arXiv:2302.02228v2 [stat.ML] 6 Jun 2023

双射因果模型的反事实可识别性

Arash Nasr-Esfahany [1] **Mohammad Alizadeh** [1] **Devavrat Shah** [1]

摘要

我们研究具有双射生成机制（BGM）的因果模型中的反事实可识别性。这类模型推广了文献中几种广泛使用的因果模型。我们证明了在三种常见的、存在未观测混杂的因果结构中，这类模型具有反事实可识别性，并提出一种实用的学习方法，将 BGM 的学习转化为结构化生成建模。学习得到的 BGM 能够实现高效的反事实估计，并可通过多种深度条件生成模型获得。我们在一项视觉任务中评估了所提出的技术，并展示了其在现实世界视频流模拟任务中的应用。

1. 引言

如果克娄巴特拉的鼻子短一些，整个世界的面貌都会改变。[2]（*Blaise Pascal，1669*年）

Pearl（因果阶梯（2018年）由三个不同层次构成，分别表示不同类型的概念：关联层（$\mathcal{L}_1$）、干预层（$\mathcal{L}_2$）以及反事实（$\mathcal{L}_3$），大致分别对应于观察、行动和想象。$\mathcal{L}_1$ 涉及被动观察到的事实信息，例如服用阿司匹林的患者康复的概率。$\mathcal{L}_2$ 涉及主动干预或行动的效果，例如，如果给患者服用阿司匹林，有多大比例的患者会康复？$\mathcal{L}_3$ 涉及世界可能存在的其他状态，包括可能与世界当前状态相矛盾的情形。例如，如果患者服用了阿司匹林并康复了，那么假如他们没有服用阿司匹林，头痛还会消失吗？这三个层次之所以不同，是因为通常无法仅凭较低层次的信息，唯一确定较高层次的问题的答案。

[1]麻省理工学院（MIT）电气工程与计算机科学系（EECS），美国马萨诸塞州剑桥市，邮编 02139。联系邮箱：Arash Nasr-Esfahany<arashne@mit.edu>。

发表于 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年，作者所有。[2]摘自Pearl 与 Mackenzie（2018年，第 8 章）

（Bareinboim 等人，2022年，定理 27）。特别是，尽管我们可以通过开展实验并主动干预世界来确定 $\mathcal{L}_2$ 信息，但通常无法回答 $\mathcal{L}_3$ 问题，即使进行实验也不例外。换言之，这些问题可能是不可识别（Ibeling 和 Icard，2020年），或者是欠规范的（D'Amour 等人，2020年）。

尽管如此，“为什么——？”和“如果——会怎样？”这类反事实查询有助于界定伤害（Richens 等人，2022年）、公平性（Kusner 等人，2017年；Zhang 与 Bareinboim，2018年）、归因（Mesnard 等人，2021年）、遗憾和责任归咎等基本概念。它们在工程领域也有应用，例如根因分析（Gan 等人，2022年；Budhathoki 等人，2022年）、轨迹驱动仿真（Alomar 等，2023；Bothra 等人，2022年），以及样本高效的强化学习（Lu 等人，2020年；Agarwal 等人，2021年）。

本文介绍一类新的因果模型，称为双射生成机制（*BGM*）。BGM 涵盖了文献中研究的若干模型类别，例如非线性加性噪声模型（ANM）（Hoyer 等人，2008年）、位置尺度噪声模型（LSNM）（Immer 等人，2022年）以及后非线性因果模型（PNL）（Zhang& Hyv¨arinen，2009年）（§4）。假设我们对底层系统有粗略的了解，即已知其因果图，我们证明了在三种有意义的因果结构下，BGM 具有反事实可识别性（§5）。

我们的可识别性结果给出了易于操作的判据，用于寻找底层 BGM （在等价意义下，见定义 6.1 ），从而能够进行反事实估计。我们将寻找满足指定判据的 BGM 转化为使用结构化生成网络进行密度估计（§6）；这在机器学习文献中已得到广泛研究（Kingma &Welling，2013年；Goodfellow et al.，2014年；Dinh etal.，2016年；Arjovsky et al.，2017年；Song & Ermon，2019年；Ho et al.，2020年）。

一旦学得底层的双射生成模型，我们便利用它进行高效反事实估计，并提供相应保证（§7）。这与符号反事实识别方法（Shpitser 与 Pearl，2007年；Correa 等人，2021年）形成对比；这类方法用干预层和观测分布来表示可识别的反事实查询，但这些表示未必具有计算可行性。此外，我们的工作推动了将机器学习方法应用于反事实估计这一不断发展的研究领域。不同于该领域大多数以往工作

不进行可识别性分析（Pawlowski 等人，2020年；Sanchez 与 Tsaftaris，2021年），仅在某些受限情形下例外（例如离散域（Shalit 等人，2017年；Oberst 与 Sontag，2019年）或不存在未观测混杂因素（Khemakhem 等人，2021年；Sanchez-Martin 等人，2021年）），我们的结果支持连续变量，允许存在未观测混杂因素，并推广到更一般的生成机制类别。

我们使用视觉任务（§8.1）验证我们的方法，并展示其在现实世界系统模拟任务（§8.2）中的应用。我们的代码可在github.com/counterfactual-BGM/cf2获取。本工作不会带来任何负面的社会影响。

2. 预备知识

我们简要介绍本文所使用的（主要是）因果概念。有关更多细节，请参阅 Pearl（2009a）、Pearl（2009b）以及 Peters 等人（2017年）。

记号：我们用大写字母表示随机变量，例如 X；用小写字母表示其取得的值，例如 x；用 P_X表示其概率密度函数（PDF）；用粗体表示一组随机变量，例如 $\boldsymbol{X}=\{X_1,\dots,X_n\}$。 $J_{f(\cdot)}$ 表示 $f(\cdot)$ 的雅可比矩阵。

结构因果模型与因果图：我们采用结构因果模型（SCM）的框架（Pearl，2009a，第7章）。一个SCM $\mathcal{M}$ 由内生变量 $\boldsymbol{V}$（由模型中的其他变量决定）、外生变量（也称潜变量或背景变量） $\boldsymbol{U}$ （其分布 $P_{\boldsymbol{U}}(\cdot)$ 由模型外部的因素决定；每个内生变量对应一个外生变量）以及生成机制 $\mathcal{F}$组成。 $\mathcal{F}$ 为每个变量 V_i 包含一个函数 f_i，该函数将内生父节点 $\text{Pa}(V_i)$ 和外生变量 U_i 映射到 V_i。整个 $\mathcal{F}$ 定义了从 $\boldsymbol{U}$ 到 $\boldsymbol{V}$的结构化映射。外生变量的先验分布 $P_{\boldsymbol{U}}(\cdot)$与生成机制（$\mathcal{F}$）共同决定内生变量的观测（$\mathcal{L}_1$）分布，我们将其记为 $P_{\boldsymbol{V}}^{\mathcal{M}}(\cdot)$。每个 $\mathcal{M}$ 都会诱导出一个因果图 $\mathcal{G}$，其中每个 $V_i \in \boldsymbol{V}$都是一个顶点；对于每个$V_i \in \boldsymbol{V}$，若 $V_j \in \text{Pa}(V_i)$，则存在一条有向箭头（$V_j \to V_i$）；对于每对 V_i、$V_j \in \boldsymbol{V}$，若 U_i与 U_j不独立，则存在一条虚线双向箭头（$V_j \dashleftarrow\!\dashrightarrow V_i$）。双向箭头表示存在未观测混杂因素。若对于每个 V_i、$V_j \in \boldsymbol{V}$，相应的 U_i 与 U_j 相互独立，则该因果模型满足马尔可夫假设。换言之，其诱导的因果图（$\mathcal{G}$）中不存在双向箭头。

干预与 *do*-算子：给定一个结构因果模型 $\mathcal{M}$以及一组内生变量 $\boldsymbol{X} \subseteq \boldsymbol{V}$，干预$do(\boldsymbol{X}:=\boldsymbol{x})$ 对应于将生成机制 $\mathcal{F}$ 替换为 $\mathcal{F}_{\boldsymbol{x}}=\{f_i\colon V_i \notin \boldsymbol{X}\}\cup\{\boldsymbol{X}:=\boldsymbol{x}\}$。换言之，$\mathcal{F}_{\boldsymbol{x}}$ 是通过从 $\mathcal{F}$ 中删除所有函数而形成的 f_i 这些函数对应于集合 $\boldsymbol{X}$ 中的成员，并将其替换为常函数集合 $\{\boldsymbol{X}:=\boldsymbol{x}\}$。我们将改变后的结构因果模型记为 $\mathcal{M}_{\boldsymbol{x}}$，将内生变量的干预层（$\mathcal{L}_2$）分布记为 $P_{\boldsymbol{V}}^{\mathcal{M}_{\boldsymbol{x}}}(\cdot)$，或简写为 $P_{\boldsymbol{V}_{\boldsymbol{x}}}(\cdot)$。干预分布（或实验分布）有助于我们分析采取行动的效果，即如果我们实施某项干预，会发生什么？

反事实：给定一个结构因果模型 $\mathcal{M}$、两组内生变量 $\boldsymbol{X}$、$\boldsymbol{E}\subseteq\boldsymbol{V}$、 $\boldsymbol{E}$的观测实现值 $\boldsymbol{e}$ （证据），以及一个干预 $\text{do}(\boldsymbol{X}:=\boldsymbol{x})$，反事实（$\mathcal{L}_3$）分布 $P_{\boldsymbol{V}}^{\mathcal{M}_{\boldsymbol{x}}}(\cdot|\boldsymbol{E}=\boldsymbol{e})$，简写为$P_{\boldsymbol{V}_{\boldsymbol{x}}}(\cdot|\boldsymbol{E}=\boldsymbol{e})$，对应于使用外生变量的后验分布 $P_{\boldsymbol{U}|\boldsymbol{E}}(\cdot|\boldsymbol{e})$，由$\mathcal{M}_{\boldsymbol{x}}$ 蕴含的分布。对于确定性反事实（δ分布），我们称之为 $\boldsymbol{V}_{\boldsymbol{x}}|\boldsymbol{E}=\boldsymbol{e}$。Pearl（2009a，第 7 章）提出了以下用于反事实估计的三步程序。i）溯因：利用 $\boldsymbol{e}$ 更新 $P_{\boldsymbol{U}}$，以获得 $P_{\boldsymbol{U}|\boldsymbol{E}}$。ii）行动：将结构因果模型 $\mathcal{M}$更新为 $\mathcal{M}_{\boldsymbol{x}}$。iii）预测：使用更新后的外生变量分布和更新后的结构因果模型来估计反事实分布。有关此定义的推广，例如嵌套反事实，请参见 Correa 等人（2021年）。反事实分布使我们能够设想这样的假想世界：除干预外，一切均保持不变。

可识别性：在知识状态不完整的情况下评估因果查询，是一个微妙的问题。若某个 $\mathcal{L}_i$ 查询的答案可以完全基于 $\mathcal{L}_j$分布（i、$j\in\{1,2,3\}$）得出，则该查询可利用 $\mathcal{L}_j$ 信息识别。形式上，设 $Q(\mathcal{M})$ 是结构因果模型 $\mathcal{M}$ 的一个 $\mathcal{L}_i$ 查询。在结构因果模型类 $\mathbb{M}$ 中，若对于来自 $\mathbb{M}$ 的任意一对结构因果模型 $\mathcal{M}_1$ 和 $\mathcal{M}_2$，只要它们在所有 $\mathcal{L}_j$ 查询上都一致（Pearl，2009a，定义 3.2.3），就有 $Q(\mathcal{M}_1)=Q(\mathcal{M}_2)$，则称 Q 可识别。这在 $1\le i\le j\le 3$ 时总是成立。针对 $1\le j<i\le 3$，刻画这一条件何时成立，一直是广泛研究的课题（Spirtes 等人，2001年）。

3. 相关工作

干预层 ($\mathcal{L}_2$) 识别与估计：假设我们对底层系统有粗略的了解，即掌握其因果有向无环图，那么文献已对干预查询的识别进行了广泛研究（Pearl，2009a年，第 3 章），其中包括将*do*-演算 (Pearl，1995a) 作为一种通用解决方案。在无法获取完整因果图的情况下，另一类研究则聚焦于利用观测数据对其进行（部分）识别。尽管这听起来颇具吸引力，也适合数据驱动和机器学习的研究实践，但众所周知，因果图只能根据观测数据识别到其马尔可夫等价类（Spirtes 等人，2001）。因此，利用等价类研究可识别性也受到关注（Zhang，2008年；Perkovi´c 等人，2018年；Jaber 等人，

2019年）。Witty 等人（2021年）利用结构因果模型参数化表示上的先验分布，借助概率程序（Goodman 等人，2008年）研究了一种基于模拟的可识别性概念。一旦确认干预查询具有可识别性，就可以采用多种方法估计因果效应，包括用于后门情形的倾向评分（Rubin，1978；Kennedy，2019年），以及适用于更宽松情形的其他统计方法（Fulcher 等人，2020年；Jung 等人，2020年）。如果干预查询不可识别，则可使用部分识别方法估计有意义的界（Manski，1990；Balke 与 Pearl，1997；Zhang 与 Bareinboim，2021年；Li 与 Pearl，2022a年）

反事实 ($\mathcal{L}_3$) **识别**：使用反事实图，Shpitser 和 Pearl (2007年) 提出了一种算法，用于根据干预数据确定反事实查询的可识别性。Correa 等人(2021年) 在此基础上进一步扩展，针对给定的因果图，提供了从一组观察分布和干预分布中识别（嵌套）反事实的充分且必要的图形条件。在可识别的情况下，这些算法将反事实查询表示为观察量和干预量的组合。然而，对这一表达式进行估计可能并不可行。 Shah 等人 (2022年) 证明了特定因果图中反事实的可识别性，并提供了一种可行的估计算法，其假设是外生变量和内生变量的联合分布属于指数族。与之相反，我们不对变量的联合分布作任何限制。

与干预情形类似，针对不可识别的反事实查询，人们提出了部分识别方法（Balke & Pearl, 1994; Tian & Pearl, 2000; Finkelstein & Shpitser, 2020; Zhang 等人, 2022; Gresele 等人, 2022）。此外，人们还单独研究了特定反事实查询的识别，例如，对接受治疗者的治疗效应（Shpitser & Pearl, 2009）、路径特定效应（Shpitser & Sherman, 2018; Zhang & Bareinboim, 2018）以及因果概率（Pearl, 1999）。

