 2.4 的证明

该证明直接来自可识别性的定义：若查询 Q 可由观测数据 $P(V)$ 和因果图识别，则所有在 $P(V)$ 上匹配的模型都必须在 Q 上匹配。现在，CGL-SCM 是 GL-SCM 的子类；如果一个 Q 可由 $P(V)$ 和因果图识别，那么它在 Q 中也必须匹配。

C. EM步骤的推导

$\mathbf{U}$ 和$\mathbf{X}$的分布可写为 [1]

$$\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X} = B^T \boldsymbol{\mu} + B^T C^T \boldsymbol{U} + B^T \boldsymbol{\varepsilon}$$

[1]在这一部分，假设 ε的方差为 $\mathbf{I}$。对于一般的 $\boldsymbol{\Psi}$，推导过程类似。

由此，我们可以在计算均值和方差后，推导出 (U,X) 的联合分布 e

$$\mathbb{E}[\mathbf{U}] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{X}] = B^T\boldsymbol{\mu}$$

现在，方差 $\Sigma_{UU} = I$（根据定义）。Σ_{UX} 和 Σ_{XX} 可按如下方式计算：

$$\begin{aligned}\Sigma_{UX} &= \mathbb{E}[(\mathbf{U} - \mathbb{E}[\mathbf{U}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \mathbb{E}[\mathbf{U}(B^T\boldsymbol{\mu} + B^TC^T\mathbf{U} + B^T\varepsilon - B^T\boldsymbol{\mu})^T] \\ &= \mathbb{E}[\mathbf{U}\mathbf{U}^TCB + \boldsymbol{\varepsilon}^TB] = \mathbb{E}[\mathbf{U}\mathbf{U}^T]CB + \mathbb{E}[\boldsymbol{\varepsilon}^T]B = CB\end{aligned}$$

同样，Σ_{XX}

$$\begin{aligned}\Sigma_{XX} &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \mathbb{E}[(B^TC^T\mathbf{U} + B^T\boldsymbol{\varepsilon})(B^TC^T\mathbf{U} + B^T\boldsymbol{\varepsilon})^T] \\ &= \mathbb{E}[B^TC^TUU^TCB + B^TC^T\mathbf{U}\boldsymbol{\varepsilon}^TB + B^T\boldsymbol{\varepsilon}\mathbf{U}^TCB + B^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^TB] \\ &= B^TC^T\mathbb{E}[\mathbf{U}\mathbf{U}^T]CB + B^T\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]B = (CB)^TCB + B^TB\end{aligned}$$

联合分布（$\mathbf{U}$，$\mathbf{X}$）可写为

$$\begin{bmatrix}\mathbf{U} \\ \mathbf{X}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0} \\ B^T\boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\mathbf{I} & CB \\ (CB)^T & (CB)^TCB + B^TB\end{bmatrix}\right)$$

在 E 步中，我们计算 $\mathbf{U}^i \mid \mathbf{x}^i$ 的分布；$\boldsymbol{\mu}$，B，$C \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}, \Sigma_{\mathbf{U}^i|\mathbf{x}^i})$ 其中，

$$\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i} = CB((CB)^TCB + B^TB)^{-1}(\mathbf{x}^i - B^T\boldsymbol{\mu}), \qquad \Sigma_{\mathbf{U}^i|\mathbf{x}^i} = \mathbf{I} - CB((CB)^TCB + B^TB)^{-1}B^TC^T$$

$\mathbf{X}^i \mid \mathbf{u}^i$ 的均值和方差；$\boldsymbol{\mu}$，B，C 也可以类似地计算，

$$\boldsymbol{\mu}_{\mathbf{X}^i|\mathbf{u}^i} = B^T\boldsymbol{\mu} + B^TC^T\mathbf{u}^i, \qquad \Sigma_{\mathbf{X}^i|\mathbf{u}^i} = B^TB$$

在 M 步中，我们希望最大化

$$\begin{aligned}&\sum_{i=1}^{N}\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}[\log p(\mathbf{x}^i \mid \mathbf{U}^i; \boldsymbol{\mu}, B, C)] \\ &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|B^TB| - \frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\Big[\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^TC^T\mathbf{U}^i\big)^T\big(B^TB\big)^{-1}\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^TC^T\mathbf{U}^i\big)\Big]\end{aligned}$$

关于 B，C，$\boldsymbol{\mu}$

为了对此进行优化，我们必须计算第三项并求其期望

$$\begin{aligned}&\sum_{i=1}^{N}\mathbb{E}\Big[\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^TC^T\mathbf{U}^i\big)^T\big(B^TB\big)^{-1}\big(\mathbf{x}^i - B^T\boldsymbol{\mu} - B^TC^T\mathbf{U}^i\big)\Big] \\ &= \sum_{i=1}^{N}(\mathbf{x}^i)^T\big(B^TB\big)^{-1}\mathbf{x}^i - 2(\mathbf{x}^i)^TB^{-1}\big(\boldsymbol{\mu} + C^T\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}[\mathbf{U}^i]\big) + \boldsymbol{\mu}^T\boldsymbol{\mu} + 2\boldsymbol{\mu}^TC^T\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}[\mathbf{U}^i] + \mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\Big[\big(\mathbf{U}^i\big)^TCC^T\mathbf{U}^i\Big]\end{aligned}$$