因果估计的神经方法[3]：将神经方法用于高效估计因果查询的研究日益增多。一系列广泛的工作致力于估计干预层($\mathcal{L}_2$) 查询，其中包括Kocaoglu 等人(2021年); Xia 等人(2021年);Ze˘cevi´c 等人(2021)，仅举几例。然而，我们关注的是反事实($\mathcal{L}_3$)的估计。

在反事实方面，一系列研究采用机器学习技术进行估计，但并不保证反事实查询具有可识别性（Pawlowski 等人，2020年；Khemakhem 等人，2021年；Sanchez-Martin 等人，2021年；Sanchez 与 Tsaftaris，2021年；Geffner 等人，2022年）。与之相反，我们在§5。此外，现有的大多数技术都对因果结构或生成机制作出了限制性假设。例如，Johansson 等人（2016）；Shalit 等人（2017）；Yao 等人（2018）；Oberst 与 Sontag（2019）；Lorberbom 等人（2021）；Xia 等人（2022）研究了离散型（有限取值）内生变量领域中的反事实估计。其他一些研究假设因果结构具有马尔可夫性（Pawlowski 等人，2020年；Khemakhem 等人（2021）；Sanchez-Martin 等人，2021年；Geffner 等人，2022年），即不存在未观测混杂因素（因果图中的双向边）。Hartford 等人（2017）；Khemakhem 等人（2021）；Geffner 等人（2022）将生成机制限制为非线性加性噪声模型（ANM）（Hoyer 等人，2008年）或位置尺度噪声模型（LSNM）（Strobl 与 Lasko，2022）。我们考虑的模型比这些先前研究更为一般。它们支持连续内生变量，允许存在未观测混杂因素，并允许采用双射生成机制；非线性加性噪声模型、位置尺度噪声模型以及后非线性因果模型（PNL）（Zhang 与 Hyv¨arinen，2009年）均是其特例。

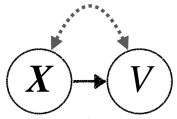
解耦：独立成分分析（ICA）（Hyv¨arinen 和 Oja，2000）研究如何从观测混合信号 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 中恢复统计独立源信号$\boldsymbol{S} = (S_1, \ldots, S_n)$，其中 $\boldsymbol{X} = f(\boldsymbol{S})$，并假设未知的 f（混合函数）是可逆的。与在线性混合函数下满足某些条件即可使该问题具有可识别性的情况不同，已知非线性混合函数下该问题不可识别（Hyv¨arinen 和 Pajunen，1999年）。近期，Hyvarinen 等人（2019年）证明，在存在辅助变量且这些变量使源信号条件独立的情况下，该问题具有可识别性。这一结果已用于从高维数据中解耦语义有意义特征（Locatello 等人，2019年；Khemakhem 等人，2020年）。

我们的研究受到非线性 ICA 框架的启发，尤其是 §5.3 节中的研究。不过，我们所考虑的问题有着根本不同：我们关注的是从已知属性中分离出未知变异来源的总贡献，而不是分离各个未知来源的影响。近期一项研究（Shaham 等人，2022年）探讨了如何从观测属性中分离潜变量，并证明重ˆ构的来源 $\boldsymbol{S}$ 与真实来源具有相同的熵分布。然而，我们关注的是逐点保证，即 ˆ $\boldsymbol{s} = g(\boldsymbol{s})$ 对于某个可逆函数$g(\cdot)$，这比诸如ˆ这样的分布性质更强 $\mathcal{H}(\boldsymbol{S}) = \mathcal{H}(\boldsymbol{S})$。

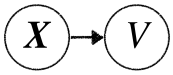
4. 问题表述

我们假设已知因果图 $\mathcal{G}$，其中可能包含未观测混杂因素（不假设马尔可夫性），并且可以获取观测数据分布 ($\mathcal{L}_1$)。我们

[3]附录A 提供了更详细的综述。

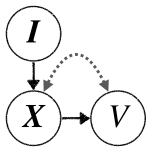


(a) 非马尔可夫

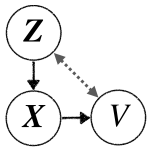


(b) 马尔可夫

图1. 与生成 V 相关的 $\mathcal{G}$ 因果图子图



（a）工具变量



(b) 后门准则

图2。 工具变量和后门准则示例。

希望学习数据生成机制 $\mathcal{F}$，以便进一步用于反事实（$\mathcal{L}_3$）估计。Pearl 的因果层级定理（Bareinboim 等人，2022年，定理 27）表明，在一般情况下无法进行跨层推断，因此我们需要作出进一步假设。

双射生成机制（BGM）：如 §2中所述，每个内生变量 V_i 在结构因果模型框架中按以下方式生成：

$$V_i \coloneqq f_i\Big(\mathrm{Pa}(V_i), U_i\Big) \tag{1}$$

我们假设，对于 $\mathrm{Pa}(V_i)$ 的每种取值，函数 f_i 都是从 U_i 到 V_i的双射映射，因此称为双射生成机制（*BGM*）。换句话说，从外生变量到内生变量的转换不会丢失任何信息。我们将生成机制的逆函数记为 $f_i^{-1}(\mathrm{Pa}_i, \cdot)$，即，

$$U_i = f_i^{-1}\Big(\mathrm{Pa}(V_i), V_i\Big). \tag{2}$$

请注意，先前工作中的非线性加性噪声模型（Peters 等人，2014年）、LSNM（Immer 等人，2022年）和 PNL（Zhang&Hyv¨arinen，2009年）模型都是双射生成模型的特例。具体来说，非线性加性噪声模型，

$$f_i\Big(\mathrm{Pa}(V_i), U_i\Big) = g_i\Big(\mathrm{Pa}(V_i)\Big) + U_i \tag{3}$$

LSNM（Immer 等人，2022年）模型，

$$f_i\Big(\mathrm{Pa}(V_i), U_i\Big) = l\Big(\mathrm{Pa}(V_i)\Big) + s\Big(\mathrm{Pa}(V_i)\Big)U_i \tag{4}$$

具有 s 一个严格正函数 $\mathbb{R}$，以及 PNL 因果模型（Zhang&Hyv¨arinen，2009年），

$$f_i\Big(\mathrm{Pa}(V_i), U_i\Big) = h_i\Big(g_i\Big(\mathrm{Pa}(V_i)\Big) + U_i\Big) \tag{5}$$

具有 h_i 一个可逆函数，在给定 $\mathrm{Pa}(V_i)$ 的条件下，都是双射的。

5. 可识别性

不失一般性，我们聚焦于识别特定内生节点 f_i 的生成机制 V_i，该机制如方程（1）所示。为便于说明，我们省略下标 i，并将 $\mathrm{Pa}(V)$ 记为 $\boldsymbol{X}$。与 $\mathcal{G}$ 的生成相关的子图 V 如图 1a所示。虚线双向箭头表示可能存在未观测混杂因素。所有证明均见附录 B。

我们给出了关于 f 及其底层因果结构的三组约束，在给定观测到的（$\mathcal{L}_1$）数据分布（$\mathcal{D}$）时，这些约束意味着反事实可识别性；也就是说，给定 $\mathcal{D}$，每组约束都能识别 f，但不影响反事实查询的不确定性除外。

U 和 V 通常可以是单维或多维的。§5.1 和 §5.2 中的结果针对标量情形，而 §5.3 中的结果适用于多维 U 和 V。

5.1. 马尔可夫情形

如果与 U 相关联的 V 外生变量独立于其父节点 $\boldsymbol{X}$，我们就得到图1b所示的因果图，其中 $\boldsymbol{X} \perp\!\!\!\perp U$。[4]

定理5.1。 若满足$P_{\boldsymbol{X},\ V}$，则双射生成模型 f 反事实可识别，当且仅当

1. （马尔可夫） $U \perp\!\!\!\perp \boldsymbol{X}$。
2. 对所有 $\boldsymbol{x}$，$f(\boldsymbol{x}, \cdot)$ 要么是严格递增函数，要么是严格递减函数。

该定理与 Lu 等人（2020年，定理1）类似，此处提及是为了完整性。在此结果中，U，$\boldsymbol{X}$，V可以是离散型或连续型。请注意，仅凭U 与 $\boldsymbol{X}$ 相互独立，并不足以保证可识别性；附录B.3.1中的反例和附录E.1.1中的实验均对此作了说明。如何将单调性条件推广到具有多维的贝叶斯图模型，目前尚不明确；这也是马尔可夫因果结构中一个已知的问题（Nasr-Esfahany 与 Kiciman，2023年）。

5.2. 工具变量（IV）

即使存在未观测混杂因素，只要我们能找到相对于变量对（$\boldsymbol{X}$，V）的工具变量，就可以根据观测（$\mathcal{L}_1$）数据建立反事实可识别性。我们将工具变量定义为

[4]通常会从因果图中省略外生变量 U。

一组独立于 U、且仅通过 $\boldsymbol{X}$影响 V的变量。图 2a展示了一个示例，其中 $\boldsymbol{I}$ 是相对于 $(\boldsymbol{X},\ V)$的工具变量。[5]

以下定理形式化了在此设置下，对于离散取值的 $\boldsymbol{X}$ 和 $\boldsymbol{I}$，即$\boldsymbol{X} \in \mathbb{X} \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$和 $\boldsymbol{I} \in \mathbb{I} \triangleq \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n\}$的反事实可识别性。[6]

定理 5.2. 若满足$P_{\boldsymbol{X}}$、V、$\boldsymbol{I}$，则双射生成模型 f 是反事实可识别的，条件是

1.（工具变量）
$\boldsymbol{I} \perp\!\!\!\perp U$。

2. 对所有 $\boldsymbol{x} \in \mathbb{X}$，$f(\boldsymbol{x},\ \cdot)$ 和 $f^{-1}(\boldsymbol{x},\ \cdot)$ 都是严格递增或严格递减且二阶可微的函数。

3. 对每个 $\boldsymbol{i} \in \mathbb{I}$和 $\boldsymbol{x} \in \mathbb{X}$，$P(\boldsymbol{i},\ \boldsymbol{x},\ \cdot)$ 都是可微的。

4. *(Positivity)* $\forall u, \boldsymbol{x} \in \mathbb{X} : P_{U,\boldsymbol{X}}(u, \boldsymbol{x}) > 0$.

5.（变异性）$\forall u$：$|\det \boldsymbol{M}(u, \mathbb{I})| \geq c > 0$，其中

$$\boldsymbol{M}(u, \mathbb{I}) \triangleq \begin{bmatrix} P(\boldsymbol{x}_1|u,\boldsymbol{i}_1) & \ldots & P(\boldsymbol{x}_n|u,\boldsymbol{i}_1) \\ \vdots & \ddots & \vdots \\ P(\boldsymbol{x}_1|u,\boldsymbol{i}_n) & \ldots & P(\boldsymbol{x}_n|u,\boldsymbol{i}_n) \end{bmatrix}$$

除了证明所需的技术条件外，定理 5.2 中的变异性条件可作如下解释：工具变量必须取足够多的不同值（至少与 $\boldsymbol{X}$可能取值数量一样多），并且工具变量必须对 $\boldsymbol{X}$有很强的影响。正性意味着 U 的支撑集与 $\boldsymbol{X}$ 无关。

5.3. 后门准则（BC）

在存在潜在混杂的情况下，我们可以从观察性（$\mathcal{L}_1$）数据确立反事实可识别性的第二种情形是：存在一组变量（$\boldsymbol{Z}$），它们相对于（$\boldsymbol{X}$，V）满足后门准则（BC），即若 $\boldsymbol{Z}$ 阻断了 $\boldsymbol{X}$ 与 V之间每一条包含指向 $\boldsymbol{X}$的箭头的路径。[7] 直观来说，我们希望 $\boldsymbol{Z}$ 能够解释 $\boldsymbol{X}$ 与 U之间所有的虚假相关（虚线双向边）。图 2b 展示了一个例子，其中 $\boldsymbol{Z}$ 相对于（$\boldsymbol{X}$，V）满足 BC。

在下文中，假设 U，$V \in \mathbb{R}^d$。

定理 5.3。 若满足$P_{\boldsymbol{X}}$、V、$\boldsymbol{Z}$，则 *BGM* f 反事实可识别

1.（后门准则）$U \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{Z}$。

2. $\forall \boldsymbol{x} : \nabla_{\boldsymbol{x}} |\det \boldsymbol{J}_{f(\boldsymbol{x},\cdot)}|$ *and* $\nabla_{\boldsymbol{x}} |\det \boldsymbol{J}_{f^{-1}(\boldsymbol{x},\cdot)}|$ *exist.*

3.（变异性）$\forall u$：存在实例 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}$，使得 $|\det \boldsymbol{M}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1})| > 0$，其中

$$\boldsymbol{M}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}) \triangleq \begin{bmatrix} P(u|\boldsymbol{z}_1) & \nabla_u P(u|\boldsymbol{z}_1) \\ \vdots & \vdots \\ P(u|\boldsymbol{z}_{d+1}) & \nabla_u P(u|\boldsymbol{z}_{d+1}) \end{bmatrix}$$

在上述定理中，$\boldsymbol{Z}$ 可以是离散型或连续型。变异性条件意味着**Z** 必须对 U产生足够强的影响。

6. BGM 的高效学习

给定一个观测数据分布（$\mathcal{D}$），本节的目标是找到一个与数据背后的 BGM f 反事实等价的 BGM f。首先，我们形式化定义等价的概念：

定义 6.1。 *(*等价*)* BGM f_1 和 f_2 等价，当且仅当存在一个可逆函数 $g(\cdot)$，使得

$$\forall \boldsymbol{x}, v : f_1^{-1}(\boldsymbol{x}, v) = g\Big(f_2^{-1}(\boldsymbol{x}, v)\Big) \text{ or alternatively} \quad (6)$$

$$\forall \boldsymbol{x}, u : f_1(\boldsymbol{x}, u) = f_2\Big(\boldsymbol{x}, g^{-1}(u)\Big). \quad (7)$$

命题 6.2。 *BGM* f_1 和 f_2 产生相同的反事实，当且仅当它们等价。

ˆ如果我们找到一个 f 与数据背后的 BGM f 等价的模型，就可以使用 f 执行溯因-行动-预测（§2），从而估计任何反事实量。根据命题 6.2，该模型保证能产生与真实 BGM 相同的反事实。

§5 中的定理（另见附录 B.2中的引理）给出了一个易于处理的目标函数，用于学习与真实 BGM 等价的 BGM，以进行反事实估计。现在，我们将介绍一种高效求解这一学习问题的方法。

ˆ为了表示 f 的搜索空间，我们使用参数化的 ˆ ˆ 双射变换 $f_\theta(\boldsymbol{x},\ \cdot)$ 将 U 映射到 V，条件为$\boldsymbol{X}$。这类似于条件生成模型（CGM）中使用的变换。此外，根据具体情形，我们要求 ˆ在 f_θ 上满足某些约束，例如严格单调或可微；在选择适当的模型族和参数化方式时，我们会将这些要求纳入考虑。例如，IV 和后门准则情形都要求变换可微。条件归一化流（CNF）（Papamakarios 等人，2021年）模型非常适合学习此类函数，因为它们通常采用具有可微逆映射的可微映射（微分同胚）构建。尽管我们的方法适用于

[5]因果文献中还有其他几种工具变量定义，表达的是同一概念，例如：Bowden 等人（1990年）；Pearl（1995b）；Angrist 等人（1996）；Heckman 和 Vytlacil（1999年）。关于此问题的讨论，见 Pearl（2009a，第 7.4.5 节）。[6]如果 $|\mathbb{I}| > n$，任选一个包含 n 个成员的子集。[7]Pearl（2009a，定义 3.3.1）还要求 $\boldsymbol{Z}$ 是 $\boldsymbol{X}$ 的非后代节点，但我们不需要这一要求，因为我们的目标不是用 $\boldsymbol{Z}$ 进行调整。

具有所需属性的任何条件生成模型，我们在实验中使用条件归一化流。附录D对条件归一化流进行了更详细的讨论。

每组约束都包含一项重要的分布要求（第一项条件），其形式为（条件）ˆ变量之间的独立性 U（方程（2））以及与其他变量之间的独立性。为了引导学习过程实现此类条件独立性，一种方法是使用对抗学习进行分布匹配（Li 等人，2017年）。不过，换个角度看待这个问题，可以得到一种更优雅的解决方案。我们不再将（$\mathbf{X}$ V）$\in \mathcal{D}$样本输入ˆf_θ^{-1}（方程（2）），并优化变换以满足（条件）ˆ独立性，而是可以按照每种情况所需的（条件）独立性来采样 U，并优化ˆ f_θ（方程（1）），以生成观测到的条件分布 $P_{\mathcal{D}}(V|\boldsymbol{X})$。这有两点好处：i）（条件）独立性通过构造即可得到保证。ii）训练目标简单地转化为密度建模，而这方面已在文献中得到深入研究。

§5.1中的马尔可夫情形提供了一个简单示例来说明学习方法。我们将严格递增的ˆ条件变换用作 f_θ的搜索空间。我们ˆ分别从高斯分布和数据集 $\mathcal{D}$中独立采样 U 和 $\boldsymbol{X}$。这保证了定理5.1的第一个条件。我们优化 f_θ 的参数ˆ，使其匹配条件分布 $P_{\mathcal{D}}(V|\boldsymbol{X})$。为此，我们可以自由选择生成建模领域多年研究所提供的各种目标函数（Mohamed & Lakshminarayanan, 2016），例如变分损失（Kingma & Welling, 2013）、对抗损失（Goodfellow et al., 2014）、似然最大化（Pa-pamakarios et al., 2021）、分数匹配（Song & Ermon, 2019）或去噪扩散（Ho et al., 2020）等。

另一方面，在**工具变量**(§5.2)和**后门准则**(§5.3)的情况下，乍看之下，训练并不那么简单。我们不再能ˆ独立地采样 U 和 $\boldsymbol{X}$，因为它们不一定相互独立。此外，我们还需要满足各个情形下的条件独立性约束。

为高效编码条件独立性约束，我们使用有向图模型（DGM）（Koller 与 Friedman，2009年）来构建结构化生成网络，这些网络以 CGM 为构建模块，并继承 DGM 中编码的所有条件独立性。[8]图 3展示了用于满足工具变量和后门准则情形下条件独立性的 DGM，以及根据这些 DGM 构建的结构化生成模型。请注意，图中所示的 DGM 和结构化生成网络并非唯一满足所需条件独立性属性的方案。在后门准则情形下，存在若干种不同的 DGM，它们编码了相同的条件

[8]这里，我们将深度生成模型视为纯粹的统计对象，不赋予其任何因果语义。

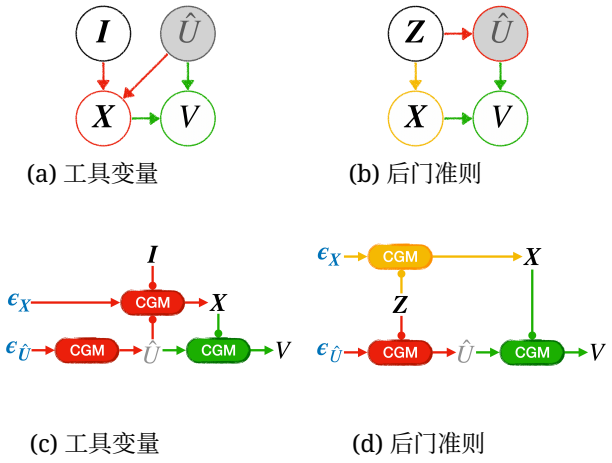


图 3. 图模型（上）编码所需的（条件）独立性。用于学习贝叶斯图模型的结构化生成网络（下）是根据相应的图模型构建的。蓝色变量从高斯分布中独立采样，黑色变量则是在数据集中观测到的。

独立性，且每一种都定义了一种不同的结构化生成网络，可用于学习ˆ f（见附录C）。ˆ还需注意，用于表示 f_θ 的 CGM（绿色的那个）有单调性和可微性等要求，而其他模型没有这些限制。

训练： 我们独立地对根变量进行采样：$\boldsymbol{\epsilon_X}$，$\epsilon_{\hat{U}}$，$\boldsymbol{I}$ 见图 3c；以及 $\boldsymbol{\epsilon_X}$，$\epsilon_{\hat{U}}$，$\boldsymbol{Z}$ 见图 3d。epsilon 变量（蓝色）从相互独立的高斯分布中采样，其余变量则从数据集 $\mathcal{D}$中采样。我们对整个结构进行端到端训练，目标是匹配 $\mathcal{D}$的分布。这个过程类似于对ˆ进行约束搜索 f，其中约束在构造时就嵌入了搜索空间。目标函数由所选的条件生成模型决定。我们甚至可以混合不同类型的条件生成模型，并结合不同的目标函数进行端到端优化。

训练的最终目标是学习绿色的条件生成模型。因此，任何训练不会影响绿色条件生成模型的条件生成模型都可以从网络中移除，以简化网络。图 3d 中的黄色条件生成模型就是这样一个例子，因为它被黑色变量包围，而这些变量是直接从 $\mathcal{D}$中采样的，因此会阻止其梯度反向传播并影响其他条件生成模型的训练。

7. 高效反事实估计

一旦我们使用第 §6 节中的技术学得一个双射生成模型，估计反事实查询就成了溯因-行动-预测程序（§2）的直接应用。我们用两个例子来说明。

假设我们关注一个反事实查询

$V_{\boldsymbol{x}'}|\{\boldsymbol{X}=\boldsymbol{x}, V=v\}$，它旨在确定因果关系的必要性。例如，已知一名患者在接受治疗 $X=1$ 后康复（$V=1$），如果没有接受治疗，她是否也会康复（$X=0$）。一般而言，仅凭观测数据（$\mathcal{L}_1$）和干预层数据（$\mathcal{L}_2$）无法识别这一查询（Pearl，1999年；Tian & Pearl，2000；Li & Pearl，2022b年）。然而，如果我们能将 V 的生成机制近似为双射生成模型（BGM），并且 §5 中的某个可识别性条件成立，那么就有希望。具体而言，我们首先从观测数据中学习双射生成模型（§6）。然后，通过求逆双射生成模型执行溯因步骤，即 $\hat{u}=\hat{f}^{-1}(\boldsymbol{x}, v)$，再将反事实 $\boldsymbol{x}'$ 输入双射生成模型，得到反事实估计，即 $v'=f(\boldsymbol{x}', \hat{u})$。

作为另一个例子，考虑 $P(V_{\boldsymbol{x}_2}|\boldsymbol{X}=\boldsymbol{x}_1)$，它探讨了对已接受治疗者的治疗效应（Shpitser& Pearl，2009年）。在溯因步骤中，我们对所有被分配到$\boldsymbol{x}_1$的观测样本（例如患者）反演学习得到的双射生成模型 $\hat{}(f)$，以获得外生后验分布 $P_{\cdot|\boldsymbol{x}_1}$ 的样本，即 $\hat{u}=\hat{f}^{-1}(\boldsymbol{x}_1,\ v)\ v\sim P(V|\boldsymbol{x}_1)$。在预测步骤中，我们将样本 $\hat{u}$输入带有$X=\boldsymbol{x}_2$的双射生成模型，从而获得感兴趣的反事实分布 $\hat{}\ v'=f(\boldsymbol{x}_2,\ \hat{u}),\ \hat{}$ $u\sim P_{\cdot|\boldsymbol{x}_1}$ 的样本。

上述示例仅考虑了因果图中特定节点 $V\in\boldsymbol{V}$的生成函数。如果某个结构因果模型（$\mathcal{F}$）的所有生成函数都能由双射生成模型很好地近似，并且也能在 §5所述的设置中学习得到，那么我们就能以类似方式回答每一个反事实查询。不过，回答某个特定的反事实查询，并不要求所有生成机制都是双射生成模型或已知。例如，对于我们进行干预的变量的祖先节点，无需学习其双射生成模型；对于既未出现在证据中，也没有指向证据的虚线双向边的变量，同样无需学习其双射生成模型。

8. 实验

我们在两种场景中评估所提出的技术：（i）一个简单任务，可用于直观展示我们的方法和先前的基线方法；（ii）现实世界视频流模拟任务。附录E 提供了实现细节。

8.1. 反事实椭圆生成

标准椭圆由两个参数确定：半长轴和半短轴（$a,\ b$）。如果进一步指定一个角度，就能得到椭圆上的一个点。设 $U\in\mathbb{R}^2$ 为椭圆的两个参数，$X\in(0,\ 2\pi)$为指定单个点的角度，$V\in\mathbb{R}^2$ 为该点的笛卡尔坐标，f 为根据 U 和 X计算这些坐标的函数。

假设我们获得了按如下方式生成的数据。我们对$z\sim P_Z$、$u\sim P_{U|Z=z}$ 和 $x\sim P_{X|Z=z}$进行采样，并输出数据元组（$z,\ x,\ v:=f(u,\ x)$）。这里，P_Z、$P_{U|Z}$和 $P_{X|Z}$

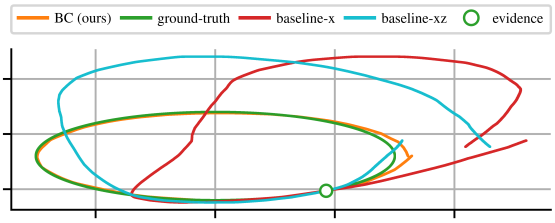


图 4。后门准则（我们的方法）是唯一能够生成椭圆的方案。

表 1. 反事实椭圆生成准确率

方案	BC	BASELINE-X	BASELINE-XZ
MAPE	1%	6607%	6582%

是三个预先定义的分布（详情见附录E.1 ）。重要的是，在给定 Z的条件下，椭圆参数和角度相互独立（$U\perp\!\!\!\perp X|Z$），但 U 和 X 在无条件时并不独立。对于在数据集中观测到的特定一对（$x,\ v$）（证据），我们的目标是绘制点 v 所属的完整椭圆。这可以通过估计反事实查询$V_{x'}|\{X=x, V=v\}$来实现，针对 $x'\in(0,\ 2\pi)$。

此任务对应于我们的 **后门准则** 设置，其中 Z 是后门变量。我们通过视觉方式（图4）和定量方式（表1）评估我们的方法在后门准则下的表现。如你所见，与基线相比，该方法取得了很高的准确率。两个基线都是单一的条件生成模型，没有考虑因果结构，这在先前的研究中很常见（Lample 等人，2017年； Zhu 等人，2017年； He 等人，2019年）。**基线-x** 模型$P(V|X)$和**基线-xz** 模型 $P(V|X,\ Z)$。

8.2. 案例研究：视频流模拟

视频流客户端使用自适应码率（ABR）算法，根据网络状况持续调整视频流的比特率。ABR 算法会从有限的可用选项中，为每个数秒长的视频数据块选择比特率。这些算法对用户体验（例如视频质量和卡顿）有重大影响，因此一直是广泛研究的主题（Tian 和 Liu，2012；Huang 等人，2014；Yin 等人，2015；Sun 等人，2016；Mao 等人，2017；Akhtar 等人，2018；Spiteri 等人，2020）。

轨迹驱动仿真是设计和评估自适应比特率算法的常用方法。在这种方法中，人们会收集真实视频流会话的轨迹，每条轨迹都记录了每个视频块的观测网络吞吐量时间序列（也可能包含其他播放器指标）。在仿真时，会重放这些轨迹（以表示网络行为），同时模拟视频客户端在新自适应比特率算法下的动态变化。然而，仅仅重放吞吐量轨迹可能会导致模拟结果产生偏差（Bartulovic 等人，2017年；

表 2。 视频流模拟中不同方案的反事实估计准确率及其训练时间

方案	归一化均方误差 (%)	时间 (s)
马尔可夫	4.1 ± 2.1	28 ± 0
IV	10.0 ± 0.7	127 ± 1
CausalSim	9.3 ± 1.0	1091 ± 1
BC	10.4 ± 7.8	37 ± 8
工具变量+后门准则	6.4 ± 2.4	49 ± 1

Alomar 等人，2023年），因为改变自适应比特率算法可能会影响同一视频流会话中原本能够达到的吞吐量。

Alomar 等人 (2023年) 将这类模拟中的偏差消除表述为一个反事实估计问题：对于每个数据块的下载，V 表示实现的吞吐量，X 表示所选比特率，U 表示未观测到的瓶颈链路容量，而 $V = f(U, X)$。请注意，在这个问题中，f 对于 X 的每个取值都严格递增，因为更高的瓶颈链路容量（U）会提高实现的吞吐量（V）。他们证明，若数据是在采用足够多样的 ABR 算法的随机对照试验（RCT）中收集的，且底层双射生成模型具有低秩结构，则该反事实查询具有可识别性。他们采用配备对抗学习的确定性自动编码器（Ghosh 等人，2019年），通过分布匹配进行反事实估计，并给出了实验结果（包括一项真实世界 ABR 设计案例研究），表明与标准的（有偏）轨迹驱动模拟器相比，该方法显著提高了模拟准确度。