项 $\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\left[(\mathbf{U}^i)^T CC^T\mathbf{U}^i\right]$ 可以按如下方式计算:

$$\mathbb{E}_{\mathbf{U}^i|\mathbf{x}^i}\left[\left(\mathbf{U}^i\right)^T CC^T\mathbf{U}^i\right] = \mathbb{E}\Big[\sum_j\sum_{l,k} c_{lj}c_{jk}\cdot U_l^iU_k^i \mid \mathbf{x}^i\Big] = \sum_j\sum_{l,k} c_{lj}c_{jk}\mathbb{E}[U_l^iU_k^i \mid \mathbf{x}^i] = \sum_{l,k}\sum_j c_{lj}c_{jk}\mathbb{E}[U_l^iU_k^i \mid \mathbf{x}^i]$$
$$= \sum_{l,k}\mathbb{E}[U_l^iU_k^i \mid \mathbf{x}^i]\sum_j c_{lj}c_{jk} = \mathbb{E}[\mathbf{U}^i\mathbf{U}^{iT} \mid \mathbf{x}^i]\cdot CC^T = \left(\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}\boldsymbol{\mu}_{\mathbf{U}^i|\mathbf{x}^i}^T + \Sigma_{\mathbf{U}^i|\mathbf{x}^i}\right)\cdot CC^T$$

D. 实验

D.1. 前门图

以下是前门 CGL-SCM $M = \langle(\boldsymbol{U},\ \boldsymbol{\varepsilon}),\ \mathbf{X}\ \mathbb{P},\ \mathbf{F}_{\mathbf{X}}\rangle$ 的向量化形式:

$$\mathbf{U} = \begin{bmatrix}U_4\end{bmatrix} \sim N(0,1) \qquad \boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\end{bmatrix} \sim \mathcal{N}(0,\mathbf{I}') \qquad \mu = \begin{bmatrix}0.3\\ 0.1\\ 0.2\end{bmatrix} \qquad T = \begin{bmatrix}0 & 0.5 & 0\\ 0 & 0 & 0.9\\ 0 & 0 & 0\end{bmatrix}$$

$$C = \begin{bmatrix}-0.2 & 0 & 0.3\end{bmatrix} \qquad B = I + \textstyle\sum_{i=1}^2 T^i = \begin{bmatrix}1 & 0.5 & 0.45\\ 0 & 1 & 0.9\\ 0 & 0 & 1\end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix}X_1\\ X_2\\ X_3\end{bmatrix} = B^T\boldsymbol{\mu} + B^TC^T\boldsymbol{U} + B^T\boldsymbol{\varepsilon}$$

运行算法 1 后，我们得到以下关于 B、C 和 μ 的估计值

$$\hat{\mu} = \begin{bmatrix}0.3002\\ 0.0963\\ 0.2007\end{bmatrix} \qquad \hat{C} = \begin{bmatrix}-0.2149 & 0 & 0.2626\end{bmatrix} \qquad \hat{B} = \begin{bmatrix}1 & 0.5012 & 0.4491\\ 0 & 1 & 0.9011\\ 0 & 0 & 1\end{bmatrix}$$

运行算法2后，我们得到了关于 T的以下估计值

$$\hat{T} = \begin{bmatrix}0 & 0.5012 & 0\\ 0 & 0 & 0.9011\\ 0 & 0 & 0\end{bmatrix}$$

D.2. 餐巾图

以下是餐巾 CGL-SCM $M = \langle(\boldsymbol{U},\boldsymbol{\varepsilon}),\mathbf{X}\ \mathbb{P},\mathbf{F}_{\mathbf{X}}\rangle$ 的向量化形式:

$$\mathbf{U} = \begin{bmatrix}U_5\\ U_6\end{bmatrix} \sim N(0,\mathbf{I}) \qquad \boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\end{bmatrix} \sim \mathcal{N}(0,\mathbf{I}') \qquad \mu = \begin{bmatrix}0.8\\ -0.9\\ 0.01\\ -0.5\end{bmatrix} \qquad T = \begin{bmatrix}0 & 0.1 & 0 & 0\\ 0 & 0 & 0.9 & 0\\ 0 & 0 & 0 & 0.8\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$C = \begin{bmatrix}-0.2 & 0 & 0.1 & 0\\ 0.3 & 0 & 0 & 0.4\end{bmatrix} \qquad B = I + \textstyle\sum_{i=1}^3 T^i = \begin{bmatrix}1 & 0.1 & 0.09 & 0.072\\ 0 & 1 & 0.9 & 0.72\\ 0 & 0 & 1 & 0.8\\ 0 & 0 & 0 & 1\end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix}X_1\\ X_2\\ X_3\\ X_4\end{bmatrix} = B^T\boldsymbol{\mu} + B^TC^T\boldsymbol{U} + B^T\boldsymbol{\varepsilon}$$

运行算法 1 后，我们得到了以下关于 B、C 和 μ

$$\hat{\mu} = \begin{bmatrix} 0.7994 \\ -0.9044 \\ 0.0113 \\ -0.4894 \end{bmatrix} \quad \hat{C} = \begin{bmatrix} -0.1271 & 0 & 0.1665 & 0 \\ 0.3389 & 0 & 0 & 0.4113 \end{bmatrix} \quad \hat{B} = \begin{bmatrix} 1 & 0.1024 & 0.0920 & 0.0566 \\ 0 & 1 & 0.8985 & 0.7134 \\ 0 & 0 & 1 & 0.7945 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

运行算法2后，我们得到了以下关于 T的估计

$$\hat{T} = \begin{bmatrix} 0 & 0.1024 & 0 & 0 \\ 0 & 0 & 0.8985 & 0 \\ 0 & 0 & 0 & 0.7945 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$