接下来，我们将展示可识别性（§5）的每组条件如何对应到这一现实世界问题，并使用 6 中的实用算法，通过 Alomar 等人 (2023年) 提供的 ABR 模拟器验证其有效性。我们使用该模拟器为每种设置生成轨迹，并利用这些轨迹学习双射生成模型。该模拟器为我们提供了用于误差计算的真实反事实。我们将本方法的反事实估计值的均方误差（MSE）归一化为标准（有偏）视频流模拟器的均方误差；该模拟器假设所选择的比特率不会影响实际达到的吞吐量。所有结果见表 2，其中均值和标准差是在十个随机种子上训练至训练损失收敛后计算得到的。

马尔可夫情形：该问题中的因果结构不是马尔可夫的，因为 X（所选比特率）和 V（观测到的吞吐量）都受到潜在网络状况（潜在混杂因素）的影响。然而，如果用于收集轨迹的 ABR 算法不时随机选择比特率，并且我们只选取包含这些随机决策的轨迹子集作为 $\mathcal{D}$（图 1），就可以消除混杂效应。该方案实现了最低的表 2 中的误差，因为它能够获取最纯净的数据（通过侵入式随机化彻底消除混杂）。

工具变量（IV）：视频服务提供商通常会针对各种 ABR 算法开展随机对照试验，以便进行比较。如果轨迹采集自随机对照试验，其因果结构自然符合图 2a，其中工具变量（I）是算法标识符。这与 **CausalSim** (2023年) 所研究的设置相同，因此我们采用其对抗学习方法作为基线。我们针对十种不同的给定算法收集随机对照试验数据。我们的工具变量方法达到了与 CausalSim 几乎相同的准确率（轻微差异在统计上并不显著），但由于不需要双层（对抗式）优化，因此收敛 $8.6\times$ 更快。

后门准则（BC）：基于缓冲区的自适应比特率算法仅利用客户端当前的播放缓冲区水平来做出比特率决策。尽管简单，这类算法仍是强大的 ABR 算法，并在实践中得到广泛应用（Yan 等人，2020年）。如果轨迹是使用基于缓冲区的算法收集的（并且包含播放缓冲区观测值），其因果结构就遵循图 2b，其中 Z 是用于选择比特率的缓冲区。因此，它满足后门准则。表 2 显示，即使无法获取随机对照试验数据，将我们的 BC 方法应用于基于缓冲区的轨迹也相当准确（尽管其表现略逊于 CausalSim 和工具变量）。

工具变量 +BC：如果问题具有适当的因果结构，则可以结合 §5 中探讨的设置来进一步提高准确性。例如，在视频流模拟中，如果 $\mathcal{D}$ 是在针对基于缓冲区的算法进行的随机对照试验中收集的，则后门准则和工具变量情形会同时适用。我们将这种情形称为工具变量+BC。与工具变量情形类似，算法标识符是工具变量，而元组（算法标识符、缓冲区）满足后门准则。附录 E.2.1 介绍了该情形的结构化生成网络。我们将其应用于使用随机对照试验在两种基于缓冲区的算法下收集的轨迹。工具变量+BC 与 BC 方案相比，可将平均误差降低近 38%。

9. 总结

在本文中，我们定义了双射生成机制（BGM），这是一类包含文献中多种广泛使用的因果模型的模型（§4）。我们针对三种著名的因果结构确立了其反事实可识别性（§5），并提出了一种实用的学习方法，将双射生成机制的学习转化为结构化生成建模（§6）。我们在一项视觉任务中评估了该方法，并展示了其在现实世界视频流模拟任务中的应用（§8）。未来值得探索的方向包括对可识别性定理进行有限样本分析、扩展可识别的因果结构，以及将所提出的方法应用于各领域的现实问题，例如计量经济学、计算机系统、

ca果机器学习等。 ork.

致谢

我们感谢 Behrooz Tahmasebi 的讨论，这些讨论促成了定理 5.2的证明。本工作得到美国国家科学基金会（NSF）项目 1751009、CSAIL SystemsThatLearn（STL）计划的资助，以及谷歌、英特尔和亚马逊作为 MIT 数据系统与人工智能（DSAIL）实验室合作项目的一部分所提供的支持。

参考文献

Agarap, A. F. 使用修正线性单元（relu）进行深度学习。*arXiv*预印本 *arXiv:1803.08375*，2018年。

Agarwal, A.、Alomar, A.、Alumootil, V.、Shah, D.、Shen, D.、Xu, Z. 和 Yang, C. Persim：通过个性化模拟器，利用异构智能体实现数据高效的离线强化学习。神经信息处理系统进展，34:18564–18576，2021年。

Akhtar, Z., Nam, Y. S., Govindan, R., Rao, S., Chen, J., Katz-Bassett, E., Ribeiro, B., Zhan, J. 和 Zhang, H. Oboe：根据网络状况自动调优视频自适应码率算法。载于*2018*年 *ACM* 数据通信特别兴趣小组会议论文集，第 44–58 页，2018年。

Alomar, A., Hamadanian, P., Nasr-Esfahany, A., Agarwal, A., Alizadeh, M. 和 Shah, D. CausalSim：用于无偏轨迹驱动仿真的因果框架。载于第 *20* 届 *USENIX*网络系统设计与实现研讨会（*NSDI 23*），第 1115–1147 页，马萨诸塞州波士顿，2023年4月。USENIX协会。ISBN 978-1- 939133-33-5。网址：https://www.usenix.org/conference/nsdi23/presentation/alomar。

Angrist, J. D., Imbens, G. W., 和 Rubin, D. B. 使用工具变量识别因果效应。美国统计学会期刊，91(434): 444–455，1996年。

Arjovsky, M., Chintala, S., 和 Bottou, L. Wasserstein 生成对抗网络。载于国际机器学习会议，第 214–223 页。PMLR，2017年。

Balke, A. 和 Pearl, J. 反事实概率：计算方法、界与应用。载于不确定性会议论文集 *1994*，第 46–54 页。爱思唯尔，1994年。

Balke, A. 和 Pearl, J. 不完全依从研究中的治疗效应界。美国统计学会期刊，92(439):1171–1176，1997年。

Bareinboim, E.、Correa, J.D.、Ibeling, D. 和 Icard, T。关于 *Pearl* 层级与因果推断，第 507–556 页。美国计算机协会，纽约州纽约市，美国，第 1 版，2022年。ISBN 9781450395861。URLhttps://doi.org/10.1145/3501714.3501743。

Bartulovic, M.、Jiang, J.、Balakrishnan, S.、Sekar, V. 和 Sinopoli, B。数据驱动网络中的偏差，以及如何应对这些偏差。载于 第 *16* 届 *ACM* 网络热点研讨会论文集，第 192–198 页，2017年。Bingham, E.、Chen, J. P.、J ankowiak, M.、Obermeyer, F.、Pradhan, N.、Karaletsos, T.、Singh, R.、Szerlip, P.、Horsfall, P. 和 Goodman, N. D。Pyro：深度通用概率编程。《机器学习研究杂志》，2018年。

Bothra, C., Gao, J., Rao, S., 和 Ribeiro, B. Veritas：从视频流追踪数据中回答因果查询。*arXiv* 预印本 *arXiv:2208.12596*，2022年。

Bowden, R. J., Turkington, D. A., 等。工具变量。剑桥图书，1990年。

Budhathoki, K., Minorics, L., Bl¨obaum, P., 和 Janzing, D. 基于因果结构的异常值根因分析。发表于国际机器学习会议，第 2357–2369 页。PMLR，2022年。

Carroll, R. J., Ruppert, D., Stefanski, L. A., 和 Crainiceanu, C. M. 非线性模型中的测量误差：现代视角。Chapman and Hall/CRC，2006年。

Chen, R. T., Behrmann, J., Duvenaud, D. K., 和 Jacobsen, J.-H. 用于可逆生成建模的残差流。神经信息处理系统进展，32，2019年。

Correa, J., Lee, S., 和 Bareinboim, E. 基于任意替代实验的嵌套反事实识别。神经信息处理系统进展，34：6856–6867，2021年。

Dinh, L.、Sohl-Dickstein, J. 和 Bengio, S. 使用 real nvp 进行密度估计。*arXiv* 预印本 *arXiv:1605.08803*，2016年。

Dolatabadi, H. M.、Erfani, S. 和 Leckie, C. 使用线性有理样条进行可逆生成建模。载于人工智能与统计国际会议，第 4236–4246 页。PMLR，2020年。

Durkan, C.、Bekasov, A.、Murray, I. 和 Papamakarios, G. 神经样条流。神经信息处理系统进展，第 32 卷，2019年。

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., 等。《欠规范》提出了

现代机器学习可信度面临的挑战。机器学习研究期刊，2020年。

Finkelstein, N. 和 Shpitser, I. 利用反事实之间的逻辑关系推导界和不等式约束。载于人工智能不确定性会议，第 1348–1357 页。PMLR，2020年。

Fulcher, I. R.、Shpitser, I.、Marealle, S. 和 Tchetgen Tchetgen, E. J. 总体间接因果效应的稳健推断：广义前门标准。英国皇家统计学会杂志：*B*辑（统计方法论），82(1):199–214，2020年。

Gan, Y.、Liang, M.、Dev, S.、Lo, D. 和 Delimitrou, C. 借助无监督学习实现实用的云性能调试。*SIGOPS* 操作系统评论，56(1):34–41，2022年6月。ISSN 0163-5980。doi: 10. 1145/3544497.3544503。URLhttps://doi.org/10.1145/3544497.3544503。

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N. 等。深度端到端因果推断。*arXiv*预印本 *arXiv:2202.02195*，2022年。

Ghosh, P., Sajjadi, M. S., Vergari, A., Black, M. 和 Scholkopf, B. 从变分自编码器到确定性自编码器。载于国际学习表征会议，2019年。

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. 和 Bengio, Y. 生成对抗网络。载于*NIPS*，2014年。

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K. 和 Tenenbaum, J. B. Church：生成模型语言。载于第二十四届人工智能不确定性会议论文集，第220–229页，2008年。

Gresele, L., K¨ugelgen, J.V., K¨ubler, J., Kirschbaum, E.,Sch¨olkopf, B., and Janzing, D. 因果推断通过结构因果边际问题实现。载于 Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. 和 Sabato, S.（编），第*39*届国际机器学习会议论文集，第162卷，《机器学习研究论文集》，第7793–7824页。PMLR，2022年7月17–23日。网址:
h
t
tps://proceedings.mlr.press/v162/gresele22a.html。

Hartford, J., Lewis, G., Leyton-Brown, K., 和 Taddy, M. Deep iv：一种灵活的反事实预测方法。载于国际机器学习会议，第 1414–1423 页。PMLR，2017年。

He, Z., Zuo, W., Kan, M., Shan, S., 和 Chen, X. Attgan：只改变你想改变的部分，实现面部属性编辑。*IEEE*图像处理汇刊，28(11)：5464–5478，2019年。Heckman, J. J. 和 Vytlacil, E. J. 用于识别和界定治疗效应的局部工具变量与潜变量模型。美国国家科学院院刊，96(8)：4730–4734，1999年。Hendrycks, D. 和 Gimpel, K. 高斯误差线性单元（GELU）。*arXiv* 预印本 *arXiv:1606.08415*，2016年。

Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. 和 Sch¨olkopf, B. 基于加性噪声模型的非线性因果发现。神经信息处理系统进展，21，2008年。

Huang, T.-Y., Johari, R., McKeown, N., Trunnell, M. 和 Watson, M. 基于缓冲区的方法实现码率自适应：来自大型视频流媒体服务的证据。载于*2014*年 *ACM SIGCOMM* 会议论文集，第 187–198 页，2014年。

Hyv¨arinen, A. 和 Oja, E. 独立成分分析：算法与应用。神经网络，13：411–430，2000。

Hyv¨arinen, A. 和 Pajunen, P. 非线性独立成分分析：存在性与唯一性结果。神经网络，12(3)：429–439，1999年。

Hyvarinen, A.、Sasaki, H. 和 Turner, R. 使用辅助变量和广义对比学习的非线性独立成分分析。载于第*22*届人工智能与统计国际会议，第859–868页。PMLR，2019年。

Ibeling, D. 和 Icard, T. 因果层级中的概率推理。载于*AAAI*人工智能会议论文集，第34卷，第10170–10177页，2020年。

Immer, A.、Schultheiss, C.、Vogt, J. E.、Sch¨olkopf, B.、B¨uhlmann, P. 和 Marx, A. 关于因果位置-尺度噪声模型的可识别性与估计。*arXiv*预印本 *arXiv:2210.09054*，2022年。

Jaber, A.、Zhang, J. 和 Bareinboim, E. 马尔可夫等价下的因果识别：完备性结果。载于国际机器学习会议，第2981–2989页。PMLR，2019年。

Johansson, F.、Shalit, U. 和 Sontag, D. 学习用于反事实推断的表征。载于国际

机器学习会议，第 3020–3029 页。PMLR，2016年。

Jung, Y.、Tian, J. 和 Bareinboim, E. 通过加权经验风险最小化学习因果效应。神经信息处理系统进展，33：12697–12709，2020年。

Kennedy, E. H. 基于增量倾向评分干预的非参数因果效应。美国统计学会杂志，114(526)：645–656，2019年。

Khemakhem, I.、Kingma, D.、Monti, R. 和 Hyvarinen, A. 变分自编码器与非线性 ICA：一个统一框架。载于人工智能与统计国际会议，第 2207–2217 页。PMLR，2020年。

Khemakhem, I.、Monti, R.、Leech, R. 和 Hyvarinen, A. 因果自回归流。载于人工智能与统计国际会议，第 3520–3528 页。PMLR，2021年。

Kingma, D. P. 和 Ba, J. Adam：一种用于随机优化的方法。载于*ICLR*（海报），2015年。

Kingma, D. P. 和 Dhariwal, P. Glow：采用可逆1×1卷积的生成流。神经信息处理系统进展，31，2018年。

Kingma, D. P. 和 Welling, M. 自编码变分贝叶斯。*arXiv*预印本 *arXiv:1312.6114*，2013年。

Kobyzev, I.、Prince, S. J. 和 Brubaker, M. A. 归一化流：当前方法的介绍与综述。*IEEE*模式分析与机器智能汇刊，43(11):3964–3979，2020年。

Kocaoglu, M.、Snyder, C.、Dimakis, A. G. 和 Vishwanath, S. 因果GAN：通过对抗训练学习因果隐式生成模型。载于国际学习表征会议，2018年。

Koller, D. 和 Friedman, N. 概率图模型：原理与技术。麻省理工学院出版社，2009年。

Kuroki, M. 和 Pearl, J. 因果推断中的测量偏倚与效应恢复。*Biometrika*，101(2):423– 437，2014年。

Kusner, M. J., Loftus, J., Russell, C. 和 Silva, R. 反事实公平性。神经信息处理系统进展，30，2017年。
Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L. 和 Ranzato, M. Fader网络：通过滑动属性操控图像。神经信息处理系统进展，30，2017年。

Lee, S. 和 Bareinboim, E. 使用矩阵方程进行因果识别。神经信息处理系统进展，34:9468–9479，2021年。Li, A. 和 Pearl, J. 因果效应界限及其在高维数据中的应用。载于*AAAI*人工智能会议论文集，第36卷，第5773–5780页，2022a年。Li, A. 和 Pearl, J. 因果概率：观测数据的作用。*arXiv*预印本 *arXiv:2210.08874*，2022b年。

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R. 和 Carin, L. Alice：迈向理解用于联合分布匹配的对抗学习。神经信息处理系统进展，30，2017年。

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Sch¨olkopf, B. 和 Bachem, O. 挑战无监督学习解缠表示中的常见假设。载于国际机器学习会议，第 4114–4124 页。PMLR，2019年。

Lorberbom, G., Johnson, D. D., Maddison, C. J., Tarlow, D. 和 Hazan, T. 学习广义Gumbel-max因果机制。神经信息处理系统进展, 34:26792–26803, 2021年。

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. 和 Welling, M. 基于深度潜变量模型的因果效应推断。神经信息处理系统进展, 30, 2017年。

Lu, C., Huang, B., Wang, K., Hern´andez-Lobato, J. M., Zhang, K. 和 Sch¨olkopf, B. 通过基于反事实的数据增强实现样本高效的强化学习。*arXiv*预印本 *arXiv:2012.09092*, 2020年。

Lu, Y. 和 Huang, B. 通过条件生成流进行结构化输出学习。载于人工智能 *AAAI* 会议论文集，第34卷，第5005–5012页，2020年。Manski, C. F. 治疗效应的非参数界限。《美国经济评论》, 80(2):319–323, 1990年。

Mao, H.、Netravali, R. 和 Alizadeh, M. 使用 Pensieve 实现神经自适应视频流。载于 *ACM* 数据通信特别兴趣小组会议论文集，第 197–210 页，2017年。

Meng, C., Zhou, L., Choi, K., Dao, T., andErmon, S. 蝴蝶流：使用蝴蝶矩阵构建可逆层。在国际机器学习会议，第 15360–15375 页。PMLR，2022年。

Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T. S., Heess, N., Guez, A., 等。无模型强化学习中的反事实信用分配。在国际机器学习会议，第7654–7664 页。PMLR，2021年。

Miao, W.、Geng, Z. 和 Tchetgen Tchetgen, E. J. 使用未测量混杂因素的代理变量识别因果效应。*Biometrika*，105:987–993，2018年。

Mohamed, S. 和 Lakshminarayanan, B. 隐式生成模型中的学习。*arXiv* 预印本 *arXiv:1610.03483*，2016年。

Monteiro, M., Ribeiro, F. D. S., Pawlowski, N., Castro, D. C., and Glocker, B. 衡量反事实图像模型的公理健全性。发表于第十一届国际表征学习会议，2023年。Nasr-Esfahany, A. 和 Kiciman, E. 学习得到的结构因果模型的反事实（不可）识别性，2023年。网址：https://arxiv.org/abs/2301.09031。Oberst, M. 和 Sontag, D. 使用Gumbel-max结构因果模型进行反事实离策略评估。发表于国际机器学习会议，第4881–4890页。PMLR，2019年。Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S. 和 Lakshminarayanan, B. 用于概率建模与推断的归一化流。机器学习研究期刊，22(57):1–64，2021年。Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. 等。PyTorch：一种命令式风格、高性能的深度学习库。神经信息处理系统进展，32，2019年。Pawlowski, N., Coelho de Castro, D. 和 Glocker, B. 用于可处理反事实推断的深度结构因果模型。神经信息处理系统进展，33:857–869，2020年。Pearl, J.智能系统中的概率推理：合理推断网络。Morgan Kaufmann，1988年。Pearl, J. 面向实证研究的因果图。*Biometrika*，82(4):669–688，1995a。Pearl, J. 关于含潜在变量和工具变量的因果模型的可检验性。发表于第十一届人工智能不确定性会议论文集，第435–443页，1995b。Pearl, J. 因果概率：三种反事实解释及其识别。*Synthese*，121(1):93–149，1999年。

Pearl, J. 因果关系。剑桥大学出版社，2009a。

Pearl, J. 统计学中的因果推断：概述。统计学综述，3:96–146，2009b。

Pearl, J. 和 Mackenzie, D. 《为什么之书：因果关系的新科学》。基础图书出版社，2018年。

Perkovi´c, E.、Textor, J. 和 Kalisch, M. 祖先图的马尔可夫等价类中调整集的完整图模型刻画与构造。《机器学习研究杂志》，19:1–62，2018年。

Peters, J., Mooij, J. M., Janzing, D. 和 Sch¨olkopf, B. 连续加性噪声模型的因果发现。《机器学习研究杂志》，15：2009–2053，2014年。Peters, J., Janzing, D. 和 Sch¨olkopf, B. 《因果推断要素：基础与学习算法》。麻省理工学院出版社，2017年。Richens, J. G., Beard, R. 和 Thompson, D. H. 反事实伤害，2022年。网址：https://arxiv.org/abs/2204.12993。Rubin, D. B. 因果效应的贝叶斯推断：随机化的作用。《统计年鉴》，第34–58页，1978年。Sanchez, P. 和 Tsaftaris, S. A. 用于反事实估计的扩散因果模型。载于第一届因果学习与推理会议，2021年。Sanchez-Martin, P., Rateike, M. 和 Valera, I. Vaca：用于干预层和反事实查询的变分图自编码器设计。*arXiv*预印本 *arXiv:2110.14690*，2021年。Shah, A., Dwivedi, R., Shah, D. 和 Wornell, G. 关于未观测混杂下的反事实推断。载于*NeurIPS 2022* 年现实世界影响因果关系研讨会，2022年。Shaham, U., Svirsky, J., Katz, O. 和 Talmon, R. 单一独立潜变量的发现。载于神经信息处理系统进展，2022年。Shalit, U., Johansson, F. D. 和 Sontag, D. 个体治疗效应估计：泛化界与算法。载于国际机器学习会议，第3076–3085页。PMLR，2017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，第352–359页，2007年。

Shpitser, I. 和 Pearl, J. 对已接受治疗者的治疗效应：识别与泛化。载于第二十五届人工智能不确定性会议论文集，第514–521页，2009年。

Shpitser, I. 和 Sherman, E. 识别与因果路径相关的个性化效应。载于人工智能不确定性：……会议论文集。人工智能不确定性会议，第 2018 卷。美国国立卫生研究院公共访问，2018年。

Song, Y. 和 Ermon, S. 通过估计数据分布的梯度进行生成建模。神经信息处理系统进展，32，2019年。

Spirtes, P.、Glymour, C.、Scheines, R. 等。因果关系、预测与搜索。麻省理工学院出版社图书，1，2001年。

Spiteri, K.、Urgaonkar, R. 和 Sitaraman, R. K. Bola：在线视频的近最优码率自适应。*IEEE/ACM*网络汇刊，28(4)：1698–1711，2020年。

Strobl, E. V. 和 Lasko, T. A. 使用异方差噪声模型识别患者特异性根本原因。*arXiv* 预印本 *arXiv:2205.13085*，2022年。

Sun, Y.、Yin, X.、Jiang, J.、Sekar, V.、Lin, F.、Wang, N.、Liu, T. 和 Sinopoli, B. Cs2p：利用数据驱动的吞吐量预测改进视频码率选择与自适应。在*2016*年*ACM SIGCOMM*会议论文集中，第272–285页，2016年。

Tian, G. 和 Liu, Y. 面向动态HTTP流媒体中灵活流畅的视频自适应。在第八届新兴网络实验与技术国际会议论文集中，第109–120页，2012年。

Tian, J. 和 Pearl, J. 因果概率：界与识别。数学与人工智能年刊，28(1)：287–313，2000年。

Trippe, B. L. 和 Turner, R. E. 使用贝叶斯归一化流进行条件密度估计。*arXiv*预印本 *arXiv:1802.04908*，2018年。

Wang, Y. 和 Blei, D. M. 多重原因的益处。美国统计学会杂志，114(528)：1574–1596，2019年。

Winkler, C.、Worrall, D.、Hoogeboom, E. 和 Welling, M. 使用条件归一化流学习似然。*arXiv* 预印本 *arXiv:1912.00042*，2019年。

Witty, S., Jensen, D. 和 Mansinghka, V. Sbi：用于贝叶斯因果推断的基于模拟的可识别性检验，2021年。网址https://arxiv.org/abs/2102.11761。

Wu, Z.、Pan, S.、Chen, F.、Long, G.、Zhang, C. 和 Philip, S. Y. 图神经网络综合综述。*IEEE* 神经网络与学习系统汇刊，32(1)：4–24，2020年。Xia, K.、Lee, K.-Z.、Bengio, Y. 和 Bareinboim, E. 因果-神经网络联系：表达能力、可学习性与推断。神经信息处理系统进展，34：10823–10836，2021年。Xia, K.、Pan, Y. 和 Bareinboim, E. 用于反事实识别与估计的神经因果模型。*arXiv* 预印本 *arXiv:2210.00035*，2022年。

Yan, F. Y.、Ayers, H.、Zhu, C.、Fouladi, S.、Hong, J.、Zhang, K.、Levis, P. 和 Winstein, K. 原位学习：一项关于视频流的随机实验。载于第 *17* 届 *USENIX* 网络系统设计与实现研讨会（*NSDI 20*），第 495–511 页，2020年。

Yao, L.、Li, S.、Li, Y.、Huai, M.、Gao, J. 和 Zhang, A. 基于观测数据的治疗效应估计表征学习。神经信息处理系统进展，31，2018年。

Yin, X., Jindal, A., Sekar, V., 和 Sinopoli, B. 一种基于控制理论的 HTTP 动态自适应视频流方法。载于*2015*年*ACM*数据通信特别兴趣小组会议论文集，第325–338 页，2015年。

Ze˘cevi´c, M., Dhami, D. S., Veli˘ckovi´c, P., 和 Kersting, K. 图神经网络与结构因果模型的关联。*arXiv*预印本 *arXiv:2109.04173*，2021年。

Zhang, J. 关于潜在混杂因素和选择偏差存在时因果发现定向规则的完备性。人工智能，172(16-17)：1873–1896，2008年。

Zhang, J. 和 Bareinboim, E. 决策中的公平性——因果解释公式。载于*AAAI*人工智能会议论文集，第 32 卷，2018年。

Zhang, J. 和 Bareinboim, E. 连续结果上的因果效应界定。载于人工智能促进协会年会论文集，第 35 卷，第 12207–12215 页，2021年。

Zhang, J.、Tian, J. 和 Bareinboim, E. 基于观测数据和实验数据的部分反事实识别。载于国际机器学习会议，第 26548–26558 页。PMLR，2022年。

Zhang, K. 和 Hyv¨arinen, A. 关于后非线性因果模型的可识别性。载于第*25*届人工智能不确定性会议（*UAI 2009*），第647–655页。AUAI出版社，2009年。

Zhu, J.-Y.、Park, T.、Isola, P. 和 Efros, A. A. 使用循环一致对抗网络进行非配对图像到图像转换。载于*IEEE*国际计算机视觉会议论文集，第2223–2232页，2017年。

A. 因果估计的神经方法

Xia 等人 (2021年) 使用基于观测数据训练的神经因果模型（NCM），在已知底层因果图（可能是非马尔可夫的）及离散内生变量的前提下，对干预层 ($\mathcal{L}_2$) 查询进行识别和估计。

Kocaoglu 等人 (2018年) 和 Ze˘cevi´c 等人 (2021年) 分别采用对抗训练和图神经网络（GNN）（Wu 等人，2020年），在已知因果图并满足马尔可夫假设的前提下，从观测数据中学习隐式结构因果模型。随后利用学习到的结构因果模型回答干预查询 ($\mathcal{L}_2$)；已知在马尔可夫性假设下，这些查询是可识别的（Bareinboim 等人，2022年，推论 2）。

Pawlowski 等人 (2020年) 和 Sanchez-Martin 等人 (2021年) 使用各种形式的深度条件生成模型，并依据已知因果图构建模型，从观测数据中学习结构因果模型，假设不存在未观测混杂（马尔可夫结构因果模型）。学习到的结构因果模型还用于干预和反事实估计。 Sanchez 与 Tsaftaris(2021年) 使用扩散去噪概率模型，学习给定带标签属性时图像的条件分布，并进一步用于反事实图像生成。然而，他们没有进行任何可识别性分析。事实上，Nasr-Esfahany 与 Kiciman (2023年) 指出，在马尔可夫设定下，无法根据观测数据识别多维变量的生成机制。为评估马尔可夫结构因果模型中不可识别的图像反事实的质量，Monteiro 等人 (2023年) 通过衡量反事实的复合性、可逆性和有效性，重新审视了反事实的公理化定义。

Xia 等人(2022年) 使用生成对抗网络（GAN）（Goodfellow 等人，2014年）从观测数据和干预数据中学习代理结构因果模型，假设已知（非马尔可夫）因果图，且内生变量为离散型。他们利用代理结构因果模型进行反事实($\mathcal{L}_3$) 识别与估计。Gumbel-Max结构因果模型（Oberst 与 Sontag，2019年）及其推广形式（Lorberbom 等人，2021年）已用于根据观测数据估计分类变量的反事实。不过，我们主要关注连续域。

假设结构因果模型具有加性噪声（ANM）（Hoyer 等人，2008年）且满足马尔可夫性，Geffner 等人(2022年) 从观测数据中学习潜在因果结构 $\mathcal{G}$ 和结构因果模型，并用它们估计干预查询和反事实查询。不过，他们的可识别性分析仅限于干预查询。非线性加性噪声模型是本文所考虑的结构因果模型类的一个特例。此外，我们允许存在未观测混杂因素，并证明了我们模型的反事实可识别性。Hartford 等人(2017年) 使用两阶段监督学习方法，基于工具变量（Angrist 等人，1996）估计反事实查询，并假设结构因果模型具有加性噪声。Alomar 等人 (2023年) 在假设生成机制为低秩的前提下，证明了可以从随机对照试验（RCT）数据中识别反事实查询；随机对照试验是工具变量情形的一种特例。此外，他们采用配备对抗学习的确定性自编码器（Ghosh 等人，2019年），通过分布匹配实现高效反事实估计。我们的方法不需要对生成机制的秩作任何假设。

Khemakhem 等人 (2021年) 使用仿射因果自回归流（Dinh 等人，2016年）学习潜在因果结构 $\mathcal{G}$，并假设不存在未观测混杂因素。他们在假设噪声服从位置尺度噪声模型（LSNM）（Strobl 与 Lasko，2022年）的条件下，证明了干预查询的可识别性。此外，他们提出使用学得的结构因果模型进行反事实估计，但未对其可识别性进行分析。我们允许存在未观测混杂因素，不将生成函数限制为 LSNM，并且也证明了反事实可识别性。

Louizos 等人（2017）使用变分自编码器（VAE）（Kingma 和 Welling，2013）来估计反事实查询，适用于存在充足的未观测混杂因素代理变量（Carroll 等人，2006年；Kuroki 和 Pearl，2014年；Miao 等人，2018年；Wang 和 Blei，2019年；Lee 和 Bareinboim，2021年），从而实现可识别性的情况。Johansson 等人（2016）；Shalit 等人（2017）；Yao 等人（2018）在假设强可忽略性和二元处理的前提下，学习用于估计个体治疗效应（ITE）的表示。我们的工作不局限于离散域，也不局限于 ITE 等特定反事实量。

B. 证明

B.1. 命题 6.2

$$\text{双射生成模型 } f_1,\ f_2 \text{ 产生相同的反事实} \iff f_1,\ f_2 \text{ 是等价贝叶斯图模型} \tag{8}$$

证明。首先，我们证明 $\Leftarrow$。考虑任意反事实查询 $V_{\boldsymbol{x}'}|\boldsymbol{X}=\boldsymbol{x},\ V=v$。在溯因步骤中，两个贝叶斯图模型都使用证据 $\boldsymbol{X}=\boldsymbol{x},\ V=v$来推断外生变量 U。我们将 f_1 推断出的外生变量称为u_1，将 f_2 推断出的外生变量称为 u_2。利用方程（6），可得

$$u_1 = g(u_2). \tag{9}$$

在预测步骤中，f_1和 f_2分别给出 $f_1(\boldsymbol{x}',\ u_1) = f_1\Big(\boldsymbol{x}', g(u_2)\Big)$ 以及 $f_2(\boldsymbol{x}',\ u_2)$作为各自的反事实估计值。利用方程（7）可得

$$f_1\Big(\boldsymbol{x}', g(u_2)\Big) = f_2(\boldsymbol{x}', u_2). \tag{10}$$

因此，估计得到的反事实相等。

接下来，我们证明 $\Rightarrow$。由于 f_1 和 f_2 都是贝叶斯图模型，我们可以很容易地验证它们之间满足以下关系：

$$\forall \boldsymbol{x}, u_1 : f_1(\boldsymbol{x}, u_1) = f_2\Big(\boldsymbol{x}, g^{-1}(\boldsymbol{x}, u_1)\Big) \tag{11}$$

其中

$$\forall \boldsymbol{x}, u_1 : g^{-1}(\boldsymbol{x}, u_1) = f_2^{-1}\Big(\boldsymbol{x}, \big(f_1(\boldsymbol{x}, u_1)\big)\Big). \tag{12}$$

现在，假设我们同时使用 f_1 和 f_2 来估计反事实查询 $V_{\boldsymbol{x}'}|\boldsymbol{X}=\boldsymbol{x}, V=v$。$f_1$的估计值为$f_1(\boldsymbol{x}', u_1)$。利用公式（11），该估计值等于 $f_2\Big(\boldsymbol{x}', g^{-1}(\boldsymbol{x}', u_1)\Big)$。使用 f_2 估计同一查询时，它在溯因步骤中推断出的外生变量为 $g^{-1}(\boldsymbol{x}, u_1)$。在预测步骤中，其反事实估计值为$f_2\Big(\boldsymbol{x}', g^{-1}(\boldsymbol{x}, u_1)\Big)$。由于这两个反事实估计值相等，因此有

$$\forall \boldsymbol{x}', \boldsymbol{x}, u_1 : f_2\Big(\boldsymbol{x}', g^{-1}(\boldsymbol{x}', u_1)\Big) = f_2\Big(\boldsymbol{x}', g^{-1}(\boldsymbol{x}, u_1)\Big). \tag{13}$$

利用双射生成模型的可逆性 f_2，我们得到

$$\forall \boldsymbol{x}', \boldsymbol{x}, u_1 : g^{-1}(\boldsymbol{x}', u_1) = g^{-1}(\boldsymbol{x}, u_1) \rightarrow \forall \boldsymbol{x}', \boldsymbol{x}, u_1 : g\Big(\boldsymbol{x}', g^{-1}(\boldsymbol{x}, u_1)\Big) = u_1. \tag{14}$$

最后一个　　个等式要对所有可能的 $\boldsymbol{x}, \boldsymbol{x}'$ 成立，唯一的可能性是 g^{-1} 不依赖于其第一个自变量，即

$$g^{-1}(\boldsymbol{x}, u_1) = g^{-1}(u_1), \tag{15}$$

这意味着 f_1 和 f_2 等价。 □

注 B.1. 这种不确定性 $g(\cdot)$ 的部分原因是，外生变量 $\Big(P(U)\Big)$ 的先验分布未知。对该先验分布的每种选择都会导致不同的 $g(\cdot)$。

B.2. 反事实等价引理

在本节中，我们给出三个引理，它们对于证明§5中的反事实可识别性结果至关重要。

B.2.1. 马尔可夫情形

引理B.2. 产生相同分布 $P_{\mathcal{D}}(\boldsymbol{X}, V)$ 的贝叶斯图模型 f和 f等价，当且仅当

1. （马尔可夫）$U \perp\!\!\!\perp \boldsymbol{X}$和 $U \perp\!\!\!\perp \boldsymbol{X}$。
2. 对所有 $\boldsymbol{x}, f(\boldsymbol{x}, \cdot)$和 $f(\boldsymbol{x}, \cdot)$，它们都是严格递增函数或严格递减函数。

证明。 我们将证明，贝叶斯图模型 f 和 f 会产生相同的反事实。根据命题 6.2，可得二者等价。假设我们关注反事实查询 $V_{\boldsymbol{x}'}|\boldsymbol{X}=\boldsymbol{x},\ V=v$。不失一般性，仅考虑满足 $P_{\boldsymbol{X},\ V}(\boldsymbol{x},\ v) > 0$的 $\boldsymbol{x},\ v$ 个样本。令 $F(\boldsymbol{x},\ v) := P(V \le v|\boldsymbol{X}=\boldsymbol{x})$表示条件累积分布函数 (CDF)，$F^{-1}(\boldsymbol{x},\ \alpha)$ 表示分位数函数；当 $P_{\boldsymbol{X},\ V}(\boldsymbol{x},\ v) > 0$时，该分位数函数存在。在溯因

步骤中，在反事实估计过程中，若二者分别都单调递增或单调递减，则贝叶斯图模型 f和 $\hat{f}$会分别返回其对应外生分布 $\Big(P(U|\boldsymbol{X})$ 和 $P(\hat{U}|\boldsymbol{X})$ 的 $F(\boldsymbol{x},\ v)^{\text{th}}$ 或 $\Big(1-F(\boldsymbol{x},\ v)^{\text{th}}\Big)$ 分位数。$\hat{}$ 这些分位数实际上可能是两个不同的值。由于 U，$U \perp\!\!\!\perp \boldsymbol{X}$，给定 $\boldsymbol{X}$ 时，U，U 的分位数函数独立于 $\hat{\boldsymbol{X}}$，且等于边缘 U，U 的分位数函数。因此，行动步骤不会改变估计的分位数。在预测步骤中，如果递增，两个贝叶斯图模型 f和 $\hat{f}$都会估计出相同的值 $F^{-1}\Big(\boldsymbol{x}', F(\boldsymbol{x},\ v)\Big)$；如果递减，则会得到其他类似的分位数变化。

□

B.2.2. 工具变量（IV）

引理 B.3. 对于 $\boldsymbol{X} \in \mathbb{X} \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ 和 $\boldsymbol{I} \in \mathbb{I} \triangleq \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n\}$，产生相同分布$P_{\mathcal{D}}(\boldsymbol{X}, V)$ 的贝叶斯图模型 f和 $\hat{f}$等价，当且仅当

1. （工具变量） $\boldsymbol{I} \perp\!\!\!\perp U$ 和 $\boldsymbol{I} \perp\!\!\!\perp U$。
2. 对所有 $\boldsymbol{x} \in \mathbb{X}$, $f^{-1}(\boldsymbol{x}, \cdot)$ 以及 $\hat{f}(\boldsymbol{x}, \cdot)$，它们要么严格递增，要么严格递减，且二阶可微。
3. $P_{\hat{U}}(\cdot)$ 可微。
4. $P_{\mathcal{D}}(\boldsymbol{i}, \boldsymbol{x}, \cdot)$ 对每个 $\boldsymbol{i}, \boldsymbol{x}$ 都可微。
5. *(Positivity)* $\forall u, \hat{u}, \boldsymbol{x} \in \mathbb{X} : P_{U,\boldsymbol{X}}(u, \boldsymbol{x}) > 0$ *and* $P_{\hat{U},\boldsymbol{X}}(\hat{u}, \boldsymbol{x}) > 0$.
6. （变异性） $\forall u$： $|\det \boldsymbol{M}_{\mathcal{D}}(u, \mathbb{I})| \geq c$，其中 c 是一个正数，并且

$$\boldsymbol{M}_{\mathcal{D}}(u, \mathbb{I}) \triangleq \begin{bmatrix} P_{\mathcal{D}}(\boldsymbol{x}_1|u, \boldsymbol{i}_1) & \ldots & P_{\mathcal{D}}(\boldsymbol{x}_n|u, \boldsymbol{i}_1) \\ \vdots & \ddots & \vdots \\ P_{\mathcal{D}}(\boldsymbol{x}_1|u, \boldsymbol{i}_n) & \ldots & P_{\mathcal{D}}(\boldsymbol{x}_n|u, \boldsymbol{i}_n) \end{bmatrix} \tag{16}$$

$\hat{}$ 证明。定义累积分布函数（CDF） $k(u) = P(U \leq u)$，$k(\hat{u}) = P(U \leq \hat{u})$。我们使用累积分布函数将随机变量变换为 0 到 1 之间的均匀分布。定义 $Z = k(U)$，$Z = k(U)$。根据概率积分变换，我们有 Z, $Z \sim Unif(0, 1)$。此外，$\forall \boldsymbol{x} \in \mathbb{X}$： $Z = s(\boldsymbol{x}, Z)$，其中 $s(\boldsymbol{x}, \cdot) = k(\cdot) \circ f^{-1}(\boldsymbol{x}, \cdot) \circ \hat{f}(\boldsymbol{x}, \cdot) \circ \hat{k}^{-1}(\cdot)$。因为 $U \perp\!\!\!\perp \boldsymbol{I}$（第一个条件），并且因为 Z 是 U 的确定性函数，我们得出结论 $Z \perp\!\!\!\perp \boldsymbol{I}$。运用类似的论证， $Z \perp\!\!\!\perp \boldsymbol{I}$。

$$\forall \hat{z}, \boldsymbol{i} : P_{\hat{Z}|\boldsymbol{I}}(\hat{z}|\boldsymbol{i}) = 1 \tag{17}$$

$$\rightarrow \forall \hat{z}, \boldsymbol{i} : \sum_{\ell=1}^{n} P_{\hat{Z},\boldsymbol{X}|\boldsymbol{I}}(\hat{z}, \boldsymbol{x}_\ell|\boldsymbol{i}) = 1 \tag{18}$$

$$\rightarrow \forall \hat{z}, \boldsymbol{i} : \sum_{\ell=1}^{n} P_{\hat{Z}|\boldsymbol{X},\boldsymbol{I}}(\hat{z}|\boldsymbol{x}_\ell, \boldsymbol{i}) P_{\boldsymbol{X}|\boldsymbol{I}}(\boldsymbol{x}_\ell|\boldsymbol{i}) = 1 \tag{19}$$

根据条件 2、3、4、6，我们知道 $\forall \boldsymbol{x} \in \mathbb{X}$： $s(\boldsymbol{x},\ \cdot)$ 严格递增且二阶可微。因此，使用变量替换公式可得：

$$\rightarrow \forall \hat{z}, \boldsymbol{i} : \sum_{\ell=1}^{n} P_{Z|\boldsymbol{X},\boldsymbol{I}}(s(\boldsymbol{x}_\ell, \hat{z})|\boldsymbol{x}_\ell, \boldsymbol{i}) P_{\boldsymbol{X}|\boldsymbol{I}}(\boldsymbol{x}_\ell|\boldsymbol{i}) \frac{\partial s(\boldsymbol{x}_\ell, \hat{z})}{\partial \hat{z}} = 1 \tag{20}$$

使用 $Z \perp\!\!\!\perp \boldsymbol{I}$，我们有：

$$P_{Z|\boldsymbol{X},\boldsymbol{I}}(z|\boldsymbol{x}, \boldsymbol{i}) P_{\boldsymbol{X}|\boldsymbol{I}}(\boldsymbol{x}|\boldsymbol{i}) = P_Z(z) P_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}|z, \boldsymbol{i}) \tag{21}$$

$Z \sim Unif(0,\ 1)$ 因此

$$\rightarrow P_{Z|\boldsymbol{X},\boldsymbol{I}}(z|\boldsymbol{x}, \boldsymbol{i}) P_{\boldsymbol{X}|\boldsymbol{I}}(\boldsymbol{x}|\boldsymbol{i}) = P_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}|z, \boldsymbol{i}) \tag{22}$$

现在我们将方程（20）与方程（22）结合起来：

$$\rightarrow \forall \hat{z}, \boldsymbol{i} : \sum_{\ell=1}^{n} P_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_\ell | s(\boldsymbol{x}_\ell, \hat{z}), \boldsymbol{i}) \frac{\partial s(\boldsymbol{x}_\ell, \hat{z})}{\partial \hat{z}} = 1 \tag{23}$$

我们可以将方程（23）写成向量乘积形式：

$$\forall \hat{z}, \boldsymbol{i} : \begin{bmatrix} P_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_1 | s(\boldsymbol{x}_1, \hat{z}), \boldsymbol{i}) & \dots & P_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_n | s(\boldsymbol{x}_n, \hat{z}), \boldsymbol{i}) & -1 \end{bmatrix} \begin{bmatrix} \frac{\partial s(\boldsymbol{x}_1, \hat{z})}{\partial \hat{z}} \\ \vdots \\ \frac{\partial s(\boldsymbol{x}_n, \hat{z})}{\partial \hat{z}} \\ 1 \end{bmatrix} = 0 \tag{24}$$

现在，我们将所有 n 个此形式的方程组合成矩阵形式：

$$\forall \hat{z} : \begin{bmatrix} p_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_1 | s(\boldsymbol{x}_1, \hat{z}), \boldsymbol{i}_1) & \dots & p_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_n | s(\boldsymbol{x}_n, \hat{z}), \boldsymbol{i}_1) & -1 \\ \vdots & \vdots & \vdots & \vdots \\ p_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_1 | s(\boldsymbol{x}_1, \hat{z}), \boldsymbol{i}_n) & \dots & p_{\boldsymbol{X}|Z,\boldsymbol{I}}(\boldsymbol{x}_n | s(\boldsymbol{x}_n, \hat{z}), \boldsymbol{i}_n) & -1 \end{bmatrix} \begin{bmatrix} \frac{\partial s(\boldsymbol{x}_1, \hat{z})}{\partial \hat{z}} \\ \vdots \\ \frac{\partial s(\boldsymbol{x}_n, \hat{z})}{\partial \hat{z}} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{25}$$

如上文所述， $\forall \boldsymbol{x} \in \mathbb{X}$： $s(\boldsymbol{x}, \cdot)$是严格递增且二次可微的。结合正性假设，我们得到

$$\forall \boldsymbol{x} \in \mathbb{X} : s(\boldsymbol{x}, 0) = 0 \tag{26}$$

考虑25式在$\hat{z} = 0$中的形式。第一个矩阵的秩为 n，这是由变异性条件保证的，并且有 $n+1$ 列。因此，其零空间具有$k \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^\intercal$的形式。这意味着

$$\forall \boldsymbol{x} \in \mathbb{X} : \frac{\partial s(\boldsymbol{x}, \hat{z})}{\hat{z}}|_{\hat{z}=0} = 1. \tag{27}$$

接下来，我们将 $[0, 1]$ 划分为 N 个部分。我们通过对 m 进行归纳法证明，对于足够大的 N

$$\forall m \in \{1, \dots, N\}, \boldsymbol{x} \in \mathbb{X} : |s(\boldsymbol{x}, \frac{m}{N}) - \frac{m}{N}| \leq \frac{m}{2N^2} B \tag{28}$$

其中 $B = \max_{\boldsymbol{x},\ \hat{z}} \frac{\partial^2 s}{(\hat{z},\ \boldsymbol{x})\ \partial \hat{z}}$。

对于 $m = 1$，根据泰勒定理，我们有

$$\exists \xi \in [0, \frac{1}{N}] : s(\boldsymbol{x}, \frac{1}{N}) = s(\boldsymbol{x}, 0) + \frac{1}{N} \frac{\partial s(\boldsymbol{x}, \hat{z})}{\partial \hat{z}}|_{\hat{z}=0} + \frac{1}{2N^2} \frac{\partial^2 s(\boldsymbol{x}, \hat{z})}{\partial \hat{z}}|_{\hat{z}=\xi} \tag{29}$$

利用方程（26）和（27）

$$\rightarrow \exists \xi \in [0, \frac{1}{N}] : s(\frac{1}{N}, \boldsymbol{x}) = \frac{1}{N} + \frac{1}{2N^2} \frac{\partial^2 s(\boldsymbol{x}, \hat{z})}{\partial \hat{z}}|_{\hat{z}=\xi} \tag{30}$$

$$\rightarrow |s(\boldsymbol{x}, \frac{1}{N}) - \frac{1}{N}| \leq \frac{B}{2N^2} \tag{31}$$

现在假设方程 (28) 对 m成立，我们将证明它对 $m+1$ 也成立。我们知道 $s(\boldsymbol{x}, \cdot)$ 是可微的。结合假设 6 可知，矩阵 M 的每个元素都关于 $\hat{z}$可微。此外，行列式是关于 M 所有元素的多项式函数，并且对每个元素都可微。因此， M 的行列式关于 $\hat{z}$可微。这意味着，如果我们将 $\hat{z}$ 的每个元素扰动 M 一个足够小的量， M 将保持满秩。因此，对于足够大的 N，方程 (25) 左侧矩阵的秩仍为 n，其零空间仍为一维。这意味着

$$\forall \boldsymbol{x} \in \mathbb{X} : \frac{\partial s(\boldsymbol{x}, \hat{z})}{\hat{z}}|_{\hat{z}=\frac{m}{N}} = 1. \tag{32}$$

根据泰勒定理，我们有

$$\exists \xi \in [\frac{m}{N}, \frac{m+1}{N}] : s(\boldsymbol{x}, \frac{m+1}{N}) = s(\boldsymbol{x}, \frac{m}{N}) + \frac{1}{N}\frac{\partial s(\boldsymbol{x}, \hat{z})}{\partial \hat{z}}|_{\hat{z}=\frac{m}{N}} + \frac{1}{2N^2}\frac{\partial^2 s(\boldsymbol{x}, \hat{z})}{\partial \hat{z}}|_{\hat{z}=\xi} \tag{33}$$

使用方程（28）对于 m 以及方程（32），我们得出结论：

$$|s(\boldsymbol{x}, \frac{m+1}{N}) - \frac{m+1}{N}| \leq \frac{m+1}{2N^2}B \tag{34}$$

由归纳法可知，对于所有 $m \in \{1, \cdots, N\}$ 的取值，方程（28）均成立，这就完成了该方程的证明。换言之，函数 $s(\boldsymbol{x}, \cdot)$ 在所有点 $\frac{m}{N}$ 上都可以任意逼近恒等函数。结合这一点以及 $\forall \boldsymbol{x} \in \mathbb{X}:\ s(\boldsymbol{x}, \cdot)$ 可微这一事实，可推出 $\forall \boldsymbol{x} \in \mathbb{X}:\ s(\boldsymbol{x}, \hat{z}) = \hat{z}$。因此，$g(\cdot) = k(\cdot) \circ \hat{k}^{-1}(\cdot)$，证明得证。 □

B.2.3. 后门准则（BC）

引理 B.4. 贝叶斯图模型 f 和 f 若产生相同的分布 $P_{\mathcal{D}}(\boldsymbol{X}, V)$，则它们等价，当且仅当

1. （后门准则）$U \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{Z}$ 和 $\hat{U} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{Z}$。
2. 对每个 $\boldsymbol{x}:\ \nabla_{\boldsymbol{x}}|$ 行列式 $\boldsymbol{J}_{f^{-1}(\boldsymbol{x},\cdot)}|$ 和 $\nabla_{\boldsymbol{x}}|$ 行列式 $\boldsymbol{J}_{\hat{f}(\boldsymbol{x},\cdot)}|$ 均存在。
3. （变异性）$\forall u:$ 实例 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}$ 存在，使得 $|$行列式 $\boldsymbol{M}_{\mathcal{D}}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1})| > 0$，其中

$$\boldsymbol{M}_{\mathcal{D}}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}) \triangleq \begin{bmatrix} P_{\mathcal{D}}(u|\boldsymbol{z}_1) & \nabla_u P_{\mathcal{D}}(u|\boldsymbol{z}_1) \\ \vdots & \vdots \\ P_{\mathcal{D}}(u|\boldsymbol{z}_{d+1}) & \nabla_u P_{\mathcal{D}}(u|\boldsymbol{z}_{d+1}) \end{bmatrix} \tag{35}$$

证明。定义 $g(\boldsymbol{x}, \cdot) \triangleq f^{-1}(\boldsymbol{x}, \cdot) \circ \hat{f}(\boldsymbol{x}, \cdot)$。利用变量替换公式，我们得到

$$\forall \boldsymbol{x}, \hat{u}, \boldsymbol{z} : P_{\hat{U}|\boldsymbol{Z},\boldsymbol{X}}(\hat{u}|\boldsymbol{z}, \boldsymbol{x}) = P_{U|\boldsymbol{Z},\boldsymbol{X}}\Big(g(\boldsymbol{x}, \hat{u})|\boldsymbol{z}, \boldsymbol{x}\Big)|\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}| \tag{36}$$

$$\text{(BC)} \Rightarrow \forall \boldsymbol{x}, \hat{u}, \boldsymbol{z} : P_{\hat{U}|\boldsymbol{Z}(\hat{u}|\boldsymbol{z})} = P_{U|\boldsymbol{Z}}\Big(g(\boldsymbol{x}, \hat{u})|\boldsymbol{z}\Big)|\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}| \tag{37}$$

根据导数链式法则，我们知道 $|$行列式 $\boldsymbol{J}_{g(\boldsymbol{x},\cdot)}| = |$行列式 $\boldsymbol{J}_{f^{-1}(\boldsymbol{x},\cdot)}||$行列式 $\boldsymbol{J}_{\hat{f}(\boldsymbol{x},\cdot)}|$，它关于 $\boldsymbol{x}$ 可微（根据条件 2）。对方程（37）关于 i^{th} 中属于 $\boldsymbol{x}$ (x_i) 的元素求导，我们得到

$$\nabla_u P_{U|\boldsymbol{Z}}\Big(g(\boldsymbol{x}, \cdot)|\boldsymbol{z}\Big) \begin{bmatrix} \frac{g_1(\boldsymbol{x},\hat{u})}{\partial x_i} \\ \vdots \\ \frac{g_d(\boldsymbol{x},\hat{u})}{\partial x_i} \end{bmatrix} |\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}| + P_{U|\boldsymbol{Z}}\Big(g(\boldsymbol{x}, \hat{u})|\boldsymbol{z}\Big)\frac{\partial |\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}|}{\partial x_i} = 0 \tag{38}$$

$$\Rightarrow \begin{bmatrix} \nabla_u P_{U|\boldsymbol{Z}}\Big(g(\boldsymbol{x}, \cdot)|\boldsymbol{z}\Big) & P_{U|\boldsymbol{Z}}\Big(g(\boldsymbol{x}, \hat{u})|\boldsymbol{z}\Big) \end{bmatrix} \begin{bmatrix} \frac{g_1(\boldsymbol{x},\hat{u})}{\partial x_i} \\ \vdots \\ \frac{g_d(\boldsymbol{x},\hat{u})}{\partial x_i} \\ \frac{\partial |\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}|}{\partial x_i} \end{bmatrix} = 0 \tag{39}$$

将关于 $\boldsymbol{z}_1$、.....、$\boldsymbol{z}_{d+1}$ 的方程堆叠起来，我们得到

$$\boldsymbol{M}_{\mathcal{D}}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}) \begin{bmatrix} \frac{\partial |\det \boldsymbol{J}_{g(\boldsymbol{x},\cdot)}|}{\partial x_i} \\ \frac{g_1(\boldsymbol{x},\hat{u})}{\partial x_i} \\ \vdots \\ \frac{g_d(\boldsymbol{x},\hat{u})}{\partial x_i} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{40}$$

由于变异性条件，方阵是满秩的，因此向量的所有元素都必须为零。这意味着$g(\boldsymbol{x}, \hat{u})$不依赖于 x_i。对所有 $i \in \{1, \ldots, d\}$ 重复同样的论证，我们得出 $g(\boldsymbol{x}, \hat{u})$不依赖于 $\boldsymbol{x}$，由此证明得证。

□

B.3. 定理 5.1

若满足以下条件，则给定$P_{\boldsymbol{X},\ V}$时，双射生成模型 f 是反事实可识别的：

1. （马尔可夫） $U \perp\!\!\!\perp \boldsymbol{X}$。
2. 对所有 $\boldsymbol{x}$、 $f(\boldsymbol{x},\ \cdot)$都是严格递增函数或严格递减函数。

ˆ证明。 设 $\mathbb{F}$ 为满足定理条件的贝叶斯图模型类。考虑任意两个贝叶斯图模型 f、$f \in \mathbb{F}$，它们产生相同的分布 $P_{\mathcal{D}}(\boldsymbol{X},\ V)$。利用引理 B.2，我们可以得出它们等价。因此，它们产生相同的反事实（命题 6.2），根据第 §2节中对可识别性的定义，这就确立了可识别性。 □

B.3.1. 独立性　　仅靠独立性假设不足以实现反事实可识别性

在本节中，我们通过一个简单的例子说明，仅有独立性假设（不包含单调性假设）不足以识别双射生成模型。此例取自Nasr-Esfahany 与 Kiciman (2023年，第 3 节）ˆ考虑以下两个简单的贝叶斯图模型 f 和 f:

$$X \sim Bernoulli(0.5),\ U \sim Unif(0,1),\ X \perp\!\!\!\perp U,\ f = \begin{cases} U, & X = 1 \\ U-1, & X = 0 \end{cases},\ \hat{f} = \begin{cases} U, & X = 1 \\ -U, & X = 0 \end{cases} \tag{41}$$

ˆ注意， f和 f生成相同的分布 $P_{\mathcal{D}}(\boldsymbol{X},\ V)$，并且满足第一个（独立性）约束。然而，它们对反事实查询给出的答案不同。考虑以下反事实查询： $V_1|X=0,\ V=v$。 fˆ和 f给出的答案分别是 $v+1$ 和 $-v$。

B.4. 定理5.2

对于 $\boldsymbol{X} \in \mathbb{X} \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$和 $\boldsymbol{I} \in \mathbb{I} \triangleq \{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n\}$，在给定$P_{\boldsymbol{X},\ V,\ \boldsymbol{I}}$的条件下，双射生成模型 f是反事实可识别的，当且仅当

1.（工具变量）
$\boldsymbol{I} \perp\!\!\!\perp U$。

2. 对所有 $\boldsymbol{x} \in \mathbb{X}$, $f(\boldsymbol{x},\ \cdot)$ 和 $f^{-1}(\boldsymbol{x},\ \cdot)$，它们要么严格递增，要么严格递减，且二阶可微。
3. 对每个 $\boldsymbol{i} \in \mathbb{I}$， $\boldsymbol{x} \in \mathbb{X}$， $P(\boldsymbol{i},\ \boldsymbol{x},\ \cdot)$ 都是可微的。
4.（正性） $\forall u,\ \boldsymbol{x} \in \mathbb{X}$： $P_{U,\ \boldsymbol{X}}(u,\ \boldsymbol{x}) > 0$。
5.（变异性） $\forall u$： $|\det \boldsymbol{M}(u,\ \mathbb{I})| \geq c > 0$，其中

$$\boldsymbol{M}(u, \mathbb{I}) \triangleq \begin{bmatrix} P(\boldsymbol{x}_1|u,\boldsymbol{i}_1) & \ldots & P(\boldsymbol{x}_n|u,\boldsymbol{i}_1) \\ \vdots & \ddots & \vdots \\ P(\boldsymbol{x}_1|u,\boldsymbol{i}_n) & \ldots & P(\boldsymbol{x}_n|u,\boldsymbol{i}_n) \end{bmatrix}$$

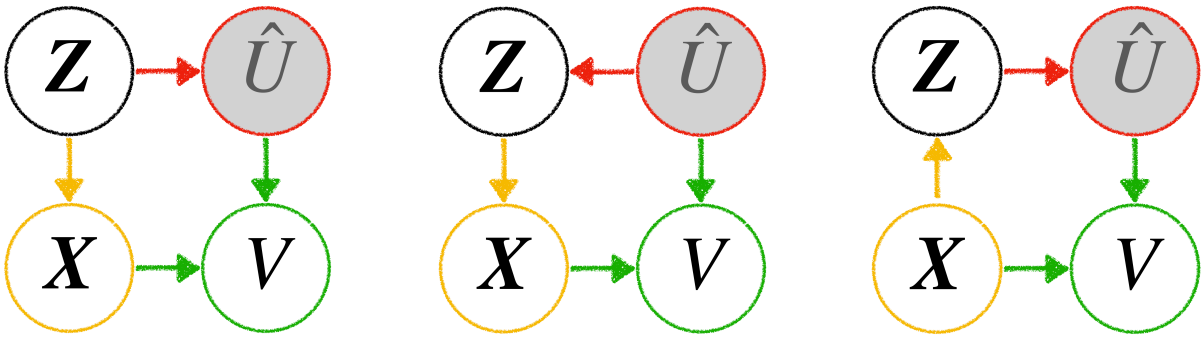
ˆ证明。 设 $\mathbb{F}$ 为满足定理条件的双射生成模型类。考虑任意两个满足定理条件的贝叶斯图模型 f， $f \in \mathbb{F}$，它们产生相同的分布 $P_{\mathcal{D}}(\boldsymbol{X},\ V)$。利用引理 B.3，我们可得它们等价。因此，它们产生相同的反事实（命题 6.2），这根据 §2 中的定义确立了可识别性。 □

B.5. 定理5.3

双射生成模型 f 在给定$P_{\boldsymbol{X},V,\boldsymbol{Z}}$ 时反事实可识别，当且仅当

1.（后门准则） $U \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{Z}$。

2. $\forall \boldsymbol{x}$： $\nabla_{\boldsymbol{x}}|\det \boldsymbol{J}_{f(\boldsymbol{x},\cdot)}|$和 $\nabla_{\boldsymbol{x}}|\det \boldsymbol{J}_{f^{-1}(\boldsymbol{x},\cdot)}|$存在。



(a) (b) (c)

图5。 后门准则（BC）情形下有向图模型的马尔可夫等价类

3.（变异性） $\forall u$： 存在实例 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}$， 使得 $|\det \boldsymbol{M}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1})| > 0$, 其中

$$\boldsymbol{M}(u, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}) \triangleq \begin{bmatrix} P(u|\boldsymbol{z}_1) & \nabla_u P(u|\boldsymbol{z}_1) \\ \vdots & \vdots \\ P(u|\boldsymbol{z}_{d+1}) & \nabla_u P(u|\boldsymbol{z}_{d+1}) \end{bmatrix}$$

ˆ证明。 令 $\mathbb{F}$ 为满足定理条件的双射生成模型（BGM）类。考虑任意两个双射生成模型（BGM） f， $f \in \mathbb{F}$ 它们产生相同的分布 $P_{\mathcal{D}}(\boldsymbol{X}, V)$。利用引理 B.4 可得它们等价。因此，它们产生相同的反事实（命题 6.2），根据 §2 中的定义，这就确立了可识别性。 □

C. 有向图模型

我们可以使用 d-分离，从联合分布的有向图模型中确定（条件）独立性。

C.1. 后门准则（BC）

在本例中，我们所要求的分布条件（§5.3）是 $\boldsymbol{X} \perp\!\!\!\perp U|\boldsymbol{Z}$。我们为本例构建的结构化生成网络（图 3d）与图 3b 中所示有向图模型的结构相似。因此，它继承了根据有向图模型通过 d-分离检验读出的所有条件独立性（Pearl，1988年；Koller 与 Friedman，2009年）。在 U 与 $\boldsymbol{X}$之间有两条路径： $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow U$ 在给定 $\boldsymbol{Z}$的条件下被阻断；而$\boldsymbol{X} \rightarrow V \leftarrow U$ 由于 V处的 v-结构而被阻断，因此 $\boldsymbol{X} \perp\!\!\!\perp U|\boldsymbol{Z}$。

其他可行的图模型：值得强调的是，用于构建结构化生成网络的有向图模型只是一个统计对象，而非因果有向无环图；其唯一目的是为结构化生成网络在各种情况下赋予相应的分布约束。因此，图3b 所属马尔可夫等价类中的所有有向图模型（即编码相同条件独立性的有向图模型）都是有效的，都可用于构建其他结构化生成网络。

要构造马尔可夫等价类，我们应固定图的骨架，并翻转边，同时不创建新的 v-结构，也不移除已有的 v-结构。图 5 展示了三种可能的选项，每种都可用于构造有效的结构化生成网络。

C.2. 工具变量（IV）

此处的结构化生成网络如图 3c所示。该网络旨在遵循图 3a 中的有向图模型，其中 $\boldsymbol{I}$ 与 U、 $\boldsymbol{I} \rightarrow \boldsymbol{X} \leftarrow U$ 与 $\boldsymbol{I} \rightarrow \boldsymbol{X} \rightarrow V \leftarrow U$ 之间的两条路径被开放的 v-结构所阻断（分别为未条件化的 X 和 $\boldsymbol{V}$），这意味着 $\boldsymbol{I}$ 与 V 相互独立。因此，结构化生成网络所产生的分布必然满足分布约束 $U \perp\!\!\!\perp V$（定理 5.2中的第一个条件），这是由其构造方式所保证的。

在这种情况下，图形　　图中所示的模型3a是其马尔可夫等价类中唯一的成员，因为翻转　　ng

任意边都会改变（条件）独立性的集合。

D. 归一化流（NF）

归一化流（NF）是一类生成模型，其分布具有可处理性，能够高效、精确地进行采样和密度估计。它们将数据（V建模为某个噪声变量（T）的变换（(U），该噪声变量从简单的基分布（P_U）中采样得到，例如高斯分布，其中 T 是一个微分同胚。[9] 这使得我们能够通过变量替换得到 V 的密度：

$$P_V(v) = P_U\Big(T^{-1}(v)\Big)|\det \boldsymbol{J}_{T^{-1}}(v)| \tag{42}$$

变换 T 可以通过可处理的优化来拟合 V的观测分布。因此，设计具有高效逆函数和雅可比行列式的表达性变换族一直是研究课题（Kingma & Dhariwal，2018年；Chen 等，2019年；Meng 等，2022年）。这一思想可以轻松扩展到条件分布建模（Trippe & Turner，2018年；Winkler 等，2019年；Lu & Huang，2020年），例如，$P_{V|\boldsymbol{X}}$ 使用条件归一化流（CNF），将变换 T参数化为条件（$\boldsymbol{x}$）的函数。关于归一化流的详细综述，请参阅 Kobyzev 等人（2020年）；Papamakarios 等（2021）。

E. 实验

实现细节：我们使用基于线性有理样条的归一化流（Dolatabadi 等人，2020年）构建所有条件生成模型，并使用 Pyro 中的实现通过似然最大化对其进行训练（Bingham 等人，2018年）。我们使用的所有样条均有 16 个区间，用于将 $(-3,\ +3)$ 映射到 $(-3,\ +3)$。我们在输入层和输出层使用仿射变换来校准范围。所有条件网络均为包含两个隐藏层的多层感知机，每层有 64 个单元。我们使用的批量大小为 2^{20}，所有实验均使用 A100 图形处理器运行。我们使用 PyTorch 中 Adam 的默认实现训练所有模型（Kingma 和 Ba，2015年）。

理论假设的经验性放宽：线性有理样条虽然可微，但不一定二阶可微。在工具变量情形（§5.2）中，根据定理 5.2 的第二个假设，需要二阶可微性。此外，我们使用的条件网络根据 $\boldsymbol{x}$ 来确定样条参数，并采用 ReLU 激活函数（Agarap，2018年），这意味着我们的条件生成模型关于 x不可微；而根据定理 5.3 的第二个假设，在工具变量情形（§5.3）中需要满足可微性。例如，使用二次样条（Durkan 等人，2019年）或 GELU 激活函数（Hendrycks 和 Gimpel，2016年）即可同时满足这两个假设。然而，第 8 节中的良好表现表明，这些技术性假设可能并非严格必要，可以予以放宽。我们将其放宽留待未来工作。

E.1. 反事实椭圆生成

我们使用以下结构因果模型生成椭圆数据集：

$$Z \coloneqq \epsilon_z, \quad \epsilon_z \sim \text{Unif}(-0.5,\ 0.5) \tag{43}$$

$$X \coloneqq (1.44254843z + 0.59701923 + \epsilon_x) \ \%\ (2\pi),\ \epsilon_x \sim \text{Normal}(0, 1) \tag{44}$$

$$U_0 \coloneqq e^{1.64985274z + 0.2656131} + \epsilon_{u_0},\ \epsilon_{u_0} \sim \text{Beta}(1, 1) \tag{45}$$

$$U_1 \coloneqq U_0(1 + \epsilon_{u_1} e^{1.61323358z - 0.18070237}),\ \epsilon_{u_1} \sim \text{Exponential}(1) \tag{46}$$

$$V_0 \coloneqq U_0\Big(2 + \sin(X)\Big) \tag{47}$$

$$V_1 \coloneqq U_1\Big(2 + \cos(X)\Big) \tag{48}$$

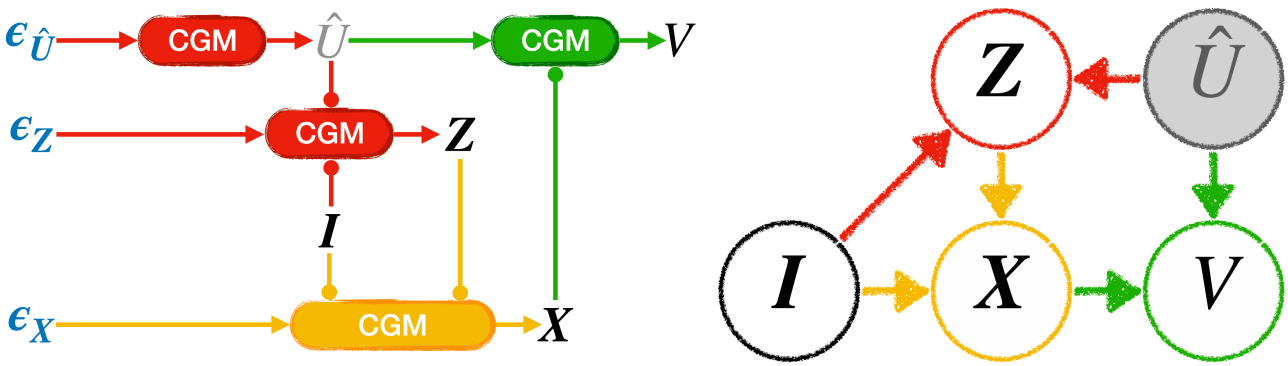
$$\tag{49}$$

所有方案均使用一系列三个带耦合的样条变换。

E.1.1. 马尔可夫情形下的失败

在椭圆生成任务中，U 和 V 都是二维的。为了通过实证评估我们能否在马尔可夫情形下学习多维贝叶斯图模型，我们通过随机打乱先前的 X 生成了第二个数据集

[9] 具有可微逆变换的可微变换。



（a）结构化生成网络　　（b）有向图模型

图 *6.* 同时利用工具变量（IV）和后门准则（BC）

数据集。我们训练了一个单一的条件生成模型（类似于 §6中的马尔可夫情形），并将其用于反事实估计，但结果失败了（平均绝对百分比误差 $= 607$）。

E.2. 视频流模拟

我们从Alomar 等人 （2023年）处获取了模拟器和 ABR 算法的实现。本研究的附录 D 说明了详细信息。我们为每个 CGM 使用一个条件样条。

E.2.1. 同时利用工具变量（IV）和后门准则（BC）

图 6b 展示了我们用于表示此情形所需（条件）独立性的有向图模型。ˆ　　ˆ在此有向图模型中，$\boldsymbol{I}$ 与 U 之间的所有路径均被阻断，因此 $\boldsymbol{I} \perp\!\!\!\perp U$，这就是工具变量情形（§5.2）中的分布目标。此外，以$(\boldsymbol{Z}, \boldsymbol{I})$ 为条件会阻断 $\boldsymbol{X}$ 与 U 之间的所有路径，因此 $\boldsymbol{X} \perp\!\!\!\perp U|(\boldsymbol{I}, \boldsymbol{Z})$。这是 BC 情形（§5.3）中的分布目标，其中变量对$(\boldsymbol{I}, \boldsymbol{Z})$ 满足后门准则。